Evaluating Transcriptional Regulation Through Multiple Lenses

by

Ivan Grubisic

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Joint Doctor of Philosopy
with University of California , San Francisco

in

Bioengineering

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Robert Tjian, Co-chair
Professor Daniel Fletcher, Co-chair
Associate Professor Tamara Alliston
Professor Michael Eisen

Spring 2014

Evaluating Transcriptional Regulation Through Multiple Lenses

Abstract

Evaluating Transcriptional Regulation Through Multiple Lenses

by

Ivan Grubisic

Joint Doctor of Philosopy
with University of California , San Francisco in Bioengineering

University of California, Berkeley

Professor Robert Tjian, Co-chair

Professor Daniel Fletcher, Co-chair

Scientific research, especially within the space of translational research is becoming increasingly multidisciplinary. With the development of each new method there is not only a need for a broad fundamental understanding of all the sciences and mathematics, but also an acute awareness of how errors propagate across methods, the limitation of the methods and what contextual frameworks need to be used for the interpretation. The ability to understand transcriptional mechanisms and the affect that subtle changes in equilibrium may have on cell fate decisions has been greatly advanced by next generation sequencing and subsequent tools that have been developed. Bioinformatic techniques can serve multiple roles. They fundamentally provide a global picture of what is happening within an experimental condition which can then be used to either confirm individual experimental findings as globally relevant, or to discover new insights to inform the next iteration of experiments. Many of the experiments are done in *in vitro* conditions and therefore I have also focused energy on trying to understand how the mechanical inputs, largely not representative of what is occurring *in vivo*, from these methods affect transcriptional regulation. Much of this research requires the switching of frameworks to understand how results from disparate data sources can be correlated. I then applied a similar thought process to the development of *Lens*. Without an effective means of communicating research findings in an elegant and streamline mannered, we are slowing down the ability for researchers to learn new frameworks to efficiently approach the next research questions. In addition to better methods of communicating, we also need more modular and simplified tools that can be applied to various experimental systems to increase the speed and efficiency of translational research.

To Family, Friends and Colleagues

The PhD has been a random walk, in part mediated by frustration, that has led me to identify new opportunities, interesting questions and unique solutions. Thank you for providing the support and guidance during this period of intellectual growth.

Contents

List of Figures

Acknowledgments

I would like to acknowledge the UC Berkeley - UCSF Graduate Group in Bioengineering for accepting me into their PhD program. Thanks to Professor Robert Tjian for providing me with an opportunity and funding in his lab to pursue my research interests. The freedom to pursue a variety of projects was challenging, but a great experience. I would also like to thank all of the members of Professor Robert Tjian's lab, all of my committee members and colleagues for consistently challenging me intellectually and forcing me to raise the level of my research methods, hypotheses and contextual analyses.

Credit also needs to be given to Professor Randy Schekman for the success of Lens. He put me in touch with Ian Mulvany, the Head of Technology at *eLife*, to continue pitching, prototyping and developing my vision for Lens. Without his willingness to help at the beginning, none of the subsequent work would have been possible.

Chapter 1

Introduction

With the technological advances in sequencing, advances in imaging and new applications of microfluidics, many are hopeful that this will spur on the next round of both scientific and medical innovations. The ability to study pluripotent embryonic stem cells (ES cells) and to induce them from various differentiated cell types (iPS cells),[1] has the potential to drastically improve the efficiency of regenerative medicine. It is a fantastic improvement that scientists are now able to evaluate the global transcriptional profile of stem cells during a differentiation program. Multiplexing and improved efficiency in high-throughput sequencing has made it both cheaper and easier to work with the same amount of material that would have been used for the targeted RT-qPCR based assay.[2,3,4] Through the use of super resolution microscopy, one can study the translational dynamics of proteins, active transcription of specific genes amongst many other things.[5] When materials are at a premium, some of the classical biochemical assays can even be run using microfluidics to optimize concentrations and therefore improve the efficiency and speed of methods like the western blot.[6] We have all of these tools at our disposal, but at this point, the likelihood that a single lab could perform all of these methods at a high level is very low. Realistically the future of science will be even more collaborative as one's area expertise gets even more specific; however, for these collaborations to work, the researchers need to have more than a superficial understanding of all emerging and established fields.

Most of my focus over the last years has been placed on trying to combine some of these tools in leveraging high-throughput sequencing to support mechanistic biochemical findings in transcription regulation of ES cells (Figure 1.1). The nature of many biochemical assays requires a targeted approach that focuses on a couple of well described genes and transcription factors that are specific to the system. There is, however, no guarantee that a functional interaction at a single locus, either transcription factor to DNA or a larger coactivator complex, will be universally true across the genome. High-throughput sequencing has made it significantly easier to assess these questions. By generating millions of sequences, these sequences can be aligned to a reference genome.[7] The entire genome is then analyzed for regions of significant enrichment.[8,9] These findings are then annotated against known transcripts to define *cis*-regulatory elements or or other elements.[10] A caveat with high-throughput sequencing, that is no different from any other PCR assay, is that an enrichment, be it a binding site in ChIP-Seq or a transcript enrichment in RNA-Seq is not necessarily functional. A binding site defined by ChIP-Seq may not have a direct role in activating transcription and a change in transcript levels in a knockout experiment does not mean that it is a primary affect. It is important to note that even though there is a breadth of new information, without combining these disparate data sources, it is very difficult to make strong mechanistic claims. It is therefore that majority of the work has gone towards using ChIP-Seq and RNA-Seq analysis to confirm the mechanistic findings in well established *in vitro* systems. We identified SCC-B, a novel coactivator complex involved in maintaining stem cell pluripotency, that was then confirmed to have a specific interaction with Oct4/Sox2 globally.[11] Continuing with the focus on stem cells, we shifted our attention to the adipogenic differentiation pathway. We found that TAF7L, a subunit of the general transcription factor core machinery TFIID, is required for adipogenic differentiation. This work hinged not only on validating the biochemical findings, but also utilizing computational techniques to inform the subsequent experiments that identified a looping interaction between TAF7L and PPARγ.[12] We then also evaluated TAF7L in mature tissue to determine if its non-canonical function was limited to only plurioptent cells, or if it interacted with additional transcription factors outside of TFIID. We found that it also had strong interactions with TRF2 regulated genes.[13] These are prime examples of how a computational biology approach can synergistically add value to the biochemical findings.

Although there is plenty of data coming through high-throughput sequencing, it is by no means guaranteed to provide meaningful results. Working within various cell types, different antibodies and potentially difficult experimental conditions, there were plenty of instances when the signal to noise was not desirable. Much of this came from either an inefficient immunoprecipitation by the antibody and/or a poor alignment.[7] The poor alignment can be due to either sequences that do not align to the genome, or are suppressed because they do not align uniquely. The multiply-aligned sequences were particularly interesting because they usually represent between 10 and 20 percent of the data which could significantly help improve the signal. The decrease in signal is not only a result of tandem repeats in the genome, but also a result of allowing for mismatches during the alignment step across highly conserved, but not identical, regions of the genome. Finding a way to efficiently utilize these sequences could also potentially uncover genomic interactions that were previously going missed, and therefore not providing a complete picture of the biological context. Historically the main tools that have addressed the problem of multiply-aligned sequences has been in RNA-Seq where batch Expected Maximization (EM) algorithms have been used to calculate the estimated likelihood of a given alignment originating from a location in the genome.[15,16,17,18] These algorithms tend to be computationally very intensive though. A new streaming online EM algorithm, eXpress, provided a manageable solution. eXpress is computationally less intensive, takes into account mismatches amongst other sequence

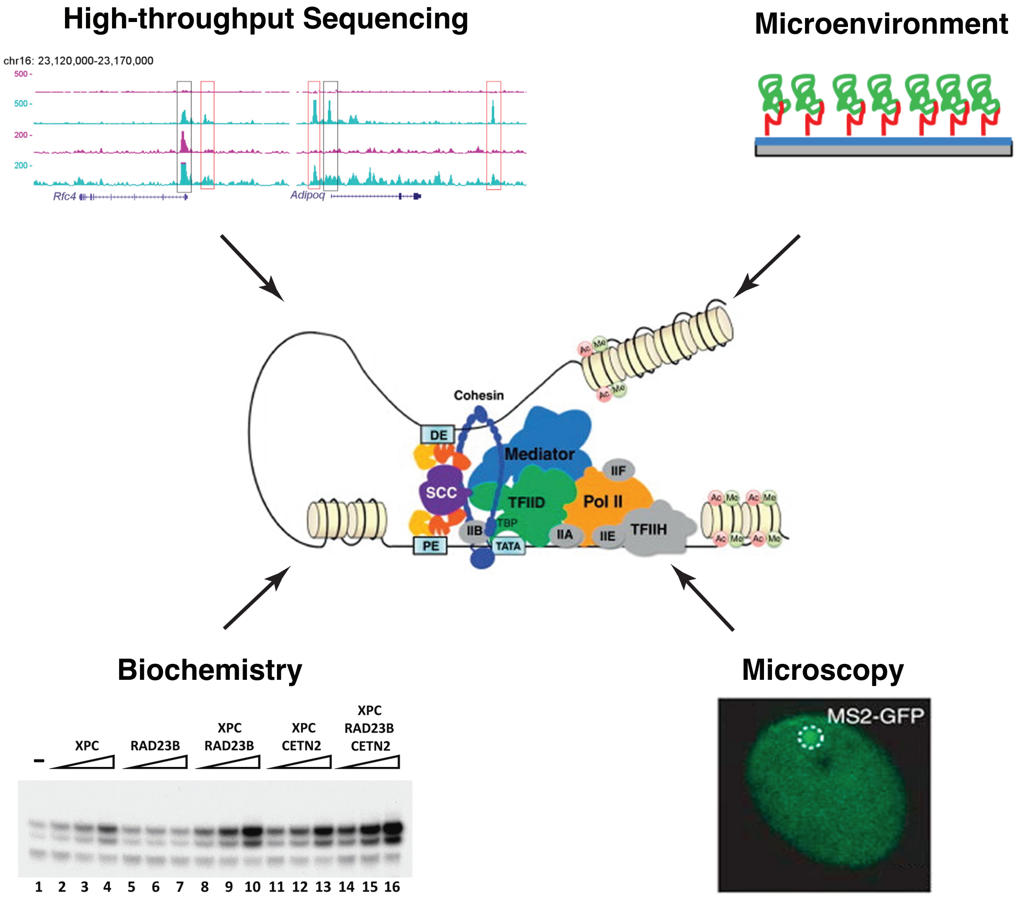


Figure 1.1: **Studying transcriptional regulation through various lenses.** There are many different ways to approach transcriptional regulation. High-throughput sequencing and microscopy are able to provide different global trends for individual transcription factors at a time.[12,5] Manipulating the microenvironment by seeding cells on different combinations of substrate stiffness and extracellular matrix can also induce different transcriptional programs that can be then further evaluated using both new methods and classical biochemistry assays.[11,14]

related parameters to increase both its sensitivity and accuracy.[17] After successfully bootstrapping eXpress to handle any genome-wide high-throughput sequencing method, regions that were either missed or underrepresented by the classical unique alignment methods can now be studied in the more complete context.

Maintaining an interest in regenerative medicine though, it continued to be unsettling that the *in vitro* experimental conditions did a very poor job of mimicking the conditions *in vivo*. There was also a growing amount of literature that was describing the manner in which the extracellular matrix (ECM) transduces both mechanical and physical cues that affect cell fate and differentiation decisions.[14,19,20] Much of the literature was focused on understanding how the changes in the microenvironment's mechanical properties would affect integrins or cytoskeletal integrity.[21,22] Ultimately the evaluation of these changes was measured through RNA or protein levels. This lead me to postulate that there must be transcriptional coactivators or families of transcription factors that would be sensitive to changes in the microenvironment. The aim was to combine a pseudo biomimetic (2D polyacrylamide gel) technique with bioinformatics and super resolution microscopy to evaluate the transcriptional dynamics in real time as a function of the cellular microenvironment. Unfortunately, scale up to produce enough gels to do ChIP-Seq or RNA-Seq was too variable to proceed with the project to its completion.

Balancing each of these research projects also required a significant time investment to stay on top of the current scientific literature. As one has to switch contextual frameworks when analyzing results for an individual experiment, the same occurs when reading research articles. Given the volume of articles, many of them are stored and read on laptops or other digital devices, making it difficult to focus on the author's argument when doing an in-depth read. More energy is expended trying to find the pertinent figures or textual context than on the author's narrative. This frustration lead me to approach *eLife* to develop a piece of open source software, *Lens*, that helps solve some of the problems with reading articles on digital screens. The reason for approaching *eLife*, an open access publisher, and publishing the code base behind *Lens* under a FreeBSD open source license, is based on the fundamental premise that scientific research and the tools that are developed around it should be readily available and free to everyone. The free flow of information, critiques, methods and innovative frameworks for approaching critical problems in both basic and translational research need to be uninhibited if the tools and discoveries will be as impactful as the authors had initially hoped.

Chapter 2

Feedback between computational and biochemical methods to evaluate transcriptional mechanisms

2.1 Motivation

The primary focus of the Tjian Lab recently has been to study transcriptional regulation in the context of pluripotent stem cells and differentiation pathways. Through very rigorous biochemical characterizations, they have been able to elucidate the transcriptional mechanisms of many transcription factors. As it became clear that the transcription factor of interest has a functional role in the experimental system, I began to collaborate with the post doctorate scientists in the lab to help inform their next round of wet lab experiments. The goal was to use high-throughput sequencing to see whether the targeted interactions held true globally. In addition to validation, the computational analysis had the ability to inform the next experiments by uncovering interactions and correlations that would have been missed in the targeted schema. I was responsible for processing the data, analyzing the results using various computational tools and then working with the post docs to contextualize the data and inform the next round of experiments.

2.2 DNA Repair complex coactivates Oct4/Sox2 to maintain ESC pluripotency

Introduction

The molecular events leading to the maintenance of pluripotency in embryonic stem (ES) cells and re-acquisition of a stem-like state in induced pluripotent stem (iPS) cells during somatic reprogramming represent mechanistically distinct processes that however converge on a set of remarkably similar transcriptional events that underpin the pluripotent state. Both ES and iPS cells depend on fundamental transcription frameworks that are governed by a common set of "core" stem cell-specific transcription factors, namely Oct4, Sox2 and Nanog.[23] These activators in turn collaborate with both ubiquitous and cell type-specific transcription factors to orchestrate complex gene expression programs that confer upon stem cells the unique ability to safeguard stemness while remaining poised to execute a broad range of developmental programs that drive lineage specification.[24,25,26,27]

Proper execution of these highly regulated processes by sequence-specific transcription factors often requires the coordinated recruitment of coactivator proteins to their cognate promoters. For example, transcriptional activators direct histone modifiers (e.g., CBP/p300) and chromatin remodelers (e.g., PBAF/BAF) to gene promoters to alter chromatin structure toward a state that is more permissive to transcriptional activation.[28] Independent of chromatin, a variety of activators recruit other classes of coactivators, such as the multi-subunit Mediator, various TBP/TAF complexes, SRC, etc, via direct protein-protein interactions to execute specific transcriptional programs. This class of coactivators often serve as molecular "adaptors" by bridging activators to the general transcription machinery thereby mediating the synergistic response by these activators (Naar et al., 1999). Interestingly, subunits of Mediator have also been shown to interact with cohesin possibly to promote DNA looping and thereby facilitate long distance interactions between enhancers and core promoters in vivo.[29] Indeed, such coactivators are often multifunctional and can activate transcription through chromatin-dependent as well as independent mechanisms. Further expanding the transcriptional repertoire of coactivator complexes, their protein levels and subunit compositions are frequently modulated in a developmental stage and cell type-specific manner.[30,31] Additionally, these protein-protein driven coactivator-activator transactions are often critical nodes in various signal transduction pathways and can serve as molecular "sensors" by integrating cell intrinsic and extrinsic cues thereby coupling gene networks with specific cellular responses to produce complex biological programs of gene expression.[32]

Totipotent ES cells employ these same sets of coactivators in conjunction with special activators such as Oct4 and Sox2 to regulate transcription of a large number of genes including Nanog that form the molecular basis of pluripotency.[33,29,34,35] The transcription of Nanog is exquisitely dependent on Oct4 and Sox2.[36,37] However, co-expression of Oct4 and Sox2 failed to robustly activate a Nanog promoter reporter construct in differentiated cells like 293 or NIH3T3 cells, even though Mediator, p300/CBP and PBAF/BAF complexes remain abundantly expressed and active.[37] This led us to speculate that one or more as yet unidentified stem cell-specific cofactor may be required to activate the transcription of Nanog and other Oct4/Sox2-target genes in ES cells. Indeed, recent studies of germ cells and differentiated somatic cells revealed that even parts of the general transcriptional machinery may be radically altered in a tissue or cell-specific context.[38,39] Diversification of the transcriptional apparatus may therefore represent a fundamental strategy, particularly in ES cells, to cope with the multi-dimensional nature of transcription programs that must be precisely tuned to both maintain pluripotency and at the same time allow for lineage-specific programs of differentiation.[40]

The human Nanog promoter contains a prototypic composite oct-sox cis-acting regulatory element located immediately upstream of the transcription start site that is conserved across several mammalian species.[36,37] A Nanog promoter-GFP reporter construct containing a DNA fragment encompassing this promoter-proximal oct-sox element is sufficient to recapitulate the robust expression pattern of endogenous Nanog in ES cells in an Oct4-, Sox2-dependent manner.[36,37] Unbiased genome-wide motif searching analyses of Oct4 in both mouse and human ES cells identified an oct-sox composite consensus sequence element, confirming that Oct4 likely orchestrates an ES-specific gene expression program primarily through cooperation with Sox2.[25,41] Since the oct-sox cis-control element in the Nanog promoter represents a common configuration that is present in the promoters of many other Oct4 and Sox2-activated genes in ES cells, the well-characterized Nanog proximal promoter provided us with a useful model template for identifying potentially novel transcriptional cofactors required for Oct4 and Sox2-directed activation. Therefore, we took advantage of a fully reconstituted in vitro transcription system where one can unambiguously and systematically test and identify transcriptional cofactors that may be directly required to potentiate Oct4- and Sox2-dependent gene activation of Nanog. Here we report the biochemical purification and identification of a multi-subunit stem cell coactivator (SCC) that is required for the synergistic activation of Nanog by Oct4 and Sox2 in vitro. After extensive biochemical characterization, we surprisingly found that SCC is none other than the XPC-RAD23B-CETN2 (XPC) nucleotide excision repair (NER) complex. SCC/XPC interacts directly with Oct4 and Sox2 and co-occupies a majority of Oct4 and Sox2 targets genome-wide in mouse ES cells. Importantly, SCC/XPC is required for stem cell self-renewal and efficient somatic cell reprogramming. Thus, our findings unmask an unanticipated selective coactivator role of an NER complex in transcription in the context of ES cells and may provide a previously unknown molecular link that couples stem cell-specific transcription to DNA damage response with potential implications for enhanced ES cell genome stability.

Results

Detection of an Oct4- and Sox2-dependent Coactivator Activity in EC and ES Cells

Having chosen the Nanog promoter as our model template, we next set out to develop an in vitro reconstituted transcription assay that could recapitulate the Oct4- and Sox2-dependent trans-activation at the Nanog promoter observed in vivo. To enhance the sensitivity of the assay, we inserted four copies of the Nanog oct-sox binding sites immediately upstream of the native oct-sox element found in the human Nanog promoter. Our basal in vitro transcription assay consisted of purified recombinant TFIIA, -B, -E and -F together with immuno-affinity purified native RNA polymerase II, TFIID and TFIIH (Figure 2.2A). When purified Oct4 and Sox2 were added to this reconstituted transcription system, only a very weak activation of the Nanog promoter was detected (Figure 2.1A, lanes 1 and 2). As a control, we could show that the same complement of general transcription factors (GTFs) was able to support strong Sp1-dependent activation from a GC box-containing "generic" transcription template (G3BCAT, Figure 2.1A, lanes 5 and 6). This initial result suggested that efficient activation of Nanog by Oct4 and Sox2 may require additional cofactors to potentiate a full activator-dependent response.

We reasoned that such a putative coactivator ought to be selectively active in pluripotent cell types that express Nanog under the control of Oct4 and Sox2. For example, NTERA-2 (NT2) is a pluripotent human embryonal carcinoma (EC) cell line that expresses Oct4, Sox2, Nanog and shares with ES cells core molecular mechanisms governing self-renewal.[42] Detailed expression profiling of NT2 and bona fide human ES cell lines revealed many similarities, including robust expression of Nanog.[43,44] However, unlike human ES cells, NT2 cell culture can be more readily scaled up, a prerequisite to generating sufficient quantities of starting materials for the biochemical purification of putative Oct4/Sox2 coactivators. We therefore chose extracts derived from NT2 cells as our starting material in our efforts to develop a "biochemical complementation" assay to hunt for pluripotent stem cell selective cofactors.

We first fractionated NT2 nuclear extracts by conventional phosphocellulose ion exchange chromatography. Next, we supplemented our "basal" reconstituted transcription reactions with various salt-eluted fractions from the phosphocellulose column to see if there was any activity that could restore Oct4/Sox2-dependent activation of our Nanog promoter. This strategy allowed us to unmask an activity in the high salt phosphocellulose fraction (P1M) prepared from NT2 nuclear extracts (but not Hela extracts, Figure 2.2B) that strongly potentiated transcription of the Nanog promoter in an Oct4- and Sox2-dependent manner using either a naked (Figure 2.1A, lanes 3 and 4) or a Nanog chromatin template assembled with a crude Drosophila cytosolic extract (data not shown). This new cofactor activity is selectively required for transcription of Nanog as it had no effect on either basal or Sp1-activated transcription from a control G3BCAT template (Figure 2.1A, lanes 5-8). Importantly, this P1M fraction also stimulated the Oct4/Sox2-dependent transcription from a native Nanog promoter template (Figure 2.1B), as well as two other Oct4/Sox2-dependent templates derived from the mouse Fbxo15 promoter[45](mFbxo15CAT, Figure 2.2C, lanes 1-4) and the human HESX1 promoter[46] (HESX1CAT, Figure 2.2C, lanes 5-8). Thus, our in vitro complementation assay programmed with naked DNA templates revealed at least one potential coactivator activity that directs Oct4/Sox2-dependent activation of Nanog. We decided to pursue characterization of this cofactor that does not appear to require chromatin-based functions. To the best of our knowledge, this finding also demonstrates for the first time a fully reconstituted, in vitro transcription system that can faithfully recapitulate stem cell-specific gene activation.

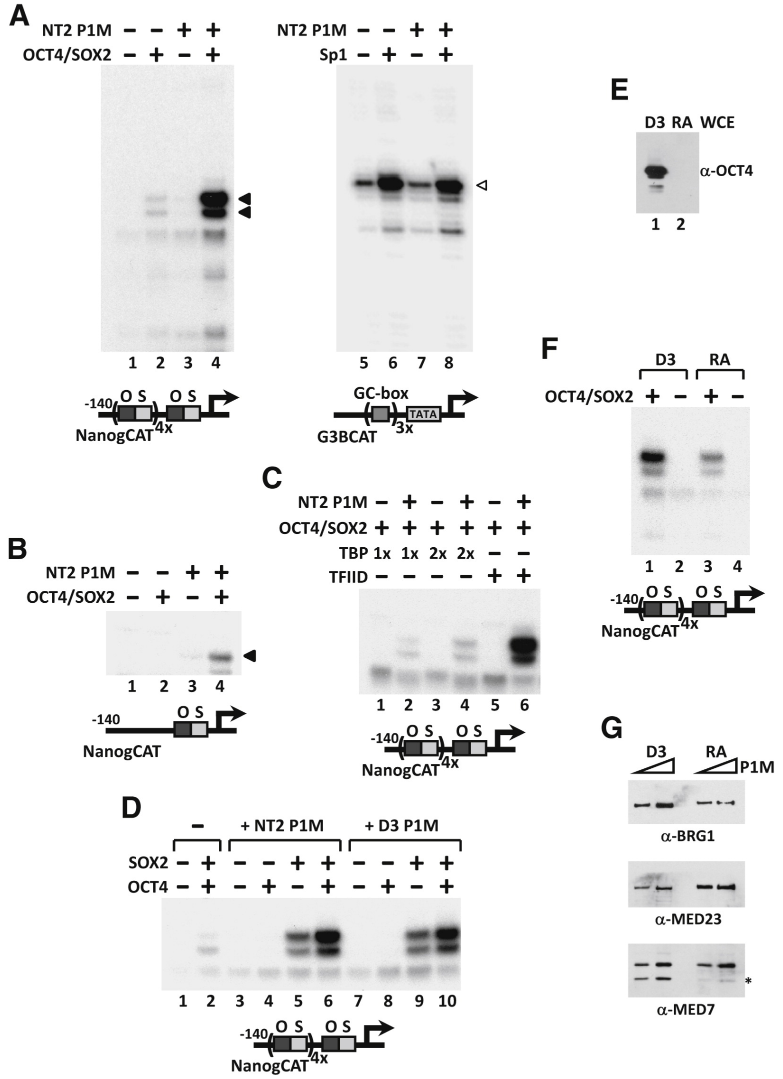


Figure 2.1: **Transcriptional Activation of *Nanog* by Oct4 and Sox2 Requires a Stem Cell-Specific Cofactor. (A)** Reconstituted in vitro transcription reactions supplemented with Oct4 and Sox2 (lanes 2 and 4) or Sp1 (lanes 6 and 8) plus a phosphocellulose 1 M KCl fraction derived from NT2 nuclear extracts (NT2 P1M, lanes 3, 4, 7, and 8) and programmed with either a *Nanog* template engineered with four extra copies of the oct-sox composite element (NanogCAT, lanes 1–4), or a GC box-containing template (G3BCAT, lanes 5–8). Oct4/Sox2, NT2 P1M-dependent transcripts are indicated by filled arrowheads and Sp1-dependent transcriptions by open arrowheads. (B) Transcription of the native *Nanog* promoter requires Oct4, Sox2, and NT2 P1M fraction (lane 4). (C) TFIID and NT2 P1M fraction are needed to potentiate Oct4/Sox2-dependent activation. Transcription reactions contain Oct4 and Sox2 (lanes 1–6), NT2 P1M fraction (lanes 2, 4, and 6) with increasing amounts of recombinant TBP (1 or 2, lanes 1–4), or TFIID (lanes 5 and 6). (D) Synergistic activation of *Nanog* by Oct4 and Sox2 requires P1M fractions prepared from NT2 or mouse ES cell line D3 nuclear extracts. In vitro transcription reactions contain equal amounts (0.7 g) of NT2 (lanes 3–6) or D3 P1M fractions (lanes 7–10), with Oct4 alone (lanes 4 and 8), Sox2 alone (lanes 5 and 9), or both activators (lanes 2, 6, and 10). (E) Immunoblotting analysis of Oct4 levels in whole-cell extracts (WCE) prepared from pluripotent D3 cells (D3, lane 1) and cells treated with retinoic acid for 6 days (RA, lane 2). (F) P1M fractions prepared from pluripotent (D3, lanes 1 and 2) and differentiated (RA, lanes 3 and 4) D3 nuclear extracts were added to transcription reactions with or without Oct4 and Sox2. (G) Western blots (2-fold titration) of P1M fractions prepared from pluripotent (D3) and differentiated (RA) D3 nuclear extracts using anti-BRG-1, anti-MED23, and anti-MED7 antibodies. Asterisk indicates a nonspecific band or a breakdown product recognized by anti-MED7 antibody.

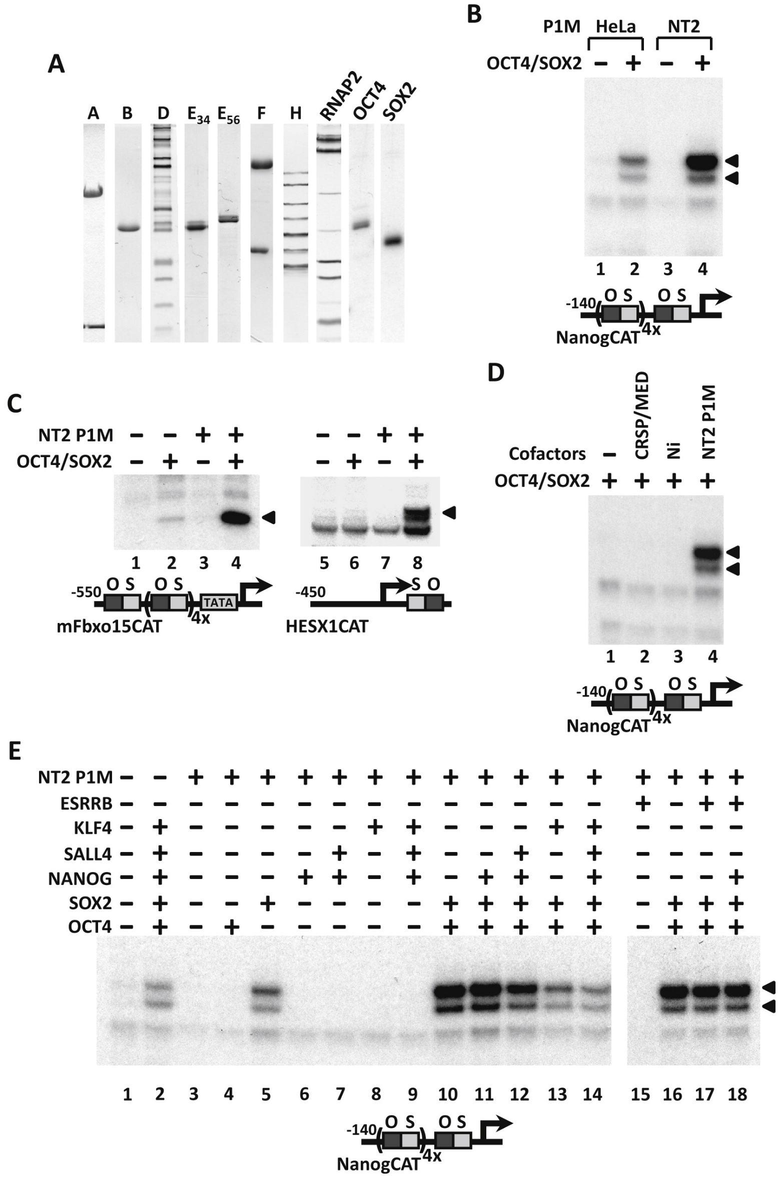

A
A B D E34 E56 F H RNAP2 OCT4 SOX2
B
P1M HeLa NT2
OCT4/SOX2 – + – +
1 2 3 4
-140 O S O S
4x
NanogCAT
C
NT2 P1M – – + + – – + +
OCT4/SOX2 – + – + – + – +
1 2 3 4 5 6 7 8
-550 O S O S TATA 4x
mFbxo15CAT
-450 S O
HESX1CAT
D
Cofactors – CRSP/MED Ni NT2 P1M
OCT4/SOX2 + + + +
1 2 3 4
-140 O S O S 4x
NanogCAT
E
NT2 P1M – – + + + + + + + + + + + + + + + +
ESRRB – – – – – – – – – – – – – – + – + +
KLF4 – + – – – – – + + – – – + + – – – –
SALL4 – + – – – – + – + – – + – + – – – –
NANOG – + – – – + + – + – + + – + – – – +
SOX2 – + – – + – – – – + + + + + – + + +
OCT4 – + – + – – – – – + + + + + – + + +
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18
-140 O S O S 4x
NanogCAT

Figure 2.2: **Transcription Factor and Cofactor Requirements for Oct4/Sox2-Dependent Activation of *Nanog*.** **(A)** Silver staining (TFII-D, -H and RNA polymerase II) and Coomassie staining (TFIIA, -B, -E34, -E56, -F, OCT4 and SOX2) of the general transcription factors and activators used in in vitro reconstituted transcription reactions. **(B)** Coactivator activity is highly enriched in pluripotent NT2 cells. Equal amounts of P1M fractions (0.7 mg) prepared from HeLa or NT2 nuclear extracts are assayed in in vitro transcription reactions with or without Oct4 and Sox2 using a *Nanog* promoter template. **(C)** Transcription of mouse *Fbxo15* and human *HESX1* promoters requires Oct4, Sox2 and NT2 P1M fraction. In vitro transcription reactions contain Oct4 and Sox2 (lanes 2, 4, 6 and 8), NT2 P1M fraction (lanes 3-4 and 7-8), and are programmed with a mouse *Fbxo15* template engineered with four additional copies of the oct-sox element (mFbxo15CAT, lanes 1-4), or a human HESX1 template (HESX1CAT, lanes 5-8). **(D)** CRSP/Mediator cannot substitute for NT2 P1M coactivator activity. Addition of GST-VP16 activation domain-purified CRSP/Mediator (CRSP/MED, lane 2), or a partially purified fraction containing CRSP activity prepared from HeLa NE (Ni, lane 3), fails to replace NT2 P1M in potentiating Nanog transcription in vitro (lane 4). textbf(E) Specificity of SCC cofactor activity. Coactivator activity supports only Oct4/Sox2-activated transcription of Nanog. Transcription reactions are supplemented with (lanes 3-18) or without (lanes 1 and 2) NT2 P1M fractions together with various combinations of affinity-purified activators as indicated by pluses (+) and minuses (-). Oct4/Sox2, NT2 P1M-dependent transcriptions are indicated by filled arrowheads.

We next investigated the relative requirements for other cofactors in our assay system. Consistent with previous studies demonstrating that TAFs in the TFIID complex are often required for transcriptional activation by a variety of activators including nuclear receptors,[47] Sp1[48] and SREBP-1,[49] substituting holo-TFIID with recombinant human TBP resulted in a near complete loss of activation by Oct4 and Sox2 (Figure 2.1C). The very weak residual activation we see using TBP (Figure 2.1C, lanes 2 and 4) is most likely due to trace amounts of TFIID present in the NT2 P1M fraction (data not shown). These findings suggest that TAFs/holo-TFIID and the putative cofactor detected in the NT2 P1M fraction are both required for optimal transcription of Nanog elicited by Oct4 and Sox2. Interestingly, in this reconstituted system, the addition of CRSP/Mediator complex was not required to obtain robust Oct4/Sox2 activation at the Nanog promoter. However, it is likely that some CRSP/Mediator is present in the P1M fraction, and it remains possible that some other component of the reconstituted system (i.e., Pol II) may have some residual amount of CRSP/Mediator contamination.[50] We found though that adding purified CRSP/Mediator instead of the NT2 P1M factor to these reactions completely failed to enhance Oct4/Sox2-dependent activation of Nanog transcription (Figure 2.2D). This finding indicates that the NT2 cofactor must be distinct from Mediator. Furthermore, addition of other transcriptional activators implicated in Nanog expression (i.e., Nanog, Sall4,[51] Klf4[52] and Esrrb[53,9]) also did not replace or enhance Oct4/Sox2-dependent transcription of Nanog in vitro (Figure 2.2E).

To confirm that this newly detected cofactor activity in NT2 cells is also present in bona fide ES cells, P1M fractions were prepared from the pluripotent D3 mouse ES cell line and assayed for transcription. We found that the D3 P1M fraction was as active as the NT2 P1M fraction in potentiating Oct4/Sox2-activated transcription of Nanog (Figure 2.1D, compare lane 2 to 6 and 10). Interestingly, the highest levels of trans-activation by the NT2 or D3 P1M fractions were observed only when both activators were added to the transcription reaction, whereas no activation was detected with Oct4 alone and a moderate level of activation was seen with Sox2 alone (Figure 2.1D, lanes 3-10). Apparently, this cofactor mediates the synergistic activation of Nanog by Oct4 and Sox2. If, as we postulated, this new coactivator functions selectively in pluripotent cells, one might expect that its presence or activity would need to be down-regulated upon differentiation, as is the case for Oct4. To investigate whether the cofactor activity is restricted to the pluripotent state of ES cells, D3 cells were induced to differentiate by removal of LIF and treatment with retinoic acid (RA). The extent of differentiation was monitored by the loss of Oct4 expression that was complete after 6 days (Figure 2.1E). Nuclear extracts and P1M fractions were then prepared from D3 cells before and after differentiation. When compared to pluripotent D3 P1M fractions, an equivalent amount of P1M fraction prepared from differentiated D3 nuclear extracts showed significantly decreased cofactor activity in our in vitro transcription assay (Figure 2.1F, compare lanes 1 and 3). This decrease is not due to a wholesale loss of transcription factors and other cofactors during stem cell differentiation because the levels of PBAF/BAF (BRG-1) and the Mediator complex (MED23 and MED7) were largely unchanged in the two extracts

(Figure 2.1G).

Purification and Identification of a Stem Cell Coactivator (SCC)

Starting with 200-400 L of NT2 cells, we were able to separate the cofactor activity into two distinct chromatographic fractions. One cofactor activity eluted from an anion exchanger (Poros-HQ) at 0.3M KCl (Q0.3, data not shown) while a second distinct activity eluted at 0.6M KCl (SCC, Figures 2.3A and 2.3B). Full synergistic Oct4/Sox2-dependant activation of Nanog required both fractions in our in vitro reconstituted transcription reactions (Figure 2.4). Using this biochemical complementation system, we sequentially purified the more robust activity, SCC, over eight chromatographic columns, resulting in >50,000-fold increase in specific activity (Figure 2.3A). Since SCC activity migrated with an apparent native molecular mass (Mr) of 600kDa during size-exclusion chromatography (Figure 2.3C), it seemed likely that this coactivator was a multi-protein complex. Accordingly, SDS-polyacrylamide gel electrophoresis (SDS-PAGE) of the most purified Mono S fractions revealed a distinct pattern of four major polypeptides (along with multiple breakdown products) that consistently co-purified with the SCC activity (Figures 2D and 2E). For the remainder of this report, we focus on the identification and functional characterization of SCC in vitro and in vivo.

To identify polypeptides comprising the SCC complex, peak Mono S-purified fractions were pooled and separated by SDS-PAGE. Surprisingly, tryptic digests of excised gel bands followed by high sensitivity mass spectrometry revealed all detectable constituents of SCC to be none other than the Xeroderma pigmentosum group C (XPC)-RAD23B-Centrin 2 (CETN2) nucleotide excision repair (NER) complex[54](Figure 2.5A). We next carried out western blot analysis with antibodies specific to XPC, RAD23B and CETN2 to confirm the identities of the purified SCC subunits (Figure 2.5B). As expected, these three polypeptides were highly enriched in the purified SCC Mono S peak fractions when compared to the crude NT2 P1M fraction (Figure 2.5B). Because identification of SCC as being identical to the XPC-NER complex was so unexpected, particularly as this repair complex has not been associated with any cell type-specific function nor linked to stem cell transcription, we next wanted to compare the relative amounts of this factor in different cell types. Consistent with the notion that SCC may be functioning in an unusual way in pluripotent stem cells, we found that these three proteins are highly enriched in ES and EC cells. For example, the levels of XPC, RAD23B and CETN2 in the NT2 P1M fraction are much higher than in an equivalent amount of P1M fraction prepared from HeLa nuclear extracts (Figure 2.5B). Accordingly, in in vitro transcription reactions, Oct4/Sox2-dependent activation of *Nanog* by HeLa P1M fraction is much lower than that of NT2 P1M fraction (Figure 2.2B). XPC and RAD23B were rapidly down-regulated upon RA-induced differentiation of mouse D3 ES cells, whereas CETN2, components of the basal transcription machinery (TBP and TFIIE) and other NER factors (XPA and XPB) decreased only slightly while the loading control -actin remained unchanged (Figure 2.5C). This finding is consistent with our previous observation that the

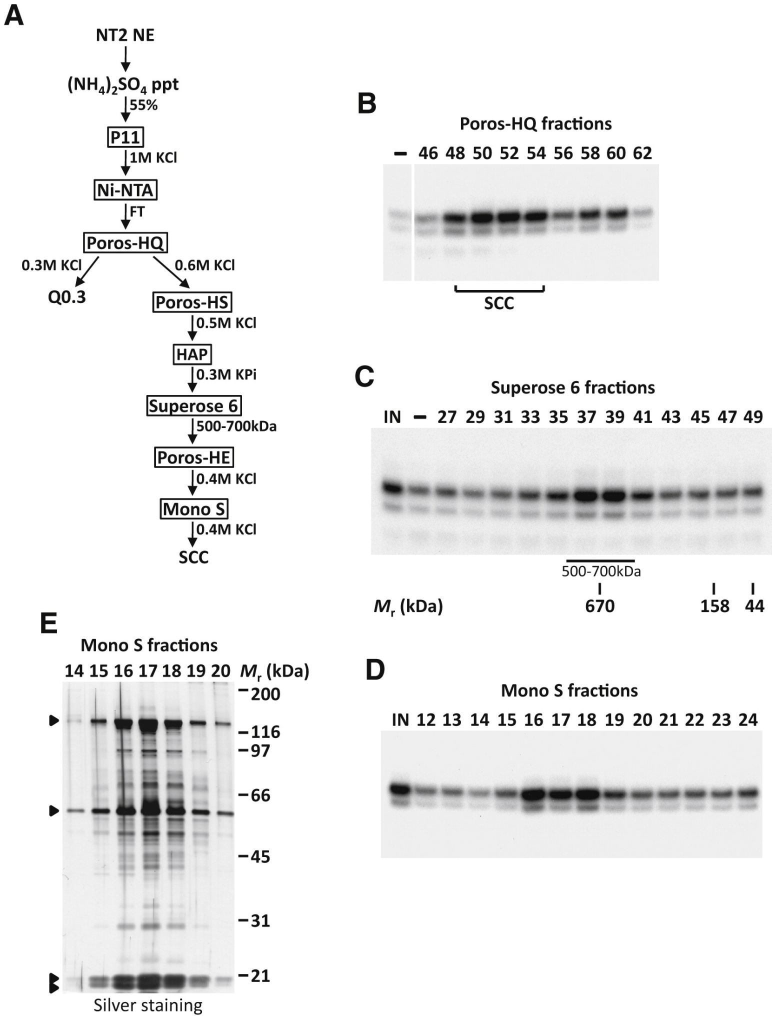


Figure 2.3: **Purification of Stem Cell Coactivator (A)** Chromatography scheme for partial purification of Q0.3 and purification of SCC from NT2 nuclear extracts (NT2 NE). NT2 NE is first subjected to ammonium sulfate precipitation (55% saturation) followed by a series of chromatographic columns as indicated. **(B)** Buffer (-) and fractions containing SCC eluted from a Poros-HQ anion exchanger (top) assayed in the presence of Oct4 and Sox2 in in vitro transcription assays. **(C)** Coactivator SCC migrates as a large complex. Input (IN), buffer (-), and Superose 6 fractions (top) assayed as in **(B)** except that all reactions are supplemented with Q0.3 **(A)**. Mobilities of peak activity (500–700 kDa) and gel filtration protein standards are shown (bottom). **(D)** Transcription profile of stem cell coactivator (SCC) activity after the final Mono S chromatography step. Reactions contain input (IN) and Mono S fractions (top) and are assayed as in **(C)**. **(E)** Silver-stained SDS-PAGE gel of the active Mono S fractions. Filled arrowheads indicate polypeptides that comigrate with SCC activity.

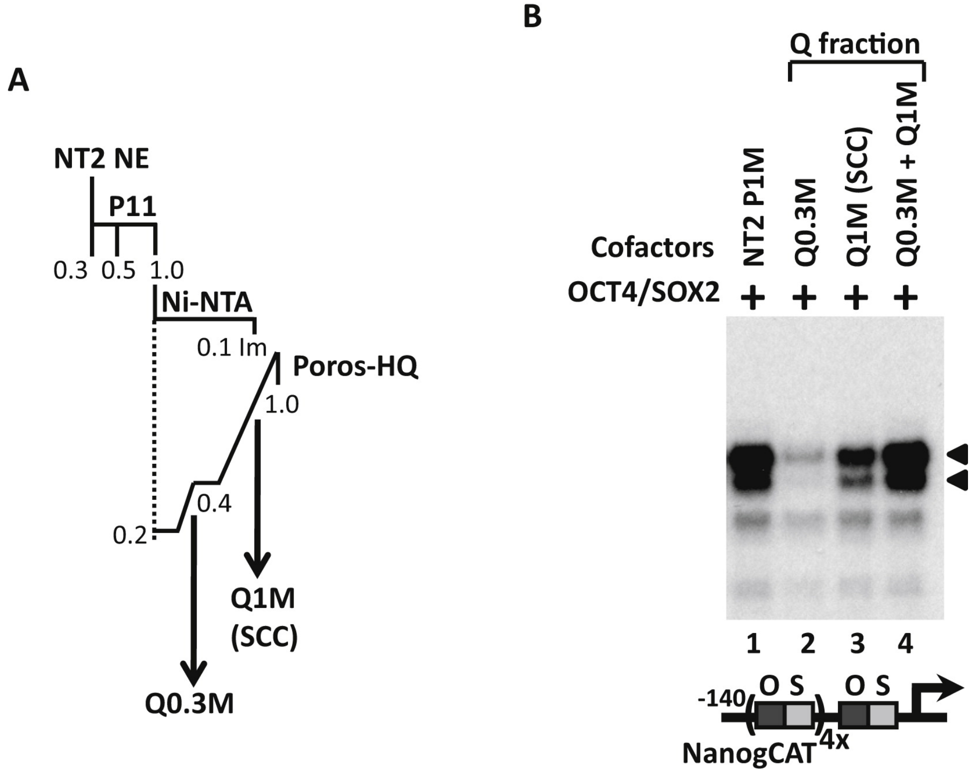


Figure 2.4: **Discovery of Cofactor Activities in NT2 Nuclear Extracts. (A)** Separation of a P1M fraction into Q0.3M and Q1M (SCC) activities by a Poros-HQ anion exchanger. NT2 P1M fraction is incubated with Ni-NTA resin and the flowthrough fraction is then applied to a Poros-HQ anion column at 0.2M KCl and step eluted at 0.4M KCl (Q0.3) and again at 1M KCl (Q1M). **(B)** Optimal activation by Oct4 and Sox2 requires both activities. Optimal transcription of *Nanog* directed by Oct4 and Sox2 requires both Q0.3M and Q1M (SCC) fractions. Transcription reactions contain NT2 P1M (lane 1), Q0.3M (lane 2), Q1M (SCC, lane 3), or both (lane 4). P1M-dependent transcriptions are indicated by filled arrowheads.

D3 P1M fraction from differentiated cells is significantly less active than the pluripotent D3 P1M fraction in potentiating *Nanog* transcription (Figure 2.1F).

Reconstitution and Mechanism of Coactivation by SCC

While we were in the process of further characterizing the role of the XPC-RAD23B-CETN2 complex in transcription, Le May et al reported that XPC and other components of the NER apparatus can be recruited to a gene promoter (e.g., RAR2) upon nuclear hormone induction.[55] Although the mechanism by which XPC and other NER factors mediate gene activation remains unclear, these recent studies and our new findings have unmasked a hitherto unknown and potentially important role for XPC that is directly linked to transcription. In our case, the most striking finding was the direct requirement for the SCC/XPC complex in selectively potentiating the transcriptional activation of Nanog by Oct4 and Sox2 in ES cell extracts. However, to more firmly establish this exciting new connection, we first needed to eliminate the possibility that trace amounts of contaminants present in our purified SCC frac-

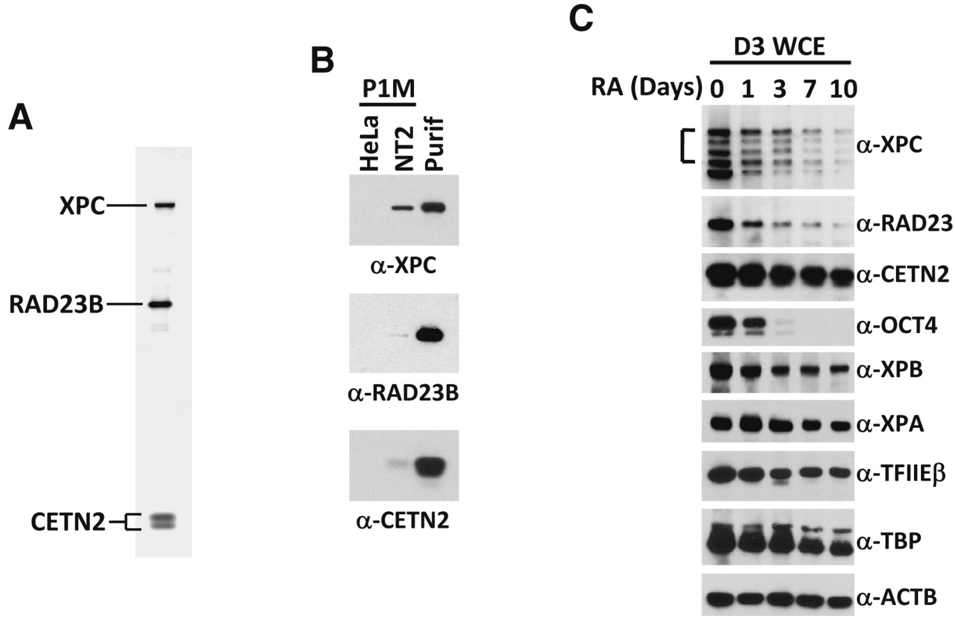


Figure 2.5: **SCC Is the XPC-RAD23B-CETN2 Nucleotide Excision Repair Complex (A)** Mass spectrometry analysis of Mono S peak activity fractions (16–18) in Figure 2.3E with protein identities indicated. **(B)** SCC is highly enriched in NT2 P1M fraction. Comparative western blot analysis of HeLa and NT2 P1M fractions (1.5 g each) and purified Mono S SCC fraction (Purif, 30 ng) using anti-XPC, anti-RAD23B, and anti-CETN2 antibodies. **(C)** Downregulation of XPC and RAD23B upon RA-induced differentiation of mouse D3 ES cells. Western blot analysis of whole-cell extracts prepared from D3 cells (D3 WCE) collected at indicated days post-RA treatment using antibodies against XPC, RAD23B, CETN2, OCT4, XPB, XPA, TFIIE, TBP, and loading control -actin (ACTB).

tion were responsible for the coactivator activity detected in our in vitro transcription assays. Therefore, we set about to reconstitute the heterotrimeric XPC-RAD23B-CETN2 complex from recombinant gene products expressed in insect (Sf9) cells following co-infection with baculoviruses expressing His-tagged XPC, FLAG-tagged RAD23B and untagged CETN2. Using an efficient two-step affinity purification procedure, we were able to purify the recombinant heterotrimeric complex to near homogeneity (Figure 2.7A). Our ability to generate pure polypeptide subunits, as well as various combinations of dimeric and trimeric complexes, allowed us to address a number of important questions, such as whether known functional domains of XPC required for NER are also necessary for the cofactor activity. It is well established that XPC's ability to interact non-specifically with DNA is essential for its NER function. Indeed, a single point mutation in the DNA binding domain (W690S) of XPC, identified in an XP patient (XP13PV), abolishes binding to damaged (and undamaged) DNA and is defective in repair in vivo and in vitro.[56,57] To address whether XPC's non-specific DNA binding activity is also important for its coactivator function, a mutant DNA-binding defective XPC (W690S) complex (that had been independently confirmed to be compromised for DNA binding in vitro, Figures S3A and S3B) was reconstituted in Sf9 cells and tested along with the wild type complex for their ability to support Oct4/Sox2-dependent transcriptional activation of Nanog in vitro. Surprisingly, both the recombinant wild-type and mutant complexes exhibited specific activities for coactivation comparable to that observed for purified native endogenous SCC from NT2 cells (Figure 2.7B). Taken together, these results confirm that the XPC-RAD23B-CETN2 complex is indeed SCC, and suggest that its DNA binding (and repair) activity is dispensable and functionally separable from its transcriptional cofactor activity at least in vitro. It has also been reported that XPC can interact directly with TFIIH[58] and thus might provide a DNA-independent mechanism by which SCC can be recruited to gene promoters. To test this possibility, a C-terminal truncation of XPC that abolishes TFIIH (and CETN2) but retains RAD23B binding (amino acids 1 to 813, C814St[59]) was used in our in vitro assay and found to have no adverse affect on the ability of a XPC (C814St)-RAD23B heterodimer to mediate Oct4/Sox2-activated transcription of Nanog (Figure 2.6C and D). We therefore speculate that SCC/XPC is most likely targeted to its cognate promoters via potential interactions with specific activators such as Oct4 and Sox2.

To probe for a potential direct interaction between SCC and Oct4 and/or Sox2, mouse SCC subunits were over-expressed with Oct4, Sox2, Klf4 and c-Myc (STEMCCA[60]) in 293T cells. SCC co-immunoprecipitated with Oct4 but not with control IgG (Figure 2.7C). To examine whether the DNA-binding property of SCC is required for its interaction with Oct4 and other activators, both the wild-type (WT) and DNA-binding defective (W683S in mouse) XPC/SCC complexes were co-expressed with STEMCCA. Immunoprecipitation of WT and mutant SCC complexes using an anti-RAD23B antibody pulled down both Oct4 and Sox2 but not Klf4 or XPA (Figure 2.7D). These data indicate a direct and specific protein-protein binding between SCC and select activators thus providing a mechanism by which SCC may serve as a transcriptional coactivator for Oct4 and Sox2 (but not Klf4, see Figure 2.2E) in potentiating Nanog transcription. These findings may also explain why the DNA-binding activity of the XPC subunit of SCC is dispensable for transcription in vitro. We were however unable to reproducibly detect a stable interaction between SCC and Oct4/Sox2 in D3 ES cell extracts. It is worth noting, though, that other coactivators implicated in Oct4/Sox2-directed transcriptional activation (e.g., Mediator and p300/CBP) have not been identified in recent "interactome" studies on Oct4, Sox2 or Nanog-associating factors,[61,62,63] supporting the idea that functional coactivator-activator interactions can often be weak and transient.

The ability to reconstitute active SCC from purified recombinant subunits also provided us with a unique opportunity to examine the contribution of individual subunits, as well as different dimeric combinations in supporting Oct4/Sox2 transcriptional activation. Purified individual subunits (XPC or RAD23B), partial dimeric complexes (XPC-RAD23B or XPC-CETN2), and holo-SCC complexes (Figure 2.7E) were assayed over a four-fold dose response range in our fully reconstituted in vitro transcription reactions containing Oct4, Sox2 and a partially purified Q0.3 fraction (Figure 2.7F). The large XPC subunit alone only slightly activated transcription above background at the highest concentrations tested (Figure 2.7F, compare lanes 1 and 4) while RAD23B alone was essentially inactive. The XPC-CETN2 dimer was slightly more active than XPC alone. By contrast, a marked gain in specific

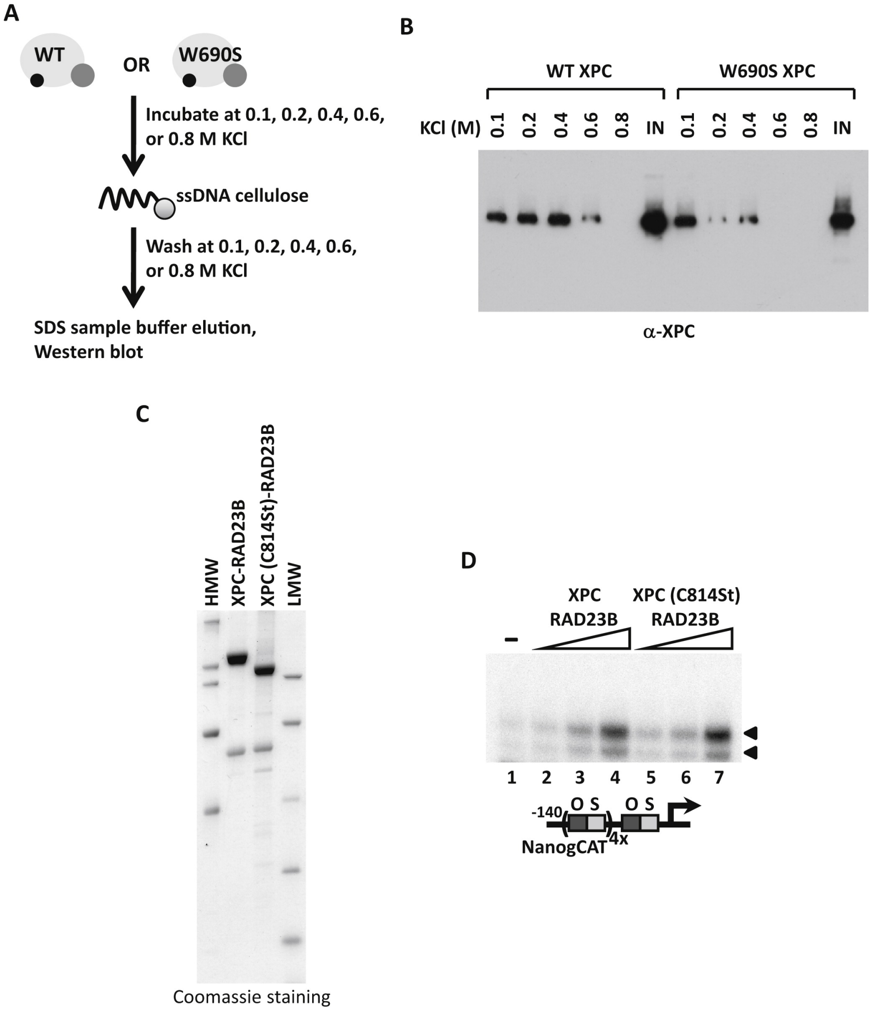

A
WT
OR
W690S
Incubate at 0.1, 0.2, 0.4, 0.6, or 0.8 M KCl
ssDNA cellulose
Wash at 0.1, 0.2, 0.4, 0.6, or 0.8 M KCl
SDS sample buffer elution, Western blot
B
WT XPC
W690S XPC
KCl (M) 0.1 0.2 0.4 0.6 0.8 IN 0.1 0.2 0.4 0.6 0.8 IN
α-XPC
C
HMW
XPC-RAD23B
XPC (C814St)-RAD23B
LMW
Coomassie staining
D
XPC RAD23B
XPC (C814St) RAD23B
-
1 2 3 4 5 6 7
-140
O S
O S
4x
NanogCAT

Figure 2.6: **DNA and TFIIH Binding by XPC Are Dispensable for SCC Activity. (A)** Recombinant mutant W690S XPC-containing SCC complex is compromised in DNA binding. Schematics of DNA pull-down experiment. W690S mutation in XPC destabilizes XPC-DNA interaction. Transcriptionally active, recombinant wild-type (WT) or mutant (W690S) XPC/SCC complexes (see Figure 2.7) reconstituted in insect Sf9 cells are incubated with single stranded calf thymus DNA (ssDNA) cellulose at increasing salt concentration (0.1 to 0.8 M KCl). **(B)** Input (IN) and bound proteins eluted by SDS are analyzed by Western blotting using anti-XPC antibody. **(C)** Coomassie-stained SDS-PAGE gel of purified full-length and C-terminally truncated (C814St) XPC-RAD23B heterodimeric complexes reconstituted in insect Sf9 cells. **(D)** SCC stimulates Oct4/Sox2-activated transcription independent of TFIIH binding. Buffer (-), increasing amount of full-length (lanes 2-4) and truncated (lanes 5-7) XPC-RAD23B heterodimeric SCC complexes are assayed together with Oct4, Sox2 and Q0.3M fractions (lanes 1-7) in in vitro transcription reactions. P1M-dependent transcriptions are indicated by filled arrowheads.

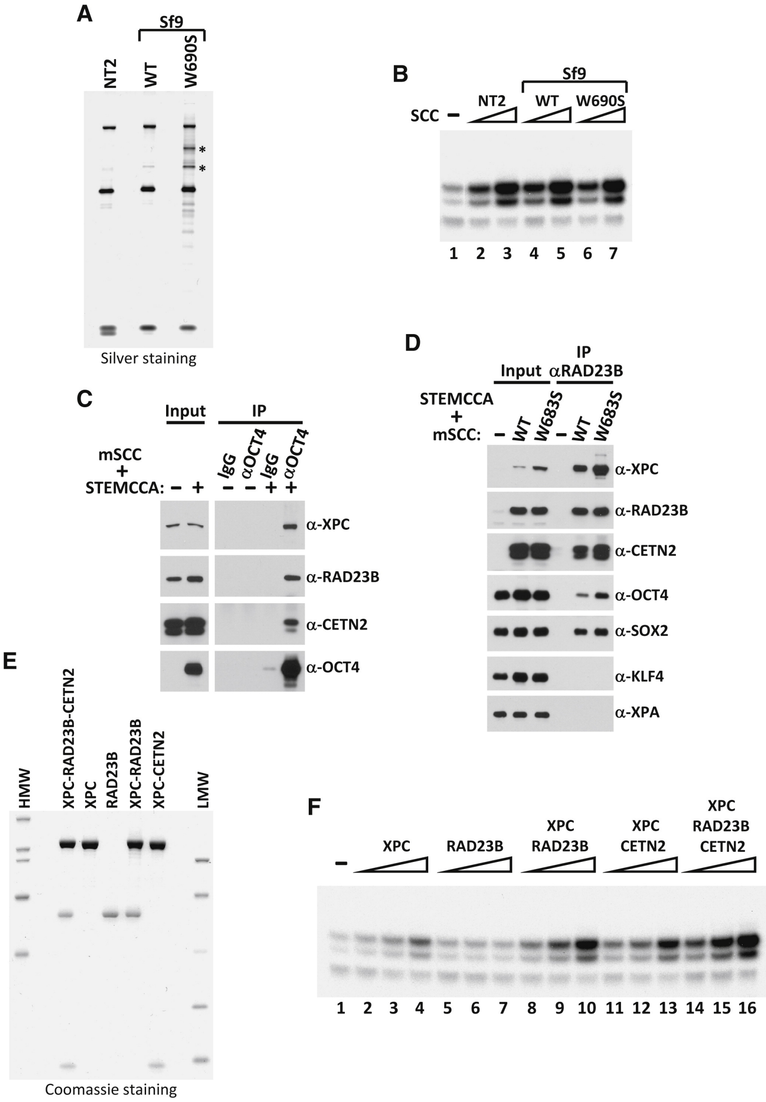


Figure 2.7: **Reconstitution of Recombinant SCC Complexes.** **(A)** Silver-stained SDS-PAGE gel of purified NT2 SCC (NT2), recombinant wild-type (WT), and DNA-binding-defective mutant (W690S) XPC-containing SCC complexes reconstituted in insect Sf9 cells by coinfection with baculoviruses expressing His-tagged XPC, FLAG-tagged RAD23B, and untagged-CETN2. Major proteolytic fragments of mutant XPC are indicated by asterisks. **(B)** Recombinant SCC complex enhances Oct4/Sox2-activated transcription of *Nanog* independent of DNA binding. Buffer (-), NT2 (Mono S peak activity fractions; lanes 2 and 3), recombinant WT (lanes 4 and 5), and W690S mutant (lanes 6 and 7) SCC complexes are assayed (over a 3-fold concentration range). All transcription reactions contain Oct4, Sox2, and Q0.3 (lanes 1–7). **(C)** Oct4 interacts with SCC. Western blot analysis of input lysates (2%) and coimmunoprecipitated proteins from extracts of 293T cells transfected with a polycistronic expression plasmid encoding all three subunits of mouse SCC (mSCC) with or without a polycistronic plasmid expressing mouse Oct4, Sox2, Klf4, and c-Myc (STEMCCA) using normal IgG or anti-Oct4 antibody. See also Figure 2.6. **(D)** SCC-B interacts directly with Oct4 and Sox2 independent of DNA binding. Control vector (-), plasmids expressing wild-type (WT), or mutant (W683S) XPC-containing mSCC complexes were cotransfected with STEMCCA into 293T cells and immunoprecipitated with anti-RAD23B antibody. Input lysates (2%) and RAD23B-bound proteins were detected by immunoblotting. **(E)** Coomassie-stained SDS-PAGE gel of purified recombinant XPC, RAD23B, dimeric (XPC-RAD23B and XPC-CETN2), and holo-SCC (XPC-RAD23B-CETN2) complexes. **(F)** Titrations (over a 4-fold concentration range) of XPC (lanes 2–4), RAD23B (lanes 5–7), XPC-RAD23B (lanes 8–10), XPC-CETN2 (lanes 11–13), and XPC-RAD23B-CETN2 (lanes 14–16) in in vitro transcription reactions supplemented with Q0.3 (lanes 1–16) and assayed as in (B).

activity was observed with the XPC-RAD23B dimeric complex that was nearly as active as the holo-complex (Figure 2.7F). These results suggest that the minimal active complex likely consists of XPC and RAD23B, while CETN2 may enhance the activity of the complex by providing structural support or stability.

SCC Coactivator Function in ES Cell Self-renewal and Somatic Cell Reprogramming

We next set out to determine the role of the SCC/XPC complex on gene expression and Nanog transcription by loss-of-function studies in ES cells. Lentiviruses containing two independent short hairpin RNAs (shRNAs) specifically targeting XPC, RAD23B and CETN2 were used to infect mouse D3 ES cells to selectively deplete SCC (Figures 2.8A, 2.9A and 2.9B). Knockdown of SCC subunits resulted in pronounced cellular morphological abnormalities and decreased alkaline phosphatase (AP) activity (Figures 2.8B and 2.9C). These knockdown cells also showed reduced proliferation rates when compared to control ES cells infected with non-target viruses, indicating that the self-renewal capacity of ES cells depleted of SCC may also be compromised (data not shown). Indeed, prolonged depletion of SCC resulted in the apoptosis of flattened, fibroblastic AP-negative cells surrounding the collapsing ES cell colonies (Figure 2.8B and data not shown). Therefore, knockdown of SCC in ES cells likely promotes differentiation followed by rapid apoptosis, two processes that are often coupled. Quantification of colony assays revealed that ES cells depleted of SCC formed fewer undifferentiated colonies with a corresponding increase in partially and fully differentiated colonies (Figure 2.8C). Consistent with the observed morphological changes associated with compromised stem cell identity, double and triple knockdown of XPC, RAD23B and CETN2 resulted in a 2-3-fold reduction in the mRNA level of Nanog (Figures 5D and S4D) as well as several other stem cell markers (Fgf4, Zfp42 and Utf1) (Figure 2.8D). Knockdown of individual subunits of SCC resulted in only mild effects on Nanog expression (Figure 2.9D). Accordingly, we did not observe overt defects in self-renewal in these single subunit knockdown ES cells (data not shown).

To further probe the molecular mechanism underpinning the function of SCC as a transcriptional coactivator for Oct4 and Sox2 in vivo, we investigated whether regulatory regions of Nanog and Oct4 might serve as direct SCC targets by performing chromatin immunoprecipitation (ChIP) assays in D3 cells using a RAD23B antibody. ChIP-qPCR analysis revealed that RAD23B (and presumably XPC/SCC) occupancy sites coincide with those of Oct4[24,25,26] and Sox2 (Figures 2.11A and 2.10A). By contrast, we failed to detect any significant enrichment of RAD23B at housekeeping genes -actin (Actb, Figure 2.11A) and dihydrofolate reductase (Dhfr, Figure 2.10B), or an intergenic region on chromosome 1 (Figure 2.10B).

To evaluate the extent to which Oct4 and Sox2 target sites overlap those of RAD23B on a genome-wide scale, we performed RAD23B ChIP assays followed by high-throughput

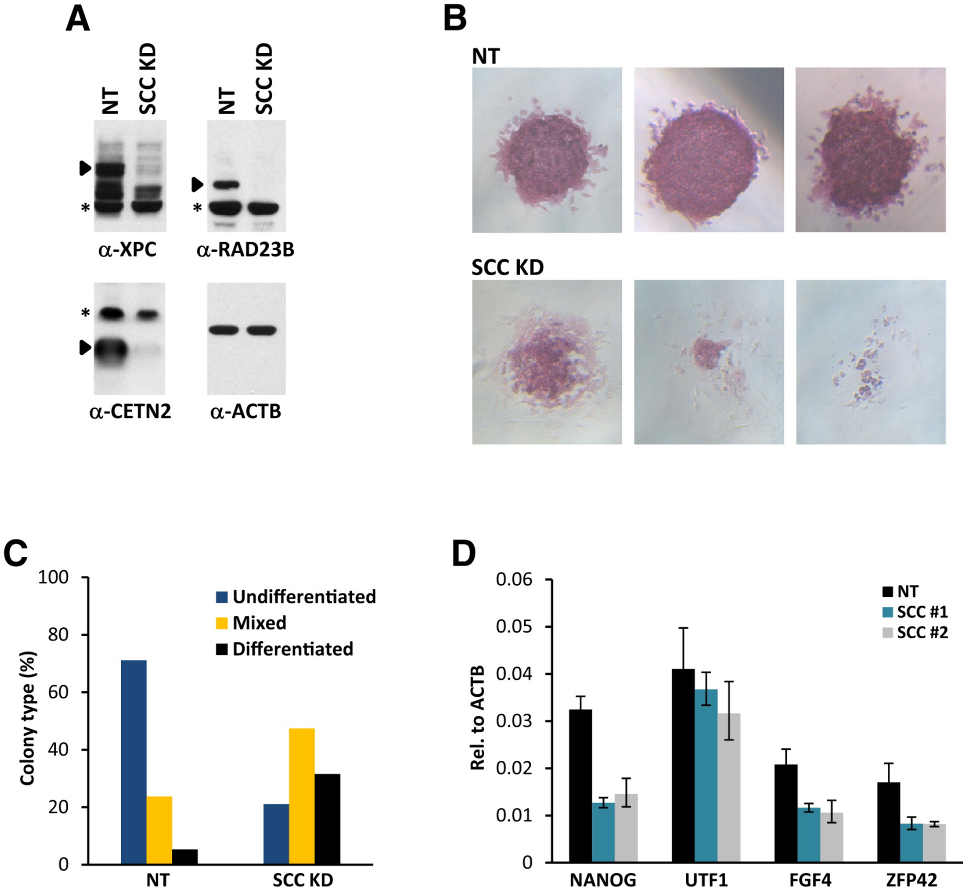


Figure 2.8: **SCC Is Required for ES Cell Maintenance. (A)** Efficiency of shRNA-mediated depletion of SCC in mouse ES cell line D3. Whole-cell extracts of mouse D3 cells infected with nontarget (NT) lentiviruses (MOI of 300) or with an equal mixture of three lentiviruses (MOI of 100 each) targeting XPC, RAD23B, and CETN2 (SCC KD) are analyzed by western blotting. Specific bands recognized by their respective antibodies are indicated by filled arrowheads. Asterisks denote nonspecific signals. **(B)** ES cell colony morphology and alkaline phosphatase (AP) activity (red) are maintained in control D3 cells (NT, top) but are compromised in SCC-depleted D3 cells (SCC KD, bottom). See also Figure 2.9C. **(C)** Clonal assays on SCC-depleted D3 ES cells. Stable nontarget (NT) and SCC-depleted (SCC KD) D3 cell pools were plated at 300 cells per well in 6-well plates, and emerging colonies were stained for AP activity. Differentiation status was scored based on AP staining intensity, ES cell morphology, and colony integrity after 6 days. **(D)** Two nonoverlapping sets of shRNAs targeting SCC (SCC 1 and SCC 2) are used to deplete SCC. Quantification of *Nanog*, *Utf1*, *Fgf4*, and *Zfp42* mRNA levels are analyzed by real-time quantitative PCR (qPCR) and normalized to *Actb*. Data from representative experiments are shown; error bars represent standard deviations. n = 3.

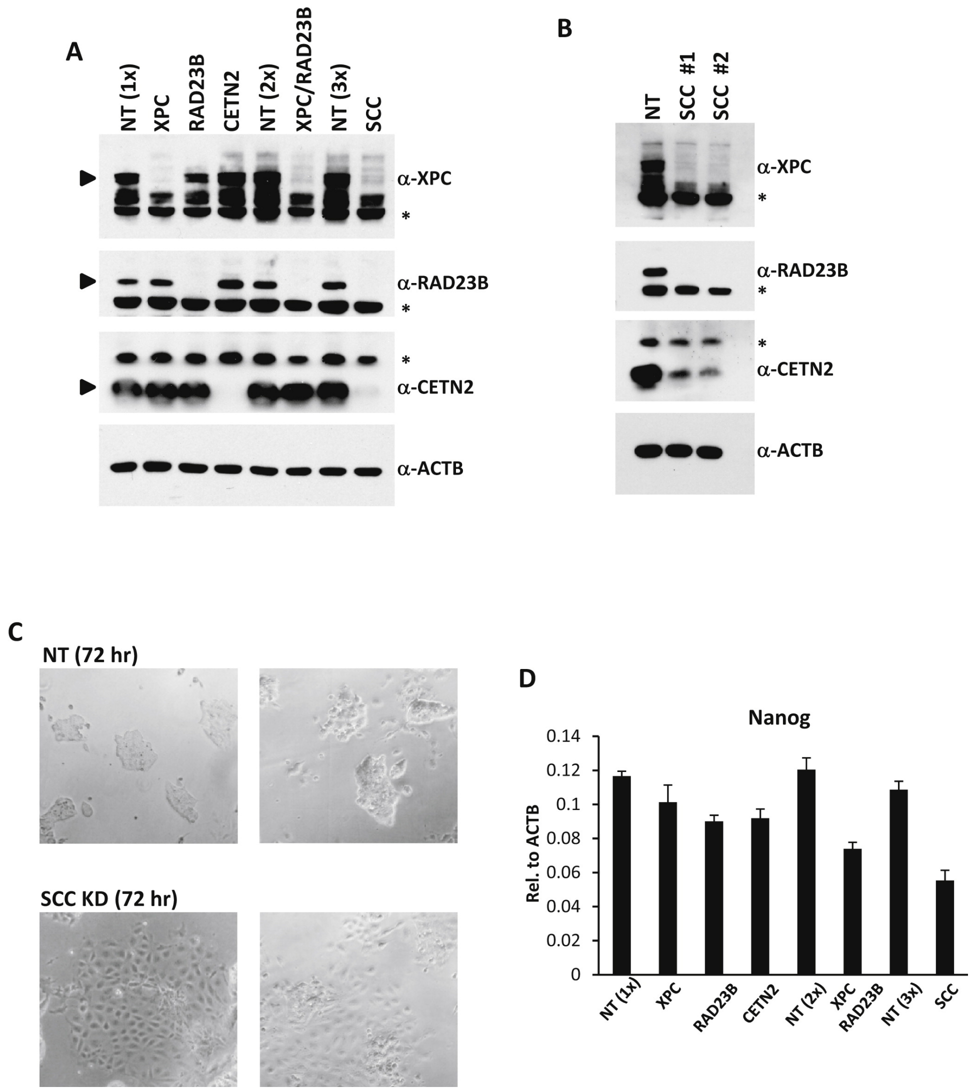


Figure 2.9: **SCC Is Required for ES Cell Maintenance.** **(A)** shRNA-mediated depletion of SCC in mouse ES cell line D3. Amount of non-target lentiviruses are adjusted to match the total viral load used to perform single, double or triple knockdown (1x, 2x and 3x, respectively) of SCC. Whole cell extracts prepared from non-target (NT), single (XPC, RAD23B, or CETN2), double (XPC/RAD23B) and triple (SCC) knockdown D3 cells are analyzed by Western blotting. Specific bands recognized by their respective antibodies are indicated by filled arrowheads. Asterisks denote non-specific signals. **(B)** Efficiency of SCC knockdown by two non-overlapping sets of shRNAs targeting SCC is monitored by Western blotting as in **(A)**. **(C)** D3 cells depleted of SCC (SCC KD) exhibit abnormal morphology at 72 hr post infection when compared to control cells (NT). **(D)** SCC depletion in D3 ES cells reduces *Nanog* expression. *Nanog* mRNA levels in single (XPC, RAD23B, or CETN2), double (XPC/RAD23B) and triple (SCC) knockdown D3 cells are analyzed by real time quantitative PCR (qPCR) and normalized to ACTB. Error bars represent standard deviations (n = 3).

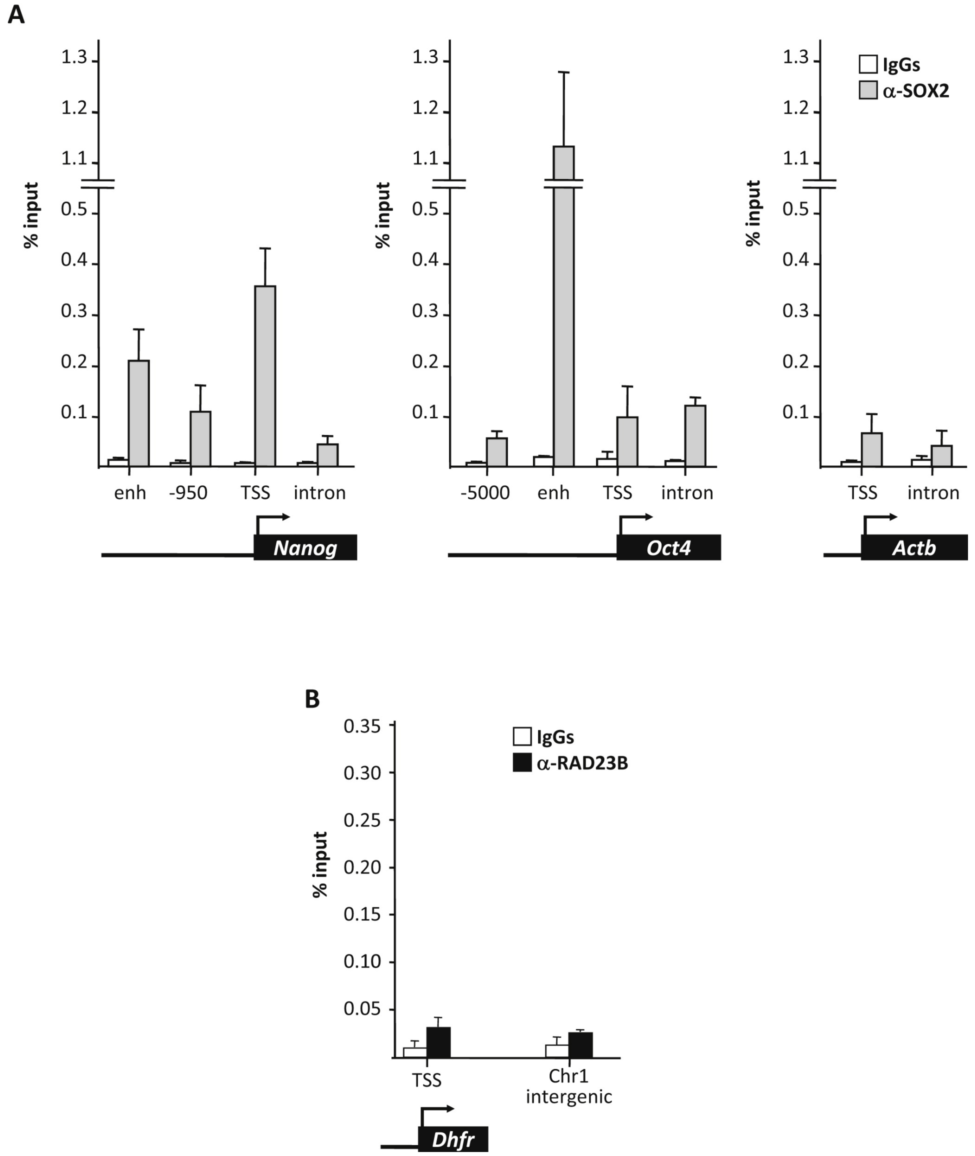
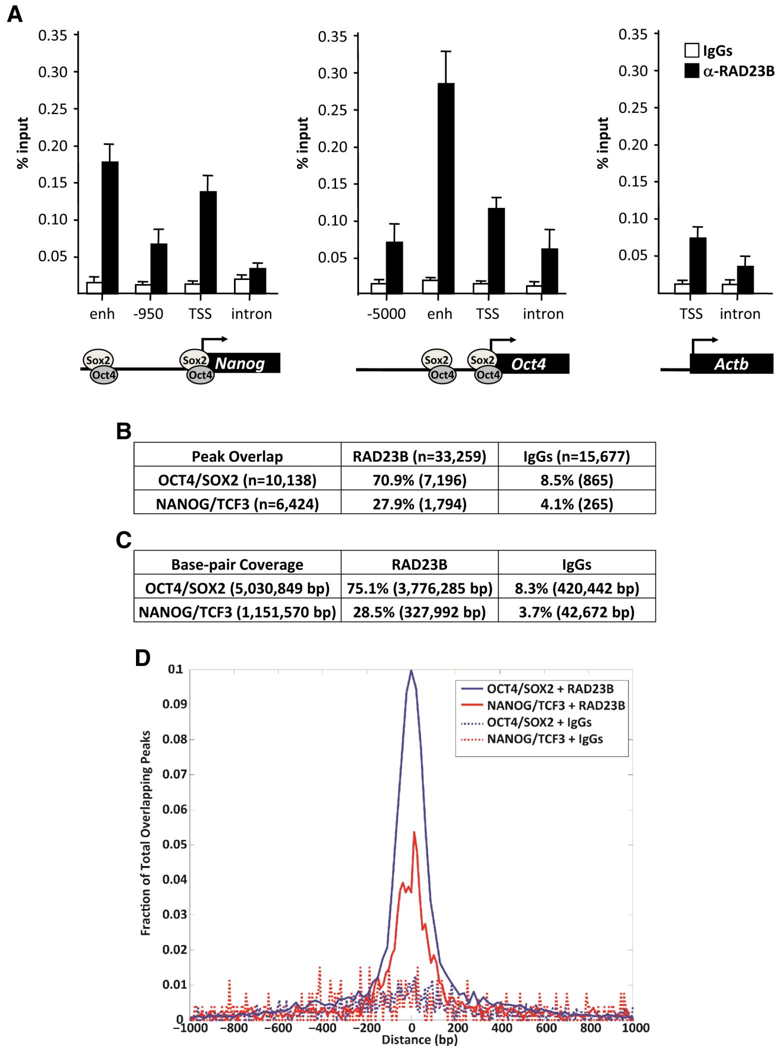


Figure 2.10: **SCC and Sox2 Co-occupy the *Nanog* and *Oct4* Promoters.** **(A)** Sox2 is enriched on the promoters of *Nanog* and *Oct4*. ChIP analysis of Sox2 occupancy on the regulatory (enh and TSS) and negative (upstream and intronic) regions of *Nanog*, *Oct4*, and *Actb*. Representative data (n > 3) showing the enrichment of Sox2 (gray bars) compared to normal IgGs (white bars) is analyzed by qPCR and expressed as percentage of input chromatin. Error bars represent standard deviations (n = 3). **(B)** RAD23B is not recruited to housekeeping gene Dhfr or an intergenic region on chromosome 1. Enrichment of RAD23B compared to IgG control is analyzed as in **(A)**.

sequencing (ChIP-seq) to identify an entire range of RAD23B/SCC-bound genomic regions in D3 cells. RAD23B ChIP-seq results were then compared with published Oct4 and Sox2 ChIP-seq data along with those of Nanog and Tcf3[27] to assess any potential bias in RAD23B occupancy in relation to these transcription factors. This analysis revealed a striking binding preference of RAD23B/SCC to genomic sites that are also co-occupied by Oct4 and Sox2, but not Nanog or Tcf3 only (70% versus 28%, p <10-15, ANOVA). This strong bias is maintained whether the ChIP-seq data sets are analyzed by the degree of peak overlap (defined by any two peaks with at least one nucleotide of overlap, Figure 2.11B) or base-pair coverage (Figure 2.11C), indicating that the majority of RAD23B/SCC binding sites align with those of Oct4 and Sox2. Importantly, the same analyses performed on ChIP-seq samples obtained from control IgG immunoprecipitations yielded only background correlation (between 4-8%), confirming the specificity of the RAD23B/SCC association. We further validated the co-localization among RAD23B/SCC, Oct4 and Sox2 by measuring the distance between overlapping RAD23B/SCC and Oct4/Sox2 peaks (See Extended Experimental Procedures). The majority of them (76%) lie within close proximity (200 base pairs) of each other (Figure 2.11D). Even though most of RAD23B/SCC-bound regions overlap poorly with those bound by Nanog/Tcf3 (28%), those that do are still largely (64%) positioned within 200 base pairs from each other but with a noticeably different distribution pattern than that of Oct4/Sox2 (p <10-15, ANOVA, Figure 2.11D). However, upon a closer look at the Nanog/Tcf3 “only” genomic coordinates that overlap with RAD23B-bound sites we found that many of them (40%) could in fact contain Oct4 and/or Sox2 when an alternative peak calling strategy (MACS) was used. Taken together, these data strongly suggest a classical coactivator function rather than a purely NER function of SCC both in vitro with naked DNA and in the context of chromatin in ES cells as XPC/RAD23B-mediated DNA damage repair generally involves transient interactions with DNA[64] that would not show either sequence or promoter specificity.

Given the importance of SCC in stem cell maintenance, we next asked whether it might also play a role in the re-acquisition of pluripotency during somatic cell reprogramming. Downregulation of either XPC or RAD23B in Oct4-GFP mouse embryonic fibroblasts (MEFs), which express some SCC albeit at significantly lower levels than ES cells, led to a dramatic reduction in the reprogramming efficiency. We observed a significant decrease in the number of AP-positive colonies, as well as a marked reduction in the percentage of partially (SSEA-1+, GFP-) and fully (SSEA-1+, GFP+) reprogrammed cells, as determined by FACS sorting (Figures 2.13A, 2.13B and 2.12A). Consistent with our in vitro reconstitution result showing that the CETN2 subunit may not be essential for the transcriptional activity of SCC (Figure 2.7F), knockdown of CETN2 had minor effects on iPS cell derivation efficiency. As expected, reprogramming efficiency using MEFs derived from XPC and RAD23B knockout (KO) mice[65] was also highly compromised. Surprisingly, RAD23A KO MEFs were nearly as efficient as wild-type or RAD23A and B double heterozygous MEFs in generating AP-positive colonies upon iPS cell induction (Figures 2.13C and 2.12B). This result may



Peak Overlap	RAD23B (n=33,259)	IgGs (n=15,677)
OCT4/SOX2 (n=10,138)	70.9% (7,196)	8.5% (865)
NANOG/TCF3 (n=6,424)	27.9% (1,794)	4.1% (265)

Base-pair Coverage	RAD23B	IgGs
OCT4/SOX2 (5,030,849 bp)	75.1% (3,776,285 bp)	8.3% (420,442 bp)
NANOG/TCF3 (1,151,570 bp)	28.5% (327,992 bp)	3.7% (42,672 bp)

Figure 2.11: **SCC Is Recruited to the *Nanog* and Oct4 Promoters and Genomic Regions Occupied by Oct4 and Sox2.** **(A)** Co-occupancy of SCC, Oct4, and Sox2 on the promoters of *Nanog* and *Oct4*. ChIP analysis of RAD23B occupancy on distal enhancers (enh), proximal promoter (transcription start site, TSS), and upstream (positions indicated by numbers) and downstream intronic regions of the *Nanog* (left), *Oct4* (middle), and *Actb* (right) gene loci. Representative data (n > 5) showing the enrichment of RAD23B (black bars) compared to normal IgGs (white bars) are analyzed by qPCR and expressed as percentage of input chromatin. Schematic diagrams of Oct4- and Sox2-binding sites on the Nanog and Oct4 regulatory regions (TSS and enhancers; see also Figure 2.10A) are indicated at the bottom. Error bars represent standard deviations. n = 3. **(B)** Percent peak overlap between RAD23B and control IgG ChIP-seq data relative to published Oct4/Sox2 and Nanog/Tcf3 peak data. **(C)** Percent base pair overlap between RAD23B and control IgG ChIP-seq data relative to Oct4/Sox2 and Nanog/Tcf3 ChIP-seq data sets. **(D)** Distribution of distance (in base pair) of RAD23B and control IgG peaks from Oct4/Sox2 and Nanog/Tcf3 peaks.

point to a non-redundant function of RAD23B in somatic reprogramming independent of its role in DNA repair as RAD23B KO (and RAD23A KO) MEFs are NER proficient.[65] Importantly, depletion of XPC (knockdown and knockout) and CETN2 in MEFs did not affect proliferation rates when compared to non-target or Oct4 knockdown MEFs. However, RAD23B-depleted MEFs displayed noticeable changes in growth rates, which may partially account for the marked reduction in reprogramming efficiency (data not shown). These data suggest that efficient reprogramming may require SCC/XPC in conjunction with Oct4 and Sox2 to re-establish ES-specific gene expression programs.

Discussion

Establishment of ground state pluripotency in embryonic stem cells represents one of the most remarkable events in development. Stem cells have evolved a subset of cell type-specific activators among a constellation of previously identified transcription factors and cofactors to resolve the dichotomy between self-renewal versus differentiation. Our de novo purification of the SCC/XPC complex as a potent coactivator for Oct4 and Sox2 was unanticipated but may in part reflect the need for stem cells to robustly expand and diversify their transcriptional repertoire while also maintaining genome integrity. Indeed, other NER factors have been shown to participate in transcriptional regulation both at the basal and activated levels. For instance, the general transcription factor TFIIH is a classic example with established roles in both transcription initiation and NER.[66] Interestingly, it has recently been reported that, in HeLa cells, the entire NER complex can be assembled onto promoters of activated genes in an XPC-dependent manner. However, XPC alone is not sufficient, as other NER components appear to be responsible for RA-activated transcription.[55] This finding in Hela cells is distinct from our observation that the XPC-NER (SCC) complex plays a direct and critical role in Nanog transcription in vitro and in ES cells. In our studies, optimal activation of Nanog by Oct4/Sox2 potentiated by SCC requires a second activity present in the Q0.3 fraction. However, preliminary mass spectrometry analyses of the partially purified Q0.3 fraction failed to detect any other XP or NER factors, or factors previously identified to co-purify with Nanog or Oct4 in ES cells[62,63] (data not shown). Therefore, the SCC/XPC complex can potentiate Nanog transcription and likely other Oct4/Sox2-directed promoters in the absence of additional XP and NER factors in vitro. Taken together, these results suggest that the mechanism by which the SCC/XPC complex coactivates transcription in ES cells may be distinct from its function in HeLa cells.

Although XPC plays a critical role in DNA lesion recognition, XPC is not universally required for NER as certain types of bulky DNA lesions (e.g., cholesterol-DNA adducts) can be repaired without XPC.[67] Intriguingly, even though XPC is recruited to gene promoters irrespective of DNA damage signals,[55] the XPC-NER complex is the only factor in the XP family that is dispensable for transcription-coupled repair (TCR).[68] Indeed, our findings suggest that the coactivator and NER duties carried out by SCC are mechanistically distinct processes as SCC can function as part of the transcriptional cofactor apparatus via a direct

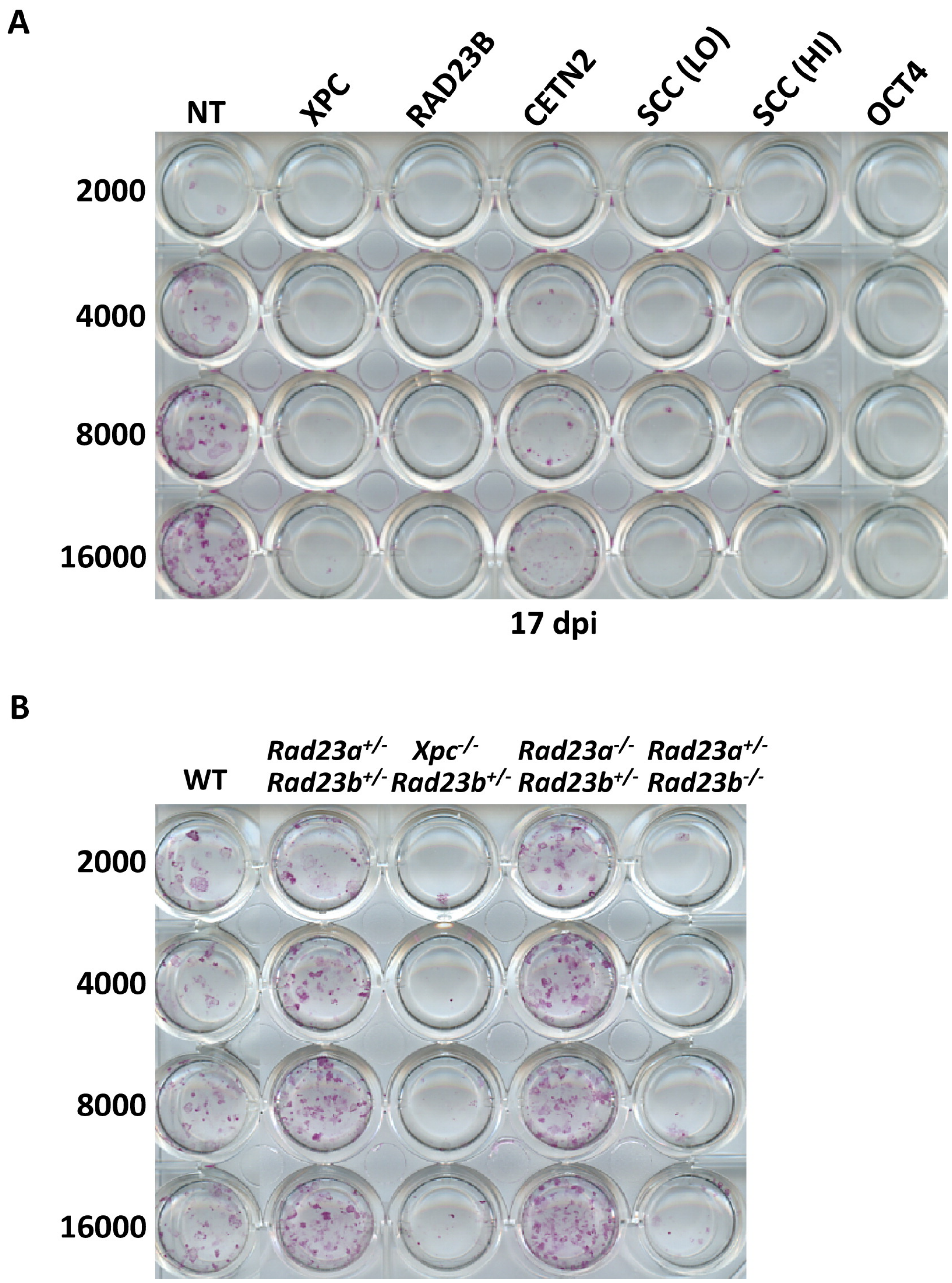

A
NT
XPC
RAD23B
CETN2
SCC (LO)
SCC (HI)
OCT4
2000
4000
8000
16000
17 dpi
B
WT
$Rad23a^{+/-}$ $Rad23b^{+/-}$
$Xpc^{-/-}$ $Rad23b^{+/-}$
$Rad23a^{-/-}$ $Rad23b^{+/-}$
$Rad23a^{+/-}$ $Rad23b^{-/-}$
2000
4000
8000
16000

Figure 2.12: **Somatic Cell Reprogramming Is Blocked by SCC Depletion.** **(A)** Induced MEFs depleted of SCC by shRNAs as described in Figure 2.13B are plated onto 24 well plates at indicated cell numbers. AP-positive (red) colonies are stained 17 days post induction (dpi). **(B)** Induced wild-type (WT), XPC, RAD23A or RAD23B knockout MEFs as described in Figure 2.13C are plated and stained as in **(A)**.

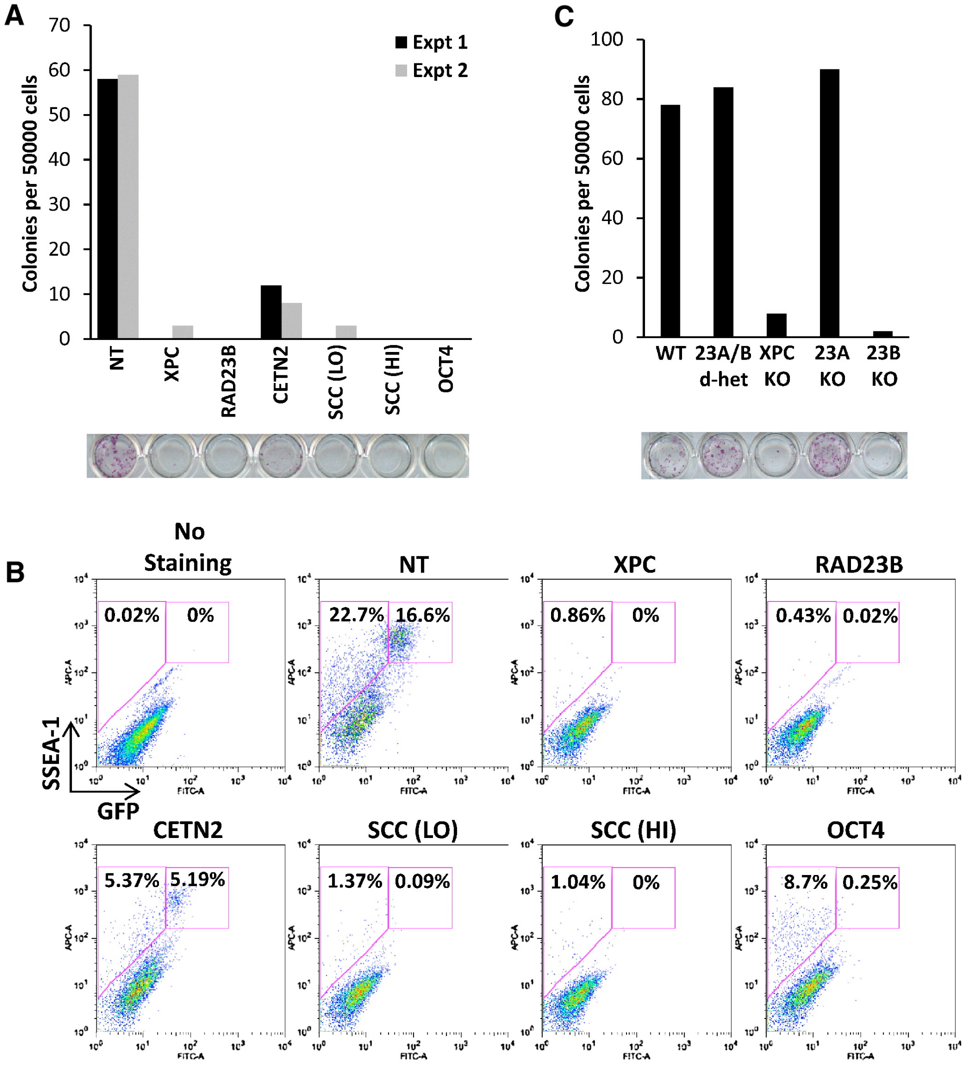


Figure 2.13: **SCC Is Required for Efficient Somatic Cell Reprogramming.** **(A)** Depletion of SCC blocks somatic cell reprogramming. Oct4-GFP mouse embryonic fibroblasts infected with lentiviruses expressing STEMCCA and rtTA together with nontarget shRNA (NT), shRNAs against Oct4, individual subunits of SCC, or all three subunits simultaneously at low or high multiplicity of infection (SCC LO or HI) are plated in 6-well plates for colony counting and FACS or in 24-well plates for AP staining. AP-positive (red) cells are stained and counted 17 days (14 days + dox, 3 days - dox) postinduction (dpi). Results from two separate experiments are shown. **(B)** Single cell suspensions of 17 dpi Oct4-GFP MEFs as described in **(A)** are stained with anti-mouse SSEA-1 antibodies and analyzed by FACS. **(C)** Wild-type (WT), RAD23A, and RAD23B double-heterozygous (23A/B d-Het) MEFs, together with XPC, RAD23A, and RAD23B knockout (KO) MEFs, are induced with STEMCCA. AP-positive colonies are stained and counted as in **(A)**.

interaction with Oct4 and Sox2 without requiring either DNA or TFIIH binding mediated by XPC.

It is worth noting that the effect of single knockdown of XPC or RAD23B was much more pronounced in the reprogramming of MEFs than in the maintenance of ES cells. We surmise that perhaps other redundant regulatory mechanisms in established ES cells can partially compensate for the loss of SCC. Such robust regulatory circuitries are likely to be less developed during the early phase of reprogramming in MEFs and are thus more susceptible to perturbation by SCC depletion. It is conceivable that SCC/XPC may also contribute to the process of chromatin reorganization and facilitate changes in the epigenetic landscape that are conducive to iPS conversion.[55]

Also in agreement with our in vitro and cell-based studies, a mouse double KO of RAD23B and its homolog RAD23A was found to be early embryonic lethal.[65] This previously puzzling phenotype can now be more readily rationalized in light of the functional role of XPC in transcriptional coactivation revealed here. These results, taken together, strongly suggest that loss of the SCC/XPC complex may indeed compromise the transcriptional integrity of pluripotent stem cells, as well as the ability of somatic cells to re-establish pluripotency. However, XPC KO mice are UV-sensitive but otherwise normal with no obvious developmental defects.[69] It has been shown that RAD23B is in vast excess relative to XPC,[70] suggesting that RAD23B may exist in other complexes independent of XPC that functionally replace SCC.

Embryonic stem cells are thought to be under strong selective pressure to maintain genome fidelity since accumulation and propagation of DNA errors to progenitor cells would be lethal during development, therefore, DNA damage response factors and pathways are often up-regulated in ES cells (e.g., XPC, RAD23B, ERCC5, etc).[71,72] Should DNA repair fail, UV-damaged ES cells can be eliminated first by repressing Nanog expression through p53 up-regulation, which in turn promotes spontaneous differentiation and efficient apoptosis.[73] It is interesting to note that upon UV-induced DNA damage in HeLa cells, recruitment of XPC to and expression of non-UV inducible genes are dramatically delayed.[55] This suggests that some sort of redistribution mechanism may redirect XPC from transcription duty at promoter targets to the NER pathway in response to DNA damage. In light of these observations, it is tempting to speculate that redistribution of XPC-RAD23B-CETN2 from Nanog and presumably other Oct4/Sox2-regulated promoters to DNA damage sites may provide an efficient sensing mechanism to perturb stem cell-specific gene expression programs and thus provide a window of opportunity for ES cells to either repair the lesions or commit to differentiation and apoptosis. The SCC/XPC complex may therefore act as a molecular link to couple stem cell-specific gene expression programs and genome surveillance in ES cells.

2.3 Dual functions of TAF7L in adipocyte differentiation

Introduction

Adipose tissue plays a central role in energy homeostasis by acting as a major lipid storage site as well as an important endocrine tissue. Excess food intake vs energy expenditure invariably leads to obesity, an increasingly prevalent condition in modern societies. Obesity is, in turn, tightly associated with an elevated risk of type-2 diabetes, hypertension, cardiovascular disease, and certain cancers, posing serious public health issues and rapidly escalating costs of health care.[74,75]

Accumulation of adipose tissue mass results from increase in both adipocyte size and number. The size of adipocytes depends on the amount of stored lipids, while an increase in adipocyte number (adipogenesis) generally results from the expansion of adult precursor cells and subsequent differentiation, an underlying cause of obesity.[74] Consequently, there is great interest in dissecting the molecular mechanisms regulating adipogenesis and adipose biology.

During the past 20 years, numerous studies have focused on the formation of adipocytes using well-established in vitro cell models.[76,77,78,79,80] The utilization of these model cells identified key adipogenic transcriptional activators such as C/EBP and PPAR.[81,82,77,78] Other transcription factors and cofactors, such as KLF15, KLF9, MED1, MED14, MED15, MED23 and TAF8, have also been reported to be involved in either adipocyte commitment or differentiation in model cell lines.[83,84,85,86,87,88] More recently, genome-wide studies using a variety of techniques (microarray, FAIRE-seq, Quanttrx and mRNA-seq) have been utilized to screen for additional putative pro-adipogenic factors based on changes in either mRNA levels or chromatin states during adipogenesis.[89,90] Specifically, adipocyte-specific expression signatures and genome-wide binding maps of pro-adipogenic activators such as PPAR, C/EBP, and RXR have been primarily determined from 3T3-L1 derived adipocytes. Other factors, including ZFP423, NF1 family proteins, and IRFs have been also implicated in adipogenesis.[91,92,93,90,94]

However, it has become evident that the full complement of key transcriptional regulators that orchestrate adipogenesis remains to be illucidated. The diversity of factors and the mechanisms driving adipocyte formation and cellular differentiation in general remain a challenge. For example, several tissue-specific components of the core transcriptional machinery were recently found to play essential roles in directing cell type-specific programs of transcription and lineage-specific differentiation. The examples of these include TAF4b which was found to be a key component in the development of the mouse ovary and spermatogenesis,[95,96] and TAF3 which was found to be required for mouse germ-layer differentiation.[40]

Additionally, a TRF3/TAF3 complex was found to be important for mouse myogenesis and zebrafsh hematopoiesis.[97,98] As increasing numbers of cell type- and tissue-specific components of the core machinery become better characterized, a somewhat different notion of how traditional sequence–specific enhancer binding factors cooperate with selective tissue-specific core factors to drive lineage-specific transcription programs has emerged.[99] With respect to adipogenesis, the only hint of such a mechanism came from previous studies of TAF8, which was reported as a component of TFIID implicated in adipocyte formation using 3T3-L1 cells.[83] Whether TAF8 operates exclusively as a subunit of TFIID or participates in other molecular transactions, in addition to its in vivo function, remain unknown.

Here we found TAF7L, a paralogue of TATA binding protein associated factor 7, is highly enriched in differentiated C3H10T1/2 adipocytes and bona fide mouse WAT. We have utilized shRNA knockdown and gene knockout strategies to determine the role of TAF7L in adipocyte formation both in vitro and in vivo. In addition, we have explored the consequences of ectopically expressing TAF7L in C2C12 myoblasts to probe its reprogramming capabilities. Further, we carried out genome-wide mRNA-seq and ChIP-seq analysis to survey its functions in adipocyte differentiation. By using a combination of cellular, biochemical, genetic, and genomic approaches, our findings suggest that TAF7L plays an integral role in adipocyte gene expression by targeting enhancers as a cofactor for PPAR and promoters as a component of the core transcriptional machinery, therefore providing new molecular insights into fat development that may prove useful for developing therapeutic strategies to treat obesity and its associated diseases.

Results

Elevated levels of TAF7L in differentiated adipocytes and WAT

To explore the regulatory mechanisms directing adipocyte formation and function, we asked whether there were significant changes to the core promoter recognition components during adipogenesis similar to what had been observed during myogenesis.[97] In particular, we set out to determine whether and which TAF subunits of the prototypic core promoter recognition complex TFIID increase or decrease in expression during adipocyte differentiation. Consistent with the previous observations from 3T3-L1 adipogenesis studies,[83] our analysis of both protein and mRNA levels revealed that TBP and most of the canonical subunits of TFIID are down-regulated during C3H10T1/2 differentiation (Figure 2.14A,B). Surprisingly however, one subunit (TAF7L) previously reported to be a component of TFIID primarily in testis[100,101] was found to be significantly up-regulated in differentiated C3H10T1/2 adipocytes (Figure 2.14A,B and Figure 2.16A) and 3T3-L1 adipocytes (Figure 2.15A,C). Importantly, this enrichment appears specific for the adipogenesis process since the mRNA abundance of *Taf7l* is downregulated to levels comparable to those of other TAF subunits during myogenesis (Figure 2.14C). To exclude the possibility that *Taf7l* enrichment reflects a cell culture artifact of C3H10T1/2 adipogenesis, we compared *Taf7l* mRNA and protein levels in bona fide mouse tissue. In concordance with previous studies, *Taf7l* is most highly expressed in testis[100] (Figure 2.14D,E). Importantly, *Taf7l* also shows significant expression in WAT and detectable expression in liver, spleen, brown adipose tissue (BAT) and kidney, but not in muscle or brain tissue (Figure 2.14D,E). By contrast, the expression of canonical TFIID subunits such as TAF4 is low in both WAT and muscle as expected (Figure 2.14E). Taken together, these data indicate that TAF7L is indeed enriched in differentiated C3H10T1/2 and 3T3-L1 adipocytes and bona fide WAT.

These findings were surprising for several reasons. First, *Taf7l* had only been well documented to be critical for directing spermatogenesis in mice, and *Taf7l*-deficient mice show an impaired male fertility phenotype but no other defects were previously reported.[102,101] Second, although earlier studies of terminal differentiation implicated specific 'atypical TAFs' in, for example, skeletal muscle, ovary and testis formation,[95,97,99] we did not anticipate *Taf7l* as a potential key player in adipogenesis. Instead, based on previous work, we expected that Taf8 would emerge as the 'cell-type specific' TAF involved in adipogenesis.[83] However, we have found *Taf7l* to be up-regulated while Taf8 mRNA is down-regulated upon induction of C3H10T1/2 or 3T3-L1 cells to form adipocytes (Figure 2.14A and Figure 2.151A). To explore this new finding, we set out to investigate the hitherto unrecognized potential role of *Taf7l* in adipogenesis.

TAF7L is required for adipocyte-specific gene expression and differentiation

To assess whether *Taf7l* is required for adipogenesis, we first knocked down TAF7L expression in C3H10T1/2 cells and then induced adipogenesis. shTAF7L and control shGFP sequences were transfected into C3H10T1/2 cells to generate puromycin resistant stable TAF7L knockdown or shGFP control cell lines (Figure 2.17). As shown by Western blot, shTAF7L significantly reduced TAF7L protein levels both pre- and more dramatically post-adipogenesis while levels of canonical TFIID subunits remained largely unaltered in control and TAF7L knockdown pre-adipogenesis cultures (Figure 2.17A, pre-). Consistent with our previous observation in terminally differentiated cells, the protein levels of the canonical TFIID subunits become largely decreased in control post-adipogenesis cultures while cells that have been depleted of TAF7L and therefore blocked from differentiation show high levels of TFIID subunits. As expected, PPAR levels increased in control shGFP cells post induction but showed markedly reduced levels in shTAF7L-treated cells suggesting that TAF7L may directly or indirectly regulate this key adipogenic transcription factor. Our results also suggest that shTAF7L efficiently reduced endogenous TAF7L levels and blocked adipogenesis without significantly affecting TFIID complex integrity (Figure 2.17A, post-). Next, C3H10T1/2 cells stably treated with shTAF7L and control shGFP were subjected to Oil red O staining to determine the efficiency of adipogenesis. Very few lipid-laden adipocytes formed in TAF7L depleted C3H10T1/2 cells, whereas over 98% of shGFP-treated cells differentiated into mature adipocytes. The few adipocytes that formed in shTAF7L-treated cells appeared smaller and exhibited abnormal morphology compared to untreated or shGFP-treated C3H10T1/2

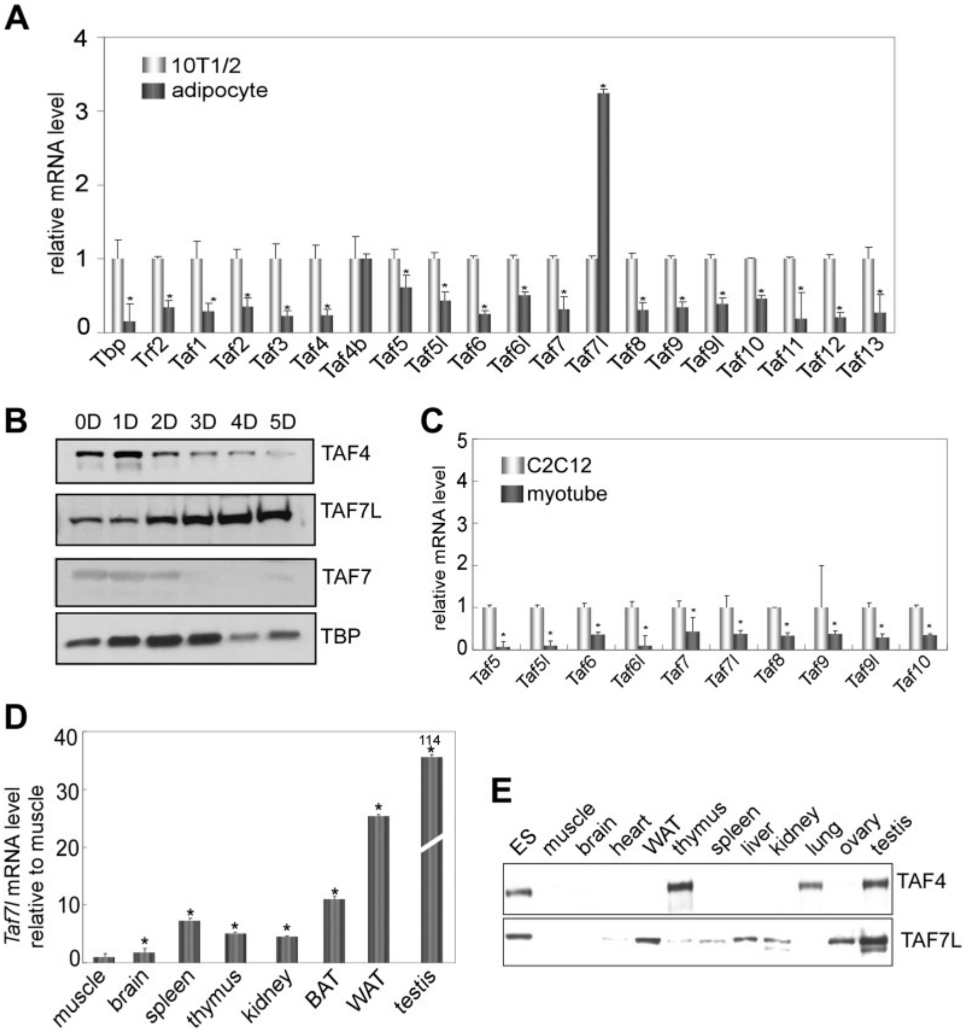


Figure 2.14: **TAF7L is enriched in terminally differentiated adipocytes and bona fide WAT.** (**A**) and (**B**) Expression of TAF7L and TFIID subunits prior to and 5 days (5D) post adipogenic induction of C3H10T1/2 cells as shown by RT-qPCR analysis (**A**) and by Western blot (**B**). (**C**) mRNA levels of TFIID subunits in C2C12 cells and myotubes. (**D**) *Taf7l* mRNA levels in different mouse tissues detected by RT-qPCR relative to muscle, whose expression level was assigned to 1 as the tissue displaying the lowest *Taf7l* mRNA levels. (**E**) Western blot analysis of mouse tissues with TAF4 and TAF7L antibodies. mRNA levels in (**A**) and (**C**) was assigned to 1 in C3H10T1/2 and C2C12 cells, mRNA levels in adipocytes and myotubes were compared with C3H10T1/2 and C2C12 cells respectively. *$p<0.05$, data is mean and s.e.m is from triplicates. RT-qPCR was normalized to the amount of total mRNA and Western blotting analysis was normalized to the amount of total protein. D, days; 10T1/2, C3H10T1/2 cells; ES, embryonic stem cell; BAT, brown adipose tissue; WAT, white adipose tissue.

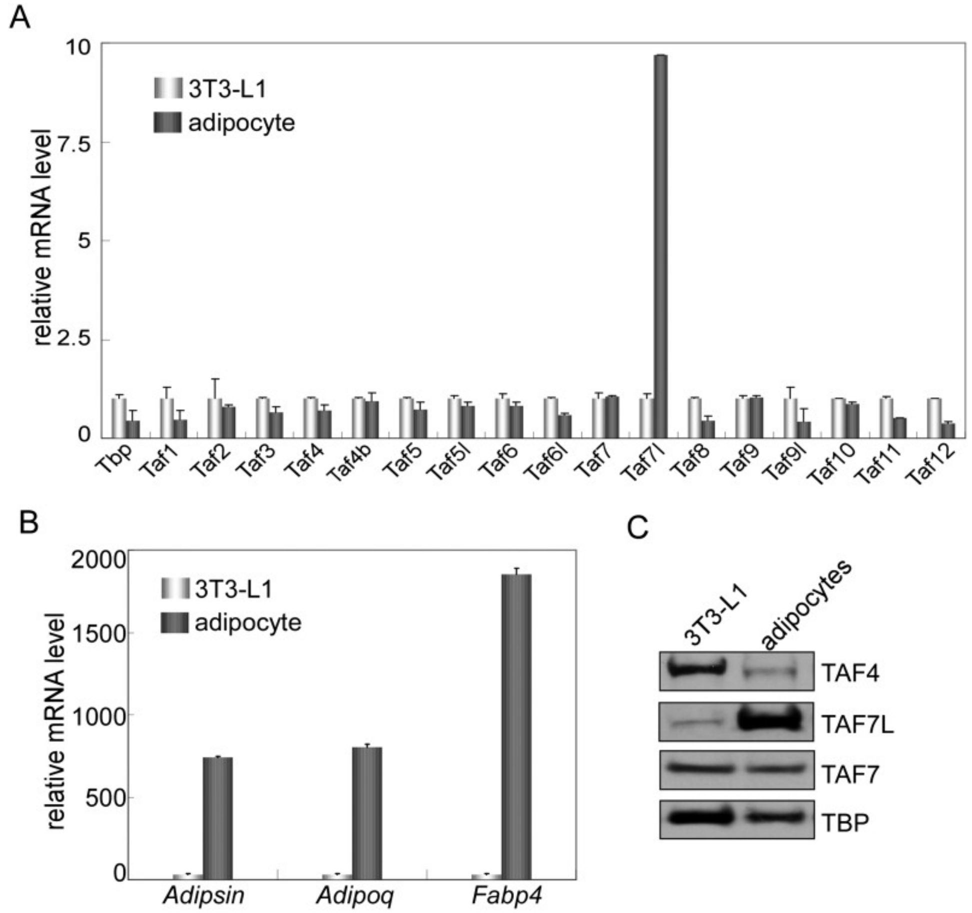


Figure 2.15: **TAF7L is enriched in 3T3-L1 differentiated adipocytes.** (**A**) Expression of*Taf7l* and TFIID subunits prior to and 7 days (7D) post adipogenic induction of 3T3-L1 cells as shown by RT-qPCR analysis (**A**) and by Western blot (**C**). (**B**) Gene expression of adipocyte marker genes*Adipsin* ,*Adipoq* and*Fabp4* of 3T3-L1 adipocytes prior to and 7 days post adipogenic induction. mRNA levels in 3T3-L1 cells were assigned to 1, mRNA levels of each gene in 3T3-L1 adipocytes were compared to 3T3-L1 cells, data is mean from triplicates.

adipocytes (Figure 2.17B,E). These results suggest that TAF7L knockdown largely compromised the adipogenic potential of C3H10T1/2 cells, thus functionally implicating TAF7L in adipocyte differentiation.

To address the possibility that the observed adipogenic defects result from off-target effects of shTAF7L treatment, we carried out 'rescue' experiments by introducing either an empty vector, a shRNA-resistant vector TAF7LmA,[103] or its control paralogue TAF7 into shTAF7L cells. The TAF7LmA expression vector contains two silent mutations in *Taf7l* cDNA which renders resistance to RNA-mediated silencing by shTAF7L. As revealed by Western blot, TAF7LmA is efficiently expressed in shTAF7L cells, similar to the TAF7 construct (Figure 2.17D). Next, we induced adipogenesis followed by Oil red O staining 5 days post induction. Our results indicate that introduction of TAF7LmA restored adipocyte

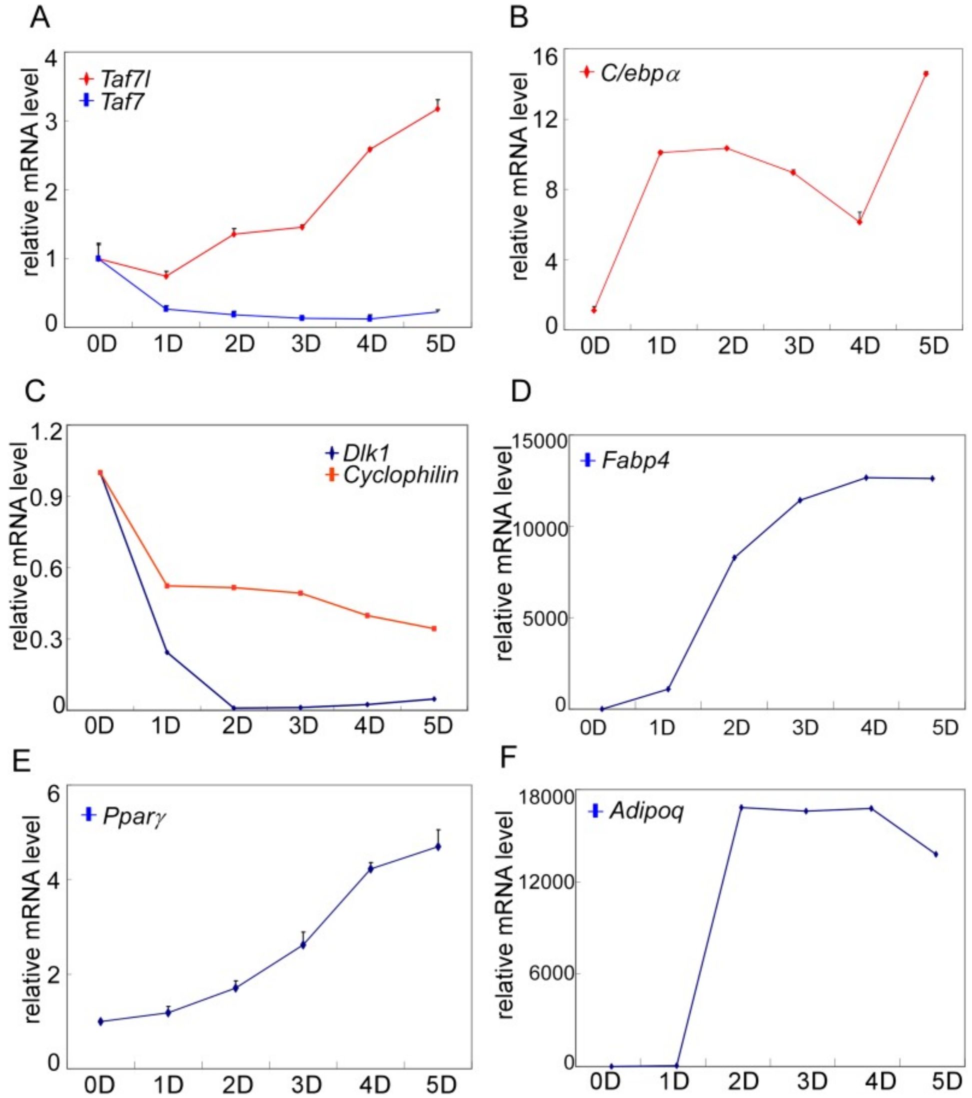


Figure 2.16: **Gene expression analysis of C3H10T1/2 cells during adipogenesis.** (**A**)–(**F**) Time course analysis by RT-qPCR analysis of *Taf7l* and *Taf7* (**A**), *C /ebp* (**B**), *Dlk1* and *Cyclophilin* (**C**), *Fabp4* (**D**), *Ppar* (**E**) and *Adipoq* (**F**) in C3H10T1/2 cells at 0D, 1D, 2D, 3D, 4D and 5D post adipogenic induction. D, days, mRNA levels in C3H10T1/2 cells at 0D were assigned to 1, mRNA levels of each gene at 0D, 1D, 2D, 3D, 4D, and 5D in C3H10T1/2 cells during adipogenesis were compared to 0D respectively, and data is mean from triplicates.

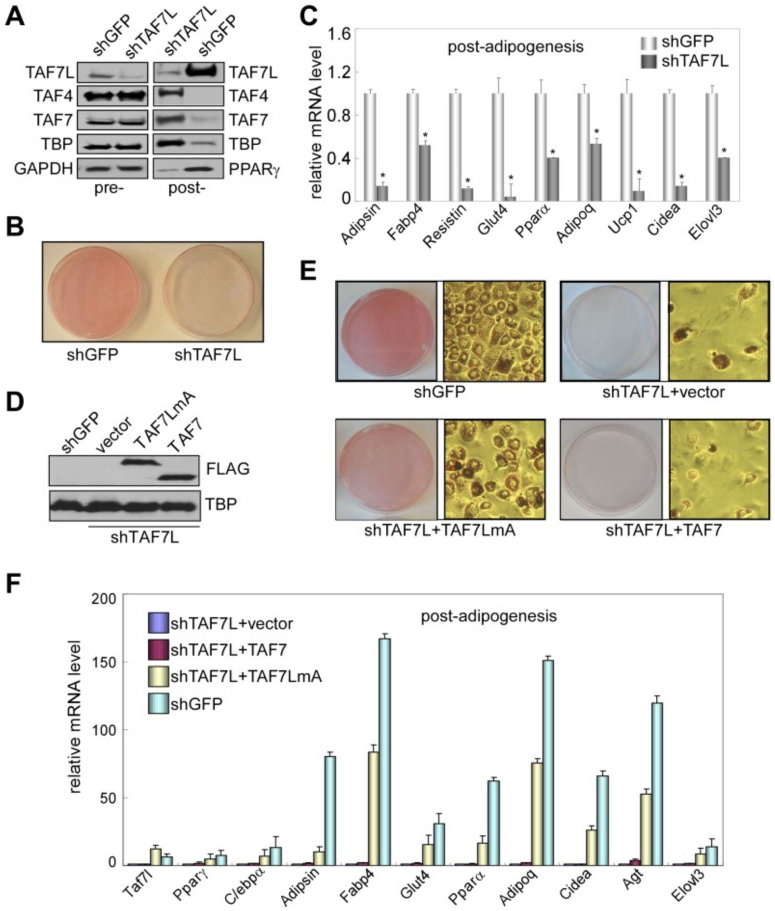


Figure 2.17: **TAF7L is required for adipogenesis in vitro.** (**A**) Western blot of TAF7L, TAF4, TAF7, TBP, GAPDH, and PPAR protein levels in C3H10T1/2 cells expressing shRNA sequence against GFP as control (shGFP) or specifically against TAF7L (shTAF7L) pre- (left panel) and post-differentiation (right panel). GAPDH protein levels serve as a loading control. (**B**) Oil red O staining in 5 days differentiated C3H10T1/2 cells stably expressing either shGFP or shTAF7L. (**C**) mRNA levels of adipocyte-specific genes by RT-qPCR on differentiated shGFP or shTAF7L C3H10T1/2 cells from (**B**), mRNA levels in shGFP cells were assigned to 1, mRNA levels of each gene in shTAF7L cells were compared to shGFP cells, *$p<0.05$, data is mean and s.e.m is from triplicates. RT-qPCR was normalized to the amount of total mRNA. (**D**) Western blot with FLAG and TBP antibodies showing the expression of FLAG-TAF7LmA and FLAG-TAF7 in shTAF7L stably transfected C3H10T1/2 cells, TBP protein levels serve as a loading control. (**E**) Oil red O staining on shGFP or shTAF7L cells ectopically expressing FLAG, FLAG-TAF7 or FLAG-TAF7LmA 5 days post adipogenesis. (**F**) mRNA levels of adipocyte-specific genes by RT-qPCR in differentiated cells from (**E**), mRNA levels in shTAF7L + vector cells were assigned to 1, mRNA levels of each gene in shTAF7L + TAF7, shTAF7L + TAF7LmA and shGFP cells were compared to shTAF7L + vector cells, data is mean from triplicates.

formation (Figure 2.17E) and elevated expression of adipocyte-specific genes compared to vector control in shTAF7L cells (Figure 2.17F). In contrast, overexpression of TAF7 failed to restore the adipogenic defects caused by the loss of TAF7L (Figure 2.17E). Collectively, these results support the notion that TAF7L is likely an important player in adipogenesis, at least in the C3H10T1/2 cell differentiation model.

To identify the full range of genes regulated by TAF7L in adipocyte differentiation, we performed mRNA-seq to profile global gene expression patterns in C3H10T1/2 cells prior to (10T1/2-pre) and after adipogenesis (10T1/2-post) (Figure 2.19A). Next, we verified our mRNA-seq results by single gene RT-qPCR assays for a handful of well-characterized adipocyte-specific genes such as Fabp4, Glut4, Adipsin, Lpl, as well as control genes such as Mef2c and Frzb (data not shown); these results confirmed high concordance between the RT-qPCR assays and the genome-wide mRNA-seq data, although RT-qPCR generally gave 3- to 4-fold higher sensitivity compared to mRNA-seq. We also surveyed a set of 2360 genes upregulated by 10-fold or more in C3H10T1/2 cells post- vs pre-differentiation (Figure 2.19); this analysis identified nearly all the well characterized adipocyte-specific genes including a large proportion of genes involved in adipocyte development and function (Figure 2.19D).

Next, we measured gene expression programs in differentiated C3H10T1/2 cells after depletion of TAF7L (10T1/2-shTAF7L-post). Importantly, genes normally up-regulated following adipogenesis (Figure 2.19A, shown in orange) showed similarly low expression levels in induced shTAF7L cells as in pre-adipocytes (Figure 2.19B, orange). Strikingly, 2083 out of 2360 genes (88%) that are highly upregulated during adipocyte differentiation failed to be induced upon adipogenic induction in the absence of TAF7L (Figure 2.19B,C). RT-qPCR analysis of several representative marker genes confirmed that TAF7L knockdown dramatically reduced their mRNA levels compared to control shGFP in post-adipogenesis cells (Figure 2.17C and Figure 2.16). These data suggest that the induction of adipocyte-specific genes is markedly compromised in the absence of TAF7L. Moreover, the overall transcriptional profile of differentiated shTAF7L-treated C3H10T1/2 cells (shTAF7L-post) largely matched the expression levels of pre-differentiated C3H10T1/2 cells (10T1/2-pre) (R = 0.98) (Figure 2.19B); while mature adipocytes (10T1/2-post) is distinct (R = 0.84) (Figure 2.19A). Thus, loss of TAF7L in C3H10T1/2 cells severely impaired its adipogenic potential rendering shTAF7L-treated cells in an undifferentiated state (Figures 2B and 3A,B). These findings were also confirmed using a different shTAF7L construct and RT-qPCR analysis (data not shown). Taken together, these results strongly implicate TAF7L in potentiating efficient adipogenesis of C3H10T1/2 cells by serving as an important regulator of adipocyte-specific gene expression.

***Taf7l* is required for WAT development in vivo**

To assess *Taf7l* function in adipogenesis in vivo, we first isolated primary adipocyte fibroblasts from WAT of *Taf7l* knockout (KO) mice and littermate controls (WT) and tested their

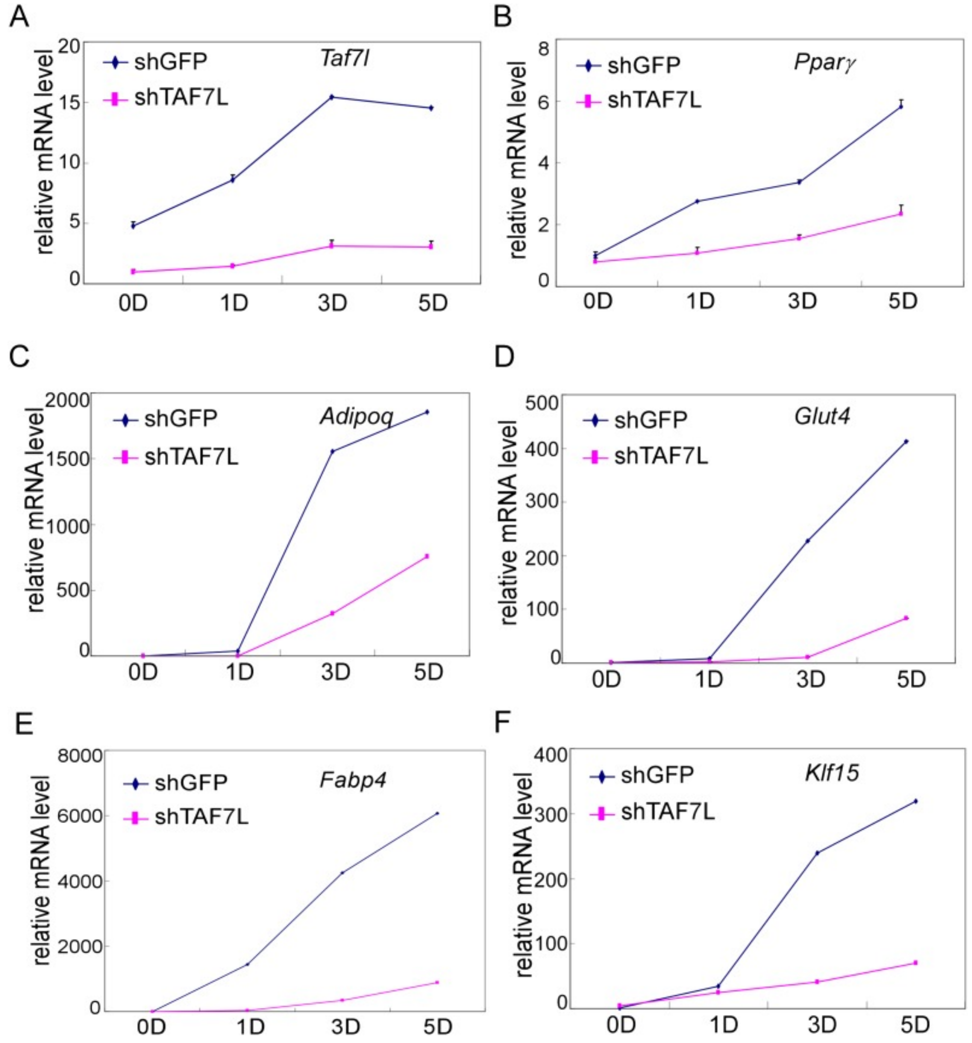
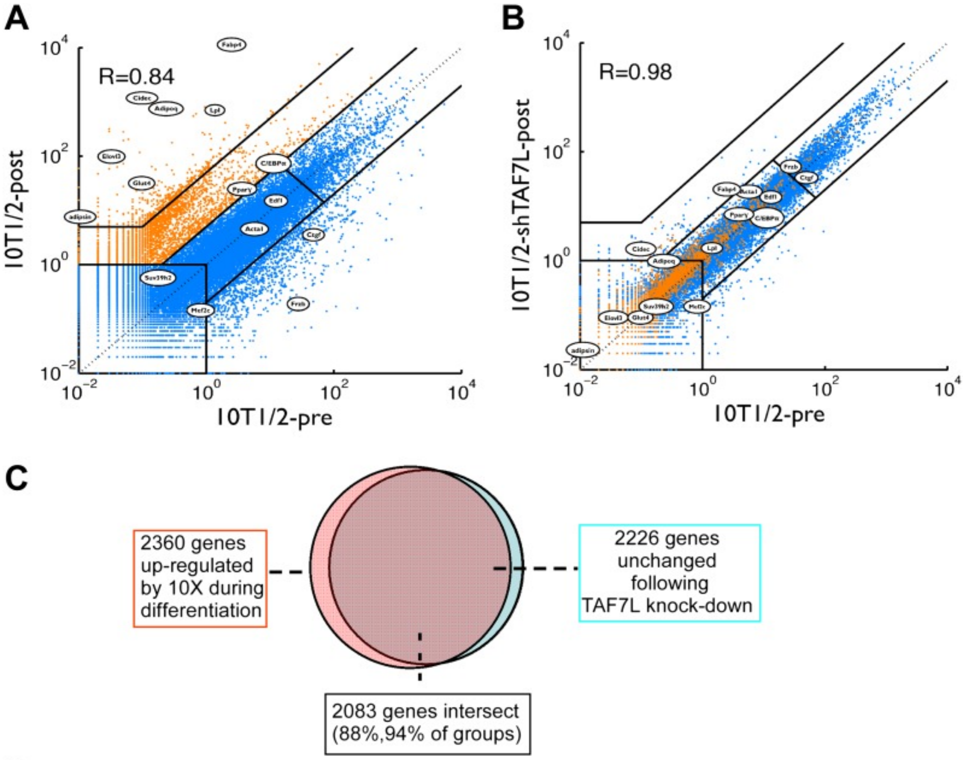


Figure 2.18: **Gene expression analysis after TAF7L knockdown in C3H10T1/2 cells.** (**A**)–(**F**) Time course of gene expression by RT-qPCR analysis of*Taf7l* (**A**),*Ppar* (**B**),*Adipoq* (**C**),*Glut4* (**D**),*Fabp4* (**E**), and*Klf15* (**F**) in C3H10T1/2 cells stably treated with shGFP or shTAF7L sequences at 0D, 1D, 3D, and 5D post adipogenic induction. D, days; shGFP, control cells; shTAF7L, TAF7L knockdown cells. mRNA levels in shTAF7L-treated C3H10T1/2 cells at 0D were assigned to 1, mRNA levels of each gene at 0D, 1D, 3D, and 5D in both shGFP and shTAF7L-treated C3H10T1/2 cells during adipogenesis were compared to mRNA levels in shTAF7L-treated C3H10T1/2 cells at 0D respectively, data is mean from triplicates.



D

gene ontology term	genes name
fat cell differentiation	*Glut4, Pparγ, C/ebpα,C/ebpβ, Klf9, Srebp1, Klf15, Pgc1β, Pparα, Ascl1, Elovl3, Fgf21*
metabolic process	*Lpl, Adipoq, Lipe, Cd36, Agt, Apoc3, Plin1, Pck1, Agpat2, Acox1, Cidec,Cidea, Ucp1 Resistin, Adipsin, Scd1, Fabp4, Adipor2*

Figure 2.19: **TAF7L is required for the expression of adipocyte-specific genes.** (**A**) and (**B**), mRNA-seq data on gene expression of C3H10T1/2 cells pre- (horizontal axis) and post-adipogenesis (vertical axis) (**A**); mRNA-seq data on gene expression in C3H10T1/2 cells pre-adipogenesis (horizontal axis) and C3H10T1/2 treated with shTAF7L post-adipogenesis (vertical axis) (**B**). Orange dots in (**A**) mark genes upregulated during adipogenesis; blue dots in (**A**) mark genes unchanged or downregulated during adipogenesis. Circled genes were tested individually in RT-qPCR analysis. R indicates the correlation of the expression programs between two compared cells (10T1/2-post vs 10T1/2-pre in (**A**), 10T1/2-shTAF7L-post vs 10T1/2-pre in (**B**)). (**C**) TAF7L knockdown blocks the upregulation of the adipocyte-specific genes which occurs during normal adipogenesis, pink circle represents 2360 genes upregulated in 10T1/2-post by 10-fold (10) from (**A**); orange circle represents 2226 genes unchanged in 10T1/2-shTAF7L-post (**B**) compared to (**A**), 2083 genes in the overlapping intersect region account for 88% of total upregulated 10 genes in (**A**). (**D**) List of gene ontology analysis hits showing a few typical adipocyte genes involved in fat cell differentiation and metabolic processes.

ability to undergo adipogenesis. As revealed by Oil red O staining, *Taf7l* KO fibroblasts produced very few, if any, lipid-filled adipocytes compared to WT cells in response to adipogenic induction (Figure 2.20A). RT-qPCR analysis confirmed that ablation of *Taf7l* also suppresses the upregulation of adipocyte-specific genes during differentiation (Figure 2.20B). These results indicate a requirement for *Taf7l* in adipocyte differentiation of primary adipocyte fibroblasts, consistent with our results from C3H10T1/2 cells. As expected, loss of *Taf7l* caused no obvious change in mRNA or protein levels of other TFIID subunits (data not shown), suggesting that deletion of *Taf7l* is unlikely to affect TFIID integrity or function in vivo, in agreement with previous observations that *Taf7l*-deficient mice appear normal except for germ cell developmental defects.[101]

To determine the influence of *Taf7l* KO in early fat development in vivo, we next examined the adipose formation in E18.5 embryos of WT control and *Taf7l* KO littermates. We utilized haematoxylin and eosin (HE) staining on transversal sections of the interscapular region to visualize the overall structure of skin and underlying tissue including subcutaneous fat, connective tissue and muscle. In particular, we used FABP4 antibody staining to localize developing subcutaneous adipose tissue. We observed that FABP4+ lipid-laden cells are significantly diminished in the subcutaneous layer of *Taf7l* KO mice compared to WT littermate controls with a concomittant increase in layers of connective-like tissue under the skin of *Taf7l* KO mice (Figure 2.20E). These results suggest that loss of *Taf7l* impairs WAT development in the later stages of embryogenesis.[104] It also appears that in *Taf7l*-deficient mice, there may be an imbalance in mesodermal-derived lineages as revealed by the appearance of a thicker layer of subcutaneous connective-like tissue.

We next examined the formation of WAT in 1-month and 4-month-old *Taf7l* KO and control WT littermates. The results revealed that 5 out of 12 *Taf7l* KO mice exhibited a noticeable reduction in both their subcutaneous and abdominal white fat pads compared to control WT animals (Figure 2.20F,G), while both groups consumed similar amounts of food and have similar growth curves (Figure 2.20C,D). Taken together, our observations with isolated primary adipose fibroblasts, E18.5 embryos, and fat pads from young mice are consistent with the notion that *Taf7l* likely functions in potentiating mouse WAT development.

Ectopic expression of TAF7L transdifferentiates C2C12 myoblast

A complementary strategy to probe the capacity of transcription factors to influence specific differentiation pathways involves the ‘reprogramming of cell fate’. For instance, transdifferentiation of C2C12 myoblasts into adipocytes by ectopic expression of PPAR and/or C/EBPa under adipogenic permissive conditions helped establish these sequence-specific enhancer binding factors as key regulators of adipogenesis.[105, 106, 81, 107] Therefore, we tested the adipogenic function of TAF7L by an analogous ‘gain of function’ approach with forced introduction of TAF7L or control vector into C2C12 myoblasts. First, we generated TAF7L-expressing (C2C12.TAF7L) or control C2C12 (C2C12.CNTL) stable cell lines by transfecting

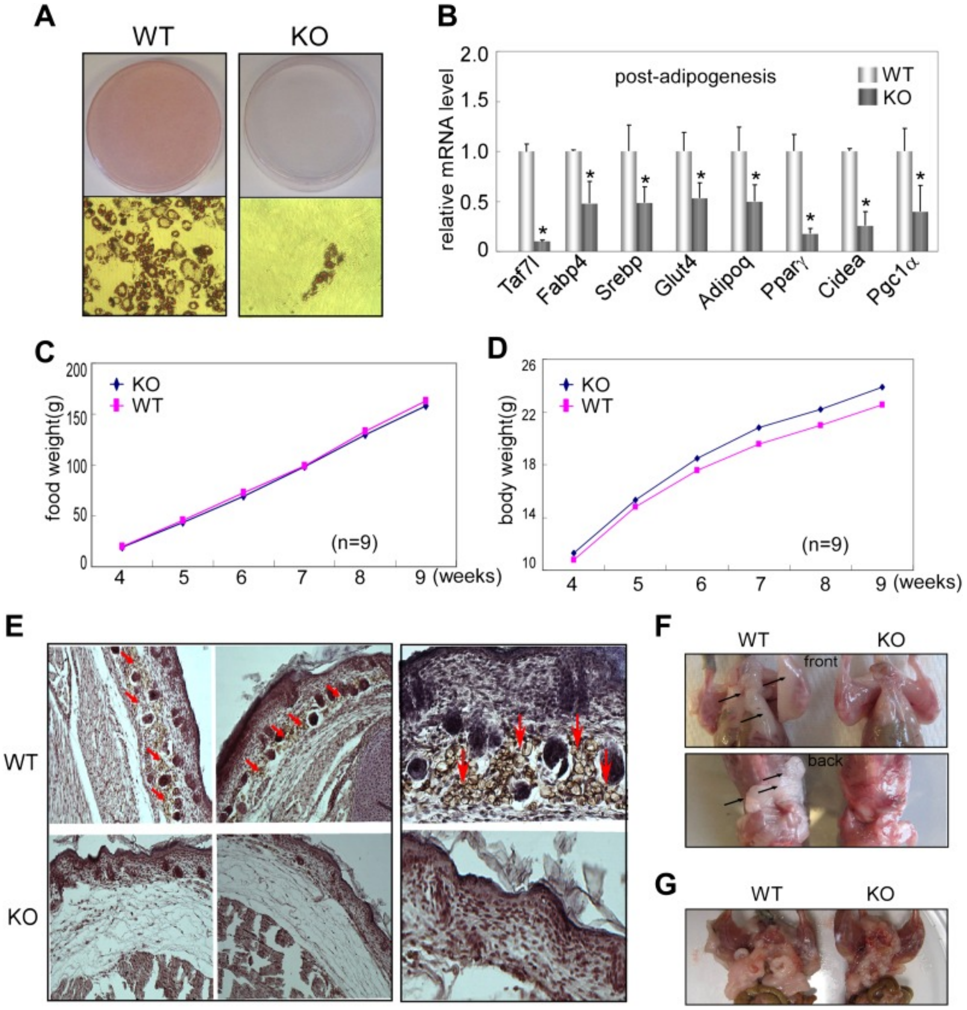
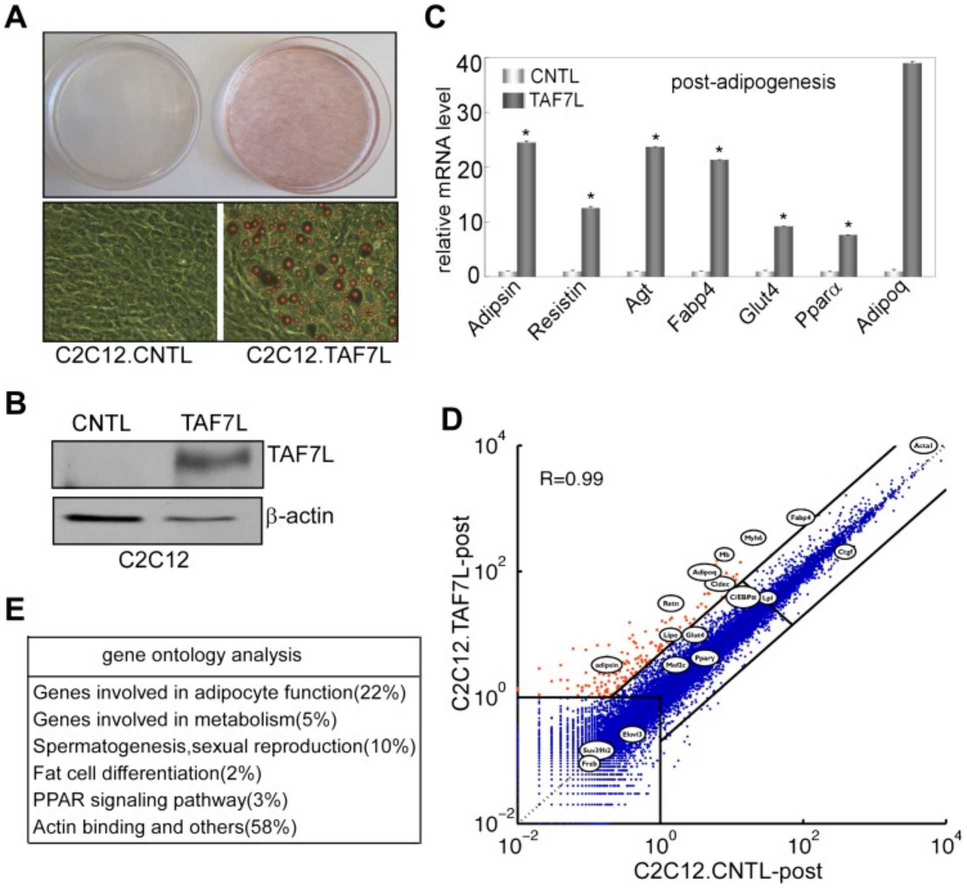


Figure 2.20: ***Taf7l* is required for WAT development in vivo.** (**A**) Oil red O staining to detect mature adipocytes from 5 day differentiated primary fibroblasts derived from adipose tissue of wild-type (WT) and *Taf7l* -deficient mice (KO). (**B**) mRNA levels of adipocyte-specific genes by RT-qPCR on WT and *Taf7l* KO primary fibroblasts post differentiation from (**A**), mRNA levels in WT cells were assigned to 1, mRNA levels of each gene in *Taf7l* KO cells were compared to WT cells, *p$<$0.05, data is mean and s.e.m is from triplicates. RT-qPCR was normalized to the amount of total mRNA. (**C**) Average food intake of WT and KO mice from week 4 to week 9 after birth. n = 9. (**D**) Average body weights of WT and KO littermates from week 4 to week 9 after birth, n = 9. (**E**) HE and FABP4 antibody stain subcutaneous fat cells in E18.5 WT and *Taf7l* KO embryos, left panel magnification, 5; right panel magnification, 20; red arrows indicate fat cells stained by FABP4. (**F**) *Taf7l* KO mice exhibits less fat tissue than WT littermate. Shown are representative photographs of 1-month-old mice with skin removed from both front and back views. (**G**) *Taf7l* KO mouse exhibits less fat formation than WT littermate. Shown is a representative photograph of 4-month-old mouse with skin removal.

either FLAG-TAF7L or empty vector followed by neomycin selection. As detected by FLAG antibody through Western blot analysis, C2C12.TAF7L stable cells achieved modestly elevated expression levels of FLAG-TAF7L protein (Figure 2.21B). Similarly, C2C12.TAF7L cells express roughly eightfold higher *Taf7l* mRNA levels than C2C12.CNTL cells (Figure 2.22). Next, we treated both C2C12.TAF7L and C2C12.CNTL stable cell lines with the four standard adipogenic inducers for 5 days and then applied Oil red O staining. A large proportion of C2C12.TAF7L cells developed into lipid-laden cells while no detectable C2C12.CNTL cells produced lipid droplets (Figure 2.21A). Gene expression analysis by RT-qPCR confirmed that C2C12.TAF7L cells have markedly increased mRNA levels of a subset of adipocyte-specific genes including Adipsin, Resistin, Ppar, C/EBP, Adipoq and Fabp4 relative to C2C12.CNTL cells post differentiation (Figure 2.21C and Figures 2.22B,C,E,F). By contrast, myoblast-gene Myf5 is downregulated in C2C12.TAF7L cells prior to and post adipogenic induction (Figure 2.22D). Furthermore, we performed mRNA-seq on differentiated C2C12.TAF7L and C2C12.CNTL cells and these genome-wide expression studies revealed that indeed, a number of adipocyte-specific genes become highly upregulated in C2C12.TAF7L cells compared to C2C12.CNTL cells post adipogenesis (Figure 2.21D). Gene ontology analysis of genes upregulated fivefold or more in differentiated C2C12.TAF7L cells vs C2C12.CNTL cells indicated that 32% of these genes are involved in either adipocyte differentiation or function. Notably, in accordance with the role of *Taf7l* in spermatogenesis that was reported previously, 10% of these up-regulated genes are involved in spermatogenesis and sexual reproduction (Figure 2.21E). Taken together, these results indicate that ectopic expression of TAF7L in C2C12 myoblasts, even at modestly elevated levels, is capable of inducing upregulation of *Taf7l* itself and other important adipogenic transcription factors including Ppar and C/ebp (Figure 2.221A–C) thereby reprogramming a significant portion of C2C12 cells into adipocytes upon induction (Figure 2.22E,F), providing further evidence for TAF7L as a pro-adipogenic regulator.

TAF7L occupancy at a majority of adipocyte-specific genes

To explore the potential mechanism by which TAF7L functions during adipogenesis, we took advantage of chromatin immunoprecipitation combined with deep sequencing (ChIP-seq) to map TAF7L, TBP, and Pol II binding profiles genome-wide prior to and after adiopgenesis. Mapped ChIP tags were analyzed by intersecting MACS and Grizzly Peak algorithms[9, 108, 8] to identify binding regions for each factor. In agreement with the low concentration of TAF7L shown in Western blots prior to differentiation (Figure 2.14A), only one significant peak was identified in C3H10T1/2 cells compared to 18,672 significant TAF7L binding peaks detected after adipocyte formation. At the same time, we found comparably large numbers of peaks for TBP (12,883 and 14,587) and Pol II (14,502 and 11,424) in pre- and post-differentiation C3H10T1/2 cells. As an example of our TAF7L ChIP-seq data, the profiles of two typical adipocyte-specific genes Adipoq and Klf15, a general highly expressed gene Rfc4, and a nonactive gene Ccdc37 are shown in Figure 2.23A,B. The expression level of each gene before and after differentiation can be deduced from the enrichment levels of



gene ontology analysis

Genes involved in adipocyte function(22%)
Genes involved in metabolism(5%)
Spermatogenesis,sexual reproduction(10%)
Fat cell differentiation(2%)
PPAR signaling pathway(3%)
Actin binding and others(58%)

Figure 2.21: **Ectopic expression of TAF7L transdifferentiates C2C12 myoblasts into adipocytes under adipogenic induction.** (**A**) C2C12 myoblasts expressing empty vector (C2C12.CNTL) or TAF7L (C2C12.TAF7L) were stained with Oil red O 5 days after inducing adipogenesis. (**B**) Western blot analysis on ectopic expression levels of FLAG-TAF7L in C2C12.7L and C2C12.CNTL cells, -actin protein level is served as a loading control. CNTL, C2C12.CNTL; TAF7L, C2C12.TAF7L. (**C**) mRNA levels of adipocyte marker genes are measured by RT-qPCR in C2C12.TAF7L cells compared with C2C12.CNTL cells 5 days post adipogenesis, mRNA levels of genes in C2C12.CNTL cells were assigned to 1. *p<0.05, data is mean and s.e.m is from triplicates. RT-qPCR was normalized to the amount of total mRNA. (**D**) mRNA-seq analyzes genes activated by TAF7L in C2C12.TAF7L compared to C2C12.CNTL post adipogenesis. Red dots represent genes upregulated in C2C12.TAF7L-post cells; blue dots represent genes unaltered or downregulated in C2C12.TAF7L-post cells compared to C2C12.CNTL-post cells after adipogenic induction. (**E**) Major gene functional groups from genes activated above fivefold by TAF7L in C2C12 cells post adipogenic induction through gene ontology analysis.

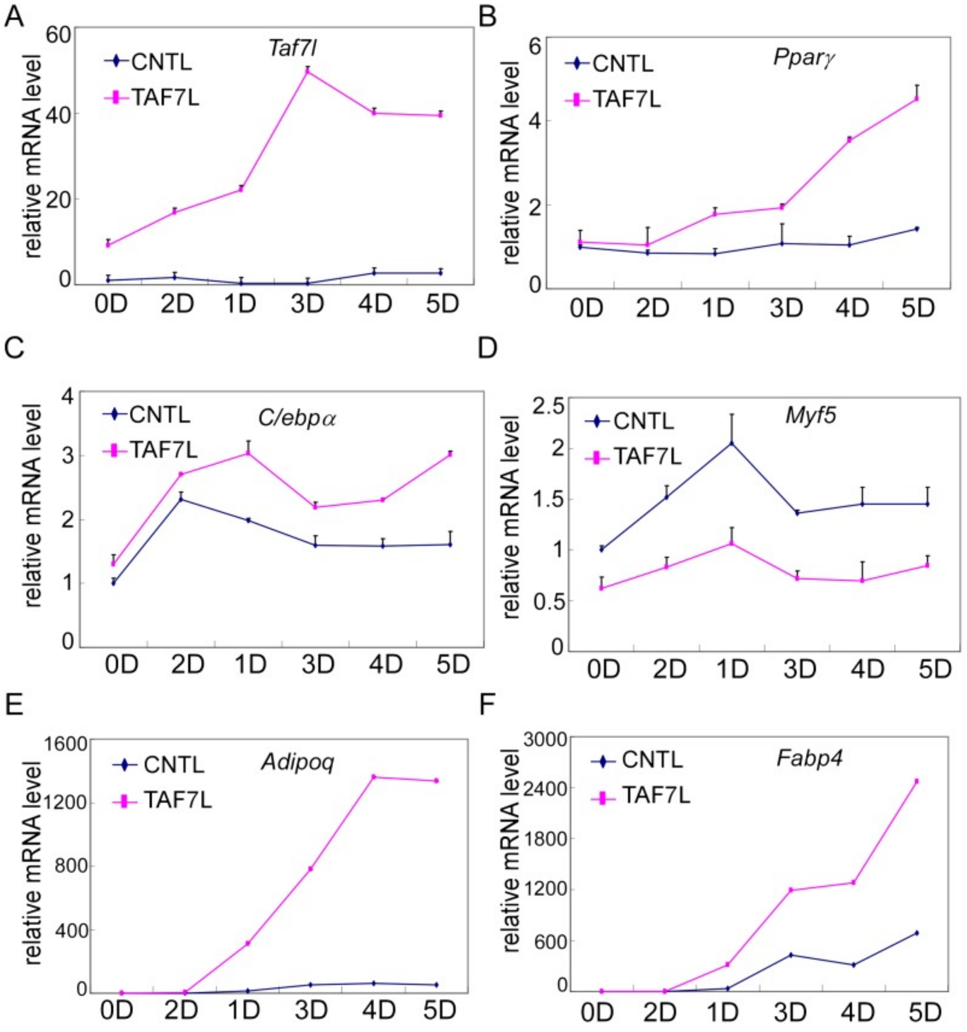


Figure 2.22: **Gene expression analysis of TAF7L-expressing C2C12 cells.** (**A**)–(**F**) Time course of gene expression by RT-qPCR analysis of *Taf7l* (**A**),*Ppar* (**B**),*C/ebp* (**C**),*Myf5* (**D**),*Adipoq* (**E**), and*Fabp4* (**F**) in C2C12.CNTL and C2C12.TAF7L cells at 0D, 1D, 2D, 3D, 4D and 5D post adipogenic induction. D, days; CNTL, C2C12.CNTL; TAF7L, C2C12.TAF7L. mRNA levels in C2C12.CNTL cells at 0D were assigned to 1, mRNA levels of each gene at 0D, 1D, 2D, 3D, 4D, and 5D in both C2C12.CNTL and C2C12.TAF7L cells during adipogenesis were compared to mRNA levels in C2C12.CNTL cells at 0D respectively, data is mean from triplicates.

Pol II.

We split the genome into three groups representing unchanged, downregulated, and upregulated (>5, >50) genes based on their expression pattern and levels before and after differentiation based on mRNA-seq data. Analysis of our ChIP-seq and mRNA-seq data suggests that among all active genes in adipocytes, TAF7L binds to >65% of adipocyte-specific genes that are highly upregulated (>50) during adipogenesis while binding much less frequently at both core promoters (<500 bp from TSS) and proximal enhancer regions (500 bp to 5 kb from TSS) near genes unaltered or downregulated during differentiation (Figure 2.23C). Next we compared the expression levels of genes in adipocytes to the binding intensity of TAF7L and found a strong positive correlation in the three expression groups. For instance, examination of 3,468 TAF7L peaks, representing 20% of the total peaks located at proximal enhancers revealed that the average TAF7L binding strength on genes upregulated >50 is twice as strong as on genes induced between 5–50 in adipocytes relative to C3H10T1/2 cells, and eight times stronger than on genes down-regulated following adipogenesis (Figure 2.23D), suggesting that TAF7L binding frequency and strength is highly correlated with upregulated genes during adipogenesis.

By mapping the genomic binding sites, we found that both TBP and Pol II display greater than 63% occupancy at transcriptional start sites (TSS) of highly expressed genes in both undifferentiated C3H10T1/2 cells and adipocytes, consistent with their roles in mediating global and general transcription functions. Interestingly, comparing the binding regions of TAF7L, TBP and Pol II in adipocytes revealed that TAF7L only partially (30%) colocalizes with TBP and Pol II at a subset of promoters, while a greater proportion (45%) of TAF7L peaks localizes to enhancer regions where TBP and Pol II are generally not found (Figure 2.24B,C). This surprising finding suggests that TAF7L may function via additional mechanisms other than as a subunit of canonical TFIID in regulating adipocyte differentiation.

TAF7L co-localizes with PPAR genome-wide

To determine mechanisms by which TAF7L operates via its occupancy at enhancers, we probed its potential association with adipocyte-specific enhancer binding transcription factors. First, we applied an unbiased genome-wide approach to identify sequence-specific transcription factors that could interact with TAF7L. We analyzed the sequences surrounding TAF7L binding sites and identified several DNA consensus sequence motifs enriched in TAF7L peaks. Next, we compared these motifs with all known sequence-specific recognition elements of transcription factors, which led to the identification of several binding motifs of well-known adipogenic transcription factors including PPAR and C/EBP[109] (Figure 2.24A).

We then used ChIP-seq to directly map the genome-wide binding profiles of PPAR and detected 4,121 and 12,893 significant peaks for PPAR in C3H10T1/2 cells pre- and post-

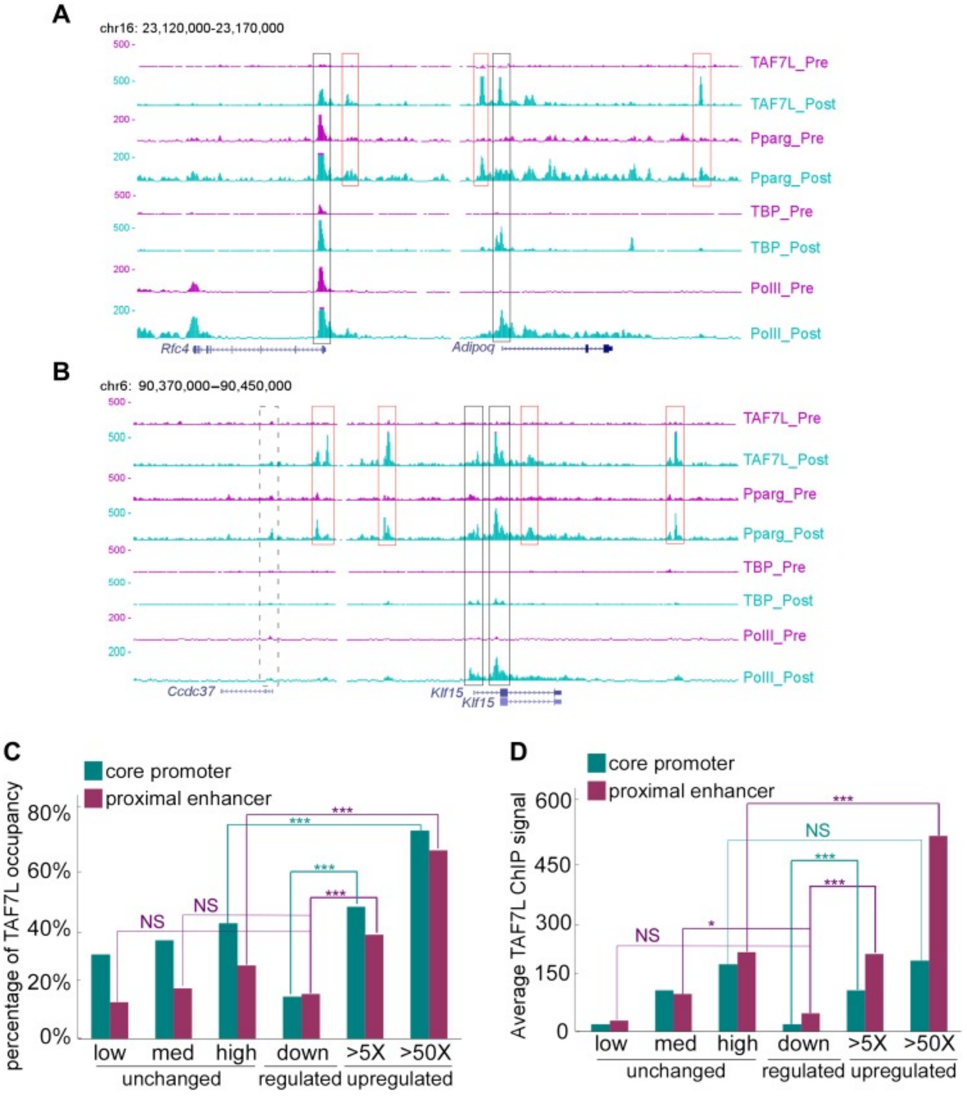
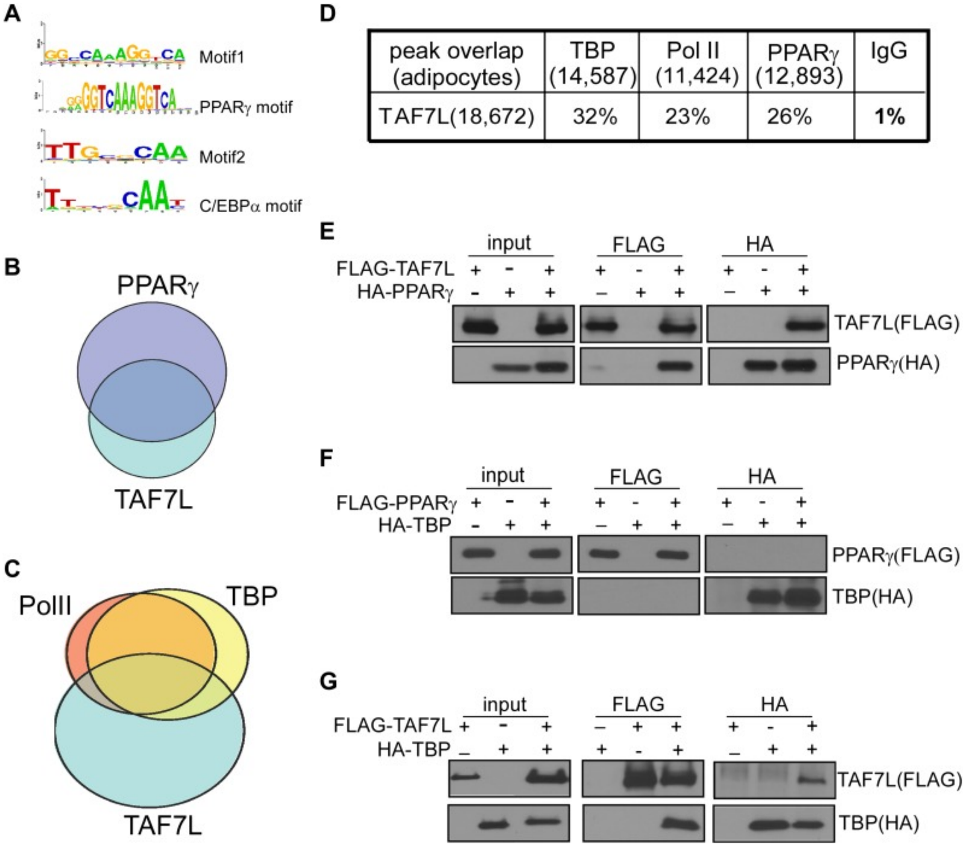


Figure 2.23: **TAF7L binds strongly on the majority of genes upregulated during adipogenesis.** (**A**) Read accumulation for eight ChIP-seq datasets including TAF7L, PPAR, TBP and Pol II before ($_{pre}$) and after ($_{post}$) adipocyte differentiation at the*Rfc4* and*Adipoq* gene loci. (**B**) The same as in (**A**) at the*Ccdc37* and*Klf15* gene loci. Vertical axis is 0–500 reads for all factors, co-localized peaks were marked with boxes, black boxes indicate promoters and red boxes indicate enhancers, solid lines denote active genes and dashed lines denote inactive gene. (**C**) Frequency (vertical axis) of TAF7L occupancy on gene expression groups (horizontal axis) including unchanged (low, med, high) (three blue dots regions from left-bottom to right-top inFigure 2.19A), downregulated (blue dots in left-bottom region inFigure 2.19A), and upregulated (>5, >50, two orange dots regions from lower to higher in Figure 2.19A). (**D**) Average TAF7L binding signal strength (vertical axis) on the core promoters (500 bp from TSS) and proximal enhancers (500 bp to 5 kb from TSS) of three major gene expression groups as in (**C**). (Regular t-test for (**C**) and (**D**), NS is no significant, *p<0.05, ***p<0.001).



peak overlap (adipocytes)	TBP (14,587)	Pol II (11,424)	PPARγ (12,893)	IgG
TAF7L(18,672)	32%	23%	26%	**1%**

Figure 2.24: **TAF7L colocalizes and associates with PPAR and TBP.** (**A**) Two top motifs (motif 1 and motif 2) were found in TAF7L binding sites. Motif1 p<2e-20) matches with PPAR binding motif and motif 2 p<3e-10) matches with C/EBP binding motif. (**B**) Overlap of PPAR peaks with TAF7L peaks in adipocytes, each circle represents the total peaks from ChIP-seq for a factor and the overlapped region represents the common binding peaks of the factors. (**C**) Similar as in (**B**); Pol II, TBP and TAF7L peaks from ChIP-seq overlap with each other in adipocytes. (**D**) Table showed the total peak numbers of each factor in adipocytes from ChIP-seq and the percentage of genome-wide peak overlapping between TAF7L and PPAR, Pol II, TBP, IgG control. (**E**) FLAG tagged TAF7L, HA tagged PPAR were overexpressed in 293T cells, immunoprecipitations were performed on both FLAG and HA antibodies and followed by Western blotting with FLAG and HA antibodies. (**F**) The same procedures were performed on FLAG tagged PPAR and HA tagged TBP. (**G**) The same procedures were performed on FLAG tagged TAF7L and HA tagged TBP as in (**E**).

differentiation, respectively. We then validated our ChIP-seq data by comparing our new data with published ChIP-seq data for PPAR, C/EBP, and RXR in 3T3-L1 derived adipocytes 6 days post-differentiation.[92, 93, 90] This extensive analysis revealed that the majority of PPAR binding sites we mapped overlap with those previously reported although some differences in specific binding sites were observed, likely due to inherent differences between 3T3-L1 and C3H10T1/2 cells. A direct comparison of TAF7L and PPAR genome-wide binding profiles revealed that 26% of TAF7L peaks were co-occupied by PPAR and reciprocally, 37% of PPAR peaks were co-bound by TAF7L (Figure 2.24B,D). Similarly, TAF7L binds to 20% of RXR bound loci genome-wide. Moreover, TAF7L also co-localizes with 25% of C/EBP binding and vice versa. Collectively, these findings provide indirect evidence for a functional association between TAF7L and adipogenic activators PPAR, RXR and C/EBP in adipocytes. Indeed, the relationship between TAF7L, PPAR, C/EBP and TBP/Pol II at a genome-wide scale suggests that TAF7L might also functions as an enhancer-associated co-activator that connects adipogenic activators with the core promoter recognition machinery to potentiate adipocyte differentiation.

TAF7L interacts with PPAR and TBP/TFIID

Prompted by the extensive co-localization between TAF7L, TBP and PPAR, we set out to examine the potential physical association between these factors. First, we constructed TAF7L, TBP and PPAR with either FLAG or HA Tags for expression in 293T cells in pair-wise combinations. Next, we performed co-immunoprecipitations (co-IP) between PPAR, TAF7L and TBP with either FLAG or HA antibodies, followed by Western blot analysis to detect the FLAG- or HA-tagged proteins. Intriguingly, these co-IP assays showed that PPAR can efficiently pull down TAF7L and vice versa with (data not shown) or without the addition of PPAR ligand rosiglitozone (Figure 2.24E); by contrast, PPAR was unable to co-IP TBP (Figure 2.24F). As expected, TAF7L and TBP can pull down each other reciprocally (Figure 2.24G), which is consistent with previous observations[100]) To confirm that the association between TAF7L and PPAR or TBP is direct and not mediated via DNA/chromatin interactions, we included benzonase treatment in our co-IP assays. Eliminating DNA in these co-IP experiments did not alter the binding interactions we observed between TAF7L and PPAR or TBP (data now shown). As expected, over-expression of TAF7L in C3H10T1/2 cells and adipocytes allows co-IP of other endogenous TFIID subunits including TBP and TAF4 (data not shown), suggesting that some tagged-TAF7L can integrate into native TFIID complexes. Taken together, these protein:protein binding assays suggest that TAF7L can physically associate with both PPAR and TBP/TFIID either directly or indirectly via presently unidentified protein factors. These studies provide a potential mechanism by which TAF7L may serve as a cofactor linking specific adipogenic activators, proximal enhancers and the core transcription apparatus.

Discussion

It is well-documented that the adult human body contains cells residing in the adipose tissue, referred to as Adipose-derived Stem Cells (ASCs).[110,111,112,113] ASCs resemble mesenchymal stem cells (MSCs) in terms of their ability to differentiate into multiple lineages including adipocytes, myotubes, osteocytes, and cartilage under appropriate developmental cues[114]). Given that increased numbers of adipocytes, a major underlying cause of obesity, are primarily derived from MSCs and/or ASCs[115]), we chose C3H10T1/2 MSCs as our cell culture model system for studying adipogenesis in large measure because MSCs efficiently recapitulate aspects of adipocyte differentiation and in vivo fat development. Using this MSC culture model as well as *Taf7l* KO mouse model for our in vitro and in vivo studies, we unexpectedly identified *Taf7l* as a key regulator of adipogenesis; adding a new piece of the molecular puzzle to the critically important regulators of fat development in mammalian organisms. We found the effect of *Taf7l* in adipogenesis to be quite robust wherein its loss led to extensive down-regulation of genome-wide adipocyte-specific gene expression in cell culture and defects in WAT development in vivo. We are particularly intrigued by the manner in which TAF7L seems to operate–serving both as an integral component of TFIID at the core promoter and as a key co-activator interacting directly with PPAR or other adipocyte-specific transcriptional factors (ATFs) at proximal enhancers of adipocyte-specific genes on genome-wide scale (Figure 2.25). Thus, a hitherto unrecognized cell-type selective core regulator with an apparent dual mechanism of action has been identified that influences the pro-adipogenic transcriptional control network. It is conceivable that TAF7L and associated regulatory factors in this newly discovered pathway may reveal potentially useful therapeutic drug targets to combat obesity and its related diseases.

It came as a surprise to find *Taf7l* playing such an essential role in adipogenesis because previous studies had primarily reported a specific role of *Taf7l* in spermatogenesis.[116,117,118] Indeed, even after being clued into the potential contribution of *Taf7l* in fat tissue development, our analysis of *Taf7l* KO mice mainly revealed defects in WAT formation at certain stages of development and there were no gross, easily-observable morphological abnormalities in young animals except when the underlying fat pads were dissected for direct inspection to reveal the partial penetrance of the lean phenotype. Also, previous studies had no reason to examine the expression of *Taf7l* in adipose tissue because *Taf7l* mRNA and protein levels are relatively low in adipose tissue compared with testis. We also note that although TAF7L protein levels become highly elevated when MSCs differentiate into adipocytes, its mRNA levels increase only modestly (2) compared to typical adipocyte marker genes such as *Fabp4* and *Glut4* during adipogensis.[119,120] It is therefore not surprising that a role for *Taf7l* in adipocytes could have been overlooked in previous studies. This cautionary tale also suggests that other key transcriptional regulators could likewise go undetected; suggesting that more studies will likely be required to take a fuller accounting of the multiple factor combinations that have evolved to orchestrate the diversity of gene regulatory pathways during differentiation and development of metazoans.

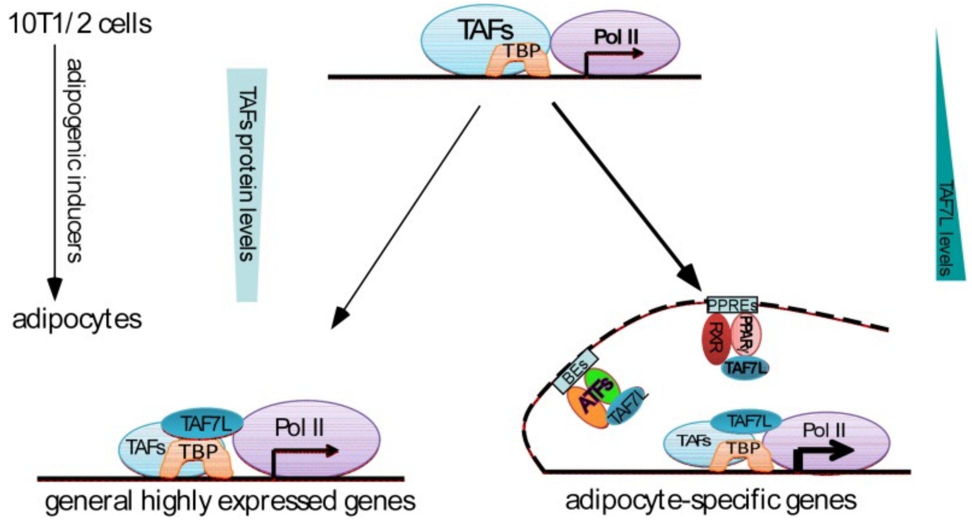


Figure 2.25: **Dual functions of TAF7L in adipocyte differentiation.** TAF7L expression is enriched during C3H10T1/2 MSCs adipocyte differentiation while other TFIID subunits (TAFs) decrease in expression. TAF7L regulates adipogenesis by associating with TBP as a component of adipocyte TFIID complex at promoters and with PPAR or other adipocyte transcriptional factors (ATFs) as a cofactor at enhancers on adipocyte-specific genes, providing the mechanisms of its dual roles during differentiation. General highly-expressed genes are those with high expression before and after adipocyte differentiation include a portion of housekeeping genes; adipocyte-specific genes are those required for adipocyte differentiation and highly upregulated during adipocyte differentiation. TAFs,TBP-associated factors; ATFs, adipocyte transcriptional factors; BEs, binding elements.

Although spermatogenesis and adipogenesis represent two biologically distinct differentiation processes, their common dependence on *Taf7l* may imply some underlying and perhaps hidden commonality. For example, one can speculate that energy storage and metabolic homeostasis are tightly related/linked to reproductive abilities since both are highly sensitive to and dependent on nutrient availability.[116,118] Another possible link involved steroid hormones (i.e. testosterone), which are essential for spermatogenesis, are derived from cholesterol.[121,122] In any case, our genome-wide mRNA-seq expression analysis unambiguously identified *Taf7l*-mediated enrichment of genes involved in both adipogenesis and reproductive processes. However, further investigation will be required to assess whether there are more direct links between adipogenesis and spermatogenesis and whether mutant *Taf7l* involved in male infertility also disturbs normal fat development and metabolism.

Our data clearly showed that TAF7L binds to and regulates the expression of a subset of white adipose genes, interestingly we also observed downregulation of some brown adipocyte genes (*Cidea*, *Ucp1* and *Elovl3*) caused by TAF7L knockdown (Figure 2.17C).[107,123] However, our incipient attempts failed to pinpoint significant defects in interscapular BAT formation in *Taf7l* KO mice. Thus, future studies will be required to determine whether *Taf7l* is also involved in BAT development. Further, our preliminary attempts to measure metabolic functions suggest that *Taf7l*-deficient mice showed changed serum glucose levels after 24 hr fasting compared with control WT littermates, it will be interesting to explore whether and how *Taf7l* KO alters energy metabolism. Therefore, further detailed investigations will be required to more rigorously delineate the potential function of *Taf7l* in energy homeostasis.

2.4 *Taf7l* cooperates with *Trf2* to regulate spermiogenesis

Introduction

Spermatogenesis is a cyclic process in which diploid spermatogonia differentiate into mature haploid spermatozoa. This process is mainly driven by two, pre- and post-meiotic, transcription waves that are tightly controlled by testis-specific transcription factors. During the pre-meiotic transcription phase, individual spermatogonia are committed to differentiating into primary spermatocytes that later undergo two meiotic divisions to generate haploid round spermatids connected by intercellular cytoplasmic bridges.[124,125,126] During the post-meiotic transcription phase of spermiogenesis, haploid round spermatids are sculptured into the elongated shape of mature spermatozoa. These latter stages are accompanied by dramatic biochemical and morphological changes including major remodeling of chromatin with protamines substituting for somatic histones to tightly pack DNA into the sperm nucleus.

Understanding the intricate mechanisms that control spermatogenesis has important implications for human health and reproduction. A key step in the regulation of spermatogenesis occurs at the level of transcription starting with the usage of distinct promoter elements[127] within uniquely reorganized chromatin[128] and driven by the action of several testis-specific transcription factors including CREM[129,130] and the core promoter recognition factors required for global or gene-specific transcription such as TFIIA/ALF (a paralogue of TFIIA),[131] Taf4b (a homologue of *Taf4*),[132] *Trf2*[133,134] and *Taf7l*.[101,100] For example, mice bearing mutant or deficient CREM showed decreased post-meiotic gene expression and defective spermiogenesis.[135] Mice deficient in *Taf4b*, a testis-specific homologue of TBP-associated factor 4 (*Taf4*) are initially normal but undergo progressive germ cell loss and become infertile by 3 months of age with seminiferous tubules devoid of germ cells. *Taf4b* depletion blunted the expression of spermatogonial stem cell genes, indicating a critical role in maintenance of spermatogonial stem cells[132]). The core promoter recognition factor *Trf2* is highly expressed in a finely regulated pattern in the mouse testis during spermatogenesis and mice lacking *Trf2* are viable but sterile due to a complete arrest of late spermiogenesis with largely normal spermatogonia and spermatocytes.[38,136] *Taf7l*, originally identified in spermatogonia, is an testis- and adipose-specific X-chromosome gene.[137] This orphan Taf is expressed throughout male germ-cell differentiation, while its intracellular localization is dynamically regulated from cytoplasm in spermatogonia to nucleus in late spermatogenesis.[100] Early studies of *Taf7l* deficient mice showed reduced fertility, abnormal sperm structure, low sperm count and weakened motility.[101] Although the spermatogenic defects of various mouse lines mentioned above have been extensively studied, the molecular mechanisms controlling testis-specific gene expression programs remain poorly understood. It also remained unclear what, if any, functional relationship exists between these testis-specific transcription factors and potential crosstalk between these various regulators has remained elusive.[38]

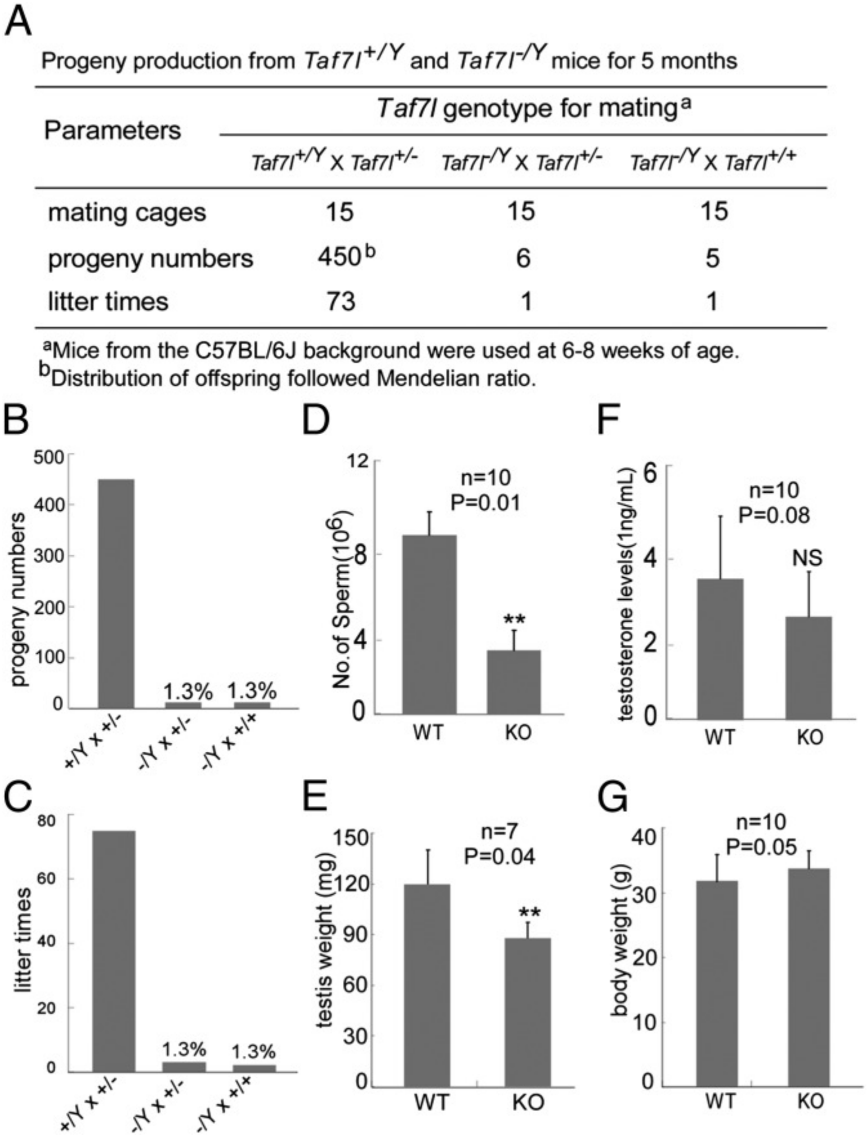
Here we report that backcrossed $Taf7l^{-/Y}$ males from *N5* to *N9* leads to infertility and $Taf7l^{-/Y}$ testes show obvious deficiencies during spermiogenesis. The more severe infertility phenotypes observed in this study relative to previous reports suggest that additional backcrossing of $Taf7l^{-/Y}$ mice may have uncovered a fuller range of *Taf7l* functions in spermatogenesis. Perhaps equally importantly, recent studies of human oligozoospermia patients have found mutations in human *Taf7l*,[102,138] providing a potential link between the critical role of *Taf7l* in mouse spermatogenesis to human infertility. These new findings give further impetus to more fully dissect the underlying molecular mechanisms by which *Taf7l* regulates spermatogenesis. In this report we explore in greater depth how *Taf7l* functions to regulate the differentiation of germ cells. We carried out genome-wide expression profiling and direct binding studies with *Taf7l* and identified many spermiogenesis-specific gene promoters targeted by TAF7L. Interestingly, *Taf7l* impairs spermatogenesis at a similar stage (spermiogenesis) to *Trf2* and is translocated into the nucleus when *Trf2* is highly expressed.[139,100] Importantly, we find that *Taf7l* regulates many known Trf2-targeted testis genes[134,136] and biochemical studies reveal that *Taf7l* interacts with *Trf2* in vitro and in testis by co-immunoprecipitation analysis. Taken together these data suggest that *Taf7l* might cooperate with *Trf2* to control transcription in the post-meiotic stage of spermatogenesis thus providing an important example of functional crosstalk between two atypical core promoter recognition factors operating coordinately to direct tissue-specific gene transcription.

Results

Taf7l Is Essential for Spermatogenesis

Initially 4 *Taf7l* heterozygous ($Taf7l^{+/-}$) females and two *Taf7l*-null ($Taf7l^{-/Y}$) males (*N6*) were obtained from the Univ. of Penn. and mating cages were set up for all 6 mice with either WT males or females. Litters were obtained from all $Taf7l^{+/-}$ females mating with WT males but no litter was born from the two $Taf7l^{-/Y}$ males mating with WT females for over a year. Further matings were carried out with the offspring ($Taf7l^{+/-}$ or $Taf7l^{-/Y}$) and again litters were only obtained from $Taf7l^{+/-}$ females mating with WT males, but none from $Taf7l^{-/Y}$ males mating with either WT or $Taf7l^{+/-}$ females. These observations indicated that N_{7-9} $Taf7l^{-/Y}$ males after 2-4 additional rounds of backcrossings might have become sterile.

To more carefully assess the fertility of $Taf7l^{-/Y}$ males, we carried out a comprehensive analysis documented in Figure 2.26A. Out of the 15 matings with $Taf7l^{-/Y}$ males (*N9*) crossed with $Taf7l^{+/-}$ females only 1 produced progeny (litter of 6). Likewise when $Taf7l^{-/Y}$ males crossed with WT females only 1 litter with 5 progeny was produced. In both



Progeny production from $Taf7l^{+/Y}$ and $Taf7l^{-/Y}$ mice for 5 months

Parameters	*Taf7l* genotype for mating[a]		
	$Taf7l^{+/Y}$ X $Taf7l^{+/-}$	$Taf7l^{-/Y}$ X $Taf7l^{+/-}$	$Taf7l^{-/Y}$ X $Taf7l^{+/+}$
mating cages	15	15	15
progeny numbers	450[b]	6	5
litter times	73	1	1

[a]Mice from the C57BL/6J background were used at 6-8 weeks of age.
[b]Distribution of offspring followed Mendelian ratio.

Figure 2.26: ***Taf7l* is essential for male reproduction abilities.** **(A)** Progeny produced by 15 mating cages with WT males ($Taf7l^{+/Y}$) and heterozygous $Taf7l^{-/+}$ females, $Taf7l^{-/Y}$ males and $Taf7l^{-/+}$ females, and $Taf7l^{-/Y}$ males and WT females ($Taf7l^{+/+}$) for 5 mo. **(B and C)** Progeny numbers and litter times of $Taf7l^{-/Y}$ males ($-/Y$) relative to WT males ($+/Y$) of mating analysis in A. **(D)** Number of sperm in WT and $Taf7^{l-/Y}$ (KO) male testis. **(E)** Testis weight of WT and KO mice. **(F)** Serum testosterone levels of WT and KO mice. **(G)** Body weights were measured for WT and KO mice. **D–G** show results with 3-mo-old mice. Values in **D–G** represent the mean ± SEM of mice (n = 7–10). Asterisks denote statistically significant differences of KO compared with WT. *P < 0.05; **P < 0.01 (Student t test).

these cases the single successful mating occurred with younger males (age 6-8 weeks) with no further pup produced over 7 months (Figure 2.26A). By contrast, WT males mated with $Taf7l^{+/-}$ females were successful 73 times producing a total of 450 progeny in 5 months. Genotyping of the offspring by PCR analysis confirmed a distribution ($Taf7l^{+/+}$, $Taf7l^{+/-}$, $Taf7l^{+/Y}$ and $Taf7l^{-/Y}$) that follows the expected Mendelian ratio. This data confirms that the reproductive potential of $Taf7l^{-/Y}$ males is severely compromised (Figure 2.26B and C), and that $Taf7l^{-/Y}$ males are essentially infertile. This severe male sterility also made $Taf7l^{-/-}$ females effectively unavailable because *Taf7l* is an X-chromosome-linked gene and heterozygous $Taf7l^{-/Y}$ behave as Taf7l-null males severely limiting our ability to assess the reproductive capacity of $Taf7l^{-/-}$ females.

As part of our phenotyping, we also measured sperm count, testis weight and serum testosterone levels as previously reported (Figure 2.26D and E).[101] $Taf7l^{-/Y}$ males also have slightly reduced testosterone levels, which may be related to our recent finding that *Taf7l* KO animals show defects in the synthesis of white adipocytes and cholesterol (Figure 2.26F).[12] Otherwise, $Taf7l^{-/Y}$ males appear normal and healthy; they showed no apparent abnormalities in major organs, no obvious differences in body weight (Figure 2.26G) and no detectable overall metabolic changes under normal feeding conditions.

Taf7l Ablation Disrupts Normal Gene Expressions in Testis

In order to gain a better understanding of the more severe phenotypes observed, we compared the genome-wide expression profiles of 3-month-old WT and $Taf7l^{-/Y}$ testes by mRNA-seq. RNAs from 6 WT and 6 $Taf7l^{-/Y}$ (KO) testes were extracted and mixed separately to avoid individual differences, RNA-seq libraries were generated for both WT and KO samples then analyzed by Illumina deep-sequencing. The genome-wide expression data identified 726 genes down-regulated (by 3-fold or more) and 894 genes up-regulated in mouse testes lacking *Taf7l* (Figure 2.29A). In good agreement with a previous study, the six genes identified as potential targets of *Taf7l* (*Cpa6*, *Adc*, *Fscn1*, *Sfmbt2*, *4732473B16Rik*, and *D1Ertd622e*) by microarray analysis[101] were also found to be reduced (2 fold) in $Taf7l^{-/Y}$ testes (Figure 2.29D; Figure 2.28). Gene ontology analysis revealed that many genes implicated in spermatogenesis and metabolism were substantially down-regulated in the testis of *Taf7l* knock-out males. By contrast, the up-regulated genes include those involved in antimicrobial activity, growth regulation, metabolism and immune system development; but not spermatogenesis (Figure 2.27).

Based on gene ontology, we have sub-divided the genes down-regulated by the loss of *Taf7l* into three classes (Figure 2.29). Note that some genes could just as well be classified within multiple sub-groups (ie.*Tssk3* gene that is involved in metabolism and sperm motility). One prominent group of genes dependent on *Taf7l* represent well-documented spermatogenic activators and markers such as *AR*, *Zfx*, *Spz1* and *Spem1*.[140,141,142,143,144] A

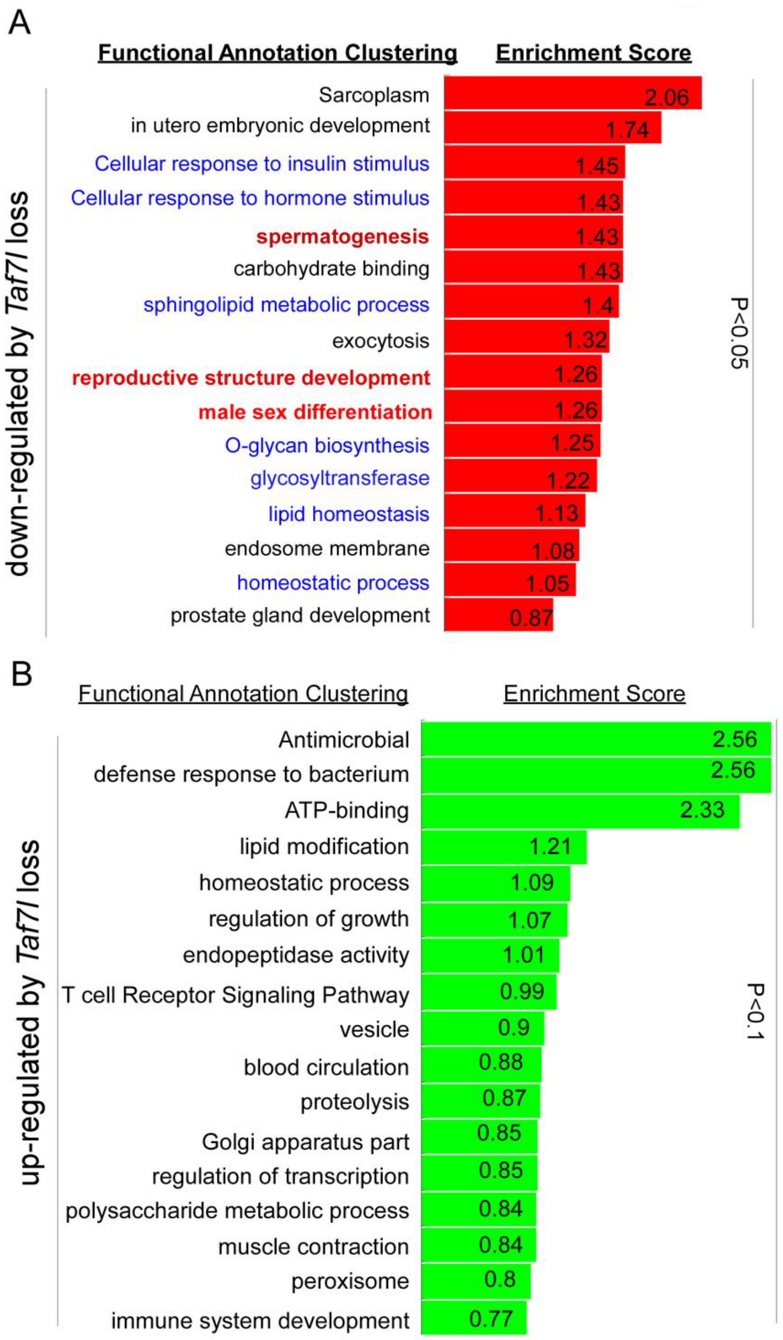


Figure 2.27: ***Taf7l* depletion down-regulates spermatogenic and metabolic genes in testis by RNA-seq analysis.** **(A)** Gene ontology analysis on genes down-regulated in $Taf7l^{-/Y}$ (KO) testis. **(B)** Gene ontology analysis on genes up-regulated in $Taf7l^{-/Y}$ testis.

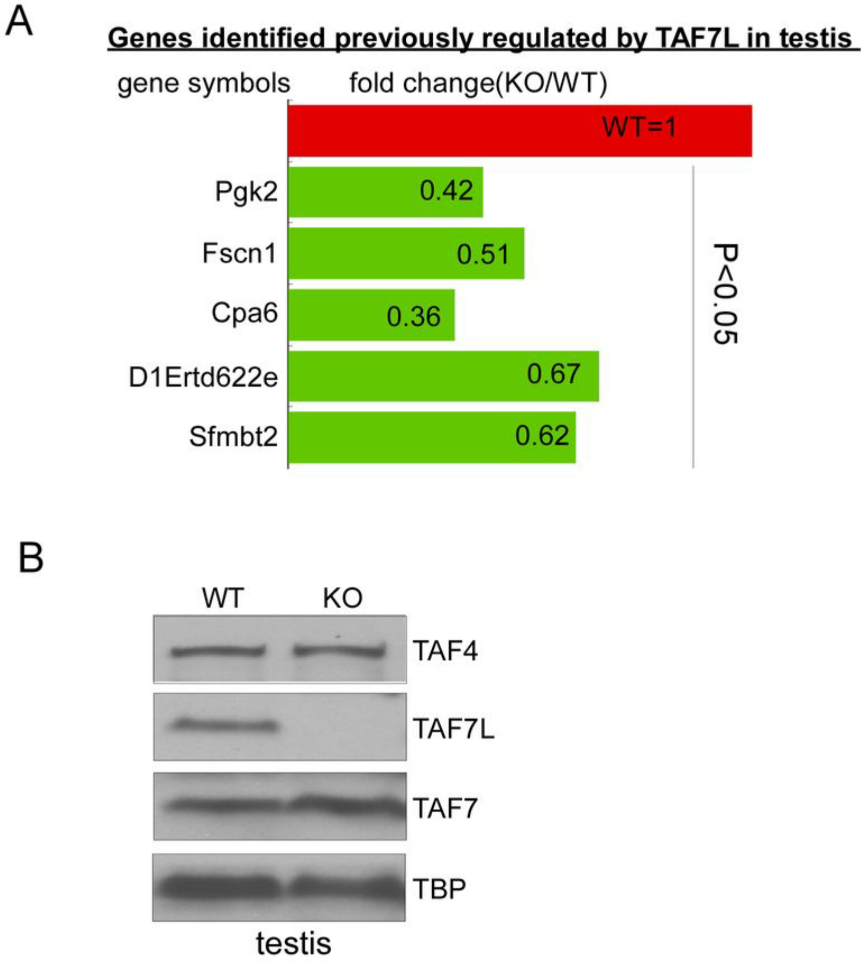


Figure 2.28: **mRNA-seq analysis confirms the previous findings from Microarray analysis. (A)** Expression of *Taf7l*-regulated genes identified by microarray analysis. Expression level of genes in WT testis were arbitrarily assigned a value of 1, their corresponding expression levels in KO testis are expressed relative to this value. **(B)** Western blots analyze TAF4, TAF7L, TAF7 and TBP protein levels in WT and $Taf7l^{-/Y}$ testis.

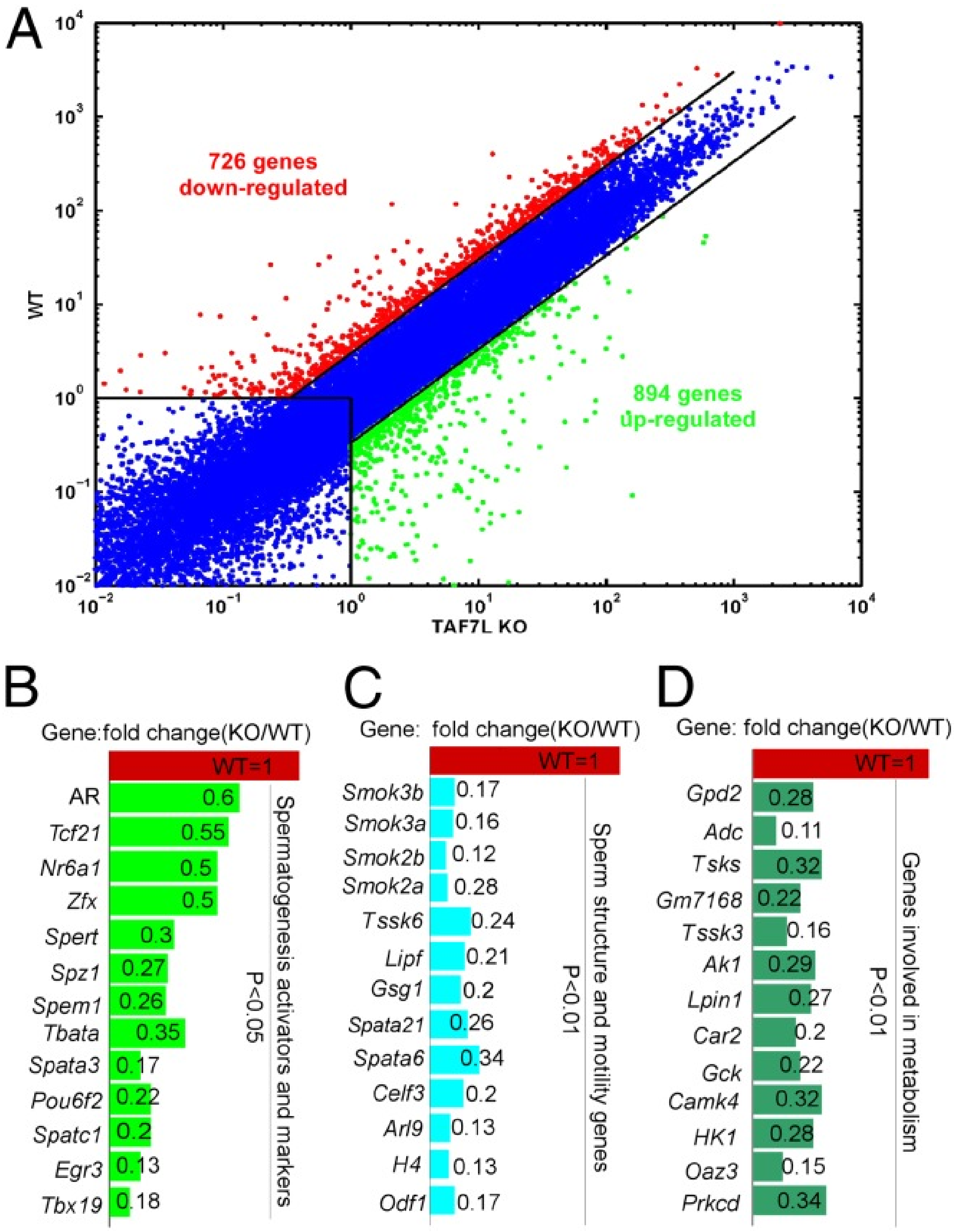


Figure 2.29: ***Taf7l* depletion dramatically deregulates testis gene expression by RNA-seq analysis. (A)** Global gene expression of testis in $Taf7^{l-/Y}$ (TAF7L KO) (horizontal axis) and WT (vertical axis) mice. Red dots represent genes down-regulated above threefold (726 genes); green dots represent genes up-regulated above threefold (894 genes); blue dots represent genes with low expression and genes with unaltered expression. **(B)** Relative expression levels of representative spermatogenic activators and markers. **(C)** Relative expression levels of representative genes important for sperm structure and motility. **(D)** Relative expression levels of genes involved in metabolism. **(B–D)** Expression level of genes in WT testis were arbitrarily assigned a value of 1; their corresponding expression levels in KO testis are expressed relative to this value.

second group of Taf7l-regulated genes (*Odf1*, *Tssk6*, *Smok2a/b* and *Smok3a/b*) have been implicated in sperm structure and motility during late stages of spermatogenesis (Figure 2.29B, C).[145,146,147,148] Additionally, a group of genes that have generally been classified as ones involved in metabolism are significantly down-regulated in the testis upon loss of *Taf7l* and likely reflect both metabolic and spermatogenesis functions. These genes include the Tssk3 or Tsks testis-specific kinases[149,150] as well as Camk4, a protein kinase involved in phosphorylating protamines,[151] Hk1 and GCK hexokinases involved in sperm glycolysis linked to motility[152] and Oaz3 (ornithine decarboxylase antizyme) that controls polyamines required for proper sperm formation[153,154] (Figure 2.29D). In order to confirm the RNA-seq results, we have carried out RT-qPCR on a handful of genes selected from these three representative classes (Figure 2.30), providing additional evidence that *Taf7l* regulates a subset of spermatogenic and metabolic genes in testis. These findings also provide some insight into the relationship between metabolism and reproduction that have been reported in the study of various animal models[155,118,156,157] linking dysregulated metabolism with sterility. Our genome-wide expression data thus points to the likely involvement of *Taf7l* in regulating both metabolic and spermatogenesis genes in testis as being at least partly responsible for the observed infertility.

TAF7L Binds to Promoters of Target Genes in Testis

Given that hundreds of genes are down-regulated or up-regulated by 3-fold or more in $Taf7l^{-/Y}$ testis, we next set out to determine whether *Taf7l* directly binds to the promoters and/or enhancers of those spermatogenic and metabolic genes in testis. To this end, we carried out ChIP-seq mapping of *Taf7l* binding sites in WT testes, using Pol II binding sites as positive controls to mark the actively transcribed genes and IgG as negative controls. First, mapped ChIP tags from deep sequencing using Bowtie were analyzed by intersecting MACS and Grizzly Peak algorithms to identify binding regions for *Taf7l* and Pol II.[8,7] This analysis identified 28,979 significant peaks for Pol II and 10,352 significant peaks for *Taf7l* (Figure 2.31C) and no significant peaks for IgG control. Next we analyzed the distances of the *Taf7l* binding peaks to transcription start sites (TSS) and found that 95% of *Taf7l* peaks are within 1kb of a TSS. Co-localization analysis between *Taf7l* and Pol II peaks revealed that almost all the *Taf7l* peaks overlap with Pol II peaks (Figure 2.31A and B), suggesting that *Taf7l* largely associates with the promoters of actively transcribed genes in the testis. In contrast to the *Taf7l* dependent genes in testis, our direct binding data revealed that nearly all the genes up-regulated by the loss of *Taf7l* in the testis (Figure 2.27) likely result from indirect secondary effects of *Taf7l* depletion. We found that many down-regulated spermatogenic and metabolic genes identified by RNA-seq analysis bear direct *Taf7l* binding sites at or near their promoters. A few representative gene loci such as *Nr6a1*, *Tssk3* and *Sgk1* clearly show that *Taf7l* co-localizes with Pol II at promoter sites (Figure 2.31D, E and F). ChIP-qPCR analysis using WT and *Taf7l*-null testes confirmed that Pol II binding becomes dramatically diminished in *Taf7l*-null testes at target promoters (*Nr6a1*, *Tssk3*, *Sgk1*, *Prm1*, and *Fscn3*), but not at a control actin intron (Figure 2.33), suggesting that the

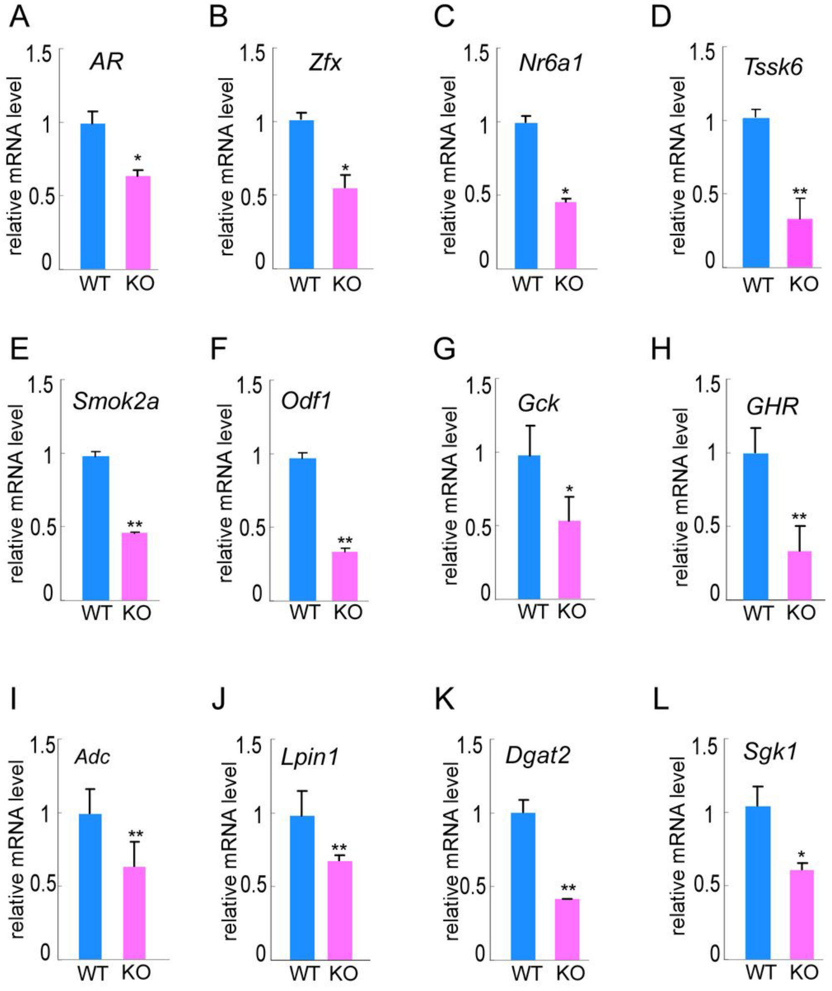


Figure 2.30: **RT-qPCR analysis verifies the results of mRNA-seq analysis. (A, B, C)** AR, *Zfx* and *Nr6a1* are activators involved in spermatogenesis; **(D, E)** *Tssk6* and *Smok2a* are genes involved in sperm motility; **(F)** *Odf1* is sperm structure gene. **(G, H, I, J, K, and L)** *Gck*, GHR, *Adc*, *Lpin1*, *Dgat2* and *Sgk1* are genes involved in lipid synthesis and metabolism. Expression level of genes in WT testis were arbitrarily assigned a value of 1, their corresponding expression levels in KO testis are expressed relative to this value. Values represent the mean ± SEM of three independent experiments. Asterisks denote statistically significant differences of KO compared to WT (student's t test, *$p<0.05$, **$p<0.01$).

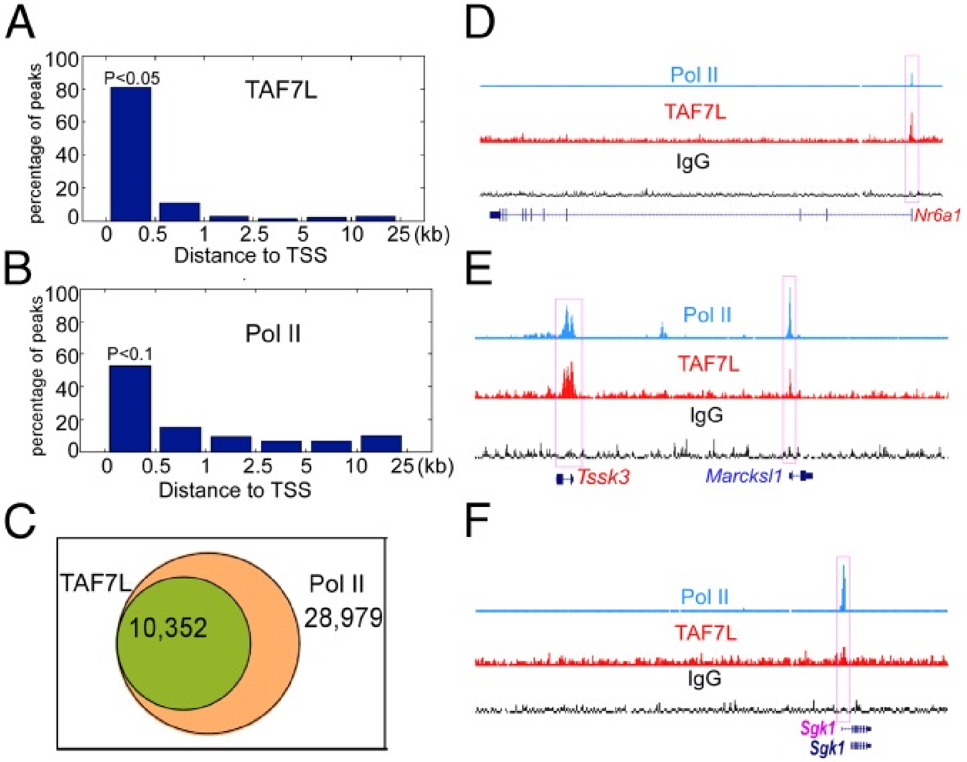


Figure 2.31: **ChIP-seq analysis identifies TAF7L and Pol II binding sites in testis.** **(A and B)** Percentage of TAF7L **(A)** and Pol II **(B)** binding peaks vs. the distance of the peaks to transcription start sites (TSS) on genome-wide scale in testis. **(C)** Overlapping between TAF7L and Pol II binding peaks in testis. **(D–F)** Read accumulation of Pol II and TAF7L were shown on the *Nr6a1* **(D)**, *Tssk3* **(E)**, and Sgk1 **(F)** gene loci. Vertical axis is 0–500 reads for each factor; colocalized peaks are marked with red solid boxes. IgG served as negative control for ChIP-seq analysis.

presence of *Taf7l* is indeed critical for the proper expression of target genes in testis. Taken together these data suggest that *Taf7l* directly binds to gene promoters and this binding is required for the expression of spermatogenic and metabolic genes in testis.

$Taf7l^{-/Y}$ Resembles $Trf2^{-/-}$ Mice and Targets Similar Postmeiotic Genes

Having found that *Taf7l* binds to a specific subset of promoter regions of genes involved in spermatogenesis, we next set out to assess the possibility that *Taf7l* associates with other testis-specific core promoter recognition factors such as *Taf4b* and/or *Trf2* to work combinatorially to target genes directing spermatogenesis. First, we examined the spermatogenic defects of $Taf7l^{-/Y}$ mice relative to $Taf4b^{-/-}$ and $Trf2^{-/-}$ mice. HE staining of testes from 3 month-old WT and $Taf7l^{-/Y}$ (KO) mice revealed dramatically decreased elongated spermatids (Figure 2.32A). The phenotype was highly reminiscent of $Trf2^{-/-}$ animals, but quite distinct from Taf4b-/- mice that showed defects in the maintenance of germ stem cells (9, 16). Next, we compared the target genes regulated by *Taf7l*, *Taf4b* and *Trf2* in testis identified by RNA-seq.[132, 136] As shown in Figure S5, *Taf7l* and *Taf4b* appear to regulate

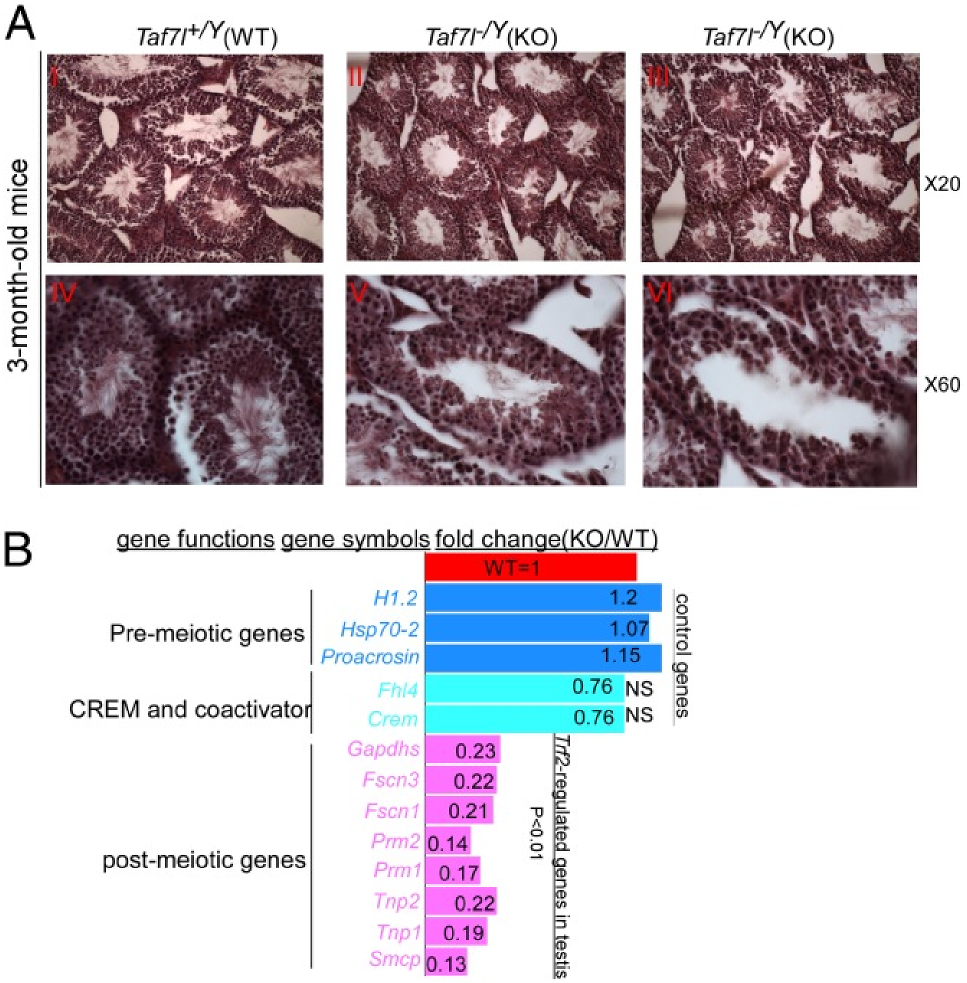


Figure 2.32: ***Taf7l* depletion blocks spermiogenesis and impairs postmeiotic gene expression in a similar way to $Trf2^{-/-}$ in testis.** **(A)** HE staining on 3-mo-old WT and KO testis at magnification 200 (I–III) and magnification 600 (IV–VI). **(B)** Expression of *Trf2*-regulated postmeiotic and control genes in $Taf7l^{-/Y}$ testis was compared with WT testis from mRNA-seq analysis. Expression levels of genes in WT testis were arbitrarily assigned a value of 1, and their corresponding expression levels in KO testis are expressed relative to this value. NS, not significant.

largely non-overlapping sets of genes expressed in the testis. For example, *GDNF*, *Stra8*, *Stag3* and *Dmc1* are robustly down-regulated by the loss of *Taf4b* but are slightly increased in $Taf7l^{-/Y}$ testes, suggesting that *Taf7l* is not required for maintenance of male germ stem cells. Instead, it appears that loss of *Taf7l* blocks spermatogenesis at a post-meiotic stage and $Taf7l^{-/Y}$ testes may actually accumulate more stem cells and meiotic cells (Figure 2.32A and B).

We also examined the influence of *Taf7l* depletion on the expression of TFIID subunits by RNA-seq and RT-qPCR analysis and found that except for *Taf7l* itself, all the other TAFs are largely unaltered or slightly up-regulated (Figure 2.35). These findings suggest that

Taf7l is a ‘non-prototypical’ testis-specific TAF subunit whose absence does not influence overall levels of TFIID. Strikingly however, loss of *Taf7l*, like the loss of *Trf2*, down-regulates the same subset of genes directing spermatogenesis. For example, *Taf7l* doesn’t influence the transcription of pre-meiotic genes such as *Proacrosin*, *H1.2* and *Hsp70-2* but dramatically down-regulates *Trf2*-regulated post-meiotic genes such as *Gapdhs*, *Tnp1*, *2*, *Prm1*, *2*, *Fscn1*, *2* and *Smcp* (Figure 2.32B). This requirement for both *Taf7l* and *Trf2* for proper expression of post-meiotic genes is also consistent with the spermiogenesis defects observed in both $Taf7l^{-/Y}$ and $Trf2^{-/-}$ testes (Figure 2.32A). These findings suggest that *Taf7l* likely operates together with *Trf2*, but not *Taf4b*, to control post-meiotic genes as deduced from our RNA-seq (Figure 2.32B; Figure 2.34), RT-qPCR (Figure 2.36) and ChIP-seq analysis (Fig. 3; Figure 2.33). We find, for instance, *Taf7l* binds efficiently to the TSS regions of *Trf2*-regulated gene *Gapdh*, but not at the promoter regions of *Taf4b*-regulated genes such as *Stra8* and *Dmc1* (Figure 2.37).

We also assessed the expression levels and nuclear localization patterns of *Taf4b*, *Trf2* and *Taf7l* during spermatogenesis to probe potential direct crosstalk between these factors based on previous studies. *Taf4b* is highly expressed in gonocytes, nuclei of spermatogonia and spermatids but not at other stages of spermatogenesis or in somatic cells.[132] By contrast, *Taf7l* is expressed in germ cells not testis somatic cells during most stages of spermatogenesis, but it remains cytoplasmically localized in spermatogonia and its expression becomes silenced at meiotic stages due to sex-chromosome inactivation (MSCI). *Taf7l* transcription becomes reactivated and translocated into nucleus of spermatids[100, 137] at a stage when *Trf2* is also specifically highly expressed.[139, 136] Taken together loss of *Taf7l* and *Trf2* exhibit a similar spermiogenesis deficiency phenotype; they appear in the nuclei of post-meiotic spermatids at similar times and bind to core promoter elements in an overlapping subset of genes and regulate their expression in the testis. These findings raise the possibility that *Taf7l* and *Trf2* may actually work together in regulating spermatogenesis.

Taf7l and *Trf2* Coregulate a Subset of Postmeiotic Genes

To further explore the possibility of interaction between *Taf7l* and *Trf2* in regulating post-meiotic gene expression during spermatogenesis, we used FLAG-tagged *Taf7l* or TAF7 to Co-IP with HA-tagged *Trf2* co-expressed in 293T cells. These co-IP experiments showed that *Taf7l* can efficiently pull down *Trf2* and vice versa (Figure 2.38A); by contrast, TAF7 (a paralogue of TAF7L) was unable to co-IP *Trf2* (Figure 2.38B). To confirm that the association between *Taf7l* and *Trf2* is directed by protein:protein interactions and not mediated via indirect DNA/chromatin interactions, we included benzonase treatment in our co-IP assays. Eliminating DNA in these co-IP experiments did not alter the binding interactions we observed between *Taf7l* and TRF2. Next, we used affinity-purified *Taf7l* antibody to co-IP endogenous *Trf2*in WT and $Taf7l^{-/Y}$ testis. The results showed that *Taf7l* can pull down endogenous *Trf2* from WT testes, but not from $Taf7l^{-/Y}$ testes (Figure 2.38C), suggesting that *Taf7l* is associated with *Trf2* in testes. The association between *Taf7l* and *Trf2* in vitro

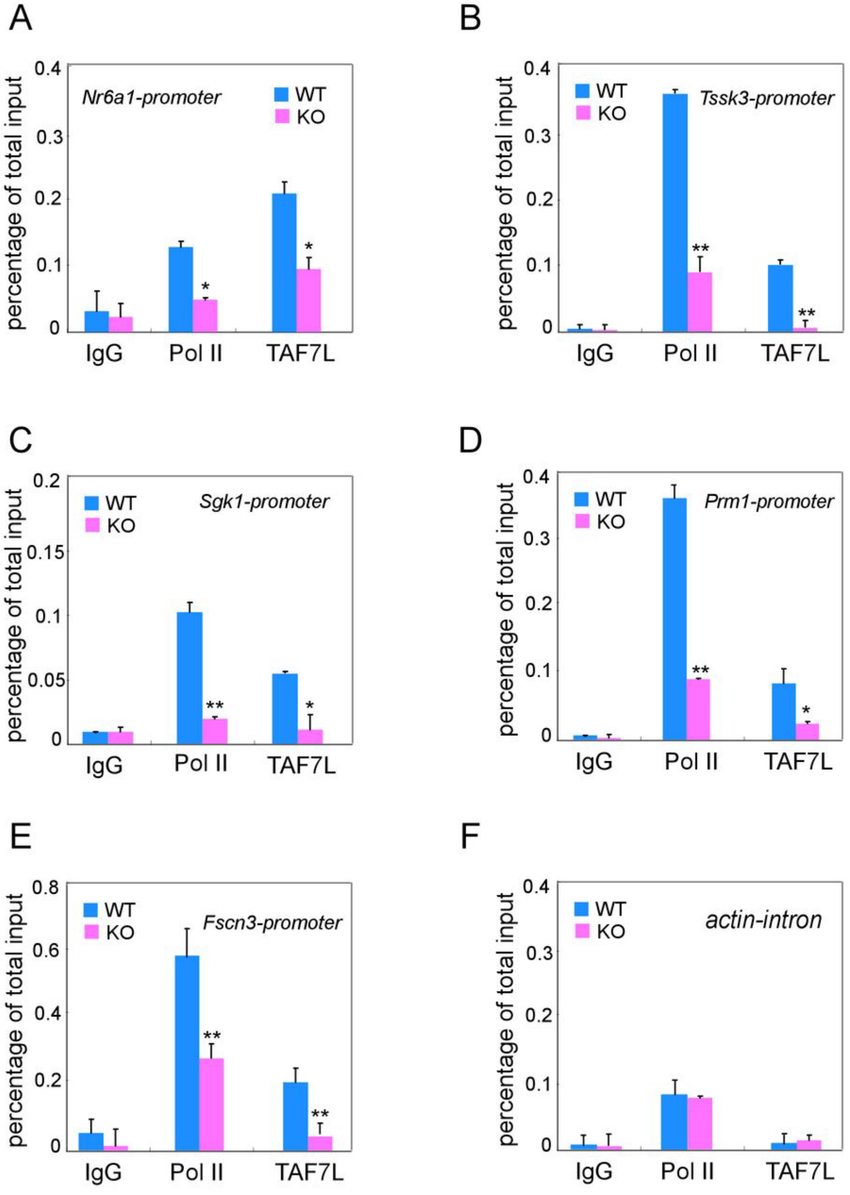


Figure 2.33: ***Taf7l* depletion greatly diminishes the binding of Pol II to target gene promoters.** ChIP-qPCR analysis of IgG, Pol II, and TAF7L on *Nr6a1* (**A**), *Tssk3* (**B**), *Sgk1* (**C**), *Prm1* (**D**), and *Fscn3* (**E**) promoters in WT and $Taf7l^{-}/Ytestis, actin{-}intron$(**F**)$isservedascontrol.Valuesrepresentthemean$± SEM of three independent experiments. Asterisks denote statistically significant differences of KO compared to WT (student's t test, *p<0.05, **p<0.01).

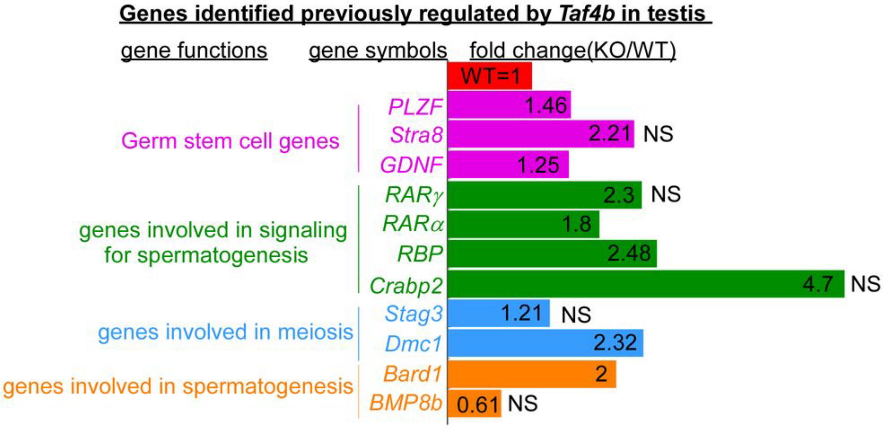


Figure 2.34: ***Taf7l* doesn't regulate *Taf4b*-regulated germ stem cell genes by RNA-seq analysis.** Expression level of genes in WT testis were arbitrarily assigned a value of 1, their corresponding expression levels in KO testis are expressed relative to this value. NS stands for no significant.

and in vivo prompted us to explore further the relationship between *Taf7l* and *Trf2* or TBP in testes. A previous study had reported that TBP can be an interacting partner of *Taf7l* in testes.[101] For these experiments we used *Taf7l* and TBP antibodies to co-IP each other in WT and $Taf7l^{-/Y}$ testes and found that only a small portion of *Taf7l* weakly associates with TBP in mouse testes (Figure 2.38D). A direct comparison of *Trf2* and TBP signals co-IPed by *Taf7l* (Figure 2.38C and D) confirmed that *Taf7l* associates more efficiently with *Trf2* than with TBP in mouse testes when all three proteins are present. These studies suggest that two atypical testis-specific core promoter recognition factors, *Taf7l* and TRF2, likely work in concert to regulate a subset of post-meiotic genes required for spermiogenesis. Ablation of either factor results in a similar blockade of spermiogenesis leading to male infertility.

Discussion

Cell-type–specific transcription is a key driver of tissue and organ formation during embryonic development. These complex expression networks are controlled by tissue-specific enhancer/promoter binding factors as well as core promoter recognition factors including Tafs, mediators, and TRFs. In several previous studies, *Taf7l* and *Trf2* have separately been found to function as testis-specific transcription factors. In this study, we provide unique evidence suggesting that these two atypical core promoter recognition factors work together to drive testis-specific gene expression.

Our model proposes that during the post-meiotic wave of transcription that occurs at the spermatogenesis-spermiogenesis transition, *Taf7l* is reactivated from MSCI and translocated into the nuclei of pachytene/round spermatids where *Trf2* is highly expressed. We postulate

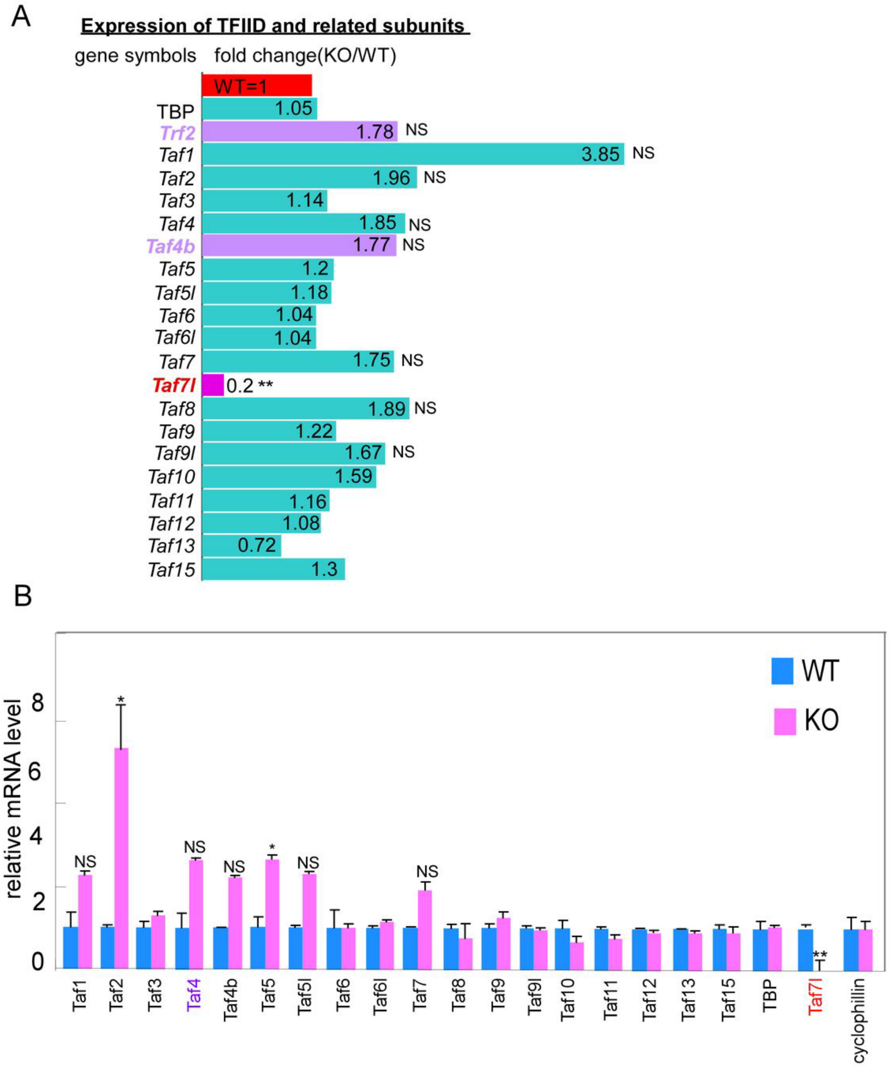


Figure 2.35: **Expression of TFIID subunits is largely unaltered in $Taf7l^{-/Y}$(KO) testis. (A)** Expression of TFIID subunits in $Taf7l^{-/Y}$ testes relative to WT testes from mRNA-seq analysis. Expression level of genes in WT testis were arbitrarily assigned a value of 1, their corresponding expression levels in KO testis are expressed relative to this value. **(B)** Expression of TFIID subunits through RT-qPCR analysis in $Taf7l^{-/Y}$ testes compared to WT testes. Values in **B** represent the mean ± SEM of three independent experiments. Asterisks denote statistically significant differences of KO compared to WT (student's t test, *p<0.05, **p<0.01), NS stands for no significant.

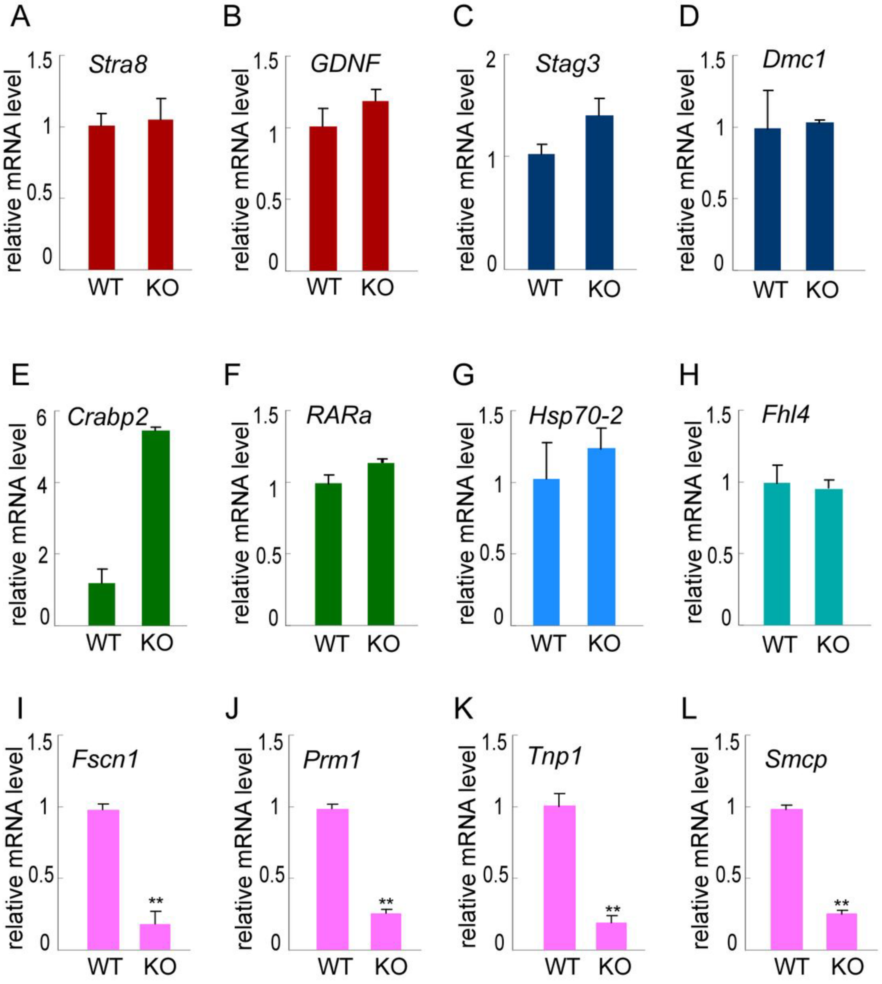


Figure 2.36: **RT-qPCR analysis shows the expression levels of *Trf2*/*Taf4b*-regulated genes in *Taf7l* KO testis.** **(A-F)** are *Taf4b*-regulated genes in testis; **(A, B)** *Stra8* and GDNF are germ stem cells genes; **(C, D)** *Stag3* and *Dmc1* are meiosis genes; **(E, F)** *Crabp2* and *RAR* are spermatogenic signaling genes. **(G, H)** *Hsp70-2* and *Fhl4* are not *Trf2*-regulated pre-meiosis genes; **(I, J, K, and L)** *Fscn1*, *Prm1*, *Tnp1* and *Smcp* are *Trf2*-regulated post-meiotic spermiogenesis genes. Expression level of genes in WT testis were arbitrarily assigned a value of 1, their corresponding expression levels in KO testis are expressed relative to this value. Values represent the mean ± SEM of three independent experiments. Asterisks denote statistically significant differences of KO compared to WT (student's t test, *p<0.05, **p<0.01).

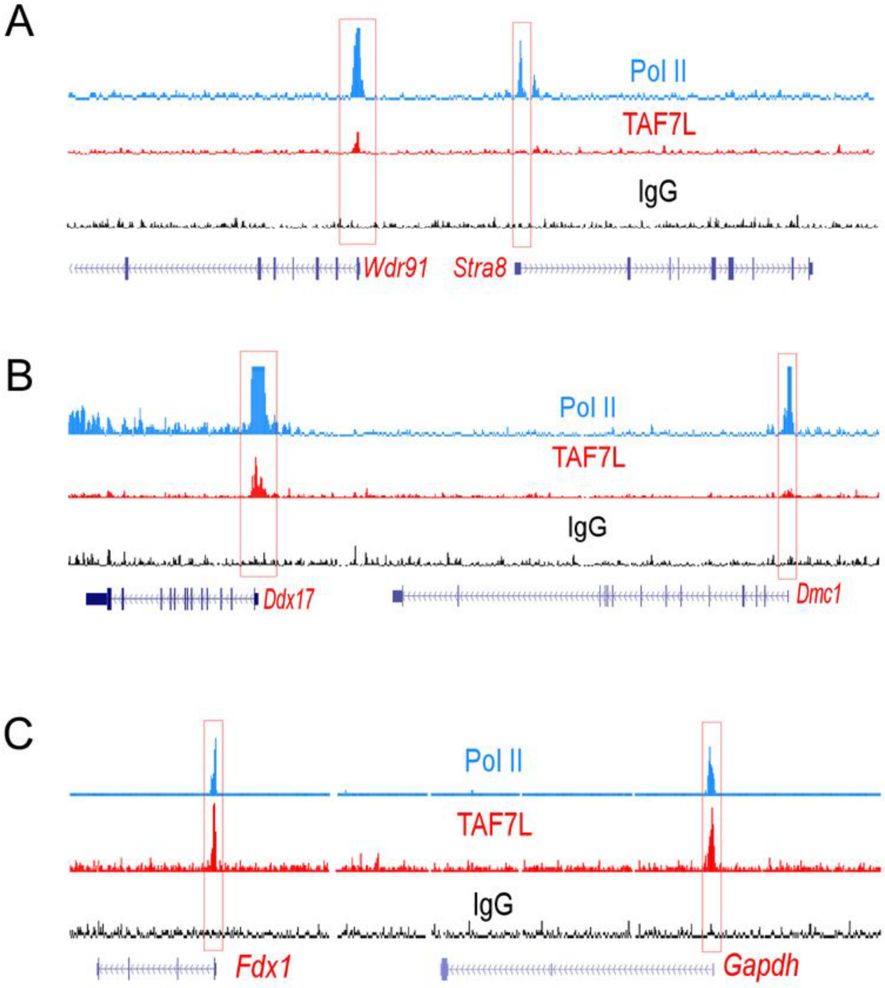


Figure 2.37: ***Taf7l* binds to promoters of post-meiotic genes but not with germ stem cell and meiotic genes.** **(A)** Read accumulation of Pol II and TAF7L was shown on the *Stra8* and *Wdr91* **(A)**, *Dmc1* and *Ddx17* **(B)**, and *Gapdh* and *Fdx1* **(C)** gene loci. Vertical axis is 0–500 reads for each factor, co-localized peaks were marked with red solid boxes; IgG is served as negative control for ChIP-seq analysis.

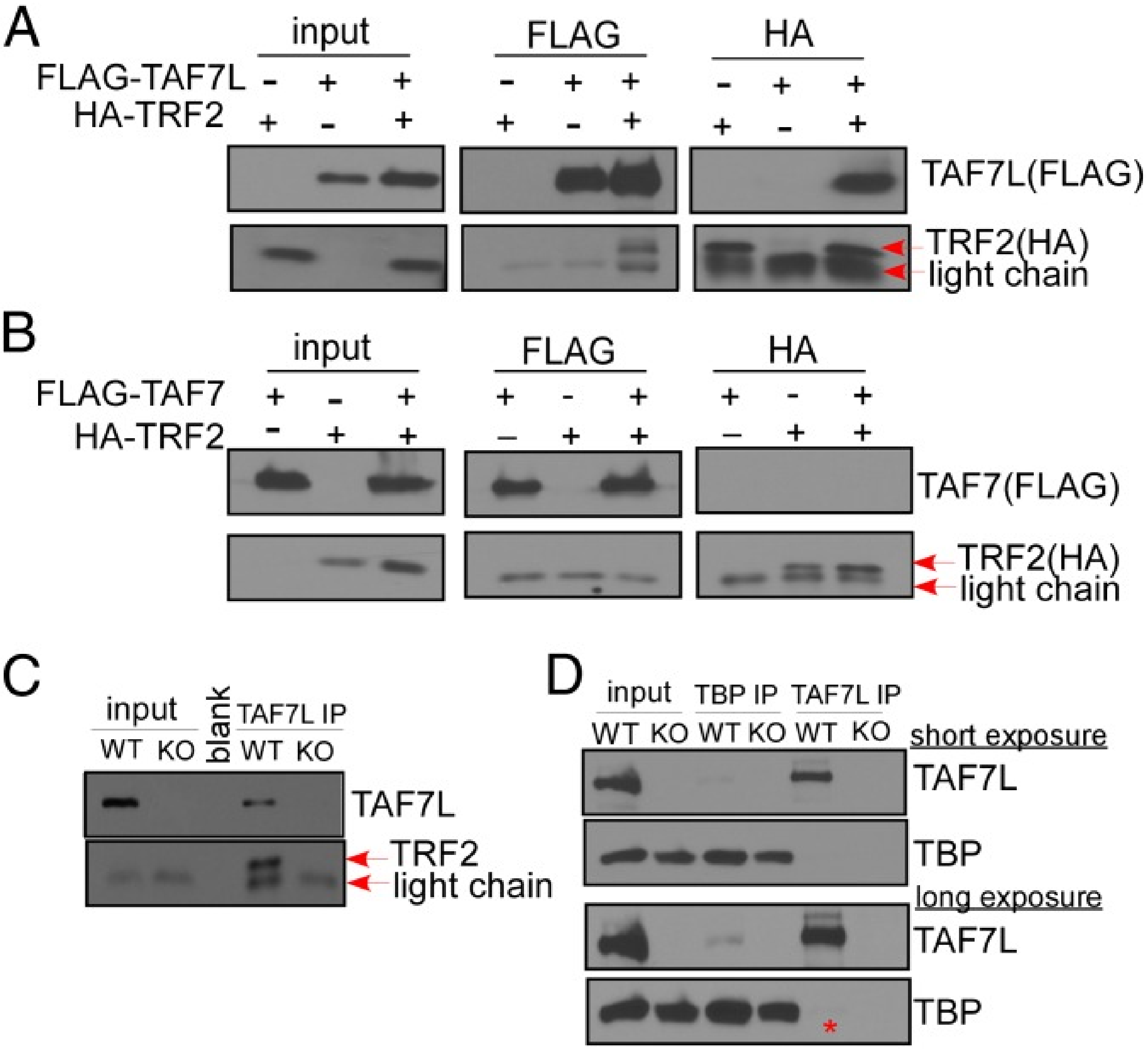
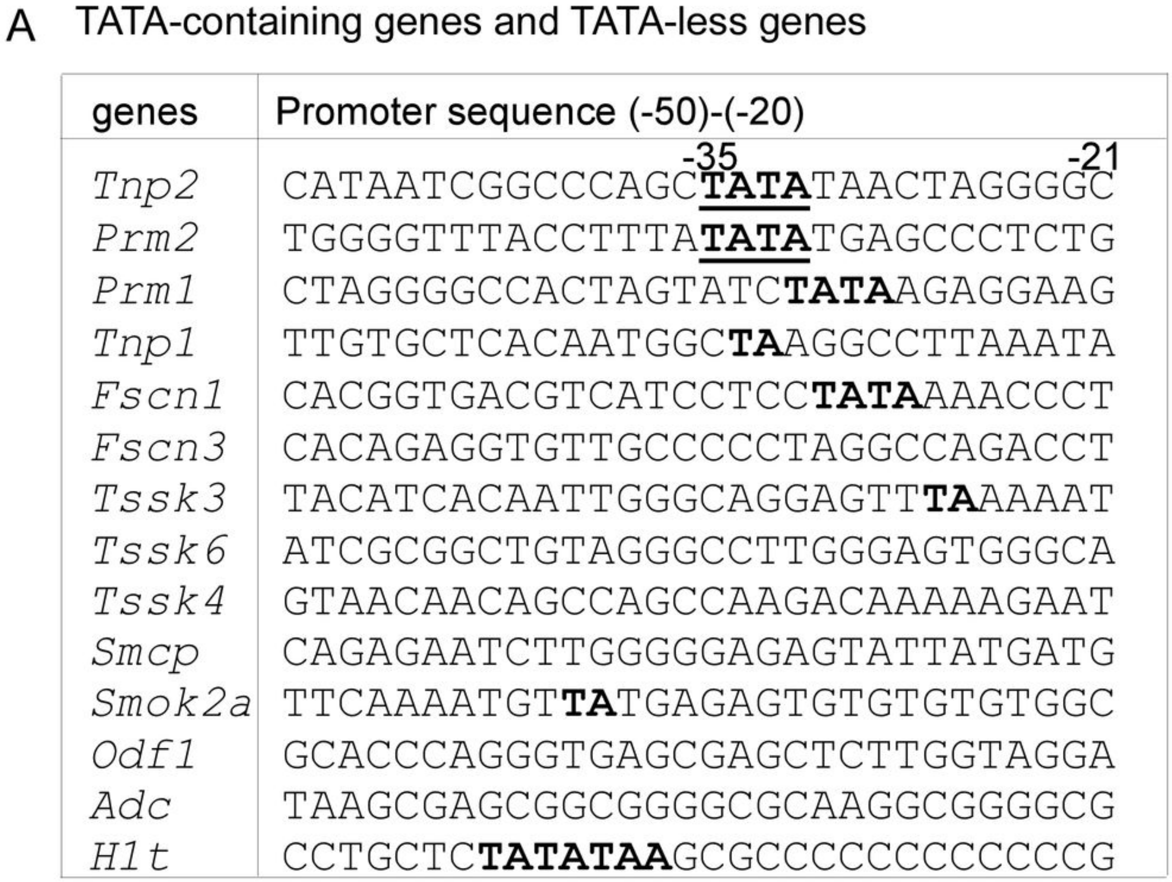


Figure 2.38: **TAF7L associates with TRF2 both in vitro and in vivo.** **(A)** FLAG–TAF7L and HA–TRF2 were overexpressed in 293T cells, and IPs were performed on both FLAG and HA antibodies, followed by Western blotting analysis with FLAG and HA antibodies. **(B)** The same procedures were performed on FLAG–TAF7 and HA–TRF2. **(C)** IPs were performed on WT and $Taf7l^{-/Y}$ testis lysates with TAF7L antibody and followed by Western blotting with TAF7L and TRF2 antibodies. **(D)** IPs were performed on WT and $Taf7l^{-/Y}$ testis lysates with either TAF7L or TBP antibody, followed by Western blotting with TAF7L and TBP antibodies. Both images are from the same Western blots with short (*Upper*) and long (*Lower*) exposure time. Red star marks the TBP signal from TAF7L IP.



A TATA-containing genes and TATA-less genes

genes	Promoter sequence (-50)-(-20)
	-35 -21
Tnp2	CATAATCGGCCCAGC**TATA**TAACTAGGGGC
Prm2	TGGGGTTTACCTTTA**TATA**TGAGCCCTCTG
Prm1	CTAGGGGCCACTAGTATC**TATA**AGAGGAAG
Tnp1	TTGTGCTCACAATGGC**TA**AGGCCTTAAATA
Fscn1	CACGGTGACGTCATCCTCC**TATA**AAACCCT
Fscn3	CACAGAGGTGTTGCCCCCTAGGCCAGACCT
Tssk3	TACATCACAATTGGGCAGGAGTT**TA**AAAAT
Tssk6	ATCGCGGCTGTAGGGCCTTGGGAGTGGGCA
Tssk4	GTAACAACAGCCAGCCAAGACAAAAAGAAT
Smcp	CAGAGAATCTTGGGGGAGAGTATTATGATG
Smok2a	TTCAAAATGT**TA**TGAGAGTGTGTGTGTGGC
Odf1	GCACCCAGGGTGAGCGAGCTCTTGGTAGGA
Adc	TAAGCGAGCGGCGGGGCGCAAGGCGGGGCG
H1t	CCTGCTC**TATATAA**GCGCCCCCCCCCCCG

Figure 2.39: **Majority of *Taf7l*-regulated genes contain a TATA-less promoter.** Promoter sequence 20(-20) to 50-50bases upstream of TSS of genes involved in spermatogenesis and regulated by TAF7L in testis.

that a testis-specific transcription pre-initiation complex (PIC) containing both *Taf7l* and *Trf2* is formed and targeted to a subset of "TATA-less" promoters to regulate post-meiotic *Trf2*-dependent gene expression[158,159,160] (Figure 2.39). At the same time, prototypic TBP-containing core promoter recognition complexes can operate to direct TBP-dependent housekeeping genes and possibly some testes-specific genes. TAF7L, *Trf2* and TBP all seem to be required for spermatogenesis and each may provide some non-redundant testis-selective transcription function. Previous studies found that *Trf2* can associate with multiple proteins to form a complex of >500 kDa.[133] It will be interesting to see whether there is a TRF2-TAF complex of similar size in testes and eventually identify the other associated factors to more fully dissect the molecular mechanisms of *Taf7l* and *Trf2*-driven testis-specific transcription. Also beyond its roles in testis-specific transcription control, it will be interesting to test whether *Taf7l* is required for maintaining chromatin architecture in round spermatids.

Given that *Taf7l* is highly expressed in germ cells, but not somatic cells in testis, we analyzed the global influence of *Taf7l* depletion in the whole testis. As a result, both up- and down-regulated genes found in this study by RNA-seq or RT-qPCR were probably the direct result of *Taf7l* KO. It is also likely that diminished elongated spermatids observed are due to blockade of spermiogenesis upon loss of *Taf7l.* Although our ChIP-seq data supported a role for *Taf7l* in regulating testis-specific genes by direct binding to their promoters, future studies with stage-specific fractionated germ cells will be helpful to elucidate a more detailed molecular mechanism.

We recently reported that *Taf7l* associates with the adipocyte-specific transcription factor PPAR during adipogenesis.[12] Here in the context of testis-specific transcription *Taf7l* teams up with TRF2, another testis-specific core promoter recognition factor, to direct male sperm formation. The association of an orphan TAF and a TRF is reminiscent of the TRF3/TAF3 scenario reported for myogenesis.[97] These studies, taken in aggregate, suggest that various cell lineages take advantage of diversified core promoter factors in a combinatorial fashion to regulate tissue specific programs of transcription.

Due to adverse side effects of female contraceptive medicines and the lack of male ones, an effective and safe therapeutic target has become a focus in the male germ cell research field. Thus far, hundreds of genes have been found to influence spermatogenesis when mutated or depleted, including core promoter recognition factors *Trf2* and Taf4b, leptin ob and leptin receptor db genes, and other genes such as *Daz* and *Ddx4*.[161,162,163,156] However, most of these genes have relatively high expression levels and function in tissue or organs other than testes.Therefore, their depletion or inhibition often results in side effects and unacceptable complications that can range from minor to severe. For example, depletion of ob/db gene results in obesity and diabetes.[156] By contrast, *Taf7l* is more highly expressed in testis than all other tissues that were sampled including adult WAT and, not surprisingly, depletion of *Taf7l* causes >98% male infertility in mice. Most importantly, ablation of *Taf7l* leaves most of the male germ stem cells intact, suggesting that the effect of contraceptives targeting *Taf7l* will likely be reversible upon removal of the treatment. It is also interesting to note that *Taf7l* is an X-chromosome linked gene, males carry only a single copy and therefore the chances of introducing random mutations in *Taf7l* is twice as high as autosomal regulatory genes such as Taf4b and Trf2. This may explain why multiple *Taf7l* mutations have been found in human infertile patients with oligozoospermia (low concentration of sperm). Our mouse studies thus may also help identify a potential candidate target gene affecting human male infertility.

2.5 Methods

DNA Constructs, Cell Lines, and Cell Culture

Construction of in vitro transcription templates and protein expression plasmids are described in Extended Experimental Procedures. HeLa, 293T, NTERA-2 (NT2), and mouse ES cell line D3 were maintained in standard conditions. Differentiation of D3 ES cells was carried out by LIF removal followed by retinoic acid treatment (5–10 M, Sigma).

Purification and Identification of SCC

Nuclear extracts from 400 l of NT2 cells were purified over eight chromatographic steps to homogeneity. Methods for purification and mass spectrometry analyses of SCC are detailed in Extended Experimental Procedures. subsectionWestern Blotting, Immunoprecipitation, and Affinity Purification

Antibodies used are described in Extended Experimental Procedures. Transcriptional activators were purified from transiently transfected HeLa cells followed by affinity purification using anti-FLAG (M2) agarose (Sigma) as described in Extended Experimental Procedures. Recombinant SCC complexes were purified from Sf9 cells infected with baculoviruses (BAC-to-BAC system, Invitrogen) expressing N-terminal His6-tagged or FLAG-tagged XPC, N-FLAG-tagged RAD23B, and untagged CETN2. Sf9 cells were harvested 48 hr after infection, and protein complexes were purified by incubating cell lysates with Ni-NTA resin (QIAGEN), anti-FLAG (M2) agarose (Sigma), and elution by the FLAG peptides.

shRNA-Mediated Knockdown of SCC by Lentiviral Infection

Control nontarget and pLKO shRNA plasmids targeting XPC, RAD23B, and CETN2 (Sigma) were transfected with packaging vectors into 293T cells using FuGENE 6 (Roche). Supernatants were concentrated by ultracentrifugation and resuspended in PBS. Viral titer was determined by a QuickTiter Lentivirus Titer Kit (Cell Biolabs). SCC knockdown was performed by incubating lentiviral concentrates with D3 cells in the presence of 8 g/ml polybrene followed by puromycin selection (1.5 g/ml).

Gene Expression Analysis and ChIP

Total RNA from shRNA-mediated knockdown D3 ES cells was isolated using RNeasy Plus Kit (QIAGEN) and analyzed by qRT-PCR. Chromatin immunoprecipitation (ChIP) assays were performed in D3 cells as described in Extended Experimental Procedures. Precipitated DNA was measured by qPCR or sequenced using an Illumina HiSeq 2000 sequencing platform. Methods for gene expression and ChIP analyses are detailed in Extended Experimental Procedures.

Somatic Cell Reprogramming

Oct4-GFP MEFs (The Jackson Laboratory) were infected with lentiviruses containing STEMCCA and rtTA, followed by infection with pLKO shRNA lentiviral supernatants targeting SCC. Oct4, Sox2, Klf4, and c-Myc expressions were induced by doxycycline, and SCC knockdown MEFs were selected with puromycin. Reprogrammed cells were either detected by alkaline phosphatase activity or stained with anti-SSEA-1 antibodies conjugated to Alexa Fluor 647 (BioLegends) and analyzed by FACS. XPC, RAD23A, and RAD23B knockout MEFs were generous gifts from Dr. Hoeijmakers (Rotterdam, The Netherlands).

Vectors and plasmids

Mouse *Taf7l* full length cDNA was isolated from mouse testis tissue, amplified by PCR and then sequenced. Full length *Taf7l*, *Taf7*, and *Ppar* cDNAs were inserted into pCMV 3FLAG-10 vector to construct pCMV 3FLAG-TAF7L/TAF7/PPAR. *Taf7l*, *Tbp*, and *Ppar* full length cDNAs were cloned into the pCS2+ vector with either HA or FLAG tag at their N-terminus. pLKO.1 shGFP and pLKO.1 shTAF7L vectors were purchased from Open Biosystems. shTAF7L-resistant pCMV 3Flag-TAF7LmA was made by site-directed mutagenesis to introduce two silent mutations in shTAF7L targeted TAF7L cDNA sequence (Ding et al., 2008).

Antibody production and purification, antibodies

A fragment of the *Taf7l* cDNA corresponding to residues 400–600 was cloned into the pGEX-4T-1 vector carrying a GST tag. GST-TAF7L (a. a. 400–600) was expressed and purified from E.coli and injected into rabbits by Covance (Covance Research Products Inc.,Denver, PA). Bleeds were collected after three boosts and TAF7L antibodies were tested and confirmed by in vitro transcribed and translated TAF7L protein and whole protein extracts from mouse testis of WT and Taf7l KO mice (data not shown). The antisera obtained were affinity-purified using antigen immobilized on Affigel 10/15 resin (Bio-rad, Hercules, CA). For Pol II antibody, monoclonal anti-Pol II (8WG16) was concentrated from hybridoma supernatant with Protein A Sepharose Beads (GE Healthcare, Piscataway, NJ).

Antibody information: anti-TAF4 (BD 612054), anti-TBP (abcam 62126), anti-FLAG (Sigma, F3165), anti-HA (abcam 9110), anti--actin (Sigma, A2228), anti-TAF7 (abnova H00006879-M01), anti-FABP4 (abcam 66682), Pol II (monoclonal 8GW16, protein-A purified), PPARγ (sc-7196), mouse and rabbit IgG (prepared in-house and concentrated with Protein A Sepharose Beads).

Cells culture, stable cell line establishment

C3H10T1/2, 3T3-L1, C2C12, HeLa, and 293T cells were cultured in high glucose DMEM with 10% fetal bovine serum at 10% CO2.

C3H10T1/2 shGFP and shTAF7L lentiviral shRNA knockdown stable cell lines were established by transfecting pLKO.1 shGFP or pLKO.1 shTAF7L into C3H10T1/2 cells and then subjecting to puromycin selection for 3 weeks.

C3H10T1/2 shTAF7L cells were transfected with pCMV 3Flag vector, pCMV 3Flag-TAF7 or pCMV 3Flag-TAF7LmA and then subjected to G418 and puromycin double selection for 3 weeks to establish shTAF7L+vector, shTAF7L+TAF7 or shTAF7L+TAF7LmA cell lines used in ‘rescue’ experiments in Figure 2E,F.

C2C12.CNTL and C2C12.TAF7L were established by transfecting pCMV 3Flag vector or pCMV 3Flag-TAF7L vector into C2C12 cells and then underwent G418 selection for 3 weeks.

Adipocyte differentiation, Oil red O staining and C2C12 myogenesis

For adipogenesis, 3T3-L1 and C3H10T1/2 cells were grown in high glucose DMEM supplemented with 10% fetal bovine serum. At confluence, cells were exposed to induction medium containing dexamethasone (1 M), isobutylmethylxanthine (IBMX, 0.1 mM), insulin (5 g/ml), rosiglitazone (1 M), and 10% FBS. 3 days later, cells were further cultured in high glucose DMEM containing insulin (5 g/ml) and rosiglitazone (1 M) until they were ready for harvest. C3H10T1/2 cells form mature adipocytes 5 days post induction; 3T3-L1 cells require 7–8 days to form adipocytes.

For Oil red O staining, pre- and post-differentiated C3H10T1/2 and 3T3-L1 cells, WT and Taf7l KO adipose-derived primary fibroblasts, shGFP and shTAF7L-treated C3H10T1/2 cells, C2C12.CNTL and C2C12.TAF7L cells were washed once in PBS and fixed with freshly prepared 4% formaldehyde in 1PBS for 30 min, followed by standard Oil red O staining method described previously.[164]

For C2C12 myogenesis, C2C12 cells are cultured in maintenance media until confluence was reached. 2 days post confluence, cells were switched to differentiation media comprised of low glucose DMEM, 2% horse serum, and 5 g/ml insulin, 3 days later, change fresh differentiation media and culture cells for additional 2 days, differentiated myotubes were harvested and purified by collecting the suspended cells after splitting and reseeding the cells for 1 hr.

RNA isolation and real-time PCR analysis

Total RNA from cultured cells or mouse tissues was isolated using QIAGEN RNeasy Plus mini columns according to the manufacturer's instructions (Qiagen Inc., Germantown, MD). For RT-qPCR analysis, 1 g total RNA was reverse transcribed using cDNA reverse transcription kit (Invitrogen, Carlsbad,, CA). SYBR green reactions using the SYBR Green PCR Master Mix (Applied Biosystems, Warrington, UK) were performed according to the manufacturer's instruction using an ABI 7300 real time PCR machine (Applied Biosystems, Foster City, CA). Relative expression of mRNA was determined after normalization to total RNA amount. Student's t-test was used to evaluate statistical significance.

Western blot analysis, immunoprecipitation

Whole cell extracts were prepared from cells by homogenization in lysis buffer containing 50 mM Tris–Cl, pH 8.0, 500 mM NaCl, and 0.1% Triton X-100, 10% glycerol and 1 mM EDTA, supplemented with protease inhibitor cocktail (Roche, Indianapolis, IN) and phenylmethylsulphonyl fluoride (PMSF). Fifteen micrograms (g) of whole-cell lysates were separated by SDS-PAGE and transferred to nitrocellulose membrane. For immunoblotting, membranes were blocked in 10% milk, 0.1% Tween-20 in TBS for 30 min, and then incubated with TAF7L, TAF4, TAF7, FLAG, -actin, PPAR and TBP antibodies for 2 hr at room temperature; detailed Western blotting procedure was performed as previously described (Zhou et al., 2006).

500 g whole-cell extracts from 293T cells transfected with FLAG-TAF7L and HA-PPAR, FLAG-PPAR and HA-TBP, or FLAG-TAF7L and HA-TBP were immunoprecipitated with FLAG or HA antibodies at 4C for overnight under the conditions of 0.3 M NaCl and 0.2% NP-40, 30 l protein A/G beads were added and incubated for additional 2 hr at 4C, after extensive washing with buffer containing 0.15 M NaCl and 0.1% NP-40, remaining beads were subjected to 10% SDS-PAGE and followed by western blotting analysis with FLAG and HA antibodies to detect tagged-proteins in the inputs and IPs as previously described (Ding et al., 2008).

Animals and genotype analysis

The derivation of *Taf7l*-knockout mice has been previously described.[101] All animal experiments were performed in strict accordance with the recommendations in the Guide for the Care and Use of Laboratory Animals of the National Institutes of Health. All of the animals were handled according to approved animal use protocols (R007) by Animal Care and Use Committee (ACUC) of the University of California, Berkeley. Mice were maintained on a standard rodent chow diet with 12 hr light and dark cycles. Taf7l KO mouse line was maintained on a C57/Bl6 background. Genotyping was performed by PCR as previously described.[101]

Preparation of primary fibroblast and induction of adipogenesis

Fresh inguinal adipose tissues were removed from 3 week old euthanized WT and *Taf7l* KO mice and finely minced, digested with 0.25% trypsin for 30 min at 37C, and centrifuged for 5 min at 2,000g. The pellet was resuspended in culture media before plated on gelatin coated plates. Cells were cultured at 37C in high glucose DMEM supplemented with 20% FBS. Adipocyte differentiation and staining were followed the same procedure as C3H10T1/2 cells.

Immunohistochemistry

For histological analysis on interscapular tissue of E18.5 embryos from WT and *Taf7l* KO mice, freshly-harvested mouse embryos were genotyped and then interscapular regions of embryos were transversally dissected and then fixed in 10% formaldehyde for 24 hr at 4C; tissue was embedded in paraffin using the microwave method and then sectioned into 8–10 m sections to mount on slides. This method and the following immunohistochemistry by haematoxylin and eosin (HE) staining were performed using the method described previously by Steven Ruzin,[165] and FABP4 immunostaining method was modified from the one described previously.[119, 104]

ChIP, ChIP libraries preparation, and deep sequencing (ChIP-seq)

Fix C3H10T1/2 cells and differentiated adipocytes with 1% formaldehyde for 10 min at room temperature then use 0.125 M glycine to stop the crosslinking for an additional 5 min. Collect cells and extract nuclei with extraction buffer. The chromatin obtained from C3H10T1/2 cells and adipocytes was fragmented to sizes ranging from 175 to 225 bp using a Covaris-S2 sonicator (Covaris, Inc., Woburn, MA) for a total processing time of 40 min (20 s on, 20 s off). 900 g of the sonicated chromatin was used in each immunoprecipitation reaction as previously described[40] with the Pol II, TBP, PPAR, and TAF7L antibodies, mouse and rabbit IgG were used as negative controls respectively. Preparation of the sequencing libraries on the DNA samples of the immunoprecipitation from antibodies and IgGs precisely followed the instructions from Illumina (Illumina Inc., San Diego, CA), qualities of the libraries were assessed by 2100 Bioanalyzer (Functional Genomics Laboratory, Berkeley,CA) and then subjected to ultra-high throughput sequencing on an Illumina HiSeq 2000 sequencer (GSL core facility, Berkeley, CA) as previously described

Chapter 3

Utilizing Multiply-aligned Sequences Reveals Overlooked Gene Regulatory Features

3.1 Abstract

High throughput analysis of gene regulation depends on optimized experimental conditions, efficient library sequencing, accurate read mapping, and robust identification of functional DNA sequence elements. Here we focus on the alignment step, which is frequently compromised by tandemly repeated DNA elements and sequencing errors that are discarded as multiply-aligned sequences. In order to more accurately preserve and utilize the information in such repeated sequences, we have extended eXpress–a tool designed primarily for probabilistically assigning ambiguous reads in RNA-Seq–to an analysis pipeline suitable for any genome-wide assay. We tested our approach on ChIP-Seq and MeDIP-Seq experiments to reveal new functionally relevant regions that are missed by standard unique alignment programs.

3.2 Introduction

High throughput sequencing enables unbiased genome-wide analysis of gene regulation, with new modalities–so-called ∗Seq assays–that have been developed to evaluate a range of questions, such as RNA production and protein:nucleic acid interactions.[2,3,4] One application of ∗Seq assays is to define significantly enriched regions and subsequently compare and contrast those targeted regions with RNA transcripts and previously defined cis-regulatory or other annotated elements.[8,9,10] Experimental inefficiencies inevitably lead to a loss of information, but computational and algorithmic bottlenecks can also degrade signal. During the alignment step, between 10 and 20 percent of the data is lost on average due to multiply-aligned sequences.[166,7,167] Teasing apart ambiguous alignments is of particular interest when studying sequence-specific transcription factors and co-activator complexes that target core promoters or cell type specific enhancer elements.[12,40] Assigning multiply-aligned sequences can identify new regulatory elements in portions of the genome that are poorly annotated or highly conserved within gene families and further enhance the enrichment of regions with significant coverage from unique alignments, providing a more detailed picture.

RNA-Seq data processing requires the use of multiple alignments to estimate gene abundances in the presence of alternative splicing.[15,16,17,18] For this reason, methods for handling multiply-aligned sequences have been highly developed in that setting. However, support for other genome-wide ∗Seq experiments has been limited primarily to tools utilizing the batch Expectation Maximization (EM) algorithm.[166,168] The batch EM algorithm is computationally intensive when determining the likelihood of an individual alignment. eXpress, a new streaming online-EM algorithm for RNA-Seq analysis, is capable of combining various steps within the pipeline to decrease the computational requirements.[17] For RNA-Seq, eXpress estimates the likelihood for alignments based on the local abundance of sequences within a transcript. Other *Seq experiments, however, do not have well annotated functionally relevant target regions, making it necessary to divide the reference genome into "bins". To more comprehensively capture a representative view of all binding regions, we combine the parameter estimates of neighboring bins. Extending the functionality of eXpress to full genome application required modifications to ensure that regions straddling bin edges are handled correctly, the establishment of appropriate bin sizes and overlaps, investigation of the extent and effect of sequence-specific biases that derive from experimental idiosyncrasies, and the development of useful outputs for downstream processing.

In this work we demonstrate solutions to the above problems and present an effective eXpress-based pipeline for evaluating multiply-aligned sequences across a complete reference genome. The results suggest that eXpress' robust statistical model and computational efficiency improve the ability to restore previously overlooked functionally significant sites in a variety of ∗Seq modalities.

3.3 Results

Validating eXpress parameters

To establish the appropriate set of parameters for incorporating multiply-aligned sequences in genome-wide ∗Seq experiments (Figure 3.1A), we explored a RNA polymerase 2 (Pol2) ChIP-Seq data-set derived from mouse embryonic stem (mES) cells. We tested the ability of eXpress to recover both peaks and the genome-wide distribution of reads once ambiguity was introduced into the read sequences (Figure 3.2 A). eXpress was run by sampling the output, an extra round of EM, various bin sizes and neighbor inclusion.

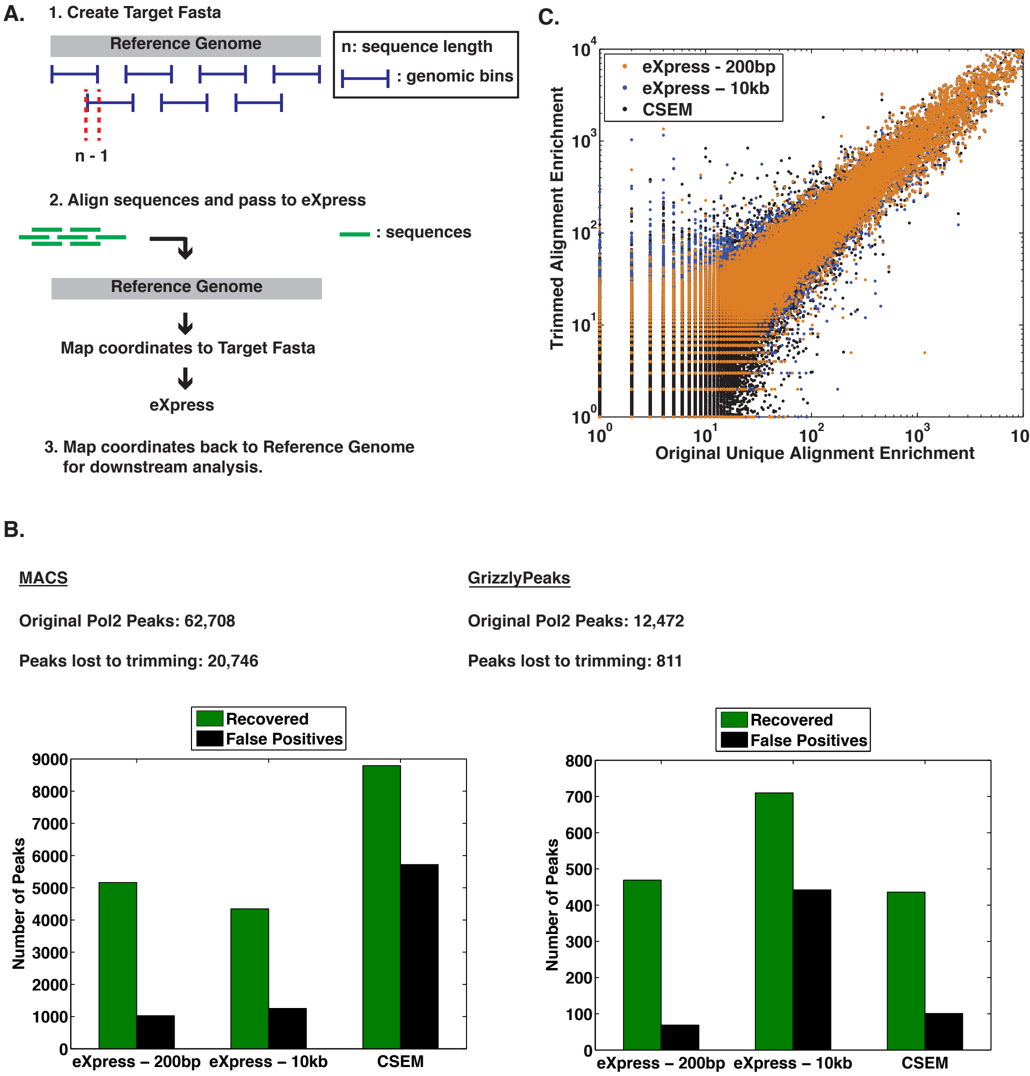
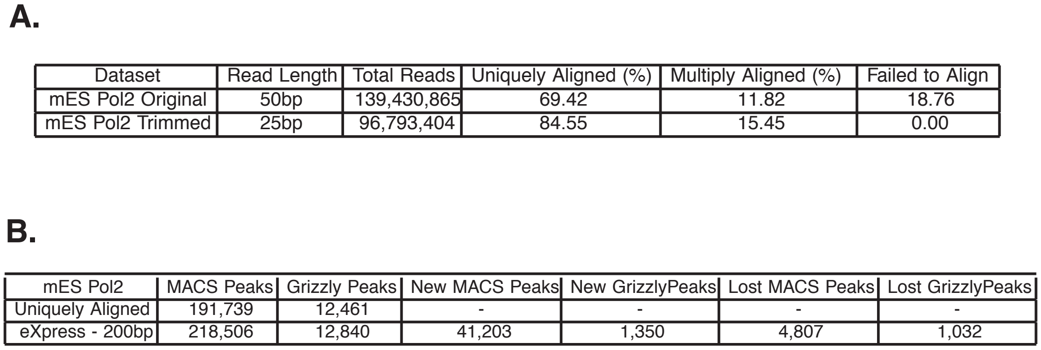


Figure 3.1: **eXpress recovers trimmed unique alignments in ChIP-Seq most effectively at a 200bp binning A.** Illustration of the new eXpress pipeline for handling multiply-aligned reads with the full genome as a reference. **B.** Processing uniquely aligned trimmed mES Pol2 ChIP-Seq with eXpress and CSEM[166] to evaluate the ability of each to recover peaks lost via either MACS (33% peaks lost) or GrizzlyPeaks (7% peaks lost). eXpress - 200bp, taking into account the neighbors and overlap between the bins, is equivalent to the length of most ChIP-Seq peaks, whereas 10kb represents a significant decrease in memory requirements. **C.** The multiply-aligned enrichments at a 250bp resolution across the entire genome with relation to the original uniquely aligned distribution.



A.

Dataset	Read Length	Total Reads	Uniquely Aligned (%)	Multiply Aligned (%)	Failed to Align
mES Pol2 Original	50bp	139,430,865	69.42	11.82	18.76
mES Pol2 Trimmed	25bp	96,793,404	84.55	15.45	0.00

B.

mES Pol2	MACS Peaks	Grizzly Peaks	New MACS Peaks	New GrizzlyPeaks	Lost MACS Peaks	Lost GrizzlyPeaks
Uniquely Aligned	191,739	12,461	-	-	-	-
eXpress - 200bp	218,506	12,840	41,203	1,350	4,807	1,032

Figure 3.2: **Summary of alignment and peak calling results. A.** Results of the alignment when mES Pol2 sequence reads are trimmed from 50bp to 25bp using bowtie on the mm9 annotation. **B.** Summarizing the new putative binding regions as defined by MACS and GrizzlyPeaks after processing the multiply aligned mES Pol2 sequences with eXpress using a 200bp binning, sampling the reads after an extra round of batch EM and using one neighbor.

Independent of peak caller, eXpress with 200bp binning is more specific at recovering peaks than eXpress at 10kb or CSEM, one of the existing EM methods[166] (Figure 3.1B). eXpress at 200bp recovers nearly 6 fold more peaks over generated false positives, a 2-3 fold improvement over CSEM and eXpress at 10kb. eXpress at 200bp is also better correlated globally than CSEM and eXpress using a 10kb binning to the original alignment (Figure 3.1C). The superior specificity and better correlation indicate that a smaller binning of the reference genome leads to more accurate assignment of multiply-aligned sequences.

Sequence bias correction decreases the memory efficiency of eXpress once the size of the reference genome gets larger. We tested if disabling sequence bias correction would improve performance and affect the results. Figure 3.3A shows that the ChIP-Seq sequence bias parameters are consistent with the expected value, whereas the sequence bias correction for RNA-Seq is clearly important. Running the recovery from Figure 1 with sequence bias correction disabled also provided very similar results (Figure 3.3B-C).

Identifying enrichments in ChIP-Seq overlooked by unique alignments

Having established a useful set of parameters for analyzing ChIP-Seq data, we next evaluated what new information may be revealed by our method. To do this, we ran the previously mentioned Pol2 ChIP-Seq dataset from mES cells, along with the corresponding IgG control, through our pipeline. Peak results are summarized in Figure 3.2B.

To evaluate potential new Pol2 target genes, we annotated the new MACS peaks against Ensembl. We focused on a region within 5kb of the transcriptional start site (TSS) (14,520

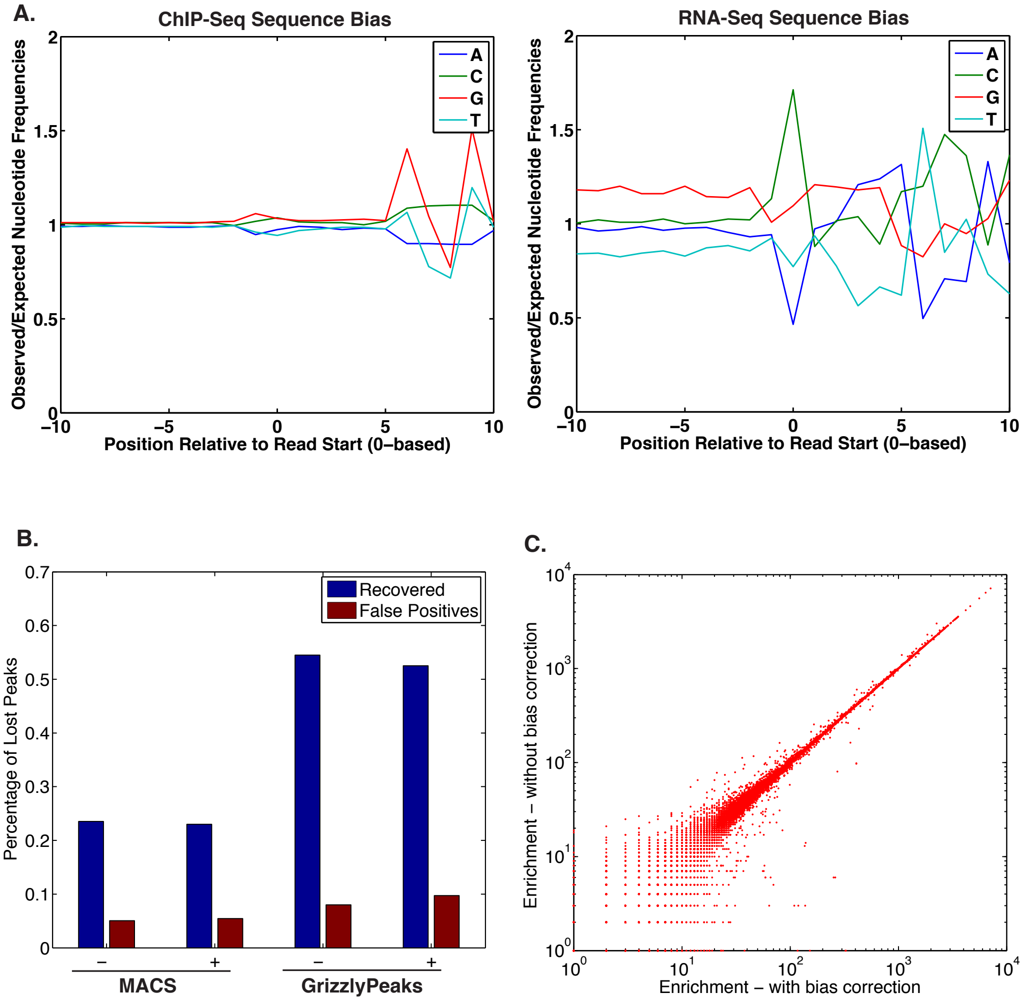


Figure 3.3: **eXpress performs comparably in recovering trimmed unique alignments with or without sequence bias correction. A.** Sequence bias as it refers to ∗Seq experiment. The observed and expected nucleotide frequencies as a function of base pair position were defined within the param.xprs output from eXpress. The ChIP-Seq experiment is the mES Pol2 experiment discussed here. The RNA-Seq experiment was done as described in Roberts and Pachter, 2013. **B.** Running eXpress - 200bp with (+) and (-) without bias correction shows an equivalent percentage of recovery and false positives irrespective of which peak calling method is used. **C.** Comparing the read enrichments globally also shows a strong correlation between both conditions.

MACS peaks annotated against 10,432 genes). Approximately 15% of these genes are lincRNAs, snRNAs, miRNAs, snoRNAs, rRNAs, pseudogenes and retrotransposons compared to 8% that we see when focusing only on unique alignments. A few of the new targets were then validated using ChIP-qPCR (Figure 3.4A). At each locus there was a significant enrichment, at least 3-fold, of the Pol2 signal over the IgG control. These findings suggest that incorporating multiply-aligned sequences into ChIP-Seq analyses identifies previously overlooked and/or underrepresented signals at certain genomic loci.

In addition to evaluating new gene targets for transcription factors, we also see enhanced enrichments in peaks that were previously identified by unique alignments. A small subset of these peaks showed more than a 3 fold increase in maximum summit value. Many of these genes are involved with nucleosome assembly and gene silencing (Figure 3.4B). All 50 of the GrizzlyPeaks binding sites were located within 5kb of the TSS, indicating a high degree of conservation at these promoter and proximal elements. Enhancing the signal can be particularly important when experimental inefficiencies may mask identifying such regions as potential functional binding regions.

Recovering a methylation signal on the *Oct4* distal enhancer

Specific promoter DNA methylation is an important stable epigenetic mark of gene silencing and represents a key mechanism regulating cell fate determination.[169] Regions rich in methylated CpG islands have a propensity to sustain sequencing errors and mismatches. We note that MeDIP-Seq data of the Oct4 distal enhancer lacks a signal when using unique alignments in human dermal fibroblasts (HDFs), (Figure 3.4C). Our pipeline was able to recover part of the signal assigned to such loci including at Oct4. Bisulfite sequencing confirmed that the distal Oct4 enhancer is in fact methylated in HDFs (Figure 3.4C) further confirming the disadvantages of restricting data coming through the alignment step.

3.4 Discussion

In this work we have established the parameters needed to handle multiply-aligned sequences using eXpress for various genome-wide *Seq experiments. We show the value of including ambiguous alignments, normally discarded, by applying our computational tool to ChIP-Seq and MeDIP-Seq analysis. In the case of ChIP-Seq, our pipeline identified many more specific transcription factor binding sites and transcript sequences across a wide range of gene classes including protein coding sequences, snRNAs, lincRNAs, miRNAs, snoRNAs, rRNAs, pseudogenes and retrotransposons. By developing analytical tools to capture more of the inherent *Seq signal, we also minimize the impact that arises from various intrinsic and inevitable experimental inefficiencies that often plague biological/molecular assays (Figure 3.4C).

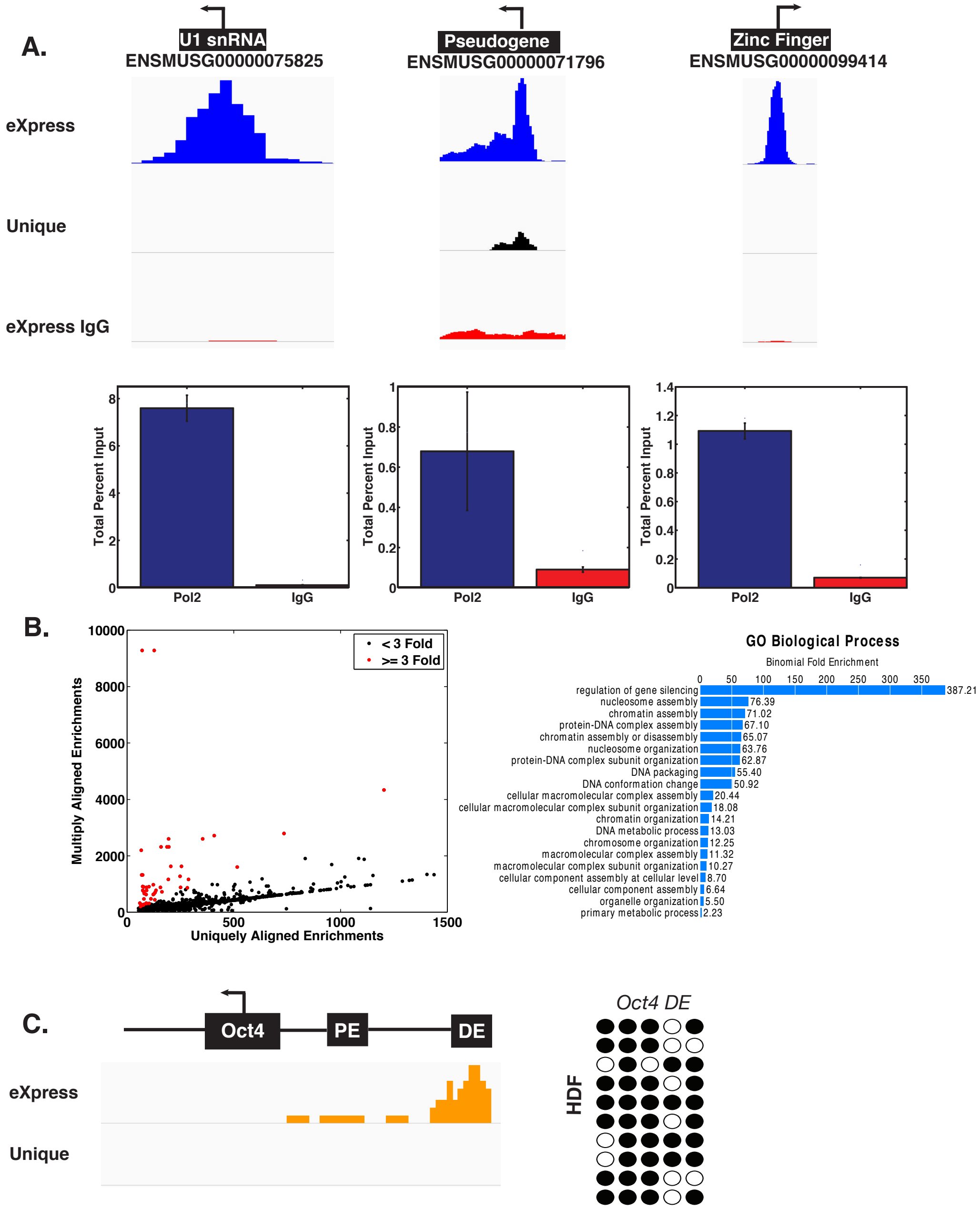


Figure 3.4: **Multiply-aligned sequences identify genomic regions via ChIP-Seq and MeDIP-Seq that were previously going unnoticed. A.** Three representative loci that were identified as new mES Pol2 ChIP-Seq peaks when utilizing multiply-aligned sequences and eXpress. ChIP-qPCR bar graphs present the Pol2 and IgG control immunoprecipitation enrichments in terms of a percentage of the total input DNA library. **B.** Comparison of the summits assigned assigned by GrizzlyPeaks that overlap between the multiply-aligned and unique cases identifies 50 peaks that have summit values increased by 3-fold or more when allowing for multiply aligned sequences. The GO-terms[10] identify the targeted genes to function primarily in nucleosome assembly and gene silencing. **C.** Limited to only unique alignments in the MeDIP-Seq experiment on human dermal fibroblasts (HDF) fails to reveal methylation at the *Oct4* distal enhancer. This distal enhancer is recovered when multiply-aligned sequences are processed by eXpress. The alignments (genomic tracks, left panel) and the bisulfite sequencing (right panel, dark circles are methylated) confirm that this region is in fact heavily methylated.

Most often, deep mechanistic questions in molecular biology require a suit of genome-wide ∗Seq experiments to be applied in combination. An additional advantage of our pipeline is that we are now able to process all of the data using the same tool. Deploying the same handling of mismatches, likelihood assessments and biases, helps manage the propagation of errors across various computational methods. By increasing the accuracy of handling multiply-aligned sequences for genome-wide ∗Seq experiments, we substantially improve the ability to interpret gene regulatory mechanisms by providing a more complete context for large sets of disparate data. As with any computational tool, there are limitations, and a complete analysis of typical ∗Seq studies will continue to require rigorous confirmation by various direct biochemical, genetic and molecular reconstitution experiments combined with other computational methods.

3.5 Methods

Aligning the sequences

The sequences, obtained in FastQ format from either Gene Expression Omnibus (GEO) or the sequencing facility, were mapped to the reference genome (mm9 or hg19) using Bowtie.[7] Two mismatches were allowed in both the unique and multiply-aligned cases. Up to 100 possible alignments were stored in the multiply-aligned BAM[170] file by using the -k 100 option in Bowtie.[?]

Full genome binning

eXpress calculates the local abundance and subsequent likelihood for each alignment over defined regions of the genome. With its initial application to RNA-Seq, those regions were transcripts. When the target sequence is the entire genome, it is necessary to bin the genome into uniquely mappable bins to avoid compounding errors when calculating the likelihoods with eXpress. The bins overlap by 1bp less than the length of the sequence. This was done using a custom python script: BinnedFasta.py. The required parameters were the read length (-l), the bin length (-b) and the reference genome (-g). Available at http://github.com/ivangrub/express-full-genome.

Converting coordinates between reference genome and binned genome

eXpress requires the same coordinate space to map the aligned reads to the binned genome. BinMapping.py takes as input the read length, bin length and reference genome that were used to create the binned genome. Additionally, the BAM file that is output by Bowtie is passed as input. Available at http://github.com/ivangrub/express-full-genome.

Running eXpress

Most genome-wide sequencing methods have reads distributed around enriched regions. To account for this, we calculated "local abundances" by evaluating read mappings to each local bin along with its two neighbors. To do so, we added a new option to eXpress (–num-neighbors) that combines a given number of neighboring target sequences on either side into a single bin when evaluating the alignment. The average fragment length was set to be the length of the read with a standard deviation of two. The output alignment file was saved via either selecting a single alignment by sampling from the posterior probabilities (–output-align-samp) or outputting the probability of each alignment (–output-align-prob). In either case, an extra round of batch EM was run (-B 1).

Converting coordinates back to reference genome from eXpress

eXpress2wiggle.py converts the eXpress outputted alignment's coordinates back to the reference genome and generates wiggle and bedGraph files for data visualization. To save the updated BAM file, the -bo y option is used. Available at http://github.com/ivangrub/express-full-genome.

Peak calling

The peak calling functions used were dependent on whether eXpress outputs the probability for each alignment or samples an individual alignment for each input sequence. If the probabilities were saved, then a shape based algorithm was required. Some examples include GrizzlyPeaks or Mosaics.[8,?] When the reads were sampled, the processed BAM file could be inputted into any traditional peak caller (e.g. MACS v1.4, SPP, etc.)[9,171]The default parameters were used for the different peak callers.

Data visualization and annotations

The IGV browser[172] was used to visualize the bigWig files generated by using bedGraph2bigWig from UCSC.[173] It converts the eXpress2wiggle.py generated bedGraph files to bigWig. Annotating the peaks to a gene reference (e.g. Refseq, Ensembl, UCSC) was done using custom python or Mathworks MATLAB scripts. Graphs were generated either with MATLAB or python with the matplotlib module.

Gene ontology

Gene ontology was performed using GREAT.[10] The background was defined as the full genome of genes coming from the mm9 mouse annotation. Associations with genes were defined by selecting the single closest gene within 5kb of the peak.

ChIP experiments

D3 mouse embryonic stem (mES) cells were cultured in 10x 150-mm2 dishes until 80% confluency and then fixed with 1% formaldehyde/PBS for 10 min at room temperature. Fixation was stopped by adding glycine at 0.125 M final concentration. Cells were washed twice with ice-cold PBS and collected by scraping and centrifugation. Cell pellets were lysed in PBS containing 0.2% Triton X, 1mM PMSF, and 1x protease inhibitor (Roche) for 20 min on ice. Nuclei were recovered by centrifugation and lysed in Lysis buffer (50mM Hepes/KOH ph7.6, 140mM NaCl, 1mM EDTA, 0.5mM EGTA, 1% Triton X100, 0.1% Na deoxycholate, 0.5% Sarcosyl, 0.1% SDS, 1x protease inhibitors and 1mM PMSF) for 10 min on ice. Nuclear lysates were sonicated to shear chromatin to an average fragment size of 500 base pairs. Sonicated samples were cleared by centrifugation, quantified by measuring A260, and 1mg of DNA/protein was subjected to immunoprecipitation by incubating with the following antibody-bead mixtures: Pol II (8WG16) and mouse IgG. 100 ul of Protein G-Dynabeads (Invitrogen) and 10 ug of each antibody were pre-incubated for 40 min at room temperature before added to the samples. After overnight incubation at 4C, immunoprecipitated samples were washed 3x with ChIP wash buffer (10mM TrisHCl pH8, 1mM EDTA, 140mM NaCl, 1% Triton X100, 0.2% Sarcosyl, 0.1% Na deoxycholate), 3x with ChIP wash buffer supplemented with 500 mM NaCl, 2x with LiCl wash buffer (10mM TrisHCl pH8, 1mM EDTA, 250mM LiCl, 1% NP-40, 1% Na deoxycholate), and 2x with TE buffer at room temperature. DNA elution, reverse crosslinking, and DNA purification were performed as described in the Abcam protocol (www.abcam.com/ps/pdf/protocols/x$_c$*hip*$_p$*rotocol.pdf*).

ChIP-Seq library Preparation

Purified ChIP DNA was prepared with NEBNext ChIP-Seq Sample Prep Master Mix Set 1 (E6240, New England BioLabs) according to the manufacturer's instructions. After the library preparation for each sample, size, purity, and concentration of the DNA libraries were checked by Agilent Technologies 2100 Bioanalyzer and quality- certified samples were submitted to sequencing (VCGSL facility, UC Berkeley). Sequencing was carried out with the Illumina HiSeq 2000 sequencing platform (single end-reads, 50 bp long).

ChIP-qPCR

ChIP-qPCR was performed using the ABI Biosystems 7500 qPCR machine. The ChIP DNA was amplified using the KAPA SYBR mix from KAPA Biosystems. PCR primers were designed to flank the enriched regions. The U1 snRNA primers were CCAGTGACCTGACAAACCCA and GCTAGTGGTTTGGAGAGGGG. The lincRNA primers were TTACAGAACAACAACAAAAGGTGT and GGCAGGGGACTTCTCAGGA. The zinc finger primers were CACCCACTTCCGCTTCTCAT and CACCCTCTCAAGGCTTCTGG.

MeDIP-Seq library preparation

Normal human dermal fibroblasts (Lonza) were cultured under standard conditions in DMEM supplemented with 10% FBS (HyClone) and GlutaMAX (Life Technologies). DNA was purified using the DNeasy Blood and Tissue Kit (Qiagen) and sonicated to an average fragment size of 150 base pairs using the Covaris S2, according to the manufacturer's protocol. Fragmented DNA was then end-repaired and ligated to Illumina TruSeq adaptors using the NEBNext End Repair, dA-Tailing, and Quick Ligation Modules (NEB). Adaptor-ligated DNA was heat denatured at 95C for 10 minutes, rapidly cooled on ice, and immunoprecipitated with 1 g mouse monoclonal antibody to 5-methylcytosine (Epigentek) per microgram of DNA, overnight at 4C in cold IP buffer (10 mM NaH2PO4, pH 7.0, 140 mM NaCl, 0.05% Triton X-100). To recover the immunoabsorbed DNA, 10 l precleared sheep anti-mouse IgG Dynabeads (Invitrogen) were added to each sample and incubated for an additional 2 hr at 4C. Each IP was washed eight times in cold IP buffer and DNA was then recovered by treatment with 2 l Proteinase K (Qiagen) at 56C for 30 minutes in digestion buffer (20 mM HEPES, pH 7.9, 1 mM EDTA, 0.5% SDS). Sixteen cycles of PCR were performed for library amplification using Kapa HiFi HotStart DNA polymerase (Kapa Biosystems) and the Illumina TruSeq PCR primers. Size selection for 220-420 base pair fragments was then performed by electrophoresis using a 2% agarose gel. The size, purity, and concentration of the DNA libraries were QC'd by Agilent Technologies 2100 Bioanalyzer and submitted to sequencing (VCGSL facility, UC Berkeley). Sequencing was carried out with the Illumina HiSeq 2000 sequencing platform (single end-reads, 50 bp long).

Bisulfite sequencing

Bisulfite conversion was performed on purified DNA using the EpiTect Fast Bisulfite Kit (Qiagen) according to the manufacturer's protocol. Bisulfite-converted DNA was then amplified using previously published primers for the Oct4 distal enhancer from,[174] and cloned into the pGEM-T Easy Vector System (Promega). 10 independent clones were sequenced and analysed for their methylation status. Closed circles represent methylated CpG dinucleotides, while open circles represents unmethylated CpG dinucleotides; each row indicates an independent clone.

Chapter 4

Identify the transcriptional link to mechanotransduction

4.1 Proposed Projects

Qualifying Exam Proposal

Specific Aims

The extracellular matrix (ECM) not only serves as a scaffold that supports the cells and a variety of biological and chemical cues, it also transduces physical signals to the cells. The mechanism of mechanotransduction is currently understood as a function of the cell's focal adhesions and cytoskeletal tension. The physical properties of the ECM do more than just change the morphology and cytoskeletal organization of the cells, but also tune the cell's transcriptome to the physiologically relevant stiffness. How these physical signals manifest themselves as transcriptional regulatory factors to achieve this tuned expression is the unknown question. My analysis will provide insights about whether there are specific transcriptional factors, epigenetic changes facilitated by histone modifying proteins and/or post-translational modifications of general transcription factors that are responsible for this transcriptional tuning.

- **Aim 1**: Use RNA-seq to identify transcriptional regulatory factors involved in the stiffness dependent tuning of myogenic and osteogenic differentiation programs.

- **Aim 2**: Verify via a siRNA screen whether the RNA-seq candidates knock out the ability of the cell to tune the transcriptome to the physiologically relevant stiffness.

Background

Introduction

Cells are capable of responding to both chemical and biological cues along with physical cues that come from the ECM. These physical cues are mediated by integrin receptors which connect to actin through the combination of vinculin and talin.[22] Changes in the physical environment's stiffness introduce a strengthening or weakening of contractile forces on the cytoskeleton.[21] This leads to a cascade of events that include the activation of the Rho GTPase Rac which activates actin polymerization and the strengthening of focal adhesions. Subsequent activation of Rho and the downstream Rho-associated kinase, induces the formation of actin stress fibers via myosin motor activation.[175] These stress fibers stabilize the cell's shape in the new physical environment. The degree of cell spreading is then entirely dependent on the substrate stiffness, because as the cell contracts, it is attempting to deform the ECM. In the event that the cell cannot deform the substrate, the cytoskeletal reorganization will settle on a spread out cell shape.

In addition to visible changes in cell morphology, the changes in the environment's stiffness also facilitate changes in gene expression. When mesenchymal stem cells (MSCs) are cultured on a stiffness similar to a particular physiological tissue, it tunes the transcriptome towards that tissue's gene expression profile.[14] Recent work has also shown that under competitive differentiation conditions, the physical cues are able to synergistically work with the chemical cues,[176] and potentially block them if the soluble cues are added later.[14] Whether the substrate stiffness[19] or the cell shape[20] is the causal parameter for the genetic tuning is still unclear; it seems to be at least somewhat dependent on the strength of the focal adhesions[176] and the cell's cytoskeletal tension.[177]

The mechanism by which contractile forces are translated into the nucleus, however, is still largely unknown. The ability for cells to respond to mechanical changes in the ECM is important in both disease and development.[178,179] Since there are both cell fate and cell maintenance decisions involved, there may be multiple pathways by which the mechanosensing is translated into gene expression. One proposed mechanism is the binding of Rho GTPase associated proteins to transcription factors involved in regulating expression of a gene of interest. An example involves the binding of p190RhoGAP, an inhibitor of RhoA, to the antagonistic transcription factors TFII-I and GATA2 to regulate the expression of VEGF receptor *VEGFR2*.[180] The elastic dependent expression of the VEGF receptor has a direct effect on the ability for angiogenesis to occur. Other proposed models include a physical connection linking the plasma membrane to the nucleus, and subsequently chromatin, via the cytoskeleton[181,182] or the presence of stiffness-dependent, specifically targeting, histone modifying proteins that would change the epigenetic structure of the chromatin.[183] Changes in nucleosome organization can increase the accessibility of transcriptional start sites for the basal transcriptional machinery and thereby effect the cell transcriptome.

I am proposing to use C3H10T1/2 cells, which have mesenchymal stem cell properties, to study the transcriptional mechanisms of myogenic[184] and osteogenic[185] tuning as a function of substrate stiffness. The cells will be cultured on type I collagen coated polyacrylamide gels of varying stiffness. I will start identifying the sensing window, defined as the time it takes for the early progenitor markers, Pax7[186] and Runx2[187] (myogenic and osteogenic markers, respectively) to become up-regulated. The progenitor markers define the sensing window because they indicate a fate decision, while avoiding the noise that will come from up-regulated transcripts associated with the differentiation program. I will then utilize RNA-seq and bioinformatic techniques to identify candidate transcriptional regulatory factors that are responsible for executing the tuned gene expression. A siRNA screen will be used to verify the candidates which will eventually lead to further single gene experiments and ChIP pull-downs to elucidate the actual mechanism of action.

Impact

Understanding the transcriptional machinery or mechanisms by which changes in the ECM's stiffness are translated can elucidate how certain disease states spread. This can be applied to cancer studies along with muscular dystrophy and several other illnesses.[178] It can also improve the current strategies in tissue engineering and regenerative medicine for stem cell therapies.

Research Methods

Aim 1

Use RNA-seq to identify transcriptional regulatory factors involved in the stiffness dependent tuning of myogenic and osteogenic differentiation programs.

Before RNA-seq experiments can be performed, acrylamide and bis-acrylamide concentrations, and type I collagen concentration need to be optimized for the appropriate stiffness and functionalization of the gels.[188] The polyacrylamide gels will be approximately 10 kPa and 35 kPa for the myogenic and osteogenic pathways, respectively, to fit into the their physiological ranges.[14] The stiffness of the gels will be characterized using atomic force microscopy (AFM)[20] and a curve fitting script in MATLAB. As a negative control, the C3H10T1/2 cells will be cultured on a type I collagen coated tissue culture dish. Tissue culture dishes have a stiffness on the order of megapascals. Figure 4.1 illustrates the theoretically predicted difference in qRT-PCR data that should occur after culturing the C3H10T1/2 cells on the two substrate stiffnesses. The time that it takes for the respective marker to be up-regulated will be defined as the end of the sensing window (T_f).

Establishing T_f will serve as a positive control that I am capable of reproducing the physiological tuning phenomenon with the C3H10T1/2 cells. The myogenic and osteogenic lineages also have different activities of RhoGTPases. Rac1 activity is specific for the myogenic[189]

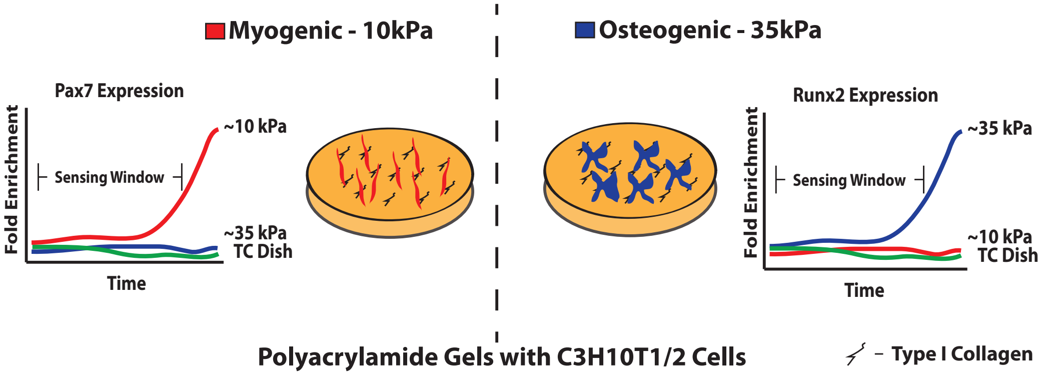


Figure 4.1: **Illustration of the expected qRT-PCR results.** C3H10T1/2 cells will be cultured on 10 kPa gels for the myogenic tuning (red) and 35 kPa gels for the osteogenic tuning (blue). The tissue culture (TC) dish, coated with type I collagen, will be the control (green). The expected qRT-PCR data will show an upregulation of Pax7 and Runx2 for myogenic and osteogenic pathways, respectively.

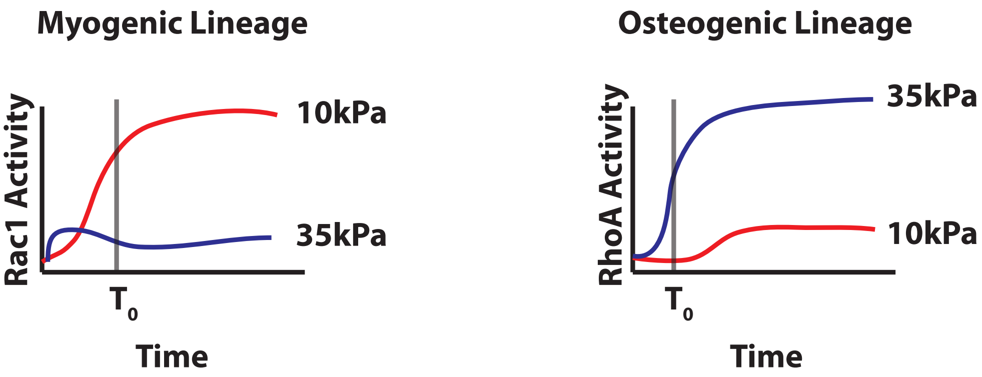


Figure 4.2: **Determine T_0.** Defining T_0 for the myogenic and osteogenic differentiation programs using Rac1 and RhoA activity levels as the earliest markers that indicate cell adhesion to the substrate and the first series of posttranslational modifications that are occuring in response to the substrate stiffness. These graphs are cartoon renditions of expected results.

differentiation, while RhoA activity is required for osteogenic[177] differentiation from mesenchymal stem cells. Since the RhoGTPases are functionally linked to the integrins and subsequently the initial stiffness response, I will define T_0 (Figure 4.2) as the time point that Rac1 and RhoA activity are being increased. Keeping in mind that transcriptional regulation occurs on a time frame of about 4 hours, T_1 will be taken 4 hours after T_0. If additional time points are needed, they will be taken in between 2 and 4 hour windows to increase the temporal resolution. The need for more data points will be a function of how successful the RNA-seq analysis ends up being.

The RNA-seq data from these time points will be mapped back to the genome using TopHat[190] and then analyzed using Cufflinks.[191] Both of these programs are provided by the Center for Bioinformatics and Computational Biology. The processed data will elucidate both the bulk differences in the transcriptome amongst the conditions but also significant differences in the transcriptional kinetics between time points.

The large amount of data that is collected by the RNA-seq will require multiple filtering steps to make the analysis feasible. One of the techniques that I plan to use, is to superimpose the RNA-seq data onto previously published ChIP-seq data of factors that are known to bind to the regulatory elements of myogenic and osteogenic genes. The overlay of this data will allow me to identify interactions that will help identify the sequence of transcriptional events at early time points. Using a variance based analysis technique like principal components along with k-means clustering could uncover some potentially novel interactions. Many of these interactions though will indicate secondary or tertiary effects. To be able to discover some of the primary effects, it would be necessary to look at the whole transcriptional regulatory network and to *a priori* take the basally expressed machinery into account as well in an effort to study the upstream effectors. The analysis will select several candidate transcriptionally related factors and I will verify them using an siRNA screen as described in the second aim.

Potential Issues

The physical properties of the polyacrylamide gels are sensitive to the efficiency of the polymerization reaction, which in part requires the use fresh ammonium persulfate every time. The presence of bubbles in the gel needs to be addressed by degassing the polyacrylamide mixture before the addition of TEMED. Figure 4.3 illustrates some of the inherent variability in the making of the gels.

Another concern would be the importance of the cell shape on the differentiation process. Confluent C3H10T1/2 cells can still have a slightly elongated shape which is considered to be required for the myogenic differentiation.[189] Osteoblasts on the other hand need to be spread out which will not be possible in confluent cultures.[177]

Previous researchers have looked to simplify the problem by not allowing the cells to reach confluency. This could be done by treating the cells with either mitomycin C or aphidicolin, a DNA polymerase inhibitor, to stop the cells from proliferating.[14,177] In both cases extensive controls would have to be done to make sure that the treatment is not effecting other cellular processes. This is also an important step to take because many of the differentiation programs require the cells to stop proliferating so it depends on either contact inhibition via confluency or drug induced cellular arrest.

Independent of which cell arrest mechanism is used, heterogeneous populations following differentiation could always be a concern. One potential solution would be to create

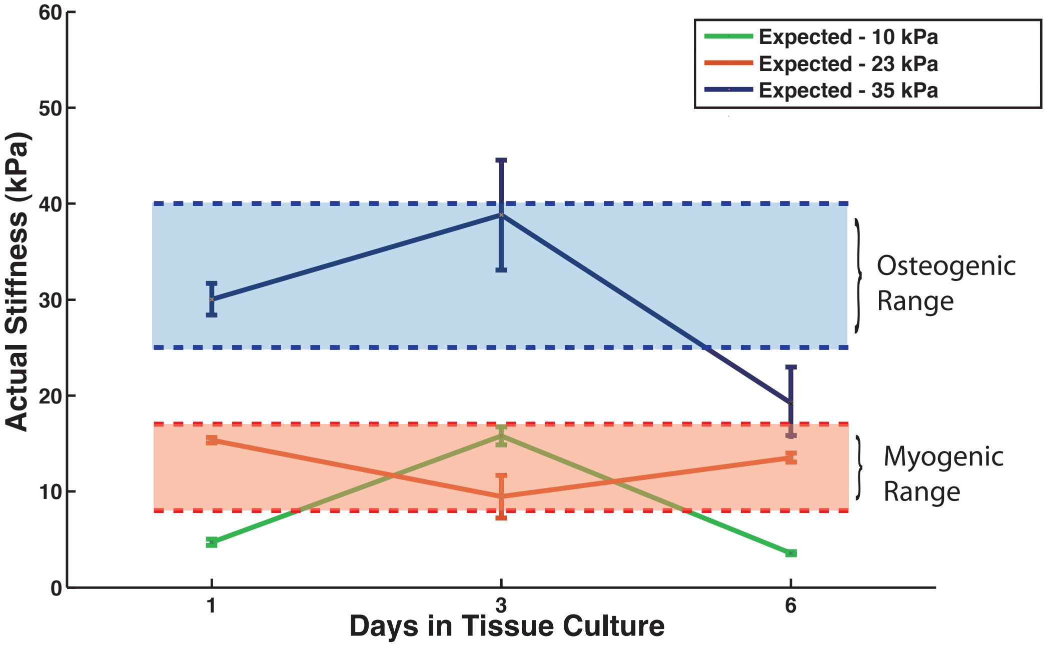


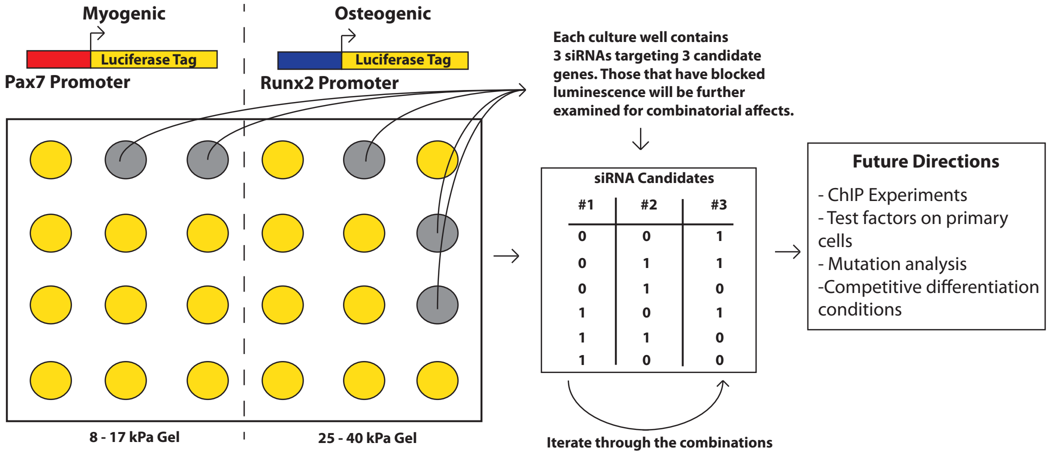
Figure 4.3: **Time-course AFM validation of polyacrylamide gels.** AFM data illustrating the consistency of the polyacrylamide gels over 6 days of tissue culture in 10% serum media conditions. Each point indicates a single gel with 3 indentation measurements.

constructs with Pax7 and Runx2 promoters[192, 193] that are connected to a human transmembrane protein's gene. This transmembrane protein would be designed to have a truncated cytoplasmic domain so as to minimize any effects on the cell.[194] A column coated with an antibody specific for the transmembrane protein could then be used to sort the cells. This would homogenize the cellular population and decrease the background for both the initial qRT-PCR and subsequent RNA-seq experiments.

Aim 2

Verify via a siRNA screen whether the RNA-seq candidates knock out the ability of the cell to tune the transcriptome to the physiologically relevant stiffness

The goal is to verify which transcriptionally related factors are required for the tuning of the gene expression to the physiologically relevant stiffness. I will continue to use Pax7 and Runx2 as the early progenitor markers;[195] however, instead of running a qRT-PCR analysis of each siRNA screen, I will create a luciferase tagged construct (Figure 4.4). I will use the Dual-Luciferase Reporter Assay System from Promega, inserting the Pax7 and Runx2 promoters in front of the luciferase gene. An additional construct containing *Renilla* luciferase, under the control of a ubiquitous promoter (e.g. SV40), will be co-transfected as an internal control[196] to normalize for transfection efficiency. The cells will be cultured on


Myogenic
Luciferase Tag
Pax7 Promoter
Osteogenic
Luciferase Tag
Runx2 Promoter
Each culture well contains 3 siRNAs targeting 3 candidate genes. Those that have blocked luminescence will be further examined for combinatorial affects.
8 - 17 kPa Gel
25 - 40 kPa Gel
siRNA Candidates
Iterate through the combinations
Future Directions
- ChIP Experiments
- Test factors on primary cells
- Mutation analysis
-Competitive differentiation conditions

#1	#2	#3
0	0	1
0	1	1
0	1	0
1	0	1
1	1	0
1	0	0

Figure 4.4: **Illustration of the luciferase assay.** The siRNA luciferase assay and illustrating the iterative method that will be used to identify the required candidates and how those lead into the future directions.

the myogenic and osteogenic tuned gels described in Aim 1 and I will assay which cultures do not bioluminesce on the same time course.

siRNA pools will be used verify the candidates that are capable of turning off the tuning phenomenon. I will assay multiple iterations with the siRNA candidate cultures that do not express the luciferase tag. By iterating the siRNA candidates it will be possible to increase the resolution of the screen to where there are only a couple of candidates to study further. Figure 4.4 illustrates this mixing in the event that the pool of siRNA included three different candidates. This can also be expanded for more siRNA where the maximum number of iterations would be $2^n - 2$, where n is the number siRNA candidates in the initial pool.

Potential Issues

A constant concern with siRNA experiments is whether there are any off-target effects. Once I am able to decrease the number to only a couple of the most important candidates, it would be important to attempt a rescue experiment. In the rescue experiment I would take cDNA that encodes for the candidate factors but has siRNA resistant silent mutations. If I am able to restore the tuning function, that would then further verify that the candidate is in fact required for the tuning and that there were no off-target effects.

Since the RNA-seq will only provide information about the transcriptome it will overlook any post-translational modifications that may be occurring. It is therefore possible that none of the siRNA for the candidate factors are successful in blocking the tuned expression. This would require an additional analytical iteration of the RNA-seq data to take into account the upstream basally expressed transcription factors and to then select new candidate factors.

Summary and Future Directions

The field of mechanotransduction is growing quickly and while the understanding of how the ECM's stiffness may affect the adhesion complexes in cells is becoming well characterized, the reason for the downstream transcriptome changes is still unknown. In my proposal I have outlined how I attempt to study the earliest mechanosensing events at the transcript level. By combining both experimental and computational techniques, I will use bioinformatics to inform my candidate selections that will be tested via siRNA screens. It is currently unclear if the mechanosensing event is an unique pulse event, a sustained signal or even a periodic expression of individual or multiple transcription factors. Much of this project will hinge on my ability to have a strong signal to noise ratio in my RNA-seq data and that is why the first part of Aim 1 is especially critical. In addition to having clean data, the bioinformatic techniques will be crucial in identifying candidates. To address this concern I will put the RNA-seq data through multiple filtering iterations. My analysis will provide progress towards understanding mechanism of a current hypothesis of how mechanotransduction works or propose an alternative mechanism of action.

Later experiments that will be done with the confirmed siRNA candidates will include tests in primary cells, ChIP experiments and mutational analysis of the promoter to try and identify the transcriptional mechanism of action.

CIRM Fellowship Proposal

Background and Significance

Articular cartilage-derived chondrocytes represent a particularly interesting system to study the transcriptional mechanisms of mechanosensing. Chondrocytes are suspended in an avascular extracellular matrix (ECM), therefore they are exposed to minimal soluble cues, and largely responsive to physical cues from the extracellular microenvironment and from mechanical loading, such as the cyclic strain that they receive in the joint during locomotion.[197] Chondrocytes maintain cartilage homeostasis by balancing their anabolic and catabolic functions. The anabolic state entails active proliferation and the production of large amounts of collagen 2 alpha 1 (Col2a1) and aggrecan.[198] The catabolic phenotype includes the production of connective tissue growth factor (CCN2)[199] and matrix metalloprotein proteases (MMPs), MMP13, that remodel the ECM1. Culturing chondrocytes in vitro on softer substrates mimic the anabolic state, whereas stiffer substrates mimic the catabolic state.[200] Acute injuries perturb tissue homeostasis, shifting the balance towards the catabolic phenotype, promoting hyperactive degradation of the ECM and disturbing a larger population of chondrocytes, ultimately leading to osteoarthritis. An ideal therapeutic would promote the anabolic function so as to repopulate the ECM with enough proteins to support the neighboring chondrocytes and prevent large-scale degradation. Much like in an acute injury where the ECM is torn, ECM degradation might affect chondrocyte function by forcing their transition from a rounded shape to a spread out morphology. Cell spreading can cause the cytoskeleton to reorganize and increase its tension via stress fibers, a process regulated by RhoA and ROCK.[201,202] Changes in cytoskeletal signaling are known to affect cell differentiation and cell maintenance.[203,204,205] The transcriptional mechanisms by which physical cues are translated into gene expression changes are only now being elucidated with the recent discovery of the first mechanosensitive transcription factors YAP/TAZ.[206] Interestingly, Sox9, the master transcriptional regulator of chondrocyte function,[207] can be phosphorylated by ROCK at serine 181, enhancing its nuclear localization and DNA binding affinity.[208] Thus Sox9 may represent a yet unidentified link to the mechanosensing signaling cascade. It is known that Sox9 targets cis-regulatory elements of both anabolic and catabolic genes,[209,210,211] but its mechanism of regulation and action are still largely unknown. The importance of Sox9 expression and its ability to regulate chondrocyte function is highlighted in growth disorders like campomelic dysplasia.[212] Studying the differential binding of Sox9, in the context of the microenvironment, will begin to provide a mechanistic insight into how Sox9 regulates the balance between anabolic and catabolic function. By further exploring what drives the redistribution of Sox9 among cis-regulatory elements, be it the phosphorylation of S181 or the binding to other transcriptional coregulators, it will drastically improve the understanding of its role in the maintenance of chondrocytes in articular cartilage, or its equally important role in growth plate chondroctyes during endochondral bone development. Understanding the mechanism of how physical cues regulate Sox9 binding and the switches between anabolic and catabolic function, would open the potential to manipulate the cell into expressing either program independent of the physical cues. A deeper understanding of the transcriptional nodes associated with mechanosensing will also simplify the development of biomaterial scaffolds for tissue specific applications by making it possible to force the cell into its native state even in an artificial microenvironment under a wide range of physical conditions.

Specific Aims

My overall goal is to study how the physical microenvironment mediates transcriptional responses. My specific aims are to evaluate the effect of physical cues on 1) the differential binding of Sox9, 2) determine the functional role of Sox9 phosphorylation in mechanosensing, 3) to transcriptionally recover the chondrocyte's anabolic state in a catabolic microenvironment and 4) to recapitulate these results in a more in vivo relevant 3D environment. These studies will be done with primary murine articular cartilage chondrocytes.

Proposed Research and Anticipated Results

Aim 1: The goal is to identify genomic regions that are differentially enriched for Sox9 between the anabolic and catabolic states, dependent on physical inputs from the microenvironment. The genomic distribution of Sox9 will be evaluated using ChIP-exo.[213] Whether the binding of Sox9 has a functional role will be evaluated by correlating the ChIP-exo results with RNA-seq data. Over the last two years I have been collaborating on bioinformatics projects examining transcriptional mechanisms with the Tjian lab by correlating genome-wide binding data to expression data. I will manipulate the microenvironment by culturing the cells on polyacrylamide gels that are conjugated with Col2a1. The softer polyacrylamide gels (0.5MPa) will stimulate the anabolic state, whereas tissue culture plastic (1GPa) will induce the catabolic state.

Aim 2: Since ROCK has an effect on both Sox9 phosphorylation and on chondrogenic function,[200] the goal is to determine if ROCK-phosphorylated Sox9 is directly stimulating the catabolic state versus the anabolic state. Combining a knockdown of the endogenous Sox9 with the overexpression of either Sox9-S181D or Sox9-S181A, phospho-mimetic and phospho-null respectively, should force either a catabolic or an anabolic state independent of the microenvironment. Additionally, Sox9 enriched regions from Aim 1 will be screened for the phosphorylated species of Sox9 by ChIP-qPCR. This will determine what the microenvironment dependent role is towards phosphorylated Sox9 binding and the activation of a transcriptional program.

Aim 3: The aim is to rescue the anabolic state by using a candidate drug approach. If the phosphorylation of Sox9 at serine 181 proves to be a functional switch between the anabolic and catabolic states, a screen of kinase inhibitors will be conducted to rescue the anabolic phenotype in a catabolic environment. Otherwise a DNase1 assay[214] can be used to identify transcriptional coregulators when the data is correlated with the Sox9 ChIP-exo data. Activators of these coregulators would then be tested to evaluate their ability to rescue the anabolic state independent of the microenvironment.

Aim 4: The goal is to determine if the 2D in vitro rescue results successfully translate into the more in vivo-relevant 3D model. The cells will be suspended in 3D hyaluronic acid hydrogels with varying mechanical properties.[207] If this work is successful, a similar hydrogel model can be used to test the efficacy of this as a therapeutic in an acute injury model in mice.

Preliminary Results: In collaboration with Dr. Alliston's lab at UCSF I have established the cell culture system to study the anabolic and catabolic function of chondrocytes (Figure 4.5). Culturing chondrocytes on substrates softer than plastic enhances expression of anabolic genes, while down-regulating catabolic ones, although further optimization is needed to reach the same levels of stiffness-mediated induction of the anabolic state seen in Allen et al. The antibody is capable of immunoprecipitating Sox9, at least when over-expressed in 293T cells (Figure 4.6). I am now evaluating whether the same antibody can pull down endogenous Sox9 in chondrocytes. Small-scale ChIP-qPCR experiments on known Sox9 target genes will further verify the efficiency of the antibody before proceeding to ChIP-exo in Aim 1. I have also successfully cloned the necessary Sox9 constructs for Aim 2.

Expected Progress: In the next year I expect to completely evaluate the function role of Sox9 binding in soft and stiff environments, to mimic the anabolic and catabolic states, respectively. A mechanistic understanding of how the microenvironment affects the transcriptional machinery of chondrocytes can provide important insights into osteoarthritis and Sox9-related developmental disorders, while improving our general understanding of how the

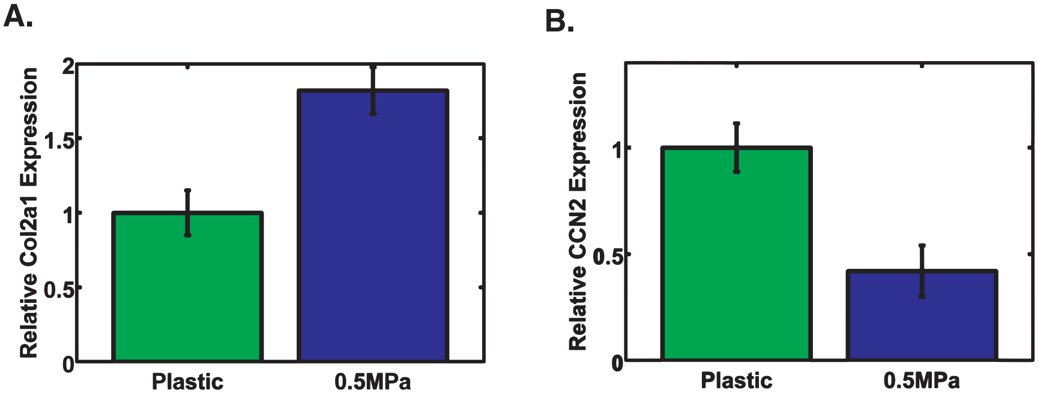


Figure 4.5: **Substrate stiffness mediates the switch between anabolic and catabolic function in chondrocytes.** In **A** and **B** expression levels were internally normalized to Rpl19, and then compared to the tissue culture plastic condition, which was set to one. **A.** Col2a1, a marker of anabolic function in chondrocytes, is enhanced on the 0.5MPa substrate after two days of culture. **B.** CCN2, a marker of catabolic function in chondrocytes, is down-regulated on the 0.5MPa substrate after two days of culture.

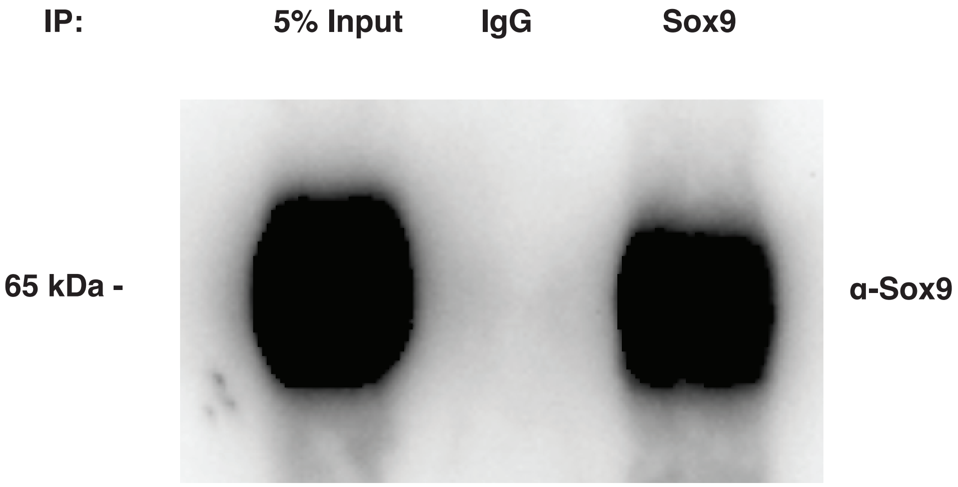


Figure 4.6: **Sox9 antibody efficiently immunoprecipitates.** Immunoprecipitation of Sox9 from total lysates of Sox9-overexpressing 293T cells.

microenvironment affects gene expression in cell differentiation and cell maintenance.

4.2 Unresolved Issues

Instead of attempting to resolve the which transcription factors are responsive to mechanical cues, more effort was placed in understanding how the redistribution of Sox9 under different mechanical conditions, influences the switch between catabolic and anabolic function in chondrocytes. Most polyacrlyamide based mechanotransduction studies are focused on utilizing small thin gels on glass cover slips which has its advantages in polymerization efficiencies. For the purpose of growing enough cells to do a ChIP-Seq experiment though (10-20 million cells), it was necessary to create 8cm gels. While attempting to establish the optimum experimental conditions though, there was difficulty in generating the reported signal over the plastic background. I believe that variability and decreased sensitivity were results of a non-uniform polymerization of the gels and inefficient conjugation of Col2a1 to the gel surface. A non-uniform polymerization would not allow the gel to have a uniform stiffness, causing sub-populations of cells to experience a different mechanical inputs. Given that the cells would only adhere to portions of the gel that were conjugated with Col2a1, any variability in the coverage could cause the cells to be more tightly clustered, forcing a change in cell shape and thereby changing their microenvironment.

Through discussions with the Weitz lab at Harvard, they believed that changing the cell shape through increased solute concentrations in the media would have a similar affect on the transcriptional program as seeding cells on gels with different substrate stiffnesses. They saw that shrinking cells on a soft substrate gel would induce a similar transcriptional program as what the cells would have on a plastic tissue culture dishes. For the purpose of studying the anabolic function of chondrocytes in a manner without needing to use polyacrylamide gels, it was necessary to swell the cells. Instead of increasing the solute concentration through PEG, water was added to the media at percentages ranging between 1 and 5% of the total volume. The media was then also supplemented with the necessary amino acids, nutrients and serum to bring those concentrations back to the initial media concentrations. Although there seemed to be a slight increase in anabolic function, Col2a1 versus MMP13 transcription, the signal was nowhere close to the expected 4 to 5 fold increase. One potential explanation for the decreased signal is that the focal adhesions had already matured on the tissue culture plastics. Focal adhesion strength has been strongly correlated with the affect of mechanical inputs and therefore it would be feasible that the changes in cell shape were not sufficient to reorganize the focal adhesions, cytoskeleton and downstream signaling cascades.

One of the plans for this project was to test the localization of Sox9 and the transcriptional activation of Col2a1 through a combination of a DNA array and MS2 repeats. The technical requirements for conducting the super resolution microscopy required that the cells be no more than 10μm off of the glass surface. To establish a reproducible substrate that would be capable of providing the appropriate mechanical inputs though would require further development and a move away from polyacrylamide gels. This was deemed to be impractical due to both deficiencies in time due to the bioinformatic projects and in the context of biomaterial development.

Once the technical issues of the surface structure and chemistry are solved, I believe that microscopy will be the best way to study how the microenvironment influences the transcriptional program. Microscopy has the advantage of being being able to evaluate both small populations of cells and single cells. Unlike with bioinformatic or biochemical methods, where we are limited to population averages that may cover a low signal to noise ratio, microscopy provides the opportunity to collect enough individual data points to develop meaningful statistics. It would also be appropriate to argue though, depending on the tissue, that the cell's microenvironment *in vivo* is not comprised entirely by single cells that lack any cell to cell contacts and therefore this is an inappropriate model system.

Chapter 5

Lens

5.1 Introduction

Working with digital documents has been difficult for the most part, because they come in presentation-centric formats, optimized for print and with the intent to have the same display across multiple devices. Ultimately, the display of these documents is preserved to allow the user to print the exact same document across any device. Content today, however, is no longer being printed out readily; instead, it is read on a variety of platforms spanning computers and mobile devices. The limitations presented by different screen sizes, the lack of tactile feedback that comes from flipping between pages and inability to purely focus on the author's arguments are problems present in all disciplines.

The web browser provides a unified platform for viewing content. Instead of binding the content to a presentation focused format, we can view the content as data. Thereby making the content readily accessible by utilizing a defined data structure. This data structure can then be processed in similar ways to databases – allowing users to query the content, create new content and build new tools.

We have decided to focus our initial efforts on applying the data-centric representation of content to the scientific literature. The Open Access scientific community has standardized much of their content into a formally annotated XML format, making it easier to gain access to a large library of articles. The content of scientific articles is a combination of text, figures, tables, videos and references which are used to form the author's argument. Scientific arguments are also rarely linear in nature, making it difficult to follow an argument when all of the content is not readily viewable on a digital device. The large amount of available data and non-linear format of the articles, provided us with the ideal start to test the flexibility of our data-centric approach. With the release of Lens, we would like to not only promote the data-centric representation of scientific content, but that of any content, which would be optimized for viewing on web clients.

5.2 The Lens Article

We are convinced that there is a need to break with traditional presentation-centric formats, by considering content as data and making it accessible in new ways. Therefore with the release of Lens, we'd like to promote a data-centric representation of scientific content, optimized for consumption by web-clients.

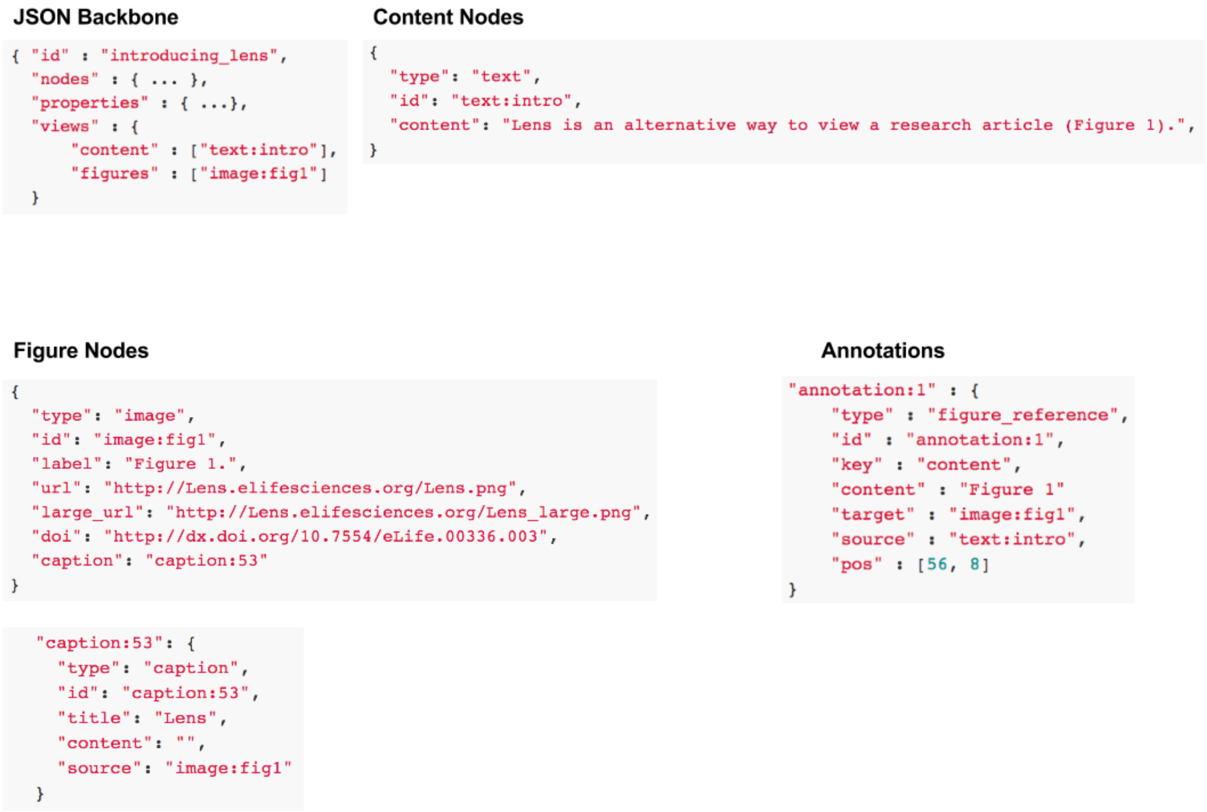
Lens can display any document that conforms to a simple JSON representation. JSON representations are flexible and can be adapted to any type of content. The initial JSON object provides the framework for organizing all of the content in an article. The *id* and *properties* keys provide global access if many JSON articles are to be queried for an initial round of filtering. The *nodes* key contains all of the content and the *view* key defines indices describing the order in which the nodes will be displayed. The nodes are then linked together via annotations. Annotations represent either font styling or an explicit reference to a target content node that is contained in its source content code. Figure 5.1 outlines a simple example of how to use the JSON representation to compose an article for Lens.

The creation of an article's JSON depends on a browser-based converter that runs through the XML tags, turning them into a smart data structure to power our viewer. The XML tags define various types of content nodes. Each node contains *type*, *id* and *content* keys that the viewer uses to define the rendering. All of the keys have their own definitions with regard to how Lens renders them. The translation from XML to a structured JSON creates a consistent presentation scheme for the content.

5.3 Presentation of Lens

When designing Lens, we worked to provide a flexible experience that supports a variety of use cases. On the far left, the document map Figure 5.2A identifies the reader's position in the context of the entire article. It also outlines each paragraph in the article. The left panel includes all of the textual content of the article (Figure 5.2B). The right panel includes the resources (Figure 5.2C). The resources include the table of contents, figures, tables, videos, supplemental data, references and information about the article. By separating the content into individual panels, we allow the reader to focus on multiple bits of content at the same time.

When reading scientific articles, it is rare for a reader to read straight through the text and figures in order. Instead, they may look to get a quick overview of the paper by looking at the figures first. By separating the figures from the text, the reader can transition between the two independently of one another. This allows the reader to simultaneously view exactly the content they want to focus on at that moment.



JSON Backbone

```
{ "id" : "introducing_lens",
  "nodes" : { ... },
  "properties" : { ...},
  "views" : {
      "content" : ["text:intro"],
      "figures" : ["image:fig1"]
  }
```

Content Nodes

```
{
  "type": "text",
  "id": "text:intro",
  "content": "Lens is an alternative way to view a research article (Figure 1).",
}
```

Figure Nodes

```
{
  "type": "image",
  "id": "image:fig1",
  "label": "Figure 1.",
  "url": "http://Lens.elifesciences.org/Lens.png",
  "large_url": "http://Lens.elifesciences.org/Lens_large.png",
  "doi": "http://dx.doi.org/10.7554/eLife.00336.003",
  "caption": "caption:53"
}
```

```
  "caption:53": {
    "type": "caption",
    "id": "caption:53",
    "title": "Lens",
    "content": "",
    "source": "image:fig1"
  }
```

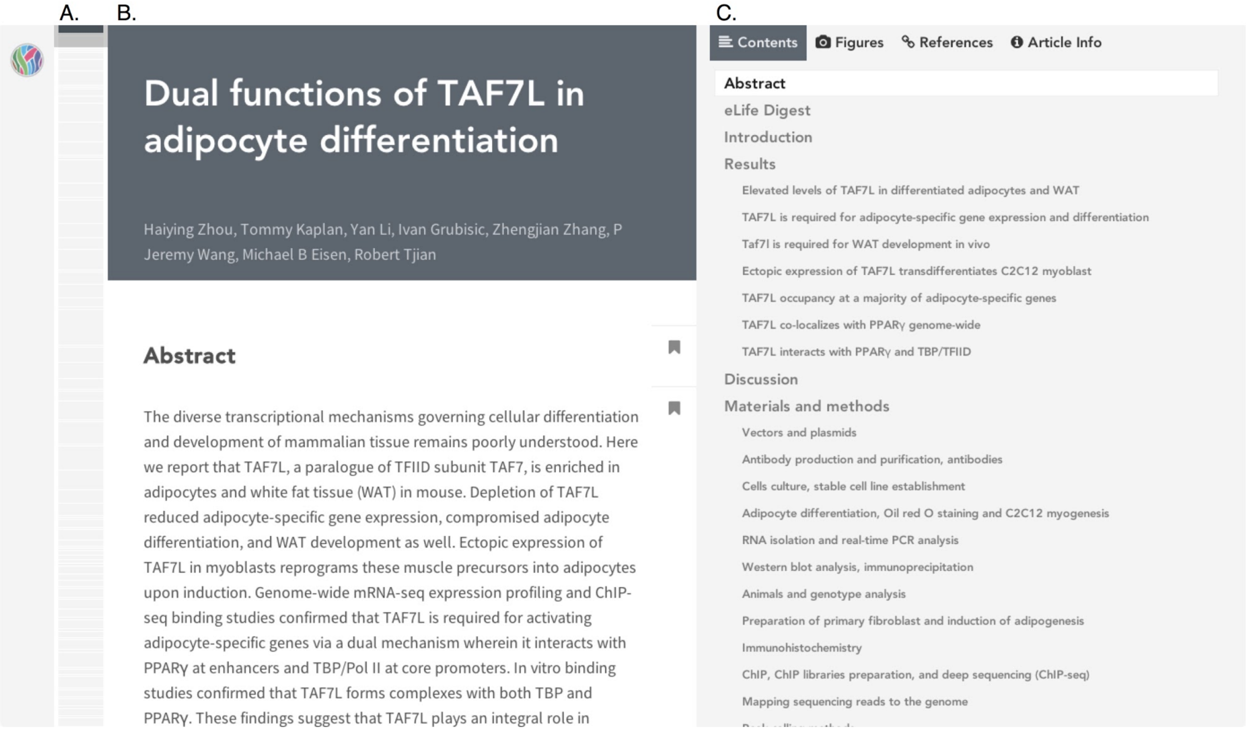
Annotations

```
"annotation:1" : {
    "type" : "figure_reference",
    "id" : "annotation:1",
    "key" : "content",
    "content" : "Figure 1"
    "target" : "image:fig1",
    "source" : "text:intro",
    "pos" : [56, 8]
}
```

Figure 5.1: **The Lens article JSON representation.** An example of how to begin creating content for Lens. The backbone organizes all of the data in a readily accessible manner. The text node defines the paragraph that may reference figures or publications. The figure node defines the image that will be displayed while the annotation links the figure to the associated text node that references the figure. See the Lens Article Format for a more detailed explanation of all the possible node types.

We initially supported two viewing modes: text-centric and resource-centric viewing. Text-centric viewing creates a microscale reading experience by limiting the viewable content to a singular content node. When a figure or publication label is selected within a text node, only the figures or publications that are referenced within that paragraph will be displayed in their appropriate resource sections (Figure 5.3, Figure 5.4). Selection of any figure or publication will color the paragraphs in the document map (Figure 5.2A) that include a reference to the selected resource.

The presentation of the article information has also been redesigned so as to organize the information in a uniform way. Each publisher will have placed information such as an author's impact statement, article keywords, major dataset links, etc. in different positions of their HTML or PDF versions of an article. With Lens, the aim is to distill all of the article information into organized subsections that can be displayed together on individual cards.



Dual functions of TAF7L in adipocyte differentiation

Haiying Zhou, Tommy Kaplan, Yan Li, Ivan Grubisic, Zhengjian Zhang, P Jeremy Wang, Michael B Eisen, Robert Tjian

Abstract

The diverse transcriptional mechanisms governing cellular differentiation and development of mammalian tissue remains poorly understood. Here we report that TAF7L, a paralogue of TFIID subunit TAF7, is enriched in adipocytes and white fat tissue (WAT) in mouse. Depletion of TAF7L reduced adipocyte-specific gene expression, compromised adipocyte differentiation, and WAT development as well. Ectopic expression of TAF7L in myoblasts reprograms these muscle precursors into adipocytes upon induction. Genome-wide mRNA-seq expression profiling and ChIP-seq binding studies confirmed that TAF7L is required for activating adipocyte-specific genes via a dual mechanism wherein it interacts with PPARγ at enhancers and TBP/Pol II at core promoters. In vitro binding studies confirmed that TAF7L forms complexes with both TBP and PPARγ. These findings suggest that TAF7L plays an integral role in

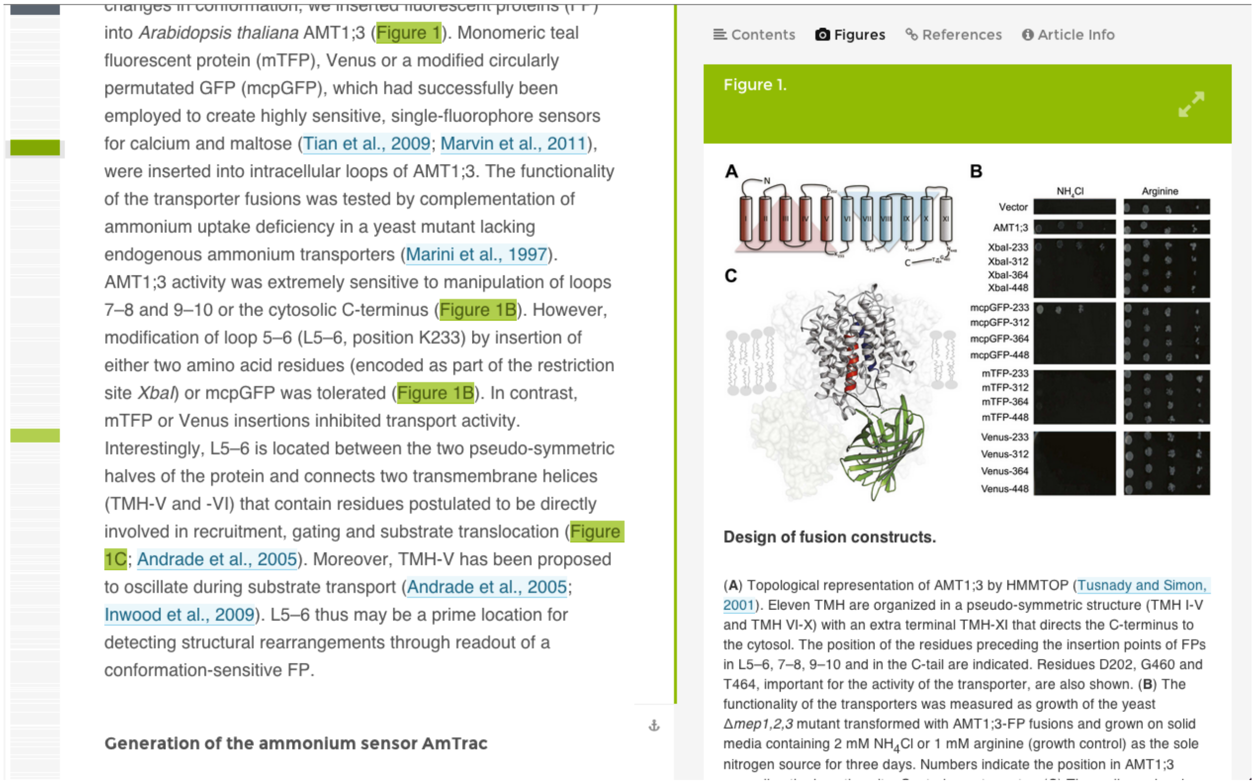
Abstract
eLife Digest
Introduction
Results
- Elevated levels of TAF7L in differentiated adipocytes and WAT
- TAF7L is required for adipocyte-specific gene expression and differentiation
- Taf7l is required for WAT development in vivo
- Ectopic expression of TAF7L transdifferentiates C2C12 myoblast
- TAF7L occupancy at a majority of adipocyte-specific genes
- TAF7L co-localizes with PPARγ genome-wide
- TAF7L interacts with PPARγ and TBP/TFIID

Discussion
Materials and methods
- Vectors and plasmids
- Antibody production and purification, antibodies
- Cells culture, stable cell line establishment
- Adipocyte differentiation, Oil red O staining and C2C12 myogenesis
- RNA isolation and real-time PCR analysis
- Western blot analysis, immunoprecipitation
- Animals and genotype analysis
- Preparation of primary fibroblast and induction of adipogenesis
- Immunohistochemistry
- ChIP, ChIP libraries preparation, and deep sequencing (ChIP-seq)
- Mapping sequencing reads to the genome

Figure 5.2: **The Lens visual layout. A.** The document map. Each gray bar identifies an individual paragraph within the article. **B.** The article's text content. **C.** All of the associated resources. This includes the Table of Contents, figures, references and additional information, including meta-data, about the article.

These information cards create a uniform viewing experience, making it easy to quickly find what the reader is looking for.

In addition to simplifying the reading experience of research articles, Lens also provides useful tools to share content with others. Each of the content nodes in the data structure are deep-linked. This means that each state is defined by a given URL. Sharing a specific URL will take the new reader to the exact same state in their browser which will facilitate discussions about the content.



changes in conformation, we inserted fluorescent proteins (FP) into *Arabidopsis thaliana* AMT1;3 (Figure 1). Monomeric teal fluorescent protein (mTFP), Venus or a modified circularly permutated GFP (mcpGFP), which had successfully been employed to create highly sensitive, single-fluorophore sensors for calcium and maltose (Tian et al., 2009; Marvin et al., 2011), were inserted into intracellular loops of AMT1;3. The functionality of the transporter fusions was tested by complementation of ammonium uptake deficiency in a yeast mutant lacking endogenous ammonium transporters (Marini et al., 1997). AMT1;3 activity was extremely sensitive to manipulation of loops 7–8 and 9–10 or the cytosolic C-terminus (Figure 1B). However, modification of loop 5–6 (L5–6, position K233) by insertion of either two amino acid residues (encoded as part of the restriction site *XbaI*) or mcpGFP was tolerated (Figure 1B). In contrast, mTFP or Venus insertions inhibited transport activity. Interestingly, L5–6 is located between the two pseudo-symmetric halves of the protein and connects two transmembrane helices (TMH-V and -VI) that contain residues postulated to be directly involved in recruitment, gating and substrate translocation (Figure 1C; Andrade et al., 2005). Moreover, TMH-V has been proposed to oscillate during substrate transport (Andrade et al., 2005; Inwood et al., 2009). L5–6 thus may be a prime location for detecting structural rearrangements through readout of a conformation-sensitive FP.

Generation of the ammonium sensor AmTrac

Design of fusion constructs.

(**A**) Topological representation of AMT1;3 by HMMTOP (Tusnady and Simon, 2001). Eleven TMH are organized in a pseudo-symmetric structure (TMH I-V and TMH VI-X) with an extra terminal TMH-XI that directs the C-terminus to the cytosol. The position of the residues preceding the insertion points of FPs in L5–6, 7–8, 9–10 and in the C-tail are indicated. Residues D202, G460 and T464, important for the activity of the transporter, are also shown. (**B**) The functionality of the transporters was measured as growth of the yeast Δ*mep1,2,3* mutant transformed with AMT1;3-FP fusions and grown on solid media containing 2 mM NH_4Cl or 1 mM arginine (growth control) as the sole nitrogen source for three days. Numbers indicate the position in AMT1;3

Figure 5.3: **Focusing on a figure.** Selecting a figure label within the text will bring that figure into view within the resources panel. The document map (Figure 5.2A) will identify all of the other references to that figure by highlighting the paragraphs green.

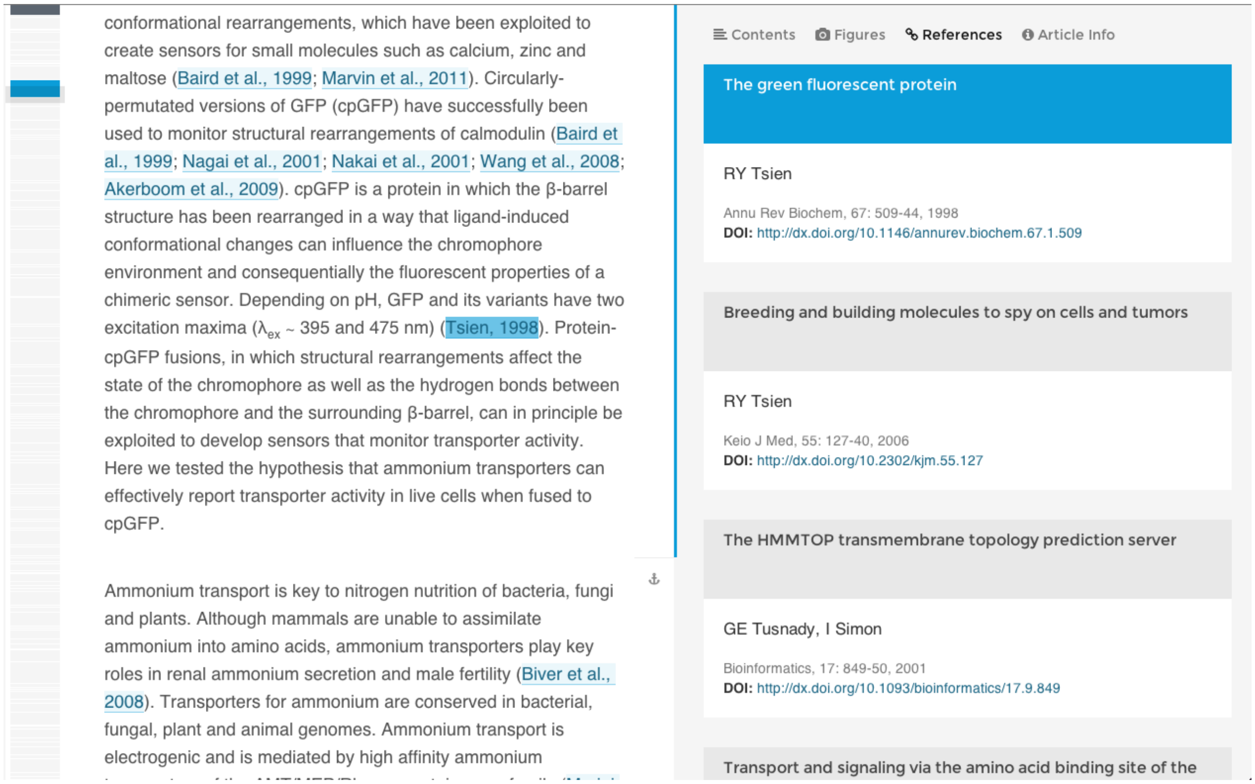
5.4 Open Source Development

5.5 The Future of Lens

oa-sandbox.org

Having looked at a variety of NLM XML files from different publishers, we noticed that even though they are all using the same tags, the manner in which they are implemented (e.g. position and logical organizations) are variable both across publishers and within a publisher's corpus. This ends up making it quite difficult to create a normalized set of articles for new analytical methods, search improvements and other tools. The Lens format can either be a clean intermediate in this process or even the basis of a future platform. It would also be trivial to create a standardized NLM XML file off of the Lens JSON representation for those that are already developing tools that run off of a XML input.

There should ultimately be very few research articles that are completely isolated (and if they are, that is probably indicative of a dead field, or a poor article), because each article



conformational rearrangements, which have been exploited to create sensors for small molecules such as calcium, zinc and maltose (Baird et al., 1999; Marvin et al., 2011). Circularly-permutated versions of GFP (cpGFP) have successfully been used to monitor structural rearrangements of calmodulin (Baird et al., 1999; Nagai et al., 2001; Nakai et al., 2001; Wang et al., 2008; Akerboom et al., 2009). cpGFP is a protein in which the β-barrel structure has been rearranged in a way that ligand-induced conformational changes can influence the chromophore environment and consequentially the fluorescent properties of a chimeric sensor. Depending on pH, GFP and its variants have two excitation maxima (λ_{ex} ~ 395 and 475 nm) (Tsien, 1998). Protein-cpGFP fusions, in which structural rearrangements affect the state of the chromophore as well as the hydrogen bonds between the chromophore and the surrounding β-barrel, can in principle be exploited to develop sensors that monitor transporter activity. Here we tested the hypothesis that ammonium transporters can effectively report transporter activity in live cells when fused to cpGFP.

Ammonium transport is key to nitrogen nutrition of bacteria, fungi and plants. Although mammals are unable to assimilate ammonium into amino acids, ammonium transporters play key roles in renal ammonium secretion and male fertility (Biver et al., 2008). Transporters for ammonium are conserved in bacterial, fungal, plant and animal genomes. Ammonium transport is electrogenic and is mediated by high affinity ammonium

The green fluorescent protein

RY Tsien

Annu Rev Biochem, 67: 509-44, 1998
DOI: http://dx.doi.org/10.1146/annurev.biochem.67.1.509

Breeding and building molecules to spy on cells and tumors

RY Tsien

Keio J Med, 55: 127-40, 2006
DOI: http://dx.doi.org/10.2302/kjm.55.127

The HMMTOP transmembrane topology prediction server

GE Tusnady, I Simon

Bioinformatics, 17: 849-50, 2001
DOI: http://dx.doi.org/10.1093/bioinformatics/17.9.849

Transport and signaling via the amino acid binding site of the

Figure 5.4: **Focusing on a citation.** Selecting a citation label within the text will bring that citation into view within the resources panel. The document map (Figure 5.2A) will identify all of the other references to that citation by highlighting the paragraphs blue.

fits into an ecosystem of feedbacks that help move research forward. These feedbacks include citations, article views, comments, news articles, blog posts and any number of other metrics that have yet to be determined. When reading an article though, this information is very loosely attributed to the article through ALMs. Altmetric, ImpactStory and PLOS, to name a few, are some of the leaders in this space. Ultimately though, all of this information is not centrally linked to the article. I see the Lens JSON as a data structure that can fulfill the function as the living version of the article. The PDF or XML versions of the articles are legacy items that will remain in the case things go wrong, but in the ideal world, as new information or metrics are available for an article, they would be presented to the reader immediately.

Authoring and Peer Review

The integration to Substance has now positioned Lens to be part of a larger publishing platform. We have enabled drag and drop functionality to Lens so that any NLM XML file can be automatically converted into the Lens format. The only caveat at the moment is that the images might not render if that publisher's definitions have not been added to Lens. We currently support all public access articles that are available on PubMed Central (PMC).

By encouraging authors and publishers to use and evolve open software components and a lightweight exchange format usable by web-clients, we reduce the obstacles that originate from economic competition and pave the way to true open publishing and an open exchange of information.

5.6 In the Press

The Back Story - eLife Blog - By Ivan Grubisic

My motivation for developing Lens came from my general frustration of not being able to see the figures and references while, at the same time, reading scientific articles. As you know, PDF files require the reader to manually scroll from page to page, while most HTML pages create anchor points that jump to that figure or reference, losing the original spot in the text. Creating a split view experience was the first step to managing this problem. I pitched this idea to eLife and then with support from Ian Mulvany, who helped build a great team including Michael Aufreiter, Graham Nott and Ian Hamilton, we began to quickly iterate to get to a functional minimum viable product.

By splitting apart the figures and references from the main text, it allows the reader to see all of the pertinent content at once. Scientists will often times approach research articles in different manners depending on their purpose for reading it. If they want to understand all of the details, they will focus on the text. This led us to implement an automatic scrolling behavior that would scroll the relevant figures automatically into view based off the paragraph that is being read. Alternatively, some readers want to initially focus on the figures to get the purpose of the article. The figure in view would then also trigger the same automatic scrolling behavior, bringing the first paragraph that references that figure into view.

Automatically triggering changes in the viewing experience, however, is not a trivial problem. We therefore reverted to manual triggering of the focused views. We attempted to do this at first by switching the view directly between the text and figures, all the while displaying only the paragraphs that are pertinent to the selected figure. After user testing, we noticed that people were confused about where they were located in the space of the article.

This then lead us to develop the document map. The map is integrated into the scroll bar and outlines each paragraph in the article. The simplicity of the map allows the reader to either scroll through the text or to click on a specific paragraph. The document map now highlights the selected figures and references so that it is easy to jump to all of the occurrences of them in the text (but never losing your place).

eLife Lens is the first step to simplifying and unifying the reading experience. The goal is to improve the ability to focus on the author's arguments by minimizing any distractions with minimalistic fonts, a simple color scheme and intuitive navigation flow. We are continuing to work on making the interface as clean as possible and continue to experiment with a variety of new features to improve the native function of Lens: to provide a novel way of viewing, and subsequently, interacting with scientific content.

The Fake Open Access Scam - PubChase Blog - By: Lenny Teytelmann

We have just enabled Lens-viewing of open access articles on PubChase, in collaboration with Ivan Grubisic[1]. Lens is an extraordinary step forward in visualization of research. Not only is it infinitely superior to PDFs, but it is even better than reading manuscript printouts. Figures are next to the text and you no longer need to hop around the articles between the text and references, constantly losing your place[2]. Alas, there is a wrinkle. We had hoped to Lensify all Pubmed Central free content, but turns out that we cannot because only a fraction of PMC content is truly open access; free to read does not mean open access.

The PMC content that we can legally display in the Lens format on PubChase is that which is under the Creative Commons Licenses. Most of these papers are from the PLOS, BiomedCentral, and Hindawi publishers. Unfortunately, almost 90% of PMC articles are free to read as PDFs, but are under restrictive publisher copyrights that make it illegal for PubChase to reformat them. Even author-submitted manuscripts in compliance with the NIH Public Access Policy are subject to the publisher copyright and we cannot display them.

This shocked me. While there has recently been much buzz about scams by new OA journals, especially with the Science Sting by Bohannon, the biggest scam is the one by subscription journals. Many erroneously assume that only open access journals charge a fee for publication, while subscription journals only charge for access. Far from it. My recent paper in PNAS cost $3,500 to publish with the following fees (excerpt from PNAS acceptance e-mail):

[1] Lens is an open access project that was initially sponsored by eLife. We are helping Ivan extend it beyond eLife to as much content as possible

[2] Please note this is still in Beta. Because of lack of uniformity in XML formats submitted to PMC, Ivan has to handle the Lensification of articles separately for each publisher, and some links may not work yet depending on the article

"Payment of the page charge of $75 per printed page will be assessed from all authors who have funds available for that purpose. Payments of $300 per article for up to five pages of Supporting Information (SI), $600 per article for six or more pages of SI, and $350 per color figure or table will be assessed. Authors of research articles may pay a surcharge of $1,350 to make their paper freely available through the PNAS Open Access option. If your institution has a current Site License, the open access surcharge is $1,000. Payment by authors of the following additional costs is expected: $150 for each replacement or deletion of a color figure or table, $25 for each replacement of a black-and-white or SI figure, and $25 for manuscript file replacement. Proofs should be returned within 48 hours."

This is way more than the cost of publishing in PLOS One, or even PLOS Biology, not to mention PeerJ, and after publication PNAS would still charge for access to the paper. But the part that upsets me most is that on top of these fees, PNAS charges $1,350[3] to publish an article as "open access", and it now turns out that it's not even open access and we cannot display it in Lens on PubChase.

While the scams of the shady OA journals are irritating, they are largely irrelevant. On the other hand, the scam by the subscription journals is outrageous and seriously damaging to science.

Mendeley Blog - By Ivan Grubisic

When I set out to develop Lens my main aim was to improve the way researchers interact with articles. The narrative within scholarly content is rarely linear, which is in direct conflict with our classical views of presenting articles in PDF or HTML. I accomplished this by decoupling the resources (e.g. figures and references) from the main text so that the reader can view multiple bits of content at any given moment instead of only focusing on one at a time.

Initially I was taking the article's HTML or NLM XML file and dumping the contents into a static HTML file that linked the text to the resources. With the help of Michael Aufreiter, we moved to a JSON representation of the article. Similar to the initial process of dumping content into a new HTML, Lens generates the HTML from the JSON. The main advantage now though is that there is a normalized data structure for each research article. Research articles live within an ecosystem of other articles and having an easily queriable database for each article would allow for researchers to create explicit links to figures or textual content for another article and have it rendered in their article. The initial premise of Lens was the desire to view the pertinent content with ease, and not have to fight against formatting or searching. The JSON data structure helps us do that in a much more efficient manner.

[3] The PNAS charge of $1350 is exactly what it costs to publish the entire article in PLOS One!!!

Having looked at a variety of NLM XML files from different publishers, I noticed that even though they are all using the same tags, the manner in which they are implemented (e.g. position and logical organizations) are variable both across publishers and within a publisher's corpus. This ends up making it quite difficult to create a normalized set of articles for new analytical methods, search improvements and other tools. The Lens format can either be a clean intermediate in this process or even the basis of a future platform. It would also be trivial to create a standardized NLM XML file off of the Lens JSON representation for those that are already developing tools that run off of a XML input.

Lens is still a beta product primarily because the meta information is difficult to parse out in a standard manner. Once this problem is solved, I could confidently say that Lens is ready for the mainstream for all Public Access articles; assuming of course that the full text XML is available. The code for Lens is distributed under a FreeBSD license so that anyone is free to use it. I would be happy to help out with the implementation as I did with PubChase recently.

5.7 Credits

Lens was developed in collaboration between myself and eLife Sciences ltd. I was supported by Michael Aufreiter from Substance, who helped with the design and implementation of the tool. Samo Korošec and Ian Hamilton (ripe) gave advice on the UX/UI design. Graham Nott implemented the deployment workflow and eLife provided support for the project and the initial corpus of documents against which the tool was developed. Rebecca Close, integrated Lens at Landes Bioscience and helped with the redesign of the converter, that now supports all NLM-compatible content.

The continued work on Lens has been supported by a flash grant from the Shuttleworth Foundation in the value of $5,000 dollars which came via a nomination from Dan Whaley at Hypothes.is. Jennifer Lin, a product manager at PLOS, has been central to organizing conversations with appropriate parties in the Open Access community to bring Lens to all of the Open Access articles via PubMed Central.

Chapter 6

Conclusions

Through the various research projects focused on evaluating transcriptional regulation in different systems, I have demonstrated the value that comes from evaluating the research question through multiple lenses. Bioinformatics has proven to be a very good tool for identifying a global trend for the small scale mechanistic findings and for elucidating new interactions by combining data sources that could then be explicitly tested. Additionally, my initial experience with the tools opened new questions that lead me to develop a set of tools that could be used in conjunction with existing methods to analyze any genome-wide sequencing assay. Although my foray into mechanotransduction was not very successful, it does highlight the need for either more creative solutions and a simpler approach to test transcriptional mechanisms in tissue culture conditions. The goal was to evaluate, what I initially felt was a relatively simple *in vitro* assay, with both a combination of ChIP-Seq and RNA-Seq to then inform the design of the DNA array and conclude the findings using microscopy. The hope was through combining both population and single cell based assays, that we could establish a complete picture for the chondrocyte's experimental system.

By attempting to apply multiple lenses to get the full context of an experimental system, we end up assessing the problem at different scope levels. These scopes can refer to the difference between population average assays, like high through-put sequencing, versus single cell imaging or single gene assays, time scales and length scales to name a few. The common thought process has been that if multiple techniques, at various scope levels, provide the same trends, then the confidence level of it not being an artifact increases. While imperfect, this tends to work well as long as there are no directly contradictory assumptions that go into establishing the experimental setup. From my perspective, the problem is that not enough emphasis is placed on evaluating the limitations of a method, understanding what types of claims can be drawn from the method and what are all of the prerequisite controls. This may seem trivial for a method that is a staple of the lab, but could prove to be tricky in a setting where there is little to no technical area expertise.

It is unrealistic, nor a good thing, to expect researchers to be an experts in every method. Currently the staple method to overcome lapses in technical area expertise is to seek out collaborations. These collaborations can be very productive and excellent learning experiences for all parties involved. The main bottleneck for them comes from either an inability to do efficient knowledge transfer and/or time restraints for one of the parties involved. Ultimately this can be distilled down to clear communication. For research to progress at a good pace, we need to communicate our ideas in a clear and direct manner so that even those without the area expertise can understand the purpose and fundamentals of a work.

Open Access publishing and many players in the scholarly communication space have been the primary movers in trying to improve the manner in which researchers interact with scientific content and one another. It is a shame that it has taken this long for us to get to a stage that anyone can read an article without a subscription. Readily accessible datasets are more difficult to come by, but even that area is improving. I hope that this space continues to evolve and that researchers begin publishing short, simple stories more often. With more well defined and well presented reports, researchers can begin to receive faster feedback from the community and findings can be validated faster. Both of these points will make things more efficient in terms of time and finances. There does seem to be a psychological block amongst unestablished graduate students, post-docs and even some faculty to fully embrace this model because of the fear that it will not be looked upon approvingly by funding agencies. There needs to be greater transparency regarding how researchers are evaluated for funding (e.g. fellowships, grants, etc.). Instead of hoping that the transparency occurs on its own, we need to find more robust ways of evaluating the impact of an individual article in the context of both the subject matter in which it resides and what it means for the scientific community. By providing funding agencies with more additional information, they will have no choice but to begin considering the new data that is being presented. This is not an easy battle, but one that I believe needs to happen.

Bibliography

1. K. Okita, T. Ichisaka, S. Yamanaka, Nature 448, 313 (2007).
2. L. J. Core, J. J. Waterfall, J. T. Lis, Science 322, 1845 (2008).
3. A. Valouev, D. S. Johnson, A. Sundquist, C. Medina, Nature (2008).
4. T. A. Down, et al., Nature Biotechnology 26, 779 (2008).
5. X. Darzacq, et al., Nature Structural Molecular Biology 14, 796 (2007).
6. A. J. Hughes, A. E. Herr, Proceedings of the National Academy of Sciences 109, 21450 (2012).
7. B. Langmead, C. Trapnell, M. Pop, S. L. Salzberg, Genome Biology 10, R25 (2009).
8. M. M. Harrison, X.-Y. Li, T. Kaplan, M. R. Botchan, M. B. Eisen, Plos Genetics 7, e1002266 (2011).
9. Y. Zhang, et al., Genome Biology 9, R137 (2008).
10. C. Y. McLean, et al., Nature Biotechnology 28, 495 (2010).
11. Y. W. Fong, et al., Cell 147, 120 (2011).
12. H. Zhou, et al., eLife 2, e00170 (2013).
13. H. Zhou, et al., Proceedings of the National Academy of Sciences 110, 16886 (2013).
14. A. J. Engler, S. Sen, H. L. Sweeney, D. E. Discher, Cell 126, 677 (2006).
15. M. Garber, M. G. Grabherr, M. Guttman, C. Trapnell, Nature Methods 8, 469 (2011).
16. A. Roberts, C. Trapnell, J. Donaghey, J. L. Rinn, L. Pachter, Genome Biology 12, R22 (2011).
17. A. Roberts, L. Pachter, Nature Methods (2013).
18. S. L. Salzberg, B. J. Wold, L. Pachter, Nature (2010).

19. K. Saha, et al., Biophysical journal (2008).

20. C. S. Goldsbury, S. Scheuring, L. Kreplak, Current Protocols in Protein Science pp. 1–19 (2009).

21. D. Mitrossilis, J. Fouchard, D. Pereira, PNAS pp. 1–6 (2010).

22. R. O. Hynes, Cell 110, 673 (2002).

23. R. Jaenisch, R. Young, Cell 132, 567 (2008).

24. L. A. Boyer, et al., Cell 122, 947 (2005).

25. X. Chen, et al., Cell 133, 1106 (2008).

26. J. Kim, J. Chu, X. Shen, J. Wang, S. H. Orkin, Cell 132, 1049 (2008).

27. A. Marson, et al., Cell 134, 521 (2008).

28. A. M. Näär, B. D. Lemon, R. Tjian, Annual review of biochemistry 70, 475 (2001).

29. M. H. Kagey, et al., Nature 467, 430 (2010).

30. R. G. Roeder, FEBS letters 579, 909 (2005).

31. D. J. Taatjes, M. T. Marr, R. Tjian, Nature Reviews Molecular Cell Biology 5, 403 (2004).

32. M. G. Rosenfeld, V. V. Lunyak, C. K. Glass, Genes & Development 20, 1405 (2006).

33. X. Gao, et al., Proceedings of the National Academy of Sciences 105, 6656 (2008).

34. B. L. Kidder, S. Palmer, J. G. Knott, Stem cells (Dayton, Ohio) 27, 317 (2009).

35. A. V. Tutter, et al., Journal of Biological Chemistry 284, 3709 (2009).

36. T. Kuroda, et al., Molecular and cellular biology 25, 2475 (2005).

37. D. J. Rodda, et al., Journal of Biological Chemistry 280, 24731 (2005).

38. J. A. Goodrich, R. Tjian, Nature Reviews Genetics 11, 549 (2010).

39. F. Müller, A. Zaucker, L. Tora, Current opinion in genetics & development 20, 533 (2010).

40. Z. Liu, D. R. Scannell, M. B. Eisen, R. Tjian, Cell 146, 720 (2011).

41. Y.-H. Loh, et al., Nature Genetics 38, 431 (2006).

42. R. Pal, G. Ravindran, Cell proliferation 39, 585 (2006).

43. C. M. Schwartz, et al., Stem cells and development 14, 517 (2005).

44. J. M. Sperger, et al., Proceedings of the National Academy of Sciences 100, 13350 (2003).

45. Y. Tokuzawa, et al., Molecular and cellular biology 23, 2699 (2003).

46. H. Chakravarthy, et al., Journal of cellular physiology 216, 651 (2008).

47. B. Lemon, C. Inouye, D. S. King, R. Tjian, Nature 414, 924 (2001).

48. S. Ryu, S. Zhou, A. G. Ladurner, R. Tjian, Nature 397, 446 (1999).

49. A. M. Näär, et al., Genes & Development 12, 3020 (1998).

50. A. M. Näär, D. J. Taatjes, W. Zhai, E. Nogales, R. Tjian, Genes & Development 16, 1339 (2002).

51. J. Zhang, et al., Nature Cell Biology 8, 1114 (2006).

52. J. Jiang, et al., Nature Cell Biology 10, 353 (2008).

53. D. L. van den Berg, et al., Molecular and cellular biology 28, 5986 (2008).

54. M. Araki, et al., Journal of Biological Chemistry 276, 18665 (2001).

55. N. Le May, et al., Molecular Cell 38, 54 (2010).

56. O. Maillard, S. Solyom, H. Naegeli, PLoS Biology 5, e79 (2007).

57. G. Yasuda, et al., Molecular and cellular biology 27, 6606 (2007).

58. A. Uchida, et al., DNA repair 1, 449 (2002).

59. B. M. Bernardes de Jesus, M. Bjørås, F. Coin, J.-M. Egly, Molecular and cellular biology 28, 7225 (2008).

60. C. A. Sommer, et al., Stem cells (Dayton, Ohio) 27, 543 (2009).

61. E. Engelen, et al., Nature Genetics 43, 607 (2011).

62. D. L. van den Berg, et al., Cell Stem Cell 6, 369 (2010).

63. J. Wang, et al., Nature 444, 364 (2006).

64. U. Camenisch, et al., The EMBO Journal 28, 2387 (2009).

65. J. M. Ng, et al., Genes & Development 17, 1630 (2003).

66. L. Schaeffer, et al., Science 260, 58 (1993).

67. D. Mu, D. S. Hsu, A. Sancar, Journal of Biological Chemistry 271, 8285 (1996).

68. J. Venema, A. van Hoffen, A. T. Natarajan, A. A. van Zeeland, L. H. Mullenders, Nucleic Acids Research 18, 443 (1990).

69. A. T. Sands, A. Abuin, A. Sanchez, C. J. Conti, A. Bradley, Nature 377, 162 (1995).

70. K. Sugasawa, et al., Molecular and cellular biology 16, 4852 (1996).

71. R. B. Cervantes, J. R. Stringer, C. Shao, J. A. Tischfield, P. J. Stambrook, Proceedings of the National Academy of Sciences 99, 3586 (2002).

72. M. Ramalho-Santos, S. Yoon, Y. Matsuzaki, R. C. Mulligan, D. A. Melton, Science 298, 597 (2002).

73. T. Lin, et al., Nature Cell Biology 7, 165 (2005).

74. E. D. Rosen, O. A. MacDougald, Nature Reviews Molecular Cell Biology 7, 885 (2006).

75. E. D. Rosen, B. M. Spiegelman, Nature 444, 847 (2006).

76. A. N. Hollenberg, et al., Journal of Biological Chemistry 272, 5283 (1997).

77. Q.-Q. Tang, T. C. Otto, M. D. Lane, Proceedings of the National Academy of Sciences of the United States of America 100, 850 (2003).

78. J.-W. Zhang, D. J. Klemm, C. Vinson, M. D. Lane, Journal of Biological Chemistry 279, 4471 (2004).

79. Q.-Q. Tang, et al., Proceedings of the National Academy of Sciences of the United States of America 102, 9766 (2005).

80. P. Tontonoz, B. M. Spiegelman, Annual review of biochemistry 77, 289 (2008).

81. E. Hu, P. Tontonoz, B. M. Spiegelman, Proceedings of the National Academy of Sciences 92, 9856 (1995).

82. R. P. Brun, J. B. Kim, E. Hu, S. Altiok, B. M. Spiegelman, Current Opinion in Cell Biology 8, 826 (1996).

83. M. Guermah, K. Ge, C.-M. Chiang, R. G. Roeder, Molecular Cell 12, 991 (2003).

84. T. Mori, et al., Journal of Biological Chemistry 280, 12867 (2005).

85. K. Matsumoto, et al., Journal of Biological Chemistry 282, 17053 (2007).

86. W. Wang, et al., Developmental Cell 16, 764 (2009).

87. L. Grøntved, M. S. Madsen, M. Boergesen, R. G. Roeder, S. Mandrup, Molecular and cellular biology 30, 2155 (2010).

88. H. Pei, Y. Yao, Y. Yang, K. Liao, J. R. Wu, Cell Death and Differentiation 18, 315 (2011).

89. R. K. Gupta, E. D. Rosen, B. M. Spiegelman, Cell metabolism 14, 739 (2011).

90. H. Waki, et al., Plos Genetics 7, e1002311 (2011).

91. J. Eguchi, et al., Cell metabolism 7, 86 (2008).

92. T. S. Mikkelsen, et al., Cell 143, 156 (2010).

93. S. F. Schmidt, et al., BMC genomics 12, 152 (2011).

94. M. Boergesen, et al., Molecular and cellular biology 32, 852 (2012).

95. R. N. Freiman, et al., Science 293, 2084 (2001).

96. E. Voronina, et al., Developmental Biology 303, 715 (2007).

97. M. D. E. Deato, R. Tjian, Genes & Development 21, 2137 (2007).

98. D. O. Hart, M. K. Santra, T. Raha, M. R. Green, Developmental Dynamics 238, 2540 (2009).

99. J. A. D'Alessio, K. J. Wright, R. Tjian, Molecular Cell 36, 924 (2009).

100. J.-C. Pointud, et al., Journal of Cell Science 116, 1847 (2003).

101. Y. Cheng, et al., Molecular and cellular biology 27, 2582 (2007).

102. O. Akinloye, J. Gromoll, C. Callies, E. Nieschlag, M. Simoni, Andrologia 39, 190 (2007).

103. N. Ding, et al., Molecular Cell 31, 347 (2008).

104. R. K. Gupta, et al., Nature 464, 619 (2010).

105. P. Tontonoz, E. Hu, R. A. Graves, A. I. Budavari, B. M. Spiegelman, Genes & Development 8, 1224 (1994).

106. P. Tontonoz, E. Hu, B. M. Spiegelman, Cell 79, 1147 (1994).

107. S. Kajimura, et al., Nature 460, 1154 (2009).

108. J. Feng, T. Liu, Y. Zhang, Current Protocols in Bioinformatics pp. 2–14 (2011).

109. K. Ge, et al., Molecular and cellular biology 28, 1081 (2008).

110. J. P. DeLany, et al., Molecular & cellular proteomics : MCP 4, 731 (2005).

111. B. Fève, Best Practice & Research Clinical Endocrinology & … (2005).

112. F. J. Gonzalez, Cell metabolism 1, 85 (2005).

113. J. C. Gerlach, et al., Tissue Engineering Part C: Methods 18, 54 (2011).

114. S. N. Gornostaeva, A. A. Rzhaninova, D. V. Gol'dstein, Bulletin of experimental biology and medicine 141, 493 (2006).

115. R. R. Bowers, M. D. Lane, Cell cycle (Georgetown, Tex.) 7, 1191 (2008).

116. R. G. Vernon, R. A. Clegg, D. J. Flint, Hormone and metabolic research 18, 308 (1986).

117. J. D. Armstrong, M. T. Coffey, K. L. Esbenshade, R. M. Campbell, E. P. Heimer, Journal of animal science 72, 1570 (1994).

118. S. Hashmi, et al., 3 Biotech 1, 59 (2011).

119. S. Kajimura, et al., Genes & Development 22, 1397 (2008).

120. P. Seale, et al., Nature 454, 961 (2008).

121. A. R. Chowdhury, A. K. Mukherjee, Indian journal of experimental biology 14, 701 (1976).

122. I. Y. Mahmoud, R. V. Cyrus, T. M. Bennett, M. J. Woller, D. M. Montag, General and comparative endocrinology 57, 454 (1985).

123. S. Kajimura, P. Seale, B. M. Spiegelman, Cell metabolism 11, 257 (2010).

124. M. Dym, D. W. Fawcett, Biol Reprod 4, 195 (1971).

125. D. W. Fawcett, Biol Reprod 2, Suppl 2:90 (1970).

126. M. P. Greenbaum, T. Iwamori, G. M. Buchold, M. M. Matzuk, Cold Spring Harbor perspectives in biology 3, a005850 (2011).

127. S. Kimmins, N. Kotaja, I. Davidson, P. Sassone-Corsi, Reproduction 128, 5 (2004).

128. P. Sassone-Corsi, Science 296, 2176 (2002).

129. M. Lamas, et al., Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences 351, 561 (1996).

130. F. Nantel, P. Sassone-Corsi, Front Biosci 1, d266 (1996).

131. P. J. Wang, D. C. Page, Human molecular genetics 11, 2341 (2002).

132. A. E. Falender, et al., Genes & Development 19, 794 (2005).

133. M. D. Rabenstein, S. Zhou, J. T. Lis, R. Tjian, Proceedings of the National Academy of Sciences 96, 4791 (1999).

134. D. Zhang, T.-L. Penttila, P. L. Morris, M. Teichmann, R. G. Roeder, Science 292, 1153 (2001).

135. F. Nantel, et al., Nature 380, 159 (1996).

136. D. Zhang, T.-L. Penttila, P. L. Morris, R. G. Roeder, Mechanisms of development 106, 203 (2001).

137. P. J. Wang, D. C. Page, J. R. McCarrey, Human molecular genetics 14, 2911 (2005).

138. A. Sediva, et al., J Clin Immunol 27, 640 (2007).

139. I. Martianov, et al., Development 129, 945 (2002).

140. J. Bao, J. Zhang, H. Zheng, C. Xu, W. Yan, Molecular and cellular endocrinology 327, 89 (2010).

141. S.-H. Hsu, H.-M. Hsieh-Li, H. Li, Experimental Cell Research 294, 185 (2004).

142. S.-W. Luoh, et al., Development 124, 2275 (1997).

143. A. M. Wikstrom, C. E. Hoei-Hansen, L. Dunkel, E. Rajpert-De Meyts, Journal of Clinical Endocrinology {& Metabolism 92, 714 (2007).

144. X. Zhou, A. Kudo, H. Kawakami, H. Hirano, The Anatomical Record 245, 509 (1996).

145. P. S. Chaudhry, S. Creagh, N. Yu, C. J. Brokaw, Cell motility and the cytoskeleton 32, 65 (1995).

146. J. Chen, Y. Wang, X. Xu, Z. Yu, Y. T. Gui, Zhonghua Nan Ke Xue (2009).

147. Y. Li, et al., Molecular human reproduction 17, 42 (2011).

148. K. Yang, et al., Molecular and cellular biology 32, 216 (2012).

149. Z. Hao, et al., Molecular human reproduction 10, 433 (2004).

150. B. Xu, et al., Soc Reprod Fertil Suppl 63, 87 (2007).

151. J. Y. Wu, et al., Nature Genetics 25, 448 (2000).

152. N. Nakamura, H. Shibata, D. A. O'Brien, C. Mori, E. M. Eddy, Molecular reproduction and development 75, 632 (2008).

153. A. Ike, S. Yamada, H. Tanaka, Y. Nishimune, M. Nozaki, Gene 298, 183 (2002).

154. K. Tokuhiro, et al., Plos Genetics 5, e1000712 (2009).

155. C. de Luca, et al., Journal of Clinical Investigation 115, 3484 (2005).

156. K. Mounzih, R. Lu, F. F. Chehab, Endocrinology 138, 1190 (1997).

157. S. Vijayaraghavan, et al., Biology of Reproduction 54, 709 (1996).

158. R. Catena, et al., FEBS letters 579, 3401 (2005).

159. T. Ohbayashi, et al., Nucleic Acids Research 31, 2127 (2003).

160. M. Teichmann, et al., Proceedings of the National Academy of Sciences 96, 13720 (1999).

161. D. E. Hickford, S. Frankenberg, A. J. Pask, G. Shaw, M. B. Renfree, Biology of Reproduction 85, 733 (2011).

162. M. M. Matzuk, D. J. Lamb, Nature medicine 14, 1197 (2008).

163. D. B. Menke, G. L. Mutter, D. C. Page, American journal of human genetics 60, 237 (1997).

164. Q.-Q. Tang, T. C. Otto, M. D. Lane, Proceedings of the National Academy of Sciences of the United States of America 101, 9607 (2004).

165. D. Schichnes, J. Nemson, L. Sohlberg, S. E. Ruzin, Microscopy and microanalysis 4, 491 (1998).

166. D. Chung, et al., PLoS Computational Biology 7, e1002111 (2011).

167. H. Li, R. Durbin, Bioinformatics 25, 1754 (2009).

168. D. Newkirk, J. Biesinger, A. Chon, K. Yokomori, X. Xie, Journal of computational biology : a journal of computational molecular cell biology 18, 1495 (2011).

169. M. Weber, et al., Nature Genetics 37, 853 (2005).

170. H. Li, et al., Nature 460, 1136 (2009).

171. P. V. Kharchenko, M. Y. Tolstorukov, P. J. Park, Nature Biotechnology (2008).

172. H. Thorvaldsdóttir, J. T. Robinson, J. P. Mesirov, Briefings in bioinformatics 14, 178 (2013).

173. R. M. Kuhn, D. Haussler, W. J. Kent, Briefings in bioinformatics 14, 144 (2013).

174. C. T. Freberg, J. A. Dahl, S. Timoskainen, P. Collas, Molecular biology of the cell 18, 1543 (2007).

175. C. K. Choi, M. T. Breckenridge, C. S. Chen, Trends in Cell Biology 20, 705 (2010).

176. K. A. Kilian, B. Bugarija, B. T. Lahn, M. Mrksich, Proceedings of the National Academy of Sciences 107, 4872 (2010).

177. R. McBeath, D. M. Pirone, C. M. Nelson, K. Bhadriraju, C. S. Chen, Dev Cell 6, 483 (2004).

178. D. Jaalouk, J. Lammerding, Nature Reviews Molecular Cell Biology (2009).

179. M. Wozniak, C. Chen, Nature Reviews Molecular Cell Biology (2009).

180. A. Mammoto, et al., Nature 457, 1103 (2009).

181. K. N. Dahl, A. J. S. Ribeiro, J. Lammerding, Circulation Research 102, 1307 (2008).

182. S. Khatau, D. Kim, C. Hale, R. J. Bloom, D. Wirtz, Nucleus (Austin, Tex.) 1, 337 (2010).

183. D. T. Butcher, T. Alliston, V. M. Weaver, Nature reviews. Cancer 9, 108 (2009).

184. F. Auradé, C. Pinset, P. Chafey, F. Gros, D. Montarras, Differentiation; research in biological diversity 55, 185 (1994).

185. C. M. Shea, C. M. Edgar, T. A. Einhorn, L. C. Gerstenfeld, Journal of Cellular Biochemistry 90, 1112 (2003).

186. M. Goulding, A. Lumsden, A. J. Paquette, Development 120, 957 (1994).

187. F. Otto, et al., Cell 89, 765 (1997).

188. J. R. Tse, A. J. Engler, Current protocols in cell biology / editorial board, Juan S. Bonifacino ... [et al.] Chapter 10, Unit 10.16 (2010).

189. L. Gao, R. Mcbeath, C. S. Chen, Stem cells (Dayton, Ohio) pp. 564–572 (2010).

190. C. Trapnell, L. Pachter, S. L. Salzberg, Bioinformatics 25, 1105 (2009).

191. C. Trapnell, et al., Nature Biotechnology 28, 516 (2010).

192. Y. V. Syagailo, et al., Gene 294, 259 (2002).

193. C. J. Lengner, et al., Mechanisms of development 114, 167 (2002).

194. M. M. Biotec pp. 1–6 (2005).

195. M. John, A. Geick, P. Hadwiger, H.-P. Vornlocher, O. Heidenreich, Current Protocols in Molecular Biology pp. 1–14 (2003).

196. Promega pp. 1–26 (2010).

197. K. Lahiji, A. Polotsky, D. S. Hungerford, In Vitro Cell Dev Biol Anim (2004).

198. T. Aigner, S. Söder, P. M. Gebhard, A. McAlinden, J. Haag, Nature Clinical Practice Rheumatology 3, 391 (2007).

199. S. OMOTO, Osteoarthritis and Cartilage 12, 771 (2004).

200. J. L. Allen, M. E. Cooke, T. Alliston, Molecular biology of the cell (2012).

201. C. C. DuFort, M. J. Paszek, V. M. Weaver, Nature Reviews Molecular Cell Biology 12, 308 (2011).

202. B. D. Hoffman, C. Grashoff, M. A. Schwartz, Nature 475, 316 (2011).

203. A. J. Engler, S. Sen, H. L. Sweeney, D. E. Discher, Cell 126, 677 (2006).

204. A. J. Keung, E. M. de Juan-Pardo, D. V. Schaffer, S. Kumar, Stem cells (Dayton, Ohio) 29, 1886 (2011).

205. K. Saha, et al., Biophysical journal 95, 4426 (2008).

206. S. Dupont, et al., Nature 474, 179 (2011).

207. C. Chung, J. A. Burdick, Tissue Engineering Part A 15, 243 (2009).

208. D. R. Haudenschild, J. Chen, N. Pang, M. K. Lotz, D. D. D'Lima, Arthritis & Rheumatism 62, 191 (2010).

209. H. Akiyama, Genes & Development 18, 1072 (2004).

210. V. Y. L. Leung, et al., Plos Genetics 7, e1002356 (2011).

211. G. Zhou, V. Lefebvre, Z. Zhang, H. Eberspaecher, B. de Crombrugghe, Journal of Biological Chemistry 273, 14989 (1998).

212. J. W. Foster, et al., Nature 372, 525 (1994).

213. H. S. Rhee, B. F. Pugh, Cell 147, 1408 (2011).

214. S. Neph, et al., Cell 150, 1274 (2012).