Dissertation

submitted to the
Combined Faculty of Natural Sciences and Mathematics
of the
Ruperto Carola University Heidelberg, Germany
for the degree of

Doctor of Natural Sciences

Presented by
Nadeshda Abazova
M.Sc. Molecular Biosciences
born in: Sofia, Bulgaria

Oral examination on July 5, 2018

Dynamic proteome remodeling during differentiation and reprogramming

Referees:
Dr. Kyung-Min Noh
Prof. Dr. Frank Lyko

ABSTRACT

Mammalian cell-fate transitions are fundamentally important processes shaping evolution and development. *In vitro* differentiation and reprogramming of pluripotent stem cells are valuable models to study these processes. Until recently, interrogating the protein regulatory networks involved in cell-fate transitions was hampered by technological limitations. Using the latest mass-spectrometry-based technologies available to date, combined with innovative biochemistry, this thesis presents three projects exploring the proteome rearrangements which occur during neuronal differentiation and reprogramming.

First, we characterized the global proteome dynamics during neuronal differentiation of embryonic stem cells (ESCs) and identified a co-expression protein cluster with high functional enrichment for neurogenic processes. A predictive bioinformatic analysis pointed out Sox2 as a top regulator of this protein group. Like most transcription factors, Sox2 forms complexes with other proteins which influence its target selection. We interrogated the chromatin-associated protein interactome of Sox2 at the beginning and the end of differentiation and found that it undergoes a remarkable stem cell- to neuronal transition. Integrative multi-omic analysis of our interactome data with transcriptomic and chromatin accessibility assays suggest that the joint genome association of Sox2 and selected interactors has a regulatory effect on hundreds of genes involved in embryonic development. Interestingly, this effect can be activating or repressing dependent on the differentiation stage.

The second study explores the potential of epigenetic memory and its proteomic manifestation in the context of induced pluripotency. We demonstrate that unlike primary neurons, neuronal cultures derived from induced pluripotent stem cells (iPSCs) retain the capacity to reprogram back to a pluripotent state. To interrogate the potential proteomic manifestation of epigenetic memory, we compared the initial iPSCs to the neuron-derived iPSCs and found that while their overall proteome composition is highly similar, the neuron-derived iPSCs retain distinct neuronal signatures in addition to the pluripotent ones. We further investigated the spatio-temporal progression of neuronal differentiation within embryoid bodies. As expected, we found that many more proteins change in the embryoid body rim, which is exposed to differentiation-inducing signals compared to the core, which is protected and retains pluripotent proteomic characteristics until the late differentiation stages. Surprisingly, however, we found that key epigenetic and developmental switches involved in pluripotency exit are initiated very early in the embryoid body core, suggesting very fast and efficient cross-communication between the cells layers in these spheroid structures.

Finally, we examined the protein regulatory network associated with the promoter of *c-Myc*, a gene critically involved with differentiation, development, pluripotency establishment and maintenance. Using a novel technique developed in our lab, we successfully isolated the *c-Myc* promoter on a single-locus level and were able to characterize the proteins associated with it. These include known *c-Myc* regulators, as well as many novel candidates. Our study provides a unique and valuable foundation for functional analysis of potential new *c-Myc* regulators.

In sum, we employed novel biochemical and mass-spectrometry based techniques in different cellular context and successfully expanded the protein regulatory networks driving differentiation and reprogramming.

ZUSAMMENFASSUNG

Entscheidungen über das Zellschicksal sind von fundamentaler Bedeutung für die Evolution und Entwicklung der Säugetiere. Die *in vitro* Differenzierung und Reprogrammierung pluripotenter Stammzellen sind wertvolle biologische Modelle zum Erforschen dieser Vorgänge. Technologische Mängel haben bis vor Kurzem die Forschung an die Protein-gesteuerten regulatorischen Mechanismen, welche in diesen zellulären Umwandlungen involviert sind, gehindert. Diese Doktorarbeit profitiert von den neusten Entwicklungen der Massenspektrometrie und Biochemie und wendet sie in drei Projekte an, welche es sich zum Ziel machen, die Umstrukturierung des Proteoms während der neuronalen Differenzierung und Reprogrammierung zu erforschen.

Wir haben das globale Proteom während der neuronalen Differenzierung embryonaler Stammzellen untersucht und einen co-exprimierten Protein-Cluster identifiziert, bei dem neuronale Prozesse stark angereichert waren. Voraussagende bioinformatische Analyse wies auf Sox2 als top Regulator dieser Proteingruppe hin. Wie die meisten Transkriptionsfaktoren, bildet Sox2 Komplexe mit anderen Proteinen, welche die Selektion seiner Zielgene beeinflussen. Wir untersuchten das interaktive Proteinnetzwerk von Sox2 am Anfang und am Emde der Differenzierung und stellten einen bemerkenswerten Übergang von Stamzell- zu neuronalen Proteinen fest. Eine integrative multi-omics Analyse unseres Interaktionsdatensatzes zusammen mit RNA-seq und ATAC-seq Daten deuteten an, dass das gemeinsame Binden von Sox2 und bestimmten Interaktionspartnern einen regulatorischen Effekt auf hunderten von Genen hat, welche in der embryonalen Entwicklung involviert sind. Interessanterweise kann dieser Effekt aktivierend oder hemmend sein, abhängig vom Differenzierungsstadium.

In der zweiten Studie wurde das Potenzial der epigenetischen Erinnerung und ihr Effekt auf das Proteom im Kontext der induzierten Pluripotenz untersucht. Wir zeigen dass im Unterschied zu primären Neuronen, iPSC-induzierten neuronalen Kulturen die Kapazität behalten, zurück zu pluripotenten Zellen umgewandelt zu werden. Um das Potenzial der epigenetischen Erinnerung zu untersuchen haben wir die anfänglichen mit den aus Neuronen reprogrammierten iPSCs verglichen und stellten fast, dass ihre Proteome insgesamt zwar sehr ähnlich sind, jedoch die aus Neuronen reprogrammierten iPSCs klare neuronale Merkmale beibehalten, zusätzlich zu ihren pluripotenten Zeichen. Weiterhin wurde der raumzeitliche Ausdruck der neuronalen Differenzierung innerhalb von Embryoid bodies.untersucht. Als erwartet stellten wir fest, dass sich die Exprimierung von viel mehr Proteinen am äußeren Rand verändern, als im Inneren Teil, welcher von der Zellkultur-Umgebung geschützt ist und bis in den späten Differenzierungsphasen pluripotente Eigenschaften beibehält. Überraschenderweise stellten wir aber fest, dass wesentliche Entwicklungs- und Epigenetikschalter sehr früh und im Kern der Embryoid bodies eingeschaltet werden, was auf äußerst effiziente Kommunikation zwischen den unterschiedlichen Zellschichten hindeutet.

Schließlich untersuchten wir das regulatorische Netwerk, assoziiert mit dem Promoter von *c-Myc* – ein Gen kritisch involviert in Differenzierung, Entwicklung, Pluripotenzentstehen und -aufrechterhaltung. Wir wandten eine innovative Technik an um den einzelnen Genort (der Promoter von *c-Myc*) zu isolieren und die Proteine zu bestimmen, welche daran gebunden waren. Darunter waren viele bekannte Regulatoren von *c-Myc*, aber auch viele neue Kandidaten. Unsere Studie bietet eine einzigartige und wertvolle Basis für funktionelle Analyse potenzieller neuer *c-Myc* Regulatoren.

Zusammengefasst setzten wir innovative biochemische und auf Massenspektrometire-basierten Techniken an, in unterschiedlichen zellulären Kontexten, und erweiterten das proteomische regulatorische Netwerk, welches die Differenzierung und Reprogrammierung vorantreibt.

Contents

1. Introduction

1.1 Cellular differentiation in evolution and development

When the primary forms of life emerged on Earth around 3.7 billion years ago, they resembled little more than an isolating compartment, made of phospholipids, and a self-replicating molecule inside, RNA (Cooper 2000). From this basic stage onwards, living matter took on a long journey guided by the core principles of evolution and development, thereby reshaping the surface and atmosphere of the planet and creating a biosphere of vast complexity and variety.

This evolutionary road is marked by several key milestones. It took about 1-1.5 billion years for living cells to develop a nucleus (as well as other cellular compartments) and another billion years for multicellular structures, initially cellular aggregates, to form (Cooper 2000). At last, the mechanism of cellular specialization - or differentiation - emerged. This was the fuel that finally drove the development of manifold organisms of unprecedented complexity. By taking on highly specialized functions, cells could form different tissues which together assemble into entities with a wide spectrum of biological capacities. The number of cell types within an organism is often used as a proxy for its complexity and in the present day biosphere, mammals sit on top of this evolutionary hierarchy (Arendt 2008; Hall & Olson 2006).

Several key aspects of the evolutionary milestones ultimately leading to the formation of higher-order organisms like mammals are undergone during their embryonic development*. It is initiated with the formation of a single diploid cell (zygote), which then forms an aggregate of identical cells (morula), which proceeds to undergo series of differentiation steps to form a complete organism

* not to be confused with the largely discredited XIX century Recapitulation theory, which postulates that during development, animals go through phases representing the evolutionary stages of their remote ancestors.

with hundreds of highly specialized cell types. Clearly, the process of cellular differentiation runs like a thread through the fundamental forces shaping biology and studying it is therefore paramount to understanding how life works and evolves.

1.2 Cell-fate transitions in nature and research

Mammalian embryonic development, much like evolution, is a unidirectional process (Hochedlinger & Plath 2009). It involves transition through a number of cell-fate decisions, during which the cellular developmental and differentiation potential progressively decreases (Hochedlinger & Plath 2009). The highest potential bare the so called totipotent cells (from Latin: *totus* – whole and *potens* – power), which are the zygote and the blastomeres of the early morula (Kelly 1977). They can give rise to an entire organism. The second-highest potential have the so called pluripotent cells (from Latin: *pluri*- very, many). Those are the embryonic stem cells (ESCs) derived from the inner cell mass (ICM) of a blastocyst. They have the potential to differentiate into any cell type, but cannot alone give rise to an organism. Down the process of cellular specialization, the differentiation potential of the stem cells gets more restricted to multipotency (multipotent stem cells can give rise only to cell types within their lineage; from Latin: *multi*- many) and unipotency (from Latin: *unus* – one) (Mitalipov & Wolf 2009).

The fact that the genetic information is preserved during development and remains identical in all adult cell types explains the necessity of "outer-genetic" mechanisms which account for the vast differences in gene expression patterns and subsequent morphology and functionality of the different cell types. The term "epigenetic" (from Greek ἐπι- outer, above) was first coined in the mid XX century by the British developmental biologist Conrad Hal Waddington, who established the very foundations of evolutionary developmental biology. Originally, the term "epigenetic" used to describe the poorly understood regulatory processes related to the formation of an adult organism from a fertilized zygote (Waddington 1953). This definition has evolved along with the significant increase in our

understanding of the mechanisms behind gene expression regulation and is currently used to describe the heritable modifications which impact gene expression and are not based on changes in the DNA sequence (Riggs & Porter 1996). However, the history of epigenetics has always been intimately linked to the study of evolution and development; as the famous molecular biologist and epigenetics expert Gary Felsenfeld described it, "Although the definition that we choose for epigenetics has changed to accommodate our increasing knowledge, it is important to remember that the original problem was: How can a single fertilized egg give rise to a complex organism with cells of varied phenotypes?" (Felsenfeld 2014).

C.H. Waddington famously depicted the relationship between gene regulation and development in his "epigenetic landscape model" (Fig. 1.1) (Waddington 1957). In it, developmental restrictions are depicted as marbles (cells) rolling down a landslide into one of several valleys (cell fates). At the earliest developmental stages, cells start with the highest developmental potential and end "down on the bottom" with their fully differentiated state, which bares no further potential.

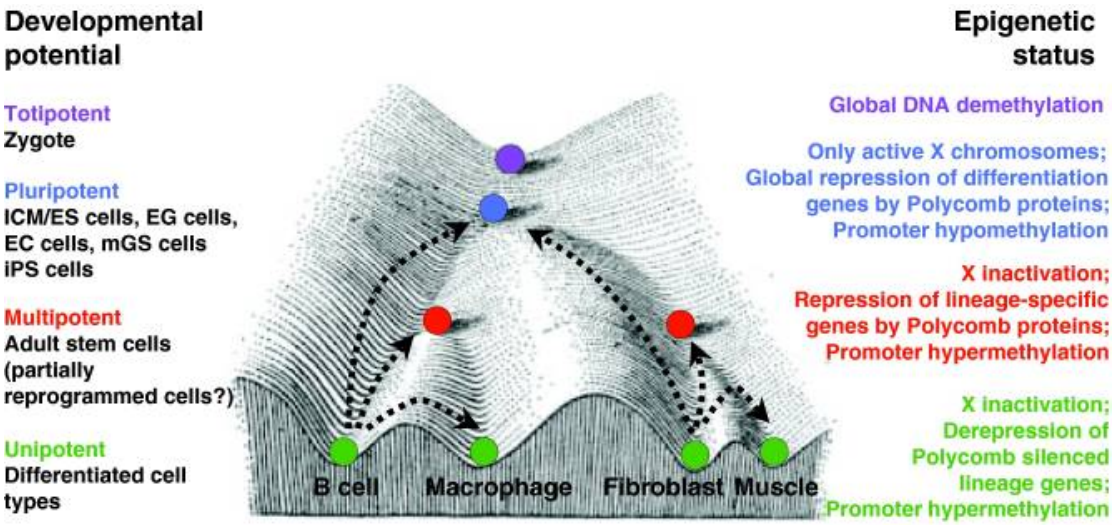


Figure 1.1 Epigenetic landscape at different stages of development. Adaptation of C. H. Waddington's epigenetic landscape model (Waddington 1957) by K. Hochedlinger and K. Plath (Hochedlinger & Plath 2009) showing cell populations with different developmental potentials (left) and their respective epigenetic states (right). Developmental restrictions are depicted as marbles (cells) rolling down a landslide into one of several valleys (cell fates). Colored marbles correspond to different differentiation states. Examples of reprogramming are shown by dashed arrows.

The past 70 years of bioresearch have provided the insight that the progressive loss of cellular developmental potential is neither unidirectional, as previously thought, nor irreversible. The developmental program of the cells can be turned backwards towards pluripotency or totipotency under certain experimental conditions - a process known as reprogramming (Stadtfeld & Hochedlinger 2010). There are several methods to achieve this, such as somatic-cell nuclear transfer (Briggs & King 1952; Wilmut et al. 1997; Hochedlinger & Jaenisch 2002; Liu et al. 2018), cell fusion (Tada et al. 1997) and overexpression of defined transcription factors (Takahashi & Yamanaka 2006). All these methods ensure that the epigenetic landscape of the somatic cell nucleus is reset back to a state which largely resembles a pluripotent or totipotent state (Fig. 1.1) (Apostolou & Hochedlinger 2013).

This remarkable cellular plasticity has vast implications, both for fundamental science and for medical research. In the context of the former, it provides an invaluable insight into the developmental processes that lead to the formation of complex organisms. By studying the molecular changes which occur during cellular differentiation and reprogramming, we gain understanding in how biological complexity emerges.

In the context of medical research, stem cells bare a tremendous therapeutic and diagnostic potential (Watt & Driskell 2010). Different stem cell therapies are already a common medical practice – most notably hematopoietic stem cell transplantations (Watt & Driskell 2010). Owing to their capacity to differentiate into any cell type, pluripotent stem cells hold a great potential in replacement therapies and offer the advantage that they can be differentiated into various lineages under strictly controlled conditions *in vitro*. Furthermore, Takahashi and Yamanaka's establishment of the so called induced pluripotent stem cells (iPSCs) (Takahashi & Yamanaka 2006), which ultimately led to Yamanaka's Nobel Prize award only 6 years later, offers the significant advantages that (1) the cells can be generated from patient-own material, thereby omitting the need for immunosuppression and (2) omit the necessity of destroying human embryos, which poses a significant ethical issue for a large number of people (Inoue et al.

2014; Wu & Hochedlinger 2011). Another important medical implication is owed to the fact that pluripotency emergence and cancer formation are processes which share a multitude of molecular properties (Meissner et al. 2008), explaining the stem cells' relevance to cancer research. Last but not least, pluripotent stem cells can be used as personalized diagnostic tools, due to their capacity to indefinitely renew themselves (Egashira et al. 2013).

Taken together, these factors explain the immense scientific interest in stem cell biology in general and the processes of differentiation and reprogramming in particular. While significant progress has been made in uncovering the molecular mechanisms behind these processes, there are still a lot of unknowns, particularly in the cases where the necessary technologies have either not existed or not been sufficiently effective in unraveling the remaining open questions. Addressing them is paramount for overcoming the hurdles hampering the way of this technology to the clinic and for advancing our understanding in how cell-fate transitions and complexity emergence work.

1.3 Proteomics of differentiation and reprogramming

In the past decades, major technological and bioinformatic advances have turned large-scale studies into indispensible components of bioresearch. The reduced cost and increased efficiency of sequencing, for example, has given the study of genomes and transcriptomes a new position of importance which spans through virtually every aspect of biology. Methods which combine biochemical assays and sequencing, such as chromatin immunoprecipitation combined with sequencing (ChIP-seq), have also greatly increased our understanding of the regulatory mechanisms behind a multitude of cellular functions and transformations. Stem cell research makes no exception when it comes to benefitting from these advances. Our current understanding of reprogramming, for example, is to a large extent based on 'global' studies of the dynamic changes in the transcriptome, genome methylation patterns and chromatin conformation, occurring during reprogramming of somatic cells to iPSCs (Polo et al. 2012; Meissner et al. 2008).

While the importance of transcriptome studies is undeniable, ultimately the key functional entities in the cells are proteins. Based on the "central dogma of molecular biology", it had been long assumed that there is a high correlation between transcriptomic and proteomic changes (Haider & Pal 2013). However, we now know that different factors, such as post-transcriptional regulations or varying protein half-lives, can lead to a discrepancy between the transcriptome and proteome (Haider & Pal 2013), underlying the importance of integrated studies.

Compared to genomics and transcriptomics, proteomics was slower to enter the big arena of big data. This was due to specific technological challenges (put bluntly and in oversimplified fashion, nucleic acids can be amplified, proteins can not, resulting in much lower coverage). That being said, the tremendous recent advances in mass spectrometry-based technologies, as well as the invention of novel biochemical methods and tailor-made bioinformatic solutions have moved proteomics to the forefront of stem cell research (Abazova & Krijgsveld 2017). The increasing impact in the field now spans both global proteome analysis and sub-proteomes such as membrane proteins (enabling the novel discovery of cell surface markers), phosphoproteomics (essential for signal transduction studies), and protein interaction networks (allowing the elucidation of protein complexes and regulatory networks) (Fig. 1.2) (Abazova & Krijgsveld 2017).

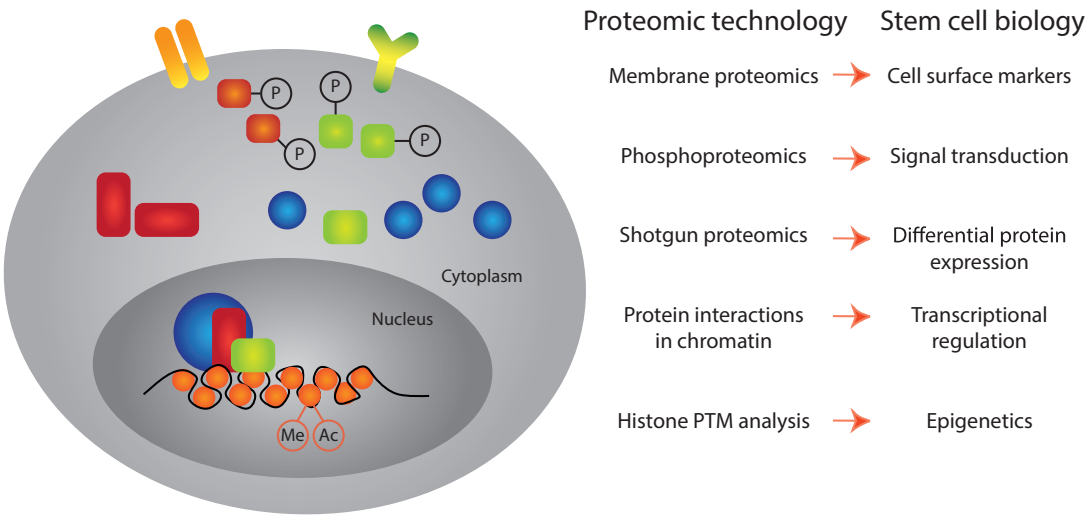


Figure 1.2 Usage of different proteomic technologies to study stem cell biology. Adapted from (Abazova & Krijgsveld 2017).

There have been a number of mass spectrometry-based applications focusing on the molecular transformations during cell fate transitions. Prominent examples include the proteome analysis of some of neuronal differentiation paths (Shoemaker & Kornblum 2016), as well as lineage-specification studies such as the transition to the extraembryonic endoderm (Mulvey et al. 2015). Induced pluripotency has also been explored from a proteomic perspective. Our lab performed the first deep, dynamic proteome profiling of fibroblasts reprogramming to iPSCs and showed that most proteome expression changes occur in the early and late phases of this transition (Hansson et al. 2012). These findings were later mirrored in the groups of Nagy and Heck in a similar system (Benevento et al. 2014). Several studies have focused on the earliest pluripotent cell-fate transitions by comparing the proteome composition of the early ("naïve") and late ("primed") states of pluripotency, which resemble the pre- and post-implantation epiblast. They identified differences in glycolysis (Taleahmad et al. 2015) and chromatin regulation (Song et al. 2012; Taleahmad et al. 2015), which supports the notion that pluripotency is controlled epigenetically (Jia et al. 2012).

Rather than solely focusing on expression changes, proteomics has also been employed to study (regulatory) interaction networks in stem cells. The core pluripotency factors Oct4, Sox2 and Nanog (OSN) have been the subject of particular attention and plentitude of interaction studies (Do et al. 2014; Pardo et al. 2010; van den Berg et al. 2010; Mallanna et al. 2010; Lai et al. 2012; Costa et al. 2013; Gagliardi et al. 2013). Based on them, interaction partners with key regulatory functions have been identified, such as TET1 and TET2, which rely on Nanog for their role in pluripotency establishment (Costa et al. 2013). Our lab has further expanded the circuitry of pluripotency by examining the chromatin-bound OSN interaction network and how it changes between the naïve and ground states of pluripotency (Rafiee et al. 2016).

In sum, proteomics is making an increasing impact in stem cell research and continues to shape our understanding of how the molecular architecture transforms during cell-fate transitions.

1.4 Objectives

This thesis aims to harness the power of some of the major technological advances in mass spectrometry-based proteomics and apply them to expand the current knowledge and understanding of differentiation and reprogramming.

In Chapter 2, we present a large-scale analysis of the dynamic proteome changes taking place during neuronal differentiation of pluripotent stem cells. We identified Sox2 as a top regulator of a cluster of proteins with similar temporal expression profile and high enrichment for neurogenesis-related processes and characterized the chromatin-associated Sox2 interactome at the beginning and end of neuronal differentiation. Finally, we used multi-omics integrative analysis with transcriptomic and chromatin accessibility datasets to explore the transcriptional regulatory effects of the associated binding of Sox2 and its newly identified interaction partners during differentiation.

Chapter 3 explores the potential and proteomic manifestation of the so called "epigenetic memory" in the context of neuronal differentiation and reprogramming. We demonstrate that unlike primary neurons, neuronal cultures generated from pluripotent stem cells retain sufficient epigenetic traits which allow them to be transformed back to a pluripotent state using only overexpression of the four "Yamanaka factors" (Oct4, Sox2, Klf4 and c-Myc). We characterized the proteomes of the initial and final pluripotent stem cells and found that while they are overall highly similar in both morphology and expression patterns, the final culture has a dual pluripotent/neuronal proteomic signature, possibly owing to epigenetic memory retention. Finally, we interrogated the spatio-temporal resolution of the proteomic shifts driving the transition from pluripotent stem cells to neural progenitors within embryoid bodies and found that, surprisingly, important epigenetic and expression switches driving neuronal differentiation become activated very early inside the embryoid body core, which expresses high levels of many pluripotency markers like Oct4 and Nanog and is largely isolated from the signal-inducing cell culture environment.

In Chapter 4, we expand the transcriptional regulatory network in pluripotent stem cells by presenting the first unbiased characterization of the proteome associated with the *c-Myc* promoter in ESCs. Using a novel technique developed in our lab, we successfully isolated the *c-Myc* promoter region at single-locus level, along with the proteins bound to it. We employed mass spectrometry to analyze them and identified over 250 proteins associated with the promoter *c-Myc*, the majority of which are novel, thus warranting further exploration of their potential regulatory function in future.

Each chapter is introduced and discussed individually.

2. Proteome characterization and dynamic Sox2 interaction network during neuronal differentiation of ESCs

2.1 Introduction

From all differentiation lineages which pluripotent cells can assume, the neuronal is among the most intriguing ones, for basic and clinical research alike. The conversion of pluripotent cells to terminally differentiated neurons represents a dramatic cellular transformation in terms of morphology (highly proliferative round cells compacted in colonies transform into an intricate network of post-mitotic cells connected with axons and dendrites) and functionality (loss of pluripotency and gain of the capacity to transmit electrical signals via neurotransmitters). It requires drastic chromatin conformation and modification changes, ultimately leading to the gene expression switch associated with this far-reaching cell fate transition.

In the context of medical research, generation of neurons from pluripotent cells has been widely seen as a potential promising treatment option for patients suffering from neurodegenerative diseases, such as Alzheimer's and Parkinson's disease. Since neurons do not divide, stem-cell-based replacement therapy remains the most likely and hopeful solution. Indeed, studies in animal models have already demonstrated the power of stem cell transplantation therapy, where ESC- and iPSC-derived neuron progenitors and neurons have improved the clinical outcome and life expectancy of diseased animals (Kim et al. 2013). It is thus clear that understanding the molecular underpinnings of this cellular transition – in addition to being scientifically interesting – is of utmost importance for bringing the therapy safely to the clinic.

There are various systems for neuronal differentiation which span a multitude of neural cell types (cortical neurons, astrocytes, motor neurons, to name but a few) and time segments of the process (e.g. pluripotent cells to neural progenitors; neural progenitors to neurons). The vast majority of studies focus only on a segment of the differentiation process (either until or from the neural progenitor stage), rather than the full timeline from pluripotent cells to fully differentiated neurons. Proteomics in particular has been rarely implied for studying a full neuronal differentiation transition owing to technical challenges and the until recently insufficient depth of proteome coverage. The last time-course study of full neuronal differentiation of ESCs using a similar differentiation system was performed in 2009 and the authors identified only 1200 proteins - a depth far below the demands of an integrated multi-omics analysis (Chaerkady & Kerr 2009). This "proteomics gap" has resulted in a lack of complete, multi-level molecular characterization of the neuronal differentiation process. To address this knowledge gap, we used the latest mass-spectrometry-based technologies to generate a dataset of much deeper coverage (470%), thereby obtaining a more comprehensive overview of the dynamic transition of pluripotent ESCs to fully differentiated neurons. To study the level at which the protein expression regulation occurs (tanscriptional, post-transcriptional, translational), a transcriptome dataset (RNA-seq) was generated in the same neuronal differentiation system and compared the expression changes on both levels. Our integrative analysis showed that the protein expression changes are nearly perfectly mirrored in the RNA-seq data, indicating that they are regulated on transcriptional level.

A detailed bioinformatic analysis of the factors which change significantly during neuronal differentiation revealed a co-expression cluster of 98 proteins which are highly enriched for neurological processes. A motif-recognition-based predictive analysis identified Sox2 (Sex determining region Y-box 2) as one of the top regulators of this protein cluster. We therefore focused our further analysis on Sox2.

Sox2 is a transcription factor fulfilling key regulatory roles in different biological contexts. It is most well-known as one of the three 'core pluripotency factors': together with Oct4 and Nanog it constitutes the center of a complex regulatory network involved in pluripotency maintenance and induction (Zhang & Cui 2014; Takahashi & Yamanaka 2006). Its expression is detected already in the early morula stage of embryonic development and is afterwards localized to the inner cell mass (ICM) of the blastocyst, from which ESCs are derived (Avilion et al. 2003). The indispensible role of Sox2 during early embryonic development is demonstrated by the fact that its zygotic deletion causes failure to form pluripotent epiblast and is therefore embryonically lethal (Avilion et al. 2003). Moreover, Sox2 was a part of the first set of genes used to generate induced pluripotent stem cells from somatic cells, further showcasing its essential role in the pluripotency network (Takahashi & Yamanaka 2006).

Interestingly, the core pluripotency factors Sox2 and Oct4 have also been shown to play important roles in cell lineage specification (Wang et al. 2012). Rather than being principal repressors of differentiation, they are also involved in a complex dose-, context-, and co-factor-dependent regulatory system which controls the cellular specification into different developmental fates (Wang et al. 2012). While Oct4 promotes the mesendoderm lineage, Sox2 represses it and induces a neuroectoderm fate instead (Zhao et al. 2004; Thomson et al. 2011; Wang et al. 2012). Indeed, while the expression of Sox2 declines during embryogenesis, it continues to be expressed throughout mouse nervous system development and into adulthood, even in some mature neurons (Ferri et al. 2004; Cavallaro et al. 2008). When the expression of Sox2 is reduced to 30% in adult mice, the mutants display severe neural stem cell proliferative defects as well as dead neurons in different regions of the brain (Ferri et al. 2004). Remarkably, it has also been shown that ectopic expression of Sox2 alone is sufficient to directly reprogram human fibroblasts to induced neural stem cells (iNSCs) (Ring et al. 2012). Taken together, this underlies the critical role of Sox2 in neuronal differentiation and development. This notion is also supported by our data, which points to Sox2 as a top regulator of neural differentiation.

Sox2 has the particular property that it is present and active both in ESCs and neurons, where it fulfills different functions. During neuronal differentiation, it begins to target different genes and while not all mechanisms behind this have been elucidated, it is clear that Sox2 relies on its varying interaction partners for target selection (Kondoh & Kamachi 2010; Zhang & Cui 2014). The cooperation with a different binding partner alters the Sox2 target gene specificity (Kondoh & Kamachi 2010). The best studied interaction is between Sox2 and Oct4 in pluripotent cells. Together, they form a dimer which regulates the expression of thousands of genes genome-wide – activating genes involved in pluripotency maintenance and repressing genes inducing lineage specification (Boyer et al. 2005). In neural development, Sox2 cooperates with various other binding partners, such as the brain-specific Brn2 in neural progenitors and Pax6 in visual system primordia during lens development (Tanaka et al. 2004; Kamachi et al. 2001). Understanding how Sox2 fulfills its varying regulatory functions during neuronal development requires identification of its changing interaction partners during the process. Albeit significant progress in the expansion of the Sox2 interactome (Gao et al. 2012; Huang & Wang 2014; Rafiee et al. 2016), a comparative analysis between its binding partners at the beginning and end of neuronal differentiation is still lacking. An important reason is that while Sox2 continues to be highly expressed in neural progenitors, its expression strongly decreases or completely disappears in most mature neurons. In our differentiation system, however, we were able to detect sufficient Sox2 expression in the terminal neurons (Suppl. Fig. 2), making it a suitable model to study the changing Sox2 interactome at the beginning and end of neuronal differentiation.

With the rapid improvement of mass-spectrometry-based technologies, the protein interactomes can be studied with much more depth and in a more comprehensive way than ever before. A milestone in this regard was the development of chromatin immunoprecipitation coupled to mass spectrometry, termed ChIP-MS, whereby proteins (and nucleic acids) are first crosslinked and then an antibody is used to pull down the protein of interest along with its interaction partners (Wang et al. 2013; Won et al. 2015). More recently, our lab developed a novel version of this technique termed ChIP-SICAP (**S**elective **I**solation

of **C**hromatin-**A**ssociated **P**roteins) allowing for the specific isolation of DNA-bound proteins from soluble protein complexes (Rafiee et al. 2016) (Fig. 2.8). Based on our data which pointed out Sox2 as an important regulator of neural differentiation, as well as on previous knowledge which indicated that its changing function in different cellular states relies on different interaction partners, we used ChIP-SICAP to establish the first-ever Sox2 chromatin-associated interactome both in ESCs and terminally differentiated neurons. Our results demonstrate a remarkable transition from stem-cell to neuronal interaction partners of Sox2 (Fig. 2.10). Furthermore, we found that Sox2 is associated with several epigenetic factors, some of which are common between the beginning and end of differentiation, some unique (Fig. 2.10).

Finally, we investigated the effects of the joint DNA binding of Sox2 and its newly identified interactors on target gene expression, as well as the biological processes in which these targets are involved. To this end, an **A**ssay for **T**ransposase-**A**ccessible **C**hromatin using **seq**uencing (ATAC-seq) dataset was generated for different time points during neuronal differentiation. We find that for some of the key Sox2 interactors we identified, the level of chromatin accessibility on their common binding sites with Sox2 correlates with the transcription levels of the most proximate genes, suggesting a possible gene expression regulatory function of the Sox2+interactor complex (Fig. 2.11). We further find that the processes in which the target genes are involved are highly dependent on the cellular differentiation state, indicating that the Sox2+interactor complex can have an activating or repressing effect dependent on the biological context (Fig. 2.12). Finally, we mined publically availabe ChIP-seq datasets for Sox2 and the interactors we identified in our ChIP-SICAP experiment. We identified the common binding sites across the genome and investigated the correlation between the expression levels of the Sox2-interactor and the target gene. We show that highly correlated and highly anti-correlated target genes are grouped in functionally distinct groups, many of which are related to neural development and epigenetic remodeling, strongly suggesting that the Sox2-interactor complex can have both an activating and a repressive effect on different genes in the same cells (Fig. 2.13).

2.2. Experimental design

We used a chemical- and matrix-based *in vitro* cell system protocol in which mESCs are transformed to terminally differentiated cortical glutamatergic neurons (Fig.2.1A,B) (Bibel et al. 2007). 129X1/SvJ ESCs were taken off leukemia inhibiting factor (LIF) and brought into 3D culture to form embryoid bodies (or "Cellular Aggregates", CAs); those were treated with retinoic acid (RA) and finally dissociated and plated out on plates covered with polyornithine (PORN) and laminin, where they formed terminally differentiated neurons (TN) 2 days later (Fig 2.1A,B; detailed protocol description in "Materials and Methods"). The quality of the neuronal population was assessed by light and fluorescent microscopy, using a neuron-specific dye (Suppl. Fig. 1).

Cells were collected every two days for proteomic analysis and every four days for transcriptomic (RNA-seq) and chromatin accessibility (ATAC-seq) analysis (Fig 2.1A). The protein interactome of Sox2 was examined at the first (ESC, day 0) and last (TN, day 10) time point of the differentiation process.

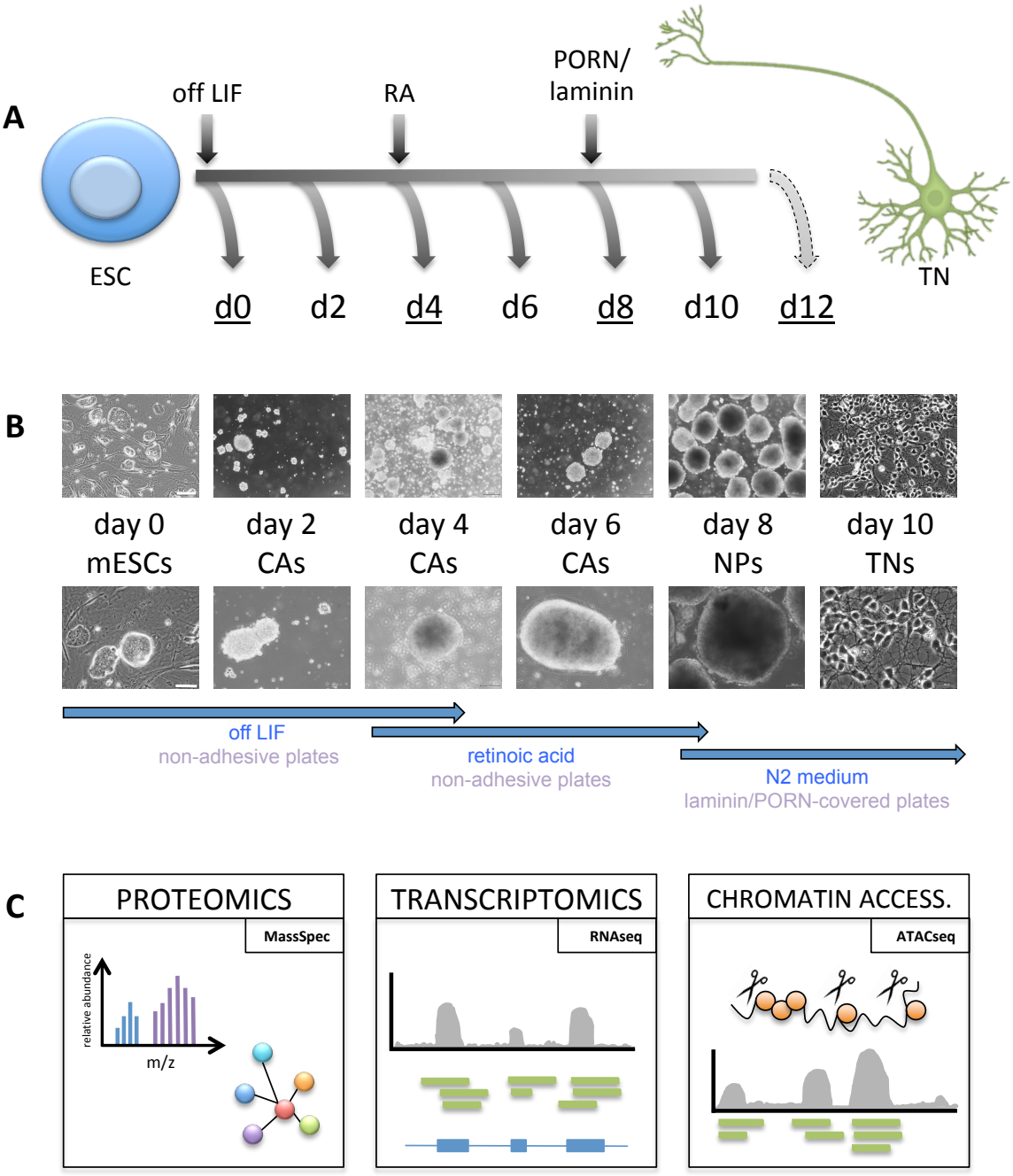


Figure 2.1 Multi-omics analysis of neuronal differentiation. (A) Schematic representation of the differentiation timeline. Cells were collected every two days between day 0 and day 10 for proteomic analysis and every four days between day 0 and day 12 for transcriptomic and ATAC-seq analysis. (B) Light microscopy and procedure outline of neuronal differentiation. Scale upper row: day0: 100 µm, days 2-8: 200 µm, day 10: 50 µm Scale lower row: 50 µm, days2-8: 100 µm, day 10: 20 µm. (C) Schematic representation of multi-omics data analysis. Whole proteome and Sox2 interactome analysis were performed with mass spectrometry, the transcriptome changes were studied via RNA-seq and ATAC-seq was used to study the chromatin accessibility changes during neuronal differentiation.

2.3 Global proteome characterization of neuronal differentiation

We used high-end mass spectrometry technology to study the dynamic protein expression changes which occur during differentiation (Fig. 2.2). Cells were collected in biological replicates every two days during differentiation and subjected to a novel method for super-sensitive proteome isolation and sample preparation (Hughes et al. 2014). After enzymatic digestion of the proteins, the peptides were labeled using a multiplexing technology called Tandem Mass Tag (TMT), allowing the subsequent pool of the samples into a single tube, thereby maximizing the quantification accuracy. The samples were fractionated using high-pH liquid chromatography (HPLC) and ran on an Orbitrap Fusion™ mass spectrometer using SPS-MS3, a synchronous precursor selection method providing the highest degree of sensitivity and accuracy available today (Fig. 2.2A) (see "Materials and Methods" for detailed method description).

We identified a total of over 5674 proteins in the first and 5695 proteins in the second replicate, 4510 of which were common, thereby achieving a much deeper proteome coverage (over 470%) than any published data in a comparable system (Fig. 2.2B). A principal component analysis (PCA) of the global proteomic data revealed segregation of the biological replicates and separation based on the differentiation stage, underlining the high reproducibility and indicating that the detected changes happen progressively as a result of the differentiation process (Fig 2.2C). The most distinct separation between neighboring time points occurs between day 8 and day 10 - an observation which is in line with the biological transformation taking place between these time points: on day 8, the neural progenitors are still packed in large embryoid bodies (>500 µm/diameter; Fig 2.1B); they are then plated out as a monolayer with neuro-inducing plate coating (PORN and laminin) and neuronal medium, thus quickly reaching a terminal differentiated state at day 10 (Fig. 2.1B).

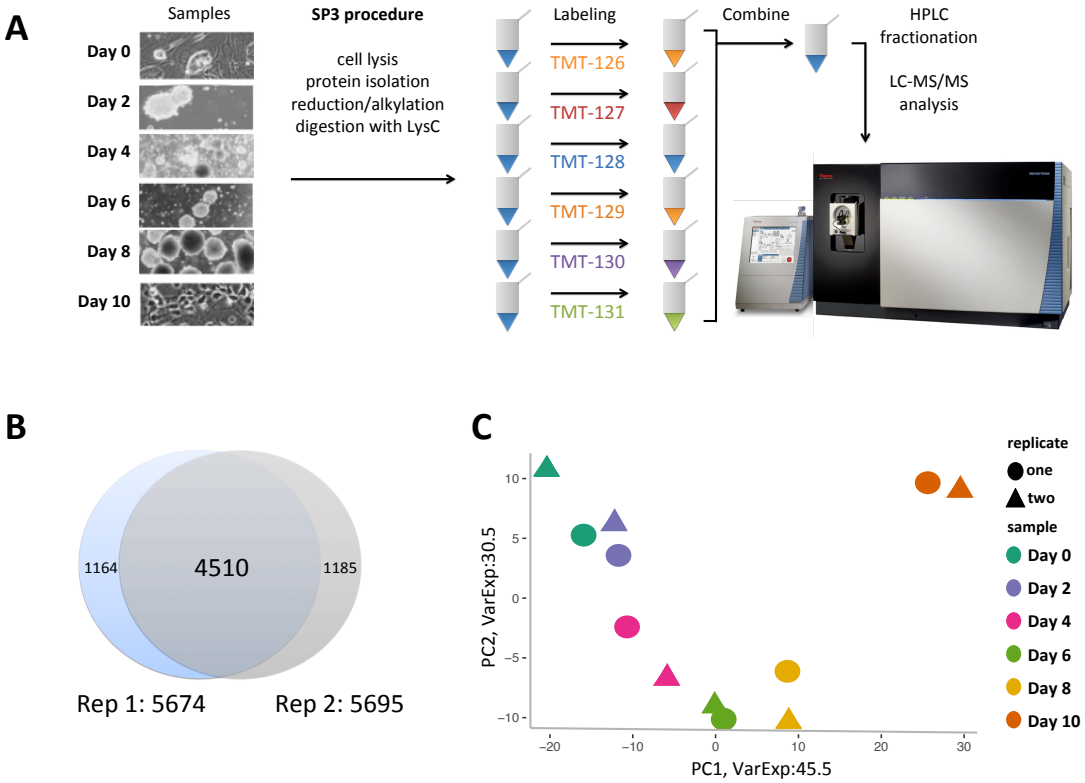


Figure 2.2 Global proteome analysis during neuronal differentiation. (A) Experimental workflow. Proteome sample preparation from cells collected every two days, peptides were labeled using TMT multiplexed labeling, pooled and fractionated with HPLC and analyzed using SPS-MS3. Full details in main text and "Materials and Methods". (B) Number of proteins identified per biological replicate. (C) PCA of global proteome data.

In order to gain insight on the proteins which are functionally relevant for the studied cellular transition, the subset of the proteome which displays significant expression changes needs to be identified. To this end, we applied a bayesian statistical analysis method called "limma", which is based on the use of linear models for assessment of differential expression and is part of the Bioconductor software project (full details in "Materials and Methods"). We find that compared to day 0, the proteomic changes increase over time and thus most significant change is between the beginning and end point (Fig. 2.3A). Our analysis revealed that overall 12% (550) of all proteins identified in both replicates display a significant expression change during differentiation, with the rest of the proteome remaining mostly stable throughout the process (Fig 2.3).

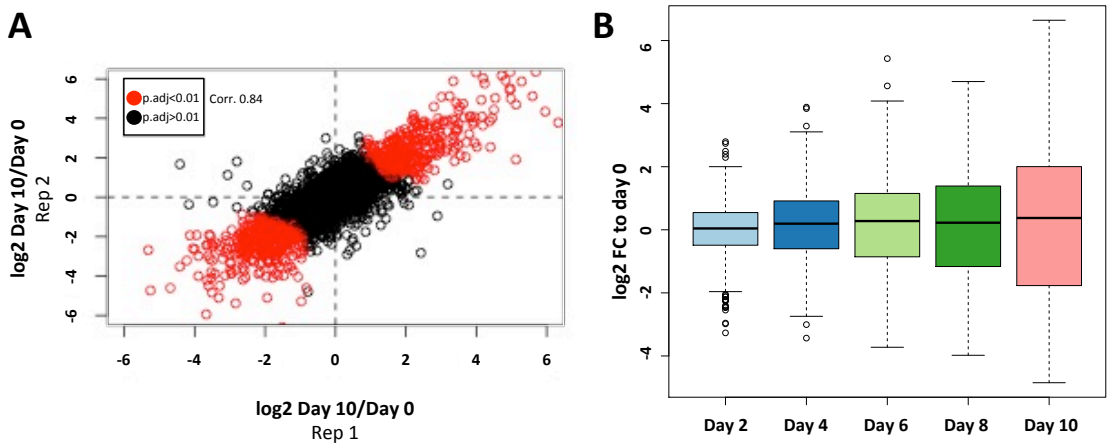


Figure 2.3 Differentially expressed proteins during neuronal differentiation. (A) Distribution Day10/Day0 ratios between biological replicates in log2. Proteins in red have passed the limma test (p.adj.<0.01) and are thus considered significantly changing. (B) Gradual expression change increase in the 550 significantly changing proteins (p.adj<0.1) shown as relative log2 fold change (FC) ratio to day 0.

2.4 Co-expression, co-regulation, co-functionality?

In order to examine the possible regulatory and functional networks in which proteins changing during neuronal differentiation are organized, we first segregated them based on their dynamic expression patterns during the differentiation time course. To this end, we used unsupervised clustering of the log2 ratios of each time point to day 0. Our analysis demonstrated that the changing proteins followed five distinct expression change patterns (Fig. 2.4). Two of the co-expression clusters are steadily upregulated throughout differentiation (Clusters 1 and 5), two are downregulated (Clusters 2 and 3) and one contains proteins which are downregulated in the early differentiation stages, but return to their initial level later in reprogramming (Cluster 4). This suggests that the neuronal differentiation program is initiated already during the very early differentiation stages and progresses steadily until the end. This is interesting because retinoic acid, the prime differentiation agent, is only added to the cells at day 4 - prior to that time point the cells are only deprived of LIF and taken in 3D culture.

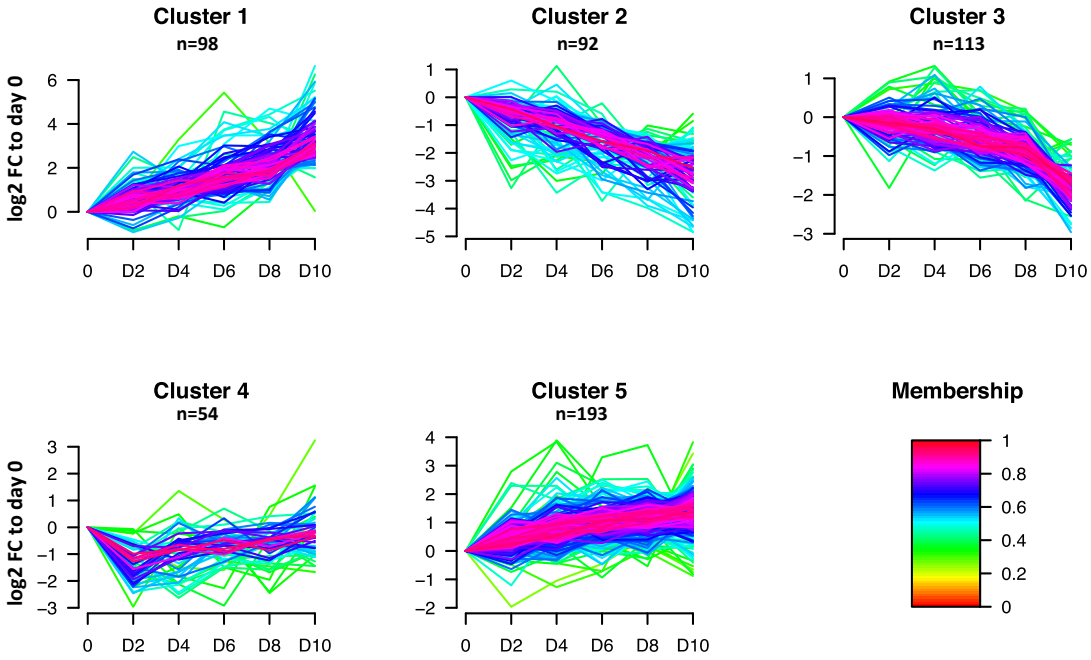


Figure 2.4 Unsupervised clustering of limma proteins. The relative log2 fold change (FC) ratio to day 0 was established for all 550 proteins after limma analysis. Each line represents a protein. "n" is the number of proteins per cluster. For inclusion into a cluster, we used an upper and lower limit of $\log_2 0.5$ and $\log_2(-0.5)$, respectively. The "membership" value and color scheme represents how well a protein fits into the average cluster profile.

Since the expression of genes to proteins is regulated on multiple levels (transcriptional, post-transcriptional, translational), we then addressed the question of the regulatory level at which the expression of the changing proteins was altered. To this end, we generated a transcriptome dataset in the same differentiation system and compared the proteome and transcriptome expression changes of the limma proteins (Fig 2.5). We report a very high correlation between the transcriptome and proteome with both levels following the same expression change patterns over the time course (Fig 2.5).

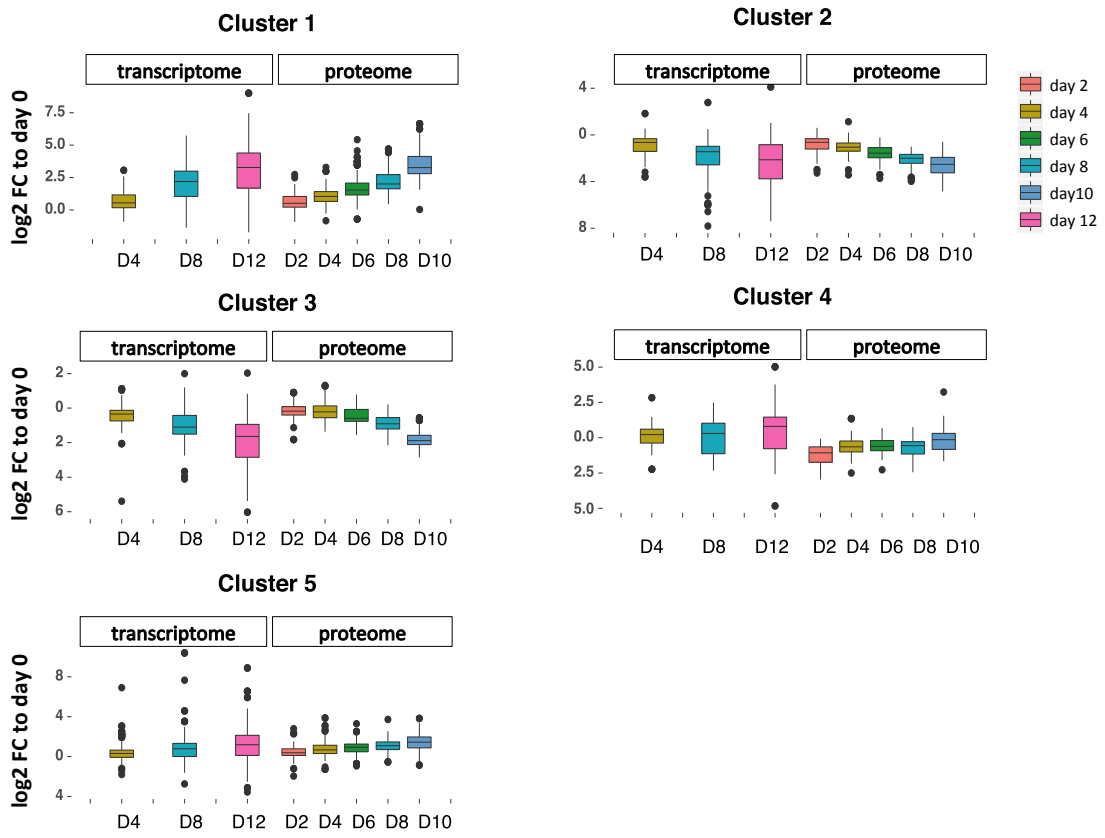


Figure 2.5 Comparative analysis of proteome and transcriptome of limma proteins. Expression change is shown as relative fold change ratio to day 0, in log2. Clear similarity of the proteome and transcriptome expression change patters visible in all five clusters. RNA-seq data by M. Gehre and D. Bunina, analysis by D. Bunina and N. Abazova.

Based on this result, we concluded that the proteomic expression changes observed during differentiation are primarily regulated at the transcriptional level, indicating a high coordination between the different regulatory systems.

To examine the link between co-expression and functionality, we applied gene ontology (GO) enrichment analysis on each of the five distinct protein coexpression clusters (Fig. 2.6). Intriguingly, one of them, Cluster 1, displayed very high overenrichment for processes involved with neurogenesis, neuronal development, neural differentiation and related processes such as cytoskeleton organization, axonogenesis and regulation of synaptic transmission (Fig. 2.6). The dynamic expression pattern that the proteins in this cluster follow is one of high and consistent increase from initiation until the last time points of differentiation,

suggesting that the neuronal lineage network becomes activated early and consistently increases over time.

Two other clusters, Cluster 3 and Cluster 5, also displayed enrichment for distinct biological processes, including cell cycle progress and morphogenesis of a branching structure, respectively. The proteins in Cluster 3 are downregulated already from day 0 onwards and decrease sharply between day 8 and day 10 (Fig. 2.4), reflecting on the decrease of the high proliferation rates in ESCs towards the post-mitotic state of the TNs. The proteins in Cluster 5 increase during differentiation and the development of branching structures is necessary for neuronal development and maturation. This result showcases the clear link between co-expression and co-functionality in the context of neuronal differentiation.

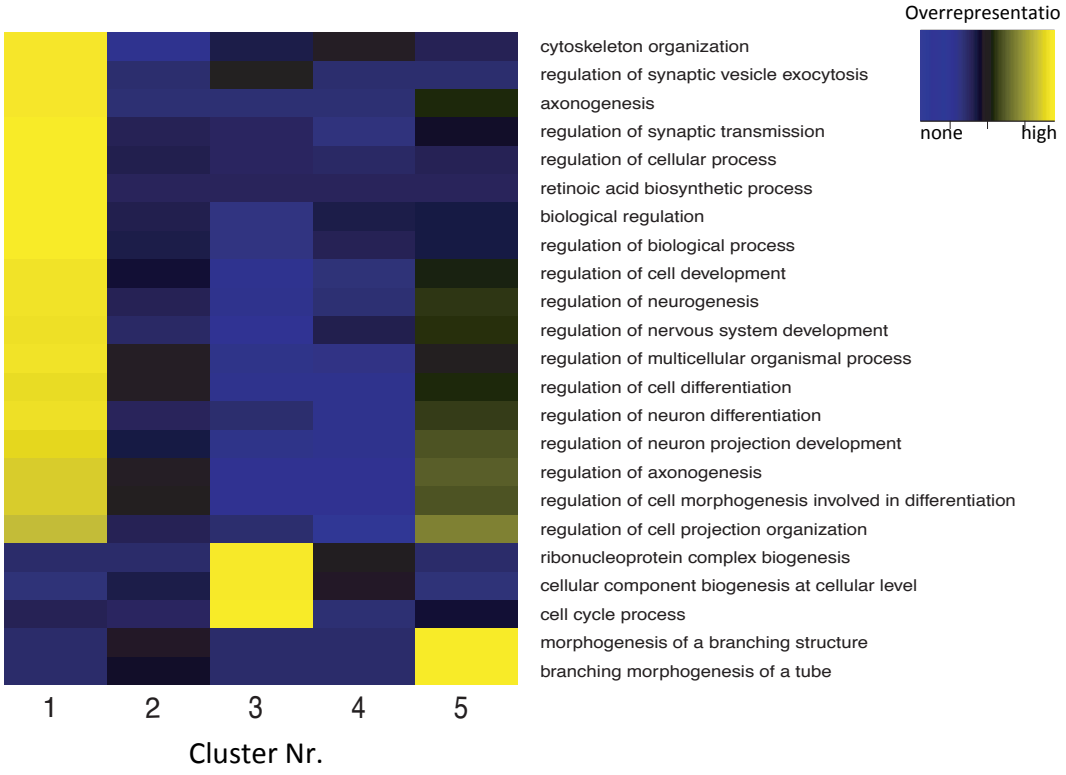


Figure 2.6 Gene ontology enrichment of differentially expressed proteins. All protein clusters from Fig. 2.4 were tested for enrichment of Biological Process GO terms compared to unregulated proteins. Cluster 1 is highly enriched for neurological and neural differentiation processes. Fischer's exact test, p.adj<0.05.

Having identified a group of proteins involved in neuronal differentiation which are both co-expressed and highly functionally related, we proceeded to investigate

whether they are also co-regulated, i.e. if they are regulated by the same transcription factors (TFs). Using a method originally developed for transcriptomic datasets, we performed an analysis based on transcription factor motif recognition, called iRegulon, which predicts whether a group of genes is regulated by the same transcription factors (Fig. 2.7)(Janky et al. 2014). Interestingly, our results indicate that the great majority (70%) of all proteins in Cluster 1 are predicted to be regulated by the same 7 transcription factor families. From all TFs belonging to these families, 27 passed our filter for direct motif similarity (Suppl. table 1) and one of the top ones was the core pluripotency factor Sox2 (Fig 2.7B). This result suggests that the proteins in Cluster 1 not only share the same dynamic expression pattern and neuron-related functionality, but are also likely regulated by the same transcription factors, one of which is Sox2.

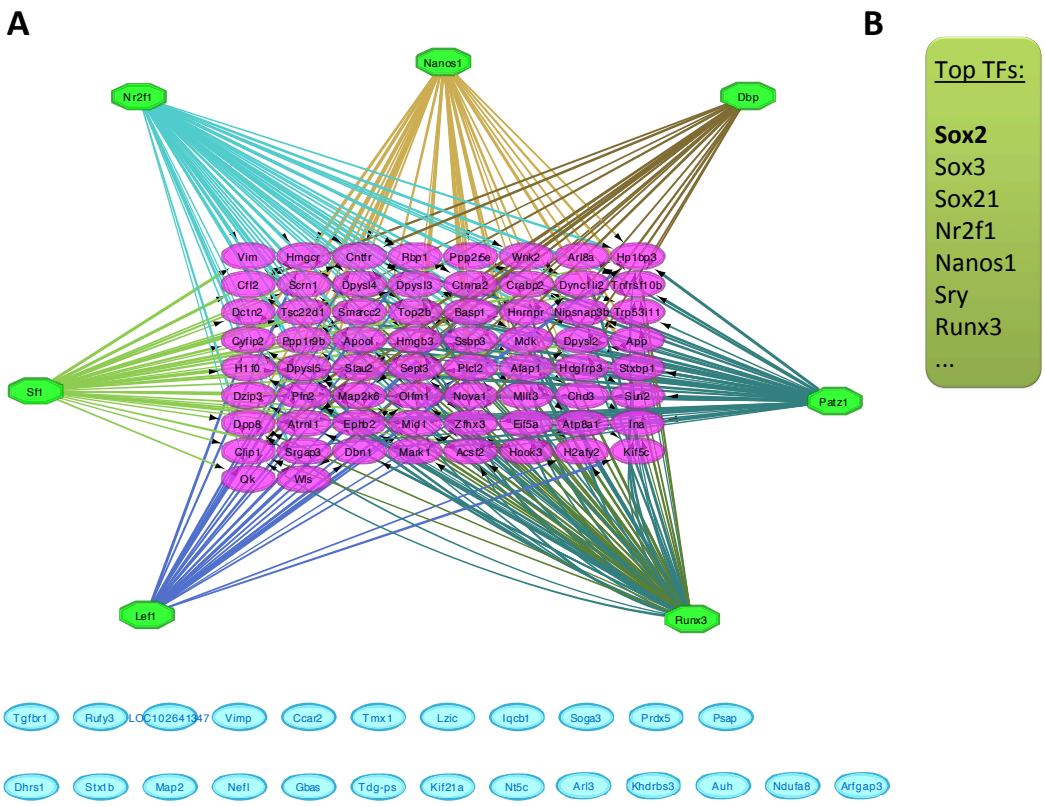


Figure 2.7 iRegulon analysis of differentially expressed proteins in Cluster 1. (A) 70% of all tested proteins (in pink; the rest 30% is in blue) are predicted to be regulated by 7 TF families (green). (B) Sox2 is among the top predicted regulators of the Cluster 1 proteins. For list see Suppl. table 1. False discovery rate<0.001, normalized enrichment score = 3. Full description of the iRegulon tool in (Janky et al. 2014).

2.5 Chromatin-associated Sox2 interaction network in ESCs and neurons

We found that in our differentiation system, Sox2 is present both in ESCs and terminal neurons (TNs) (Suppl. Fig. 2) and our data suggests that it is a key regulator of genes involved in neuronal differentiation (Fig. 2.7). In order to elucidate the mechanisms which allow Sox2 to fulfill its different functions in ESCs and terminal neurons, we performed an experiment to determine its protein interaction partners in both cellular states, while it is bound to DNA. For this, we used ChIP-SICAP (selective isolation of chromatin-associated regions), a novel technique developed in our lab (Rafiee et al. 2016). The exact procedure is described in the "Materials and Methods" section of this thesis. Briefly, cells were crosslinked using 1.5% formaldehyde under standard ChIP conditions and the nuclei were extracted and lysed after which the chromatin was sheered to a fragment size of ~500 base pairs (Suppl. Fig. 3). Sox2 was first immuno-precipitated using a specific antibody and then the ends of the DNA fragments were biotinylated using terminal deoxynucleotidyl transferase (TdT) and pulled down using streptavidin-coated beads (Fig. 2.8). This ensures the specific isolation of the chromatin-associated Sox2 interactome. As a negative control, the same procedure was done using an unspecific IgG antibody. Finally, the isolated proteins were subjected to proteolytic digestion and measured on LC-MS/MS.

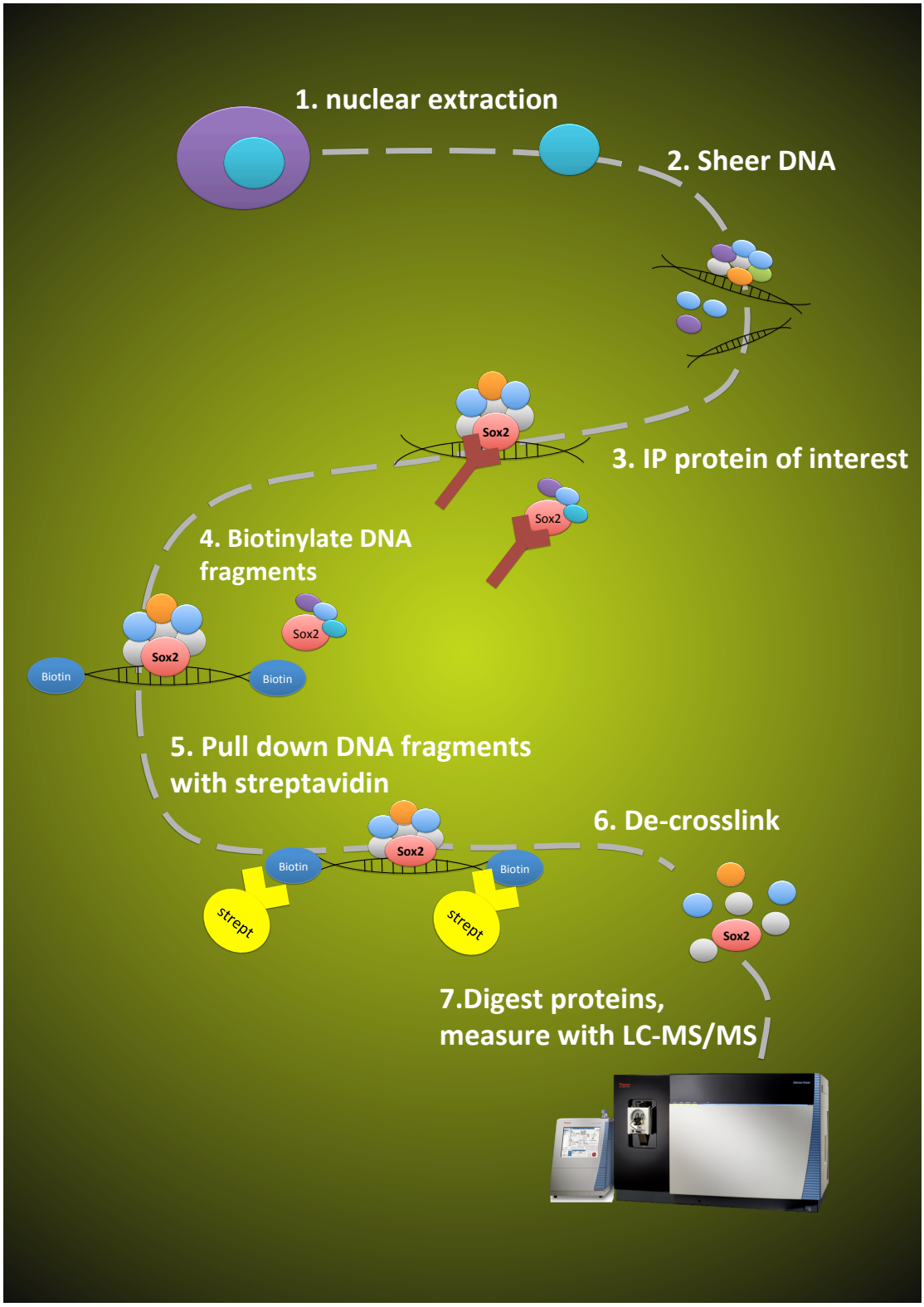


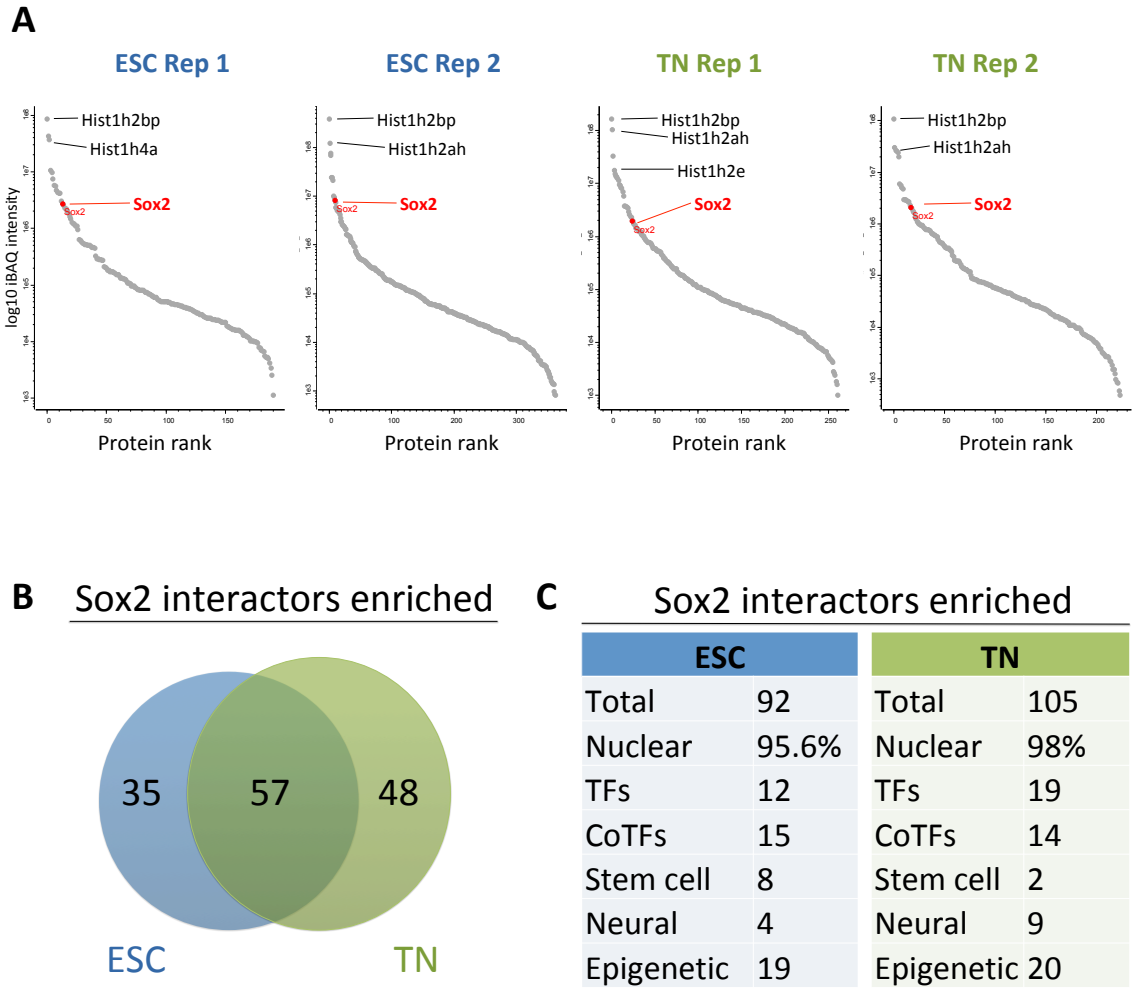
Figure 2.8 Schematic illustration of ChIP-SICAP for Sox2.

In order to examine the enrichment efficiency, we ranked all identified proteins based on their protein iBAQ intensity (Fig. 2.9A), in each sample and replicate. In all cases, Sox2 was among the most highly enriched proteins, indicating a high

efficiency of the ChIP pull-down. There were few proteins with higher intensity than Sox2, most notably histones, which is expected given their high abundance and the chromatin isolation (Fig. 2.10A).

To ensure that only true Sox2 interactors and no ubiquitous proteins are included in the analysis, the experiment was performed again with an unspecific antibody, IgG, and only proteins which were either exclusively present in the Sox2 pull-downs or displayed a minimum of 4X enrichment over the negative control in both biological replicates were included in the analysis.

We identified a total of 92 proteins in ESCs and 105 in TNs of which 57 were common between the two cell types (Fig. 2.10B,C). Over 95% of all identified proteins are nuclear, showcasing the high specificity of the method (Fig. 2.10C). Interestingly, 13% of all Sox2 interactors in ESCs and 19% in TNs are transcription factors (TFs), and over 13% co-transcription factors (CoTFs), underlining the regulatory essence of the Sox2-centred network (Fig. 2.10C).



ESC		TN	
Total	92	Total	105
Nuclear	95.6%	Nuclear	98%
TFs	12	TFs	19
CoTFs	15	CoTFs	14
Stem cell	8	Stem cell	2
Neural	4	Neural	9
Epigenetic	19	Epigenetic	20

Figure 2.10 Sox2 enrichment and interaction proteins in ESCs and TNs. (A) All identified proteins are ranked based on their intensity (X axis, protein rank with the most intense protein being 1, second-most intense 2 and so on; Y axis is log10 of the protein intensity). Sox2 is among the most highly enriched proteins in each sample and raplicate. (B) Number of enriched Sox2 interactors over the negative control in ESCs and TNs. (C) Overview of the identified enriched Sox2 interactors in each cell type. Color code: blue=ESCs, green=TNs.

Going through a detailed analysis of the Sox2 interactome in both cell types, we observed a remarkable transition from plutipotency- to neurobiology-related Sox2 interaction partners (Fig. 2.10). As expected, one of the most highly enriched interactors present exclusively in ESCs is Oct4. Other prominent stem cell factors identified solely in ESCs include Sall4 and Rif1 - the former is a known Sox2 interactor, the latter is novel. The proteins identified as Sox2 interactors in TNs encompass a multitude of neuronal genes, including Fabp7, which has been shown to play an essential role for neurogenesis in vivo (Watanabe et al. 2007), the topoisomerase Top2b which plays a critical role in forebrain development and

neuronal migration (Yang et al. 2000)(Lyu & Wang 2003) and the CoTF Ctbp2, which is a coactivator of retinoic acid signaling (Bajpe et al. 2013) and is thus essential for neuronal differentiation. The majority of the proteins identified in TNs have not previously been described as Sox2 interactors, possibly owing to the fact that virtually all interactome studies thus far have been focused on pluripotent or neural stem cells, but not on differentiated neurons.

For all shared Sox2 interactors between ESCs and TNs, we calculated an association rate relative to Sox2 between the two cell types in order to determine if the association of the respective interactor is preferential in ESCs or TNs (Fig. 2.10). Remarkably, we observe a stem cell- to neuronal- factor transition here as well: the known pluripotency-related proteins Trim24 and Trim28 preferentially bind to Sox2 in ESCs and the neuronal factors Adnp and Myef2 increase their Sox2 association rate in TNs.

Notably, a significant part of the identified Sox2 interactors are proteins involved in different levels of epigenetic remodeling. For example, in ESCs we report the histone deacetylase Mta2, as well as Hells - a factor involved in DNA methylation during development. Among the shared factors between the two cell types are the DNA methyl transferase Dnmt1 and the histone methyl transferase Ehmt1. In TNs we identify the multifunctional chromatin regulator Hmgb1 and Wiz - a factor linking histone methyltransferases to Ctbp, the factor involved in retinoic acid signaling which we also identified in TNs (Fig. 2.10) (Ueda et al. 2006). The high representation of epigenetic factors (20 in ESCs and 19 in TNs) might be directly related to the fact that the ESC-to-TN transition is associated with dramatic chromatin and DNA methylation reorganization.

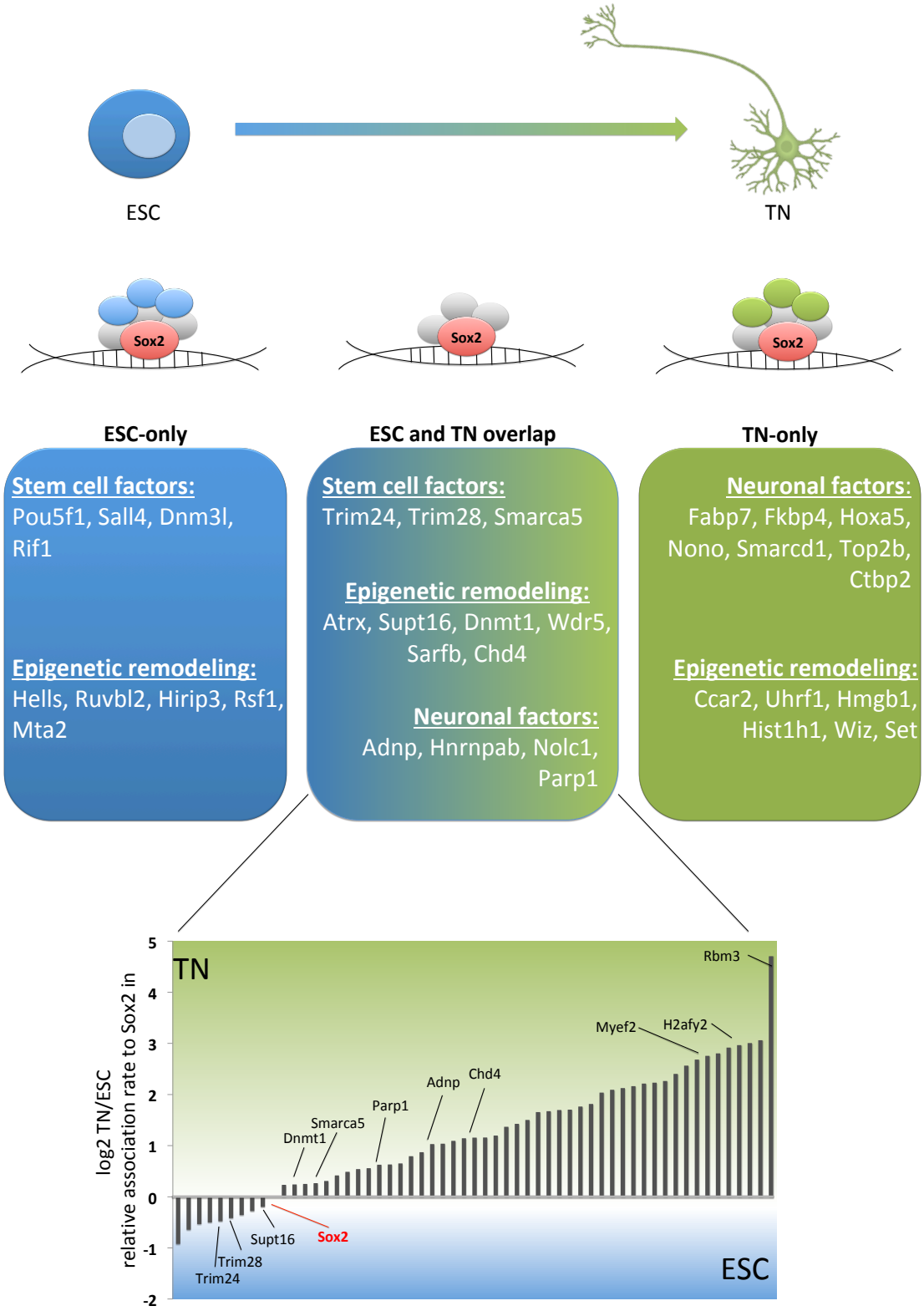


Figure 2.10 Chromatin-associated Sox2 interactome in ESCs and TNs. We report three groups of Sox2 interactors: present exclusively in ESCs, present in both ESCs and TNs and present exclusively TNs. There is a clear transition from stem cell to neuronal interactors. This is also recapitulated in the Sox2 association rate of the shared interactors between ESCs and TNs. The association rate is defined as intensity ratio between the interactor and Sox2 in TNs over ESCs.

Taken together, our data suggests that Sox2 is involved in the process of neuronal differentiation of ESCs and it relies on a highly dynamic interaction network which displays a remarkable stem cell-to-neuronal transition and is likely responsible for altering the Sox2 target genes during this dramatic cellular transformation.

2.6 Sox2 and its interactors: effects on target gene expression and biological processes

To examine the effect of the associated binding of Sox2 and its newly identified interaction partners, we first studied the chromatin accessibility change in the common binding sites across the genome during neuronal differentiation and correlated it to the transcriptional change of the genes in close proximity to these binding sites. This was done in collaboration with Dr. Bunina, a shared member of the labs of Dr. Noh and Dr. Zaugg. To this end, **A**ssay for **T**ransposase-**A**ccessible **C**hromatin using **seq**uencing (ATAC-seq) was performed at four time points during differentiation. ATAC-seq provides quantitative information about the genomic regions which are more "open" (i.e. accessible) or "closed" (i.e. inaccessible) (Buenrostro et al. 2013). Using available databases for transcription factor DNA binding motifs, ATAC-seq allows the estimation of transcription factor activity increase/decrease between cellular states: if there is an overall enrichment increase of the motifs that a particular TF binds to, this is an indication of TF activity increase (for full method description see (Buenrostro et al. 2013)).

For the following analysis, in addition to the chromatin-associated Sox2 interactome described in Fig 2.10, we also included Sox2 interactors which we identified in a ChIP-MS experiment (i.e. without additional selection for chromatin-bound proteins). We did this in order to include Sox2-centered complexes which might bind on chromatin only transiently or very dynamically and therefore be missed by the ChIP-SICAP approach.

We first identified the ATAC-seq peaks containing motifs for both Sox2 and its interactors. We then identified the genes which are positioned in the vicinity of the

common ATAC-seq peaks and are therefore putative targets of the Sox2-interactor complex. We then calculated a ratio between the ATAC-seq peak intensity in TNs (day 12) over ESCs (day 0) and correlated it to the target gene expression change between these time points (Fig. 2.11). We observed a correlation with five Sox2 interactors, the highest of which, 0.24, is between the Sox2-Fubp1 ATAC-seq peak intensity change and their target expression (Fig. 2.11). Unexpectedly, in the case of Sox2-Oct4, this correlation appears lower, 0.20. However, this is possibly explained by the fact that at day 12, Oct4 is not expressed at all and therefore its motifs are occupied by other factors.

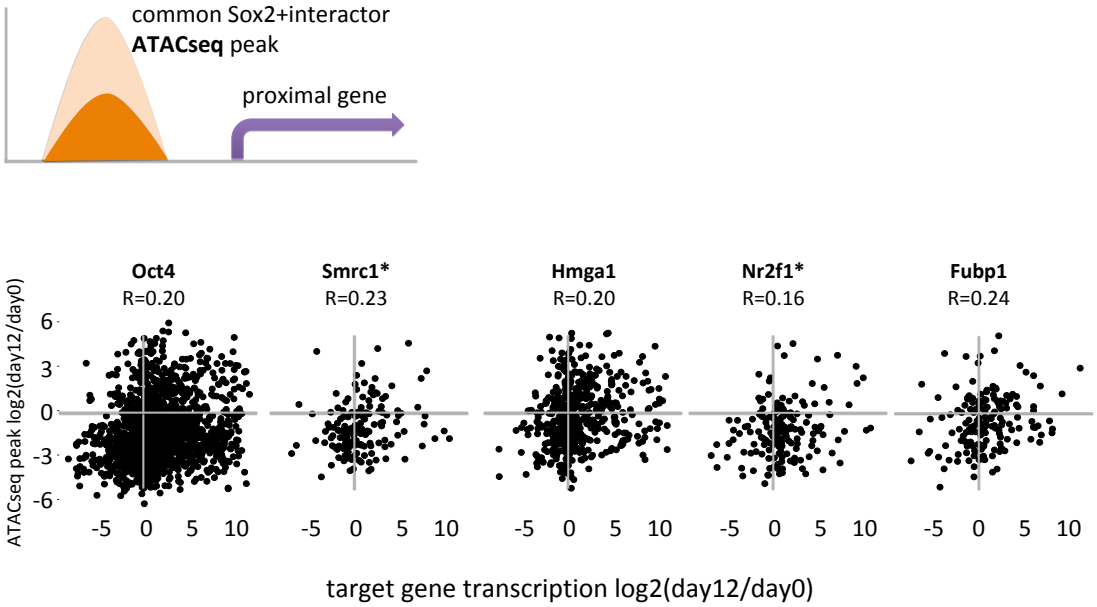


Figure 2.11 Spearman correlation between common ATAC-seq peaks of Sox2 and its interactors and their target genes expression. Genes marked with * were derived from our ChIP-MS dataset, the others from ChIP-SICAP. ATAC-seq data and analysis by D.Bunina, RNA-seq data by M.Gehre.

We went on to investigate the biological processes in which the common targets of Sox2 and its interaction partners are involved. To this end, we first examined the differential chromatin accessibility of the common motifs regions (contained within common ATAC-seq peaks) of Sox2 and its interaction partner between day 12 and day 0 (Fig. 2.12). We focused on the regions with differential chromatin accessibility in either TNs or ESCs (Fig. 2.12A). We subjected the genes in the vicinity of these regions (the putative target genes) to GO term enrichment analysis and found significant enrichment for the targets of four of the Sox2

interactors: Nr2f1, Hmga1, Smrc1, and Oct4 (Fig. 2.12B). Remarkably, we observe enrichment for targets which at the respective time point are silent (like neural genes in ESCs, in proximity to common Sox2 and Oct4 binding sites) or active (like genes involved in blastocyst formation in ESCs, in proximity to common Sox2 and Smrc1 binding sites). This possibly relates to the "priming" effect which Sox2 is known to have together with other transcription factors, which keep genes in a silent, but "ready" state for transcription once an additional signal is received (see "Introduction" for details). Our results suggest that the chromatin regulator Smrc1, whose expression positively correlates with the accessibility of its common binding sites with Sox2 (Fig. 2.11), might cooperate with Sox2 in ESCs to maintain a stem cell state and simultaneously prime them for differentiation (see enriched terms in Fig. 2.12). Hmga1, whose common binding sites with Sox2 are more accessible in neurons than in ESCs, might act in a complex with Sox2 as activator of neurogenesis-related processes (Fig. 2.12). This is in line with our observation that the overall expression levels of their common targets correlates with the accessibility of their common binding sites (Fig. 2.11). Nr2f1 might cooperate with Sox2 as a repressor of the mesoderm lineage and Oct4 might block the exit of pluripotency and entry into the neuronal lineage by repressing the expression of neuronal genes.

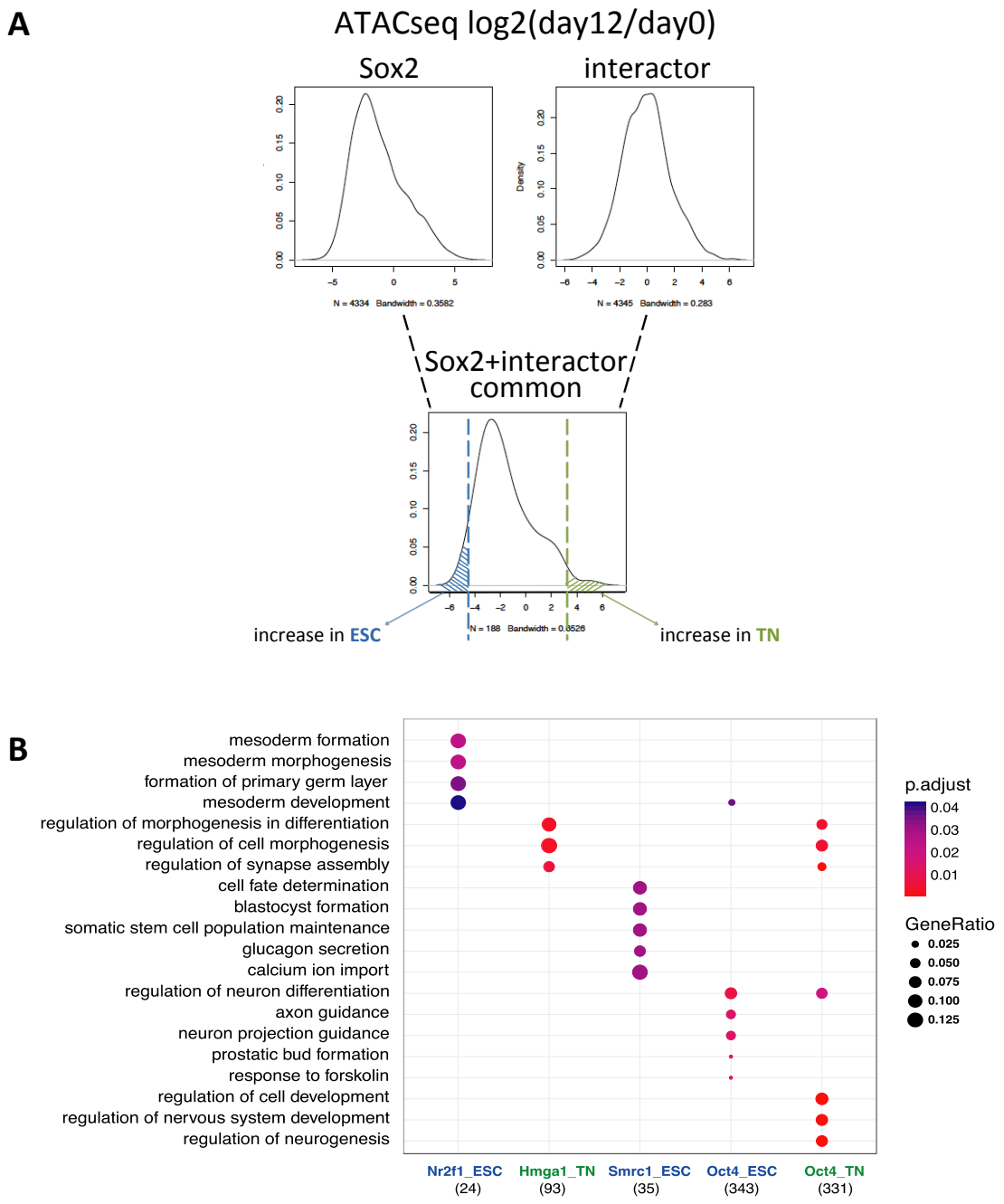


Figure 2.12 Gene Ontology enrichment of targets of Sox2-interactor complexes. (A) Schematic representation of the ATAC-seq peak distributions of Sox2, its interactors and the common peaks between them, between ESCs (day 0) and TNs (day 10). We selected the top 20% most changing peaks and performed GO enrichment for their putative targets. (B) GO enrichment for genes in the vicinity of the common ATAC-seq peaks of Sox2 and its interactors. Only genes which displayed significant GO enrichment are shown. p.adjusted<0.05. The Gene Ratio is a measure of the proportion of target genes belonging to the respective GO term. The number under each gene name indicates the number of common targets it has with Sox2. ATAC-seq data and analysis by D.Bunina.

Finally, we mined publically available ChIP-seq datasets and found data for a total of 18 TFs which we had identified as Sox2 interactors. The ChIP-seq data was obtained in ESCs. We overlayed the signals of Sox2 and each of these TFs and selected their common binding sites across the genome (Fig. 2.13A). Then, the closest target genes (up to 100kb from the common binding site) were selected and the gene expression correlation was calculated between the Sox2-interactor and its respective target, across the entire differentiation timeline (Fig. 2.13B). We focused our analysis on interactors which are either highly correlated (pearson correlation >0.5) or highly anti-correlated (pearson correlation <-0.5) to their targets (Fig. 2.13B). Intriguingly, there are distinct GO term enrichments for these two groups, belonging to the same Sox2 interactor. For example Sall4, a known pluripotency gene, is anti-correlated to genes involved in neurogenesis and neural development and correlated to genes involved in embryonic development and stem cell maintenance (Fig. 2.13C). One possible explanation is that in a complex with Sox2, it might simultaneously act as an activator of some (pluripotency) genes and repressor of other (neuronal) ones.

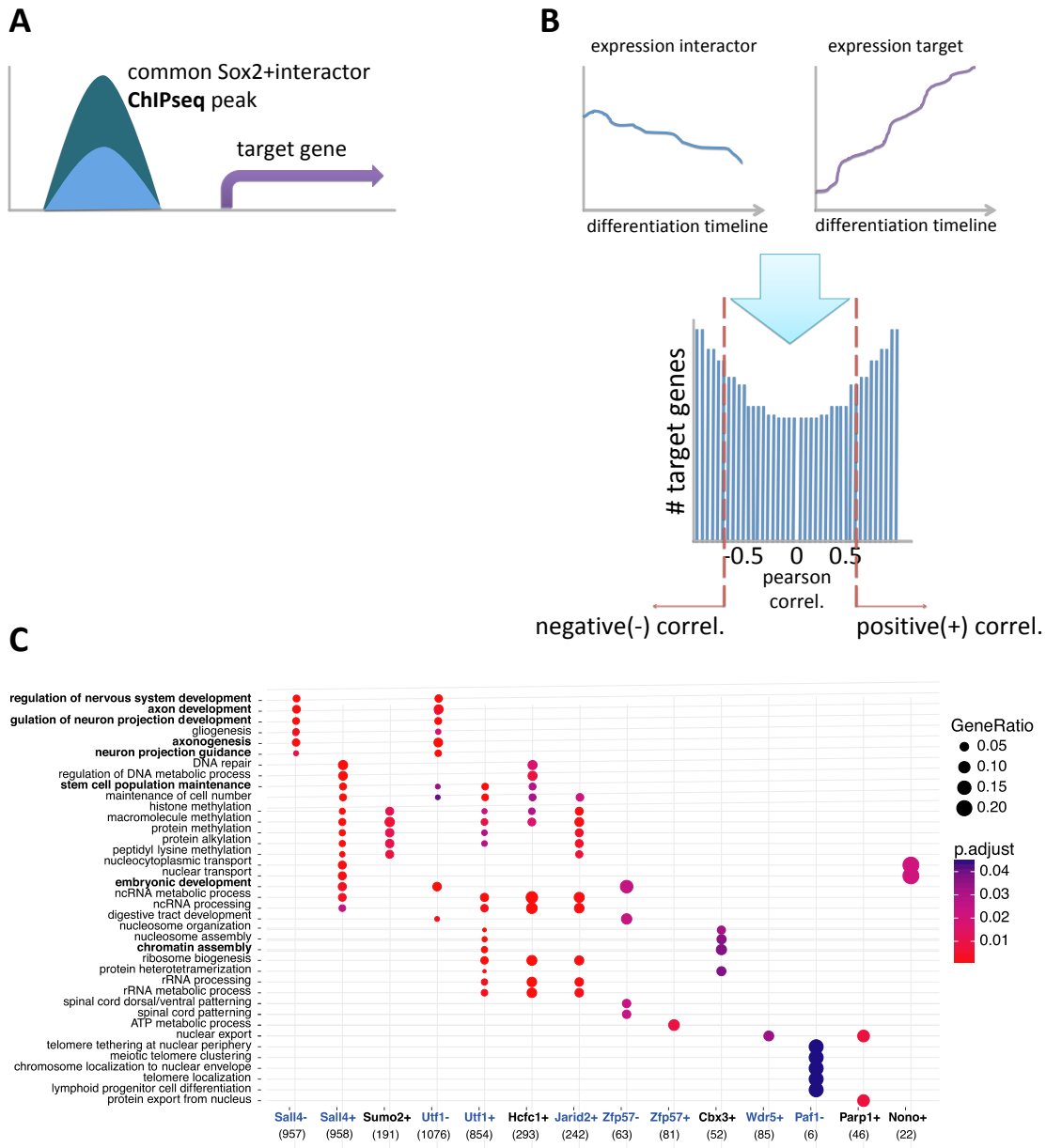


Figure 2.13 ChIP-seq-based GO term enrichment analysis correlated and anti-correlated target genes of Sox2-interactor complexes. (A) We selected common ChIP-seq peaks between Sox2 and its interectors and the closest genes to these peaks, which we refer to as target genes. (B) We correlated the dynamic expression during the course of neuronal differentiation between the Sox2 interactor and its targets. We selected the most highly correlated (pearson>0.5) and anti-correlated (pearson<-0.5) target genes for further analysis. (C) GO term enrichment analysis of the correlated (+) and anti-correlated (-) target genes. Interactors marked in blue were only shown to bind to Sox2 in ESCs, the black ones were common between ESCs and TNs. Analysis by D. Bunina.

In summary, our analysis suggests that the joint DNA binding of Sox2 and some of the key interaction partners we identified has an effect on target gene transcription (Fig. 2.11, Fig 2.13) and that their joint target genes are highly

involved in developmental regulatory processes during neural differentiation (Fig. 2.12, Fig. 2.13). Furthermore, the results suggest that the interaction between Sox2 and its partners can have an activating or repressing effect dependent on the biological context (ESCs vs TNs) (Fig. 2.12, Fig. 2.13).

2.7 Discussion

The process of complete transition from embryonic stem cells to terminally differentiated neurons entails great molecular, morphological and functional changes which need to be elucidated in order to bring the technology to the clinic. In this study, we presented a high throughput analysis of the dynamic proteome changes taking place during this major cell fate transition. Our dataset greatly exceeds the depth of any previously existing proteomic study, thereby serving as a valuable resource for future research and enabling integrative analysis in a multi-omics context.

We identified a cluster of proteins with highly dynamic and coordinated expression increase during the differentiation time-course, which was enriched for processes related to nervous system development and neuronal differentiation. Based on predictive bioinformatic analysis, we identified Sox2 as a key factor involved in the regulation of this group of proteins. This was not surprising, since Sox2 is known to be involved both in pluripotency maintenance and neuronal development (Avilion et al. 2003; Ferri et al. 2004; Cavallaro et al. 2008).

The fact that Sox2 relies on its protein interaction partners in order to change its target specificity in different cellular contexts has prompted various studies which aimed to determine the Sox2 protein interactome (Kondoh & Kamachi 2010; Zhang & Cui 2014; Gao et al. 2012; Huang & Wang 2014; Rafiee et al. 2016). However, the majority of this work was focused on ESCs, where Sox2 fulfills a very different function compared to neurons. In fact, the comprehensive protein interactome network of Sox2 has never been established in differentiated neurons before. The main reason is that while Sox2 is still highly abundant in neural stem

cells, where it maintains their progenitor identity, it completely disappears in most mature neurons and remains present in only few of them (Cavallaro et al. 2008). This left an important question open: how does the Sox2 interactome change between pluripotent and fully differentiated neuronal cells? The fact that in our differentiation system Sox2 was still present in the terminal neurons (Suppl. Fig. 2) allowed us to address this knowledge gap by generating the first comparative Sox2 interactome dataset at the beginning and end of neuronal differentiation. We identified a "core" Sox2 interactome which was present both in ESCs and TNs, as well as subsets specific to either cell type (Fig. 2.10). Remarkably, we observe that the composition of the Sox2 interactome undergoes a stem-cell-to-neuronal transition, with key pluripotency-related proteins like Oct4, Sall4 and Rif1 present only in ESCs and proteins involved in neurogenesis, like Fabp7 and Top2b present only in TNs (Fig. 2.10). Interestingly, we observe similar functional transition even in the "core" Sox2 interactome: we find that the Sox2 association rate of proteins with known stem cell-related functions is higher in ESCs (Trim24, Trim28) and the association rate of proteins involved in neuronal differentiation is higher in TNs (Adnp) (Fig. 2.10). It should be noted that many of these proteins are multifunctional and a strict distinction between a "stem cell" and "neuronal" proteins is not always possible. For example, Smarcd1, which we identified as a Sox2 interactor in TNs, is a known component of the neuron-specific chromatin remodeling complex nBAF and the neural progenitor-specific complex npBAF, which are key regulators of the epigenetic remodeling mechanism occurring during the progenitor-to-neuron transition (Lessard et al. 2007). However, the same protein was also shown to play an important role in pluripotent cells, were its knockdown impairs the self-renewal capacity of the stem cells (Gao et al. 2012). That being said, when taken together, the Sox2 interactors we identified in both cellular states form a clear pattern of pluripotent-to-neuronal functional transition.

Our interrogation of the putative effects of cooperative binding between Sox2 and a subset of its newly identified interaction partners suggests that it might have a regulatory effect on the gene expression of their common target genes (Fig. 2.11). Based on our data, this effect can be either activating or repressive, dependent on the cellular context (ESCs or TNs). Our results indicate that the chromatin

regulator Smrc1 may cooperate with Sox2 in ESCs to maintain their stem cell character while at the same time prime them to exit pluripotency (a major enriched term is "cell fate determination"). "Priming" here refers to a repressive binding at a target promoter, which can be quickly activated upon association of additional factors. The dual function of activating pluripotency-associated genes while at the same time repressing genes involved in lineage commitment has been described for the cooperative binding of Oct4 and Sox2 in pluripotent cells (see "Introduction" for details). Another interesting Sox2 interactor we identified is the chromatin remodeling protein Hmga1. Based on the increased chromatin accessibility measured with our ATAC-seq approach, Hmga1 appears to increase its cooperative DNA binding with Sox2 in neurons compared to ESCs and their common targets are enriched for processes related to neuronal differentiation, such as regulation of synapse assembly and morphogenesis during differentiation (Fig. 2.12). Interestingly, in glioblastoma stem cells, Hmga1 has been described as a regulator of Sox2 itself, which controls its expression by modifying the chromatin architecture at the Sox2 promoter (Lopez-Bertoni et al. 2016). Given that Sox2 is also self-regulatory, it is thus possible that Hmga1 and Sox2 act as a complex entangled in an expressional feedback loop involved in different stem-cell related settings: regulation of stemness in glioblastoma and promotion of neuronal differentiation in CNS development. Exploring this possibility calls for functional experiments *in vivo* and *in vitro* which test the phenotypic manifestation of a disrupted binding between the two factors.

It should be noted that ATAC-seq-based analysis offer a proxy of possible cooperative binding of different transcription factors which is solely based on prior information of the genome-wide recognition motifs of each factor and the accessibility of the regions containing these motifs. Stringently investigating the cooperative binding between Sox2 and selected interaction partners requires a different approach, such as performing ChIP-seq for each of them and comparing their binding sites. However, this is only feasible for selected interaction partners (a high-throughput ChIP-seq experiment for hundreds of interaction partners is very time- and resourse-consuming). The ATAC-seq based analysis can serve as a valuable foundation for selecting of the most promising candidates. At the same

time, combined with our RNA-seq data, ATAC-seq offers a more comprehensive overview of the relationships between chromatin accessibility and transcriptional regulation at different sites.

As a next step, we are currently working on the generation of a Sox2 ChIP-seq dataset in both ESCs and TNs. While ChIP-seq data in pluripotent cells exists in wide abundance, there is currently no publically available data in glutamatergic neurons. Adding this information to our current dataset will expand our understanding about the relationship between the changing Sox2 interaction partners and Sox2 target genes during neuronal differentiation. Combined, the ChIP-SICAP and ChIP-seq datasets from ESCs and TNs can also serve as valuable resource for future functional analyses. For example, the knock-down effect of distinct Sox2 interactors can be interrogated in the context of Sox2 target specificity in neuronal cells. This is a field in which the current knowledge is very sparse and our integrated multi-omic datasets lay a solid foundation for its expansion.

3. Epigenetic memory and spatio-temporal signature during neuronal differentiation and induced pluripotency

3.1 Introduction

In the last decades, a vast body of research has supported the notion that development is under epigenetic control (Kiefer 2007). DNA methylation, the complex histone modification code and exchange of histones are all key elements of the epigenetic rearrangements which drive cellular specialization and determine cell identity (Kiefer 2007). Notably, these epigenetic rearrangements are also crucially involved in the reverse process, namely cellular reprogramming, when the developmental program of the cells is turned backwards towards pluripotency or totipotency (Apostolou & Hochedlinger 2013). In fact, the first reprogramming experiments using somatic cell nuclear transfer (SNCF) simultaneously served as the first and ultimate demonstration that cellular differentiation is caused by epigenetic and not by genetic changes (Briggs & King 1952). SNCF is based on enucleating an egg and inserting the nucleus of a more differentiated cell. This reprogramming technology was a milestone in bioresearch and demonstrated that by resetting the epigenetic landscape of the chromatin, the cells can re-gain the necessary developmental potential to give rise to (cloned) organisms. SNCF was used to generate the first mammal from reprogrammed adult cells ("Dolly the sheep")(Wilmut et al. 1997) and later on from fully differentiated cells (Hochedlinger & Jaenisch 2002).

Despite the indisputable success of SNCF technology, the cloned animals displayed clinical anomalies, which have been attributed to the so called "epigenetic memory" (Gao et al. 2003; Santos et al. 2003; Ng & Gurdon 2005; Ng & Gurdon 2008; Wee et al. 2006). The concept of epigenetic memory refers to residual gene expression and chromatin structure traits which have been inherited from the donor cell. Cloned animals have been shown to retain histone modification marks from their cell of origin (Santos et al. 2003; Wee et al. 2006; Ng & Gurdon 2008), as

well as gene expression signatures (Gao et al. 2003; Ng & Gurdon 2005), indicating that the reset of the epigenome back to a totipotent state was not 100% complete and the residual marks affect the developmental process.

An entire new branch of reprogramming and epigenetic studies emerged when Takahashi and Yamanaka made the groundbreaking discovery that it is possible to reverse the developmental clock and transform fully differentiated cells back to a pluripotent state only by overexpressing four defined factors, Oct4, Sox2, Klf4 and c-Myc (OKSM) (Takahashi & Yamanaka 2006). Aside from its immense promise for biomedical research (one can now create personalized pluripotent cells from the patient's own somatic cells and differentiate them to almost any needed cell type), induced pluripotency provided an invaluable tool for studying the molecular mechanisms behind cell-fate plasticity and lineage commitment. Compared to SCNT, generating induced pluripotent stem cells (iPSCs) is much less technically challenging and, importantly, allows the performance of experiments which require high cell numbers as starting material and which explore the dynamic molecular transformations during reprogramming.

Since their establishment 12 years ago, iPSC have been the foundation of a multitude of epigenetic studies, serving as a unique tool to examine the mechanisms behind cellular plasticity (Krishnakumar & Blelloch 2013). The phenomenon of epigenetic memory retention in particular has been the object of much investigation. While iPSCs generated from various cell types reach a functional and epigenetic state highly similar to embryonic stem cells (ESCs), they also retain specific epigenetic signatures which are remnants from the cell type of origin (Vaskova et al. 2013). Hochedlinger's lab showed that early-passage iPSCs generated from fibroblasts, hematopoietic and myogenic cells display distinct epigenetic and transcriptional patterns, and vary in their *in vitro* differentiation potentials (Polo et al. 2010). Importantly, these cell-of-origin traits disappeared over long-time culture of iPSCs (Polo et al. 2010). In the same year, Daley's lab reported that the retention of epigenetic memory can be observed also on DNA methylation level (Kim et al. 2010). In this study, the authors generated iPSCs from fibroblasts and hematopoetic progenitors and compared their methylation

patterns to ESCs. They report that fibroblast-derived iPSCs contained 3349 differentially methylated regions compared to ESCs, whereas only 516 were differentially methylated in iPSCs derived from hematopoietic cells. Importantly, cells involved in hematopoiesis and osteogenesis (i.e. responsible for cell specialization) were among the most highly expressed differentially methylated regions, underlining the existence and importance of epigenetic memory retention (Kim et al. 2010). In a study focused on chromatin accessibility, Benvenisty's lab demonstrated that iPSCs derived from pancreatic islet beta cells not only retained an open chromatin state at key beta-cell genes (*INSULIN*, *PDX1*, and *MAFA*), but also preferentially differentiated into insulin-producing cells (Bar-Nur et al. 2011). Remarkably, this differentiation was more efficient compared to ESCs, which are considered the "gold standard" of pluripotency (Bar-Nur et al. 2011). Taken together, these studies demonstrate the significant impact of epigenetic memory on the developmental potential, cellular plasticity and gene expression signature during reprogramming.

While iPSCs can and have been generated from almost any cell type, to date there has been no successful reprogramming of neurons using only the four "Yamanaka factors" (OKSM). Two research groups, Jaenisch and Dyer, reported attempts to do this (Kim et al. 2011; Hiler et al. 2016). Because neurons are very sensitive and difficult to manipulate, the authors of both studies made use of a so called "reprogrammable mouse" - a genetically modified mouse strain containing a genome-integrated doxycycline-inducible OKSM cassette, along with a reverse tetracycline transactivator (rtTA) and an Oct4-GFP pluripotency reporter (Stadtfeld et al. 2010). This system allows the initiation of reprogramming in primary neurons upon addition of doxycycline to the cell culture, thereby omitting the otherwise necessary viral transduction which is particularly challenging for the sensitive postnatal neurons. In 2011, Jaenisch's lab demonstrated that reprogramming of post-natal neurons was only possible if in addition to OKSM overexpression, the key tumor-suppressor p53 was inactivated (Kim et al. 2011). Later on, Dyer's lab developed a protocol which omitted the necessity of p53 inactivation; however, the reprogramming of neurons required mixed culture in which cells from "reprogrammable mice" were cultured as aggregates with wild-

type cells (Hiler et al. 2016). This protocol does not enable neuronal reprogramming using only OKSM overexpression, it requires the dispersion of the cellular aggregates and growth of individual iPSC clones, and it takes 56 days to complete (Hiler et al. 2016).

In this thesis, we present a study examining the presence, potential and spatio-temporal manifestation of epigenetic memory and its effect on the proteome in the context of neural differentiation and reprogramming. Although primary neurons cannot be reprogrammed only by OKSM overexpression, we hypothesized that neurons generated from pluripotent stem cell *in vitro* may give rise to iPSC due to retention of epigenetic memory. Indeed, we successfully differentiated (primary) iPSCs to a neuronal culture, which was then reprogrammed back to (secondary) iPSCs. In order to interrogate the potential effect of epigenetic memory on protein expression, we then used mass spectrometry to analyze and compare the global proteome composition of primary and secondary iPSCs. As a final step, we aimed to investigate the spatio-temporal dynamics of the developmental switches driving differentiation. To this end, we explored the proteomic changes which occur at different sites (rim vs. core) of embryoid bodies during differentiation of iPSCs.

3.2 Full cycle of cell-fate transitions

The "reprogrammable mouse" strain described in (Stadtfeld et al. 2010) and used in Jaenisch and Dyer's studies (Kim et al. 2011; Hiler et al. 2016) is not commercially available. Therefore, we generated it ourselves from separate strains containing each of the desired mutations (for details see Materials and Methods). Specifically, we generated two mouse lines, a double mutant *R26-**rtTA**$^{+/+}$/ Oct4-**GFP**$^{+/+}$* and *Col1a1-**OKSM**$^{+/+}$*. When mated with each other, these mice give rise to heterozygous triple-mutant embryos. Cells extracted from these embryos can be subjected to reprogramming upon addition of doxycycline to the cell culture medium.

MEFs from triple-mutant embryos at 13.5 dpc were reprogrammed to iPSCs (Suppl. Fig.5; procedural details in "Materials and Methods"). Clonal iPS cell lines were generated upon picking distinct iPSC colonies and culturing them separately.

To ensure the erasure of any residual fibroblast-specific epigenetic traits, the lines were cultured more than two weeks and split several times before initiation of neural differentiation (the epigenetic reset of long-time iPSC culturing is described in (Polo et al. 2010)). The pure iPS cell lines are referred to as primary iPSC (p.iPSC). They displayed characteristic iPSC morphology, were strongly Oct4+, AP+ and expressed other key pluripotency factors, among which Nanog, Sox2, Sall4 and Esrrb (Fig. 3.1A, Fig.3.3D, Suppl. Fig.5).

The p.iPSC were differentiated to glutamatergic neurons passing through an embryoid body stage, using Bibel's protocol (Fig.3.1A)(Bibel et al. 2007). The final neuronal culture (day 12) displayed characteristic neuronal morphology and network, was 100% Oct4–negative, 98.2% of the cells stained positive for the neuron-specific marker NeuO (Suppl. Fig. 8) and 86% of the cells were positive for the mature neuronal marker Map2 (Suppl. Fig.6). In addition, the cells expressed the glutamatergic neuronal marker glutamine synthetase, as well as the neurogenesis marker doublecortin, the synaptic vesicle glycoprotein synaptophysin and the neural cell adhesion proteins Ncam1/2 (Suppl. Fig.7).

The neuronal culture was subjected to reprogramming upon addition of doxycycline for 14 days, after which the drug was withdrawn from the culture to ensure the purely endogenous expression of OKSM. We developed a reprogramming protocol which takes a total of 32 days and is based on the step-wise removal of neuronal signals and increase of pluripotency-facilitating ones (for full details, see "Materials and Methods"). To accurately monitor the progression of pluripotency loss and gain, the cells were analyzed via flow cytometry for the pluripotency marker Oct4-GFP (Fig. 3.1A,B). We observed a very clear shift of Oct4+ to Oct4– majority populations during differentiation and vice versa during reprogramming. At day 32, secondary iPSCs (s.iPSCs) had formed. The s.iPSCs displayed characteristic iPSC morphology and strong expression of Oct4-GFP (Fig.3.1B). In addition, they expressed many pluripotency-network genes, among which Sox2, Sall4, Stat3, Lin28 as well as the stem cell marker alkaline phosphatase (Fig. 3.3B).

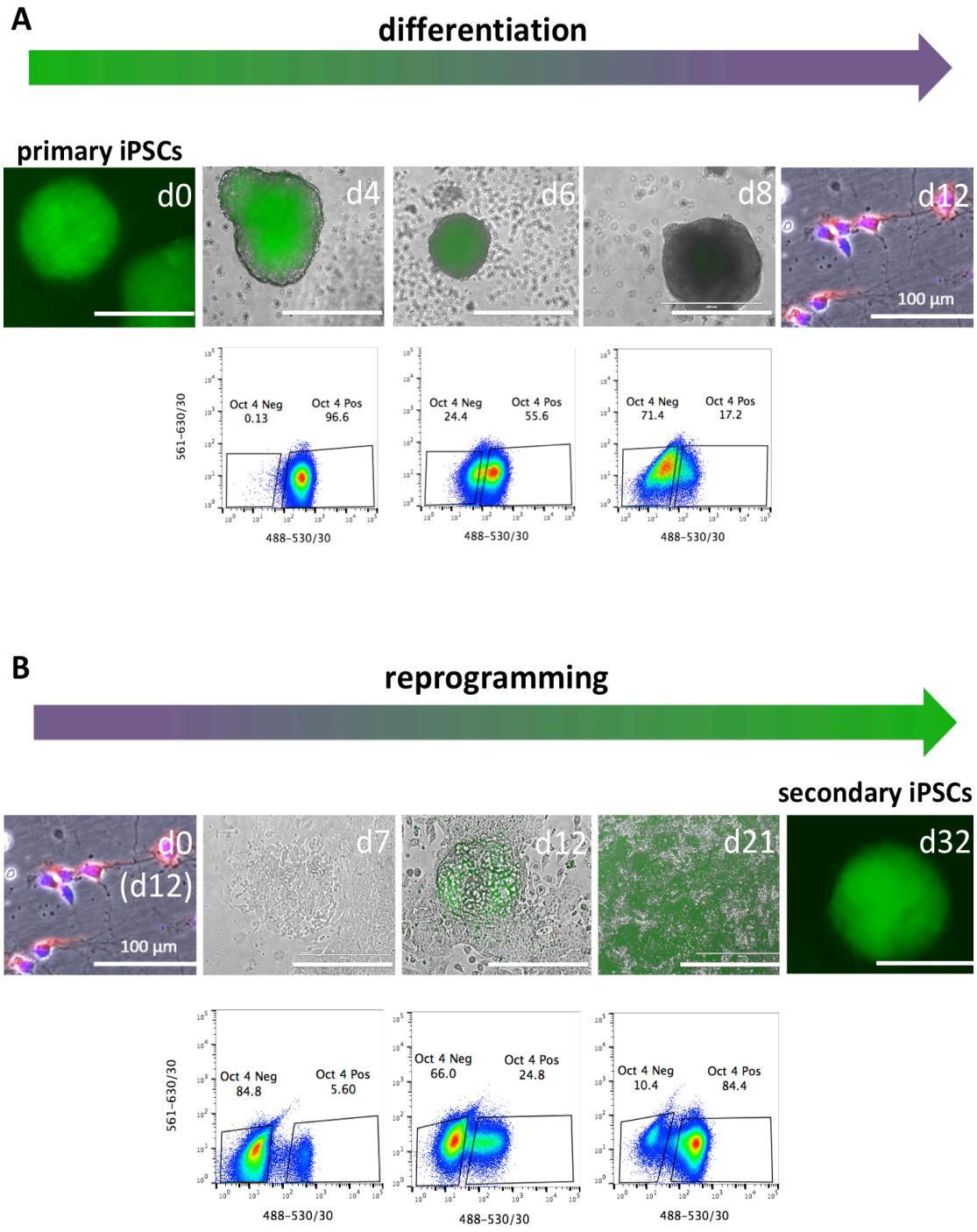


Figure 3.1 Full cycle of cell-fate transitions – iPSCs differentiated to neurons, neurons reprogrammed back to iPSC. (A) Primary iPSCs were differentiated to glutamatergic neurons in 12 days, passing through an embryoid body stage until day8. Oct4-GFP expression is progressively lost starting from the outer rim of the embryoid body and continuing towards the core. Flow-cytometry analysis at days 4-8 displays a clear shift from Oct4-GFP+ to Oct4-GFP– majority. (B) Glutamatergic neurons from (A) reprogrammed to secondary iPSCs in 32 days. Day 12 of differentiation represents day 0 of reprogramming. Flow cytometry analysis at distinct time points reveals a shift from Oct4– back to Oct4+ majority of the cells.

Taken together, these results suggest that the full cyclic conversion of pluripotent cells to differentiated neuronal culture and back to pluripotent cells was successful and required only the controlled overexpression of the four "Yamanaka factors". This supports the hypothesis that in contrast to primary neurons, which cannot be reprogrammed in this way, reprogramming of neuronal cultures generated from pluripotent cells is feasible, possibly owing to their epigenetic memory.

3.3 The dual nature of secondary iPSC

The primary iPSCs underwent a long process of differentiation to neurons and subsequent reprogramming to secondary iPSC (44 days in total; 12 days of differentiation, 32 days of reprogramming). To examine the effect of these transformations on the cells' expression signatures, we compared the proteome profiles of p.iPSCs and s.iPCSs (Fig. 3.2A). The full proteome of each of the two populations was measured in replicates on an OrbitrapFusion mass spectrometer, coupled to an LC system. In the p.iPSCs, 3242 and 3220 proteins were identified in each replicate, 3085 (>95%) of which were common between them. In the s.iPSCs - 3433 and 3353, 3231 (>96%) common (Fig. 3.2B). For both populations, the correlation between the replicates was very high (R=0.96; Fig. 3.2C), indicating high reproducibility of the data. Comparing the proteins identified in both replicates for the p.iPSC and s.iPSC populations, we find that they share 2767 proteins (~90%)(Fig. 3.2B), the expression change correlation between them is very high (R=0.85; Fig.3.2C), and only 6% of all proteins display a significant difference in expression (p.adjusted<0.05, over 2 fold; Fig. 3.3A). This underlines that the primary and secondary iPS cell populations reproducibly display highly similar proteome profiles.

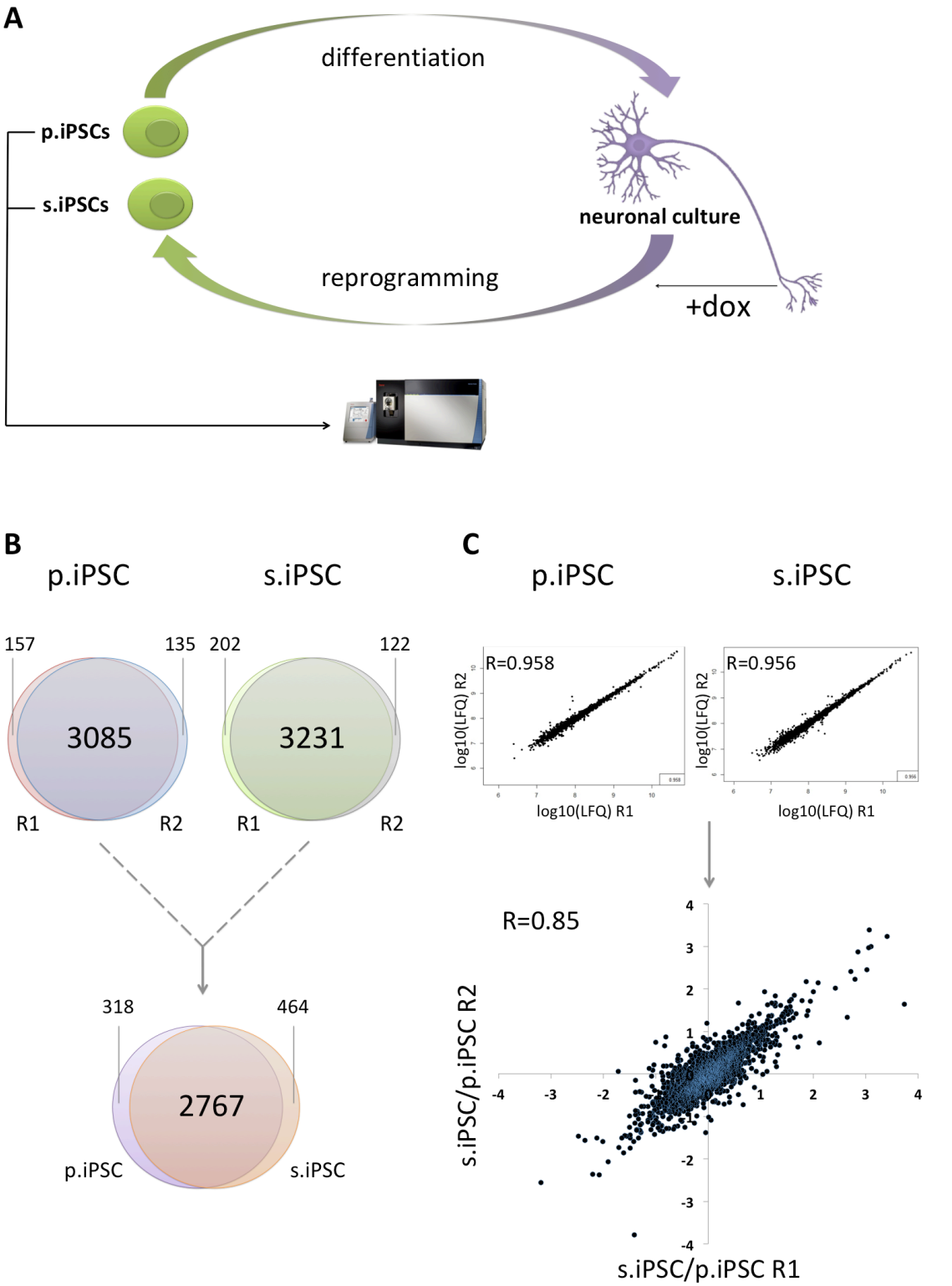


Figure 3.2 Proteome analysis of primary and secondary iPSCs. (A) Primary iPSCs were differentiated to neuronal culture, which was reprogrammed to secondary iPSCs. The two iPSC populations were collected and subjected to full proteome analysis via mass spectrometry. (B) Protein identifications in both replicates and conditions. High overlap between the replicates (>96%) and conditions (~90%). (C) Scatter plots of the log10 protein intensities between the replicates and iPSC populations reveals high experimental reproducibility. Scatter plot of the log2 expression changes between p.iPSC and s.iPSC showcases they are highly consistent between the replicates (R=0.85)

Focusing on the proteins uniquely identified or significantly enriched in either the primary or secondary iPS cell populations, we found a clear distinctive trend which separates them in "more pluripotent" (p.iPSC) and "more neuronal" (s.iPSCs). While the majority of pluripotency-network proteins are shared between the two populations (Oct4, Sox2, Sall4, Rif1, Stat3, Lin28), there are few which were only identified in the primary iPSCs (Nanog, Esrrb, Dppa4) (Fig.3.3B). Several proteins with key functions in neurogenesis were exclusively identified in secondary iPSCs, most notably the neural progenitor marker Nestin (Fig. 3.3B). To gain a more global overview of the differences between the two iPSC populations, we performed GO term enrichment analysis of the proteins uniquely identified or significantly enriched in either one. Our results demonstrate that in p.iPSCs, there is an enrichment for proteins associated with stem cell maintenance and related processes (Fig. 3.3C). For example, proteins associated with RNA processing were also highly enriched in p.iPSC, which is expected in pluripotent cells, as they are transcriptionally hyperactive and contain twice as much RNA (normalized to DNA content) compared to neural progenitors (Efroni et al. 2008). In contrast, proteins overexpressed or unique to s.iPSC were enriched for nervous system development and related processes such as morphogenesis in cell differentiation (Fig. 3.3D).

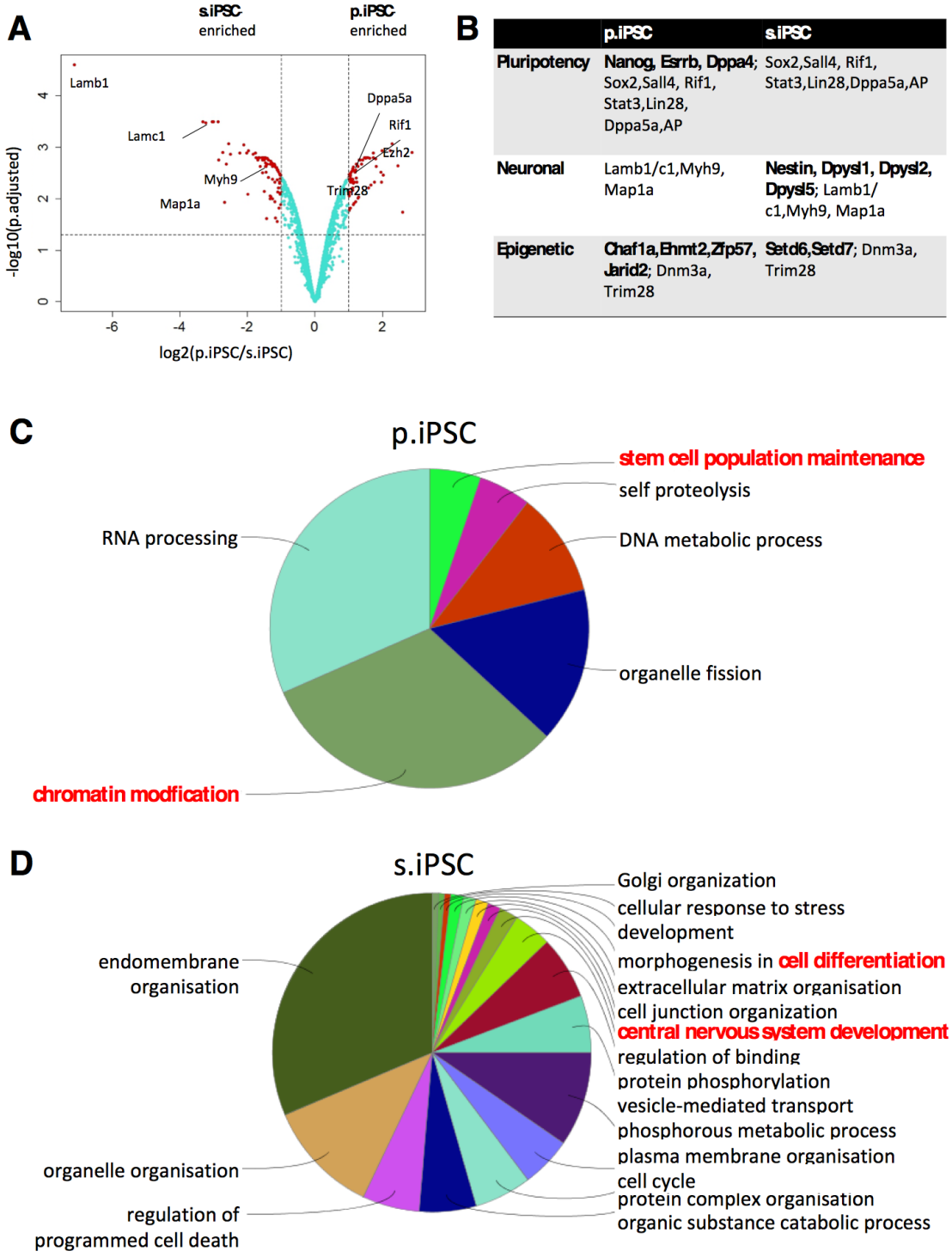


Figure 3.3 Differential expression analysis between primary and secondary iPSCs. (A) Volcano plot of common proteins between p.iPSCs and s.iPSCs. X-axis: log2 expression fold change between both populations. Y-axis: –log10 of the p.adjusted value calculated using the limma statistical test (for details see "Materials and Methods". Significantly (p.adj<0.05) differentially (>2 fold) expressed proteins are displayed in red. Only 170 (6%) of all proteins are differentially expressed. (B) Key pluripotency, neuronal and epigenetic factors identified in each population. Uniquely identified proteins in only one population are displayed in bold. (C,D) GO term enrichment analysis of proteins uniquely

identified in primary (C) and secondary (D) iPSCs. Only significantly enriched terms were included in the analysis ($p<0.05$). For a detailed disambiguation of number of proteins/term and term representation, see Suppl.Fig.8

A key difference between the primary and secondary iPSC was the presence and enrichment of distinct epigenetic regulators. Primary iPSC displayed an overall high enrichment of proteins related to chromatin remodeling (Fig. 3.3C). Specifically, an entire network of epigenetic regulators with known functions in pluripotency and development was uniquely identified in p.iPSCs (Fig. 3.3B, Fig. 3.4). An important constituent of this network are the polycomb repressive complex 2 (PRC2) catalytic subunit Ezh2, its non-catalytic partner Eed, and the PRC2 recruiter Jarid2 (Fig.3.4). PRCs play a crucial role in embryonic development; in pluripotent stem cells, the promoters of many genes involved in lineage specialization are enriched for PRCs and contain bivalent chromatin marks (i.e. involved both in transcriptional activation and repression; H3K4me3, H3K27me3) (Landeira et al. 2010). This is part of a process referred to as "transcriptional priming", which ensures that no unscheduled differentiation takes place and as the embryo development progresses, the lineage-specific gene expression is initiated correctly (Landeira et al. 2010). In the primary iPSCs, we also uniquely identified the H3K9 euchromatic methyltransferase Ehmt1, which is known to act together with PRC2 to promote gene-specific silencing during early pre-implantation development (Fig.3.4)(Maier et al. 2015; Simon & Kingston 2013).

Another protein with a key epigenetic regulatory function which we uniquely identified in primary iPSCs is Chaf1a, the major subunit of the chromatin assembly factor 1 (CAF-1) (Fig.3.3B, Fig.3.4). CAF-1 is critically involved in heterochromatin organization during early embryonic development and the depletion of its major subunit leads to developmental arrest at the 16-cell stage and disrupted heterochromatin organization, resembling a 2- to 4-cell stage (Houlard et al. 2006). In embryonic stem cells, Chaf1a depletion also leads to severe disruption (mislocalization, loss of clustering, de-condensation) of the heterochromatic regions (Houlard et al. 2006).

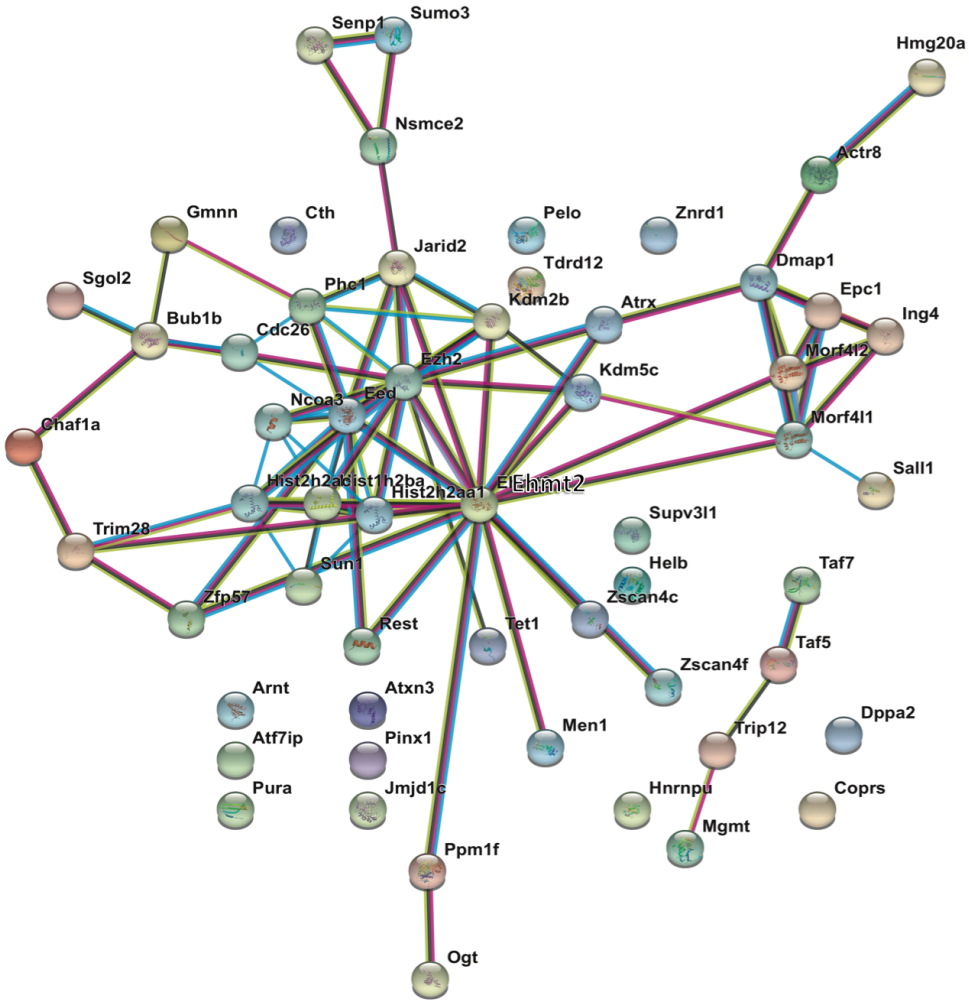


Figure 3.4 Highly interconnected epigenetic network in p.iPSCs. The network included a multitude of epigenetic regulators with key functions in pluripotent cells. Most of the connections in the network are based on experimental data. The color code of the connective edges is as follows: experimental data (pink), curated databases (blue); for predicted interactions, based on gene neighborhoods (green); gene fusions (red); gene co-occurance(blue); light green - textmining (light green); co-expression (black); protein homology (light blue).

The only epigenetic factors with established roles in pluripotency uniquely identified in secondary iPSCs were the histone methyltransferases Setd6 and Setd7 (Fig. 3.3B). Interestingly, they have been shown to fulfill opposing functions in this biological context. Setd6 is involved in embryonic stem cell self-renewal, whereas Setd7 regulates cell differentiation by impacting the silencing of Oct4 and Nanog (Binda et al. 2013; Castaño et al. 2016).

Taken together, our results demonstrate that the primary and secondary iPSCs are highly similar in morphology and expression patterns, but primary iPSCs have a stronger epigenetic and expression signature associated with pluripotency, whereas secondary iPSCs have retained certain neuron-lineage specific traits. This finding suggests that the secondary iPSCs have not completely erased the epigenetic traces of their neuronal past and that this effect translates into a distinct proteomic signature.

3.4 Spatio-temporal proteomic switches during neuronal differentiation of iPSCs

Our findings complement previous research which has showcased the importance of epigenetic memory and control during cell-fate transitions. To elucidate the spatio-temporal component of this regulatory transition, we examined the proteome changes occurring during differentiation of iPSCs to neural progenitors, dissecting the process in time (analyzing different time points) and space (sorting the cells from the Oct4– rim and Oct4+ core of the embryoid bodies) (Fig.3.5). Oct4 is considered as the most stringent molecular marker for pluripotency and is often used to distinguish between pluripotency- and lineage-committed cell populations. In our system, we observe that during neuronal differentiation, the Oct4-GFP expression gradually diminishes, starting from the outer edges of the embryoid bodies and progressing towards the core, suggesting that the differentiation itself follows the same spatial progression (Fig. 3.5B). iPS cells in biological replicates were differentiated until day 8, which represents the neural progenitor stage. Embryoid bodies were collected every 2 days, dissociated and sorted via FACS based on their Oct4-GFP expression (Fig.3.5). Cells from each collected population were subjected to full proteome characterization. To maximize the proteomic identification rate, we also collected unsorted cell populations at each stage, pooled them, isolated and fractionated the proteome and generated a library of MS/MS IDs (details in "Materials and Methods"). Each of the sorted samples was matched against this library, thereby increasing the overall identification rate to 5382 proteins.

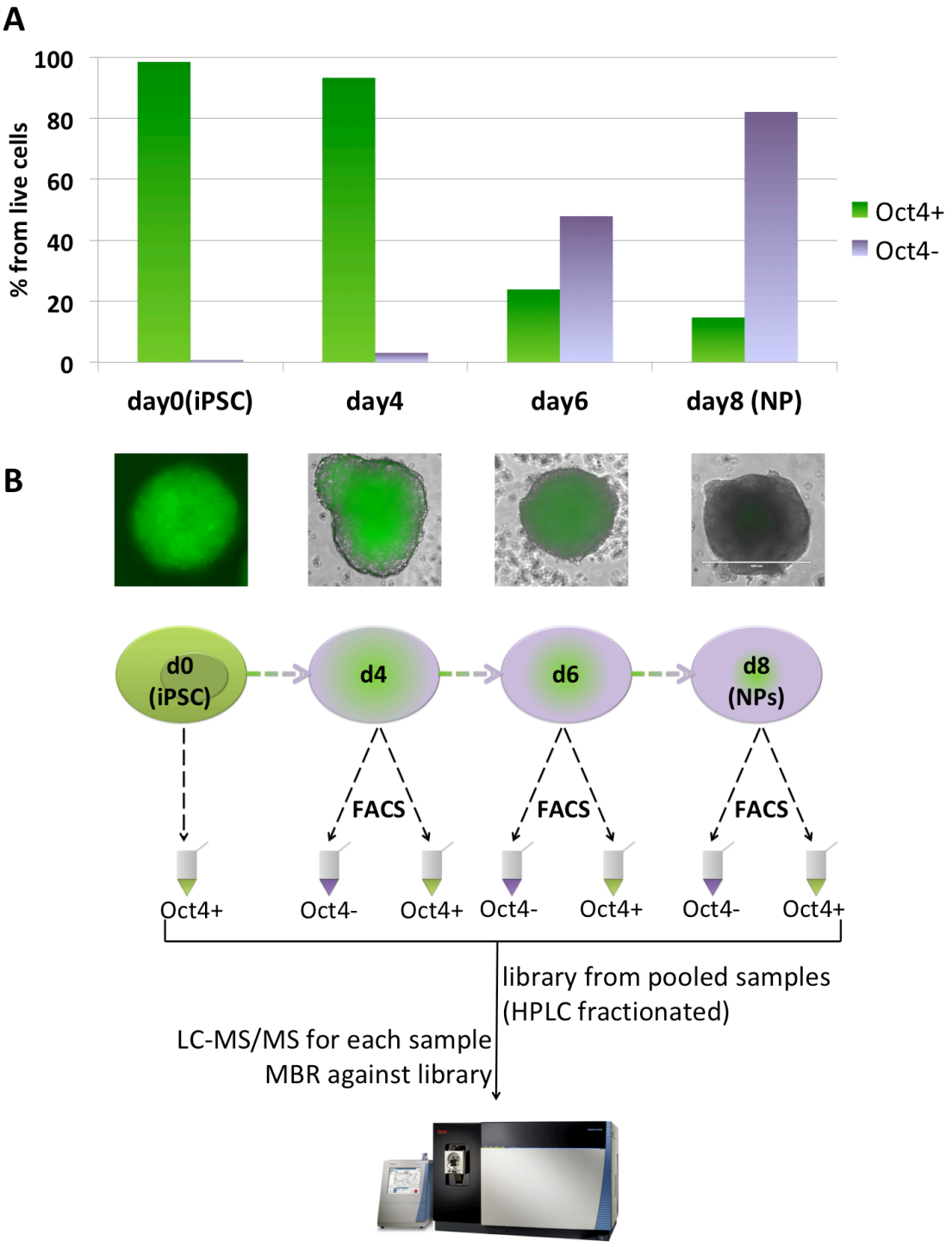
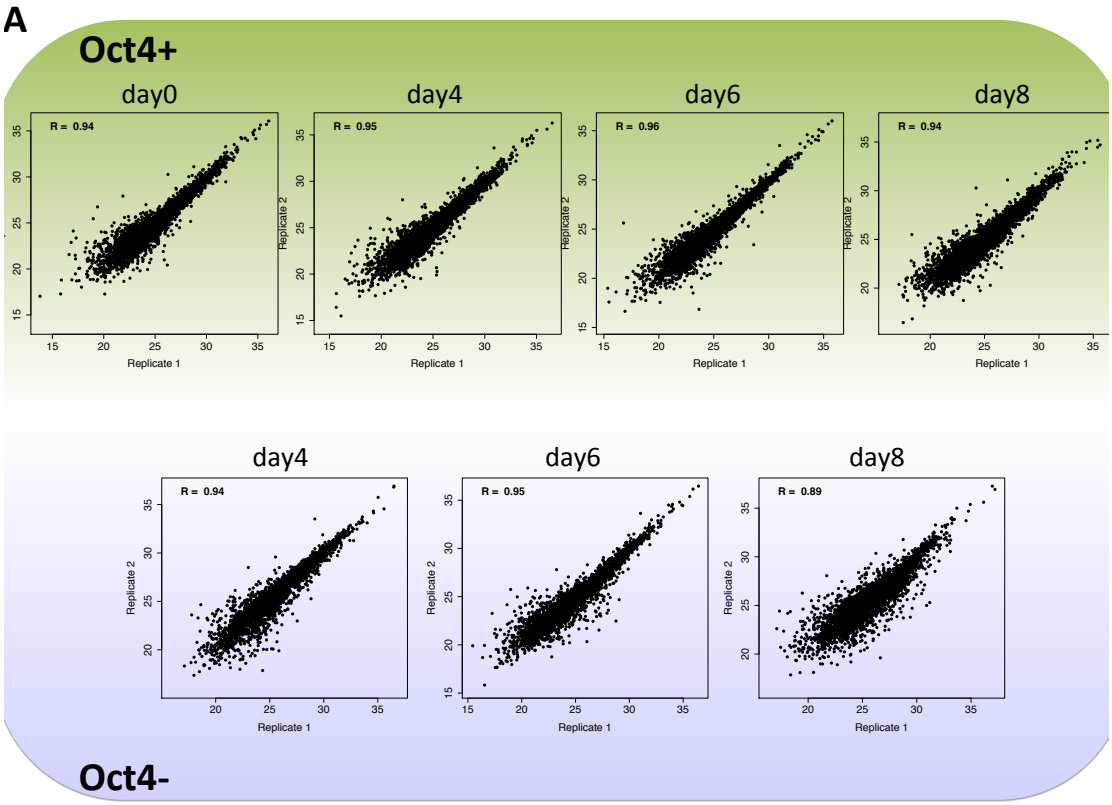


Figure 3.5 Experimental setup for sorting and mass spectrometry analysis of embryoid bodies. (A) Percentage of Oct4-GFP positive and negative cells at each differentiation time point, measured by flow cytometry. (B) iPSCs (day 0) were differentiated to neural progenitors (day 8) and sorted based on Oct4-GFP expression every two days. Each population was measured via mass spectrometry. A pooled sample from all populations was used to generate an MS/MS spectra library, against which the samples were matched. For details see "Materials and Methods".

We observed very high correlation between the biological replicates in each condition, showcasing the high reproducibility of the experiment (Fig. 3.6A). Our mass spectrometry data confirms the marker-based FACS separation in Oct4– and Oct4+ populations, as this transcription factor was only identified in the Oct4+ populations (Fig. 3.6B). In line with expectation, we found that key pluripotency factors were strongly downregulated in the Oct4– populations and showed moderate or no downregulation in the Oct4+ populations (Fig.3.6B). For example, at day 8 the pluripotency-associated protein Esrrb was downregulated nearly 5 fold in the Oct4– and only 1.3 fold in the Oct4+ population (Fig.3.6B). Like Oct4, Nanog was exclusively identified in the Oct4+ populations.

In contrast, proteins associated with neuronal lineage conversion were strongly upregulated in the Oct4– and downregulated in the Oct4+ populations (Fig.3.6B). At day 8, the neural progenitor marker Nestin was upregulated 8 fold in the Oct4– population and downregulated nearly 5 fold in the Oct4+ population (Fig. 3.6B).



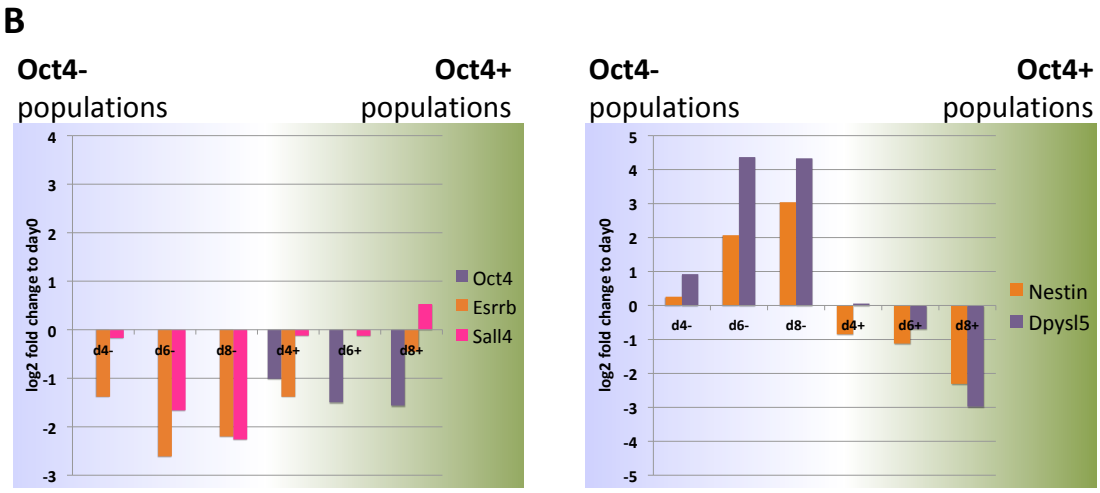


Figure 3.6 Proteome analysis of sorted populations. (A) Scatter plots of the log10 protein intensities between the biological replicates of each population reveals high experimental reproducibility (R≥0.9). (B) Key pluripotency factors (left) and neuronal factors (right) log2 expression changes compared to day 0 (iPSC) in Oct4– and Oct4+ populations.

These observations confirm the notion that the process of neuronal differentiation has a spatial component and progresses from the Oct4– outer edges of the embryoid bodies towards their Oct4+ core. To gain a global overview of the

dynamic proteome changes occurring in the Oct4+ and Oct4– populations during differentiation, we first subjected them to differential expression analysis using a limma test (for details see "Materials and Methods"). As expected, we found that compared to the initial iPSCs, many more proteins display a significant expression change in the Oct4– populations (1408) than in the Oct4+ populations (254) (p.adjusted<0.1; Fig.3.7A). Interestingly, we observed that while the average magnitude of these expression changes increases over time (3.4 fold at day 4 versus nearly 16 fold at day 8 in the Oct4– populations), there is a much higher number of proteins changing in the initial than in the late differentiation stages, implying that many proteins change only transiently (Fig.3.7A). This might be an indication that the initial differentiation signals triggered by the withdrawal of LIF and the formation of 3D cultures cause a multitude of weak and transient expression changes, and upon addition of retinoic acid at day4, the expression differences are channeled and amplified more specifically towards neuronal lineage conversion.

To examine the biological processes in which significantly changing proteins are involved, we performed GO enrichment analysis for each of the sorted cell populations (all time points compared to the initial iPSC stage) (Fig. 3.7B). Expectedly, in the Oct4– populations, we found that terms related to differentiation and neuronal development were enriched at each time point, including retinoic acid metabolism, neuronal differentiation, head development and lipid metabolic processes (Fig.3.7B). Also expected was the downregulation of RNA processing, mitotic cell division and early embryonic development in the Oct4– populations (Fig.3.7B).

Given all these results, it was very surprising to find that in the earliest differentiation stage (day4), in the Oct4+ population, key pluripotency-associated factors and epigenetic remodelers were significantly downregulated (p.adjusted<0.1). This included the pluripotency factors Dppa4 and Esrrb, which were downregulated over 3 fold as well as the epigenetic regulators Jarid2, Zfp57 and the major unit of the CAF-1 complex, Chaf1a, whose roles in pluripotency hav been discussed in detail above (Fig.3.7B). The positive c-Myc regulator Rnf17 was

also downregulated in this population, possibly reflecting slowed proliferation rates. Interestingly, the proteins upregulated in the Oct4+ populations at each time point were enriched for neurogenesis-related processes, including ceramide and sphingolipid metabolism, microtubule organisation and retinoic acid metabolism (Fig.3.7B). Among the upregulated proteins are Camkd2, which is associated with neural projection (Uboha et al. 2007) and Myosin1b, which has been shown to play a critical role in axon formation (Iuliano et al. 2018).

Taken together, our results suggest that contrary to expectation, the proteomic and developmental switches involved in the pluripotency-to-neural lineage transformation become activated very early during differentiation and do not exclusively affect the Oct4– rim of the embryoid bodies, but also their core, which is more protected from the cell culture environment and continues to express Oct4 and other key pluripotency genes like Nanog and Sall4. This finding suggests effective cross-communication between the signal-receiving outer layer of the embryoid bodies and their isolated core and sheds a new light on the spatio-temporal progression of pluripotency loss and neuronal differentiation.

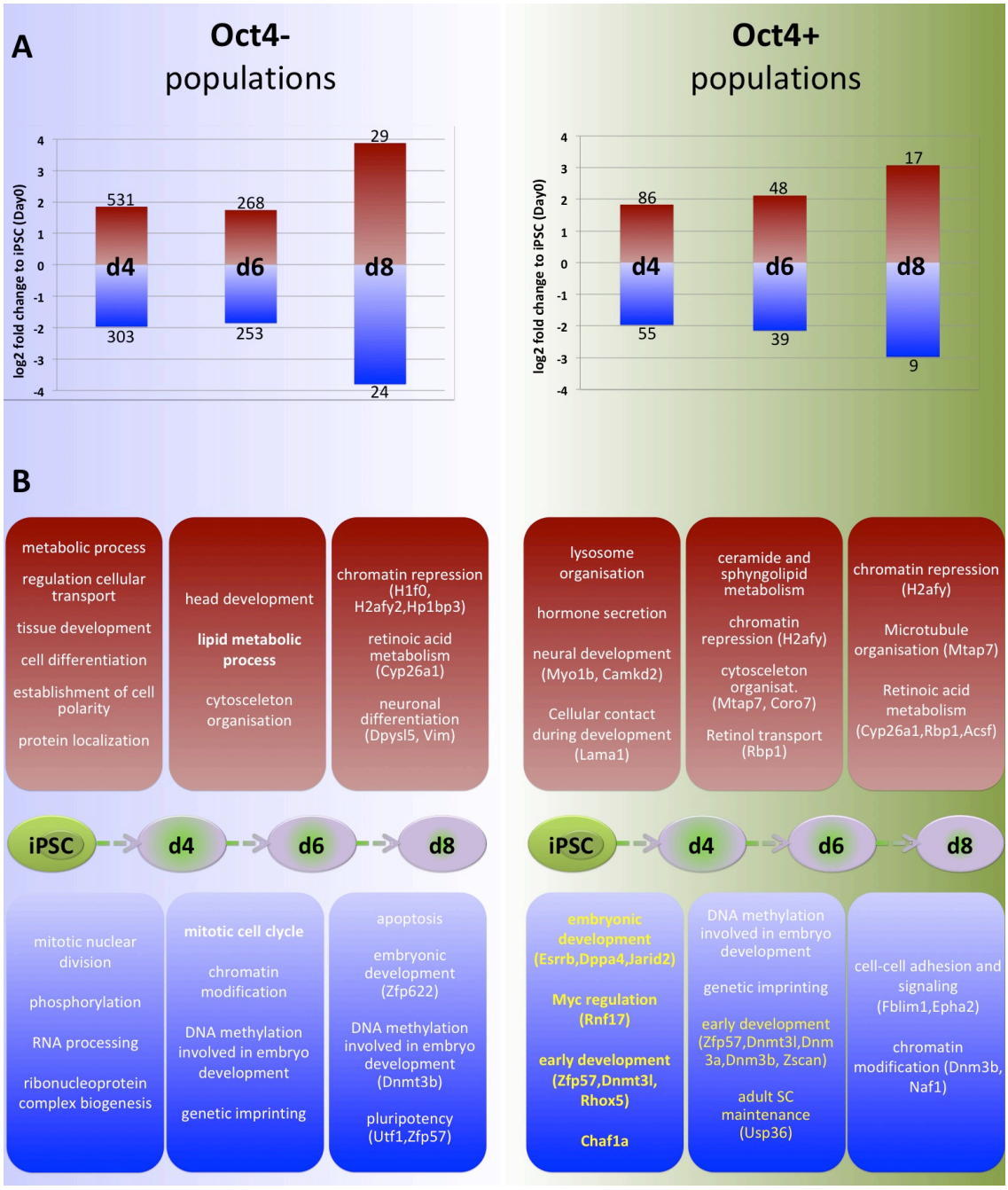


Figure 3.7 Differential expression and GO term enrichment analysis. (A) Differentially expressed proteins between each time point and day0, as determined by a limma test (p.adjusted<0.1). The differential expression is displayed in log2. Upregulated proteins in red, downregulated in blue. Numbers represent the number of up-/downregulated proteins at each time point. Oct4– populations on the left, Oct4+ on the right. (B) GO term enrichment analysis and notable proteins for each protein group in (A).

3.5 Discussion

The presented study explores the capacities and proteomic manifestation of epigenetic memory in the context of cell differentiation and reprogramming. Using an identical mouse model like the one we generated, previous studies have shown that driving primary neurons back to a pluripotent state using exclusively OKSM overexpression is impossible (Kim et al. 2011; Hiler et al. 2016). Our data suggests that unlike primary neurons, a neuronal culture generated *in vitro* from pluripotent cells retains sufficient epigenetic information to enable their reprogramming back to pluripotent cells using only OKSM. It should be noted that there are differences between these two types of neuronal cultures. Most importantly, primary neuronal cultures are free of other cell types, whereas *in vitro* neuronal differentiation is never 100% pure. That means that there is a small minority of cells present in the culture which are not mature post-mitotic neurons (Suppl. Fig.6). It cannot be excluded that they have also contributed to the generation of secondary iPSC. We made an attempt to use flow cytometry to sort viable *bona fide* neurons with the help of a fluorescent label for the neuronal marker NeuO. However, the vast majority of the neurons did not survive the procedure and from those who did, only a few single cells managed to reconstruct the morphological features they lost during the flow cytometry sort (data not shown). Therefore we performed the experiment with a bulk neuronal culture, which we determined to be completely free of the pluripotency marker Oct4 and was 98.2% NeuO+ and 86% Map2+. Thus, despite the presence of a few cells which are not strictly mature neurons, our neuronal culture was of very high purity and contained no visible traces of pluripotent cells.

Comparing the primary iPSC populations to the secondary neuron-derived iPSC, we found that while they are extremely similar in terms of both morphological characteristics and proteome expression profiles, each population had a distinct signature. Primary iPSCs expressed more pluripotency-associated proteins compared to secondary iPSCs, as well as an entire network of epigenetic remodelers with key functions in pluripotent cells. Secondary iPSCs on the other hand retained expression of several key neural-lineage specific proteins. We

attribute this "proteomic memory" effect to the phenomenon of epigenetic memory. After generating the initial iPSCs from embryonic fibroblasts, we kept them in culture for more than two weeks before initiating neuronal differentiation. The motivation was specifically to erase any residual fibroblast-specific epigenetic signatures, which is an effect that long-time culture is known to elicit (Polo et al. 2010). In contrast, secondary iPSCs were collected for mass spectrometry analysis as soon as they had formed. We expect that if subjected to long-time culture, these differences will diminish over time, which is something we plan to test in future experiments.

Studying the proteome changes which occur during neuronal differentiation of iPSCs by separating the Oct4– rims from the Oct4+ core of the embryoid bodies opened a new perspective into the spatial component of the molecular switches controlling the *in vitro* differentiation process. Several findings supported the assumption that differentiation starts at the embryoid body rim and progresses towards the core, which retains pluripotent features for many days after initiation of differentiation. First and foremost, the core continues to express Oct4, as well as other pluripotency proteins like Nanog, Sall4 and Esrrb. In the Oct4– rim, these factors are either completely absent or strongly downregulated. In contrast, neuronal proteins such as Nestin and Dpysl are strongly upregulated in the Oct4– rim. Additionally, compared to the initial iPSC stage, many more proteins change in expression in the Oct4– rim than in the Oct4+ core. Finally, it should be noted that in the cell culture system we use, the differentiation-promoting signals are delivered in the external fluid rather than in the internal structure (which is a major difference to *in vivo* embryonic development). The limitations in diffusive transport are increased by the fact that embryoid bodies tend to form shells of collagen (Sachlos & Auguste 2008). Based on all these observations, we assumed that the inside of the embryoid bodies will retain a pluripotency-like state up until the moment when Oct4 expression is completely absent. We were surprised to find that instead, key pluripotency-related factors and epigenetic regulators become significantly downregulated in the Oct4+ cells immediately upon initiation of differentiation and this effect becomes more pronounced over time (Fig.3.7). This suggests that even though only the outside of the embryoid body is exposed to

differentiation-promoting signals from the cell culture environment, the cross-communication within the 3D aggregates is very fast and efficiently transmits signals which promote the exit from the pluripotent state. Moreover, these signals preferentially elicit a neuronal-lineage entry, as in the Oct4+ embryoid body core we primarily observe significant upregulation of proteins involved in nervous system development like Myosin1b and Camkd2, as well as enrichment of related processes such as retinoic acid metabolism, retinol transport, ceramide and sphyngolipid metabolism. Our findings shed a new light on the spatio-temporal resolution of neuronal differentiation and showcase the speed and efficiency of differentiation-promoting signal transmission within the three-dimentional embryoid body structures.

4. Targeted isolation of proteins associated with the *c-Myc*- promoter

4.1 Introduction

Understanding the molecular mechanisms driving differentiation and reprogramming is to a large extent dependent on interrogating the transcriptional regulatory networks in pluripotent cells. The field relies on various approaches which aim to identify novel regulators within this transcriptional network. For example, loss-of-function screens based on RNA interference (RNAi) have led to the discovery of important pluripotency factors such as Esrrb, Tbx3 and Tcl1 (Ivanova et al. 2006). Protein-DNA interaction studies have also greatly contributed to our understanding of transcriptional regulation in pluripotency. The most prominent technique to study this is chromatin immunoprecipitation combined with sequencing (ChIP-seq). In ChIP-seq, a protein is immuno-precipitated from sheared chromatin using a high-quality antibody and subsequently the co-isolated fragments of DNA are used to determine the genome-wide binding sites of the protein of interest. This method was used to determine the binding sites of the core pluripotency factors Oct4, Sox2 and Nanog (and later many others) in embryonic stem cells (ESCs) and led to the fundamental discovery that they bind cooperatively at numerous active and silent target sites (Loh et al. 2006; Boyer et al. 2005). These studies shaped the current understanding that Oct4, Sox2 and Nanog establish and maintain a pluripotent state by simultaneously activating pluripotency-associated genes (including their own and each other's expression) and repressing lineage-specific ones (Yeo & Ng 2013).

Given that genomic regulation is to a large extent controlled by association of specific proteins to DNA, it is not surprising that ChIP-seq has become the core method in studying transcriptional regulation in pluripotent cells. It addresses the fundamental question which genes are regulated by a particular protein of interest. However, the vast majority of transcriptional regulators do not act alone, but as protein complexes, meaning that a lot of regulatory interactions remain

undescribed. In order to construct a truly comprehensive transcriptional regulatory network, one needs to interrogate the DNA-protein interactions not only from the protein perspective (which genomic sites are bound by a given protein), but also from the DNA perspective (which proteins are bound on a given genomic site).

The targeted isolation of specific genomic segments and subsequent identification of the proteins associated with them has been the subject of a large body of research (Wierer & Mann 2016). Notably, Kingston's lab developed a strategy which they named "proteomics of isolated chromatin segments" (PICh), which relies on DNA probes to retrieve specific genomic loci and identify the proteins associated with them via mass spectrometry (Déjardin & Kingston 2009). The authors were successful in characterizing the telomere-bound proteome, owing to the large abundance of these genomic segments. However, they report that using conventional DNA capture via gene-specific probes commonly used for fluorescent *in situ* hybridization resulted in low yields and contamination with unspecific proteins (Déjardin & Kingston 2009). Their strategy can therefore not be applied for the characterization of the proteome associated with a singular locus, such as a gene promoter.

Another study presented the tagged transcription activator-like - protein A fusion protein (TAL-prA), designed to recognize a specific promoter in yeast, thus enabling the usage of ChIP to obtain the specific promoter site along with the other proteins bound to it (Byrum et al. 2013). While this technique was successfully applied to isolate the particular promoter of interest, it suffers from the fact that it relies on the generation and promoter-association of an artificial protein, which possibly disrupts the natural proteome composition at the site.

Finally, the widespread expansion of technologies based on clustered regularly interspased short palindromic repeats (CRISPR) has also made its way into the proteomics field by using the capacity of guide RNA (gRNA) to target specific genomic loci and using it as a basis to identify the proteome associated with it (Fujita & Fujii 2013; Waldrip et al. 2014; Liu et al. 2017). Tackett and Fujii's labs

used affinity tag coupled to an inactive form of the endonuclease Cas9 to immunoprecipitate the genomic region of interest and subsequently identify the proteins which were co-purified (Fujita & Fujii 2013; Waldrip et al. 2014). Their studies identified histones and some transcriptional regulators as bound to the site of interest, some of which the authors could validate via ChIP-seq, demonstrating the capacity of the method to identify novel proteins binding to a site of interest. Xu's lab used *in vivo* biotyinilation of dactivated Cas9 to capture the chromatin-regulating proteome and long-range DNA interactions at targeted *cis*-regulatory elements (Liu et al. 2017). Although these CRISPR-based methods successfully identified proteins associated at targeted loci, they have several important disadvantages. First, they require transfections, which can elicit immunogenic and cytotoxic reactions and are not always possible (or desirable) in certain cell types. Second, they require the locus-binding of a (fusion) protein, thereby possibly disrupting the natural stochiometry and composition of the locus-associated protein complexes. Third, they suffer from the general concerns regarding off-target effects in the CRISPR-Cas9 system (O'Geen et al. 2015).

Recently, our lab began developing a technique which would enable the unbiased proteome characterization of any genomic locus. The technique, which is still in development, was named "targeted isolation of genomic regions" (TIGR) and aims to circumvent all above-mentioned limitations of the existing methods and fulfill the following criteria:

- not rely on exogenous protein expression;
- not involve any artificial interaction which could disrupt the natural composition of DNA-associated protein complexes;
- allow the targeted isolation of singular loci, such as gene promoters.

TIGR is based on the use of biotynilated DNA probes (~50bp) which are specific to the target region of interest (Fig 4.1). The DNA and protein content of the cells is crosslinked using standard ChIP conditions, the DNA is sheared to a fragment size of 500-1000 bp and semi-denatured to allow the binding of the probes. Streptavidin beads are used to "fish out" the biotninylated probes along with the genomic fragment to which they bind and all proteins associated with it (Fig 4.1A).

A scrambled probe, not complementary to any genomic region, is used as a negative control. To ensure the specificity and enrichment of the targeted locus, the purified DNA fragments are tested against a whole-genome input via qPCR and subjected to next generation sequencing (NGS). To test the reproducibility, the experiment is performed with two probes binding in close proximity of each other, which should therefore pull out the same fragments of sheared DNA, serving as stringent replicates. Finally, the co-isolated proteins are measured via mass spectrometry (Fig 4.1A).

In this thesis, we aimed to isolate the promoter region of *c-Myc* from pluripotent cells and identify the proteins bound to it (Fig. 4.1B). C-Myc plays a crucial role in development, stem cell self-renewal and differentiation (Meyer & Penn 2008), as well as pluripotency establishment and maintenance (Chappell & Dalton 2013). In fact, c-Myc is one of the four "Yamanaka factors" originally used to induce pluripotency in somatic cells (Takahashi & Yamanaka 2006). Approximately 30% of all active genes in embryonic stem cells are bound by both c-Myc and the core pluripotency factors Oct4, Sox2, and Nanog (Rahl et al. 2010) and it is believed that they work in a coordiated manner whereby a complex of Oct4, Sox2 and Nanog recruits RNA Pol II to selected genes and c-Myc controls gene expression by releasing a transcriptional pause (Young 2011). Additionally, c-Myc is a transcription factor critically involved in proliferation, growth and apoptosis and its overexpression is a key event in tumor formation (Levens 2008). It has been shown to be the most important contrubuting factor for tumor development in iPSC-derived mice (Nakagawa et al. 2008). It is thus clear that c-Myc sits at the crossroad of pluripotency and cancer, entangled in a complex regulatory network involved in the regulation of thousands of genes (Chappell & Dalton 2013). Elucidating the mechanisms behind its transcriptional regulation would provide an invaluble insight into these processes and characterizing the proteome composition on its promoter constitutes an essential step in this direction. To this end, we performed TIGR targeting the region upstream from the transcription start site (TSS) of the *c-Myc* promoter in mouse embryonic stem cells (mESCs), using two biotinylated probes: one is complementary to the region 369 bp

upstream of the TSS and the other 424 bp (Fig 4.1B). Hereafter the probes will be referred to as Myc-369 and Myc-424.

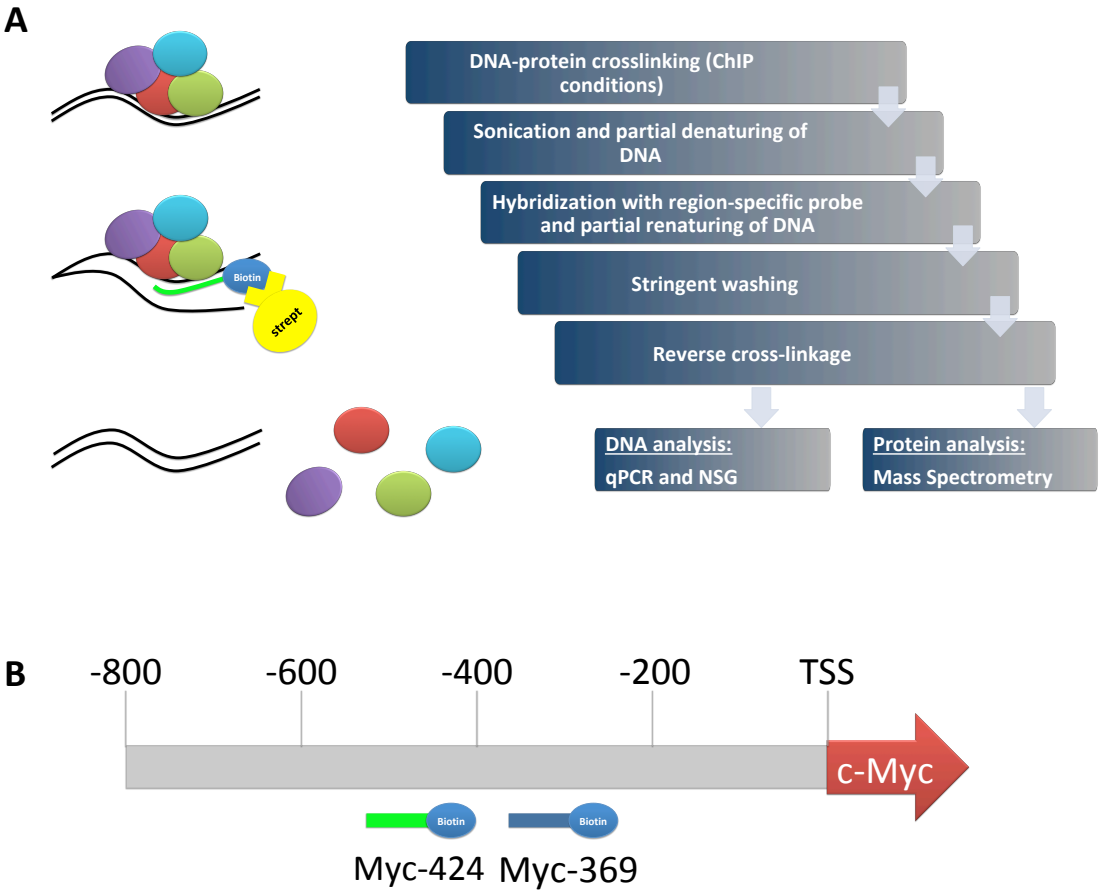


Figure 4.1 TIGR principle and probe design. (A) Summarized TIGR protocol (full details in "Materials and Methods"). (B) Two probes binding to the promoter region of *c-Myc* were designed. They have a length of 50 bases and bind 369 and 242 bases upstream from the TSS. Thus, they are designed to pull down the same DNA fragments.

4.2 Enrichment and specificity of *c-Myc* promoter isolation

In order to estimate the enrichment of the targeted region, we performed qPCR comparing the *c-Myc* promoter enrichment to the negative control (TIGR performed with a scrambled probe) and the whole genome input (Fig. 4.2). As a negative control, we tested for the enrichment of two other gene promoters, *Oct4*

and *Nanog,* as well as a non-coding region on chromosome 2. Our results demonstrate that the targeted region was highly enriched compared to the scrambled control - over 1500 and 3000 fold with each probe (Fig. 4.2A). The targeted region was enriched with high efficiency (7% and 15%, compared to the whole-genome input sample) (Fig. 4.2B). In both cases, the enrichment was significantly higher using the Myc-424 probe. None of the other regions we included in the qPCR experiments as negative controls displayed any enrichment (Fig. 4.2A,B)

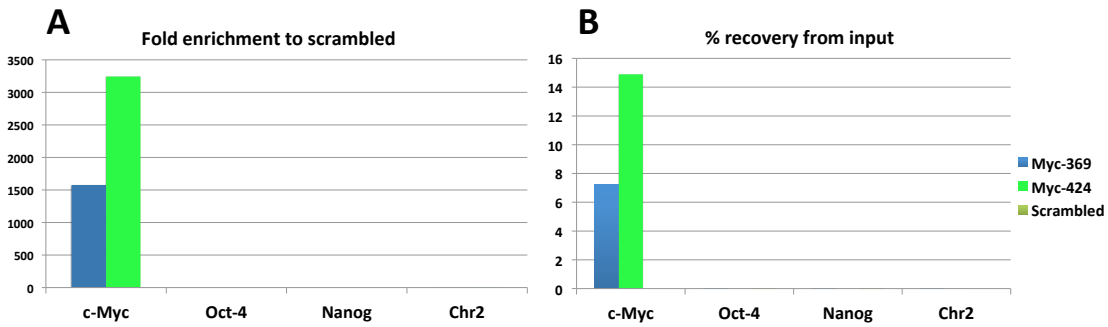


Figure 4.2 Estimation *c-Myc* promoter region enrichment via qPCR. (A) Using the Myc-369 probe, the targeted region was enriched >1500 fold compared to the scrambled control; with the Myc-424 probe over 3000 fold. There is no unspecific enrichment on any of the other sites we included as negative control. (B) Compared to the whole genome input sample, the *c-Myc* promoter was recovered with an efficiency of 7% and 15% using each of the probes.

After having ensured the enrichment of the *c-Myc* promoter, we proceeded to examine the specificity of the genomic isolation. The only comprehensive approach to do this is using next generation sequencing (NGS) and testing whether the targeted locus is the only highly enriched region across the entire genome. As before, we performed the experiment using the Myc-396 and Myc-424 probes, a scrambled probe as a negative control and whole-genome sample as input to compare to. Indeed, our results show a clear enrichment on the promoter region of *c-Myc,* which is not present in the negative control (Fig. 4.3A). Aside from the *c-Myc* promoter, there were several other regions that displayed slight enrichment over the negative control. However, upon comparing them to the target region we find that they are all significantly lower (Fig. 4.3B), indicating that the targeted isolation of the *c-Myc* promoter was highly specific. As expected, the number of

mapped reads was overall very low, possibly owing to the miniscule amount of DNA resulting from a narrowly targeted genomic isolation (Suppl. Fig. 4).

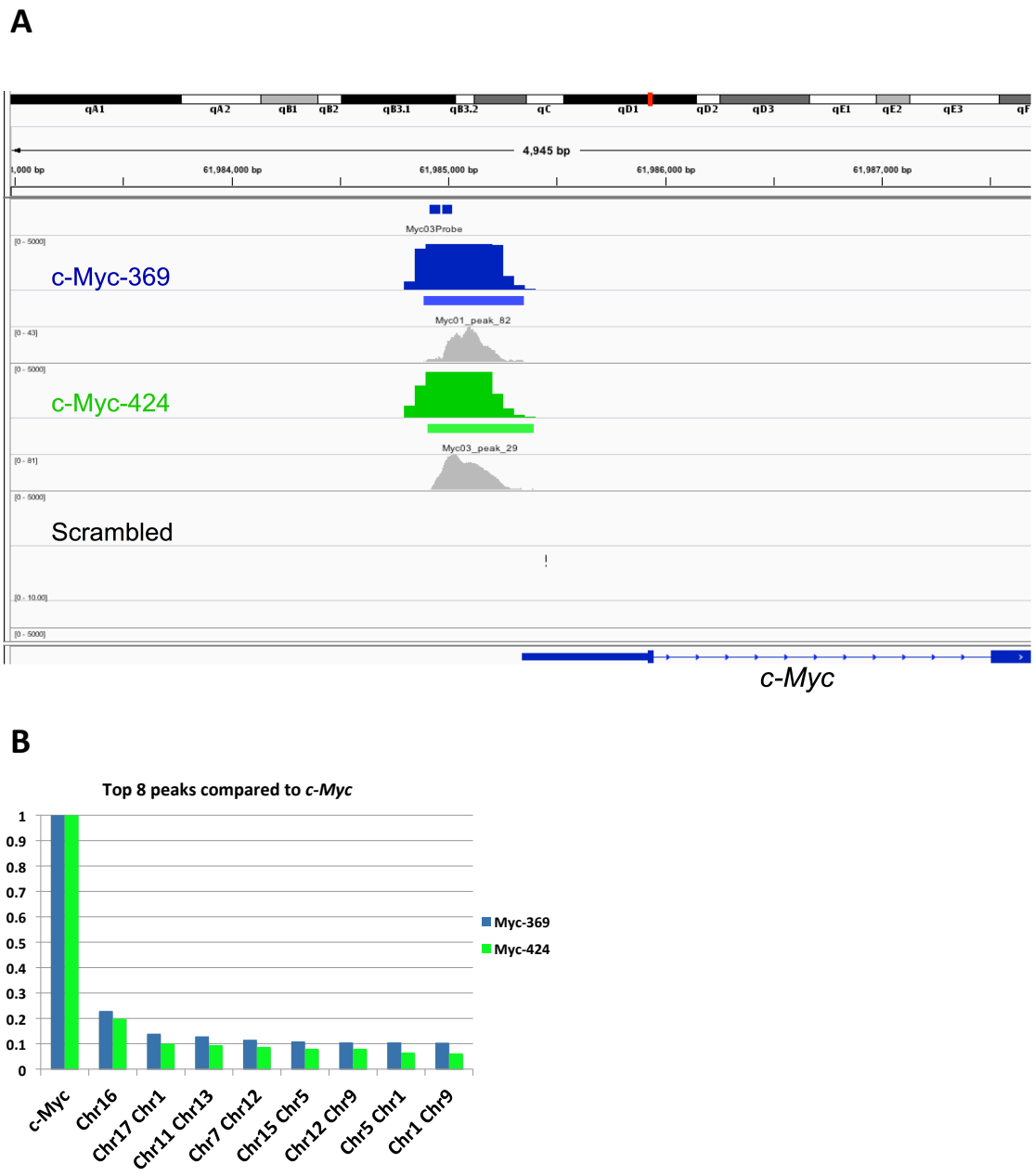


Figure 4.3 Specificity of targeted *c-Myc* promoter isolation using next generation sequencing. (A) We report highly specific enrichment on the *c-Myc* promoter compared to the negative scrambled control. The displayed peaks are the most highly enriched ones across the genome. (B) All genomic regions, which displayed any enrichment over the scrambled control, were compared to the targeted region. The y-axis represents the fold-enrichment compared to *c-Myc*. All regions are designated with their chromosomal position. *C-Myc* displays significantly higher enrichment than all subsequent regions.

4.3 Analysis of the proteome associated with the *c-Myc* promoter

ESCs in native ground state of pluripotency were cultured serum-free under 2i conditions and labeled via "stable isotope labeling with amino acids in cell culture" (SILAC). The cells treated with the negative scrambled control, the Myc-369 and Myc-424 probes were labeled as "light", "medium" and "heavy", respectively (Fig. 4.4). 24 million cells from each condition were collected, subjected to TIGR and mixed in a ratio 1:1:1. Subsequently the samples were run on a mass spectrometer, using LC-MS/MS (Fig. 4.4).

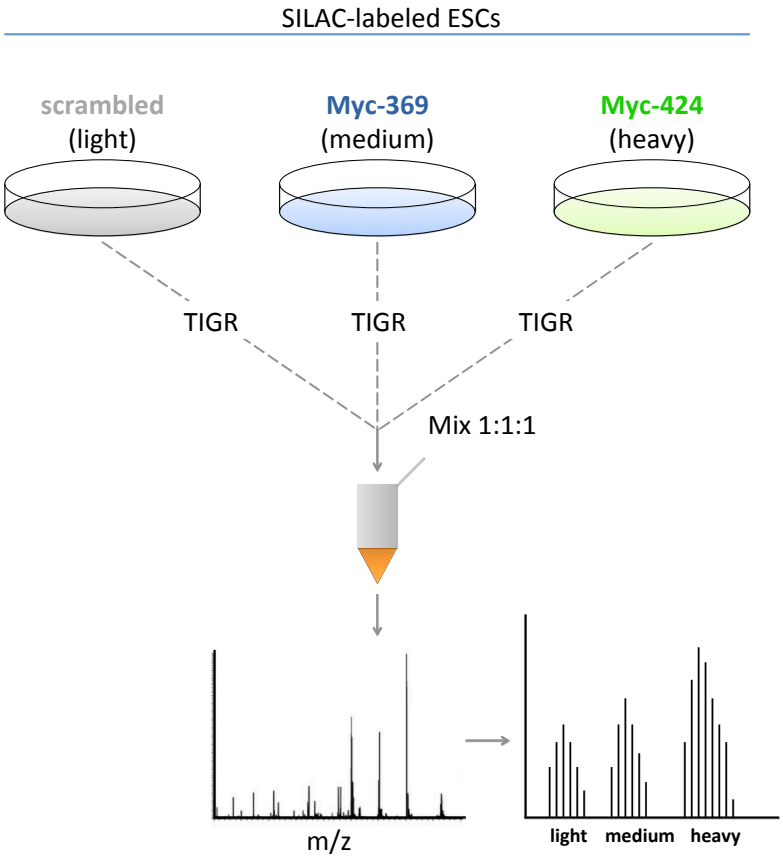


Figure 4.4 Schematic outline of the proteomic experimental setup. 24 million cells cultured in light, medium and heavy SILAC medium were crosslinked with formaldehyde and subjected to TIGR with a scrambled, Myc-369 and Myc-424 probes, respectively. Subsequently the isolated proteins were pooled together and measured via LC-MS/MS. The SILAC labeling enables the distinction between the origins and quantities of each identified protein.

A total of 338 proteins were isolated using Myc-424 and 254 proteins using Myc-369 (Fig. 4.5A). We calculated the ratios of the protein intensities with each probe

compared to the negative control and find that they display a high correlation (R=0.72; Fig. 4.5B).

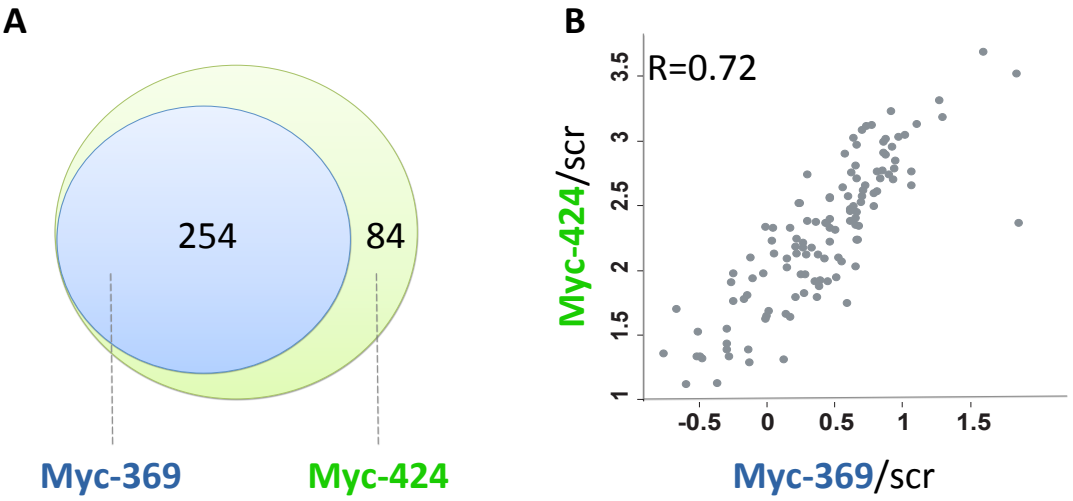


Figure 4.5 Proteome analysis of TIGR results for *c-Myc*. (A) With the Myc-424 probe, we fetched a total of 338 proteins. With Myc-369, 254, all of which were also identified with Myc-424. (B) A scatter plot of the log2 ratios over the negative scrambled control for each of the probes. The results display a high correlation (R=0.72), but an overall lower intensity in the Myc-369 pull-down.

There is a large overlap between the proteins isolated with each probe; in fact, all proteins isolated using Myc-369 were also identified in the Myc-424 sample, indicating high reproducibility. This was further confirmed by the highly correlated enrichment rates compared to the negative control. It should be noted that there is a significant difference in the overall intensities between the proteins isolated with each probe - Myc-369 was significantly less efficient compared to Myc-424, which is in line with our enrichment comparison between the probes on DNA level (Fig. 4.2A,B). This made it difficult to select highly enriched proteins compared to the negative control from both replicates. To ensure the stringency and specificity of all further analysis, we therefore focused only on proteins, which were **exclusively** present in the Myc-pull-downs, but not present in the scrambled control.

This proteomic subset encompasses a total of 90 proteins. It includes the transcription factor Ybx1, which is a known component of a complex promoting c-Myc stability and the co-transcription factor Ewing sarcoma breakpoint region 1

(Ewsr1) which, as a fusion protein with FLI, is a regulator targeting *c-Myc* (Kowalewski et al. 2011).

In order to examine more globally the biological processes in which the *c-Myc*-associated proteins are involved, we subjected them to GO term enrichment analysis, using the whole genome as a background and stringent statistical selection for enrichment significance ($p<0.05$) (Fig. 4.6). In line with expectation, we find that among the most highly enriched terms are processes related to proliferation and cell cycle progression, in which c-Myc is critically involved. Other enriched terms include biogenesis or ribonucleoprotein complexes and RNA localization (Fig. 4.6).

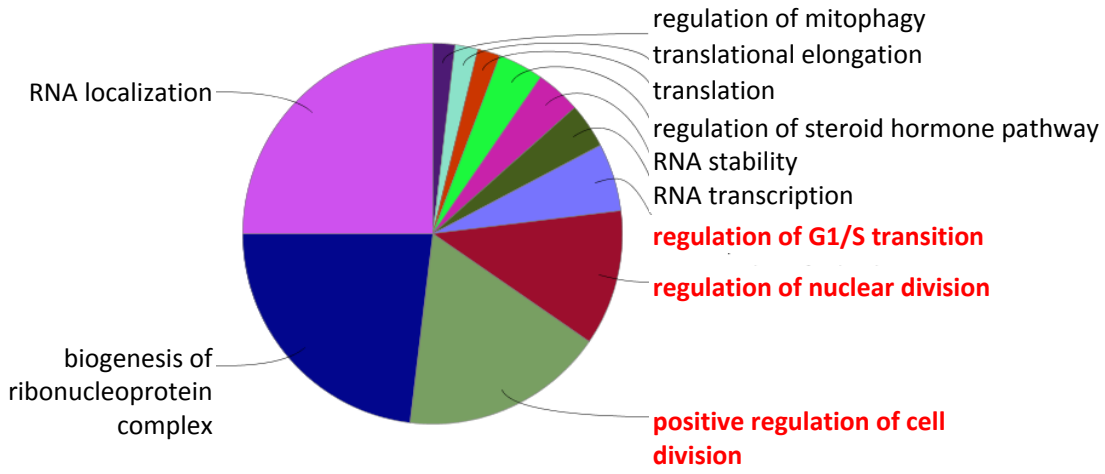


Figure 4.6 GO term enrichment analysis of proteins associated with the *c-Myc* promoter. Only proteins exclusively identified with either of the Myc-specific probes, but not with the negative control, were included. Only significantly enriched pathways were included ($p<0.05$), minimum 3 genes per term. We report high enrichment for processes related to cell cycle transition and proliferation.

In order to examine the connectivity between the proteins specifically associated with the *c-Myc* promoter, we subjected them to network analysis (Fig. 4.7). The connectivity was estimated based on 1) **known interactions** from experimental data and manually curated databases; 2) **predicted interactions** from gene neighborhood, fusion and co-occurrence; 3) text-mining, co-expression and protein homology. We find that that 91% of the proteins isolated with TIGR are highly interconnected and the majority of these connections are based on experimental data (Fig. 4.7). We further find that 38 proteins are known to be

involved in gene expression (highlighted in red), as would be expected for promoter-associated proteins.

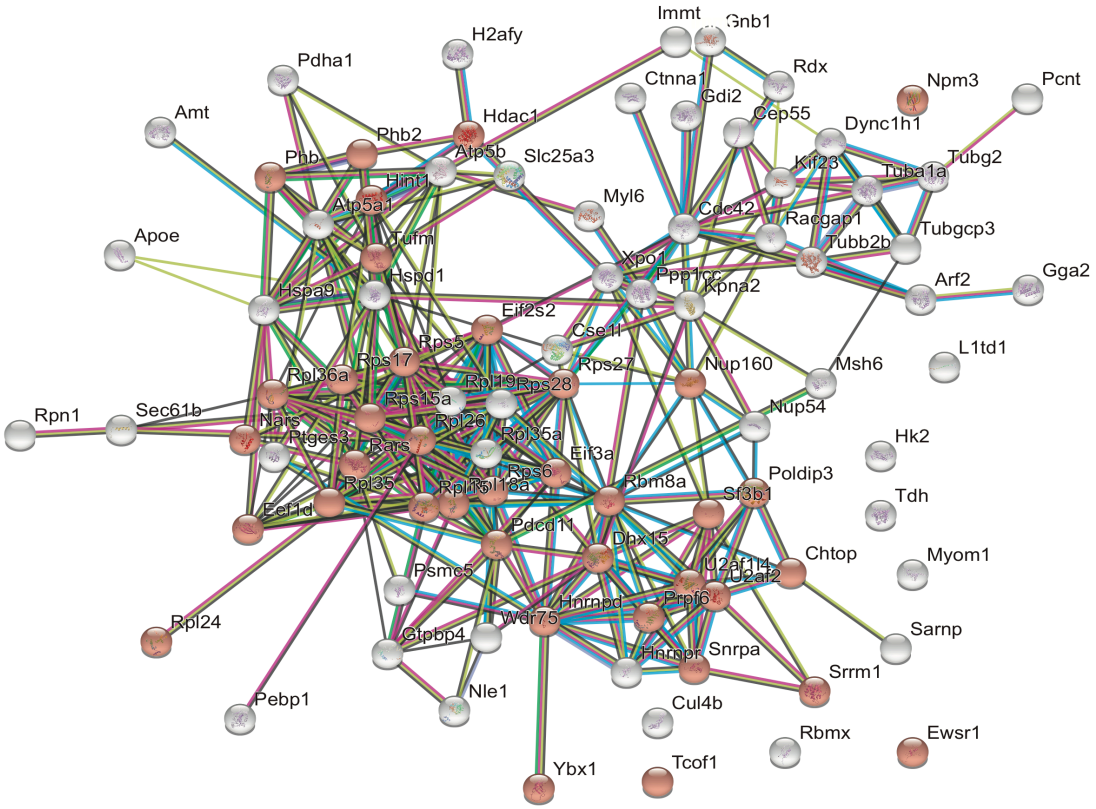


Figure 4.7 Network analysis of proteins associated with the *c-Myc* promoter. Over 90% of the isolated proteins are highly interconnected and the information is primarily based on experimental data (pink edges) or curated databases (blue edges) which represent the highest confidence levels of interactome studies. The rest of the color code refers to predicted interactions, green - based on gene neighborhoods; red - gene fusions; blue - gene co-occurance; light green - textmining; black - co-expression; light blue - protein homology.

Taken together, our data shows that the proteins we identify are enriched in processes in which c-Myc is a key factor and they are highly interconnected, which are indications supporting the hypothesis that they are specifically associated with the *c-Myc* promoter and thus involved with its regulation. This is further confirmed by the presence of known *c-Myc* regulators like Ybx1 and Ewsr1. At the same time, the data contains many interesting novel candidates such as the Chromatin Target of Prmt1 (Chtop), which is critically involved in the tumorigenicity of glioblastoma cells (Takai et al. 2014). Chtop is also predicted to play an important role in the ligand-dependent activation of estrogen receptor target genes (UniprotKB, by

similarity). This makes it a particularly interesting candidate for further exploration, since *c-Myc* is a known and important target of estrogen and regulation of *c-Myc* expression is a key step in estrogen-induced proliferation of breast cancer (Dubik et al. 1987; Dubik & Shiu 1988).

4.4 *In-vitro* versus *in-silico* proteomic composition on the *c-Myc* promoter

To make a more general comparison between our results and data from previous experiments, we used the *in silico* ChIP predictive analysis tool, which is a part of the ChIP-Atlas project (Oki & Ohta 2015). The ChIP-Atlas uses virtually all publically available ChIP-seq data from over 69 000 experiments and offers series of tools for integrative analysis. The *in silico* ChIP tool uses this comprehensive database to predict proteins bound to given genomic loci.

The analysis resulted in a total of 273 proteins predicted to localize on *c-Myc* in pluripotent stem cells (Suppl. table 2). From those, 10 were present in our extended dataset, which includes all identified proteins with higher expression compared to the negative control, in at least one replicate.

These overlapping proteins are primarily factors critically involved in epigenetic regulation, such as Smarca5, which is involved in nucleosome remodeling; the histone deacetylase Hdac1; the nucleosome assembly protein Nap1l1; and Ruvbl2, which is part of a histone acetyl-transferase complex. Other overlapping proteins include the pro-apoptotic transcription factor Parp1, the core promoter binding transcription factor Nono and Trim28, which is required to maintain a repressive state on various genes in pluripotent cells.

While some of the proteins we identified as associated to the *c-Myc* promoter have been previously described as such, the majority are novel, underlining the promising potential of TIGR to identify new gene regulators.

4.5 Discussion

In the current study, we aimed to specifically isolate the promoter of *c-Myc* and identify the proteins associated with it in the biological setting of embryonic stem cells, thereby expanding the Myc-centred gene regulatory network involved in pluripotency maintenance. We showed that by using a novel technique developed in our lab, we were able to efficiently (Fig. 4.2) and specifically (Fig. 4.3) isolate the *c-Myc* promoter region along with over 250 proteins associated with it (Fig. 4.5A), many of which specifically (i.e. not identified in the negative control). These proteins displayed enrichment for processes related to the biological activity of c-Myc, such as proliferation and cell-cycle transition (Fig. 4.6) and were highly interconnected in a common protein-association network (Fig. 4.7). Among them were known *c-Myc* regulators such as Ybx1 and Ewsr1, as well as proteins shown to bind to *c-Myc* in previous ChIP-seq experiments, such as Smarca5 and Hdac1. Importantly, our dataset contained interesting novel candidates such as the Chromatin Target of Prmt1, Chtop. Their association with the *c-Myc* promoter will be subjected to validation by ChIP-PCR and their potential functional effect on *c-Myc* expression tested by knockdown experiments.

In our experimental setup, we chose to use two different probes to isolate the *c-Myc* promoter as they represent very stringent replicates. Indeed, the overlap and correlation between the proteins isolated with each probe was very high, pointing to high reproducibility (Fig. 4.5). However, there was a two fold difference in target region enrichment efficiency between the two probes (Fig. 4.2), which ultimately translated to a difference in number and enrichment of proteins (Fig. 4.5). By only focusing on proteins isolated with both probes, we would therefore exclude many interesting candidates which are not present in the negative control (proteins co-isolated with a scrambled probe). We therefore included them in our analysis. Based on the observation we made that higher genomic enrichment efficiency ultimately results in more proteins identified (Fig. 4.2, Fig. 4.5), for future experiments we consider testing the probe efficiency first and then proceeding using only technical replicates of the most efficient probe. Another

option would be to scale the amount of input material proportionally to the efficiency of the probe.

Taken together, our data suggests that we successfully isolated the promoter of *c-Myc* and identified a multitude of novel proteins associated with it, many of which with potential regulatory function. Our data serves as a valuable basis for the expansion of the pluripotency-associated transcriptional regulatory network.

5. Concluding remarks

The expansion of biological knowledge and understanding is crucially dependent on technological advances which allow us to see the invisible. Present-day bioscience heavily relies on progress in engineering, physics, chemistry and informatics and these integrated efforts have drastically increased the speed and efficiency with which biological information is acquired. The proteomics field belongs to the most prominent examples of this. Recent advances in biochemical assays, mass spectrometry engineering and bioinformatic analytical tools have enabled the characterization of the global proteome composition, as well as specific sub-proteomes in various biological contexts. This major progress has tremendous impact on bioresearch, as proteins are the main functional entities in a cell and the characterization of their complex interplay is paramount to understanding how biological processes work.

The presented thesis aimed to expand the current understanding of the regulatory networks involved in differentiation and reprogramming by exploiting these technological advances. Indeed, we used a multitude of techniques, many of which innovative, to unravel different layers of proteomic information in this context. On a global scale, this included an interrogation of the full proteome changes during neuronal differentiation and reprogramming. We further narrowed down the profiling by investigating the spatio-temporal expression patterns of distinct cell populations during differentiation. Finally, we focused on highly specific protein interaction networks which were either associated with a targeted transcription factor (Sox2) or genomic locus (*c-Myc)*. This multi-level interrogation of the dynamic proteome architecture revealed several novel findings, including the interplay between the dynamically changing Sox2 interactome, the subsequent re-definition of its targets and the effect on neuronal protein expression. The dynamic Sox2 interactome reflects its dual functionality in pluripotency maintenance and neuronal differentiation, as it undergoes the same compositional transition. We also demonstrated that, strikingly, within the three dimentional embryoid body structures, the epigenetic and developmental proteome switches towards

differentiation become initiated immediately and within an enclosed group of cells which express many pluripotency markers. Finally, by successfully isolating a single genomic locus, we discovered a highly interconnected and specific group of proteins associated with the promoter of *c-Myc*, many of which novel. Taken together, these and other findings expand the current understanding of the highly dynamic and interconnected molecular mechanisms behind cell-fate transitions and open up new questions and follow-up perspectives.

One example of an interesting future prospect would be the application of TIGR in a different biological setting, particularly cancerous tissues. Given that the (mis)regulation of this proto-oncogene is commonly involved in the emergence of different tumors, investigating the proteins associated with the *c-Myc* promoter in this setting could potentially provide valuable insight regarding the regulatory mechanisms behind its tumorigenicity. Such insight would also benefit the use of iPSC-based therapies in the clinic, as currently the most serious impediment is that c-Myc overexpression in iPSC-derived cells leads to tumor formation.

Another interesting field of exploration would be the signaling transmission within three-dimentional embryoid bodies. Since their core is largely protected from the cell culture environment both by their position and by an external collagen shell, it is likely that the quick proteomic rearrangements we find in their core might be caused by secretion of epigenetic and develomental cues from the cell coat on the outer rim of the embryoid body. Exploring this could provide valuable insight regarding the similarities and differences in signal trasduction between early embryos and the embryoid bodies designed to model them.

In sum, by expanding our knowledge on the regulatory networks within different stem cell systems, this thesis underlines the significance of proteomics-based studies in developmental biology and lays a solid foundation for further exploration.

6. Materials and Methods

6.1 Establishment of 'reprogrammable' mouse line

'Reprogrammable' mouse embryos were obtained, containing three key components: (1) a reprogramming cassette with the four "Yamanaka" genes *Pou5f1*, *Klf4*, *Sox2*, and *Myc*, (OKSM cassette) under the control of the bi-directional tet-responsive element (*tetO*) with CMV minimal enhancer-less promoter; (2) an expression of reverse tetracycline-controlled transactivator protein (rtTA); and (3) IRES-EGFP fusion cassette downstream of the stop codon of the Oct4 (*Pou5f1*) gene. To this end, male mice homozygous for each of the three mutations were first obtained from The Jackson Laboratory:

- Col1a1-OKSM, http://jaxmice.jax.org/strain/011001.html;
- Rosa26-rtTA, http://jaxmice.jax.org/strain/006965.html;
- Oct4-eGFP, http://jaxmice.jax.org/strain/008214.html.

The mutant males were mated with wildtype (WT) females. The resulting embryos were subjected to rederivation by embryo transfer. After birth, the heterozygous population was left to reach reproductive maturity (min 7 weeks for the males and 12 weeks for the females) and mated to homozygousity. Subsequently, double-homozygous (*hom/hom)* mutants for Rosa26-rtTA and Oct4-eGFP were established and mated with *hom* Col1a1-OKSM mice.

The resulting embryos were *het/het/het* for all three mutations. Any cell type obtained from them could be subjected to reprogramming upon addition of doxycycline to the cell culture media.

6.2 Cell culture

6.2.1 Generation of STEMCCA-MEFs

Due to the presence of a stem cell cassette in the 'reprogrammable' mice, the mouse embryonic fibroblasts obtained from them are here referred to as STEMCCA-MEFs.

Pregnant female mice were sacrificed at 13.5 dpc after timed mating. The uterus was placed in a petri dish and each embryo was separated from the placenta and the embryo sac. The embryos were dissected, whereby the head, internal organs and gonads were removed and the remaining tissue was washed with PBS and minced manually in 1 ml of 0.05% Trypsin-EDTA (Thermo Fisher, 25300054) in a laminar cell culture hood. After incubation 5 minutes at 37° C, the trypsin was inactivated using 1 volume of MEF medium (DMEM high glucose (Thermo Fisher, 11965-092) with 10% fetal serum albumin (Thermo Fisher, 10270-106), GlutaMAX (1:100, Thermo Fisher, 35050-061), penicillin-streptomycin (1:100, Thermo Fisher, 15140-122), nonessential amino acids (1:100, Thermo Fisher, 11140-050), 2-mercaptoethanol (10 µM, Sigma, M7522). Cells were centrifuged at 300 g, 5 minutes, taken up in fresh media and plated in T25 filtered flasks (Thermo Fisher, 156367) coated in 0.2% gelatine (Sigma) at a density corresponding to 1 embryo per flask. Flasks were stored under low-oxygen (5%) conditions for 24 hours, after which they had reached >95% confluency. Cells were frozen with 10% DMSO (Sigma) in liquid nitrogen until further use. Upon usage, cells were routinely tested for mycoplasma infection with e-Myco™ *plus* Mycoplasma PCR Detection Kit (Intron Biotechnology).

6.2.2 Culture of ESCs and iPSCs

Self-generated iPSCs and 46C mESCs were grown feeder-free on 0.2% gelatinized plates in media with the following components: DMEM high glucose (Thermo Fisher, 11965-092), 15% fetal serum albumin (Thermo Fisher, 10270-106), GlutaMAX (1:100, Thermo Fisher, 35050-061), penicillin-streptomycin (1:100, Thermo

Fisher, 15140-122), nonessential amino acids (1:100, Thermo Fisher, 11140-050), 2-mercaptoethanol (10 µM, Sigma, M7522) and 20 ng/ml (iPSCs) or 200 ng/ml (46C cells) LIF (EMBL protein expression facility).

Cells grown under 2i conditions were cultured in DMEM/F12 medium (Pierce, 88215; product discontinued), nonessential amino acids (1:100, Thermo Fisher, 11140-050), penicillin-streptomycin (1:100, Thermo Fisher, 15140-122), GlutaMAX (1:100, Thermo Fisher, 35050-061), 2-mercaptoethanol (10 µM, Sigma, M7522), 20 ng/ml LIF (EMBL protein expression facility), 0.5mg/ml of BSA (Sigma, A3059), 1µM of PD0325901 (Reagents Direct, 39-C68), 3µM of CHIR99021 (Reagents Direct, 27-H76), 100 mg/ml Proline (Sigma, P5607). For *light* SILAC labeling, 100 mg/ml of Lysine (L8662) and 100 mg/ml of Arginine (Sigma, A6969) were used; for *medium* SILAC – 100 mg/ml of L-Lysine-2HCl, 4,4,5,5-D4 (Silantes, 211104113) and 100 mg/ml of Arginine 13C6 (Silantes 201204102); for *heavy* SILAC, 100 mg/ml of 13C6,15N2-L-Lysine HCl (Silantes, 211604102) and 100 mg/ml of 13C6,15N4-L-Arginine HCl (Silantes, 201604102).

Cells were routinely tested for mycoplasma infection with e-Myco™ *plus* Mycoplasma PCR Detection Kit (Intron Biotechnology).

Fixation: added to the 2iL-medium. For cell fixation, the cells were harvested by Stempro Accutase (Life technologies, A11105-01), and spun 5min at 200g to remove the medium. Then the cell pellet was resuspended in 1.5% formaldehyde (Pierce, 28906) in PBS. After 15 min incubation at room temperature with occasional rotations, 125mM Glycine (Merck, 56-40-6) was added to the solution to quench cross-linking. Then the cells were washed twice with PBS, counted and aliquoted, and stored at -80°C as dry cell pellets.

6.2.3 Cellular reprogramming and establishment of STEMCCA-iPS cell lines

MEFs were thawed upon placing in 37°C water bath for 5 minutes and taken up in 9 ml MEF medium (composition see above), centrifuged at 1200 rpm for 4 minutes, taken up in fresh MEF medium and expanded in T25 flasks (Thermo Fisher, 156367) in low oxygen conditions (5% O2) for 24 hours until full confluency.

The cells were split in gelatinized (0.2%) 15 cm cell culture plates in ES media containing 2 µg/ml doxycycline and the AGi (ascorbic acid (Sigma,) and GSK3b inhibitor CHIR99021 (Reagents Direct, 27-H76) small molecules combination as described by (Bar-Nur et al. 2014). The medium was changed every day and the cells were kept on doxycycline for 8 days and cultured for additional 4 days without doxycycline to ensure that the endogenous pluripotency network had been initiated. The cells were routinely monitored for their Oct4-GFP expression under a fluorescent microscope. At day 12, typical iPSCs with clear, shiny edges and pronounced GFP expression had formed. A subset of the colonies were tested and were positive for Alkaline Phosphatase (AP staining Kit, Vector Labs, SK-5100).

To generate clean, iPS-only lines, 50 colonies were cut out surgically under a light microscope (using a scalpel and a syringe needle), collected in a well of a round-bottom 96 well plate (Thermo Fisher) containing 10 µl trypsin and dissociated for 20 minutes at room temperature. The dissociated cells from each colony were subsequently plated in one well from a gelatinized (0.2%) 96 well plate. Over the following days, the cells were monitored for colony formation and only the ones which managed to re-build distinct, GFP-positive iPS colonies were expanded and kept as iPS cell lines.

6.2.4 Neuronal differentiation of pluripotent stem cells

The differentiation protocol was performed as described in (Bibel et al. 2007), with few modifications. Briefly, self-generated iPSCs and 129X1/SvJ ESCs were thawed on feeder-MEFs (for 129 cells) or on gelatinized plates (for iPSCs) and kept on for two passages in ES medium (see composition above) with 15% ES Cell Qualified EmbryoMax® FBS (Millipore, ES-009-B) and 20 ng/ml LIF (EMBL protein expression facility). The 129 cells were then plated for 2 passages on gelatinized plates. Differentiation starts upon transfer of 4e6 cells/10 cm non-adhesive plates (Sigma, P9366 Sigma) and removal of LIF from the medium, leading to the formation of embryoid bodies. The medium was changed every two days. On days four and six, retinoic acid (Sigma, R2625) at a final concentration of 5 μM was added to the medium. On day 8, the emryoid bodies were dissociated, brought in single-cell suspension, plated on PORN/laminin-coated plates and switched to N2 medium containing DMEM high glucose (Thermo Fisher, 11965-092), 1xN2 supplement (Thermo Fisher, 17502048), 1x B27 supplement (Thermo Fisher, 17504044) and penicillin-streptomycin (1:100, Thermo Fisher, 15140-122), following the exact protocol as described by Bibel (Bibel et al. 2007). At day 10 the cells had formed dense neuronal networks and represented a differentiated neuron state.

Cells were collected, centrifuged at 2000 rpm for 5 minutes and the cell pelletts were frozen at -80°C for all proteomic/biochemical experiments.

6.2.5 Reprogramming of neuronal cultures

The STEMCCA-iPSCs were subjected to neuronal differentiation described above. The differentiated neurons (day 10) were kept in N2 medium culture for additional two days (day 12). To ensure that there were no pluripotent cells left in the neuronal culture, the cells were observed under fluorescent microscope (Evos). No Oct4-GFP-positive cells were observed.

At day 12, the reprogramming was initiated upon addition of 2 µg/ml of doxycycline to the cell culture and simultaneous switch to ES media (see composition above). The entire reprogramming of the neuronal culture took 32 days after this initiation step. Doxycycline was added to the cells for a total of 14 days. The cells were kept in their PORN/laminin coated plates for 4 days after initiation of reprogramming. At day 4, the cells were split 1:3 and plated on gelatin-coated plates. The media was changed without splitting on days 2, 6, 8, 11 and 25; they were split on days 4, 9, 12, 14, 16, 19, 23, 27. To estimate the progress of reprogramming, the cells were analyzed on a cell sorter (MoFlo) on days 7, 12 and 21. On day 32, Oct4-GFP-positive colonies with characteristic iPSC morphology and size had formed.

6.3 Proteome sample preparation

6.3.1 SP3 sample preparation

The following procedure termed "Single-Pot Solid-Phase-enhanced Sample Preparation" (SP3) was adapted from (Hughes et al. 2014). It was applied for the whole proteome samples from chapter 2 and the Oct4-GFP sorted populations from chapter 3.

Cell pellets of 1e6 cells from each condition and replicate were reconstituted in 100 µl of lysis buffer (50mM Ambic, 1% SDS, 1X Protease Inhibitor Cocktail (Roche; 05892791001), 10mM TCEP and 40mM CAA) and sonicated for 12 cycles (30/30 seconds on/off) on a Bioruptor Pico (Diagenode). They were heated up at 95°C for 5 min and cooled down for 5 min at room temperature.

A 50/50% mixture of Sera-Mag Speed Beads A and B (Fisher Scientific; CAT No. 24152105050250, CAT No. 44152105050250) was rinsed in water on a magnetic stand 2 times and taken up water, in their original volume. 4µl of the bead mixture was added to the samples and immediately afterwards, 104 µl of acetonitrile (ACN) were added. The samples were left for 10 minutes at room temperature, after which they were placed on a magnetic rack and left for another 2 minutes to allow the magnetic beads to settle. The supernatant was removed, the beads were washed on the rack 2 times with 1 ml 70% ethanol and once with 1 ml ACN. The

supernatant was removed, the beads were air-dried for 1 minute and taken up in 20 µl TEAB with pH 8.5. For the whole proteome samples in chapter 2, 2 µg of the proteolytic enzyme LysC were added to the samples; for the sorted populations in chapter 3, we used Trypsin, the amount was adapted to the measured sample amount as follows sample:trypsin = 1:50. The samples were left for 16 hours at 37°C.

The following day, samples in chapter 2 were subjected to TMT labeling (see method below). The other samples were subjected to SP3 peptide clean-up as follows: 2 µl from the prepared seramagnetic bead mixture (see above) was added to each sample and vortexed. ACN was added to a final concentration of 95% and the sample incubated for 18 minutes at room temperature. The samples were spun briefly with a tabletop centrifuge and the tubes were placed on a magnetic rack. The supernatant was removed, the beads were washed with 200 µl 100% ACN once. The beads were air-dried for 1 minute. 10 µl of 0.1% formic acid was added to the dried beads, the tubes were vortexed and placed in a sonication water bath for 5 minutes. The tubes were placed on a magnetic rack, the liquid was removed and placed in fresh tubes. These samples were then run on an Orbitrap Fusion mass spectrometer (see settings below) without further clean-up.

6.3.2 In-solution sample preparation

1.5e6 cells from each condition were lysed in 300 µl lysis buffer containing 0.1% Rapigest (Waters), 50 mM Ammonium Bicarbonate pH 8.0, 10 mM Tris(2-carboxyethyl)phosphine hydrochloride (TCEP) (Sigma-Aldrich, C4706) and 40 mM CAA. The lysate was sonicated with a Benson probe sonicator for 60 seconds at 10 seconds on, 2 seconds off cycle and 20% output on ice and afterwards subjected to proteolytic digestion with 1:50 trypsin (Promega) for 16 hours at 37°C. The samples were then acidified to 1% end concentration of trifluoroacetic acid (TFA) (Sigma-Aldrich, 76-05-1), incubated for 30 minutes at 37°C and spun down at 15000g for 10 minutes. The supernatant was then subjected to clean-up and mass spectrometry measurement as described in "HPLC fractionation and mass spectrometry analysis" (without fractionation and label-free).

6.3.3 TMT peptide labeling

In order to enable the multiplexed mass spectrometry analysis of the samples, they were subjected to isobaric chemical labeling using the so called tandem mass tag (TMT, Thermo Fisher, 90061). The TMT reagents are used for labeling of the peptides and consist of a reactive group (which covalently binds to the peptide N-termini or lysine residues), a spacer arm and a reporter group. During mass spectrometry analysis, the tag fragments break off and produce unique reporter ions, thus enabling the relative protein quantification between the different samples. We used TMT 6-plex labeling (in which the reporter mass varies from 126 to 131 Da) to label the samples of each stage of differentiation (day 0, 2, 4, 6, 8 and 10).

After the proteolytic digest, the beads were put on a magnetic rack and the supernatant was transferred to new tubes. 20 µg (in 1 µl) of TMT reagent was added to each sample and left for 30 min at room temperature. After that the same amount was added again and left for 30 min at room temperature. 1 µl of quench mix (50mM ammonium bicarbonate and 10mM Lysine) was added to each simple and incubated for 5 minutes. 2 µl of bead mix (see preparation above) was added to each sample and ACN was added up to a final percentage of 95% and incubated for 10 minutes. The beads were put on a magnetic rack for 2 minutes, the supernatant was removed and the beads were washed with 100% ACN. The supernatant was removed and the beads were reconstituted in 4% DMSO in water and placed on magnetic rack. Finally the supernatant was taken in fresh tubes and formic acid was added to a final percentage of 0.1%.

The labeling efficiency was first controlled by measuring 10% of each sample separately on an Orbitrap Fusion mass spectrometer (exact settings below). The multiplexed samples were then mixed in a ratio 1:1 (based on the overall abundance estimation from the label check run), subjected to fractionation via high pH high-performance liquid chromatography (HPLC) and mass spectrometry measurement (full description and settings in "HPLC fractionation and mass spectrometry analysis").

6.3.4 MS/MS library preparation

This method was used for preparation and analysis of the Oct4-GFP sorted populations described in Chapter 3. 1e6 non-sorted cells from each time point and condition were combined and subjected to in-solution digestion as described above. The peptide concentration was measured using Pierce™ Quantitative Colorimetric Peptide Assay (Thermo Fisher) and 50 µg were subjected to HPLC fractionation into 60 fractions, which were pooled into 12, as described below. The equivalent of 1.5 µg from each fraction was measured on an OrbitrapFusion mass spectrometer using the settings described below and a 2 hours LC gradient. The same gradient and LC-MS/MS settings were used for measurment of the sorted populations, which were prepared using the SP3 protocol as described above. The library and sorted population samples were measured in technical replicates, which were combined upon raw data analysis with MaxQant software. The "match between runs" function in MQ was applied to match IDs from the library to the sorted populations.

6.4 HPLC fractionation and mass spectrometry analysis

The combined samples were fractionated with 1200 Infinity HPLC system (Agilent), using a Gemini C18 column (Phenomenex). A 60 minute gradient was used, which progresses linearly from 0 to 35% ACN in 20 mM ammonium formate, pH10. The flow rate was set at 100µl/minute. Peptide elution was detected via UV detector at 254 nm. 33 fractions were collected and pooled into 11 fractions (combination strategy: fraction 1, 12 and 23; 2, 13 and 24 etc.).

Each fraction was then measured on an Orbitrap-Fusion Quadrupole-Linear-Ion Trap-Orbitrap hybrid mass spectrometer (Thermo Fisher) coupled to EASY-nLC system (Thermo Fisher). The samples were loaded onto a 100 µm x 2 cm Acclaim Pepmap RSLC trap column (5µm particles, 100Å pores, C18) in 100% solvent A (0.1% formic acid in water, ULCMS Grade, Biosolve) and eluted onto a 75 µm x 50 cm (2µm particles, 100Å pores, C18) Acclaim Pepmap RSLC analytical column by a gradient from 3% solvent B (0.1% formic acid in 80% acetonitrile and 19.9% water, ULCMS

Grade, Biosolve) to 50% solvent B in 86 minutes at a flow rate of 300 nl/min. Eluting ions were analyzed by electrospray using a 10 µm Picotip coated fused silica emitter (New Objective) and a Nanospray-Flex ion source (Thermo) connected to an Orbitrap-Fusion Quadrupole-Linear-Ion Trap-Orbitrap hybrid mass spectrometer (Thermo Fisher). The Orbitrap was operated in positive mode generating profile spectra at a resolution of 60.000 FWHM, AGC target was $1x10^6$, maximum injection time 50 ms. The mass spectrometer was set to data-dependent mode of acquisition (top speed) and the most intense ions (threshold $5x10^3$) were selected for HCD-fragmentation using nitrogen as a collision gas (33% HCD collision energy) by the Quadrupole (1.6 m/z window) and resulting fragments were analyzed by the Linear-Ion-Trap set to rapid scan rate, first mass 120 m/z, an AGC Target of $1x10^4$, a maximum injection time of 50 ms and data type to centroid. Selected ions were excluded for reselection 60 (146 min gradient) seconds with a window of 20 ppm.

6.4 FACS

Cell sorts were performed either on a Becton Dickinson FACSMelody or a Beckman Coulter MoFlo XDP cell sorter. Both instruments were set to 20psi sheath pressure with a 100um nozzle running Becton Dickinson FACSFlow as sheath carrier. GFP was excited with 488nm (20mW or 400mW respectively) and emission was picked up with a 530/30 bandpass filter; DAPI live dead staining was done on both systems with all samples to exclude dead or compromised cells (405nm excitation, 450/50nm bandpass filter - 25mW or 240mW respectively). PMT performance and setup was checked either with BD CST beads or Spherotech Rainbow beads on the MoFlo systems. Purity of the cells was frequently assessed post sort. Sorts were performed under chilled sample conditions at 4°C.

6.5 ChIP-SICAP

Cells were cross-linked using 1.5% formaldehyde for 15 minutes and the reaction was stopped using 125mM Glycine. 24e6 fixed cells were used for each condition and replicate. Each pellet was resuspended in 5ml lysis buffer 1 (HEPES-KOH pH7.5 50mM, NaCl 140mM, EDTA 1mM, Glycerol 10%, NP-40 0.5%, Triton X100 0.25%) and rotated on a wheel for 10 minutes at 4°C. The lysate was centrifuged at 1350rpm for 5min at 4°C and the pellet resuspended with 5 ml lysis buffer 2 (Tris HCl pH 8 10mM, NaCl 200mM, EDTA 1mM, EGTA 0.5mM) and rotated on a wheel at room temperature for 10 minutes. The lysate was centrifuged at 1350rpm for 5min at 4°C and the pellet resuspended with 1.5 ml lysis buffer 3 (Tris HCl pH 8 10mM, NaCl 100mM EDTA 1mM, EGTA 0.5mM, Na-deoxycholate 0.1%, Na-lauroylsarcosine 0.5%). The chromatin was then sonicated in a cooled Bioruptor Pico (Diagenode) for 7 and 11 cycles for ESCs and TNs, respecitively, on 30s ON/30s OFF cycles. An aliquot of the cross-linked material was de-crosslinked at 95°C for 1 hour and tested on an 1.5% agarose gel (fragment size 300-500 bp). Triton was added to the chromatin samples (1% final concentration), which were then spun in a cooled centrifuge at 12700rpm, 4°C for 10 minutes. The protein amount in each sample was quantified using a Pierce BCA assay according to the manufacturer's protocol (Thermo Fisher). 12.5µg to 20µg of Sox2 antibody (R&D Systems, AF2018) was added to each sample and the same amount of IgG antibody (Santacruz) was added to the samples which served as negative control. The samples were rotated for 18 hours at 4°C. 100µl/sample Protein G magnetic beads (Invitrogen) were washed twice using 0.5% BSA (Sigma) in 1x PBS and left in the same solution on a wheel overnight at 4°C.

The following day, the beads were washed twice with PBS, added to the chromatin samples and rotated on a wheel for 3 hours at 4°C. The following washing steps were performed by resuspending the beads by flapping the tube several times, quickly spinning them down and and putting them back on the magnet. The beads were washed once using washing buffer A (SDS 0.1%, Triton 0.5%, EDTA 2mM, NaCl 150mM, Tris HCl pH 7.6 20mM), once with washing buffer B (SDS 0.1%, Triton 0.5%, EDTA 2mM, NaCl 300mM, Tris HCl pH 7.6 20mM) and once with

100mM Tris-HCl pH 7.6. The beads were then resuspended in 100 µl 1x TdT buffer (Thermo Fisher) and kept 5 minutes at room temperature, occasionally swirling the beads. The tubes were then placed on a magnetic stand, the liquid removed and the beads resuspended with fresh 88 µl TdT buffer (Thermo Fisher). 10 µl biotin-14-dCTP (Thermo Fisher) and 2 µl TdT enzyme (Thermo Fisher) were added and incubated for 30 minutes at 37°C on a thermomixer (Eppendorf), agitating at 500 rpm. After that, the beads were washed 4 times with 1 ml ice cold IP buffer (see composition above) at room temperature and resuspended in 100 µl mixture of 7.5% SDS and 200mM DTT in water. The beads were then incubated 30 minutes at 37°C on a thermomixer (Eppendorf), agitating at 750 rpm. The supernatant was collected and the beads were discarded. 0.7 ml IP buffer (see composition above) and 50 µl protease-resistent streptavidin beads (New England Biololabs; patent for protease-resistence filed by our lab; appl. Nr. CA2998549A1) were added to each sample and rotated at 4°C overnight.

The suspended beads were put on a magnetic stand, the liquid removed. The beads were then washed 3 times with SDS wash buffer (10 mM Tris-Cl pH 8, 1 mM EDTA, 1% SDS, 200 mM NaCl), 1 time with 10% isopropanol and 4 times with 20% ACN. The beads were resuspended in 80 µl ACN and transferred to 0.2 ml PCR tubes, which were placed on magnetic rack. The supernatant was removed and the beads resuspended in 14 µl digestion buffer (0.1% SDS in 50 mM Ammonium Bicarbonate). 1 µl of 100 mM DTT was added to the samples, which were then incubated at 95°C for 20 minutes to detach the proteins from the streptavidin beads. The tubes were cooled down to room temperature, 1 µl of 200 mM IAA was added to the samples and incubated 30 minutes at room temperature. The reaction was quenched upon addition of 1 µl of 100 mM DTT. The tubes were placed on a magnetic stand and the liquid transferred to fresh tubes (the beads were discarded). 300 ng Trypsin (Promega) was added to each sample and incubated at 37°C overnight. The following day, they were subjected to SP3 clean-up as desicribed above.

6.6 TIGR

For this experiment, we used the following probes (biotynilated) and primers:

Probes:	
Myc-369	[BtnTg]TGACGCGGTCCAGGGTACATGGCGTATTGTGTGGAGCGAGGCAGCTGTT
Myc-424	[BtnTg]TCTAAATTCTGTTTTCCCCAGCCTTAGAGAGACGCCTGGCCGCCCGGGACG
Scrambled	[BtnTg]AGGTGCAGCCGTGGTTAAAAGATGAATAAAGTGAAATGAGGTAAAGCCTCTT
Primers:	
cMyc-F	AACTCATTCGTTCGTCCTTC
cMyc-R	ACAGTAATAGCGCAGCATGAA
Chr2-F	TCTGAGCACCTCATCTCATTG
Chr2-R	TTTGAAAGGACTTGCCCAGAA
Oct4-F	AGAAATAATTGGCACACGAACA
Oct4-R	TCACCGGACACCTCACAAAC
Nanog-F	GAAAGCCGTGTATAAACAGAGAC
Nanog-R	CTTTACCTCATTTCACTTTATTCATC

We used fixed, SILAC-labeled cells (see preparation details in "Cell culture" above). The Myc-424 probe was applied to "heavy" labeled cells, Myc-369 to "medium" and the unspecific "Scrambled" probe to "light". For enrichment tests with qPCR, we used 4e6 cells/condition, for next generation sequencing and mass spectrometry, 24e6 cells/condition.

The cells were taken up and vortexed in 200 µl IP buffer (50mM Tris-Cl, pH 7.5; 5 mM EDTA, 1% TritonX-100) with 1x protease inhibitor (Roche). 5 µl RNaseA (10 mg/ml, Fermentas) was added and incubated at 37°C for 20 minutes in a thermomixer (Eppendorf), agitating at 750 rpm. 0.5 ml DNAzol (Life Technologies) was added to each sample, followed by vortexing and centrifugation at 5000 g for 2 minutes. The supernatant was removed and the pellet was resuspended in 500 µl DNAzol, followed by vortexing and centrifugation as above. The pellets were taken up in 300 µl of 25mM NaOH and incubated at 37°C for 30 minutes. The samples were sonicated using Bioruptor Pico (Diagenode) for 12 cycles (30 seconds on, 30 seconds off; high intensity mode). 1µl of the biotynilated probes (1 pmol/µl) were added to the respective cell lysate and the samples were placed in 30 kD Amnicon

ultrafiltration columns (Millipore). The columns were spun 5 minutes at 12000g, 4°C. 300 µl of BW buffer (Tris-HCl 10mM pH=7.5, EDTA 1mM, NaCl 1M) were added to each sample. The tubes were centrifuged 10 minutes at 12000g, 4°C. The liquid in the column was reduced to approximately 100 µl, which were transferred to 2 ml eppendorf tubes with BW buffer with additional 0.1% TritonX100. The samples were heated up at 65°C at 700 rpm agitation. 600 µl of BW buffer and 50 µl protease-resistent streptavidin beads (New England Biolabs; patent for protease-resistence filed by our lab; appl. Nr. CA2998549A1) were added to each sample and the tubes were rotated at room temperature for 1 hour. Subsequently, the beads were washed 2X with SDS-WB 1 (Tris-HCl 10mM pH=7.5, EDTA 1mM, NaCl 350 mM, SDS 0.5%), 1X with SDS-WB 2 (Tris-HCl 10mM pH=7.5, EDTA 1mM, NaCl 200 mM, SDS 1%), 2X with 20% isopropanol in 500mM NaCl), 5X with 60% ACN in 100mM NaCl.

From this step onwards, the protocol "splits" dependent on the application. The samples were either used for DNA enrichment tests (via qPCR) and next generation sequencing, or for proteome analysis using mass spectrometry.

For DNA analysis:

The beads were resuspended in 220 µl TE buffer (10mM Tris, 1mM EDTA) and incubated overnight at 65°C. Proteinase K (2 µl) were added to each sample and put at 55°C thermomixer (Eppendorf) at 700 rpm for 30 min. The tubes were placed on a magnetic stand for 2 minutes and the liquid transferred to fresh tubes. The samples were then subjected to phenol/chlorophorm isoamylalcohol DNA purification and precipitated using glycogen and 100% ethanol. The purified DNA was taken up in 30 µl elution buffer from Qiagen PCR purification kit. For enrichment analysis using qPCR, TaKaRa SYBR green mastermix and an ABI7500 real-time PCR system (Applied Biosciences) were used. Thermal conditions were set to 95°C for 10 minutes, followed by 40 cycles of 95°C for 15 seconds, 60°C for 30 seconds, 72°C for 30 seconds. For next-generation sequencing, a library was prepared starting with DNA end repairing by Klenow T4 DNA polymerase and T4 polynucleotide kinase, followed by A-tailing and ligation with NEBNext Multiplex Oligos for Illumina (Index primers set 1). The libraries were amplified by PCR under the following thermal conditions: 98°C for 30 seconds, 16 cycles of 98°C for 10 seconds, 65°C for 30 seconds, 72°C for 30 seconds and extension for 5 minutes.

The products were subjected to size selection with Ampure XP beads (0.6X-1.5X-0.98X) and eluted in 15 µl elution buffer (Qiagen). The concentration was measured with Qubit and the fragment size via Bioanalyzer. Finally the pooled libraries were sequenced using Illumina MiSeq set at 4 million reads.

For protein analysis:

The beads were resuspended in 220 µl of 0.1% Rapigest (Waters). 10 µl of 10 mM DTT were added to each sample and they were heated at 95°C for 20 minutes. 10 µl of 200 mM IAA were added to each sample and incubated at room temperature without light for 30 minutes. The reaction was quenched upon addition of 5 µl 100 mM DTT. The tubes were placed on magnetic rack and the liquid placed in fresh tubes. The protein samples were finally digested overnight at 37°C by 100 ng mixture of Trypsin and LysC (Promega). The peptides were subjected to SP3 peptide clean-up as described above and measured on a Orbitrap Velos Pro (Thermo Fisher), connected to nanoAcquity UPLC via a nanoelectrospray ion source (Waters).

6.7 Immunofluorescence staining and microscopy

The following antibodies and stains were used for immunofluorescence in chapter 3:

- Anti-Map2 (Abcam, ab5392)
- Goat Anti-Chicken IgY H&L (Alexa Fluor® 647) preadsorbed (Abcam, ab150175)
- NeuroFluor™ NeuO (Stem Cell Technologies); membrane-permeable fluorescent probe for the detection of live neurons
- DAPI (Thermo Fisher)

For detection of Map2, neurons at day 12 of differentiation were fixed with 2% paraformaldehyde in PBS, permeabelized, and blocked with 1% BSA. Primary Anti-Map2 antibody diluted 1:500 in 1% BSA was applied overnight. Cells were washed in PBS and secondary goat anti-chicken antibody (1:1000) in 1% BSA was applied. DAPI stain (1:500) was added 30 min before microscopy.

Live neurons were stained with NeuO following the manufactuer's protocol.

6.8 Bioinformatic analysis

6.8.1 MS spectra analysis

MS spectra were analyzed using the quantitative proteomics software MaxQuant (open source software by Jürgen Cox and Matthias Mann (Cox & Mann 2008)), using the standard pre-set settings with the following alterations: label-free quantification (LFQ) for p.iPSC vs. s.iPSC and GFP-sorted populations analysis from chapter 3 and SILAC-based relative quantification for TIGR spectra from chapter 4. Spectra from the ChIP-SICAP experiment in chapter 2 were analyzed using iBAQ intensity values. For the library MS/MS matching from chapter 3, "match between runs" was applied. The TMT-based experiment in chapter 2 was analyzed using or Proteome Discoverer 1.4 (Thermo Fisher). Proteins were identified using MASCOT search engine (Matrix Science) and the latest Mus Musculus database from Uniprot.

6.8.2 Statistical and GO term enrichment analysis

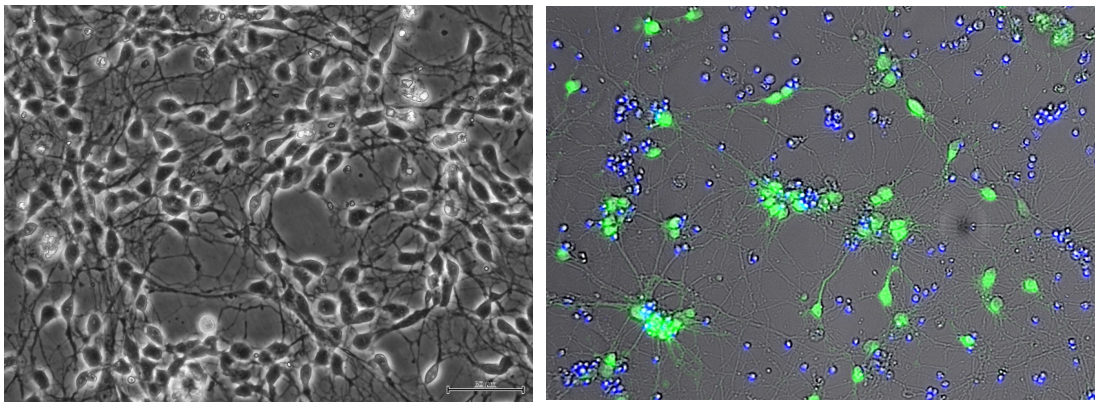
The statistical analysis was performed using the open-source computational platform for analysis of Omics data Perseus (Tyanova et al. 2016) and the open source software environment for statistical computing R (https://www.r-project.org/). Contaminants, proteins only identified by a modification site and proteins derived from the reverse part of the decoy database were filtered out. Differential expression analysis was performed using the "linear models for microarray data" (limma) R package (Ritchie et al. 2015), which is a part of the Bioconductor project (Huber et al. 2015). The dynamic expression clusters were generated using the R-based Graphical Proteomics Data Explorer (GProX) using standard settings as described in the figure legend. GO term enrichment analysis was either performed in GproX (for the protein clusters in chapter 2) or in Cytoscape with the ClueGo application (all other chapters). GO term annotations were downloaded from Perseus. Only significantly enriched terms were considered ($p<0.05$), GO levels set between 3 and 8.

7. List of Abbreviations

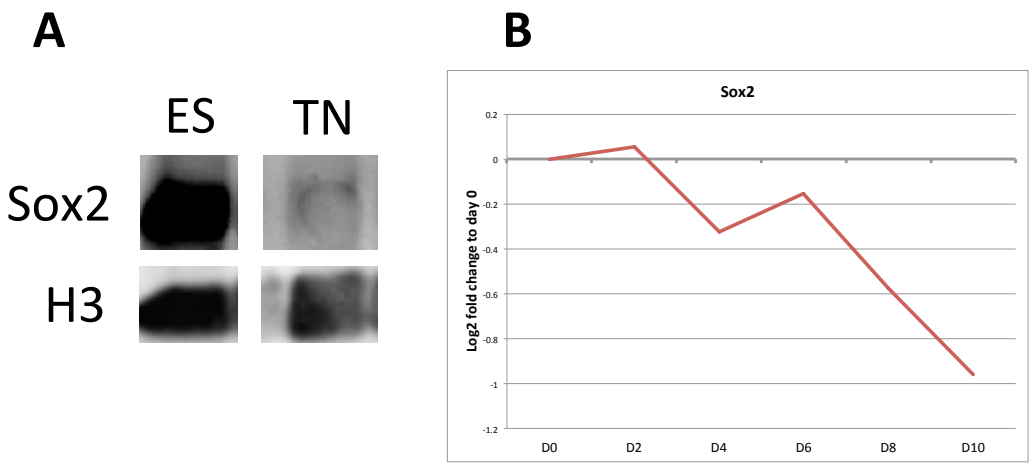
BP	Biological Process
CC	Cellular Compartment
ChIP	Chromatin Immuno-Precipitation
ChIP-MS	Chromatin Immuno-Precipitation followed by Mass Spectrometry
ChIP-SICAP	Chromatin Immuno-Precipitation with selective isolation of chromatin-associated proteins
ESCs	Embryonic Stem Cells
FBS	Fetal Bovine Serum
GO	Gene Ontology
ICM	Inner Cell Mass
iPSCs	induced Pluripotent Stem Cells
LC-MSMS	Liquid chromatography coupled to MSMS. Peptide mixtures are separated based on their biochemical properties usually hydrophobicity), thereby increasing the number of entified and quantified peptides and the proteome sampling depth.
LFQ	Label-Free Quantification. A proteomic quantification strategy which relies on bioinformatic tools post-MS measurement.
M2-rtTA	reverse tetracycline-dependent transactivator
MEFs	Mouse Embryonic Fibroblasts
MF	Molecular Function
MS	Mass Spectrometry
MSMS	Tandem Mass Spectrometry. A method by which precursor peptides are measured once to determine their overall

	mass and then fragmented again to generate fragment spectrum used for their exact identification
OKSM	Oct4, Klf4, Sox2 and c-Myc
Rosa26-rtTA	allele *Gt(ROSA)26Sor* (*Rosa26*) locus8
RT-PCR	Reverse Transcriptase Polymerase Chain Reaction
SCNT	Somatic-Cell Nuclear Transfer
SILAC	Stable Isotope Labeling with Amino acids in Cell culture
SP3	Single-Pot Solid-Phase-enhanced Sample Preparation. A method for proteomic sample preparation for ultrasensitive analysis. Particularly advantageous for limited input material.
SPS	Synchronus Precursor Selection.
STEMCCA	**Stem C**ell **Ca**sette (Col1A-tetO-OKSM, usually with R26-rtTA)
TdT	Terminal deoxynucleotidyl Transferase
TF	Transcription Factor
tetO	tetracycline operon, requires doxycycline-activated rtTA to allow gene expression
TIGR	Targeted Isolation of Genomic Regions
TMT	Tandem Mass Tag. Isobaric peptide-labeling strategy which allows multiplexing (simultaneous analysis) of up to 11 samples.

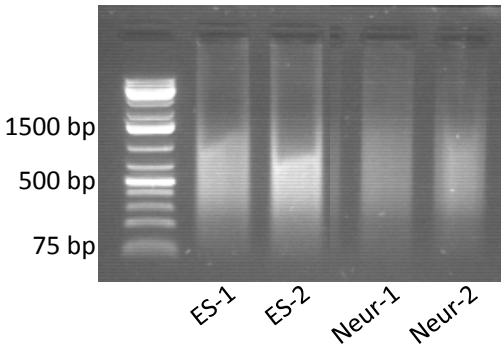
8. Supplementary information



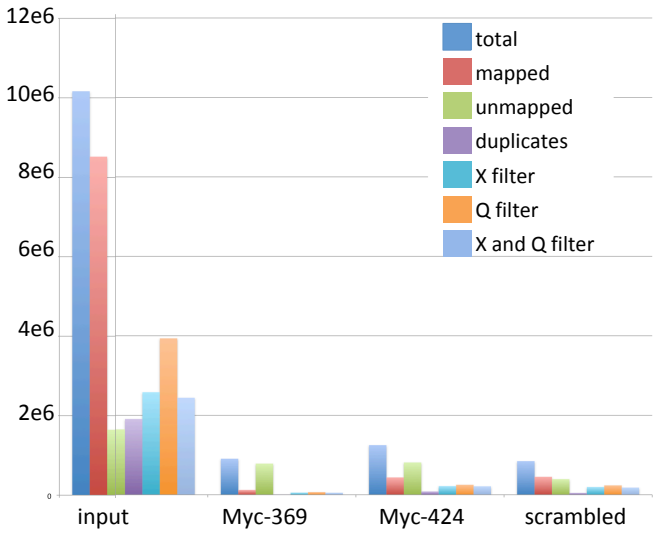
Supplementary Figure 1 Assessment of purity of neuronal population
a) Light microscopy image of neuronal culture at day 10. Neuronal network with secondary connective structures clearly visible. Scale bar 50 µm. b) Assessment of neuronal culture with NeuO (green), a viable marker for mature neurons. Dead cells are stained with DAPI (blue). 95% of cells are NeuO-positive, based on 100 frame count. Scale bar 100 µm.



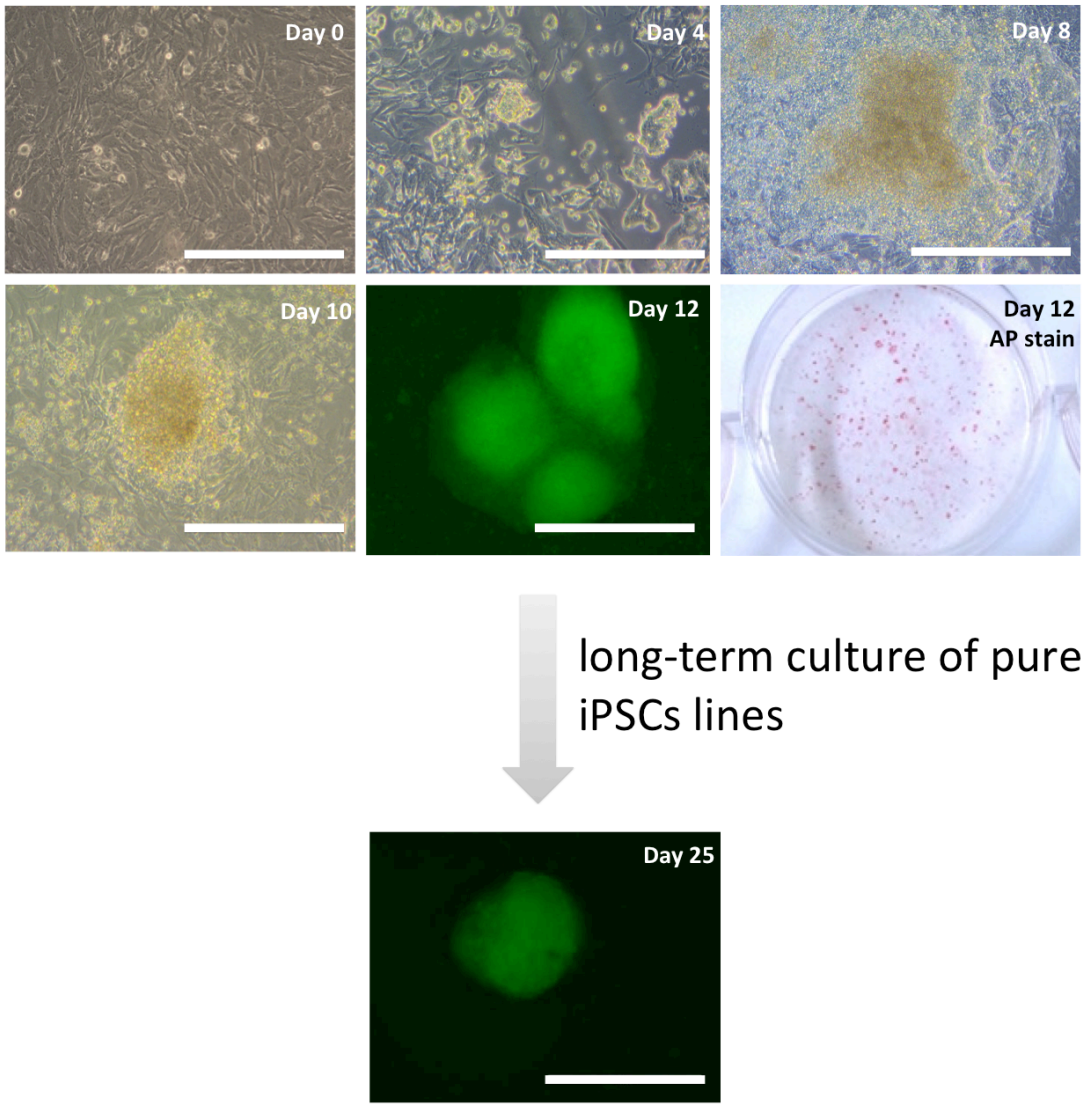
Supplementary Figure 2 Sox 2 protein expression during neuronal differentiation. (A) Western blot at day 0 (ES) and day 10 (TN). (B) Mass-Spec results for Sox2 throughout differentiation. Sox2 strongly decreases but is present until the terminal differentiation stage.



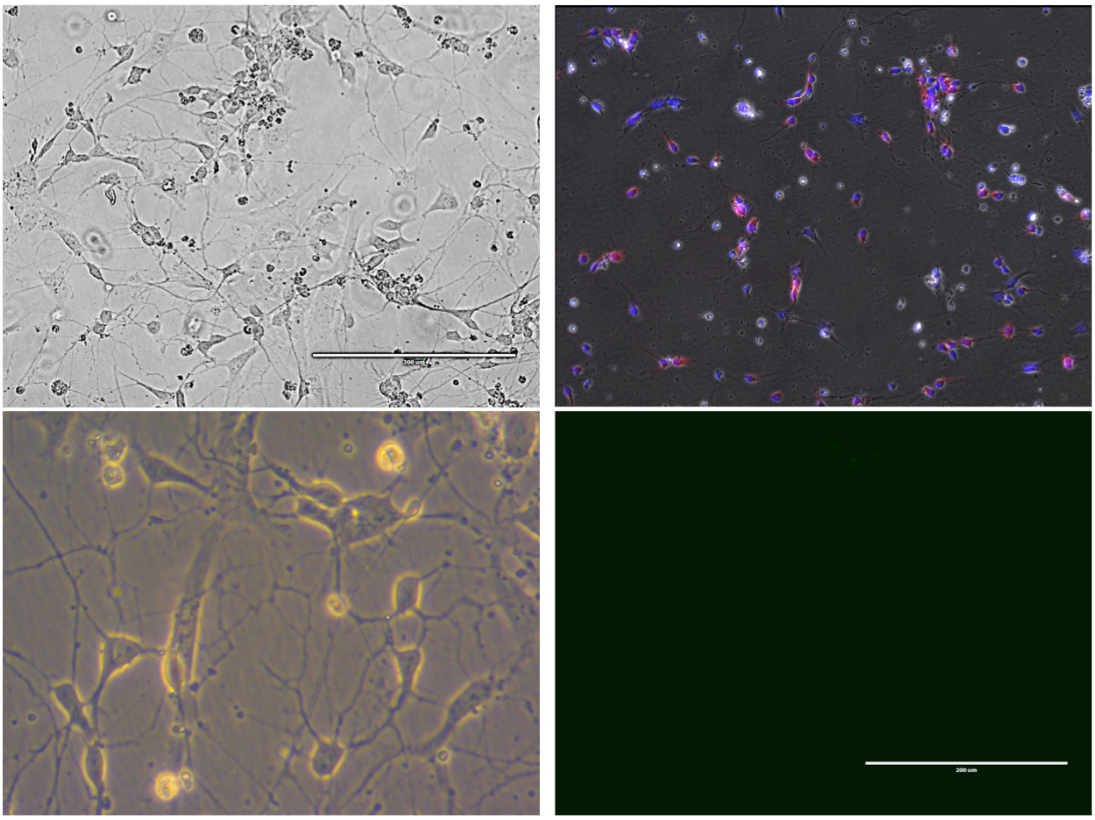
Supplementary Figure 3 Sonicated chromatin of ESCs and TNs, in biological replicates.



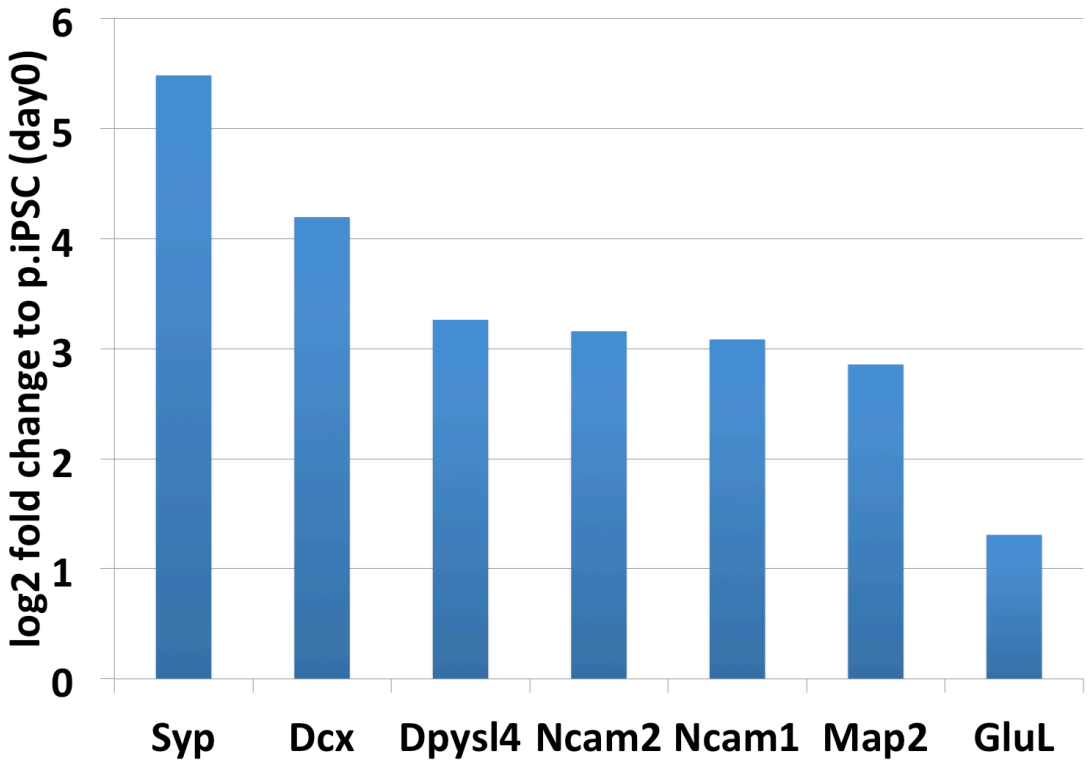
Supplementary Figure 4 Mapped reads of sequenced samples. The color scheme corresponds to the number of reads before (dark blue) and after applying different filters. Overall low number of reads from TIGR pull-downs with both probes, varying between 1e5 for Myc-369 probe and 3.5e5 for Myc-424 and the scrambled negative control.



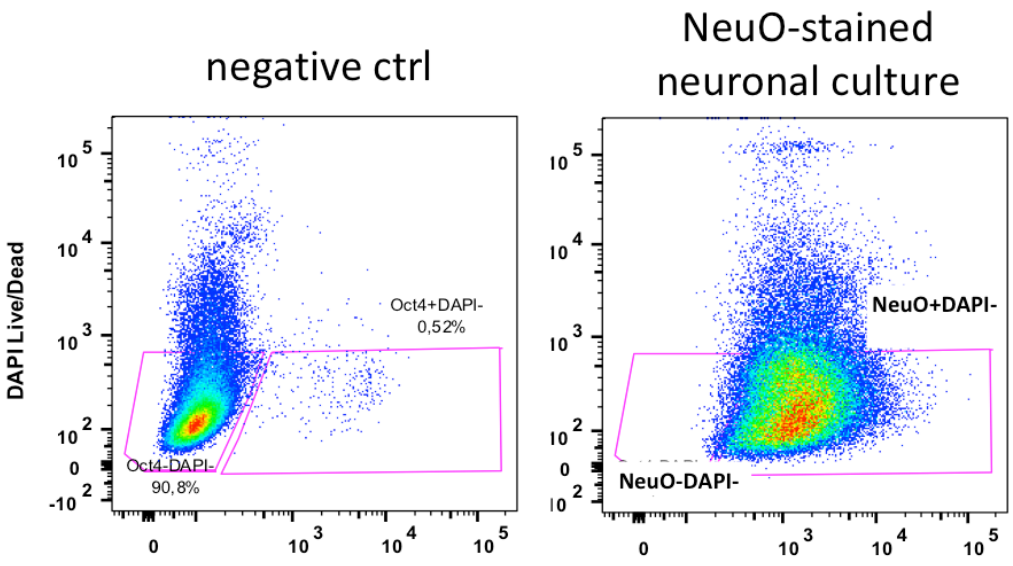
Supplementary figure 5 Generation of primary iPSC lines from MEFs. We used a "fast-track" protocol described in (Bar-Nur et al. 2014). iPSC colonies with characteristic morphology, strong Oct4-GFP expression and positive for the stem cell marker alkaline phosphatase were generated after 12 days. Single colonies were picked and cultured separately for two weeks, thereby firming pure iPS cell lines with characteristic morphology and pluripotency marker expression.



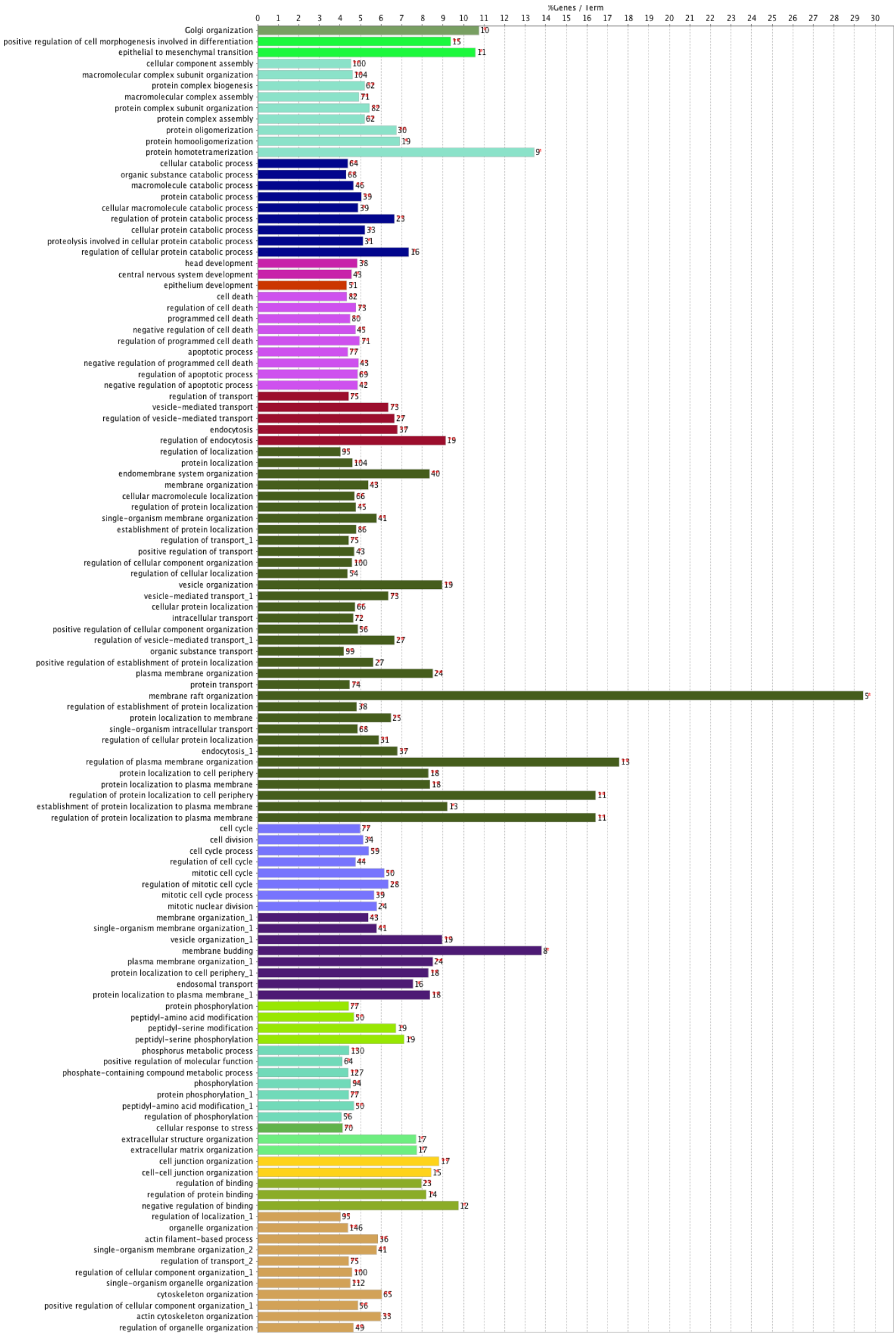
Supplementary Figure 6 Neuronal cultures generated from p.iPSCs. Left: Light microscopy images at day 12 of differentiation. Clearly visible neuronal networks, with a few non-neuronal cells. Right: Fluorescent microscopy of neuronal cultures. Upper: cells were stained with an antibody against the neuronal marker Map2 (mCherry, red) and the cellnuclei with DAPI (blue). Map2+ and Map2- cells were counted, the cultures were predominantly neuronal (86%). Lower: No Oct4-GFP expression visible. Scale: upper left, upper and lower right, 200 µm; lower left, 100 µm.



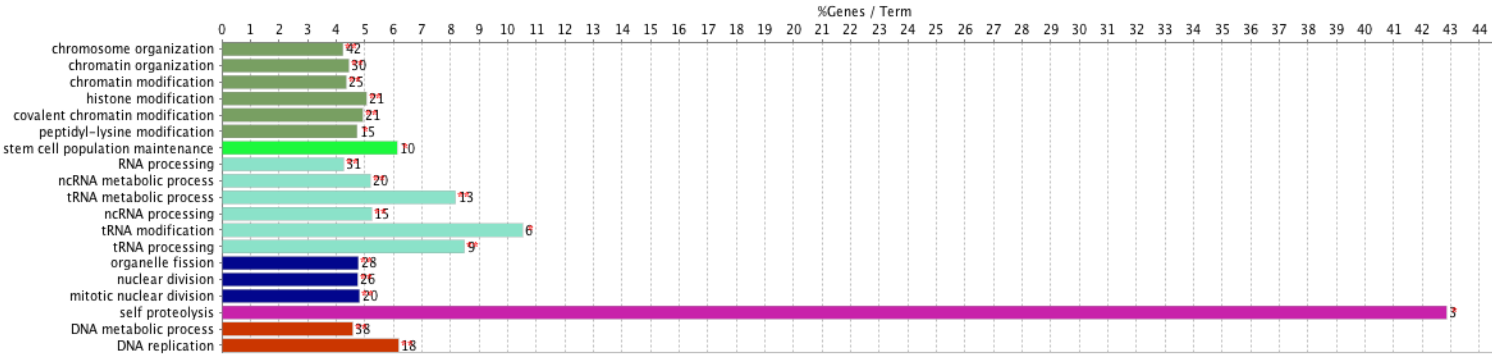
Supplementary figure 7 Expression of neuron- and neurogenesis-specific proteins in the iPSC-generated neuronal cultures (at day 12).



Supplementary Figure 8 FACS analysis of iPSCs-generated neuronal culture at day 12, stained with neuron-specific life stain NeuO. Cells were excited with 488 nm (NeuO) and 405 nm (DAPI). 98.2% of all live cells were NeuO+, 1.8% NeuO–.



Supplementary Figure 8 GO term enrichment analysis of primary (second list)



and secondary (first list) iPSCs. The number beside each column represents the number of proteins identified in this term. The column represents the percentage of all proteins in a GO term covered in our dataset. Only significantly enriched GO terms were included ($p<0.05$). Whole genome background.

Motif similarity=Direct, NES>3.0, all for which there is a TF assigned to motif
Sox2
Sox3
Sox4
Sox6
Sox8
Sox9
Sox10
Sox14
Sox21
Sry
Sf1
Nanos1
Ets1
Gli3
Mafk
Dbp
E2f1
Nr2f1
Nr2f2
Nr3c1
Runx3
Pou5f1
Tcf7
Lef1
Smad1
Patz1
Bad

Supplementary table 1 Transcription factors from iRegulon analysis significantly predicted to regulate the expression of proteins in Cluster 1.

Chd2	Sox2	Smad3	Isl1	Stat5a	Men1	Kmt2b	Pbx1	Cdk8	Nr0b1	Nfyc
Kdm1a	Tcf3	Myod1	Mbd3	Ncor1	Kmt2a	Cbx2	Aebp2	Dgcr8	Dpf2	E4f1
Rest	Tal1	Smad2	Kdm5b	Ncor2	Med23	Onecut1	Vsx2	Drosha	Klf5	Dppa2
Rcor2	Setdb1	Rfx1	Klf4	Daxx	Parp1	Smarcc1	Hcfc1	Phc1	Hira	Suv39h1
Ehmt2	Mtf2	E2f4	Meis1	Sumo2	Pcgf2	Baz1a	Zfp384	Cbx8	Kdm6b	Suv39h2
Sin3a	Jarid2	Tet1	Sox3	Hoxb4	Ncoa3	Phf5a	Mafk	Arid3a	Crebbp	Gad1
Cdyl	Kdm5a	Yy1	Sox11	Baz2a	Kat8	Phrf1	Zc3h11a	Dicer1	Thap11	Pwp1
Brf2	GFP	Mapk8	Cbx7	Ctcfl	Zfp57	Ruvbl1	Gata4	Mettl3	Mcrs1	Etv2
Gtf3c2	Nelfa	Nfya	Cebpb	Rcor1	Kmt2d	Ruvbl2	Phf19	Smarcad1	Msl2	Cdk7
Trim24	Supt5	Atf7ip	Gata2	Cbx1	Kdm6a	Sap18	Gli1	Tbx3	Kansl3	Zic2
BrdU	Ctr9	Hdac1	Lmo2	Ell	Tead4	Smarca5	Gli2	Zfp322a	Sp1	Nkx2-2
Smarca4	Kdm2a	Hdac2	Fli1	Cbfa2t2	Zic3	Srsf1	Pcgf6	Cbx5	Ep400	Nkx6-1
Nap1l1	Rara	Chd4	Gata1	Prdm14	Kdm2b	Srsf2	Usp7	T	Usp16	Hnrnpl
Kat5	Spi1	Dpy30	Gfi1	Neurod2	Ell3	Ssrp1	Aff3	Btaf1	Tcf12	Pknox1
Pou5f1	Atrx	5-hmC	Gfi1b	Biotin	Epop	Baz1b	Sall4	Ino80	Dnajc2	Hand1
Nanog	Tbp	Ctcf	Elk4	5-mC	Elob	Gtf3c1	Leo1	Dr1	Utf1	Rad23b
Epitope tags	Chd7	Runx1	Ncapd 3	Dnmt3b	Zmynd8	Neurog2	Cdc73	Rbpj	Tfap2c	Jun
Otx2	Ep300	Ell2	Ncapg	Taf1	Rybp	Lhx3	Ldb1	Ascl1	Cdx2	Fgfr1
Brd4	Smc1a	Olig2	Rad21	Nrf1	Zfp281	Ebf2	Phf20	Esrrg	Msl1	Nr4a1
Dnmt1	Smc3	Ncaph2	Cdk9	Zfp42	Gmnn	Onecut2	Kdm4c	Stat1	Nsl1	Zfp217
Tet3	Med12	Rxra	Pou3f1	Nono	Zic1	Tfe3	Myc	Zbtb2	Tgif1	Tgif2
Phf13	Med1	Rarg	Pou3f2	Paf1	Ogt	Fam60a	Sox17	Eed	Tdg	Sirt6
Ezh2	Nipbl	Trim28	Trp53	Stag2	Tet2	Kdm4a	Mbd2	Smad4	Taf9b	Brd2
Rnf2	Rbbp5	Taf3	Smad1	Stag1	Zfp143	Foxa2	Cbx3	Max	Esrrb	
Suz12	Wdr5	Raf1	Gtf2b	Bmi1	Morc3	Foxa1	Med26	Supt6	Nfyb	

Supplementary table 2 Full list of all proteins predicted to interact with *c-Myc* in mouse pluripotent stem cells. Analysis done with the *in silico* ChIP tool in the ChIP Atlas. Threshold of significance 100, antigen type "transcription factors", cell type "pluripotent stem cell".

9. References

Abazova, N. & Krijgsveld, J., 2017. Advances in stem cell proteomics. *Current Opinion in Genetics and Development*, 46, pp.149–155.

Apostolou, E. & Hochedlinger, K., 2013. Chromatin dynamics during cellular reprogramming. *Nature,* volume 502, pages 462–471

Arendt, D., 2008. The evolution of cell types in animals: emerging principles from molecular studies. *Nature*, 9, pp.868–882.

Avilion, A.A. et al., 2003. Multipotent cell lineages in early mouse development depend on SOX2 function. *Genes & Development*, 17(1), pp.126–140.

Bajpe, P.K. et al., 2013. The Corepressor CTBP2 Is a Coactivator of Retinoic Acid Receptor/Retinoid X Receptor in Retinoic Acid Signaling. *Molecular and Cellular Biology*, 33(16), pp.3343–3353.

Bar-Nur, O. et al., 2014. Small molecules facilitate rapid and synchronous iPSC generation. *Nature Methods*, 11(11), pp.1170–1176.

Benevento, M. et al., 2014. Proteome adaptation in cell reprogramming proceeds via distinct transcriptional networks. *Nature Communications*, 5, p.5613.

van den Berg, D.L.C. et al., 2010. An Oct4-Centered Protein Interaction Network in Embryonic Stem Cells. *Cell Stem Cell*, 6(4), pp.369–381.

Bibel, M. et al., 2007. Generation of a defined and uniform population of CNS progenitors and neurons from mouse embryonic stem cells. *Nature Protocols*, 2(5), pp.1034–1043.

Boyer, L.A. et al., 2005. Core Transcriptional Regulatory Circuitry in Human Embryonic Stem Cells. *Cell*, 122(6), pp.947–956.

Briggs, R. & King, T.J., 1952. Transplantation of Living Nuclei From Blastula Cells into Enucleated Frogs' Eggs. *Proceedings of the National Academy of Sciences of the United States of America*, 38(5), pp.455–463.

Buenrostro, J.D. et al., 2013. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature Methods*, 10(12), pp.1213–1218.

Byrum, S.D., Taverna, S.D. & Tackett, A.J., 2013. Purification of a specific native genomic locus for proteomic analysis. *Nucleic Acids Research*, 41(20),

pp.e195–e195.

Cavallaro, M. et al., 2008. Impaired generation of mature neurons by neural stem cells from hypomorphic Sox2 mutants. *Development*, 135(3), pp.541–57.

Chaerkady, R. & Kerr, C., 2009. Temporal Analysis of Neural Differentiation Using Quantitative Proteomics†. *Journal of proteome …*, 8(3), pp.1315–1326.

Chappell, J. & Dalton, S., 2013. Roles for MYC in the establishment and maintenance of pluripotency. *Cold Spring Harbor perspectives in medicine*, 3(12), p.a014381.

Cooper, G.M., 2000. *The cell : a molecular approach*, ASM Press.

Costa, Y. et al., 2013. NANOG-dependent function of TET1 and TET2 in establishment of pluripotency. *Nature*, 495(7441), pp.370–374.

Cox, J. & Mann, M., 2008. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature Biotechnology*, 26(12), pp.1367–1372.

Déjardin, J. & Kingston, R.E., 2009. Purification of Proteins Associated with Specific Genomic Loci. *Cell*, 136, pp.175–186.

Do, E. kyoung et al., 2014. Reptin Regulates Pluripotency of Embryonic Stem Cells and Somatic Cell Reprogramming Through Oct4-Dependent Mechanism. *STEM CELLS*, 32(12), pp.3126–3136.

Dubik, D., Dembinski, T.C. & Shiu, R.P., 1987. Stimulation of c-myc oncogene expression associated with estrogen-induced proliferation of human breast cancer cells. *Cancer research*, 47(24 Pt 1), pp.6517–21.

Dubik, D. & Shiu, R.P., 1988. Transcriptional regulation of c-myc oncogene expression by estrogen in hormone-responsive human breast cancer cells. *The Journal of biological chemistry*, 263(25), pp.12705–8.

Egashira, T., Yuasa, S. & Fukuda, K., 2013. Novel insights into disease modeling using induced pluripotent stem cells. *Biological & pharmaceutical bulletin*, 36(2), pp.182–8.

Felsenfeld, G., 2014. A brief history of epigenetics. *Cold Spring Harbor perspectives in biology*, 6(1), p.a018200.

Ferri, A.L.M. et al., 2004. Sox2 deficiency causes neurodegeneration and impaired neurogenesis in the adult mouse brain. *Development*, 131(15), pp.3805–3819.

Fujita, T. & Fujii, H., 2013. Biochemical and Biophysical Research Communications Efficient isolation of specific genomic regions and identification of associated

proteins by engineered DNA-binding molecule-mediated chromatin immunoprecipitation (enChIP) using CRISPR. *Biochemical and Biophysical Research Communications*, 439(1), pp.132–136.

Gagliardi, A. et al., 2013. A direct physical interaction between Nanog and Sox2 regulates embryonic stem cell self-renewal. *Embo J*, 32(16), pp.2231–2247.

Gao, Z. et al., 2012. Determination of protein interactome of transcription factor Sox2 in embryonic stem cells engineered for inducible expression of four reprogramming factors. *The Journal of biological chemistry*, 287(14), pp.11384–97.

Haider, S. & Pal, R., 2013. Integrated Analysis of Transcriptomic and Proteomic Data. *Current Genomics*, 14(2), pp.91–110.

Hall, B. & Olson, W., 2006. *Keywords and Concepts in Evolutionary Developmentalitle*, Harvard University Press.

Hansson, J. et al., 2012. Highly Coordinated Proteome Dynamics during Reprogramming of Somatic Cells to Pluripotency. *Cell Reports*, 2(6), pp.1579–1592.

Hochedlinger, K. & Jaenisch, R., 2002. Monoclonal mice generated by nuclear transfer from mature B and T donor cells. *Nature*, 415(February), pp.1035–1038.

Hochedlinger, K. & Plath, K., 2009. Epigenetic reprogramming and induced pluripotency. *Development (Cambridge, England)*, 136, pp.509–523.

Huang, X. & Wang, J., 2014. The extended pluripotency protein interactome and its links to reprogramming. *Current Opinion in Genetics and Development*, 28, pp.16–24.

Huber, W. et al., 2015. Orchestrating high-throughput genomic analysis with Bioconductor. *Nature Methods*, 12(2), pp.115–121.

Hughes, C.S. et al., 2014. Ultrasensitive proteome analysis using paramagnetic bead technology. *Molecular systems biology*, 10(757).

Inoue, H. et al., 2014. iPS cells: a game changer for future medicine. *The EMBO journal*, 33(5), pp.409–17.

Ivanova, N. et al., 2006. Dissecting self-renewal in stem cells with RNA interference. *Nature*, 442(7102), pp.533–538.

Janky, R. et al., 2014. iRegulon: From a Gene List to a Gene Regulatory Network

Using Large Motif and Track Collections H. J. Bussemaker, ed. *PLoS Computational Biology*, 10(7), p.e1003731.

Jia, J. et al., 2012. Regulation of pluripotency and self- renewal of ESCs through epigenetic- threshold modulation and mRNA pruning. *Cell*, 151(3), pp.576–589.

Kamachi, Y. et al., 2001. Pax6 and SOX2 form a co-DNA-binding partner complex that regulates initiation of lens development. *Genes & Development*, 15(10), pp.1272–1286.

Kelly, S.J., 1977. Studies of the developmental potential of 4- and 8-cell stage mouse blastomeres. *Journal of Experimental Zoology*, 200(3), pp.365–376.

Kim, S.U., Lee, H.J. & Kim, Y.B., 2013. Neural stem cell-based treatment for neurodegenerative diseases. *Neuropathology*, 33(5), p.n/a-n/a.

Kondoh, H. & Kamachi, Y., 2010. SOX-partner code for cell specification: Regulatory target selection and underlying molecular mechanisms. *The international journal of biochemistry & cell biology*, 42(3), pp.391–9.

Kowalewski, A.A., Randall, R.L. & Lessnick, S.L., 2011. Cell Cycle Deregulation in Ewing's Sarcoma Pathogenesis. *Sarcoma*, 2011, p.598704.

Lai, Y.-S. et al., 2012. SRY (sex determining region Y)-box2 (Sox2)/poly ADP-ribose polymerase 1 (Parp1) complexes regulate pluripotency. *Proceedings of the National Academy of Sciences*, 109(10), pp.3772–3777.

Lessard, J. et al., 2007. An essential switch in subunit composition of a chromatin remodeling complex during neural development. *Neuron*, 55(2), pp.201–15.

Levens, D., 2008. How the c-myc promoter works and why it sometimes does not. *Journal of the National Cancer Institute Monographs*, (39), pp.41–43.

Liu, X. et al., 2017. In Situ Capture of Chromatin Interactions by Biotinylated dCas9. *Cell*, 170(5), p.1028–1043.e19.

Liu, Z. et al., 2018. Cloning of Macaque Monkeys by Somatic Cell Nuclear Transfer. *Cell*, 172(4), p.881–887.e7.

Loh, Y.-H. et al., 2006. The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. *Nature Genetics*, 38(4), pp.431–440.

Lopez-Bertoni, H. et al., 2016. Epigenetic modulation of a miR-296-5p:HMGA1 axis regulates Sox2 expression and glioblastoma stem cells. *Oncogene*, 35(37),

pp.4903–4913.

Lyu, Y.L. & Wang, J.C., 2003. Aberrant lamination in the cerebral cortex of mouse embryos lacking DNA topoisomerase II. *Proceedings of the National Academy of Sciences*, 100(12), pp.7123–7128.

Mallanna, S.K. et al., 2010. Proteomic Analysis of Sox2-Associated Proteins During Early Stages of Mouse Embryonic Stem Cell Differentiation Identifies Sox21 as a Novel Regulator of Stem Cell Fate. *STEM CELLS*, 28(10), pp.1715–1727.

Meissner, A. et al., 2008. Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature*, 454(7205), pp.766–70.

Meyer, N. & Penn, L.Z., 2008. Reflecting on 25 years with MYC. *Nature Reviews Cancer*, 8(12), pp.976–990.

Mitalipov, S. & Wolf, D., 2009. Totipotency, Pluripotency and Nuclear Reprogramming. In *Engineering of Stem Cells*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 185–199.

Mulvey, C.M. et al., 2015. Dynamic Proteomic Profiling of Extra-Embryonic Endoderm Differentiation in Mouse Embryonic Stem Cells. *STEM CELLS*, 33(9), pp.2712–2725.

Nakagawa, M. et al., 2008. Generation of induced pluripotent stem cells without Myc from mouse and human fibroblasts. *Nature Biotechnology*, 26(1), pp.101–106.

O'Geen, H., Yu, A.S. & Segal, D.J., 2015. How specific is CRISPR/Cas9 really? *Current opinion in chemical biology*, 29, pp.72–8.

Oki, S. & Ohta, T., 2015. ChIP Atlas, http://chip-atlas.org. Available at: http://chip-atlas.org.

Pardo, M. et al., 2010. An Expanded Oct4 Interaction Network: Implications for Stem Cell Biology, Development, and Disease. *Stem Cell*, 6(4), pp.382–395.

Polo, J.M. et al., 2012. A molecular roadmap of reprogramming somatic cells into iPS cells. *Cell*, 151(7), pp.1617–1632.

Rafiee, M.-R. et al., 2016. Expanding the Circuitry of Pluripotency by Selective Isolation of Chromatin-Associated Proteins. *Molecular Cell*, 64(64), pp.624–635.

Rahl, P.B. et al., 2010. c-Myc Regulates Transcriptional Pause Release. *Cell*, 141(3), pp.432–445.

Riggs, A. & Porter, T., 1996. Overview of Epigenetic Mechanisms. In *Epigenetic Mechanisms of Gene Regulation*. pp. 29–45.

Ring, K.L. et al., 2012. Direct Reprogramming of Mouse and Human Fibroblasts into Multipotent Neural Stem Cells with a Single Factor. *Cell Stem Cell*, 11(1), pp.100–109.

Ritchie, M.E. et al., 2015. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7), pp.e47–e47.

Shoemaker, L.D. & Kornblum, H.I., 2016. Neural Stem Cells (NSCs) and Proteomics. *Molecular & Cellular Proteomics*, 15(2), pp.344–354.

Song, J. et al., 2012. DNA and chromatin modification networks distinguish stem cell pluripotent ground states. *Molecular & cellular proteomics : MCP*, 11(10), pp.1036–47.

Stadtfeld, M. & Hochedlinger, K., 2010. Induced pluripotency: History, mechanisms, and applications. *Genes and Development*, 24, pp.2239–2263.

Tada, M. et al., 1997. Embryonic germ cells induce epigenetic reprogramming of somatic nucleus in hybrid cells. *The EMBO journal*, 16(21), pp.6510–20.

Takahashi, K. & Yamanaka, S., 2006. Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors. *Cell*, 126, pp.663–676.

Takai, H. et al., 2014. 5-Hydroxymethylcytosine Plays a Critical Role in Glioblastomagenesis by Recruiting the CHTOP-Methylosome Complex. *Cell Reports*, 9(1), pp.48–60.

Taleahmad, S. et al., 2015. Proteome Analysis of Ground State Pluripotency. *Scientific reports*, 5(November), p.17985.

Tanaka, S. et al., 2004. Interplay of SOX and POU Factors in Regulation of the Nestin Gene in Neural Primordial Cells. *Molecular and Cellular Biology*, 24(20), pp.8834–8846.

Thomson, M. et al., 2011. Pluripotency factors in embryonic stem cells regulate differentiation into germ layers. *Cell*, 145(6), pp.875–89.

Tyanova, S. et al., 2016. The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nature Methods*, 13(9), pp.731–740.

Ueda, J. et al., 2006. Zinc Finger Protein Wiz Links G9a/GLP Histone Methyltransferases to the Co-repressor Molecule CtBP. *Journal of Biological*

Chemistry, 281(29), pp.20120–20128.

Waddington, 1953. Epigenetics and Evolution. *Symposia of the Society for Experimental Biology*, 7, pp.186–199.

Waddington, C.H., 1957. *The Strategy of the Genes*, London: Geo Allen and Unwin.

Waldrip, Z.J. et al., 2014. A CRISPR-based approach for proteomic analysis of a single genomic locus. *Epigenetics*, 9(9), pp.1207–1211.

Wang, C.I. et al., 2013. Chromatin proteins captured by ChIP–mass spectrometry are linked to dosage compensation in Drosophila. *Nature Structural & Molecular Biology*, 20(2), pp.202–209.

Wang, Z. et al., 2012. Distinct Lineage Specification Roles for NANOG, OCT4, and SOX2 in Human Embryonic Stem Cells. *Cell Stem Cell*, 10(4), pp.440–454.

Watanabe, A. et al., 2007. Fabp7 Maps to a Quantitative Trait Locus for a Schizophrenia Endophenotype T. F. C. Mackay, ed. *PLoS Biology*, 5(11), p.e297.

Watt, F.M. & Driskell, R.R., 2010. The therapeutic potential of stem cells. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 365(1537), pp.155–63.

Wierer, M. & Mann, M., 2016. Proteomics to study DNA-bound and chromatin-associated gene regulatory complexes. *Human molecular genetics*, 25(R2), pp.R106–R114.

Wilmut, I. et al., 1997. Viable offspring derived from fetal and adult mammalian cells. *Nature*, 385(6619), pp.810–813.

Won, K.-J. et al., 2015. Proteogenomics analysis reveals specific genomic orientations of distal regulatory regions composed by non-canonical histone variants. *Epigenetics & Chromatin*, 8(1), p.13.

Wu, S.M. & Hochedlinger, K., 2011. Harnessing the potential of induced pluripotent stem cells for regenerative medicine. *Nature Publishing Group*, 13(5), pp.497–505.

Yang, X. et al., 2000. DNA topoisomerase IIbeta and neural development. *Science (New York, N.Y.)*, 287(5450), pp.131–4.

Yeo, J.-C. & Ng, H.-H., 2013. The transcriptional regulation of pluripotency. *Cell Research*, 23(1), pp.20–32.

Young, R.A., 2011. Control of the Embryonic Stem Cell State. *Cell*, 144(6), pp.940–954.

Zhang, S. & Cui, W., 2014. Sox2, a key factor in the regulation of pluripotency and neural differentiation. *World journal of stem cells*, 6(3), pp.305–11.

Zhao, S. et al., 2004. SoxB transcription factors specify neuroectodermal lineage choice in ES cells. *Molecular and Cellular Neuroscience*, 27(3), pp.332–342.

List of publications

Abazova N, Bunina D, Gehre M, Zaugg J, Noh K and Krijgsveld J: **Multi-omics characterization and Sox2 regulatory interactome network in neuronal differentiation of ESCs.** *(manuscript in preparation)*

Abazova N, Todorow V and Krijgsveld J: **Proteomic memory during cyclic transition from pluripotency to neuroal fate and back.** *(manuscript in preparation)*

Abazova N and Krijgsveld J: **Advances in Stem Cell Proteomics.** *Curr. Opin. Genet. Dev,* 2017 **46:** 149-155

Cheloufi S, Elling U, Hopfgartner B, Jung YL, Murn J, Ninova M, Hubmann M, Badeaux AI, Euong Ang C, Tenen D, Wesche DJ, Abazova N, et al.: **The histone chaperone CAF-1 safeguards somatic cell identity.** *Nature* 2015, **528**: 218-224.

ACKNOWLEDGEMENTS

This thesis would not have been possible without the valuable contributions of everyone I worked with and I would like to take the opportunity to thank you all!

First, I am very grateful to my supervisor Prof. Jeroen Krijgsveld for giving me unprecedented freedom to explore the scientific paths which evoked my curiosity. Trusting me and allowing me to pursue the ideas I had has greatly enriched my independent scientific thinking and the ability to manage my projects. Jeroen was also very generous with his time and was always available to talk to, which is something I appreciated a lot.

I am grateful to my thesis advisory and defense committee members Dr. Kyung-Min Noh, Prof. Frank Lyko, Prof. Andreas Trumpp and Dr. Martin Jechlinger for their support and for the constructive discussions. Additionally, Dr. Noh provided me with a lot of valuable advice and practical directions with regards to neuronal differentiation, as well as with the 129/ESC cell line.

The work described in chapter 2 is part of a collaborative project which involves Dr. Daria Bunina, a shared post-doc between the groups of Dr. Noh and Dr. Zaugg. Daria generated the ATAC-seq data, Maja Gehre (Dr. Noh's group) generated the RNA-seq data. Daria performed the integrative bioinformatic analysis based on these datasets.

I would like to thank all present and former members of the Krijgsveld team who helped me realize this work and who made the last four years so fun. A special "thank you" goes to Vanessa Todorow, who did her Bachelor's thesis under my direct supervision. She greatly contributed to the work on reprogramming and differentiation of iPSCs in every stage of the project. Vanessa's culinary and personal contributions deserve an acknowledgement as well, as she often cooked deliciously and organized fun activities for the entire lab. Dr. Gianluca Sigismondo was a fantastic help for all ChIP-SICAP experiments and true expert of Italian-

grade coffee, without which nothing in the lab would work. Dr. Gertjan Kramer was invaluable support for everything related to mass spectrometry and a sarcastic mastermind of the finest. Dr. Dimitris Papageorgiou was very helpful with the primary and secondary iPSCs comparison experiment; he also brought the Mediterranean sunshine and warmth all the way from Greece to us. Thanks to Sophia Föhr for great technical support, for organizing and hosting fun lab parties and for always being there when I needed her. Torsten Müller helped me with some of the MS analysis; he was also my finest distraction and brought me joy even in the toughest of the working days. Thanks to Jakob Trendel for the constructive lab meeting discussions and for adding a spice to every conversation with his humor and wit. German Monogarov provided a constant supply of chocolate, a fuel of productive energy. Thanks to Bianca Kuhn for being so sweet and caring. Thanks to Dr. Daria Fijalkowska for good discussions on integrative multi-omics analysis and for her general awesomeness. Karim Aljakouch, thanks for being in science and not (yet) pursuing a career as a star pastry chef – but please continue sharing your culinary super-talent with us!

In addition, I would like to acknowledge some former members of our lab. Dr. Christian Frese helped me with MS data analysis of the neuronal differentiation samples and made every party (much) more fun; Dr. Sina Rafiee provided me with instructions on TIGR; Anna Strzeszewska was a great discussion partner on everything from biochemical applications to gender equality.

The support from different facilities and services at EMBL has been vital to this thesis. The proteomics core facility, in particular Mandy Rettel and Dr. Joanna Kirkpatrick; the genome core facility, especially Dr. Vladimir Benes and Paul Collier; the flow cytometry core facility, especially Dr. Malte Paulsen, who was extremely helpful with experimental design and data analysis of the flow cytometry experiments. A big "Thank you!" goes to Bernd Klaus from the Centre for Statistical Data Analysis for invaluable support with the analysis of the full proteome dataset in chapter 2. Matt Rogon from the Centre for Biomolecular Network Analysis helped me with the iRegulon analysis in chapter 2. Charles

Gidarot from the Centre for Biomolecular Network Analysis was indispensible for the sequencing data analysis of the TIGR experiment.

Dr. Bachir El Debs helped me with the data analysis of the sorted populations and has been an overall invaluable support during the writing stage of this thesis. His encouragement and warmth meant the world to me.

Finally, thanks to my friends and family who are the sunshine of my life. Mom and dad, every wonderful thing I have, I owe to you. Thank you for being the best parents and the best people.