Gene Regulatory Dependencies in Mouse Embryonic Stem Cells

Inaugural-Dissertation
zur
Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultät
der Universität zu Köln

vorgelegt von
Fabian Titz-Teixeira
geboren in Wermelskirchen
Köln, 2024

Berichterstatter:

Prof. Dr. Andreas Beyer

Prof. Dr. Achim Tresch

Tag der mündlichen Prüfung: 29.05.2024

Table of Contents

Acknowledgments

Even if I would like to take all credit for this thesis, it wouldn't be fair to everyone contributing to this work, directly or indirectly.

First, I would like to thank Andreas Beyer, my supervisor, who gave me the opportunity to work on this project. His enthusiasm for science and ability to maintain an overview over many different projects has never failed to impress me. His supervision style has always helped me to navigate the different stages of this journey. While I had lots of freedom to explore questions left and right, he always offered valuable input not to get lost during the process. I often utilized his open-door policy to get feedback on minor and major issues. Most importantly, I must thank him for the great atmosphere and work environment he has created in his group.

I am very thankful for the close cooperation with Martin Leeb from Vienna. His group performed most wet-lab experiments that my work is based on. The cooperation between Martin's group and us has always been very close and allowed me to see the bigger picture beyond the screen of my workstation. Andreas Lackner and I were in close exchange from the start of my PhD, which helped me a lot in the beginning. Without the experiments performed by Andreas and Michelle Huth, most of the analysis in this thesis would not have been possible. Additionally, the feedback from all three of them throughout our meetings helped me to better understand the biology behind the project and our results. Lastly, Martin provided critical input and guidance throughout the journey, whether proofreading abstracts for conferences, parts of this thesis, or being part of my thesis advisory committee.

From the start, I have always appreciated the atmosphere in our group. I really enjoyed spending time with the group, from retreats to conferences to lunch and coffee breaks. Even on bad days, there were always lunch and coffee discussions to look forward to. Whether the discussions revolved around science or something completely unrelated did not matter; they were always good for my mood. I would like to emphasize my gratitude to Dennis and Antonios, who proofread my thesis and thereby greatly improved it; Jan, who always helped me a lot, especially in the last year; Tim and Francisco, my long-term office neighbors, were always there for a quick discussion and largely contributed to keeping the mood up; Ronja, Andrew, Paula and many other current and former members of the group massively enriched the time we shared over the last years. I would also like to thank Kay for helping with endless forms and never losing her humor.

I am also deeply thankful to my family and friends. My grandparents, who built the foundation for the privilege to decide to do a PhD in my mid-twenties, my parents for their constant support, and Jeannine for taking the weight off my shoulders when needed. All the nieces and nephews always provided the best distraction, and their laughter is more rewarding than any title could be.

Finally, my sincerest thanks go to Anja. Her love and support improved my life in so many ways. She is always on my side, whether she helps me navigate life or just shares food, wine, and laughter with me on some camping ground.

Erklärung zur Dissertation

gemäß der Promotionsordnung vom 12. März 2020

„Hiermit versichere ich an Eides statt, dass ich die vorliegende Dissertation selbstständig und ohne die Benutzung anderer als der angegebenen Hilfsmittel und Literatur angefertigt habe. Alle Stellen, die wörtlich oder sinngemäß aus veröffentlichten und nicht veröffentlichten Werken dem Wortlaut oder dem Sinn nach entnommen wurden, sind als solche kenntlich gemacht. Ich versichere an Eides statt, dass diese Dissertation noch keiner anderen Fakultät oder Universität zur Prüfung vorgelegen hat; dass sie - abgesehen von unten angegebenen Teilpublikationen und eingebundenen Artikeln und Manuskripten - noch nicht veröffentlicht worden ist sowie, dass ich eine Veröffentlichung der Dissertation vor Abschluss der Promotion nicht ohne Genehmigung des Promotionsausschusses vornehmen werde. Die Bestimmungen dieser Ordnung sind mir bekannt. Darüber hinaus erkläre ich hiermit, dass ich die Ordnung zur Sicherung guter wissenschaftlicher Praxis und zum Umgang mit wissenschaftlichem Fehlverhalten der Universität zu Köln gelesen und sie bei der Durchführung der Dissertation zugrundeliegenden Arbeiten und der schriftlich verfassten Dissertation beachtet habe und verpflichte mich hiermit, die dort genannten Vorgaben bei allen wissenschaftlichen Tätigkeiten zu beachten und umzusetzen. Ich versichere, dass die eingereichte elektronische Fassung der eingereichten Druckfassung vollständig entspricht.“

Teilpublikationen:

Lackner, A. et al. Cooperative genetic networks drive embryonic stem cell transition from naïve to formative pluripotency. EMBO J 40, e105776 (2021). Doi: 10.15252/embj.2020105776

10.03.2024,	Fabian Titz-Teixeira,	Fabian Titz-Teixeira Digital unterschrieben von Fabian Titz-Teixeira Datum: 2024.06.13 09:06:29 +02'00'
Datum,	Name,	Unterschrift

Abstract

The pluripotency of embryonic stem cells (ESCs) is defined by their capacity for self-renewal and differentiation into any cell type of an organism. These two characteristics, combined with the ability to induce pluripotency, harbor great potential to apply pluripotent cells in different fields of research and medicine. In contrast to the well-studied state of pluripotency, the cellular networks that disrupt the self-renewal program and promote the exit from pluripotency are poorly characterized. This thesis aims to investigate the networks that promote differentiation and break the self-renewal circuit, thus leading to cells transitioning from naïve to formative pluripotency.

We conducted a comprehensive analysis of multiple datasets reflecting the naïve to formative transition to elucidate the mechanisms that drive differentiation. A dense 32-hour wildtype (WT) differentiation time course was utilized to quantify the differentiation delay of 73 mouse knockout (KO) cell lines exhibiting altered differentiation. The comparison of the delays of the naïve and formative networks suggested a decoupling between both networks in their exit from naïve pluripotency. Furthermore, we investigated gene regulatory dependencies between genes in the naïve to formative transition. For this purpose, we developed two methods to infer gene regulatory dependencies from the KO data and an independent single-cell (sc) WT differentiation time course. The resulting dependencies of both approaches showed high concordance and revealed that the transcriptional adaptation of the formative network is independent of the naïve network. Additionally, we observed that parts of the formative network had already adapted to differentiation despite inhibiting fibroblast growth factor/extracellular signal-regulated kinase (FGF/ERK) signaling, indicating FGF/ERK independent regulation of the formative network. Moreover, we provided potential upstream regulators of the formative network based on integrating publicly available ATAC-seq data and a database of transcription factor (TF) motifs.

In summary, we show that the transcriptional adaptation of the formative network surprisingly is independent of the naïve network. Thus, whereas the naïve pluripotency network is essential for maintaining pluripotency, genes required to establish subsequent cellular states are under distinct regulatory input.

Zusammenfassung

Die Pluripotenz embryonaler Stammzellen wird durch ihre Fähigkeit zur Selbsterneuerung und Differenzierung in jeden Zelltyp eines Organismus definiert. Diese beiden Eigenschaften, kombiniert mit der Möglichkeit Pluripotenz zu induzieren, bergen ein großes Potenzial für die Anwendung pluripotenter Zellen in verschiedenen Bereichen der Forschung und Medizin. Im Gegensatz zum gut erforschten Stadium der Pluripotenz sind die zellulären Netzwerke, die das Selbsterneuerungsprogramm stören und den Austritt aus der Pluripotenz fördern, schlecht charakterisiert. Diese Arbeit zielt darauf ab, Netzwerke zu untersuchen, die die Differenzierung fördern und den Kreislauf der Selbsterneuerung durchbrechen. Dies führt zur Differenzierung naiver Zellen in Richtung formativer Pluripotenz.

Zur Aufklärung von Mechanismen, die die Differenzierung vorantreiben, führten wir eine umfassende Analyse mehrerer Datensätze durch, welche den Übergang von naiver zu formativer Pluripotenz widerspiegeln. Für die Quantifizierung der Differenzierungsverzögerung von 73 Maus Knockout-Zelllinien mit veränderter Differenzierung, wurde ein Wildtypdifferenzierungszeitkurs über 32 Stunden genutzt. Ein Vergleich der Verzögerungen des naiven und formativen Netzwerks führte zu der Annahme einer Entkopplung beider Netzwerke während ihres Austritts aus der naiven Pluripotenz. Darüber hinaus untersuchten wir genregulatorische Abhängigkeiten zwischen Genen im Übergang von naiver zu formativer Pluripotenz. Um aus den Knockout-Daten und einem unabhängigen Einzelzell-Differenzierungszeitkurs genregulatorische Abhängigkeiten abzuleiten, entwickelten wir zwei Methoden. Die resultierenden Abhängigkeiten beider Ansätze zeigten hohe Übereinstimmungen und deuteten darauf hin, dass die transkriptionelle Anpassung des formativen Netzwerks unabhängig vom naiven Netzwerk ist. Zudem beobachteten wir, dass Teile des formativen Netzwerks sich bereits an die Differenzierung anpassten, obwohl FGF/ERK Signale inhibiert wurden. Dies weist auf eine FGF/ERK unabhängige Regulation des formativen Netzwerks hin. Des Weiteren identifizierten wir potenzielle übergeordnete Regulatoren des formativen Netzwerks auf Grundlage der Integration öffentlich verfügbarer ATAC-seq Daten und einer Datenbank mit Bindemotiven von Transkriptionsfaktoren.

Zusammenfassend zeigen wir, dass die transkriptionelle Anpassung des formativen Netzwerks überraschenderweise unabhängig vom naiven Netzwerk ist. Während das naive Netzwerk für die Aufrechterhaltung der Pluripotenz unerlässlich ist, unterliegen die Gene, die anschließende zelluläre Zustände herstellen, einer anderen regulatorischen Kontrolle.

Abbreviations

Chiron	CHIR990201
DEG	Differentially expressed gene
dKO	double knockout
E	embryonic day
EPI	epiblast
EpiLC	epiblast-like cell
EpiSC	epiblast-derived SC
ESC	embryonic stem cell
ExE	extraembryonic endoderm
Fgf	fibroblast growth factor
FGF/ERK	fibroblast growth factor/extracellular signal-regulated kinase
fPSC	Formative Pluripotent SC
GPR	Gaussian Process Regression
ICM	inner cell mass
iPSC	induced pluripotent stem cell
JAK/STAT	Janus kinase/signal transducers and activators of transcription
KO	knockout
LIF	leukemia inhibitory factor
log2FC	log2 fold change
mESC	mouse embryonic stem cell
NMF	non-negative matrix factorization
PCA	principal component analysis
PD	PD0325901
PRC2	polycomb repressive complex 2
PrE	primitive endoderm
PS	primitive streak
PSC	pluripotent stem cell
RSC	Rosette-like SC
SC	stem cell
sc	single-cell
TE	trophectoderm
TF	transcription factor
tSNE	t-distributed stochastic neighbor embedding
TSS	transcription start site
VE	visceral endoderm
WNT	wingless-related integration site
WT	wildtype
XPSC	X Pluripotent SC
ZP	zona pellucida

1 Introduction

All cells of multicellular organisms are derived from stem cells (SCs)[1,2]. In mammals, this process of deriving specific cell types from SCs starts with the fertilization of the oocyte[3–5], followed by embryogenesis. Here, all structures of the embryo, and therefore of the organism, are initiated through the differentiation of ESCs[6,7]. Malfunction of SCs is linked to diseases like different types of leukemia[8–11] or glioma[12–15]. These malfunctions can occur during differentiation or proliferation of SCs and show their phenotypic effect when the daughter cells differentiate later. In adulthood, SCs replace damaged or dying cells to maintain the entire repertoire of cells the organism needs[16,17]. This concept was adapted in the field of regenerative medicine, with the goal to replace damaged cell populations using SCs[18–20]. Replacement of malfunctional cells with functional cells derived from SCs has been used as a therapeutic strategy in diseases such as leukemia[21–23] and in the restoration of the corneal epithelium to recover vision[24–27].

While an organism's specific cell types serve diverse functions, SCs are defined by two main characteristics[2]: First, their self-renewal capacity allows the cells to undergo extensive population doublings with up to 160 doublings in some extreme cases[28]. The property of self-renewal is defined by symmetric and asymmetric cell division. In symmetric cell division, one SC produces two daughter SCs or two daughter cells that differentiate[29–32]. In asymmetric cell division, one daughter cell remains a SC while the other differentiates to a more specific cell type[33–36]. Second, the potency of the SCs defines which cell types can be derived from a stem cell. As a result, SCs are clonogenic, meaning that one SC is sufficient to give rise to many genetically identical cells and thus form colonies of cells[37–39]. The clonogenicity of SCs becomes critical in leukemia and other diseases where one stem cell gives rise to a malfunctioning colony of cells that take over a bigger part of the overall population of this cell type[40–43].

Pluripotent stem cells (PSCs) are SCs that can give rise to all cell types of the organism but lost the ability to contribute to extra-embryonic tissues[44,45] (further described in 1.2). Advances by Yamanaka and colleagues allowed to chemically reprogram SCs to regain pluripotency and introduced induced pluripotent stem cells (iPSCs)[46]. Here, differentiated cells are exposed to the factors *Oct4*, *Sox2*, *Klf4*, and *c-Myc*. The exposure to the so-called Yamanaka factors leads to the dedifferentiation of the cells and the dedifferentiated cells gain properties of PSCs. This allowed for broader use of SCs and paved the road for iPSC-derived organoids[47–49]. Organoids are tissue-engineered *in vitro* models that recapitulate aspects of the structure and function of the corresponding tissues[50]. The iPSCs either contribute to parts of the organoids or the organoids are fully iPSC derived. These organoids can be used as new systems to model developmental processes and effects of KOs in those developmental processes[47–49,51,52]. Aside from investigation of developmental processes, iPSC derived organoids enable drug screening on a large scale and personalized drug screening approaches when patient derived iPSCs are used[53–56]. In the last years, iPSCs additionally enabled the opportunity to derive embryoid

bodies from either ESCs or iPSCs[57–59]. The embryoid bodies are a promising tool to investigate embryogenesis *in vitro* as they undergo neurulation and heart development[59]. Another recently emerging field in SC research investigates the rejuvenation of SCs[60]. Here, cells are exposed to the Yamanaka factors for a limited time leading to DNA methylation and transcription patterns that resemble patterns at younger age[61,62].

With recent developments in the induction of potency states of SCs, the boundaries were pushed even further than iPSCs by establishing medium conditions to induce totipotency and keep cells in totipotency[63–65]. Totipotent SCs can contribute to all cell types of the organisms and the extra-embryonic tissues[66,67] (further described in 1.2). This work marks an essential step in establishing cell culture conditions for induced totipotent SCs, which might have a similar impact on the field of SC research as Yamanaka and colleagues had with their work. Induced totipotent SCs could be useful for further development of embryoid bodies that are currently achieved by mixing trophoblast and extraembryonic endoderm SCs with ESCs[57–59].

The most apparent process that is associated with SCs is embryonic development. Fertilization of the oocyte by a sperm results in a single-cell embryo and starts this process in mammals[3,5]. Aside from embryogenesis, SCs are crucial at further stages of life. As mentioned above, adult SCs replace dying and damaged cells of the adult organism[16,17]. These cells are also called somatic SCs, and their importance for the adult organism is highlighted by the fact that SC exhaustion in aged individuals culminates in disease formation and death[68].

This thesis will focus on the mechanisms and networks involved in the early differentiation of mouse embryonic stem cells (mESCs). The ability to self-renew and to differentiate into any cell type harbor a great potential in different fields such as regenerative medicine. In contrast to the state of pluripotency, the networks that keep the cells in pluripotency, and how to induce pluripotency of SCs, the exit from pluripotency is only poorly characterized. However, to make use of the full potential of ESCs, the interplay of networks involved in the exit from pluripotency has to be characterized adequately. Thus, this work aims to elucidate the regulatory principles that break the circuit of self-renewal and promote the exit from pluripotency.

1.1 Differentiation of mouse embryonic stem cells

To interpret the network and interplay of processes involved in the exit from pluripotency, it needs to be put in the context of embryo development of the corresponding organism. Thus, an understanding of early mouse embryogenesis is required to put this work into the context of the underlying biology. Here, we will focus on the sequence of events and timing of the corresponding processes required to result in mouse embryogenesis.

The timings of the processes involved in embryogenesis are given by embryonic days (E), starting with fertilization at E0 as a reference time point for subsequent events. Before the sperm can fertilize the egg, it must bind to the zona pellucida (ZP)

and cross it[69]. The ZP, or egg coat, surrounds mammalian oocytes and is a barrier for cross-species fertilization and polyspermy[70,71]. During fertilization, sperm and oocyte fuse and form the zygote. To complete fertilization, the haploid nuclei of sperm and egg fuse to one diploid nucleus[72]. In contrast to many other species, the fusion of the two pronuclei in mammals is not immediate, and membranes of both nuclei break down in preparation for the first mitotic division of the zygote[4,5,73]. Through rapid cell cycles and proliferation, the one-cell zygote duplicates and gives rise to multiple blastomeres, constituting the 2-, 4-, and 8-cell embryo[74]. Each cleavage corresponds to a doubling in cells and can be used as a different timing for developmental processes in early embryogenesis. Flattening and further morphogenic changes of the blastomeres[75–78] result in loss of individual cell definition and morula formation around E2.5. The overall cytoplasmatic volume remains constant over the first few cleavages and thus increases the nucleo-cytoplasmatic ratio[79]. The morula embryo, consisting of eight to sixteen blastomeres, is still surrounded by the ZP.

Blastocyst formation at around E3.5 requires allocation of cell lineages and further morphogenic changes at the 16- to 32-cell stage. The morphogenic change needed at this stage is cavitation driven by the occurrence of small local cavities in the morula that fuse to one cavity[80]. In parallel to these morphological changes, the three lineages of the pre-implantation blastocyst are established. The cells that are on the outside of the embryo after cavitation will contribute to the trophectoderm (TE) lineage. Primitive endoderm (PrE) and epiblast (EPI) lineages, however, will be derived from the cells found on the inner cell mass (ICM)[81]. The separation into the precursors of extraembryonic TE on the outside and PrE and EPI on the inside is driven by the first asymmetric cell divisions during embryogenesis[82,83]. The decision for different blastocyst lineages is not synchronized. The TE lineage occurs at the fourth and fifth cleavage, while the PrE and EPI lineage occur at the fifth and sixth cleavage[83]. After the allocation of lineages, the different lineages show different morphological behavior and speed in cell division. Cells from the PrE and EPI lineage continue with cleavage, resulting in no change in their nucleo-cytoplasmatic ratio, whereas cells from the TE lineage grow in size, decreasing their nucleo-cytoplasmatic ratio[79]. TE cells divide faster than PrE and EPI cells during the fifth and sixth cleavage. This sped-up division in TE cells is then slowed down, and PrE and EPI cells divide faster in the seventh and eighth cleavage[84].

The next step in embryogenesis is the implantation of the blastocyst into the uterus around E4.5, closely followed by gastrulation. One prerequisite for successfully implanting the blastocyst into the uterus is the hatching of the ZP[69]. As the ZP is non-adhesive, hatching is required for the TE to invade the uterine wall and thus implant into the uterus[85]. A consequence of hatching impairment of blastocysts is failure to implant into the uterus or early loss of pregnancy[86–88]. After successful implantation, cells from the ICM grow into the blastocyst cavity, forming the peri-implantation EPI. During this growth, the initially mixed EPI and PrE cells[81] sort into two distinct layers[89,90], separating the embryonic EPI cells from extraembryonic PrE cells.

Further changes in the shape and topology of the embryonic structures coupled with intensive growth during this phase are the foundation for later developmental processes and their success. The extraembryonic endoderm (ExE) and the visceral endoderm (VE) aid structural support and signals for further developmental processes[7]. The ExE is derived from the TE and positioned at the proximal part of the embryo. The VE is derived from the PrE and envelopes the ExE and the EPI. The EPI is restructured first into rosette-like structures of polarized cells, subsequently followed by the emergence of the proamniotic cavity around E5.5, also referred to as central lumen, caused by hollowing of their apical membranes[91]. After this step, the formation of the post-implantation EPI is completed, and the anterior visceral endoderm formation establishes the anterior-posterior axis[92] in the embryo around E6.0.

Around E6.5, the post-implantation EPI forms the primitive streak[93] (PS) in the posterior region of the EPI. The establishment of the primary germ layers is the first sign of gastrulation. During this phase, the embryo is slowly restructured into a cylindric shape. The cells in the PS undergo epithelial-mesenchymal differentiation linked to changes in attributes like cell shape and cell-matrix interactions. The resulting mesoderm cells then invade the space between VE and EPI. Initiation of gastrulation seems to be affected, but not solely controlled, by the total number of cells or tissue mass of the post-implantation EPI. This has been shown by splitting 2-cell embryos and subsequent differentiation into blastocysts with reduced numbers of cells in the ICM[94]. Post-implantation EPI derived from these blastocysts showed delays in various developmental processes, including gastrulation[95]. Removing only one blastomere from 4-cell embryos and deriving 3/4 embryos reduces the number of cells in the ICM and allows estimating cells required in the EPI to initiate gastrulation. Around 1000 cells were the required number of post-implantation EPI cells, allowing for gastrulation[96]. In parallel to the formation of the PS, the amnion formation is initiated during early gastrulation. This process is completed in late gastrulation, around E7.5, giving rise to the innermost extraembryonic membrane surrounding the fetus of amniotes[97]. Additionally, in late gastrulation, the node establishes the left-right symmetry of the embryo[98,99] and cells of the anterior EPI contribute to the embryonic ectoderm[93].

From E8.0 on, organogenesis establishes the foundation for the organs of the mouse embryo[100]. This is achieved by further specifying the patterns established during gastrulation accordingly. Neural tube formation and formation of somites as a result of segmentation of mesodermal cells are the processes starting organogenesis[101,102]. Additionally, during that stage, the distinct form of the embryo is established through restructuring of the cylindric cup into a form containing a head, heart, limbs, and a spinal cord[100]. With the start of organogenesis, early differentiation of the embryo from a one-cell zygote to a structure of tens of thousands of cells establishes the foundation needed for a new organism, and further proliferation and differentiation will develop the embryo into a fetus. As this work investigates very early SC differentiation, further steps are not described in more detail.

1.2 States of potency

One property that defines SCs and determines the role of a SC in the organism is its potency[103]. The potency of the cell restricts which cell types the daughter cells or the cell itself can differentiate to, thereby defining the developmental potential. Through differentiation, SCs lose that developmental potential and become more specialized[104,105]. The least restrictive potency state is totipotency, which allows differentiation towards all cell types of the organism, including extraembryonic cell types[66,67]. In the state following totipotency, named pluripotency, SCs have already lost the capability to give rise to extraembryonic tissues[44,45]. After further differentiation, SCs become multipotent and are further restricted in possible cell types they can differentiate into[106]. Functionally, these multipotent SCs replace damaged and dying cells in adulthood[107,108].

Totipotent SCs can give rise to all cell types, including extraembryonic cell types. Thus, totipotent cells have the potential to form the whole organism[66,67]. Cells in this state are also called omnipotent, and their potential to build extraembryonic tissue is gradually lost during pre-implantation[109–111]. There are arguments to use the term totipotency for organisms only and use different terms for single cells, which can differentiate into all tissues but cannot orchestrate a developmental sequence, resulting in a complete organism[112,113]. Here, totipotency is used in the broader sense and applied to single cells. In mice, both cells from the 2-cell embryo can give rise to all tissues, including extraembryonic tissue[114–116]. This is not the case in 4-cell and 8-cell embryos, where some cells are totipotent, and others have already lost totipotency[117]. The totipotency of a SC can be tested by removing a single blastomere from the pre-implantation embryo and placing it in an empty ZP to transfer it to the uterus of a pseudo-pregnant host[118–120].

Pluripotent stem cells (PSCs) can give rise to all cell types of the three primary germ layers but cannot give rise to extraembryonic tissues[44,45]. Although pluripotency is only transitory *in vivo*, embryo-derived PSCs can self-renew in defined conditions indefinitely[121]. All cells of the ICM and a subset of cells in earlier embryonic stages have already differentiated and transitioned to PSCs[122]. Pluripotency can be described by different stages (Figure 1.1) as the transcriptome and epigenome is highly dynamic between the stages[123,124]. However, the transitions between stages are fluid[121,125–129]. Naïve pluripotency, often called the ground state of pluripotency[128,130,131], is at the less restricted end of the spectrum, and primed pluripotency is at the other. Both stages are linked through formative pluripotency. The rosette-stage has been proposed as an alternative intermediate state between naïve and primed pluripotency[132], but in the context of this thesis, we will focus on naïve, formative, and primed pluripotency. PSCs from any of the pluripotent stages can still contribute to all somatic lineages (mesoderm, endoderm, and ectoderm).

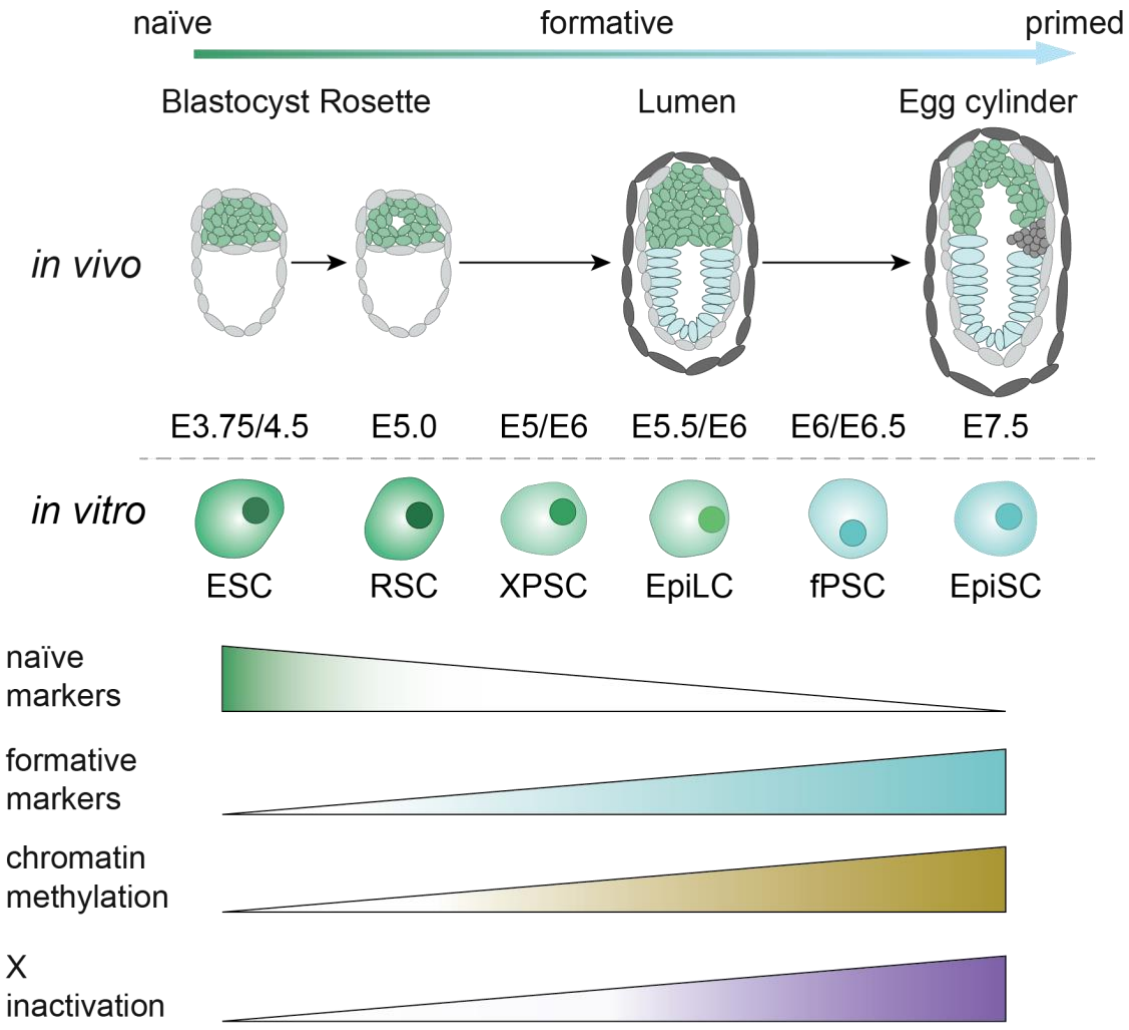

Figure 1.1: Schematic view of naïve to primed pluripotency in vivo and in vitro.

Different in vitro *models are mapped to the corresponding* in vivo *states and put on the E3.75 to E7.5 time axis. Figure is adapted from Furlan et al.*[133] *and changes of transcriptional networks and chromatin were added. Transcriptional rewiring is depicted by changes of naïve and formative markers in blue and green. Changes of chromatin are depicted in yellow and purple.*

Naïve pluripotent cells are the least restricted PSCs and, in addition to cell types of somatic lineages, can also contribute to the germline. ESCs isolated from the ICM of the pre-implantation mouse blastocyst represent the naïve state[44,45,134,135]. The resulting naïve PSCs share high similarity to the pre-implantation EPI in transcription and DNA methylation. Expression of genes specific to naïve PSCs, namely *Esrrb*, *Nanog*, *Tfcp2l1*, *Tbx3*, *Prdm14*, and *Klf4*, promote the self-renewal and serve as a marker for the naïve state of pluripotency. Culture conditions of two inhibitor medium (2i) with or without leukemia inhibitory factor (LIF) keep ESCs in naïve pluripotency *in vitro*, which serves as a ESC research standard for different mouse strains[136,137]. The 2i medium strengthens the above-mentioned naïve TFs and blocks the progression of pluripotency or loss of developmental potential (1.4). Additionally, chromatin in the naïve ground state is largely hypomethylated, making the chromatin more accessible[138–140], including imprinted regions of the genome[141]. The X-chromosome in female cells has not been inactivated in naïve SCs yet, and the cells rely on a mixture for glycolytic and mitochondrial respiration[129].

Following naïve pluripotency, SCs enter the formative pluripotency state. Formative pluripotency describes a transitory state that prepares the cells for lineage priming[124]. Even though the state is only transitory, the rewiring leading to formative pluripotency is needed in preparation for later progression. Expression of the naïve markers is repressed in this state, and expression of the formative markers *Otx2*, *Pou3f1*, *Dnmt3a*, *Dnmt3b*, and *Fgf5* is established instead[124,134,142,143]. Additionally, chromatin methylation becomes more abundant, preparing the cells for lineage priming on a transcriptional and epigenetic level. Inactivation of the X-chromosome is partially observed in formative pluripotency but not completed. This transitory state is also required for contribution to germ cells[124]. ESCs, as well as iPSCs, cannot respond to germ cell specification signals directly but have to be induced into epiblast-like cells (EpiLCs) first before permitting further induction of germ cells[144]. Transcriptionally, these EpiLCs represent E5.5 to E6.5 post-implantation EPI and are different from epiblast-derived SCs (EpiSCs). Post-implantation EPIs at E5.5 to E6.5 are very responsive to germ cell specification signals[145].

The last stage of pluripotency, namely primed pluripotency, describes the most restricted state of pluripotency, in which the cells lose the property to give rise to germ cells. However, the cells are prepared for further differentiation to mesoderm, definitive endoderm, and ectoderm. The chromatin is hypermethylated[146,147] and inactivation of the X-chromosome is completed[148]. Expression of naïve marker genes is either expressed at low levels or absent in primed PSCs[142]. As expression and DNA methylation patterns are adjusted for the different somatic lineages in primed PSCs, the cells are not homogeneous anymore. Primed PSCs are the *in vitro* counterpart to post-implantation EPI and, in contrast to naïve PSCs, cannot contribute to blastocyst chimeras[149]. In the naïve to primed transition, the cells undergo morphological changes and become more polarized after implantation in the uterus[132]. Additionally, the primed cells undergo a metabolic switch to rely only on glycolytic respiration, while naïve pluripotent cells rely on glycolytic and mitochondrial respiration[129].

Further differentiation causes the cells to transition from pluripotency to multipotency. Multipotent SCs lost the ability to contribute to all lineages' cell types as they are already restricted to specific lineages[150–152]. For example, SCs from the mesoderm can contribute to bone, muscle, and blood[153–155]; SCs from the endoderm can contribute to the gut, lungs, and liver[156–158]; and SCs from the ectoderm can contribute to the brain and skin[159,160].

1.3 Comparison of *in vitro* and *in vivo* mouse embryonic stem cells

In this work, we analyze the adaptation of mESCs *in vitro* in the context of early differentiation. However, to gain the most biological insight from *in vitro* experiments, it is important to link them to the corresponding *in vivo* events while being aware of possible differences and limitations. Therefore, we provide an overview of similarities between SCs *in vivo* and *in vitro*, distinct features separating *in vitro* and *in vivo* SCs, and map the *in vitro* state to the corresponding *in vivo* states. *In vivo*, the naïve to

primed transition in the mouse embryo occurs from E3.5 to approximately E7.5[161]. At E7.5, a part of the primed cells has already contributed to the mesoderm or ectoderm by further differentiation[106,162–166]. While corresponding pluripotency states have already been mapped to *in vivo* developmental events in the previous chapter (1.2), the focus in this chapter will be on the shared adjustments taking place *in vivo* and *in vitro* during the naïve to primed transition and differences between *in vivo* and *in vitro* cells at these stages.

In 1981, mESCs were first successfully isolated directly from the blastocyst[44,45]. However, the isolated cells were kept on feeder layers containing mouse embryonic fibroblasts, which restricted the genetic background of the blastocyst-derived ESCs to mainly one strain[167]. In addition to the requirement for the mouse embryonic fibroblasts, the medium was required to contain fetal calf serum[44,45]. However, the requirement for fetal calf serum could be removed by adding the polypeptide *Bmp4*[168] and the mouse embryonic fibroblasts were replaced by LIF[169–171].

Later, conditions based on 2i or 2i LIF culture medium were established and allowed to maintain naïve pluripotency through stabilizing self-renewal and blocking differentiation[130] while also lifting the restriction on the genetic background of the mouse strains[136,137]. One trait of the pre-implantation EPI also found in ESCs is open chromatin and a high transcriptional activity[172–175]. Additionally, female ESC lines exhibit the X-chromosomal inactivation present at this stage *in vivo*[148]. However, this is suspect to variability on the cell level even in 2i medium[176]. The TFs that characterize the *in vivo* pre-implantation EPI, such as *Nanog*, are also stably expressed in *in vitro* ESCs[177,178]. While ESCs are kept in naïve pluripotency through defined medium conditions, this state is transitory in the ICM *in vivo*. Thereby, some genes like *Esrrb* and *Tbx3* that regulate pluripotency *in vitro* are not required in EPI development[179,180]. Additionally, bivalent marks, promotors covered by both H3K4me3 and H3K27me3, differ between *in vitro* ESCs and *in vivo* embryos[181–183]. Even though ESCs are a good model for naïve PSCs in the EPI, their limitations must considered, such as differences to EPI derived naïve PSCs in epigenetic factors and the expression of many genes[184].

Although the cells in this work are isolated from the ICM at E3.5, the cells will show minor adjustments in medium conditions and correspond to cells of the EPI from the pre-implantation embryo at approximately E4.5[134,185]. Thereby, *in vitro* naïve ESCs in 2i or 2i LIF medium best describe EPI cells from the *in vivo* pre-implantation blastocyst around E4.5. Unlike ESCs, derived from the pre-implantation EPI or ICM, EpiSCs are derived from cells of the post-implantation EPI. The cells can be isolated from the EPI over a relatively long window of embryonic time points from E5.5 to E8.0[186–188]. While it is possible to derive EpiSCs from the later stages, the efficiency declines due to the gradual loss of pluripotency in those later stages[189]. Additionally, *in vitro* differentiated ESCs can generate cell populations closely resembling EpiSCs by long-term culture in EpiSC conditions[190,191].

The intermediate state of formative pluripotency is captured in EpiLCs[144] *in vitro* and is achieved by differentiating ESCs in fibroblast growth factor (FGF) and Activin-

containing medium[186,188]. These EpiLCs are most similar to the cells of the post-implantation EPI at E5.75 with a higher similarity to the EPI than EpiSCs. ESCs and EpiLCs differentiate in the presence of FGF and Activin, whereas self-renewal of EpiSCs is promoted under these conditions[186,188]. The EpiSCs correspond to primed PSCs of the EPI during late gastrulation. As this work focuses on the formative pluripotency corresponding to approximately E5.5 as an intermediate state between naïve and primed pluripotency, the intermediate *in vitro* cells mostly referred to in this chapter will be EpiLCs. However, other stages on the spectrum of states between naïve and primed pluripotency can also be captured and represented by *in vitro* counterparts such as rosette-like SCs[132] (RSCs), X pluripotent SCs[192] (XPSCs), and formative pluripotent SCs[193] (fPSCs). Here, RSCs and XPSCs represent stages of pluripotency that are less progressed than EpiLCs, while fPSCs represent formative cells poised for gastrulation[133].

While mESCs derived from the ICM of the blastocyst grow in dome-shaped colonies *in vitro*, EpiSCs grow in monolayers similar to human ESCs[129,188]. This change of cell shape *in vitro* recapitulates changes of cell shape that the cells of the *in vivo* EPI are exposed to. *In vivo*, the ball of cells from the ICM of the blastocyst transforms into a cup-shaped cylinder consisting of a monolayer of polarized cells[162]. Aside from cell shape, other features of the transitions observed *in vivo* are also observed *in vitro*. The increase in overall DNA methylation[138,140,146,147] from naïve to primed pluripotency and X-chromosomal deactivation in female ESCs[148] are observed *in vivo* and *in vitro*. However, as mentioned above, *in vitro* models also have limitations in fully representing the *in vivo* state as many genes and epigenetic factors show differences[181–184]. Thus, *in vitro* models are helpful to investigate the early differentiation of mESCs, but it should always be considered that the resulting findings do not necessarily reflect the *in vivo* differentiation.

1.4 Transition between states of pluripotency and regulatory networks in control

To be able to interpret the results of this work in the context of early differentiation, an overview of known pathways and regulatory networks that define the different states of pluripotency is important. The cells in this study represent naïve pluripotent SCs, and the experiments cover their transition towards formative pluripotency. Our experiments only partially cover the transition from formative to primed pluripotency. Thus, we will emphasize the pathways and networks involved in naïve and formative pluripotency and the transition between both states and focus on Janus kinase/signal transducers and activators of transcription (JAK/STAT) signaling, canonical wingless-related integration site (WNT) signaling, and FGF/ERK signaling, the three most important pathways to this work (Figure 1.2).

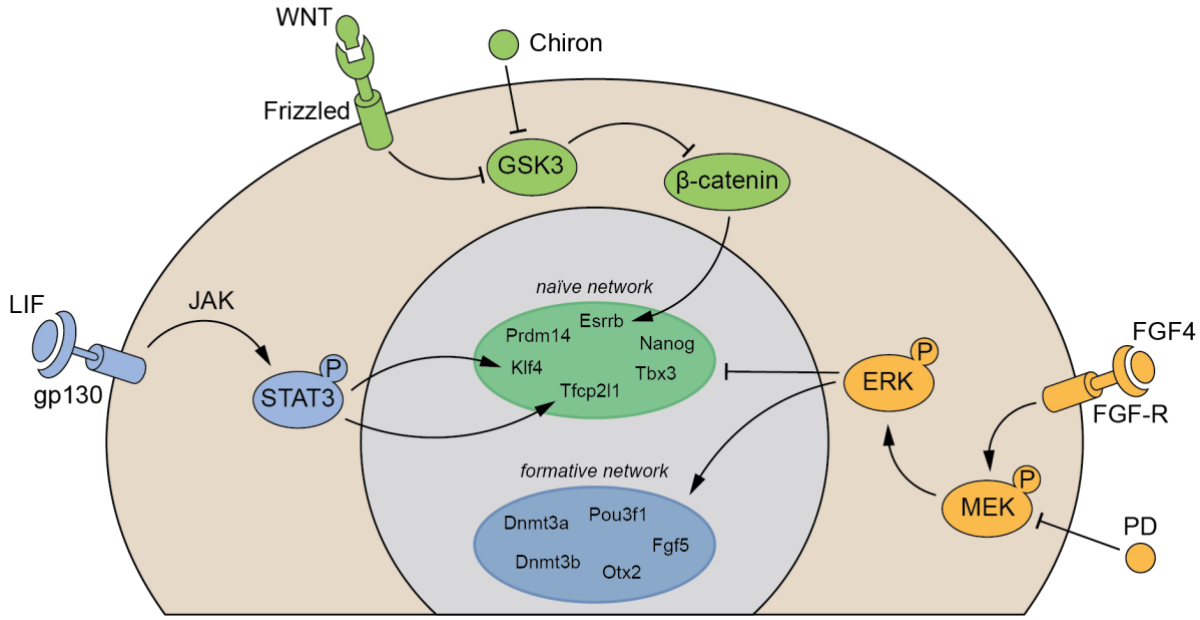

Figure 1.2: Effect of JAK/STAT (blue), canonical WNT (green), and FGF/ERK signaling (orange) on the naïve and formative network.

Lines with arrowheads indicate activation and lines with bars indicate repression. Figure is adapted from Hackett et al.[128].

JAK/STAT signaling is one signaling cascade that promotes and stabilizes the expression of parts of the naïve network[128]. Extracellular LIF also used as an inhibitor of differentiation[169,170,194,195], binds to a heterodimeric cell surface receptor consisting of LIF-R and gp130[196]. This complex leads to subsequent phosphorylation and activation of STAT3 in the cytoplasm. The expression of *Lif* in the trophectoderm and *Lifr* and gp130 in the ICM complement each other in the blastocyst[197,198]. While LIF is a crucial part of mESC culture conditions based on serum, embryos missing *Lif*, *Lifr*, and gp130 expression can still differentiate to mid-gestation stage[199–201]. *Stat3*, on the other hand, is expressed from the oocyte to the blastocysts, and loss of *Stat3* results in late blastocysts lacking cells with EPI and primitive endoderm identity[202].

Canonical WNT signaling also influences the tendency of naïve ESCs to stay in self-renewal or to start differentiating[203,204]. This is also highlighted by the small chemical compound CHIR990201[205,206] (Chiron), which mimics WNT signaling by repression of GSK3 and, in combination with other factors, is capable of blocking differentiation of naïve ESCs[130,207]. While canonical WNT signaling is active in naïve ESCs, it is shut down when cells start to differentiate[208]. The signaling is activated through binding of WNT ligands to Frizzled receptors[209]. The following signaling cascade stabilizes *β-catenin* in the cytoplasm through repression of GSK3 and results in translocation of *β-catenin* into the nucleus[167]. *β-catenin* stabilizes the expression of *Esrrb* in the nucleus through interaction with TCF7L1[204,210]. TCF7L1, also known as TCF3, is a transcriptional repressor that competes with the TFs OCT4 and SOX2 for core pluripotency genes[211,212]. Thus, the interaction between *β-catenin* and TCF7L1 in the nucleus promotes self-renewal and the expression of naïve markers. Chiron closely

mimics the WNT-dependent repression of GSK3 and thus also stabilizes *β-catenin* in the cytoplasm and nucleus[130,207].

Various processes are known to be regulated through the MAPK signaling cascade, including proliferation and differentiation[213,214]. ERK, the former name of MAPK, is activated downstream of extracellular FGF signals through phosphorylation by MEK [208]. Previously it has been shown that inhibition of FGF/ERK signaling promotes stability of the naïve pluripotent state and inhibits differentiation[215,216]. Even though inhibition of MEK signaling in the 8-cell state was shown to suppress the development of the primitive endoderm, it does not block blastocyst formation itself[217,218]. Additional studies have shown that FGF/ERK signaling is required to destabilize the naïve network and initiate differentiation of naïve ESCs[219,220]. The importance of FGF/ERK signaling in the differentiation of naïve ESCs is leveraged by PD0325901 (PD), which inhibits MEK and thereby its downstream target ERK[130,205].

The expression of different *Fgf*s and corresponding *Fgf* receptors depends on the cell state and can be used as marker genes for certain cell states[208]. *Fgf5*, for example, has been used as a marker gene for the post-implantation EPI[221]. *Fgf2* plays a role in stabilizing primed pluripotency in EpiSCs and conversion from ESCs to EpiSCs[208]. *Fgf4* is the only *Fgf* expressed in the EPI but not in the trophectoderm or the primitive endoderm[222–226]. While *Fgfr1* is expressed in all cells of the blastocyst, *Fgfr2* is more specific to the trophectoderm, only showing weak expression in the primitive endoderm and no expression in the EPI[227].

The pluripotency state functioning as the starting point in our experiments is naïve pluripotency represented by mESCs derived from the ICM of the blastocyst[134,185]. These cells show high expression of *Nanog*, *Tfcp2l1*, *Esrrb*, *Tbx3*, *Klf4*, and *Prdm14*, referred to as naïve marker genes in this work. Nanog in combination with *Sox2* and *Oct4* is considered a core pluripotency factor[228–230]. However, it is surprisingly not part of the Yamanaka factors used to induce pluripotency nor does it improve efficiency of reprogramming[46,231–234]. *Klf4*, induced by JAK/STAT signaling[235,236], is a member the Yamanaka[46] factors and an important regulator of cell cycle under conditions of DNA damage[237,238]. The naïve marker genes additionally interact with each other directly, for example, *Klf4* regulates *Nanog*[239] and *Esrrb* is a target gene of *Nanog*[240].

mESCs are held in the naïve state in cell culture by actively blocking differentiation. As described in the previous paragraphs, this is achieved by adding at least two inhibitors to the medium in which the cells are cultured[130]. These inhibitors affect WNT signaling, FGF/ERK signaling, and JAK/STAT signaling, leading to stabilized naïve pluripotency in the ESCs[130]. The first inhibitor is Chiron, which mimics WNT/*β-catenin* signaling and thereby stabilizes the expression of the naïve marker *Esrrb*[205,206]. The second inhibitor, PD, inhibits FGF/ERK signaling[130,205]. The inhibition of FGF/ERK signaling further stabilizes the naïve network by blocking factors that would promote the differentiation of ESCs[219,220]. Adding these two inhibitors to the culture medium is referred to as 2i and is sufficient to keep the cells in naïve pluripotency[130]. Additionally, LIF can be added to the medium, further stabilizing

naïve pluripotency by promoting the expression of *Klf4* and *Tfcp2l1* through JAK/STAT signaling [169,170,194,195]. LIF was initially used in medium containing *BMP4* from the serum[168], blocking the formative network. 2i, however, is sufficient to stabilize the cells in the naïve state; therefore, LIF is not always included in the medium[130]. The cells used as a reference for the ground state of pluripotency in this work are kept in medium containing 2i without LIF.

When the medium, including 2i, is replaced by differentiation permitting medium without the inhibitors, the cells start to differentiate. The removal of PD lifts the inhibition of FGF/ERK signaling and allows the upregulation of the FGF/ERK pathway[130,205]. With the removal of Chiron, the mimicking of WNT signaling is stopped and downregulation of WNT targets, such as *Tcf7l1*, is possible[130,207]. JAK/STAT signaling is downregulated from naïve to formative pluripotency. Even though LIF signaling is not additionally blocked in 2i medium, the fact that LIF and serum conditions were used to keep ESCs in naïve pluripotency before 2i medium highlights the importance of LIF signaling for naïve pluripotency[169,170,194,195]. Approximately 24 hours after changing to differentiation-permitting medium, the cells reach the formative state, and an additional 24 hours later, they become primed pluripotent SCs[124,142,241–243].

The transition from naïve to formative pluripotency is governed by extensive transcriptional rewiring represented by the downregulation of the naïve marker genes[244]. While naïve markers are downregulated, expression of genes specific to the formative state is established[122,124,142,143]. The marker genes for the formative network used in this work are *Otx2*, *Fgf5*, *Pou3f1* (*Oct6*), *Dnmt3a*, and *Dnmt3b*. These formative markers maintain expression in their transition from formative to primed pluripotency[124,143]. The *de novo* DNA methyltransferases *Dnmt3a* and *Dnmt3b* have been shown to be essential during development[245]. Other core pluripotency genes like *Oct4* and *Sox2*, which are not or only slightly downregulated in the naïve to formative transition, are downregulated when the cells differentiate towards primed pluripotency[143,244]. Additionally, the transition to primed pluripotency is defined by the upregulation of lineage factors that prime the cells for later lineage decisions[243].

Although the transcriptional networks and pathways active at different stages of pluripotency are well described, the mechanisms guiding the transition between different stages remain to be elucidated. More specifically, the processes that initiate the shift from self-renewal to differentiation in the transition from naïve to formative pluripotency, which consequently lead to the described changes in pathway activities and different transcriptional networks, remain to be elucidated.

2 Aim of the thesis

The different states of pluripotency in SCs and the different transcriptional networks that define these states are well studied. The ability of SCs to differentiate into all cell types of an organism, combined with the possibility to induce pluripotency, harbors great potential for using them in different research and medical fields. However, the mechanisms that govern the switch from self-renewal and promote the exit from pluripotency remain to be explored. In this thesis, we aim to investigate the mechanisms that disrupt the self-renewal program of the stem cells and initiate the exit from pluripotency.

For this purpose, we have analyzed multiple datasets that reflect the transition from naïve to formative pluripotency. First, we examined the differentiation delay of 73 different KO cell lines in the first 24 hours of differentiation. To determine the differentiation delay, we used a dense 32-hour WT differentiation time course as a reference to match the transcriptional changes of the KOs to the time course. The comparison of delays for different networks suggested a decoupling between the naïve and formative networks in the exit from naïve pluripotency.

We inferred gene regulatory dependencies from the KO data and an independent sc WT differentiation time course to investigate the decoupling between the two networks further. Here, we developed two methods and overlapped both approaches. The results from both approaches revealed that the formative network adapted independently from the naïve network. Analysis of the transcriptional adaptation of both networks under inhibition of FGF/ERK signaling further suggested that there must be FGF/ERK independent mechanisms initiating the upregulation of the formative network. Finally, we used publicly available TF motif and ATAC-seq data to propose potential upstream regulators that drive the upregulation of the formative network.

3 Contributions

All computational analyses and visualizations (unless stated otherwise) presented in this thesis were performed by me. The KO data used in multiple steps of the analyses was processed and analyzed by Robert Sehlke and Marius Garmhausen. Robert Sehlke performed the differential expression analysis of the KO data, and the results are used for analyses in this work.

Most experimental work in the laboratory was performed by Martin Leebs' group members in Vienna. Michelle Huth and Andreas Lackner performed the experiments on the WT differentiation time course and the KO data. The medium change experiment and the double knockouts (dKOs) were performed by Michele Huth. Christa Bücker kindly provided the scRNA-seq data used in this thesis. Experimental work and data processing were done by members of Christa Bückers' group in Vienna. The count table and assignments of cells from CellRanger were provided to us, and the following analyses were performed by me. The relevant experimental design is provided for the corresponding datasets.

Processing of the WT time course and the quantification of the KO delays is published in:

Lackner, A. et al. Cooperative genetic networks drive embryonic stem cell transition from naïve to formative pluripotency. EMBO J 40, e105776 (2021). Doi: 10.15252/embj.2020105776

Other results presented in this thesis are unpublished.

4 Methods

4.1 Wildtype differentiation time course

The WT differentiation time course covers the first 32 hours of differentiation from naïve to primed pluripotency. The cells were cultured in 2i medium to maintain naïve pluripotency. The medium was changed to N2B27 to start differentiation, and RNA-seq measurements were performed every 2 hours without replicates. The 32-hour and 2i samples were done in doublets, and an additional doublet of cells in 2i + LIF medium was sequenced. The 2i samples serve as the reference for naïve pluripotency.

Raw RNA-seq data were aligned to the mm10 genome and read counts were obtained using the STAR aligner[246] (version 2.5.2b). Gene expression for the following biotypes was quantified: protein_coding, lincRNA, processed_transcript, antisense, 3prime_overlapping_ncRNA, bidirectional_promoter_lncRNA, macro_lncRNA, miRNA, misc_RNA, lncRNA, scaRNA, scRNA, sense_intronic, sense_overlapping, snoRNA, snRNA, and sRNA. Read counts were normalized with the apeglm() algorithm[247], integrated in DESeq2[248] (version 1.40.2), and subsequently, the log2 fold changes (log2FCs) between all samples and the mean expression of the 2i samples was calculated. This resulted in expression profiles for each gene over 32 hours of differentiation.

Gaussian Process Regression (GPR) was applied to smooth the profiles of each gene using the bgp() function (R package tgp version[249] 2.4.21; settings: “bmznot” as beta prior and “expsep” for the Gaussian process correlation model). The chosen beta prior applies an independent Normal prior with mean zero, and the chosen correlation model uses a separable power exponential family. The separable power exponential family assumes a higher correlation for measurements with lower distances. Log2FCs profiles against 2i were z-transformed, and genes that changed during the time course were clustered into nine groups using hierarchical clustering. Here, we used the hclust() function (R package stats version 4.3.1) using Euclidean distance and the “ward.D2” method.

4.1.1 Induction and repression time points

Two different approaches to identify induction or repression points were applied to the log2FC profiles of the time course. The first approach was based on runs of expression gain or expression loss defined by slopes of log2FCs between time points. A run starts when the absolute value of the log2FC greater or equal to 0.1 and ends when the log2FC crosses that threshold in the other direction later. When a run ends due to the first derivative crossing the threshold in the opposite direction, a new run in that direction starts. Additionally, a point of regulation can only be followed by the opposing point of regulation, i.e., a point of induction can only be followed by a point of repression and vice versa.

The second approach was based on the first and second derivatives of expression profiles for the time course. Here, we calculated a dynamic threshold per gene based on the maximum possible change between the most highly and lowest expressed time points and the number of time points. The scaling factor is set to 16/17, thus making the denominator smaller than the total number of time points. A denominator smaller than the number of time points forces the dynamic threshold to be higher than the change from a steady state to a monotonic increase or decrease of expression.

Calculation of dynamic threshold to define induction and repression points.

$$\text{dynamic threshold} = \frac{\log2(\text{maximum TPM}/\text{minimum TPM})}{\text{number of time points}*\text{scaling factor}}$$

All time points where the absolute value of the second derivative was higher than this dynamic threshold were considered points of regulation. The following first derivative defined the type of regulation, either indicating expression gain or loss for induction or repression time points, respectively. Additionally, the same criterion as in the other approach was applied, i.e., a point of induction can only be followed by a point of repression and vice versa.

4.1.2 Differential expression in medium change

The medium change covers the early adaptation to the change to different media. The cells were cultured in 2i medium and transferred to one of four different media. The cells were transferred to unconditioned 2i medium, N2B27 medium, N2B27 medium containing PD, or N2B27 medium containing Chiron. RNA-seq measurements were performed in triplicates. The reference measurements were performed in 2i medium before changing the medium and 4 hours and 8 hours after changing the medium to one of the four different media.

The RNA-seq data was trimmed using trimgalore[250] (version 0.6.7-1) aligned to the mm10 genome and read counts were obtained using the STAR[246] aligner (version 2.9.9a). Gene expression for the following biotypes was quantified: protein_coding, lincRNA, processed_transcript, antisense, 3prime_overlapping_ncRNA, bidirectional_promoter_lncRNA, macro_lncRNA, miRNA, misc_RNA, lncRNA, scaRNA, scRNA, sense_intronic, sense_overlapping, snoRNA, snRNA, and sRNA. Read counts were integrated into DESeq2[248] (version 1.40.2) and log2FCs and adjusted p-values for the differences between all samples and WT in 2i were obtained.

4.2 Differentiation delay Knockouts

The KO data covers the naïve to formative transition in WT and 73 different KOs. The KOs were selected based on a previous screening approach[251]. RNA-seq measurements for each KO were performed in doublets from samples cultured in 2i medium and 24 hours after changing medium to N2B27 medium. WT samples were performed in doublets per batch, resulting in 14 WT samples in 2i and 14 WT samples 24 hours after the change to N2B27 medium.

4.2.1 Quantification of knockout delay

The expression profiles of 73 knockouts at 24 hours were mapped onto the differentiation axis from the time course to quantify differentiation delays per KO, i.e., we aimed to quantify how many hours the KO expression pattern is behind the expected differentiation in WT cells. This mapping to the time axis was done first by computing the log2FCs between the KO and WT at 24 hours of differentiation for each KO. The resulting log2FC profiles were compared with the log2FCs at each time point during the WT differentiation. We assumed that the time point at which the difference between the WT profile and the KO profile is minimized best reflects the molecular differentiation state of the respective KO.

We used the Euclidean distance for this purpose. The distances between the knockouts were scaled to make them better interpretable.

Calculation of scaled distances per KO over all time points.

$$\text{scaled distance} = \frac{\text{Euclidean distance} - \min \text{Euclidean distance}}{\max \text{Euclidean distance} - \min \text{Euclidean distance}}$$

The maximum and minimum distances refer to the respective maximum and minimum Euclidean distances of the respective KO across all time points, i.e., the worst and the best matching time point. Thus, the best matching time point will get a distance score of 0.

In order to further increase the time resolution, additional time points were imputed via linear interpolation. Here, we interpolated expression profiles every 15 min between the two neighboring time points of the time point with the smallest Euclidean distance. The timing of knockouts was repeated on the interpolated time points, and the new minimal distance was used to quantify the differentiation delay.

4.2.2 Dependencies between genes

First, genes of interest were restricted to constitutive genes, induced genes, and naïve associated genes, all defined based on the KO data in previous work[251]. Additionally, genes that were differentially expressed (abs(log2FC) >= 0.5 and adjusted p-value <= 0.05) in the WT between 24 hours and 2i were excluded. This led to 1203 genes of interest for the following analysis. WT completion was calculated for all 1203 genes in all 73 KOs.

Calculation of WT completion for a given gene.

$$\text{WT completion} = \frac{\log2\text{FC}(24 \text{ hours vs } 2\text{i})_{\text{KO}}}{\log2\text{FC}(24 \text{ hours vs } 2\text{i})_{\text{WT}}}$$

The distribution of differences in WT completion across all KOs was calculated for each gene pair. Subsequently, the mean of each distribution was tested for a significant difference from zero. Here, we used the two-sided t-test and corrected for multiple testing according to Benjamini and Hochberg[252]. All gene pairs with an adjusted p-value < 0.05 were kept in the analysis. The underlying distribution's mean was used to measure the dependency between gene A and gene B. Additionally, all dependencies were excluded, with less than 70 percent of all KOs showing the same sign as the mean dependency.

Calculation for the dependency of gene A on gene B. The dependency of gene B on gene A would have the opposite sign.

$$\text{dependency}(\text{gene A to gene B}) = \text{mean}(\text{WT completions}_{\text{gene A}} - \text{WT completions}_{\text{gene B}})$$

A positive dependency between A and B shows that gene A consistently shows higher WT completion than gene B. A negative dependency between A and B shows that gene B consistently shows higher WT completion than gene A. As the dependency A to B and the dependency B to A share the underlying distribution, they have the same value with opposing signs.

The analysis was repeated after excluding KOs with weak or no differentiation delay phenotype. The results based on the remaining 58 KOs were used for further analysis. Non-negative matrix factorization (NMF) (R package NMF[253] version 0.26) was applied to the gene-gene dependency matrix of the 1203 genes to cluster genes into groups with similar dependencies to other genes. We grouped the genes into six signatures, and each gene was assigned a weight to all six signatures. The genes were then grouped into the signature with the highest weight. Thus, we binarized the NMF weights to receive six clusters of genes.

4.3 Independent *in vitro* single-cell data

The sc data covers the naïve to primed transition of WT mESCs. The cells were sampled at five different time points: 0 hours of differentiation, 6 hours of differentiation, 12 hours of differentiation, 24 hours of differentiation, and 48 hours of differentiation. The data provided to us additionally included the labels “Doublet” and “Negative” for droplets that were classified as doublets and cells where demultiplexing did not work.

Cells were excluded from the analysis if less than 200 genes were measured, more than 3,500 genes were measured, more than 20 percent of the reads came from mitochondrial genes, or they were marked as a doublet. The counts were normalized using the SCTransform() function (from the R package Seurat[254] version 4.4.0), accounting for the number of genes measured, the percentage of mitochondrial reads, and the library size after excluding mitochondrial genes. Principal component

analysis (PCA) and t-distributed stochastic neighbor embedding[255] (tSNE) were calculated with the corresponding functions from the Seurat package, and the first 20 PCs were used as input for tSNE.

4.3.1 Dependencies between genes

The cells from the 48-hour time point and cells marked as negative were excluded from further analysis, and all cells from 0 to 24 hours were assigned pseudo time using the psupertime package[256] (version 0.2.6). Before calculating the gene regulatory dependencies in the SC data, genes were further filtered for genes that showed the same directional change between 6 hours, 12 hours, and 48 hours compared to the 0-hour time point. Additionally, genes were sorted out if this direction was the opposite as observed in the 32-hour time course. The remaining genes were overlapped with the 1203 genes from the KO-based regulatory dependencies and only genes in both sets were kept. The overlapping 383 genes were then tested for their gene-gene dependencies.

The first step in deriving gene regulatory dependencies from the sc data is to rank both genes from early to late. For genes that are downregulated from 0 hours to 24 hours, early ranks will correspond to high expression and late ranks to low expression. For genes that are upregulated from 0 hours to 24 hours, early ranks will correspond to low expression and late ranks to high expression. If ranks are tied, the pseudotime is used to resolve ties. Here, the earliest pseudotime will be assigned the earliest rank. Next, the rank difference for each cell is calculated.

Calculation of the size of the buffer zone. What time point or time points the cells that are considered in the calculation come from depends on the kinetics of both genes.

$$\text{buffer zone} = 1.1 * \max(\text{rank difference of all cells with count } (0,0)_{\text{time point}})$$

$$\text{time point} = \begin{cases} 24 \text{ hours if both genes are upregulated} \\ 0 \text{ hours if both genes are downregulated} \\ 0 \text{ hours and } 24 \text{ hours if one gene is upregulated and one downregulated} \end{cases}$$

A buffer zone was calculated to avoid overinterpreting rank differences from cells that are likely based on technical dropouts. The extent of this buffer zone depends on the largest rank difference in cells that have no measurement for both genes at a time point where at least one gene is expected to show high expression. If both genes were upregulated, all cells without measurements from the 24-hour time point were considered; if both genes were downregulated, all cells without measurements from the 0-hour time point were considered; and if one gene was up- and the other gene was downregulated, all cells without measurements from the 0 and the 24-hour time point were considered.

The mean rank difference of all cells outside the buffer zone was calculated and scaled by the number of cells outside the buffer zone. This mean corresponds to the gene-gene dependency of the two genes. A binomial test for the number of cells above and beneath the buffer zone was performed to test if a significant imbalance

was observed between the number of cells on each side of the buffer zone. The resulting p-values were then adjusted for multiple testing, according to Benjamini and Hochberg[252]. The resulting dependencies were only kept with an adjusted p-value < 0.05 and at least 600 cells outside of the buffer zone.

4.4 Dependency network

Gene regulatory dependencies between the KO-based and the sc-based approaches were compared. If significant dependencies shared the same signs between both approaches, they were considered to be consistent. If the significant dependencies had opposing signs between both approaches, they were considered to contradict. Subsequently, we filtered the gene regulatory dependencies from the KO-based approach by the contradicting and consistent dependencies. Here, the filter based on contradiction is less stringent because dependencies can only contradict if the genes are measurable in the sc data and the dependencies are significant. The filter based on consistency requires significant dependencies in the sc data and the same sign as the KO-based dependency, making the filter more restrictive.

Clusters were taken from the NMF-based clustering on the KO-derived dependencies. The mean dependencies between clusters were calculated using all dependencies from genes of cluster A to genes of cluster B, including zeros. Clusters were visualized as a network. All mean dependencies were viewed as an edge between two clusters for that purpose. Only positive mean dependencies were used, as each edge can be described either by a positive or negative dependency. A positive dependency from cluster A to cluster B results in a directed edge from cluster A to cluster B. Additionally, edges that can be described as a combination of edges were removed from the network.

4.4.1 Processing of differentiation blocking double knockout

The last four bases of the quant-seq data were cut due to low sequence quality. The data was then aligned to the mm10 genome and read counts were obtained using the STAR aligner[246] (version 2.7.3a). Gene expression for the following biotypes was quantified: protein_coding, lincRNA, processed_transcript, antisense, 3prime_overlapping_ncRNA, bidirectional_promoter_lncRNA, macro_lncRNA, miRNA, misc_RNA, lncRNA, scaRNA, scRNA, sense_intronic, sense_overlapping, snoRNA, snRNA, and sRNA. Counts of the remaining genes were integrated into DESeq2[248] (version 1.40.2) and log2FCs, and adjusted p-values for the differences between dKO samples and WT in 2i were obtained. Log2FCs and adjusted p-values were used for visualizations in combination with medium change results.

4.5 Potential upstream regulators of formative network

ATAC-seq peaks called from Kinoshita et al.[257] were merged over eleven replicates. Five replicates cultured in 2i LIF and six replicates of EpiLCs were merged into a common reference of open chromatin using bedops[258] (version 2.4.40). All regions covered by at least three replicates of any condition were extracted using bedops and considered open chromatin in the naïve to formative transition.

Motifs from CIS-BPs[259] database were converted to uniprobe format, and Transfac motifs were excluded from the analysis. The FIMO algorithm[260] (MEME suite version 5.5.0) was used to test for the occurrence of motifs in the open chromatin from Kinoshita et al.[257]. Positions of open chromatin enriched for a TF binding motif were then assigned to potential target genes using the "ClosestGene" method from TFTargetCaller[261] (version 0.7). The resulting TF target gene pairs were filtered for a q-value < 0.05. Additionally, TFs were selected for FPKM values >= 1 in the WT samples of the KO data.

Target genes were limited to an extended set of formative associated genes. Motif sequences of remaining TFs were clustered using the compare_motifs() function (R package universalmotif[262] version 1.18.1) with Euclidean distance. The formative markers *Otx2*, *Pou3f1*, *Dnmt3a*, *Dnmt3b*, and *Fgf5* were extended in parallel to the naïve associated genes in Lackner et al.[251] using results from the KO experiment. Multiple regression analysis was performed to quantify the similarity of each gene's expression after 24 hours of differentiation to the expression of the formative markers. Formative extended genes were defined as having a R^2 >= 0.65.

4.6 Visualization

Schematic figures of methods and models were done in Adobe Illustrator (version 28.3). Results were visualized using different R-packages. Heatmaps were mostly created using the ComplexHeatmap R-package[263] (version 2.16.0) with the exception of the motif comparison heatmap, which was created using the heatmap2() function from gplots[264] (version 3.1.3). The corrplot package[265] (version 0.92) was used to compare correlations and overlaps between sets of differentially expressed genes (DEGs) were visualized using the UpSetR package[266] (version1.4.0). All other visualizations were done using the ggplot2 R-package[267] (version 3.4.4).

5 Results

5.1 Wildtype differentiation time course

Whereas the different states of pluripotency are relatively well described, the transition between the stages still remains inadequately understood. RNA-seq data of a dense differentiation time course and RNA-seq data from 73 KO cell lines were analyzed to understand these transitions better, more precisely the transition from naïve to formative pluripotency. The dense differentiation time course was carried out to assess changes in the expression of single genes during the naïve to formative transition and to allow the quantification of differentiation delay in the KO cell lines.

The time course covers gene expression from the naïve state in 2i medium to 32 hours of differentiation in N2B27 medium in steps of two hours. Since 2i medium blocks the differentiation of cells, measurements in 2i correspond to 0 hours of differentiation. Formative pluripotency is established after approximately 24 hours and thus covered by the data. While RNA-seq measurements at 0 and 32 hours were done in duplicates, all time points from 2 to 30 hours were single RNA-seq measurements, allowing the expression measurements to be closer. Due to the very dense nature of the time course measurements, each time point also carried information about neighboring time points. The dense measurements compensated for the lack of duplicates for all time points but 0 and 32 hours. In addition to time points from 0 to 32 hours of differentiation, a duplicate of samples in 2i LIF medium was measured. These samples include LIF as an additional inhibitor to the two inhibitors Chiron and PD, thus resulting in more restrictive conditions than 2i.

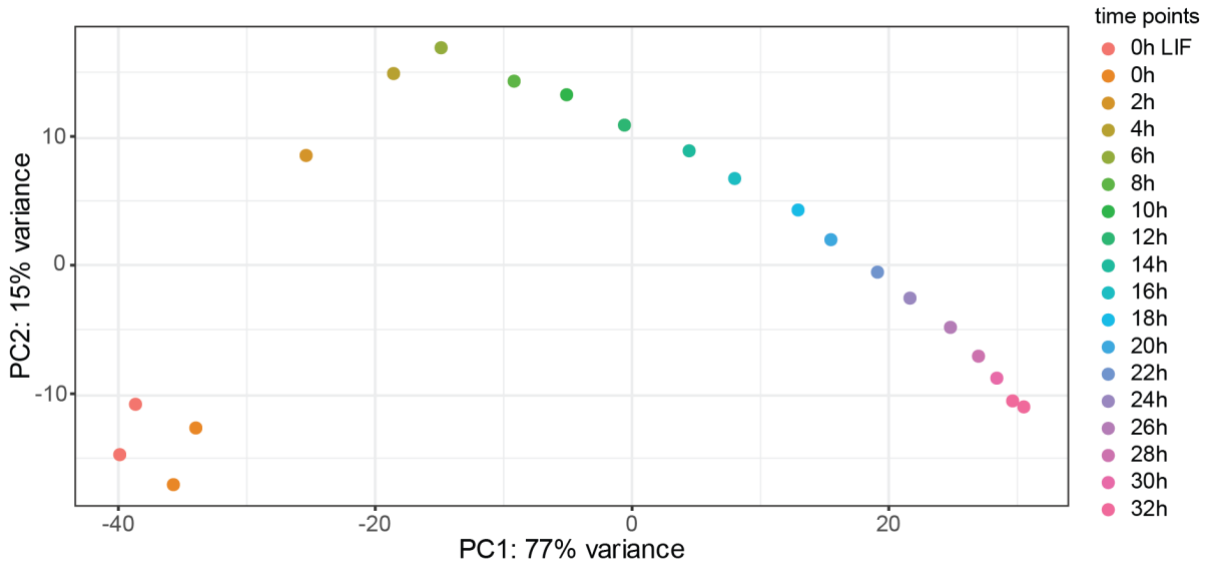

Figure 5.1: PCA plot of samples from the differentiation time course.

Time points are labeled by color. Orange and yellow corresponds to early time points; green and blue to time points in the middle; purple and red to late time points.

After applying variance stabilizing transformation, PCA separated samples by time in the first component (Figure 5.1). Negative values of the first principal component represented samples from early time points, whereas later time points were assigned positive values. According to the first principal component, the most extreme samples were from 32 hours of differentiation on the upper end and 2i LIF on the lower end. The second most extreme samples on the lower end were the 2i samples. This reflects the more restrictive medium conditions of 2i LIF compared to 2i. While the first component captured the differentiation time axis from 2i LIF and 2i to 32 hours in the correct order, the second component did not. Here, a quick increase in values from 2i LIF and 2i samples to 2 hours was observed. Up to 6 hours, the samples moved further from the 2i samples in the second component, followed by a decrease in values to 32 hours. At 32 hours, the samples were almost at the same level as 2i samples regarding the second component.

Visualizing shrunken log2FCs obtained using the apeglm() algorithm[247], integrated in DESeq2[248], depicted the different effect of noise on different genes. Even though the overall change of expression for naïve and formative genes like *Nanog*, *Tfcp2l1*, *Otx2*, and *Dnmt3a* was relatively constant (Figure 5.2), there were differences in the noise of the measurements. While log2FCs for *Otx2* were relatively consistent in their directionality compared to neighboring time points, *Tcfp2l1* showed a steep drop in log2FC from 4 to 6 hours, followed by a steep increase to 8 hours. This could either be the kinetics of the gene adapting to differentiation or a noisier measurement at 6 hours. Since the time course setup did not include replicates for all time points from 2 to 30 hours, this cannot be distinguished.

Since RNA-seq data can be noisy, especially when dealing with lower count numbers, it is advisable to use biological or at least technical replicates when sequencing different states. For the time course, we prioritized a dense coverage of time points over measuring duplicates or triplicates with two or three times the distance between time points. While this did not allow for an estimate of variability or noisiness of the data between 2 and 30 hours directly, neighboring time points could add insight. Here, GPR was used to smoothen the time course and incorporate information from neighboring time points into the outcome at each time point. Resulting representations of the expression change after GPR (Figure 5.2) nicely smoothed the noise from single measurements for genes that showed little noise, such as *Nanog* or *Otx2*. Expression changes of genes with more noise or stronger fluctuations between neighboring time points, such as *Tfcp2l1*, were also represented by a smooth function after applying GPR. While the method smoothed out sharp differences caused by a single time point, as observed in *Tfcp2l1*, general trends like a sigmoid expression change in *Nanog* were still preserved after smoothing.

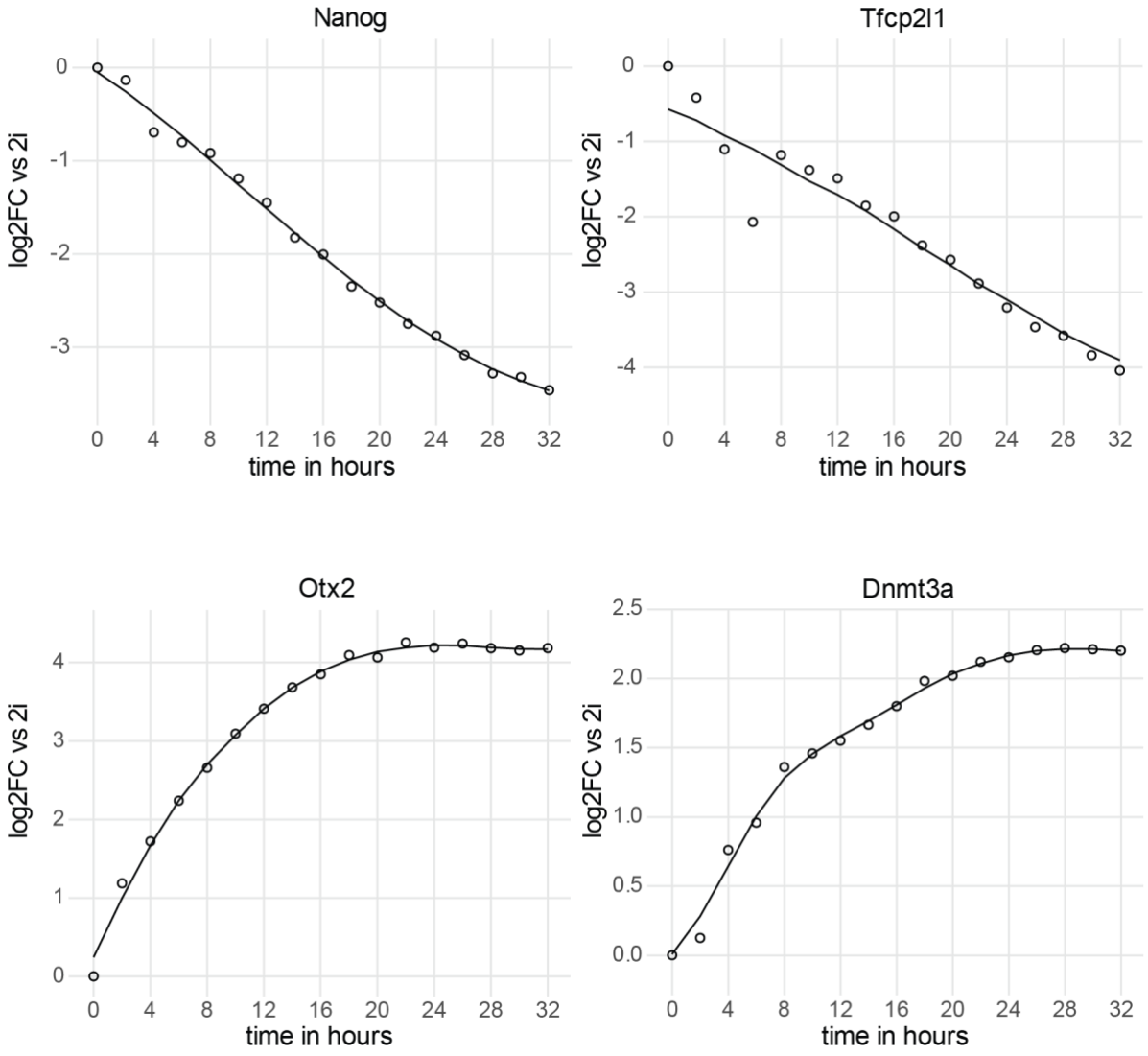

Figure 5.2: Expression changes of naïve and formative markers in WT time course.

Log2FCs vs. 2i from naive marker genes Nanog *and* Tfcp2l1 *are in the top row, and formative marker genes* Otx2 *and* Dnmt3a *are in the bottom row. Circles represent shrunken log2FCs vs. 2i, and the line represents resulting log2FCs vs. 2i from GPR.*

Overall, the time course is a valuable recourse for investigating transcriptional changes from pre- to post-implantation of mESCs as the first principal component nicely separated the samples by the time of differentiation. The decision to sample every two hours without replicates instead of duplicates or triplicates with longer distances between time points did not allow us to estimate the variability of measurements per time point directly. However, applying GPR to the time course helped estimate variability based on neighboring time points and helped smooth the expression changes over the time course accordingly. While single outlier measurements and overall noise seemed to be smoothed out, trends captured by more time points were smoothed but still preserved.

5.1.1 Kinetics of changed genes over wildtype time course

A first step to enhance our understanding of transcriptional changes in the transition from naïve to formative pluripotency is information on expression kinetics on the gene level. While the time course already provided this information after using GPR, we aimed to extend this knowledge through a systematic investigation of changes in gene expression. More precisely, we sought to pinpoint the timing of gene regulation during this transition, i.e., when genes are induced or repressed from naïve to formative pluripotency.

First, log2FCs vs. 0 hours were z-transformed and clustered to define different adaption patterns to differentiation. Z-transformation was used as the amplitude of fold change was not of interest in this first step. Here, the primary interest was to group genes by different kinetics rather than intensities of expression changes. Hierarchical clustering of transformed expression changes led to 9 clusters (Figure 5.3) of different types of kinetics.

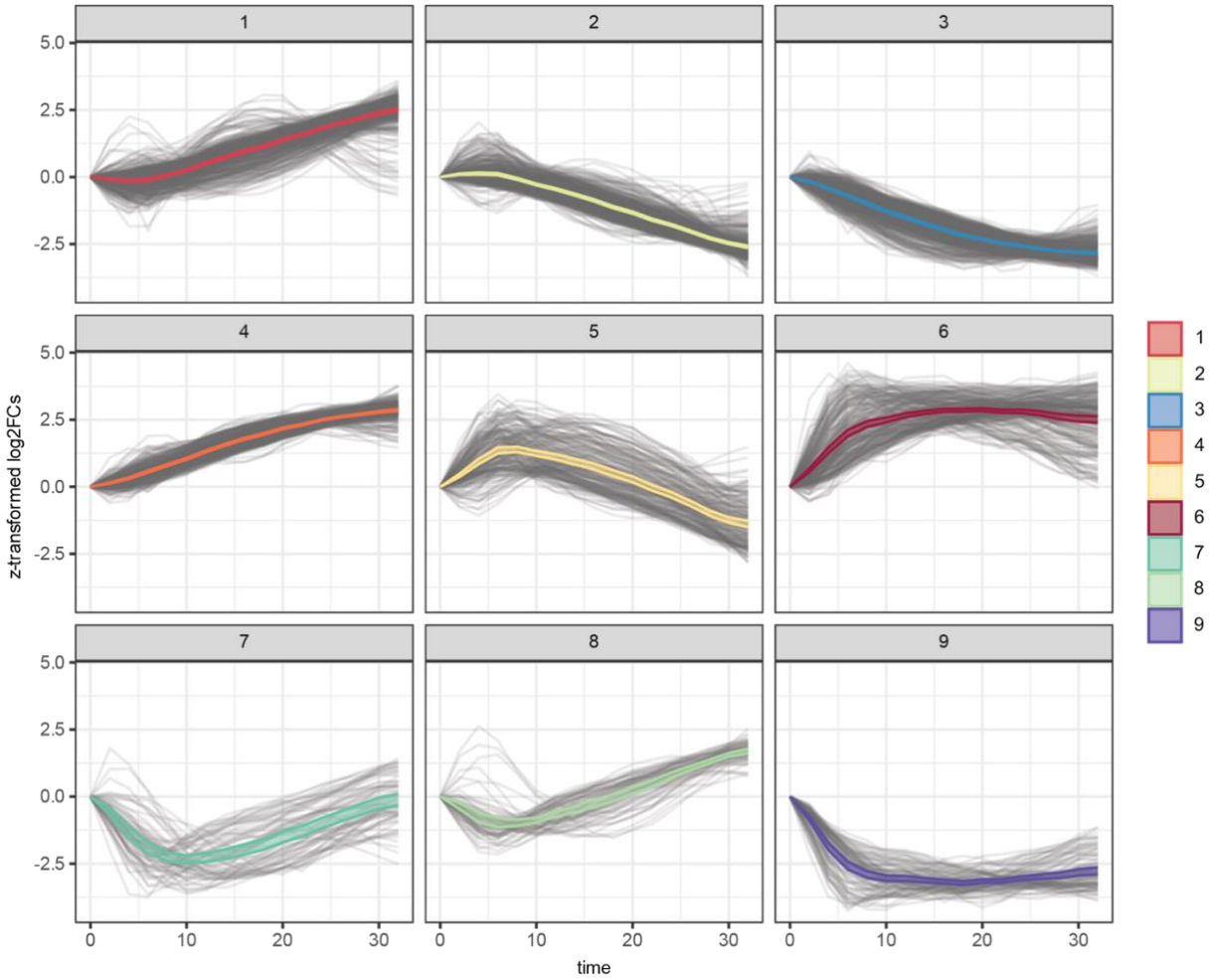

Figure 5.3: Clusters of different kinetics in WT time course.

Z-transformed log2FCs of genes that change expression over time clustered into 9 clusters. The mean changes of each cluster are shown in color, all individual genes are shown in grey.

The first four clusters contained genes that exhibited relatively constant expression gain (clusters 1 and 4) or loss (clusters 2 and 3) during the naïve to formative transition. These four clusters were the largest regarding the number of genes in them. Cluster 6 represented genes that followed a very sharp increase in expression in the first 4 to 6 hours of differentiation, followed by constant expression or a slight decrease toward the naïve expression level. Clusters 5, 7, and 8 showed an initial increase (5) or decrease (7 and 8) in the first 4 to 6 hours of differentiation, followed by a reversal in the other direction.

Instead of applying unsupervised clustering to group genes by expression behavior and inferring patterns of change, regulation points can alternatively be defined directly. To identify induction and repression time points for different genes, we first must clearly define an induction or repression time point. In this context, we defined that an induction time point follows either a loss or minimal gain of expression and is followed by a stronger gain of expression. A repression time point, by that definition, is a point that follows either a gain or very little loss of expression and is followed by a stronger loss of expression. We further extended these definitions to be helpful for the time course data.

First, as there is no previous time point before the 0-hour time point, it cannot follow other time points and, thus, cannot fulfill the criterion of following minimal gain or loss of expression. Therefore, we assumed constant expression in 2i so that expression gain or loss of a particular strength from 0 to 2 hours was considered a point of induction or repression, respectively. Second, we specified that a point of induction cannot follow a point of induction. Likewise, a point of repression cannot follow a point of repression. Even though the expression of a gene might be induced first, plateau shortly in the expression change, and then gain expression again, our interest was in those time points where the overall direction of regulation changes. Thus, we only chose the first point of regulation if a point of regulation of the same type followed.

Two different approaches to identify points of induction and repression were applied to the log2FC profiles according to the methods (4.1.1). The decision on which approach to use depended on the first and second derivatives of log2FC profiles between neighboring time points of a gene. The first approach was applied for genes with a relatively constant change of expression and a high signal-to-noise ratio. If this approach identified more than two points of regulation for a given, the second approach was used instead. The second approach was worse at detecting induction or repression at 0 hours when the expression constantly changed over time. However, it is better at handling noisier data than the first approach.

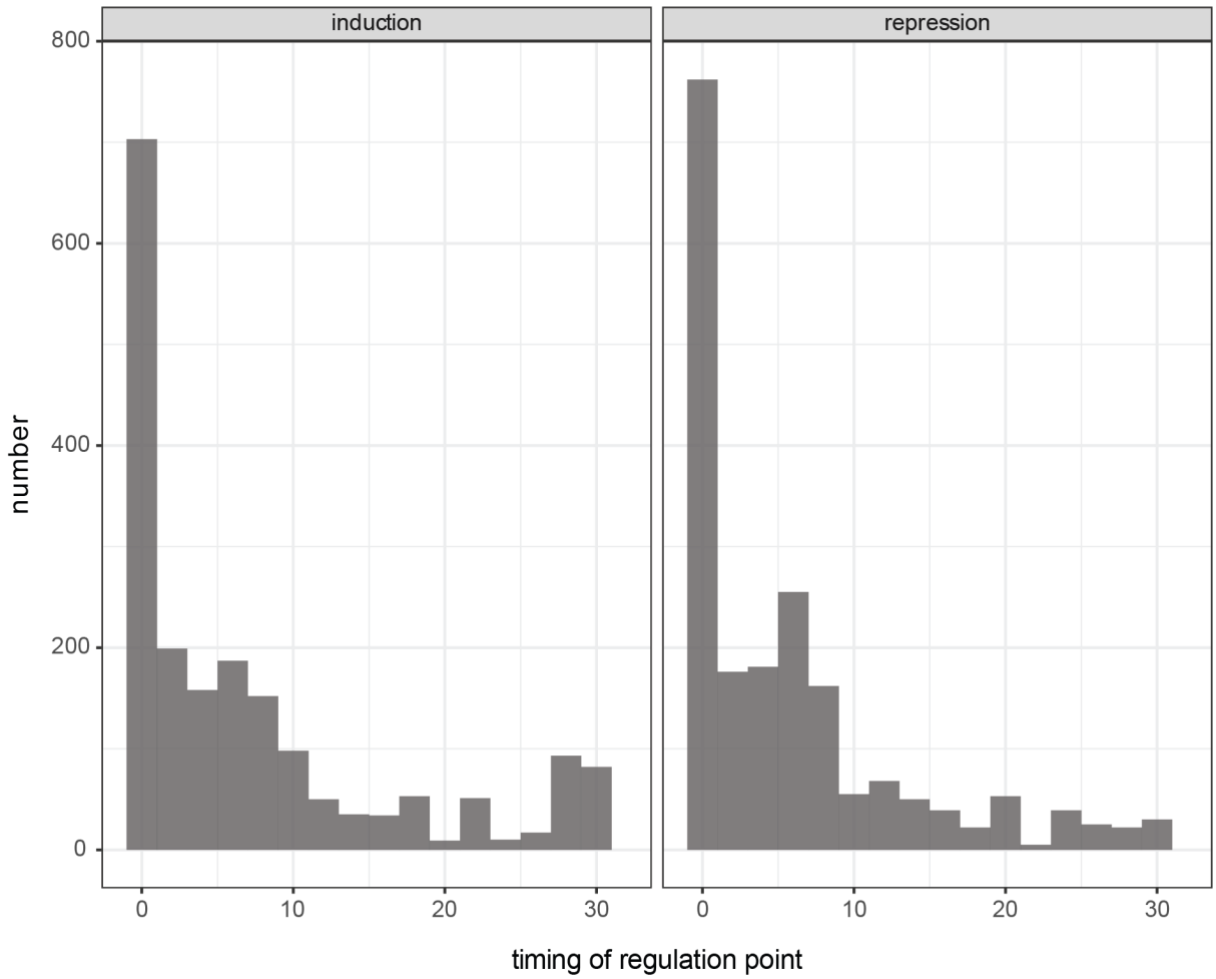

Figure 5.4: Histograms showing regulation of genes during the WT time course.

Points of induction are in the left histogram, and points of repression are on the right. A gene can have up to three points of induction or repression.

Histograms for induction and repression points (Figure 5.4) revealed two major time points of regulation during early differentiation. The time point with the most points of regulation assigned is 0 hours for induction and repression. The other time point that showed much regulation was around 6 hours of differentiation. While the 0-hour time point separated itself from the following time points in terms of the absolute number of regulation points assigned, the regulation time frame at approximately 6 hours was much broader, spanning from 2 to 10 hours. After 10 hours, time points of regulation became less frequent and only were more frequent towards 24 hours and later again. This fits the transition the cells are going through as cells at 24 hours represent cells that have transitioned from naïve to formative pluripotency. From there, additional rewiring is needed to transition towards primed pluripotency. The time course, however, does not cover 48 hours, approximately corresponding to primed pluripotency.

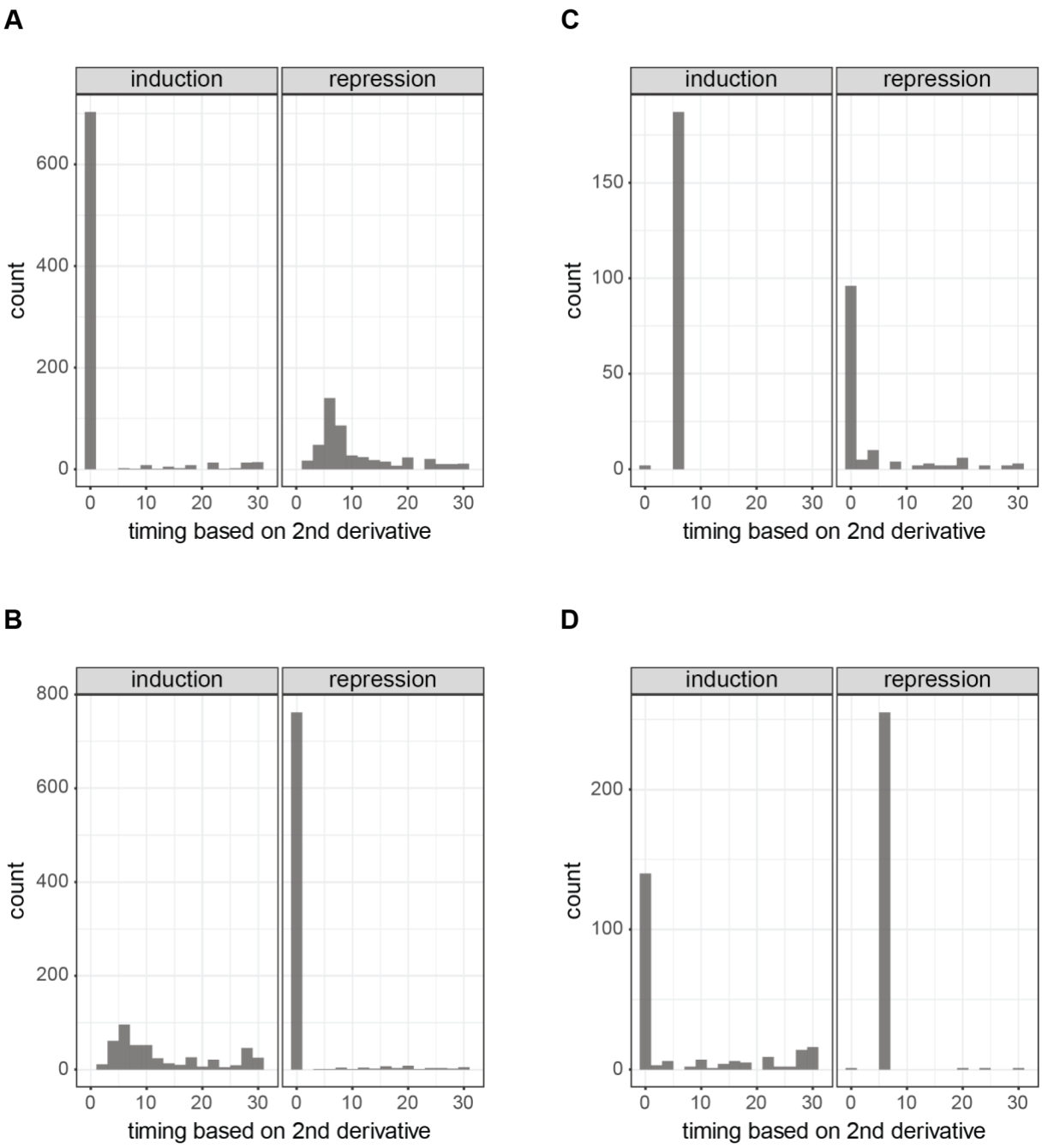

Figure 5.5: Induction and repression timings for different groups of genes.

A) Histograms of points of induction (left) and repression (right) for genes that were assigned a point of induction at 0 hours. B) Histograms of points of induction (left) and repression (right) for genes that were assigned a point of repression at 0 hours. C) Histograms of points of induction (left) and repression (right) for genes that were assigned a point of induction at 6 hours. D) Histograms of points of induction (left) and repression (right) for genes that were assigned a point of repression at 6 hours.

Investigation of genes induced (Figure 5.5 A) or repressed (Figure 5.5 B) at the beginning of the time course revealed mainly two trajectories. These genes either mostly remained in that regulatory trajectory for the rest of the time course or were exposed to a reversal of that initial trajectory around 6 hours. More strikingly, more than half of the genes induced (Figure 5.5 C) or repressed (Figure 5.5 D) at 6 hours exhibited opposite regulation at 0 hours. This showed that the regulation around 6

hours mainly was a reversal of initial adaption to differentiation cues, and about a third of initial regulation was reversed around 6 hours.

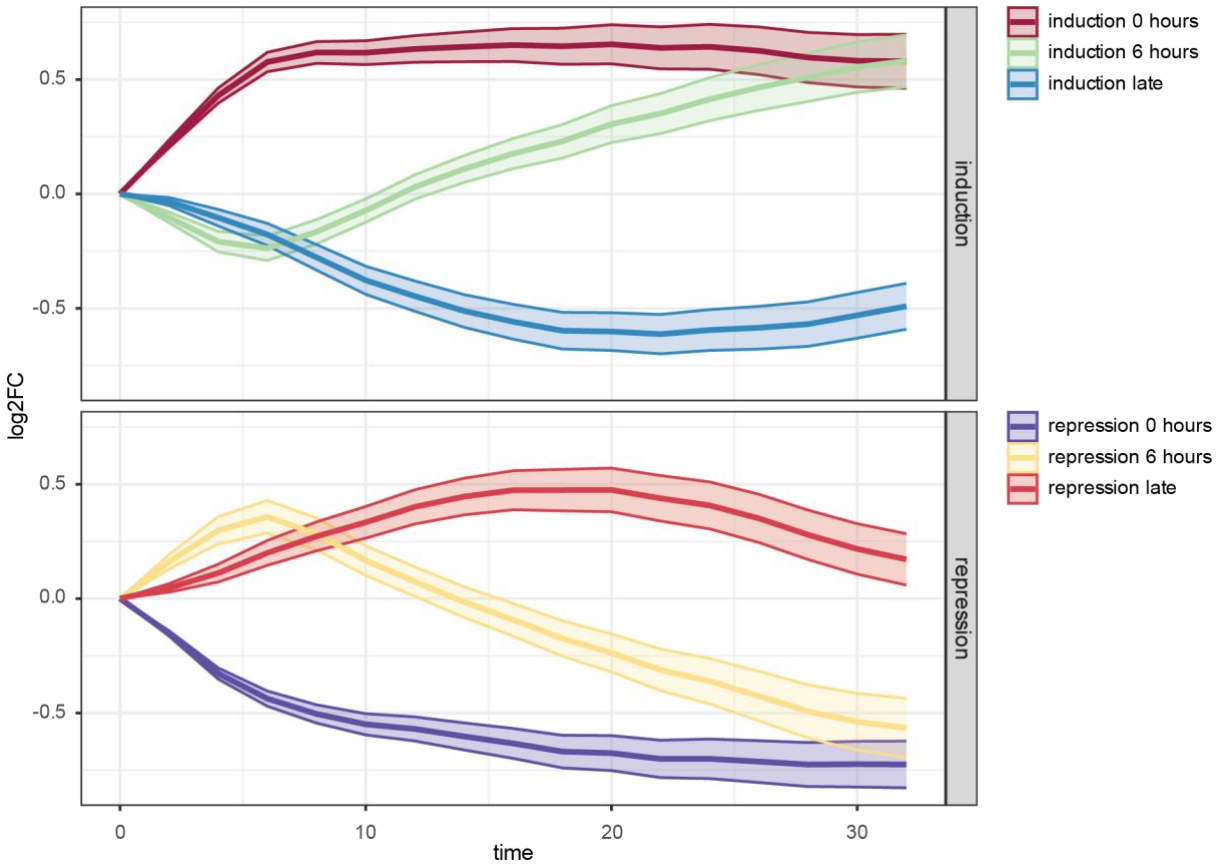

Figure 5.6: Mean changes of expression for groups of genes based on time of regulation.

Log2FCs of different groups of genes depending on when genes were induced or repressed during the time course. The color indicates the group a gene belongs to based on the time point of regulation.

These trends remained when plotting mean z-transformed log2FCs of groups of genes depending on specific time points of regulation (Figure 5.6). Here, the green and yellow ribbons represent the mean expression changes of genes induced or repressed at 6 hours, respectively. These groups clearly showed the trend of initial regulation in one direction with a reversal in the other direction at 6 hours. Initially regulated genes (dark red for initial induction; purple for initial repression), on average, showed an initial increase or decrease in expression, which plateaued at approximately 6 hours. This resulted from the later regulation of these genes, as described above. Most of the genes in this group were either monotonically up or downregulated or showed a trajectory reversal around 6 hours. On average, this led to a plateau after 6 hours in these groups. Both groups of genes defined by induction or repression at 6 hours (green and yellow, respectively) showed an initial regulation in the opposite direction. Genes induced late (blue) or repressed late (light red) were also initially regulated in the opposite direction and then showed reversal after 20 hours. This might relate to the following transition to primed pluripotency.

When comparing the groups from the supervised and unsupervised approach, groups of genes seemed inconsistent between approaches initially. The genes induced at 0 hours in the supervised approach primarily belonged to clusters 6, 4,

and 5 in the unsupervised approach. However, this is because the supervised approach did not restrict one gene to one group, as only one specific regulation point defined the groups. In this case, this was the induction at 0h, and genes in this group are a mixture of genes that monotonically gained expression, hit a plateau later, or decreased expression at later time points. Thereby, genes induced at 0h in the supervised approach were in clusters 6 (monotonic up), 4 (monotonic up), 5 (initial upregulation with reversal), and cluster 2 (initially little change followed by downregulation) in the unsupervised approach. While cluster 2 lost expression on average, it also contained genes that showed initial upregulation. Other groups, such as genes repressed at 6 hours to clusters 2 and 5, had a more intuitive mapping from the supervised to the unsupervised groups.

The genes that changed expression in our WT differentiation time course were grouped by their kinetics in two different ways. One focused more on the overall change over the whole time point, while the other focused on behavior at specific time points. As both approaches had strengths and weaknesses, the choice of approach depends on the question asked. While the supervised approach led to groups with a very defined behavior at one specific time point, it was blind to differences at other time points in this group. In the unsupervised approach, however, genes were grouped by somewhat consistent behavior over time, with the weakness that genes in the groups on specific time points might behave differently.

5.1.2 Initial regulation does not depend on the process of changing the medium

Since most genes showed their first point of regulation at 0 hours, we next asked what could be causing the initial regulation aside from the differentiation of the cells. One factor that could have caused such behavior could be removing the old medium and adding fresh medium. This switch must be part of the experiment to remove the mESCs from differentiation-blocking to differentiation-permitting conditions.

An additional experiment was carried out to determine whether the medium change affects gene expression and might be the cause for the points of regulation at 0 hours. The medium was switched from 2i to N2B27 and from 2i to 2i. Switching to 2i represented the isolated effect of removing and adding medium again while switching to N2B27 represented the same setup as in the differentiation time course. Comparison of the 2i samples after the switch to 2i samples before the switch allowed us to observe the pure effect of removing and adding fresh.

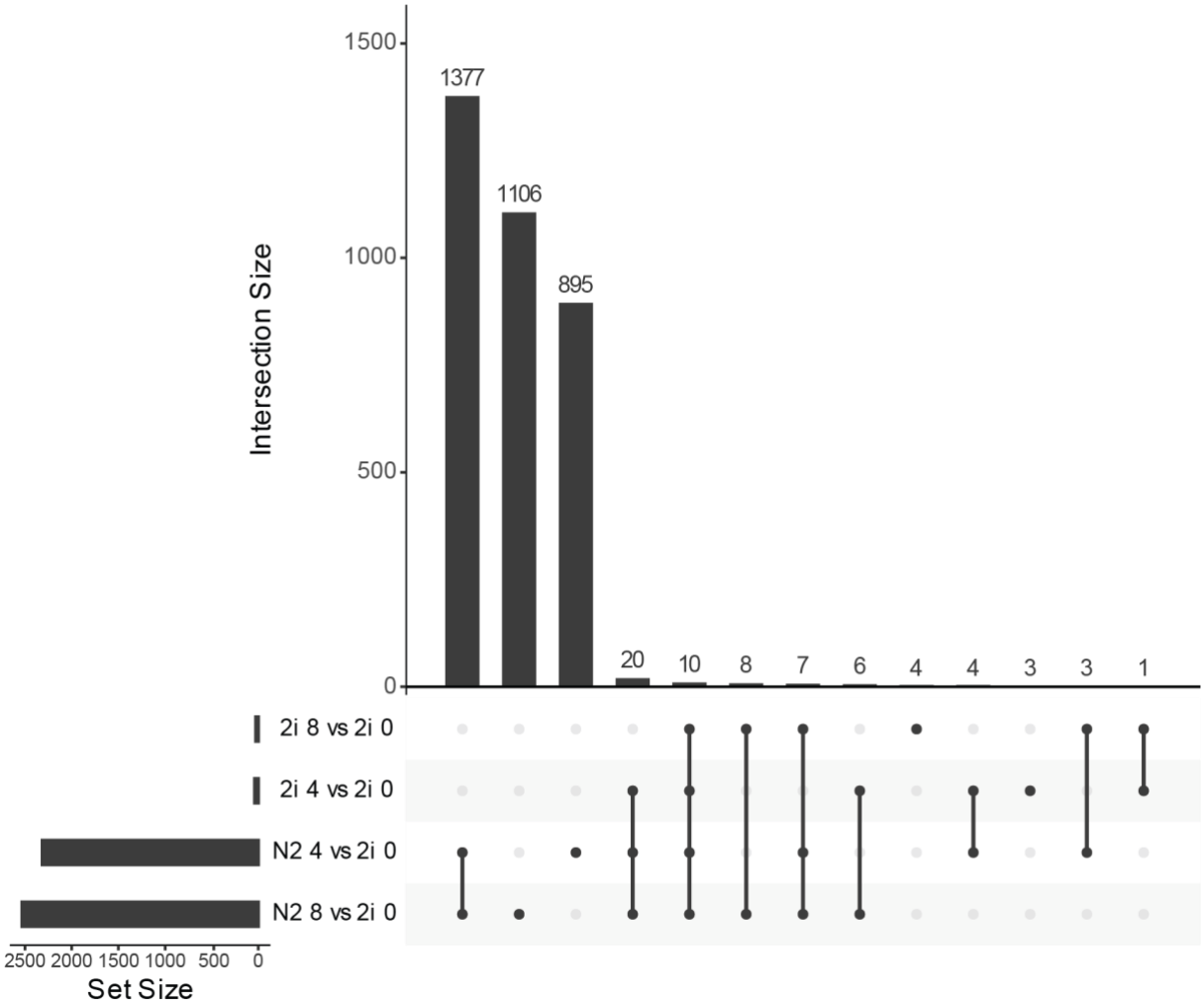

Figure 5.7: Upset plot for differentially expressed genes for medium change.

The number of differentially expressed genes in cells in 2i or N2B27 4 and 8 hours after the medium change from 2i. Overlaps are indicated by connecting lines in the lower part of the plot.

Performing differential expression analysis on the medium switches from 2i to 2i and 2i to N2B27 revealed that only a few genes changed significantly (absolute log2FC >= 0.5 and padj <= 0.05) in the 2i to 2i comparisons. 4 hours after the medium switch from 2i to 2i, 44 genes showed differential expression and 8 hours after the switch, that number decreased to 33. Of those genes, only 11 were shared between 4 and 8 hours after the change (Figure 5.7). While the switch from 2i to 2i did not result in many genes changing expression significantly, the switch to N2B27 resulted in 2,316 and 2,534 genes being differentially expressed after 4 and 8 hours, respectively. Here, 1,414 genes were shared between 4 hours of adaptation and 8 hours of adaptation to the new medium. These genes represent the adaptation of change to a different medium and the adaptation to differentiation cues as corresponding pathways are not blocked by 2i anymore.

The low number of DEGs between 2i samples before and after the medium switch demonstrated that the initial regulation of genes (5.1.1) was not caused by medium withdrawal and adding fresh medium. The high number of DEGs between 2i before the medium switch and samples differentiated in N2B27 medium indicated that the initial regulation was either caused by adaptation to a different medium or adaptation to differentiation.

5.1.3 Timing of knockouts based on naïve markers is disconnected from timing based on differentially expressed genes

In previous work[251], we used an insertional mutagenesis screen in haploid ESCs to identify potential genes and mechanisms that interfere with the differentiation of naïve stem cells. The cell lines in use contained a GFP-tag for the *Rex1* gene, which is downregulated when cells start to differentiate. If the loss of fluorescence in cells with mutations was lower than in the unaffected control cells, the mutated gene was a potential candidate to play an important role in differentiation. Through this screening approach, a set of 73 different KO target genes were identified, and KO cell lines of these genes were established using CRISPR-Cas9. The set of 73 KO genes included candidate genes from the screen and complex members of those candidates or positive controls like *Myc*. Even though the GFP-tagged cell lines allowed for identifying cells that differentiate slower than expected through the *Rex1* labeling, it did not allow quantifying the differentiation delay.

We used the differentiation time course to put the 73 KO cell lines on a WT differentiation time axis. This allowed quantifying the delay of differentiation based on different subsets of genes or networks of genes. The first gene set used was a core set of naïve marker genes to allow assessment of the delay of the naïve transcriptional network. The second group was a core set of formative marker genes to assess the delay of the formative transcription network. The last group of genes used here was the 3,068 genes showing differential expression (absolute log2FC >= 0.5 and padj <= 0.05) between WT cells at 24 hours of differentiation and in 2i medium. The set of DEGs helped estimate an overall delay of all genes adjusted in the initial transition from naïve to formative pluripotency.

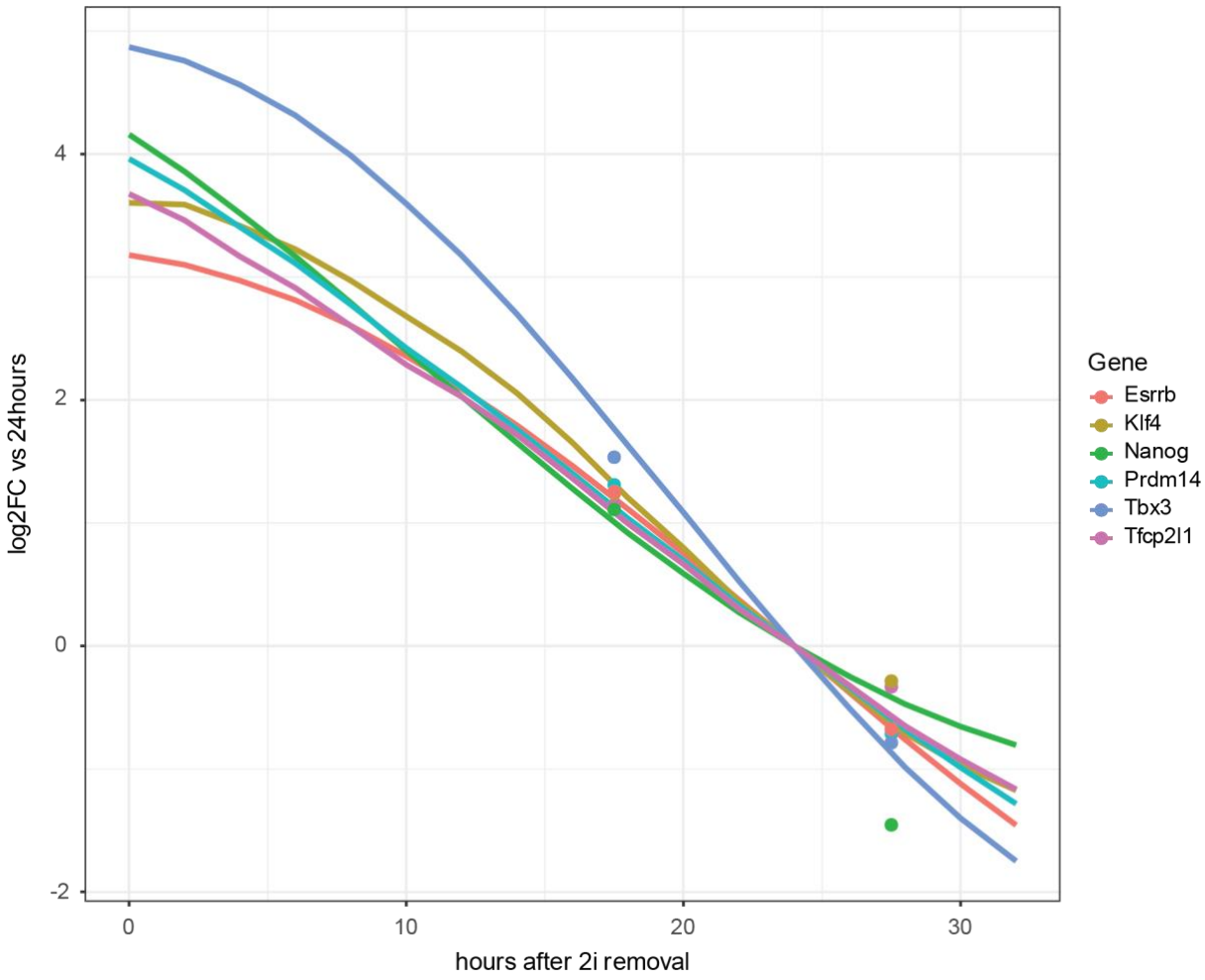

Figure 5.8: Example of identification of time point that best represents KOs using Pten and Myc.

Lines represent expression changes of naïve markers over the WT 2-hour time course (in relation to 24 hours). Points represent the corresponding KO vs. WT changes from the KO data at 24 hours. Lines and points are colored by the corresponding marker genes.

In order to put the 73 KO cell lines on the WT time course, the time course was represented in log2FCs vs. 24 hours. This allowed comparing the log2FC profile of each KO at 24 hours vs. WT at 24 hours with the log2FC profile of each time point vs. 24 hours in the WT differentiation time course. Additionally, log2FCs of the time course were scaled depending on the differences in 2i to 24 hours comparison in the KO experiment and the WT time course. Finally, the most similar time point defined the differentiation delay by subtracting the best-fitting time point from 24 hours. Therefore, a KO that did fit the expression profile of the WT at 20 hours had a delay of 4 hours. A KO best represented by the WT profile at 28 hours, on the other hand, had a delay of -4 hours or, in other words, progressed 4 hours further than in WT. We applied this procedure for expression change profiles of all KOs for the three different gene sets. For example, putting the *Pten* and the *Myc* KO on the WT time axis (Figure 5.8) led to delays of 6.5 hours and -3.5 hours, respectively. *Myc* was one of the control genes added to the 73 KO genes because it accelerates differentiation when knocked out.

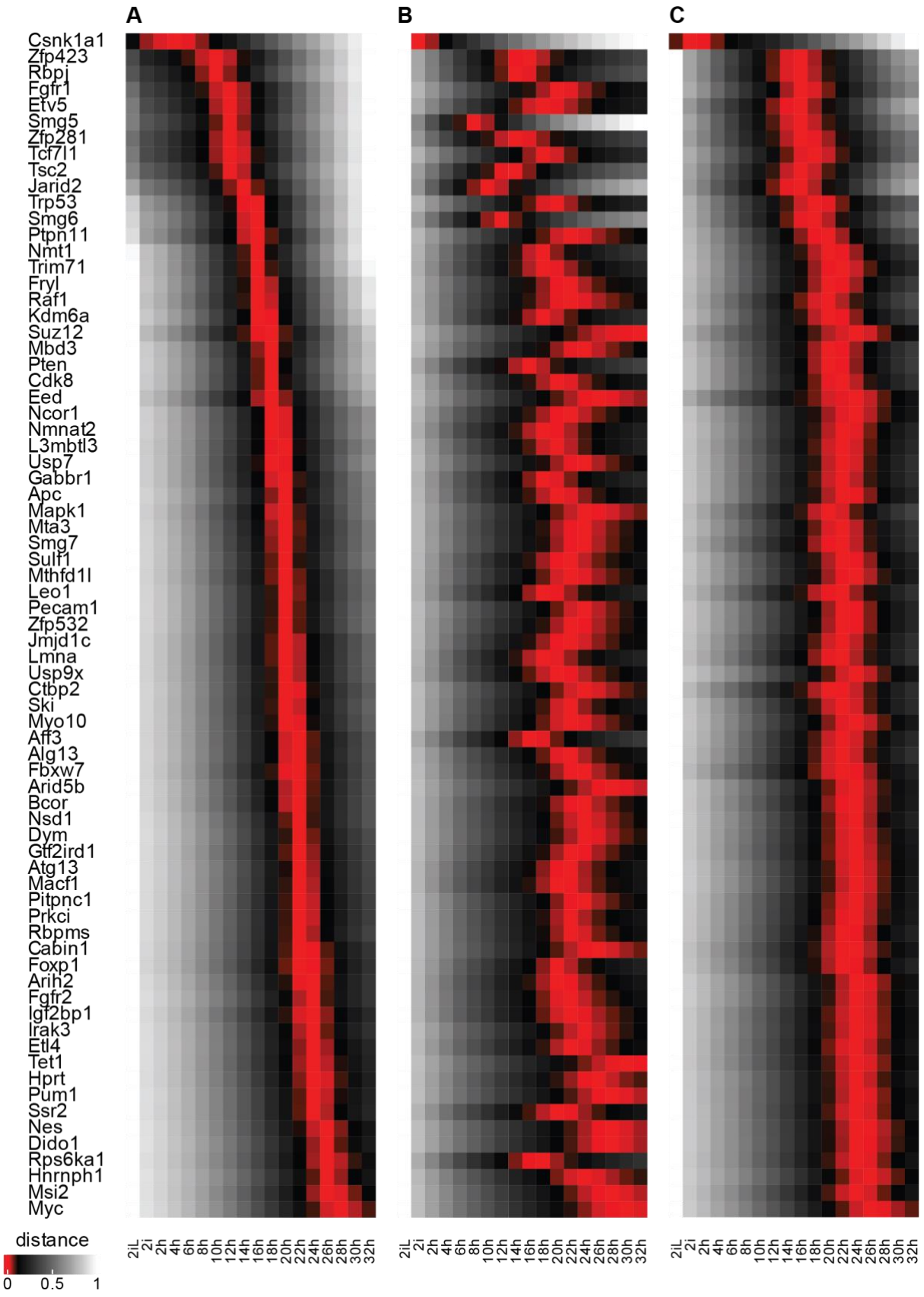

Figure 5.9: Timing of all KOs based on naïve marker genes (A), formative marker genes (B), and DEGs (C).

The color encodes for the relative distance of the KO vs. WT profile to each time point. White indicated the highest relative distance, and red the lowest relative distance. Each heatmap represents the best timing for a specific gene set. Naïve marker genes in A, formative marker genes in B, and all DEGs for C. The row order of all Heatmaps is based on the timings based on naïve markers.

When the 73 KO cell lines were put on the differentiation axis of naïve marker genes, they covered a broad span of differentiation delays from 19.75 to -3.5 hours (Figure 5.9 A). *Csnk1a1* was the gene leading to the strongest delay in differentiation, almost showing no progress along the naïve marker differentiation axis. While the differentiation delay was also observable at the cell culture level, this KO regained differentiation potential when kept in continuous culture[251]. As mentioned above, *Myc*, included as a positive control, showed an accelerated differentiation based on naïve marker genes, as expected. Using the formative marker genes (Figure 5.9 B) and DEGs (Figure 5.9 C) as gene groups to put the KOs on the differentiation axis revealed two major differences. First, the delays for both groups, formative markers and DEGs, were smaller, i.e. timings were later. Second, the time frames with lower distance to the best fit spanned more time points than in the timing based on naïve marker genes. Additionally, the DEGs showed a similar order of KO timings compared to the naïve markers, while the order based on formative markers was not as consistent with the order based on naïve markers.

While about 21 KOs only showed a mild effect on differentiation (delay of 2 to -2 hours) based on naïve marker genes, the differentiation delays were less pronounced using formative markers or DEGs. Here, 26 and 44 KO genes led to a mild effect on differentiation delay (Supplementary Table 1). This difference in differentiation delay strength was also observed when comparing the timings of KOs directly between the different reference gene sets. Comparing timings based on formative markers (Figure 5.10 A) and DEGs (Figure 5.10 B) with timings based on naïve marker genes revealed that naïve marker genes, across most KOs, show stronger delays than the other two groups. Taken together, this shows that the effect of differentiation delay depends on the gene set used as a reference. The effect on the naïve marker genes was also the most pronounced of the three groups used here.

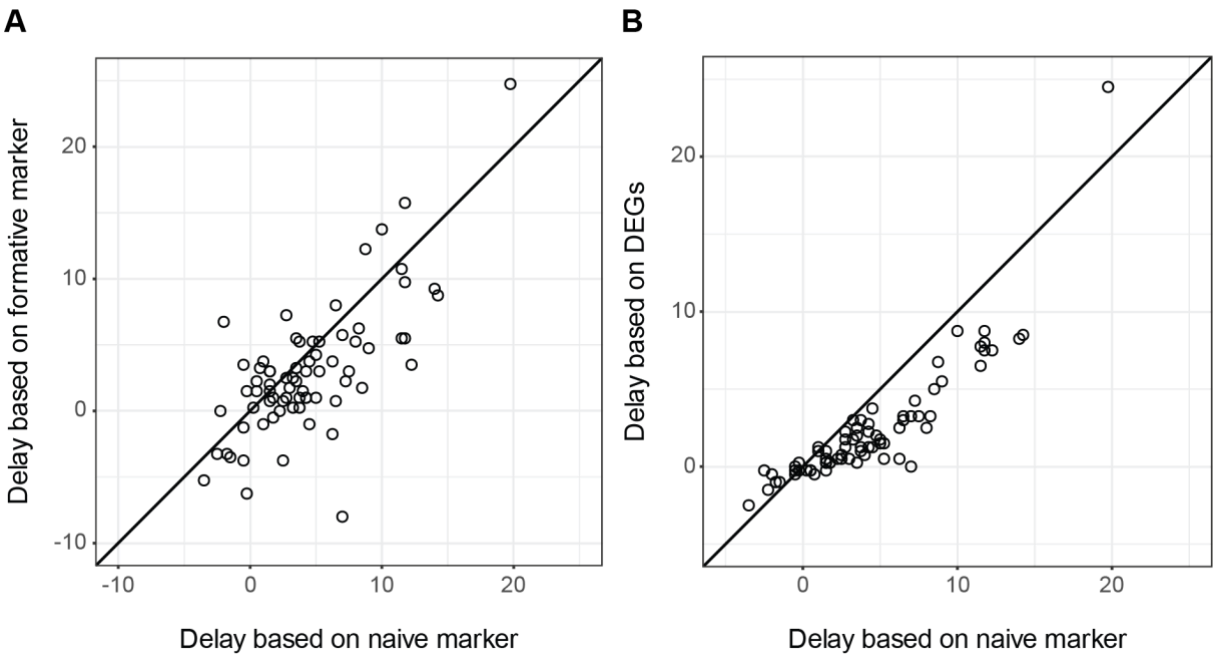

Figure 5.10: Scatterplots comparing differentiation delays based on different groups.

A) Scatterplot comparing timings of KOs based on formative marker genes vs. timings based on naive marker genes. B) Scatterplot comparing timings of KOs based on DEGs vs. timings based on naive marker genes.

5.2 Gene hierarchies in naïve to formative differentiation

As described in the previous chapter, different genes were affected differently by the differentiation delay a cell was exposed to. Thus, the differentiation delay of a cell always depends on the genes used as a reference. That observation motivated the investigation of possible gene hierarchies in early differentiation. While some processes or genes are already linked to an important role in early differentiation, this analysis can provide insight into the possible impact of basal processes or genes linked to basal functions on differentiation. Construction of a possible gene hierarchy in the naïve to formative pluripotency transition could extend understanding of the interplay of genes and processes during this strictly regulated transition.

5.2.1 Knockout-based gene regulatory dependencies

As some genes, like naïve markers, are well studied in the naïve to formative transition, we decided to first analyze possible hierarchies and dependencies on the gene level. Here, naïve markers can help interpret results as we observed that naïve markers are regulated highly similarly in differentiation (Figure 5.8).

Genes used in this analysis were limited to genes fulfilling different criteria. First, genes were required to show differential expression in the WT differentiation (absolute log2FC 24 hours vs. 2i >= 0.5 and padj <= 0.05). The assumption was that to lead to expression changes of other genes, the expression of the gene itself had to change. Second, the genes had to be part of one of three groups defined in previous work[251]. The first group, the naïve associated genes, were genes that, across the

KOs, followed an expression profile highly associated with the naïve marker genes. These genes can either be up or downregulated even though all naïve markers exhibit downregulation. The second group, the constitutive genes, were deregulated in 2i and after 24 hours in at least one KO. The third group, the induced genes, were not or only slightly deregulated in 2i but significantly deregulated after 24 hours in at least one KO. By these restrictions, 1,203 genes that changed during the transition from naïve to formative pluripotency, were deregulated in at least one of the KOs or tightly associated with the naïve marker genes.

Possible gene regulatory dependencies were calculated using the completion percentage of WT change as a measurement for each gene in each KO (Figure 5.11). This measurement reflects how much of the change to a gene in the WT after 24 hours has also happened in the KO. The completion percentage of WT change was then compared for each gene pair over the 73 KO cell lines. The differences in WT completion over all KOs result in a distribution for every pair of genes. The mean from this distribution was required to be significantly different from 0 (two-sided t-test: adjusted p-value <= 0.05 after correction for multiple testing according to Benjamini and Hochberg[252]).

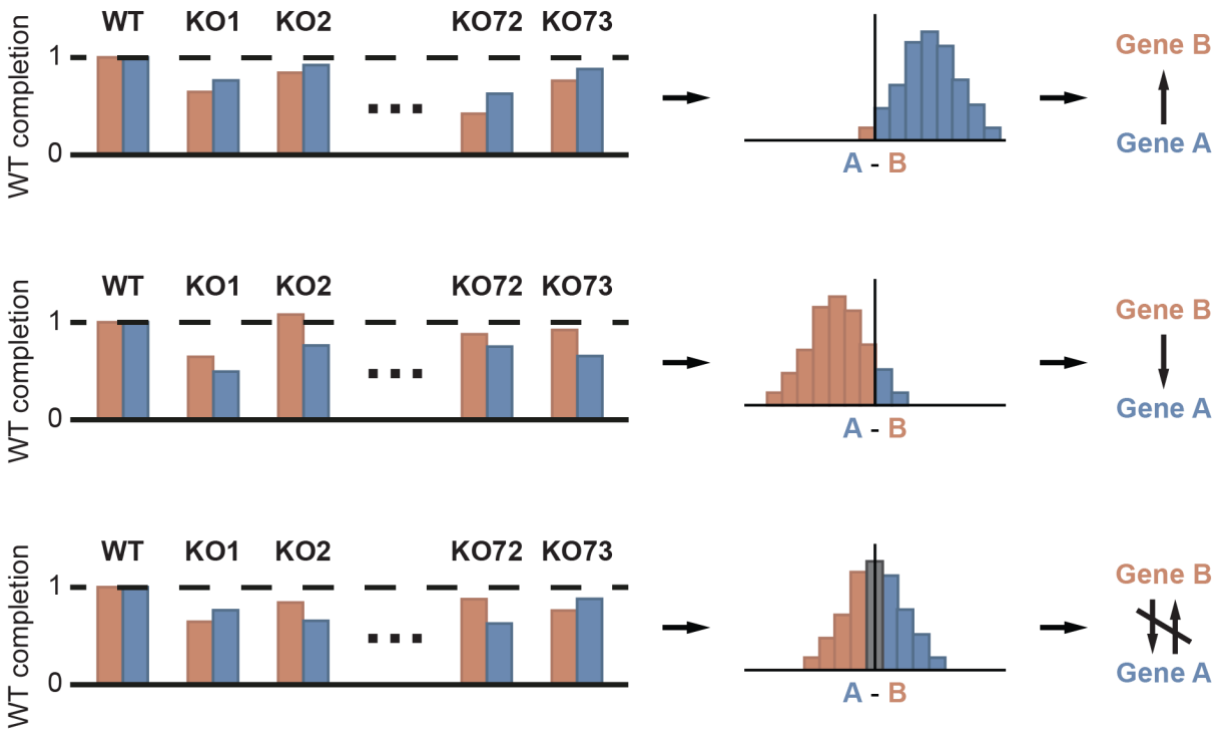

Figure 5.11: Scheme representing possible dependencies between two genes derived from the KO data.

Possible dependency of gene B on gene A, gene A on gene B, and no dependency between gene A and gene B are shown from top to bottom, respectively. Each example consists of three parts. On the left, WT completion of genes A and B are shown for the 73 different KOs. The histogram resulting from differences in WT completion between gene A and gene B is depicted in the middle. On the right, the resulting dependency between the genes is shown.

Additionally, we required the overall consistency of direction in the differences to be consistent with the mean in at least 70% of the KO cell lines. When comparing two genes, A and B, a positive dependency means that over most KOs, gene A showed higher WT completion than gene B. Thereby, gene A adapted independently of gene

B in the KO data, and gene B might depend on changes of gene A. While the independence of gene A from gene B, in that case, would be supported by our data, dependence in the other direction would need further evidence to be considered a causal dependency. In this case, the observed possible dependence of gene B on gene A might be caused by a third gene affecting either gene A and gene B or only gene B, thus leading to the observed consistency in WT completion differences between A and B. However, a negative dependency between A and B would describe the opposite case. B would adjust independently of A, and A might depend on B. The possible dependence and independence between *Dnmt3a* and *Nanog* (Figure 5.12) fulfills the aforementioned requirements and would result in a positive dependency between *Dnmt3a* and *Nanog* as the histogram was shifted towards positive values. This means that changes in *Nanog* might depend on changes of *Dnmt3a*, but more strikingly, *Dnmt3a* changes happen independently of changes of *Nanog*.

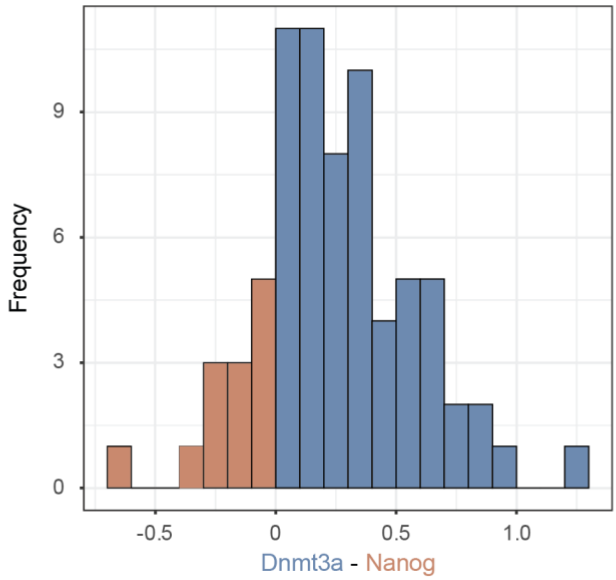

Figure 5.12: Histogram of WT completion difference between Dnmt3a and Nanog.

Bars in blue represent KOs with higher completion of Dnmt3a*, and red bars represent KOs with higher completion of* Nanog*.*

All possible combinations of gene-gene dependencies for the 1203 genes of interest would result in 0.7 million gene regulatory dependencies. Of those, approximately 23% showed a significant difference in their WT completion consistent across most of the 73 KO cell lines. The remaining 165,544 gene regulatory dependencies resulted in a matrix inversely mirrored among the diagonal. The mirrored axis with the opposing sign results from A to B and B to A, sharing the same underlying histogram with opposing signs.

While dependencies calculated from the KO data strongly support independence from one gene's adjustments to another, causal dependence is not supported. The calculated dependency between two genes, or difference in WT completion, can also result from a third gene. Thus, this direction of possible dependencies needs more support to draw further conclusions from the KO-based gene-gene dependencies.

The independence can already be used to examine processes and associated gene regulatory dependencies. However, this work will refer to the calculated difference in WT completion as dependency.

5.2.2 Knockouts with differentiation delay phenotype show higher consistency of gene regulatory dependencies

As the 73 KOs had different strengths of phenotype, i.e., delaying differentiation by a different degree, we asked whether the strength of the phenotype resulting from the KOs influenced the dependency analysis. More precisely, we asked whether weaker phenotypes were more likely to contradict the overall difference in WT completion between two genes. The timing of differentiation progress based on naïve marker genes (Supplementary Table 1) defined KOs with weak phenotypes. A KO was considered weak or a no-differentiation-delay phenotype when timed to 23 hours or later (Figure 5.9 A). KOs mapped to an earlier time point on the differentiation axis have a better matching time point in the time course than 24 hours and were thereby considered to show a differentiation delay. The KOs considered stronger phenotypes with the smallest delay in differentiation were mapped to 22.5 hours. As weak phenotypes were expected to show close to 100% WT completion for most genes, they might primarily represent noise and add noise to the analysis. Therefore, we reran the analysis, excluding KOs without differentiation-delay-phenotype, including control KOs like *Myc* (acceleration of differentiation), thereby limiting the KOs to 56. While removing 17 KOs (*Msi2*, *Rps6ka1*, *Arih2*, *Cabin1*, *Hprt*, *Tet1*, *Igf2bp1*, *Nes*, *Ssr2*, *Irak3*, *Etl4*, *Pum1*, *Myc*, *Dido1*, *Hnrnph1*, *Fgfr2*, and *Foxp1*) from the analysis, previous dependencies would not be expect to change signs as most data used was still maintained. However, the effect size (shift of the corresponding histogram from zero) and the consistency (fraction of KOs with the same sign for the dependencies as the average dependency), could change when removing KOs from the analysis. Additionally, removing KOs from the analysis will remove statistical power from the test of the shift from zero for the dependency between two genes.

Inspection of the positioning of weak phenotypes (white color) in the histograms of selected examples (Figure 5.13) revealed that weak phenotypes clustered at the tails of the histograms opposing the direction of the mean. In the case of *Dnmt3a* and *Nanog* (Figure 5.13 A), weak phenotypes accumulated at the left end of the histogram. This end represented KOs with further progression of *Nanog* compared to *Dnmt3a*, thereby contradicting the overall direction of the dependency. For the comparison of *Pou3f1* and *Klf4* (Figure 5.13 B), however, the weak phenotypes tended to be present in a second peak of the histogram. This peak was also located on the opposite side of the overall direction of the strong phenotypes. The difference to the *Dnmt3a* and *Nanog* example was that the weak phenotypes did cause a bimodal distribution instead of just blending in as extreme values of one side. Both examples shared the commonality that weak phenotypes were not spread across the whole histogram but tended to represent extreme values of the distribution that contradicted the overall direction.

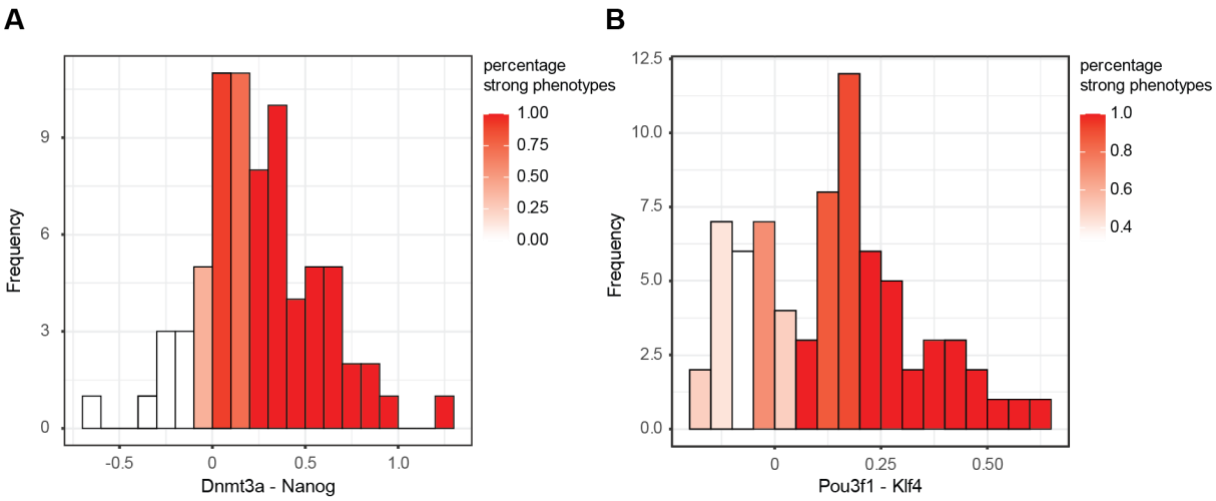

Figure 5.13: Histograms of gene regulatory dependencies for two gene pairs.

A) Dependency between Dnmt3a *and* Nanog*. B) Dependency between* Pou3f1 *and* Klf4*. The bars of the histograms are colored by the percentage of strong phenotypes.*

To assess the general validity of the observed patterns in those two examples, we tested whether weak phenotypes had different relative positions in all histograms than strong phenotypes. All positive dependencies between two genes were inspected for the relative position of genes from both groups to the mean of each histogram. The relative position ranges from -1 to 1, with 0 corresponding to the mean. All negative values represented KOs with smaller WT completion differences than the mean of the histogram. Thereby, all positive values represented KOs that showed bigger WT completion differences than the mean of the histogram. The WT completion differences were then divided by the biggest absolute distance from the mean per gene pair to allow cross-gene pair position comparison. By definition of the relative position, strong negative values were more likely to show opposing signs to the mean of a gene pair. Thus, a two-sample Kolmogorov-Smirnov test was used to test if the relative position of strong phenotypes differed from that of weak phenotypes. The two-sided test resulted in a p-value < $2.2*10^{-16}$, showing that both groups represented different distributions. Comparing the means of both underlying distributions revealed that strong phenotypes were almost centered around 0 with a mean of approximately 0.038. Weak phenotypes, however, were shifted to the negative relative positions centered around approximately -0.145. Hence, weak phenotypes were more likely to show opposing directions of dependencies than the rest of the KOs.

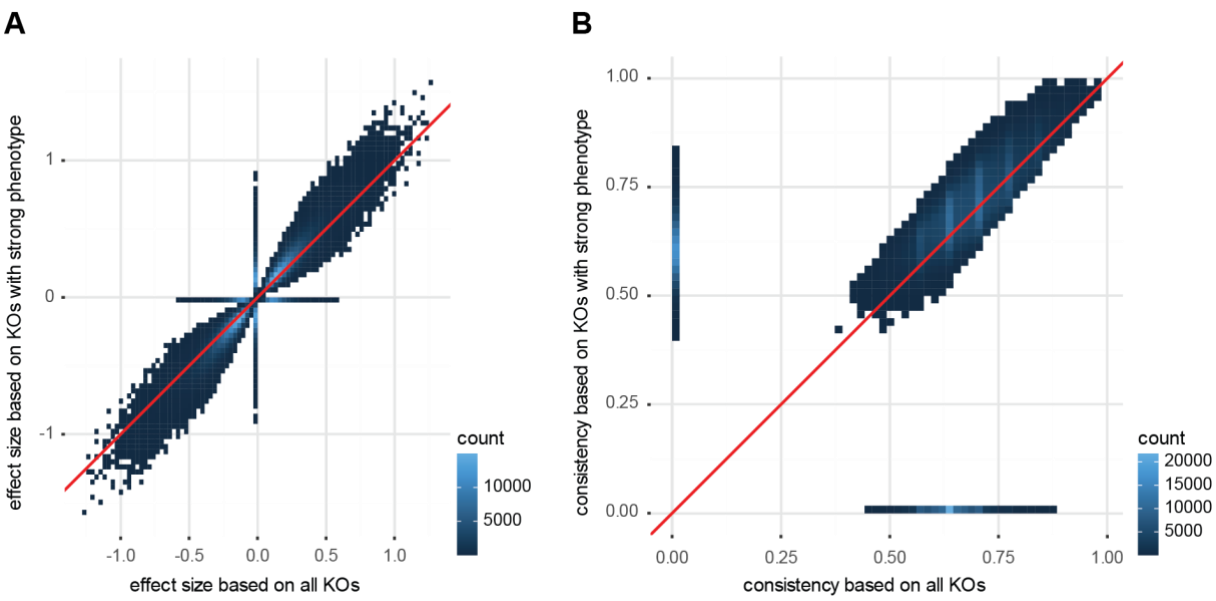

Figure 5.14: comparing dependency metrics before and after removing weak phenotypes.

A) Two-dimensional density plot of effect size based on all 73 KOs (x-axis) and the reduced set of KOs with a differentiation delay phenotype (y-axis). Bright blue color indicates higher density of gene-gene dependencies and red line marks the diagonal. B) Two-dimensional density plot of consistency of gene-gene dependencies for all 73 KOs (x-axis) and the reduced set of KOs with a differentiation delay phenotype (y-axis). Bright blue color indicates higher density of gene-gene dependencies and red line marks the diagonal.

If the hypothesis that KOs with weak phenotypes would add more noise to the analysis were correct, both metrics (effect size and consistency) would be expected to increase in the analysis based on strong phenotypes. This trend was observed for effect size (Figure 5.14 A) and consistency (Figure 5.14 B). While the effect size, including all KOs and excluding KOs with weak phenotypes, was highly correlated, a shift in the slope was observed. This was caused by more extreme effect sizes when only considering KOs with stronger phenotypes. The same trend was observed with the level of consistency for all KOs and only KOs with stronger phenotypes. Here, most gene regulatory dependencies had a higher consistency when only considering strong phenotypes compared to using all KOs in the analysis.

Additionally, the overall number of dependencies significantly different from zero (before checking for consistency across KOs) increased from approximately 420,000 to 435,000. Considering the loss of statistical power, this already hints at stronger consistency when excluding phenotypes that did not show a differentiation defect phenotype. When considering the consistency, the difference in approaches showed a bigger effect. In the case of all 73 KOs, approximately 165,000 gene-gene dependencies exhibited a consistency of at least 70% of KOs showing the same direction of WT completion differences as the mean over all KOs, while approximately 255,000 did not fulfill this requirement. When only considering the 58 KOs with a differentiation delay phenotype, this shifted to approximately 225,000 dependencies meeting the requirements and approximately 210,000 not meeting the 70% requirement. This showed that only considering differentiation-delay-phenotypes in the KO-based gene-gene dependencies resulted in higher consistency and represented more robust results. Therefore, the following steps will be based on the

dependencies derived from the 58 KOs with a differentiation-delay-phenotype. In the resulting matrix of dependencies (Figure 5.15), genes were first separated by whether they were mostly upstream of other genes (top half) or mostly downstream of other genes (bottom half). Genes downstream of other genes had mostly negative dependencies toward other genes. In contrast, genes that tended to be upstream of other genes had mostly positive dependencies towards other genes. The split was also represented in the connectedness of genes within the same branch of the split and genes from the other branch. While genes tended to have fewer dependencies to genes within the same branch (downstream genes were less connected to other downstream genes), dependencies to the other branch were more frequent (downstream genes had many negative dependencies towards upstream genes).

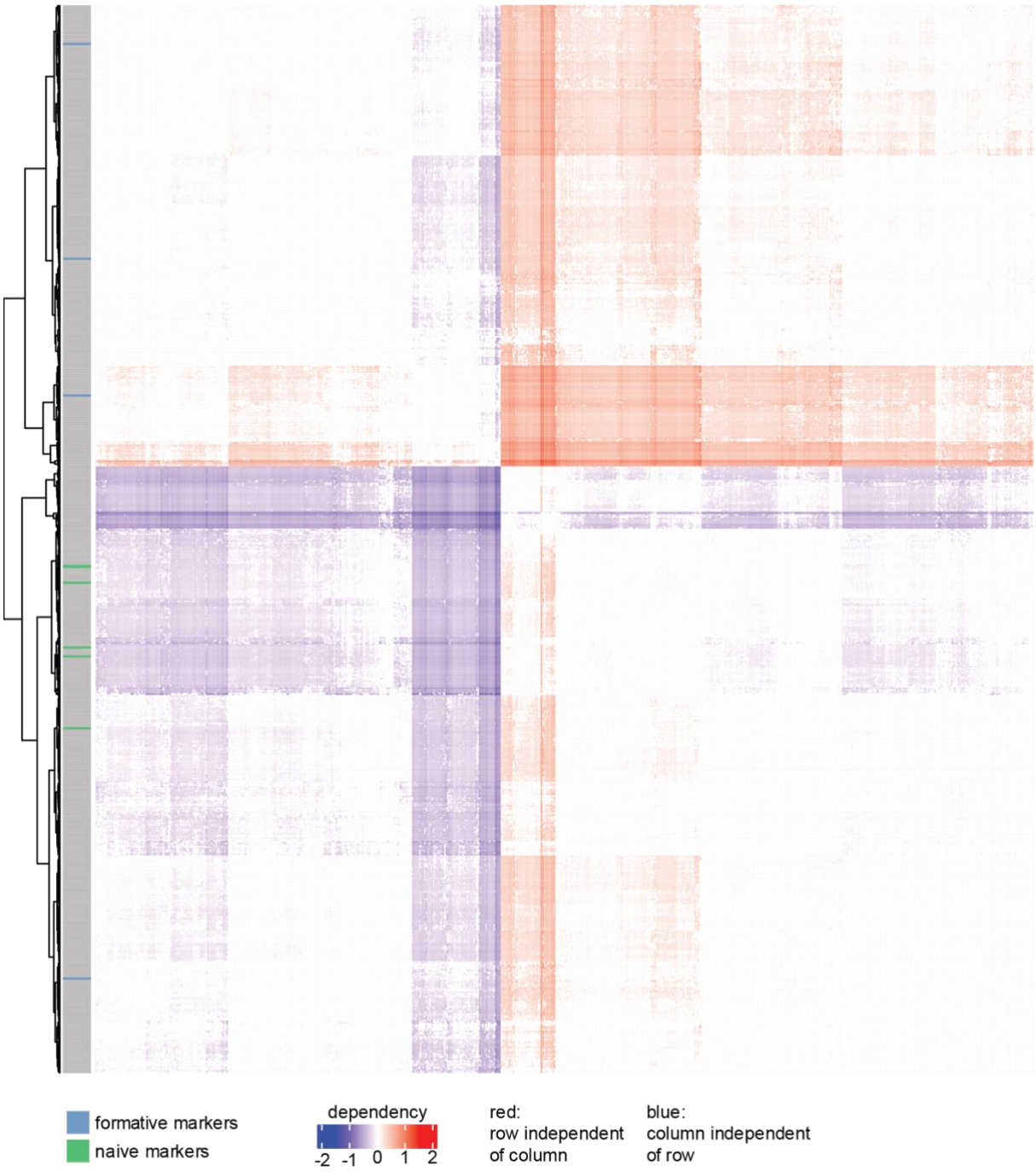

Figure 5.15: Heatmap of gene regulatory dependencies calculated from KOs with differentiation delay phenotype.

Dependencies are read row to column with positive values indicating the gene in the row adjusting independently from the gene in the column across the 58 KOs. Negative values indicate that the gene in the column adjusts independently from the gene in the row across the 58 KOs. The annotation on the left of the heatmap highlights naïve and formative marker genes in green and blue, respectively.

Subsequently, the resulting dependencies were clustered using NMF. NMF was used due to its clustering properties as well as maintaining a measure for how well genes fit into sub-optimal clusters. Each gene was assigned to a signature with a certain weight, and therefore the weights for the signatures can later be used to see how influential this gene is to other signatures. The clustering properties are achieved by assigning the genes to the signature with the highest weight for that gene. This artificially produces hard clustering, but information on the weights can still be accessed for each gene later. Here, we grouped the genes into 6 clusters (Supplementary Table 2) of comparable sizes (Figure 5.16). While formative markers spread across clusters 2, 5, and 4, naïve markers were mostly in cluster 3, with the only exception of *Tfcp2l1* in cluster 1.

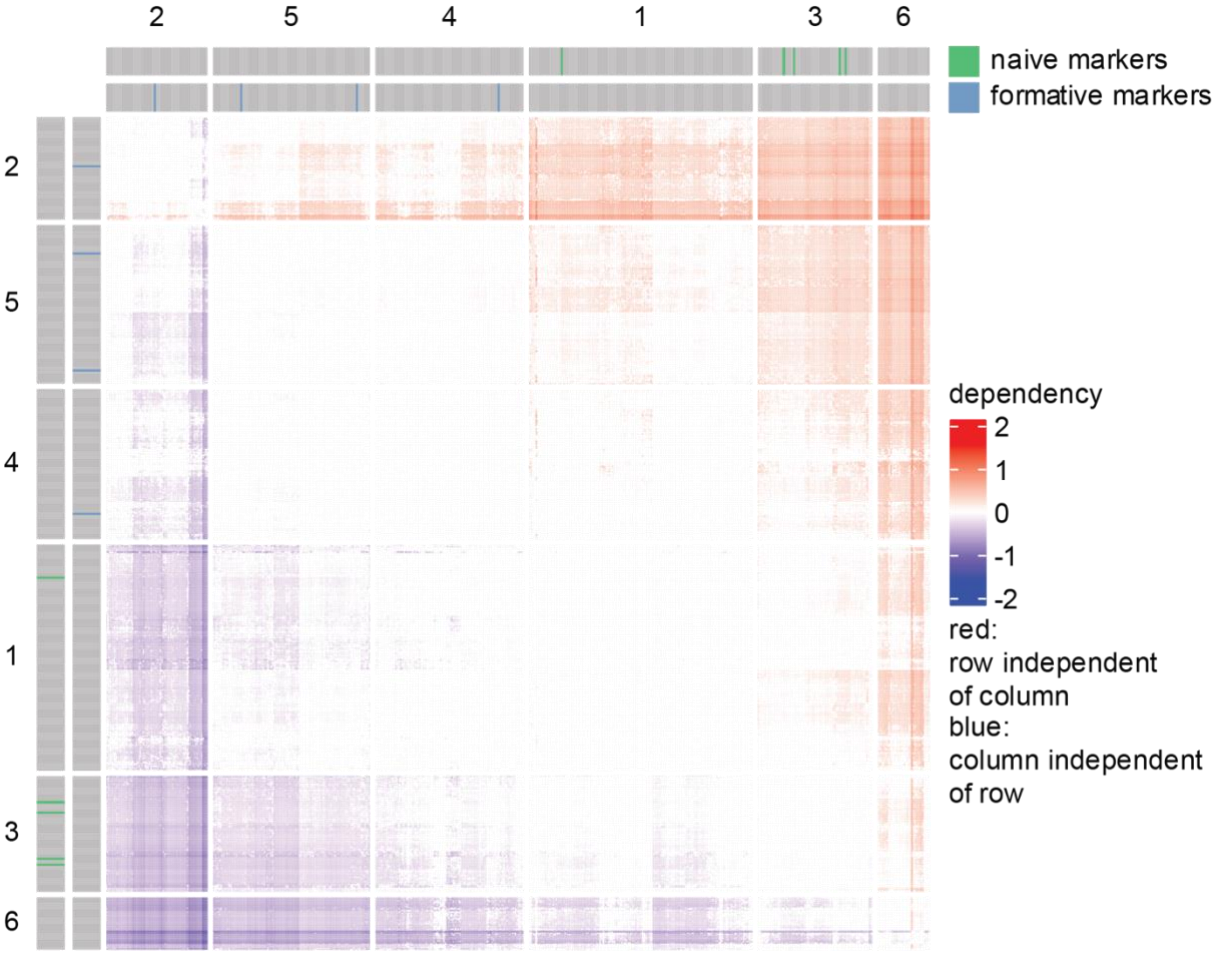

Figure 5.16: Gene regulatory dependencies based on differentiation delay phenotypes only clustered by NMF.

Clusters were additionally sorted by highest average row means, and the presence of naïve (green) and formative (blue) marker genes was annotated on the left and top of the heatmap.

5.2.3 Adaptation to differentiation of formative markers is independent from naïve markers

As shown in the previous section, the first split of the 1,203 genes did separate the remaining genes into two sets of comparable size (Figure 5.15). The first group consisted of genes that tended to be downstream of other genes, thus possibly

depending on other genes. The second group consisted of genes that tended to be upstream of other genes, thus adjusting to differentiation independently of other genes. While six naïve marker genes were located at the upper half of the heatmap and thereby belonged to the genes that were mostly downstream of other genes, formative marker genes (except for *Otx2*, which was not part of the 1,203 genes) were located at the lower half, the genes that were mostly upstream of other genes.

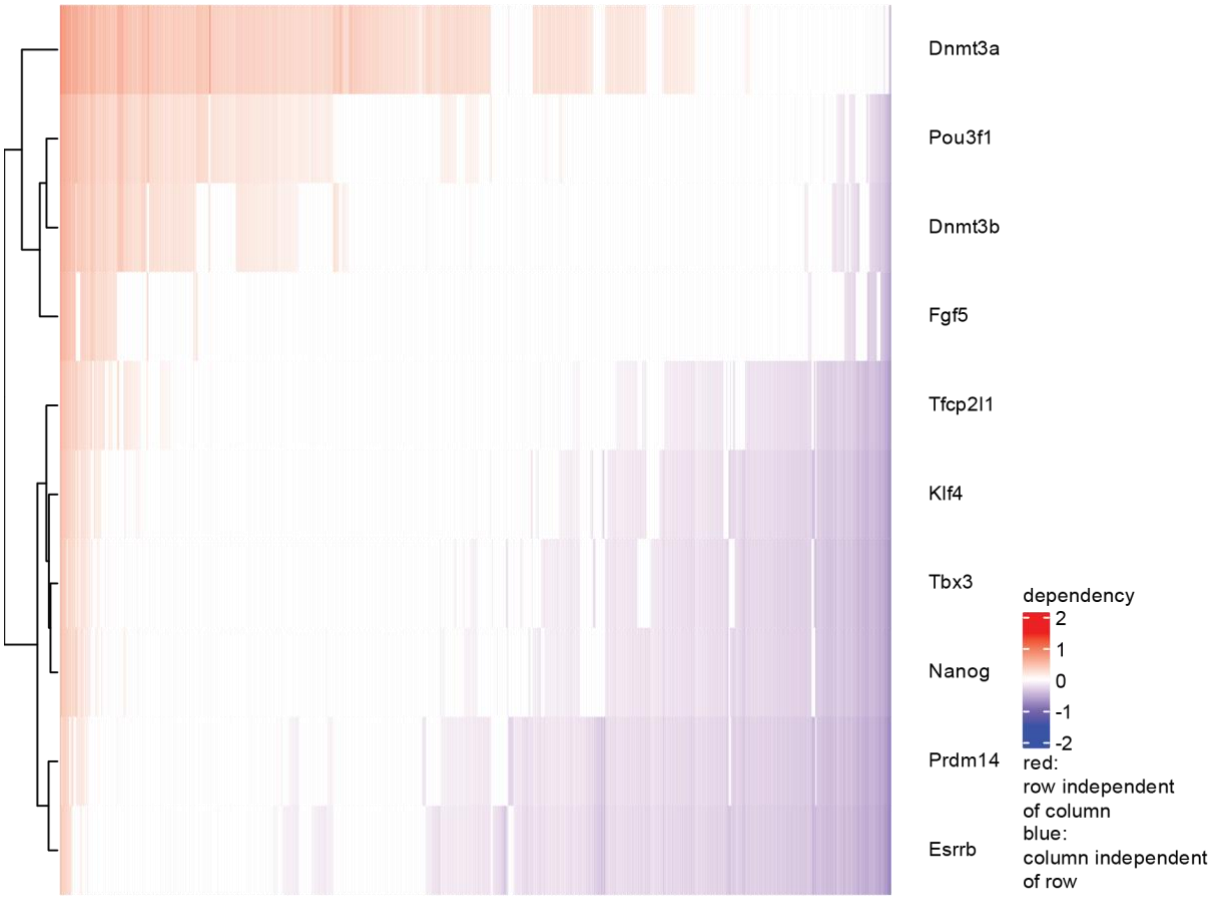

Figure 5.17: Gene regulatory dependencies of* naïve *and* formative *markers.

All gene regulatory dependencies between all included naïve and formative markers (rows) to all genes with at least one dependency to one of the markers.

Only plotting gene regulatory dependencies from naïve and formative marker genes to other genes emphasized this separation (Figure 5.17). Here, the first split of the dendrogram also separated formative (top) from naïve (bottom) marker genes. While formative marker genes showed mostly positive dependencies, all formative marker genes also included negative dependencies. The opposite applied to the naïve marker genes. Even though most dependencies from naïve marker genes to other genes were negative, all naïve marker genes also included some positive dependencies towards other genes.

Additionally, the number of positive dependencies varied between the formative markers. *Pou3f1* and *Dnmt3b* were comparable in positive and negative dependencies towards other genes. *Dnmt3a* included more positive and fewer negative dependencies, indicating a more independent adjustment to differentiation

from other genes than *Pou3f1* and *Dnmt3b*. Even though *Fgf5* still clustered with the other formative markers, it was by far the most balanced marker gene when considering the ratio of positive and negative dependencies toward other genes. The same observation of different positive to negative dependencies in formative marker genes was made for the naïve markers. *Esrrb* and *Prdm14* were the markers with the most negative and the least positive dependencies. *Tbx3*, *Tfcp2l1*, *Klf4*, and *Nanog* were less extreme when looking at the ratio of negative to positive dependencies (compared to *Esrrb* and *Prdm14*) but still showed a tendency towards negative dependencies.

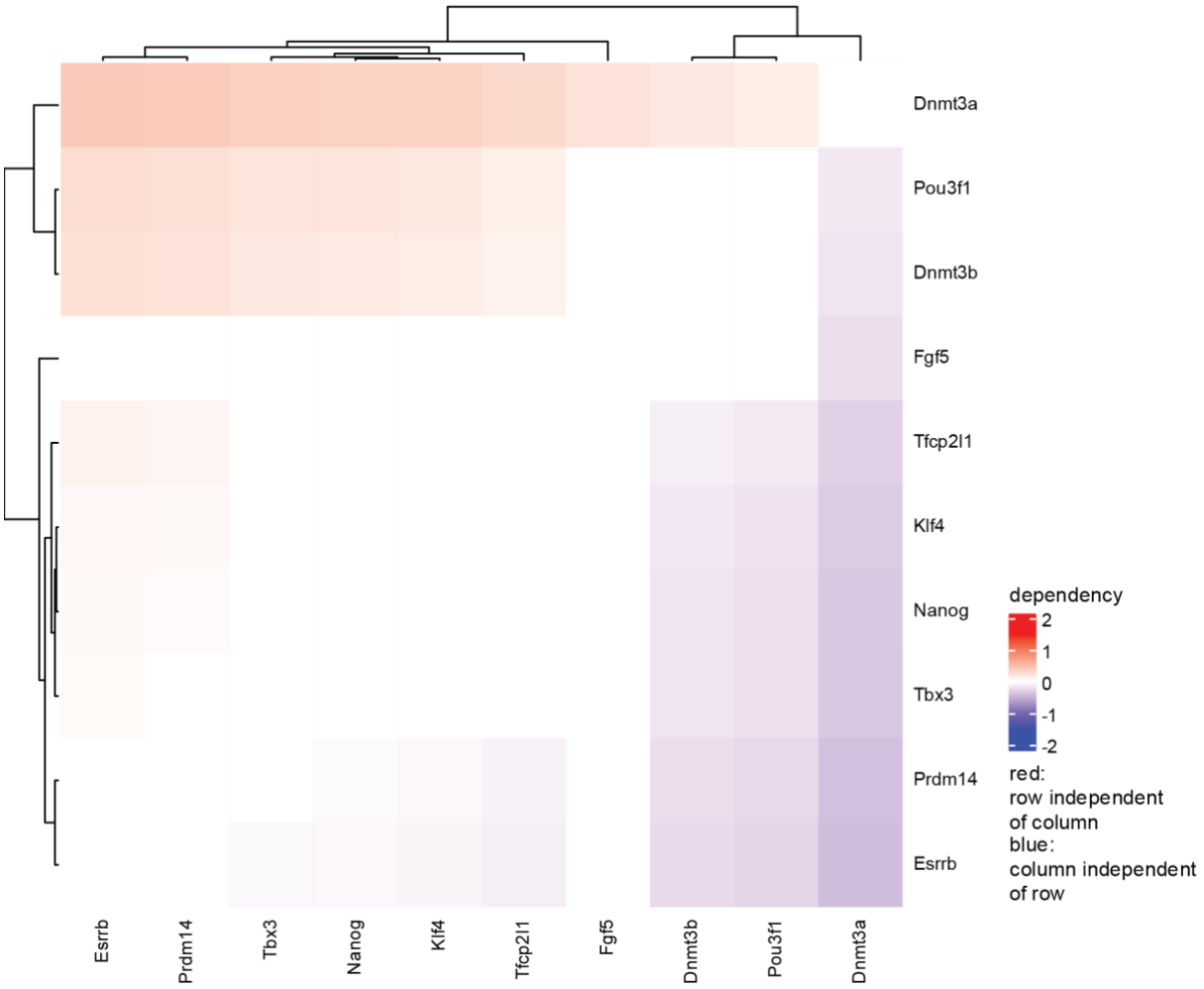

Figure 5.18: Gene regulatory dependencies between all included naive and formative markers.

Only dependencies between naïve and formative markers are depicted. Rows and columns are clustered hierarchically according to the included dependencies.

This trend was mostly conserved when only dependencies between naïve and formative markers were considered (Figure 5.18). *Dnmt3a* was the most extreme upstream gene of the included markers, with positive dependencies to all naïve markers and *Fgf5*. *Fgf5*, while belonging to the formative marker genes, clustered with the naïve marker genes, which agreed with *Fgf5* being the formative marker with the lowest number of positive dependencies towards other genes. While *Dnmt3a* had a positive dependency to *Fgf5*, another formative marker, the naïve markers *Tfcp2l1*,

Klf4, and *Nanog* contained positive dependencies to *Esrrb* and *Prdm14*. Additionally, *Tbx3* contained a positive dependency towards *Esrrb*. However, all within formative marker or within naïve marker dependencies were relatively weak in comparison to dependencies between the groups. With the only exception of *Fgf5*, the general tendency was that formative markers had positive dependencies towards naïve markers.

These results suggest that naïve marker genes mostly adapt downstream of other genes in the naïve to formative transition, while formative marker genes tend to adapt upstream of other genes. Thus, formative markers adapted independently of many other genes, including naïve markers, to the differentiation signals. Many other genes, including naïve markers, may depend on formative marker gene expression changes in the naïve to formative pluripotency transition. The naïve maker genes, however, might depend on changes of other genes, including formative markers, to adapt to differentiation. Only a limited number of genes are downstream of naïve markers. While naïve markers are required to keep the cells in the naïve pluripotent state, the transition to subsequent cell states seems to be independent of the naïve network.

5.2.4 Gene regulatory dependencies are in agreement with known dependencies

To put the potential gene regulatory dependencies identified in the previous section into a biological context and see if they make sense, the KEGG pathway annotation[268–270] for “signaling pathways regulating pluripotency of stem cells” (Figure 5.19) was assessed. Here, a broad range of pathways, such as WNT, LIF, or FGF/ERK signaling are included, thus providing a broad overview of dependencies identified in our analysis. Only considering genes found in this pathway annotation (Figure 5.20 A) provided a first overview of gene regulatory dependencies of this pathway and naïve and formative marker genes from our analysis.

The first aspect found in the network and the pathway overview are *Sox2* and *Nanog*, located downstream in both views. In the gene regulatory dependency overview, *Nanog* had negative dependencies to *Wnt5b*, *Fzd2*, *Fzd10*, and *Axin2* in the WNT branch of the pathway annotation. The negative dependencies indicated that the adaptation of *Nanog* to differentiation might depend on these genes. This agreed with these genes being positioned upstream of *Nanog* in the WNT branch of the annotation. However, the dependency of *Nanog* to *Esrrb* is positive indicating the possible dependency of *Esrrb* on changes of *Nanog*. In the KEGG pathway, *Esrrb* is located upstream and downstream of *Nanog*, partially supporting what was captured in the dependencies. Previously, it was shown that *Esrrb* is a direct target of *Nanog*[271] further supporting the dependency. While *Esrrb* strengthens the naïve marker *Nanog* in the naïve state, in the transition to formative pluripotency, *Nanog* seems to adapt independently from *Esrrb*.

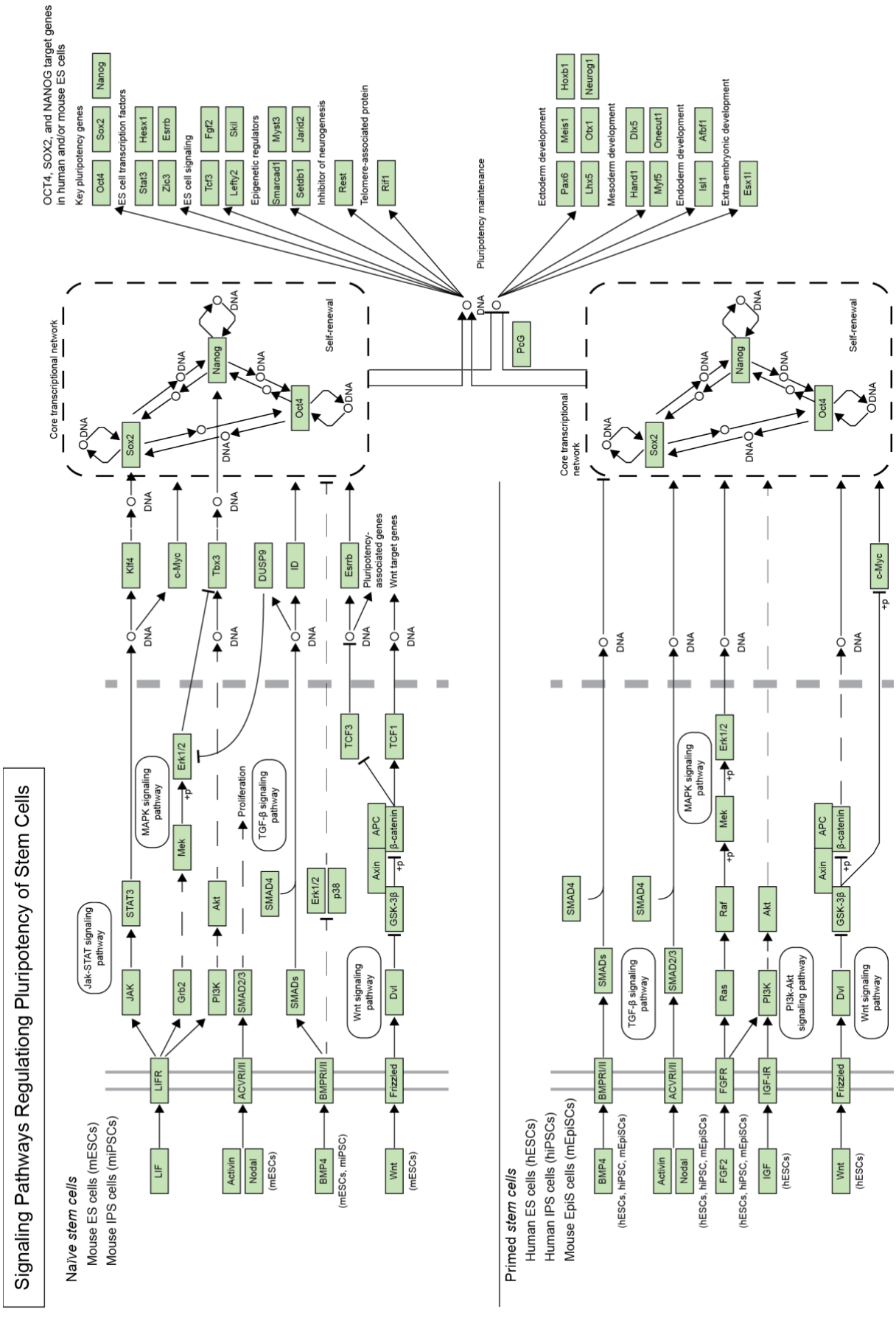

Figure 5.19: KEGG pathway mmu04550 adapted from the KEGG pathway database [268–270]***.***

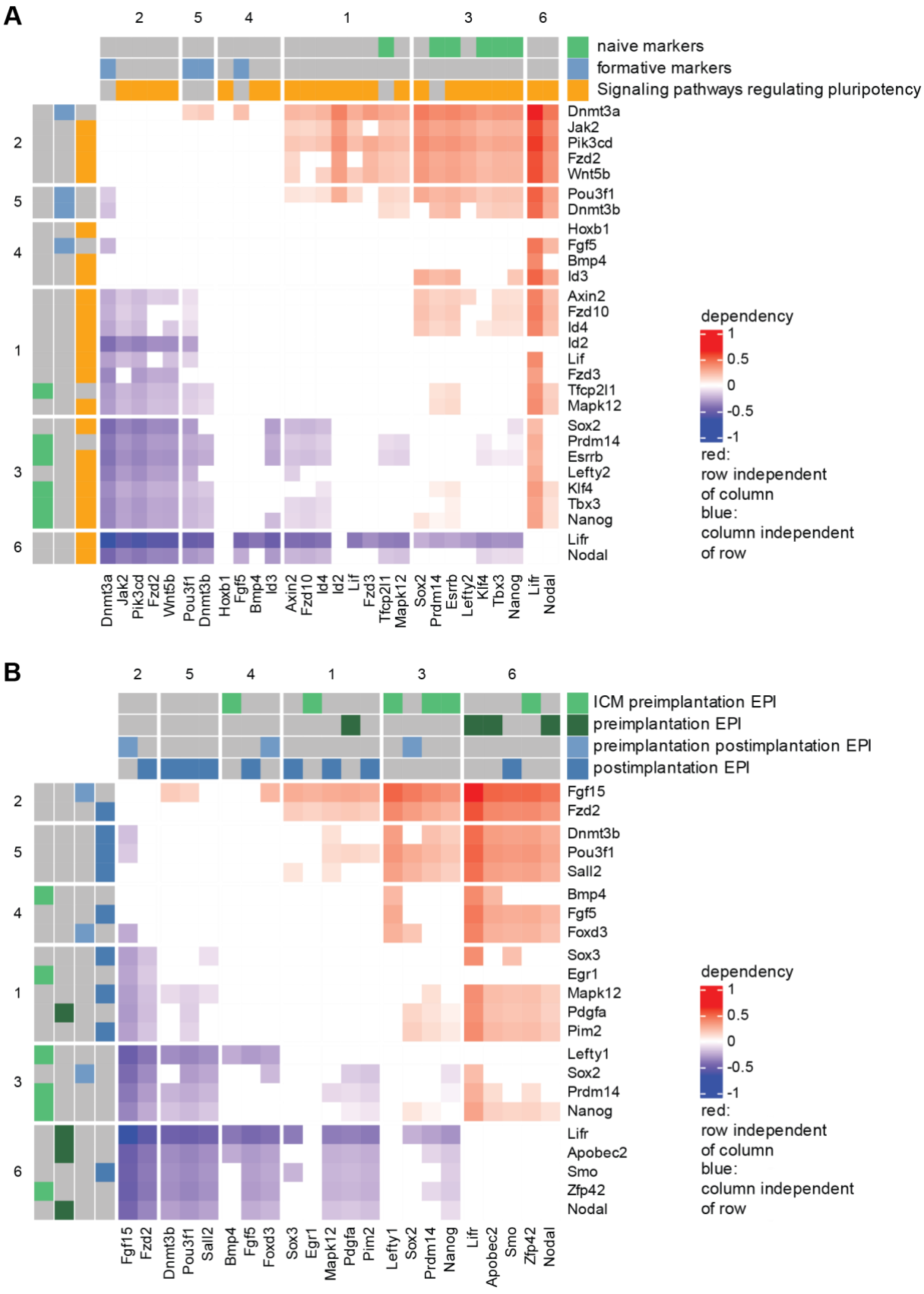

Figure 5.20: Heatmaps showing gene regulatory dependencies for different pathways and gene groups.

A) Heatmap of dependencies for all naïve markers (green), formative markers (blue), and genes assigned to KEGG pathway mmu04500 (orange). B) Heatmap of dependencies for genes defined to be expressed at different stages of differentiation. Groups are colored from early stages to late stages (light green, dark green, light blue, dark blue) and taken from Boroviak et al.[241].

The second aspect of the KEGG hierarchy captured by the KO-based gene regulatory dependencies was local structures between genes directly connected through the same pathways. *Axin2* was one gene representing a gene where gene regulatory dependencies nicely captured the actual hierarchy in their pathway. *Axin2* is located downstream of *Wnt5b* and *Fzd2* in the KEGG hierarchy which was captured by negative dependencies to those two genes. On the other side, *Axin2* also had a positive dependency on *Esrrb*. That agreed with *Esrrb* being further downstream of *Axin2* in the WNT signaling pathway. However, *Fzd10* and *Fzd3* had no dependency on *Axin2*. This could either mean that *Fzd2* is more strictly regulated in the time window of early differentiation or that the regulation of those *Fzd*s and *Axin2* is too tight to be captured by our data.

While the previous examples showed agreement between the hierarchies annotated in the KEGG database, some examples disagreed compared to the calculated dependencies. One example to investigate possible partial disagreements was LIF signaling annotated in KEGG. The LIF branch of the KEGG pathway combines MAPK, PI3K-AKT, and JAK/STAT signaling in the naïve state. While JAK/STAT signaling is not represented on the side of primed ESCs in the pathway annotations, MAPK is located downstream of FGF and PI3K/AKT downstream of FGF and IGF signals. These signaling pathways are upstream of *Nanog* and *Sox2* in the KEGG annotation. This was also represented in the gene-gene dependencies, as *Jak2*, *Pik3cd*, *Id3*, *Axin2*, and *Id4* all have positive dependencies towards *Nanog* and *Sox2*. Thus, the adjustment of this group of genes, representing different signaling pathways, happened independently of *Nanog* and *Sox2*. Lifr was an exception that had a negative dependency to both genes, *Nanog* and *Sox2*, even though it is upstream of them in the KEGG annotation. The only dependency of Lifr that agreed between the hierarchy given by KEGG and the KO-based dependencies was *Lif* to *Lifr*. *Lif* adjusted independently of *Lifr* according to the KO-based dependencies, which agreed with its upstream location in KEGG. However, the following downstream KEGG genes (*Jak2*, *Pik3cd*, *Klf4*, and *Tbx3*) all adapted independently from *Lifr* in the KO-based dependencies. Additionally, the positive dependencies of *Pik3cd* and *Jak2* to *Lif* indicated *Lif*-independent adjustment of JAK/STAT signaling and PI3K/AKT signaling in the naïve to formative transition.

Next, we explored gene regulatory dependencies of state-specific genes from literature[241] (Figure 5.20 B). Here, the KO-based dependencies were limited to those of genes linked to the ICM preimplantation EPI, the preimplantation EPI, the preimplantation postimplantation EPI, and the postimplantation EPI. The previous chapter (5.2.3) suggests that the formative network adapts to differentiation independent of the naïve network. Thus, we investigated whether genes specific to later stages of differentiation (postimplantation and preimplantation postimplantation) adapted independently of the genes specific to earlier changes (ICM and preimplantation). Even though the difference between preimplantation and preimplantation postimplantation stages might not be as clear as the naive to formative state, a trend of late state genes (blue) being more likely to be in the upstream clusters 2, 4, and 5 was observed. Additionally, the early-stage specific

genes (green) were observed more often in the downstream clusters 1, 3, and 6. This hinted at an independent adjustment of later stage-specific genes from genes of the earlier stages. Thus, the observed independent adjustment of genes needed for later stages of differentiation did not only apply to the formative network but also other specific genes of those stages. Additionally, the dependent adjustment of genes established in the naïve state did not only apply to the naïve network but also other specific genes of those stages.

While the dependencies from the KO approach mainly agreed with established pathway structures examples where dependencies did not agree with the pathway hierarchy do not necessarily mean that the dependencies are faulty. The dependencies represent an order of adjustment of different genes to differentiation. Thus, a gene downstream of another gene in the pathway can also adapt independently if pathways are rewired during state transitions. Other reasons to see a gene upstream in a pathway but mostly downstream in the dependencies, such as Lifr, could be dependence through feedback loops or lack of signaling through that pathway combined with downstream rewiring. Further, the mostly independent adjustment of later-stage specific genes from early state-specific genes was not limited to the naïve and formative markers. Taken together, our results agreed with established gene-gene dependencies, and findings from previous sections can be extended to more genes than the naïve and formative markers.

5.2.5 Gene regulatory dependencies in *in vitro* single-cell data

While a positive dependency of gene A to gene B strongly supports the independent adjustment of expression of gene A from changes in gene B, it does not support the dependence of changes in gene B on changes in gene A as strongly. To further support the possible dependencies identified from the KO data, we used a different approach on different data. Here, we used data from a sc experiment[272] kindly provided by the group of Christa Bücker that covers the transition represented in the KO data. This scRNA-seq dataset starts in 2i medium and has measurements after 6, 12, 24, and 48 hours. The sc data was also done *in vitro* and covered 48 hours of naïve to primed pluripotency transition.

The data provided by Christa Bücker and colleagues were count data for cells and assignment of cells to time points. In addition to assignment to time points, some cells were identified as "Doublet" or "Negative". These two categories are cells identified as droplets including two cells, or not assigned to a time point. While we completely excluded doublets from the analysis, we kept cells marked as negative for data visualization. When comparing the number of UMIs (Figure 5.21 A) and the number of genes (Figure 5.21 B) measured in the negative cells to all other groups, the negative cells look more like a mixture of the time points and less like the doublets. Therefore, negative cells were not excluded from the analysis. Additionally, cells were excluded if they had less than 200 genes measured, more than 3,500 genes measured, or more than 20 percent of the reads coming from mitochondrial genes. Filtering left 12,760 cells in the analysis: 2,255 cells from 0 hours, 2,559 cells

from 6 hours, 2,316 cells from 12 hours, 2,039 cells from 24 hours, 1,613 cells from 48 hours, and 1,978 negative cells.

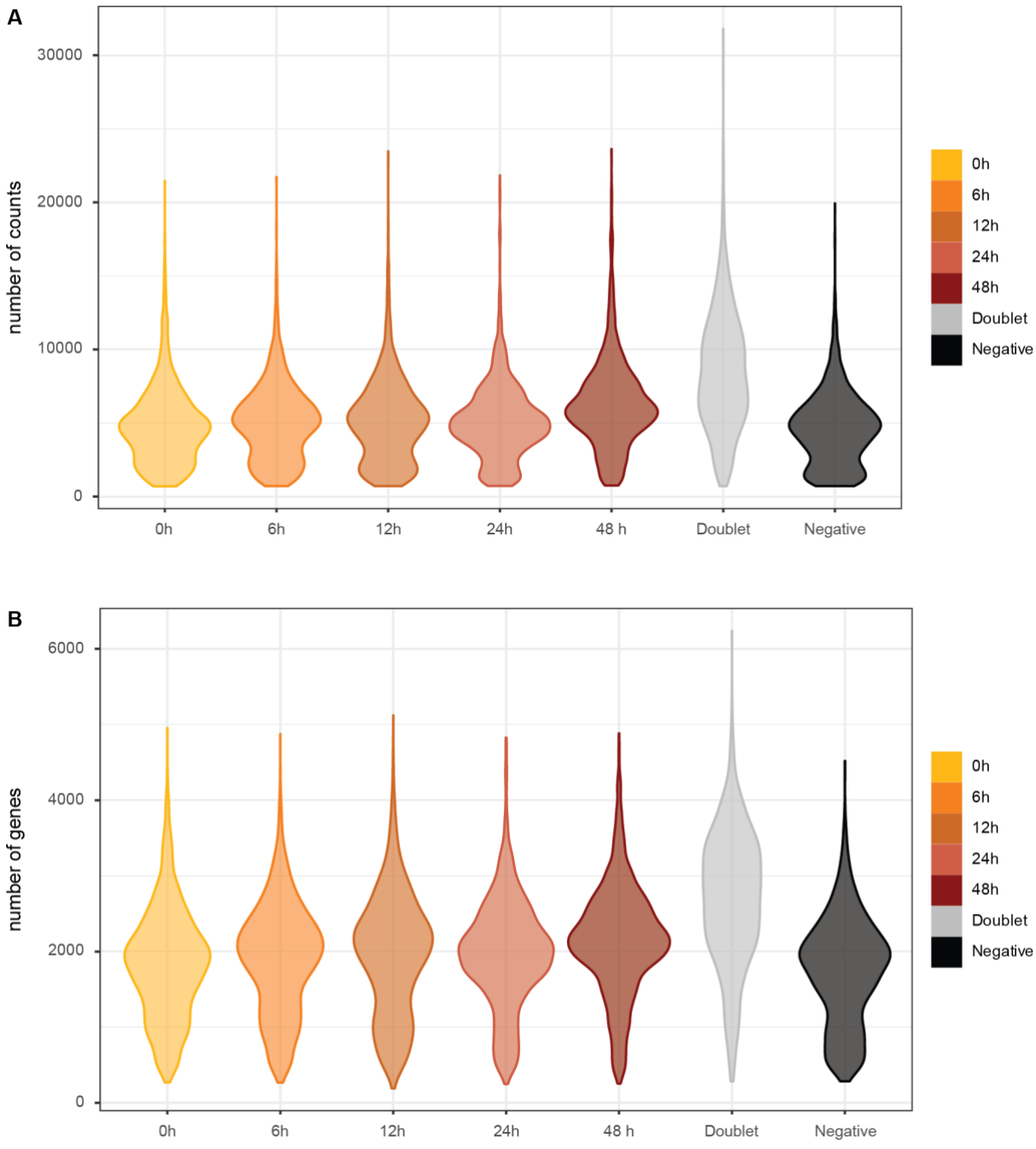

Figure 5.21: Violin plots summarizing quality control readouts for sc data.

A) Violin plot comparing the total counts per cell between different time points. Additionally, doublets and cells without mapping to one of the time points were included. B) Violin plot comparing detected genes per cell between different time points. Doublets and cells without mapping to one of the time points were included as additional groups.

Counts were normalized using SCTransform(), and resulting library size corrected counts were used in the downstream analysis of possible gene regulatory dependencies. For visualization purposes, residuals from SCTransform() were used to account for the effects of different numbers of genes detected, percentage of mitochondrial reads, and library size excluding mitochondrial reads. PCA of cells

showed that while the first principle component (Figure 5.22, Figure 5.23 A) was mainly driven by a few cells that behave differently than the rest, the second and third components (Figure 5.23 B) already led to a cloud of points that show a clockwise transition from early (top) to later time points. The first principal component additionally seemed to capture a bias in library size after removing mitochondrial reads (Figure 5.22 A) and in the number of genes measured (Figure 5.22 B). This bias, however, was not linked to specific time points. The negative cells were spread across all time points and did not seem more prevalent in a specific area of the first three components (Figure 5.23 A, Figure 5.23 B). tSNE (Figure 5.23 C) visualization of cells drew a similar picture with a counterclockwise transition from early to late time points. The PCA and tSNE plots indicated that while time points were separable to a certain degree, cells of different time points were mixed at the edges. This resulted from different adaptions to differentiation on the level of single cells. Additionally, the spread of negative cells (no time point assignment possible) across all time points showed that the missing assignments were not specific to one time point but a problem that occurs by somewhat equal chance at all time points.

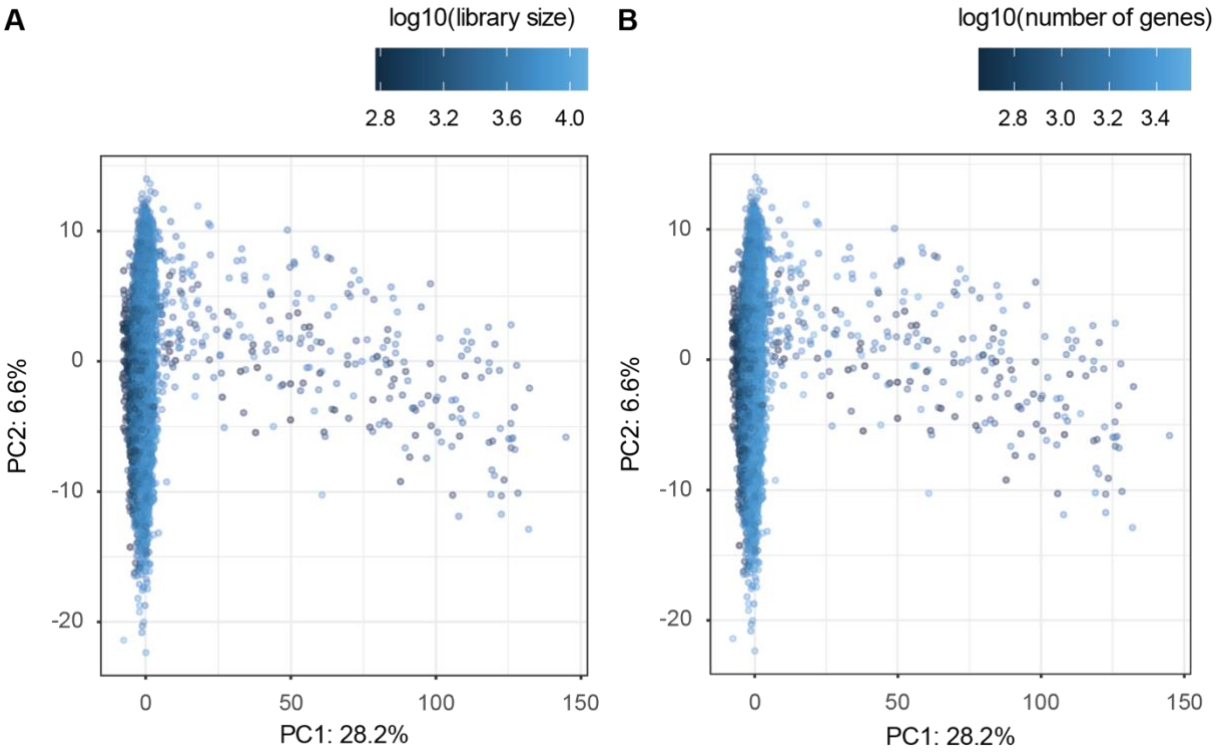

Figure 5.22: First two principal components from PCA analysis of sc time course.

Cells are labeled by A) library size after removing mitochondrial genes and B) number of detected genes.

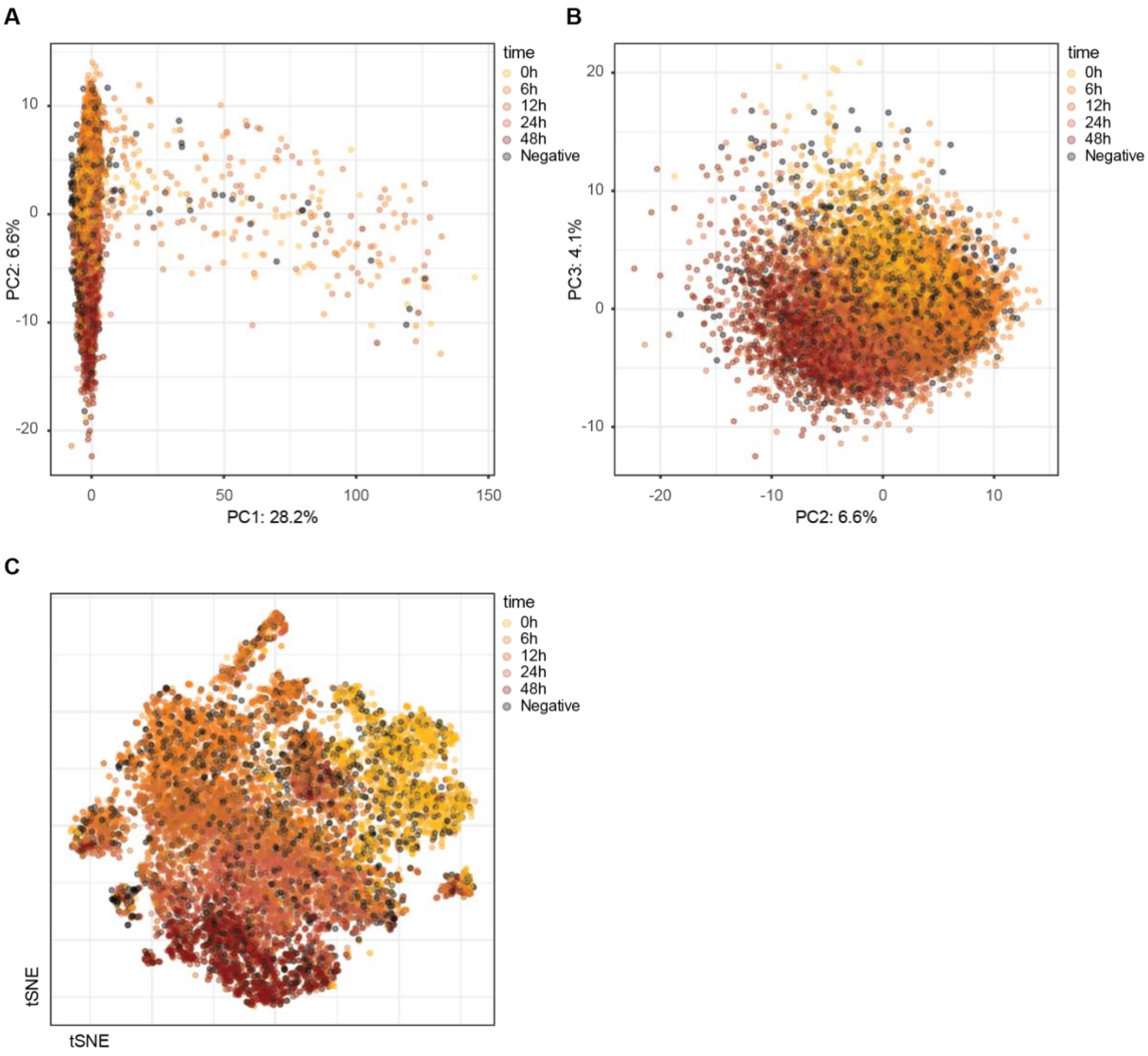

Figure 5.23: Different visualizations of dimensionality reduction approaches colored by time point.

A) All cells in the first two principal components. B) All cells in the second and third principal components. C) tSNE embedding of all cells.

Even though the sc data covered differentiation of cells from 2i to 48 hours of differentiation, we decided to exclude the 48-hour time point from our analysis. Thereby, the sc and KO data cover the same time span and represent the transition from naïve to formative pluripotency. Including the 48-hour time point would also include the following formative to primed transition, where trends of the first 24 hours might be reversed. As negative cells spread across all time points, we had to exclude them to remove all 48 hours cells from the analysis. This led to the exclusion of 3591 cells from the analysis.

Dependencies between genes derived from the KO data were based on the variability that the KOs introduced to the differentiation status of the mESCs through interference with different mechanisms. In the sc data, however, the variability was based on different stages of mESC differentiation in the individual cells. Even though the medium for the cells was changed for all cells, the cells differentiated at different speeds or started to differentiate at different time points. This effect averaged out in

bulk sequencing. Therefore, the approach used on the sc data had to differ from the one used on the KO data. Here, we based possible gene regulatory dependencies on cells that showed early expression behavior of one gene paired with late expression behavior of another. If there was a tendency of cells being more likely to show late expression of gene A with early expression of gene B than the other way around, it is likely that gene A adapts independently of gene B, and dependence of gene B and gene A is possible.

As we used the sc approach to add confidence in the possible gene regulatory dependencies from the KO data, we only considered gene pairs with a possible dependency in the KO data. Genes included in the analysis were also required to show consistent behavior in the first 24 hours of the sc time course. This meant that the average log2FC between cells from 6 hours, 12 hours, and 24 hours to cells from 0 hours had to be zero or share the same sign. Additionally, the direction of log2FCs in the sc data was required to have the same sign as the direction of change in the WT time course. Thereby, one can define early and late expression of genes more reliably as genes are either upregulated or downregulated in relation to the naïve ground state. These filtering steps reduced 1203 genes for possible gene-gene dependencies to 383 genes that can be used in both approaches, KO and sc.

Calculating possible gene regulatory dependencies in the sc data was done based on all 9169 cells from 0 to 24 hours of differentiation. Thus, the analysis was not only based on cells where both genes were measured but also on cells where only one or no gene was measured. Even though for a single cell, it cannot be defined whether no measurement is caused by a biological or technical dropout, over many cells, information about biological dropouts should be retained to a certain degree.

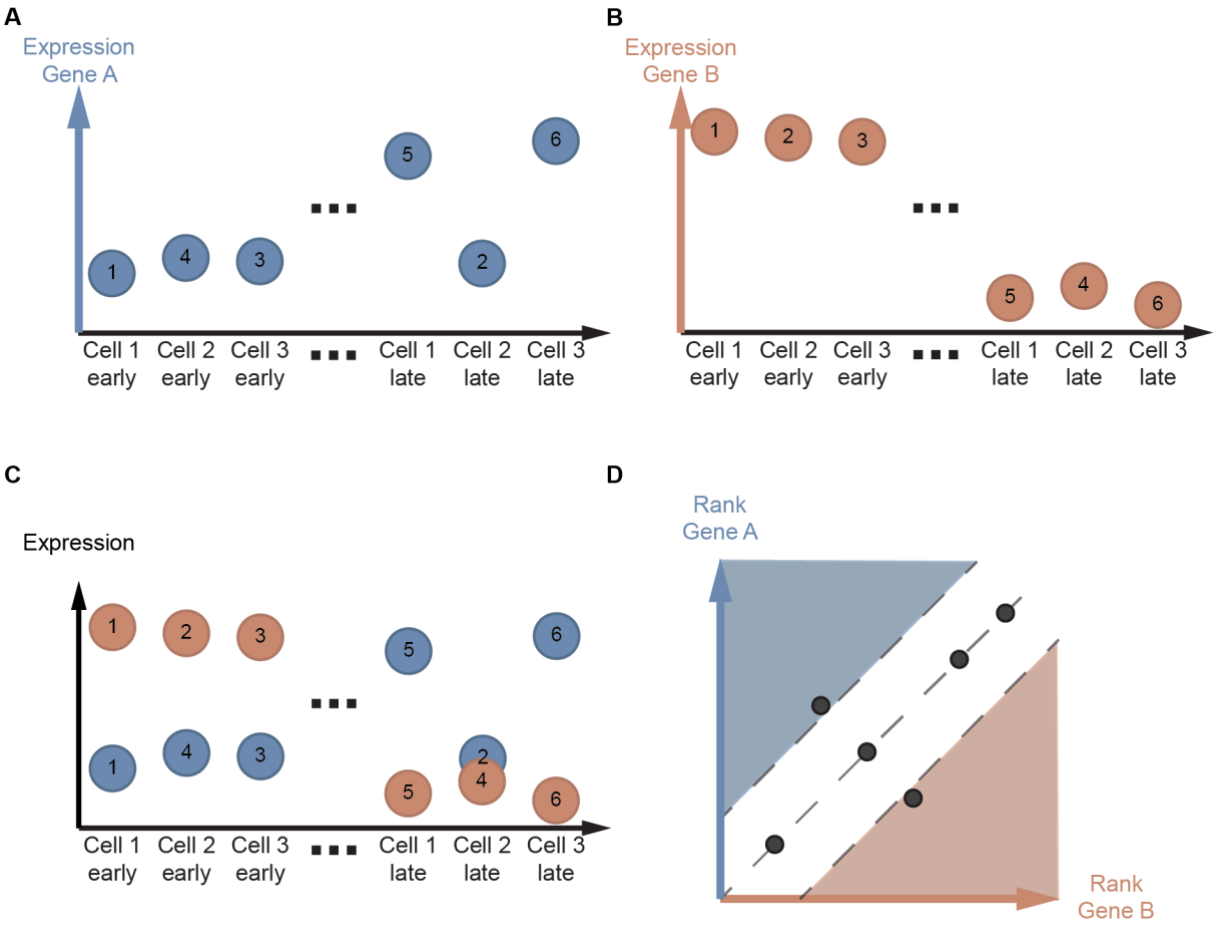

Figure 5.24: Schematic view of approach to calculate gene regulatory dependencies from sc data.

A) Ranking of a gene that is upregulated during differentiation. The cells with the lowest expression have the lowest or earliest rank. B) Ranking of a gene that is downregulated during differentiation. The cells with the lowest expression have the highest or latest rank. C) Pairs of ranks of gene A (blue) and Gene B (red) for each cell. D) Comparison of ranks per cell. Each dot represents a cell, the position on the x-axis represents gene B's rank (early to late), and the position on the y-axis represents the rank of gene A. The white area represents rank matches that are relatively close and is defined through cells with dropouts that should show high expression. Cells in the blue or red area show late expression behavior of one gene and early expression behavior of the other.

The approach used on the sc data is based on rank differences of the two genes being ranked from early to late. Thus, a gene upregulated during the naïve to formative transition had early ranks if it was lowly expressed in a cell (compared to the other cells) and late ranks if it was highly expressed (Figure 5.24 A). This was reversed for genes that were downregulated during the transition, with early ranks for cells showing high expression and late ranks for cells showing low expression of the gene (Figure 5.24 B). The differences in ranks per cell (Figure 5.24 C) were then used to identify genes with early expression behavior of one gene but late expression behavior of the other (Figure 5.24 D). We first used a method to retrieve a supervised pseudotime for all cells included in the analysis to rank genes from early to late. This would give an order from early to late, which was used to rank cells in case of ties in ranking.

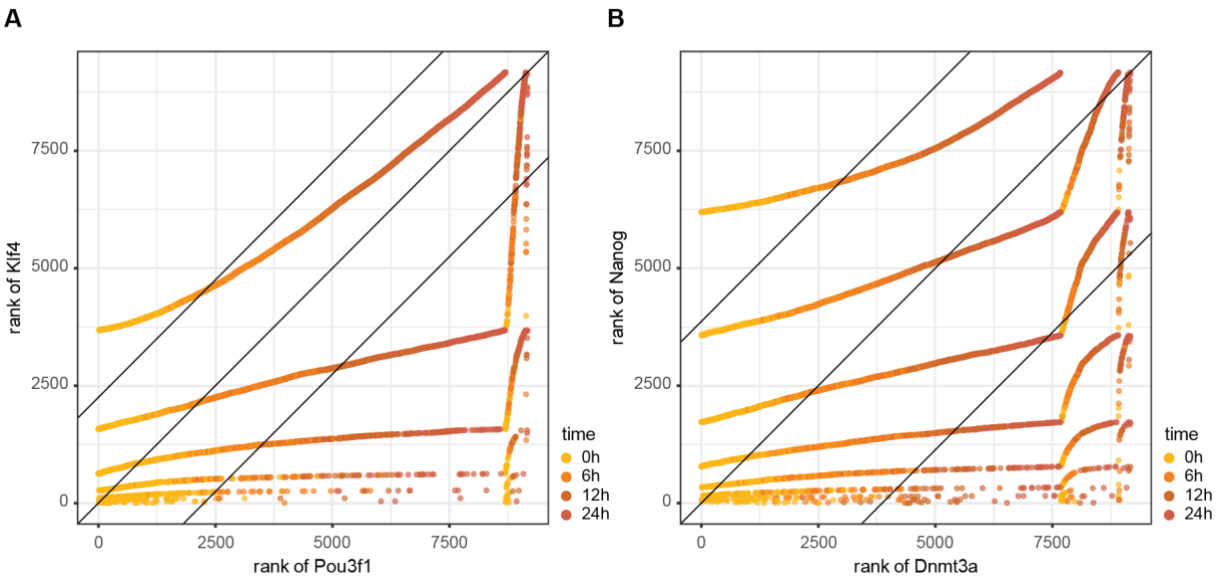

Figure 5.25: Examples of sc-based gene regulatory dependency between A)* Dnmt3a *and* Nanog *and B)* Pou3f1 *and* Klf4*.

The color indicates the time point the cell was sampled, and ranks indicate early to late rankings. The two lines parallel to the diagonal indicate the adaptive buffer zone that separates cells included in later statistical tests (outside of the buffer zone) from the other cells.

As cells with no measurement in both genes might be caused by technical dropouts, an adaptive buffer zone was applied that excluded cells inside this zone to avoid overinterpreting rank differences caused by those cells (Figure 5.24 D, Figure 5.25). This buffer zone scales with the maximal absolute rank difference for all cells that had no measurements for either gene at the time points where high expression is expected for the gene. If both genes were upregulated, all cells without measurements from the 24-hour time point were considered; if both genes were downregulated, all cells without measurements from the 0-hour time point were considered; and if one gene was up- and the other gene was downregulated, all cells without measurements from the 0 and the 24-hour time point were considered. All cells outside of this buffer zone were then used to retrieve the corresponding gene regulatory dependency. To do so, the mean of all rank differences was calculated and scaled by the number of cells in the analysis. Additionally, a binomial test for the number of cells above and beneath the buffer zone along the diagonal was performed to test if the difference was significantly imbalanced. Only dependencies with an adjusted p-value (multiple testing correction according to Benjamini and Hochberg[252]) smaller than 0.05 were considered for further analysis.

Two examples used to illustrate the method further are the dependency between *Dnmt3a* and *Nanog* (Figure 5.25 A) and the dependency between *Pou3f1* and *Klf4* (Figure 5.25 B). In both cases, the gene on the x-axis is a formative marker gene, thus upregulated during differentiation. The y-axis represents a naïve marker in both examples and thus, high ranks represent low expression. Between *Dnmt3a* and *Nanog*, the buffer zone included cells with rank differences of about 3,000. Between *Pou3f1* and *Klf4*, this zone included cells with a rank difference of about 2,000. This shows that more cells that should show high expression of *Nanog* (early time points)

or *Dnmt3a* (late time points) had dropouts compared to *Pou3f1* and *Klf4*. Both examples had more cells that show late expression behavior of the formative marker but early behavior of the naïve marker (more cells in bottom right than top left). This difference of cells in the opposing groups was found to be significant using the binomial test (p-value < 0.05) in both examples. However, after adjusting for multiple testing, the pair of *Dnmt3a* and Nanog did not show a significant difference (adjusted p-value < 0.05).

An additional criterion for a more restrictive or less restrictive analysis was the minimum number of cells required outside the buffer zone. Depending on expression levels and noise in the data, the number of cells outside of the buffer zone might be relatively low. The higher the number of cells outside the buffer zone, the higher the method's robustness. Both the mean dependency and the results of the binomial test gain reliability when based on more cells. While a more restrictive cut-off would make the outcome more reliable, the number of outcomes will be reduced by sorting out gene pairs with fewer cells outside the buffer zone.

5.2.6 High concordance between dependencies from single-cell and knockout approach

The sc dependencies were intended as an additional support layer for possible gene regulatory dependencies derived from the KO data. The need for different approaches in KO and sc analysis was based on use of different experimental designs (bulk RNA-seq vs. scRNA-seq), different sources of variation (effect of KOs on differentiation vs. variability of differentiation in single cells), and different laboratories the experiments were conducted in (laboratory of Martin Leeb vs. laboratory of Christa Bücker). Despite these differences, the consistent dependencies between the approaches can be assumed to be more likely to be causal and driven by differentiation.

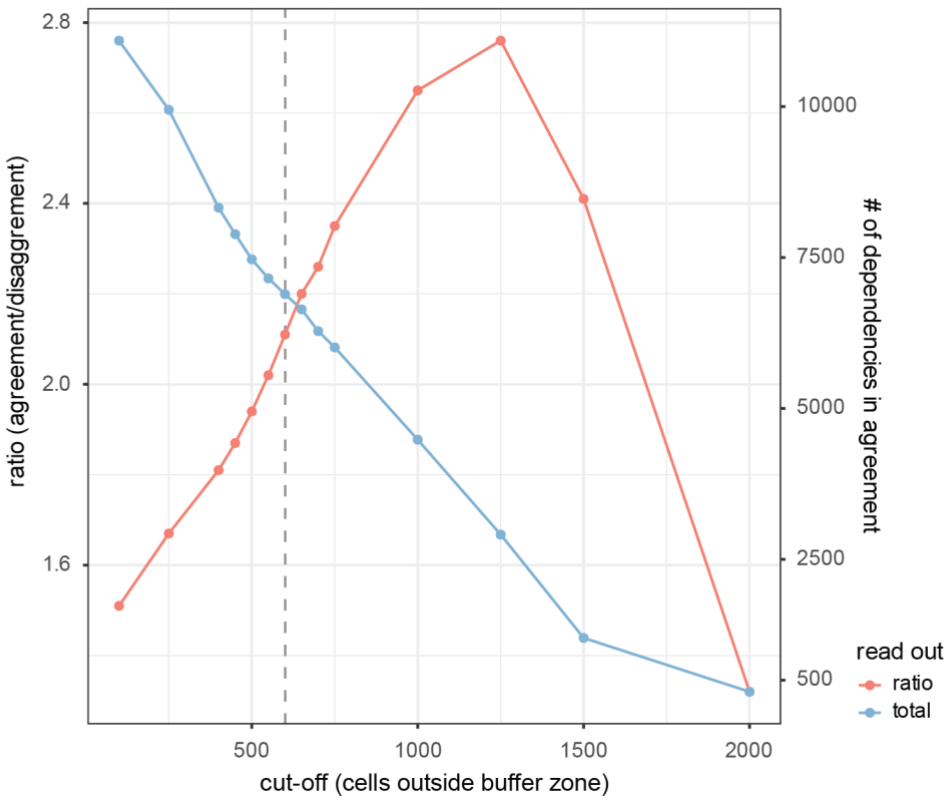

Figure 5.26: Readouts from running sc approach with different cut-offs of the minimum number of cells outside the buffer zone plotted against the used cut-offs.

The x-axis represents different tested cut-offs. The red line represents the ratio of resulting gene regulatory dependencies that agree between the sc- and the KO-derived dependencies. Values for the red line can be read off the left y-axis. The blue line represents the remaining gene regulatory dependencies that agree between approaches. Values can be read off the right y-axis.

Calculating sc dependencies with a very relaxed requirement for the minimum number of cells outside of the buffer zone already led to a bias towards agreement between dependencies from the sc approach and the KO-derived dependencies (Supplementary Table 3). The strength of the agreement was based on the minimum number of cells that had to be outside the gene-gene-specific buffer zone. To a certain point, choosing a more restrictive cut-off (more cells required outside the buffer) led to a higher agreement regarding dependencies between approaches and datasets (red line Figure 5.26). However, when the number of cells required was too high (more than 1,250 cells), this trend reversed, and the agreement between approaches decreased with more stringent requirements. Additionally, the number of included gene regulatory dependencies decreased rapidly with more stringent cut-offs (blue line Figure 5.26). While a more stringent cut-off led to more robust results in the sc data, leading to higher agreement with the KO data, the total number of measurable dependencies decreased. Although the number of dependencies decreased quite drastically with the choice of the minimum number of cells as a cut-off, the number of genes with at least one connection did not decrease as rapidly. Thus, the choice of stringency was a tradeoff between the connectivity between genes and the robustness of the sc approach. As the sc approach was an additional filter on top of the KO analysis, we aimed for a two-thirds agreement ratio, fulfilled under a cut-off of 600 cells outside the buffer.

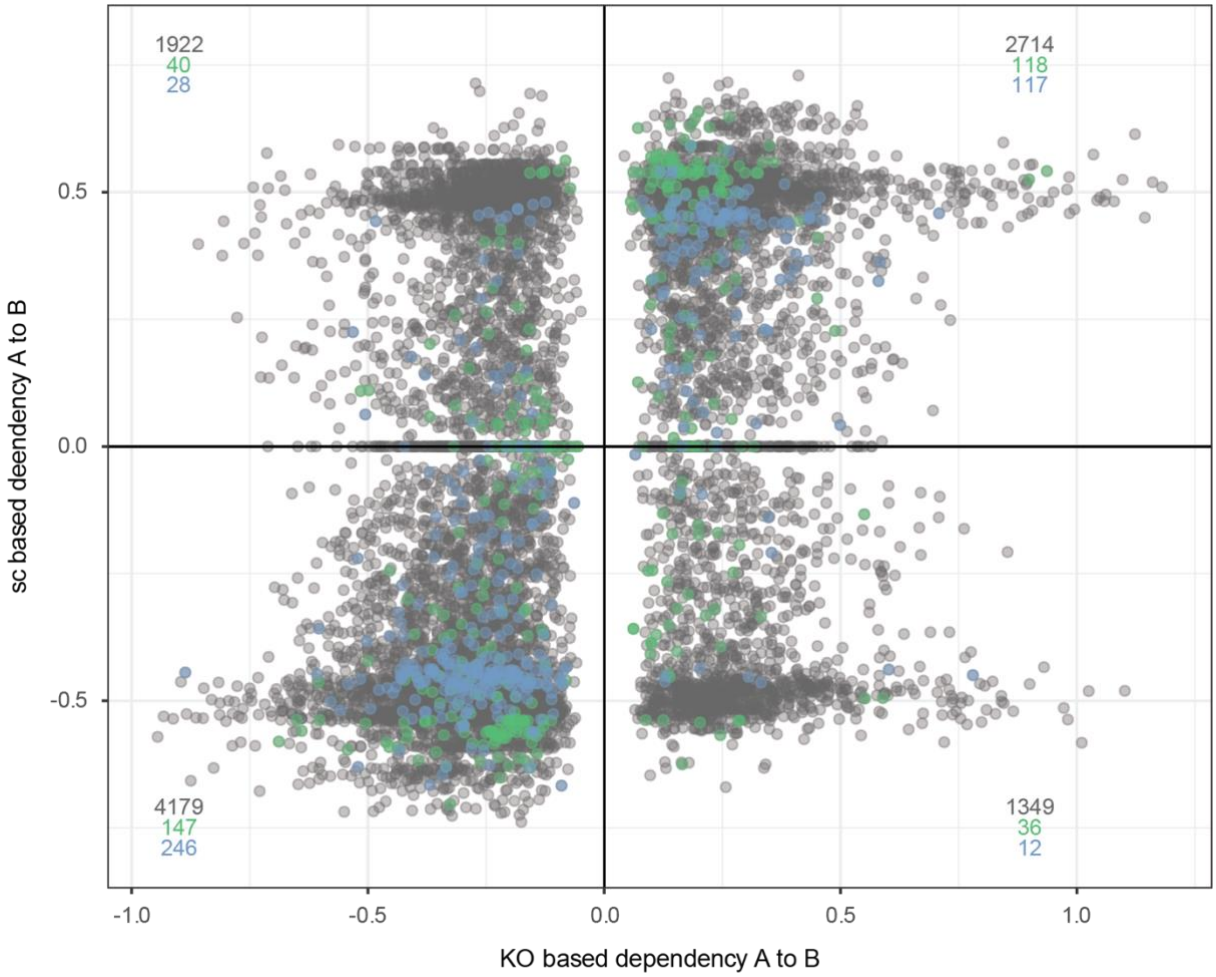

Figure 5.27: Consistency between KO (x-axis) and sc (y-axis) dependencies.

Each dot represents one gene-gene dependency. Blue Dots include at least one formative marker gene, and green dots at least one naïve marker gene. Grey dots neither include a naïve nor a formative marker. Numbers in the corner of each quadrant provide the number of gene-gene dependencies in the corresponding quadrant (If a dependency includes a naïve and a formative marker gene, it is included in the grey, green, and blue color).

Running the sc-analysis with a cut-off of at least 600 cells outside the buffer zone, approximately 60% of all gene regulatory dependencies were in the first and third quadrants, thus agreeing between approaches (Figure 5.27). The agreement between gene regulatory dependencies was even more pronounced, only considering dependencies including at least one naïve or formative marker gene (green and blue points). Here, the agreement between approaches was at approximately 84%. Thus, the 6,893 gene-gene dependencies from the first and third quadrants were considered higher confidence gene-gene dependencies that did not contradict between different approaches based on different experimental designs.

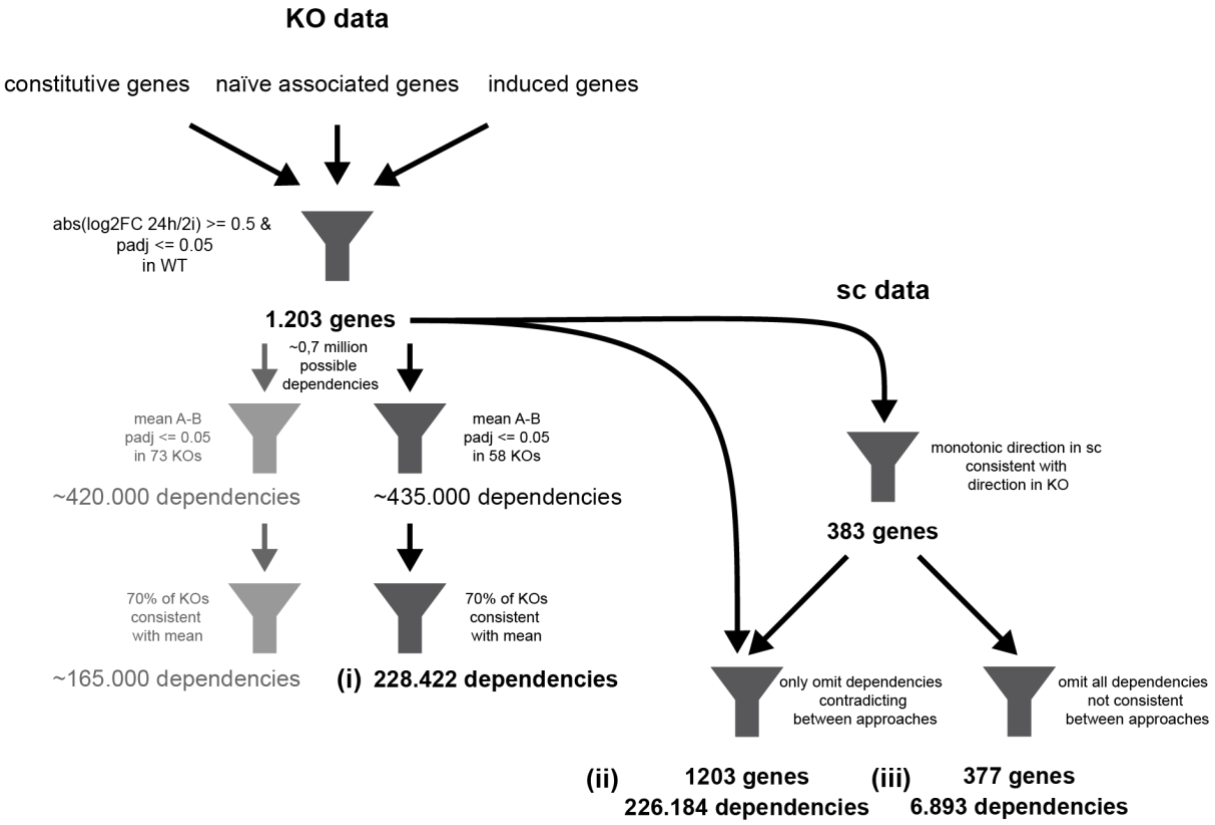

Figure 5.28: Schematic view of filtering steps for KO-based and sc-based analysis and resulting dependencies.

Grey steps on the left side correspond to KO-based analysis, including all 73 KOs. (i) is reached after excluding weak phenotypes and calculating the KO-based dependencies. (ii) is reached when removing all sc-based dependencies contradicting the KO-dependencies from all KO-dependencies. (iii) is reached when only consistent dependencies between KO-based and sc-based analysis are kept.

5.2.7 Dependency network

While gene-wise dependencies help investigate the interaction of very few genes, interpretation of the gene-wise dependencies increases drastically in difficulty with more genes. In the case of the gene regulatory dependencies established in previous sections, the information is quickly overwhelming as all 1,203 genes can be compared to all 1,203 genes. Therefore, it makes sense to condense the gene-wise information to allow for an overview on a more general level. Here, we used the resulting clusters from the NMF to define bigger groups of genes. The average dependency between those groups was used to gain an initial overview and locate genes within these groups. Even though the clusters were used to build the network, one must remember that NMF results in soft clustering and that some clusters were similar to others.

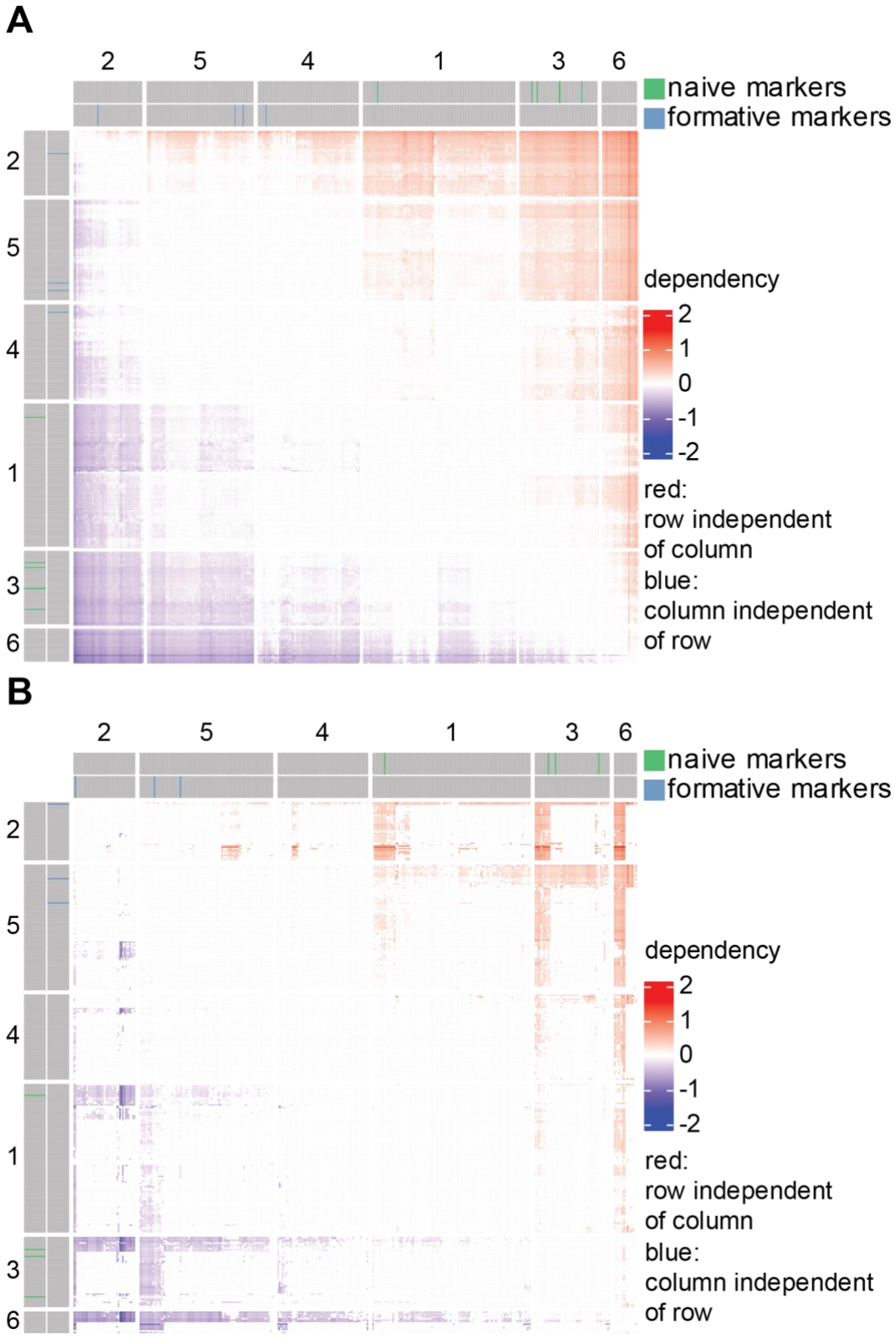

Figure 5.29: Gene regulatory dependencies from Figure 5.16 after filtering for the sc approach.

A) All dependencies that were not consistent between the sc and KO approaches were removed. B) All dependencies that cannot be detected in the sc data or are inconsistent between approaches were removed.

The clustering of genes for the network was always based on the gene regulatory dependencies retrieved from the KO approach. However, the information in the network was based on different steps of previous analysis. The least restrictive layer was the KO-based dependencies without considering the sc approach (Figure 5.16, Figure 5.28 (i)). The next layer was the dependencies left after removing those that disagreed between the KO-based and sc-based approaches (Figure 5.29 A, Figure 5.28 (ii)). Here, genes not measured consistently enough in the sc data were kept. The most restrictive layer of information was only to consider those consistent dependencies between the KO-based and sc-based approaches (Figure 5.29 B, Figure 5.28 (iii)). The difference between the last two layers is that the latter only considers the dependencies that could be tested and agreed upon in both approaches, while the less strict layer removes contradicting information. At the most restrictive layer (Figure 5.29 B), each cluster consist of two groups of genes. The first group of genes was measured in the sc data consistently and had many dependencies to other genes. The second group of genes was not measured consistently and thus is less likely to have enough cells outside of the buffer zone to be kept in the sc approach. This led to a pattern of well-measurable, thus well-connected genes on the top rows of each cluster and connections to well-measurable genes of other clusters on the left columns of each cluster.

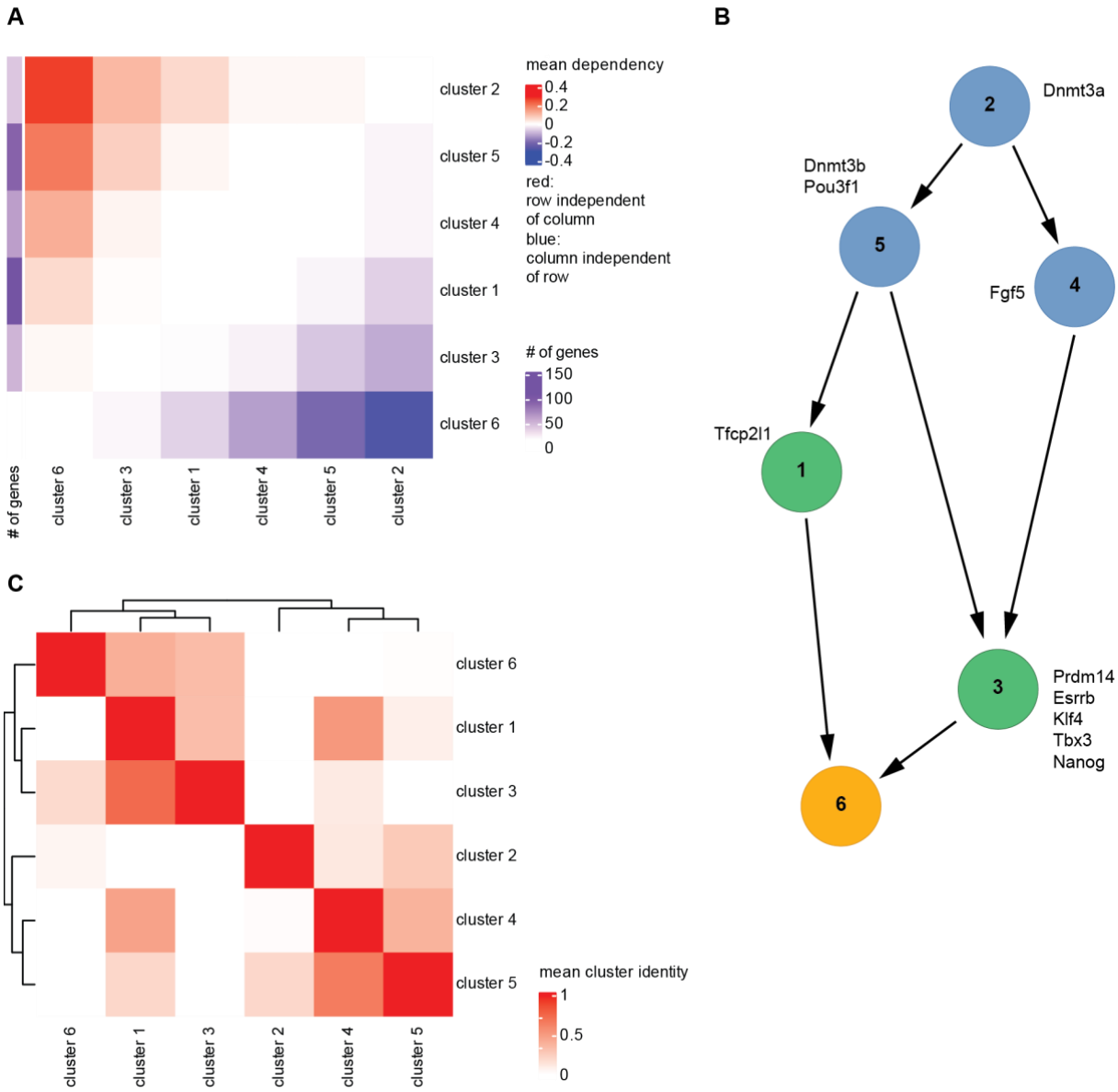

Figure 5.30: Overview of dependencies between clusters after summarizing dependencies on the gene level.

*A) Average dependencies between genes of two clusters (row to column). B) Schematic view of dependency network. The color of the node**s** indicates whether formative (blue), naive (green), or no marker genes (orange) are present in a cluster. Corresponding markers are noted next to each cluster. C) Average identity of genes from clusters in rows to clusters in columns retrieved from NMF.*

To further simplify the network, it has to be viewed on the cluster level instead of the gene level. Even though the most restrictive layer on the gene level reduced the number of genes in the network from 1,203 to 377, this layer still included 6,893 gene regulatory dependencies. In the first step, the gene regulatory dependencies from one cluster to another cluster were averaged to retrieve the mean dependency between all cluster-cluster pairs (Figure 5.30 A). The mean value of cluster-cluster dependencies per row indicated how dependent genes from a given cluster adjust to the differentiation. Cluster 2 had the most positive values. Therefore, genes in this cluster adjusted most independently from genes in other clusters. Whereas genes in

cluster 6 had the most negative values, indicating possible dependence of these genes on genes from other clusters.

We further simplified the resulting cluster-cluster dependencies to build a graph summarizing all dependencies after removing redundant edges (Figure 5.30 B). Here, all direct edges that were a path over multiple other edges were considered redundant and removed. By that, cluster 2 only had two edges, and all other edges were able to be described through paths from clusters 4 or 5 towards other clusters. This network, however, was just a schematic view as the length of edges did not precisely represent the edge weights. The position of a cluster, however, gives an idea of how far upstream or downstream a cluster is located. Cluster 1, for example, was higher than cluster 3, representing a more negative row mean of cluster 3 (Figure 5.30 A). Additionally, the location of each marker gene is represented by names next to the corresponding clusters. *Dnmt3a* was in cluster 2, the most upstream cluster. *Dnmt3b* and *Pou3f1* were in cluster 5, while *Fgf5* was in cluster 4. All naïve markers were in cluster 3 except for *Tcfp2l1*, which was in cluster 1. Thus, all naïve markers were in the intermediate downstream clusters, while formative markers were spread across the three upstream clusters.

The mean cluster identity between clusters (Figure 5.30 C) represented how similar clusters were based on gene identity towards all clusters. The cluster identity of a gene was given by the identity ratio towards a given cluster divided by the highest cluster identity (used to retrieve binary gene to cluster mapping). NMF retrieved these, and the identity of a gene towards a cluster describes how much a gene contributed to a given signature or cluster. When the resulting ratio was close to 1, the gene was considered a good fit in that cluster. If the value was close to 0, the gene did not represent that cluster well. This retrieved information on how similar signatures were in the NMF. While the diagonal, by default, had a ratio of 1, the clusters were split into two groups. The upstream group consisted of clusters 2, 4, and 5, and the downstream group consisted of clusters 1, 3, and 6. Clusters 2 and 6 were the outlier in their corresponding triplets. This similarity was also represented in the structure of the constructed graph (Figure 5.30 B). Clusters 2 and 6 were the most upstream and most downstream, respectively, while clusters 4 and 5 were intermediate upstream clusters, and clusters 1 and 3 were intermediate downstream clusters. Additionally, clusters 4 and 5 shared a high identity towards each other, and clusters 1 and 3 shared a high identity towards each other.

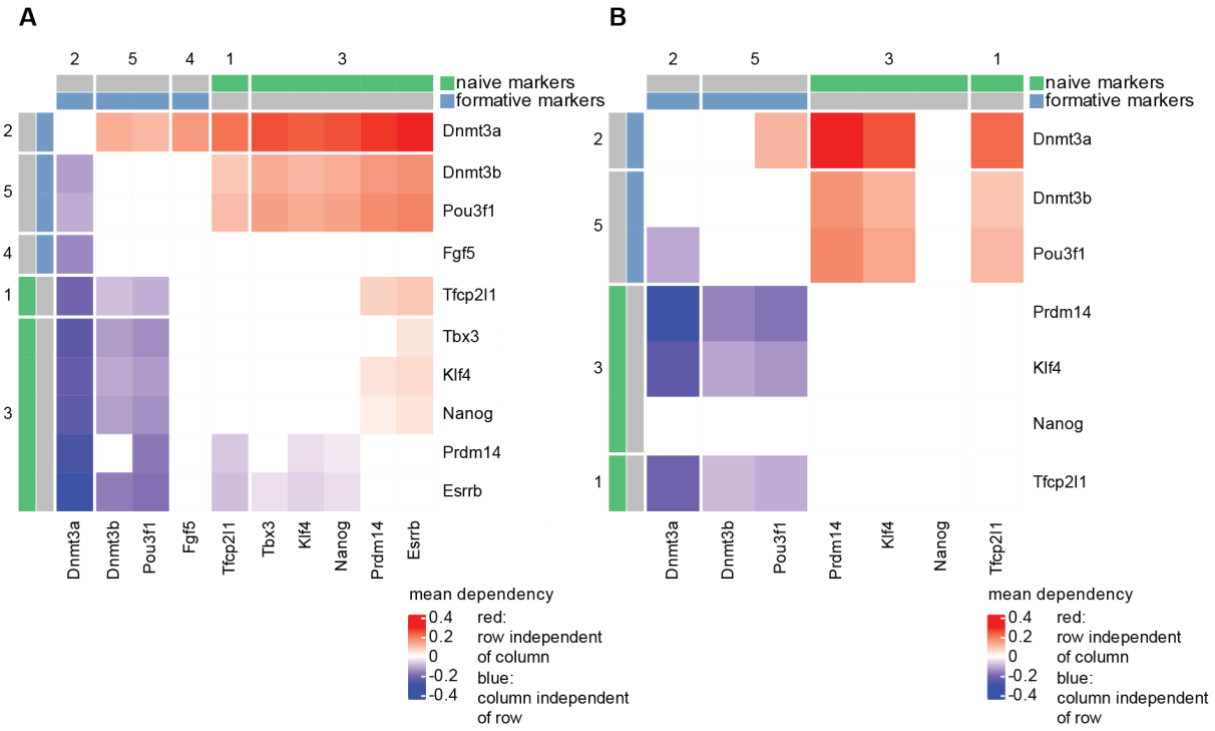

Figure 5.31: Dependencies between naive and formative marker genes that were consistent between approaches.

A) Dependencies between naïve and formative markers that do not contradict between the KO and sc approach. B) Dependencies between naïve and formative markers consistent between the KO and sc approach.

When only considering the dependencies between naïve and formative marker genes on the less restrictive layers (Figure 5.31), the split in naïve and formative markers was maintained. The formative markers still mostly had positive dependencies, while naïve markers had mostly negative dependencies. *Fgf5* was the only exception, with only one negative dependency towards *Dnmt3a* while not connected to any other marker. As clusters were retrieved from the KO-based dependencies and used in the restrictive layers, there cannot be changes in clusters for the markers. However, the overall structure of the marker gene sub-network could have changed in both layers. The medium layer (Figure 5.31 A) still contained dependencies within each set of markers, i.e., *Dnmt3a* was located upstream of *Dnmt3b*, and *Tfcp2l1* was located upstream of *Esrrb* and *Prdm14*. These dependencies were removed in the most restrictive layer (Figure 5.31 B), and the only dependency between markers of the same group maintained was *Dnmt3a* to *Pou3f1*. Additionally, *Fgf5*, *Esrrb*, and *Tbx3* had no connection to any other marker gene and were removed from that subnetwork in this layer.

The different layers and visualization of the gene regulatory dependency network allow for a usage in a broad spectrum of questions. When getting a general overview, the simplified representations allow for first insights, and when picking smaller gene sets, the reduced gene-level view can help ask more specific questions. The general trend of naïve markers showing possible dependence on other genes and formative markers adapting mostly independently from other genes was retained across different network layers.

5.2.8 Independent adjustment of the formative network from the naïve network is observed in independent experiments

The results from previous sections (5.2.3, 5.2.7) revealed that the formative network was located in the upstream part of a dependency network. Additionally, the formative network showed consistent independence from the naïve network across the KO and sc approach. This suggests that while the naïve network keeps the ESCs in naïve pluripotency, it does not promote the emergence of the formative network. To further investigate the independence of the formative marker network from the naïve network, we wanted to incorporate additional experimental data.

RNA-seq data from the medium change experiment (5.1.2) was used. The medium change of interest in this analysis is from 2i to N2B27 and changes to medium containing only one of the inhibitors. The adaption of the ESCs under the influence of only PD or Chiron will help interpret the independence of formative markers from the naïve markers. While Chiron promotes a part of the naïve network through inhibition of *β-catenin* signaling, PD blocks differentiation of the cells through inhibition of FGF/ERK signaling. Additionally, quant-seq data from a dKO after 16 hours in differentiation permitting medium, as well as long-term passaging in differentiation permitting medium, were included in the analysis. The dKO was included as the cells morphologically seemed undifferentiated while also seemingly not being differentiated based on a *Rex1*-GFP reporter system. These systems should give more insights into the changes in the naïve and formative network for additional partially or almost undifferentiated cell lines.

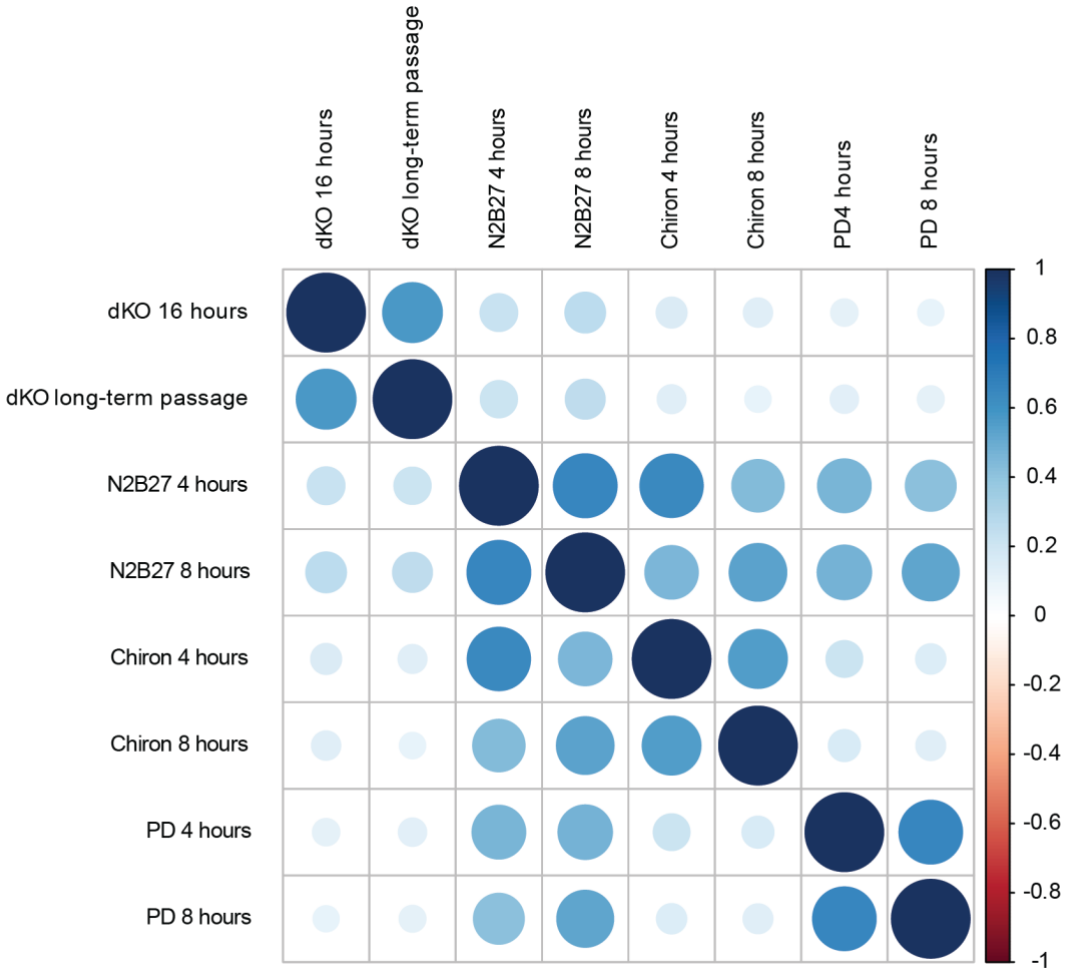

Figure 5.32: Correlation between changes for different conditions.

Size represents the strength of correlation or anticorrelation. Color encodes for the strength and direction of correlation. Blue is correlated and red anticorrelated.

A first overview of those conditions was gained by correlating expression changes of all genes in the corresponding conditions (Figure 5.32). Overall, the changes in all conditions showed weaker to stronger correlation with no example of anticorrelation on the global level of approximately 18,000 genes. However, there was a clear distinction between the dKO and the medium change examples. This is most probably caused by different sequencing techniques and being part of two separate experiments. Even though all log2FCs were calculated in reference to expression in 2i in both experiments, many factors, including using quant-seq, will have an influence. For all four conditions, the changes were most similar to those of the same condition at the second time point. In the case of the dKO, this meant that 16 hours had the highest correlation to the long-term passaged KO, while for the medium change conditions, it meant that the 4 hours and 8 hours samples of the same condition correlated best. Additionally, N2B27 at 4 hours had a similarly high correlation to Chiron at 4 hours as it had to N2B27 at 8 hours. PD and Chiron shared higher similarity with the normal differentiation in N2B27 medium than to each other. This underpins the different mechanisms both inhibitors act through. They do not just cause stronger inhibition in the same mechanism or pathway but synergistically affect different pathways, in consequence blocking differentiation.

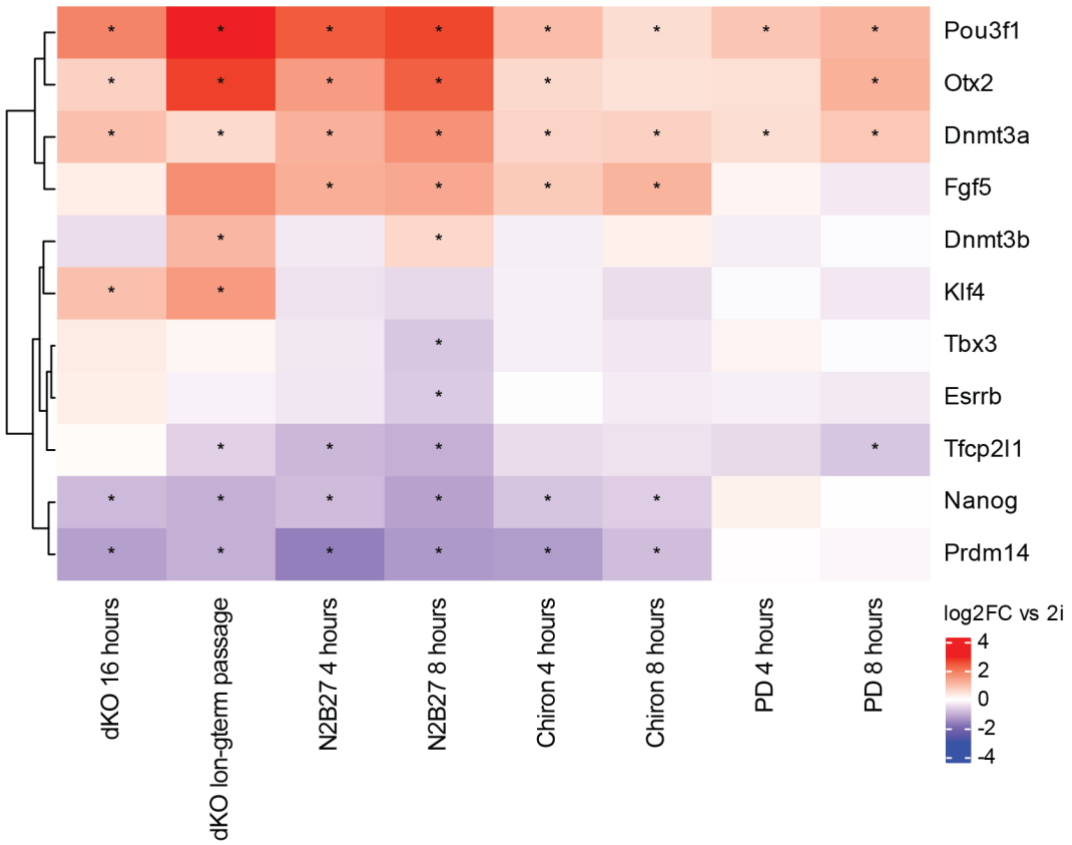

Figure 5.33: Changes of naive and formative marker genes across different conditions.

Color represents log2FC vs 2i, and stars represent significant log2FCs (adjusted p-value <= 0.05 and absolute log2FC >= 0.5).

When comparing expression changes of the naïve and formative network in the described experimental conditions to cells in 2i medium, most formative markers were nicely separated from the naïve marker genes (Figure 5.33). Naïve and formative markers were separated in the first split across the eight experimental conditions considered, except for *Dnmt3b*, which clustered with the naïve markers. This first split separated generally upregulated genes from generally downregulated genes. Even though all conditions but the switch to N2B27 medium delayed or blocked differentiation, adjustment towards differentiation for both networks would still be expected. Thus, formative and naïve markers were expected to separate in these conditions.

A closer look at the formative marker genes *Pou3f1*, *Otx2*, and *Dnmt3a* showed the most substantial upregulation. In all conditions, *Pou3f1* and *Dnmt3a* were significantly upregulated (log2FC >= 0.5 and adjusted p-value <= 0.05). *Otx2* showed significant upregulation in all conditions but 8 hours after replacing the medium with Chiron-containing medium and 4 hours after replacing the medium with PD-containing medium. *Fgf5* was significantly upregulated in the switch to N2B27 and Chiron medium but not in the PD medium and the dKO. *Dnmt3b*, as the outlier of formative marker genes, was only significantly upregulated in long-term passaging of

the dKO and after 8 hours in the switch to N2B27 medium. Most of the other conditions showed weak downregulation of *Dnmt3b* without being significant.

In naïve marker genes, the expression change did not seem to be as pronounced as in the formative markers. *Prdm14* and *Nanog* were the genes with the most significant changes, showing significant downregulation (log2FC <= -0.5 and adjusted p-value <= 0.05) in the dKO, N2B27 medium, and the Chiron medium. However, the expression of these two was almost unchanged in the PD medium. *Tcfp2l1* was the third strongest naïve marker regarding expression changes across conditions. It was significantly downregulated in long-term passaging of the dKO, in N2B27 medium, and at 8 hours in the PD medium. *Tbx3* and *Esrrb* were only significantly downregulated at 8 hours in N2B27 medium and mostly showed weak downregulation for the other conditions. The last naïve marker, *Klf4*, was only significantly regulated in the dKO conditions. Here, it was upregulated, thus in contrast to the general regulation pattern naïve markers followed during differentiation.

The more robust adaptation to differentiation of the formative network compared to the naïve network in the added experimental conditions agreed with previous sections. The number of significantly changed genes and the amplitude of change were higher in the formative marker genes. Thus, the formative network across different experimental setups and conditions adapts independently from the naïve network. Especially the examples of Chiron and PD only media showed a higher degree of adaption in the formative network, whereas N2B27 medium showed a similar degree of adaption to the naïve network. In PD medium *Nanog* showed little to no difference (slight upregulation) as PD interferes with FGF/ERK signaling, which represses *Nanog*. Additionally, FGF/ERK signaling is known to promote the differentiation to the formative state[132,273,274]. However, in the data provided here, the formative network already started to adapt despite the ongoing inhibition of FGF/ERK in the presence of PD. Thus, other mechanisms besides FGF/ERK signaling must control the formative network.

5.2.9 Upstream regulators of formative network

The previous section further supported the independence of the formative network from the naïve network in the context of differentiation by the analysis of single inhibitors and a dKO. Additionally, the adaptation of parts of the formative network was observed despite the inhibition of FGF/ERK signaling in the PD-containing medium. This hints at FGF/ERK independent mechanisms controlling the formative network. To elucidate additional potential upstream regulators of the formative network, we used ATAC-seq data to identify possible TFs that could be in control of this network.

First, we merged ATAC-seq peaks from literature[257] over different conditions. Peaks indicating open chromatin were used for the analysis if at least three samples corresponding to naïve or formative pluripotency had a peak at that position. By that, accessible chromatin was only considered if it was open in enough samples of the

either condition. Second, these regions were searched for potential TF binding sites. CIS-BPs[259] motifs were used to identify motifs in open chromatin via the FIMO algorithm[260]. Next, to link the accessible identified binding motifs of TFs to potential target genes, TFTargetCaller[261] was applied using the “ClosestGene” method. This had the disadvantage of only assigning one gene for a given motif location, but it was chosen to be more restrictive. The resulting TF-to-target gene mappings gave a list of potential interactions between a TF with a binding motif in open chromatin close to a gene's transcription start site (TSS).

As the formative marker genes only consist of five genes, we extended the formative markers by genes showing similar expression patterns across the 73 KOs from our KO data. This was done in parallel to the definition of naïve associated genes in previous work[251]. Additional genes extended the formative genes to be able to capture more potential TF target gene interactions of genes that are regulated similarly to the formative markers. Extending the formative marker set led to 100 genes including the five core formative markers *Dnmt3a*, *Dnmt3b*, *Otx2*, *Fgf5*, and *Pou3f1*. Of those 100 genes, 42 were downregulated and 58 were upregulated in WT differentiation.

Searching for TFs that potentially interact with any of the 100 genes instead of just the five marker genes led to 101 TFs as potential upstream regulators of 74 formative state genes after further filtering steps. First, TFs had to potentially target at least five of the 100 genes. Second, TFs had to surpass a FPKM value of 1 in the WT cells in 2i (KO data). This ensured that the TFs were potentially targeting more genes of this extended gene set and being expressed in the naïve state already to select for TFs more likely to be effective in the early adaption from naïve to formative pluripotency. The resulting heatmap of TF-gene interactions (Figure 5.34) showed that a subgroup of target genes (columns) was tighter connected to the TFs (rows). This group includes *Pou3f1*, *Dnmt3a*, and *Dnmt3b*. *Otx2* and *Fgf5*, however, were in the group of target genes that were less connected. Clustering of the TFs resulted in 7 different potential interaction patterns of TFs to the 74 potential target genes.

As the target calling of TF targets was performed using the “ClosestGene” method, we wanted to ensure the difference in the connection between the two groups of targets was not an artifact. The method used assigned one target gene to a given TF motif site. This gene was always the gene with the closest TSS to the motif's occurrence. Thus, genes in very gene-dense regions would be less likely to be assigned as a target. To exclude that this bias caused the connectivity pattern we observed, we performed a Kolmogorov-Smirnov test in the distance of the TSSs of the 74 target genes towards the closest TSS. The means of the corresponding distances were 43,532 base pairs and 37,749 base pairs in the lowly connected genes and the highly connected genes, respectively. Performing the two-sample Kolmogorov-Smirnov test (two-sided) on both distributions resulted in a p-value of 0.98, indicating that the distributions were not likely to come from different underlying distributions.

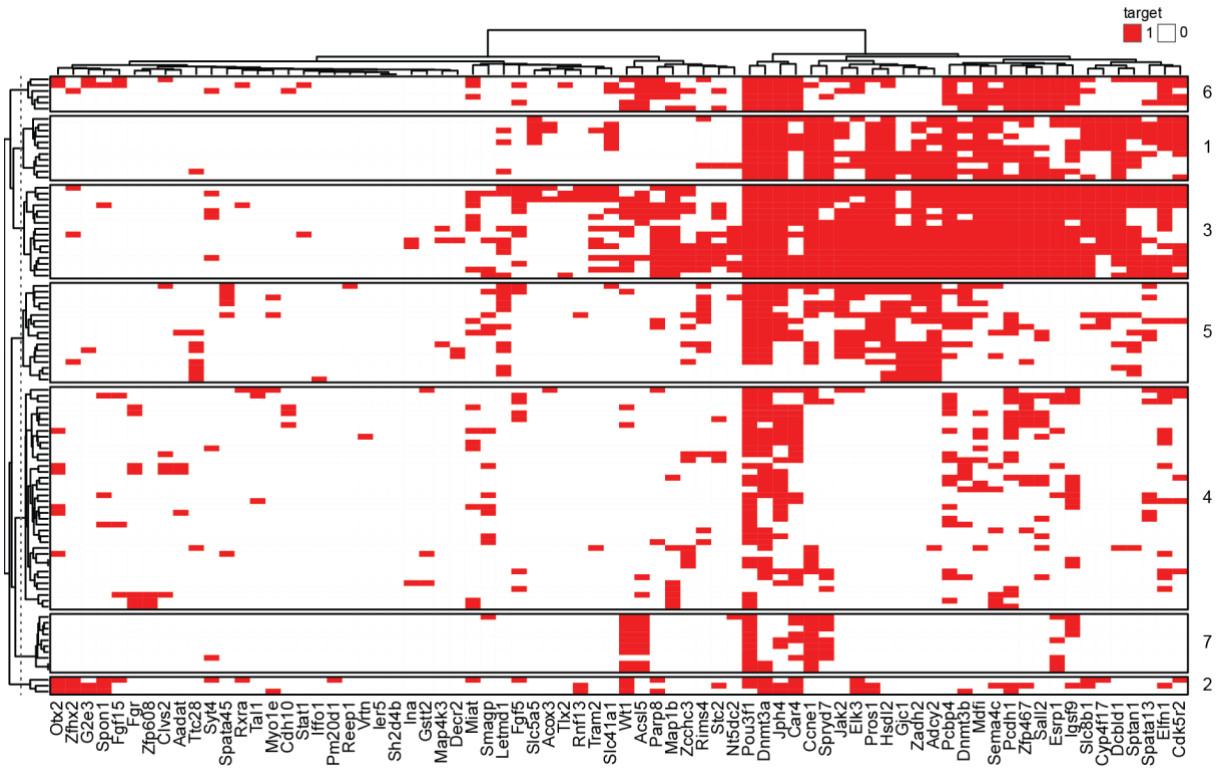

Figure 5.34: Heatmap of potential TF-target pairs.

A target (columns) and TF (rows) pair is marked in red if the TF motif is found in an ATAC-seq peak and the target is assigned as the closest gene.

As the TF-target interaction depended on the binding motifs of TFs and the occurrence of these motifs closely to TSSs of the targets, clusters of similar binding patterns could be confounded by the similarity of binding motifs between different TFs. To test this, all motifs assigned to at least one of the remaining 74 TFs were compared and clustered by similarity (Figure 5.35 A). This resulted in 8 clusters of different sizes. Cluster 1 (turquoise) only had very few red motif-motif pairs and thus was the least homogenous cluster. Clusters 8 and 5 (pink and blue) and clusters 2 and 3 (yellow and purple), on the other hand, were highly similar to each other. Adding the motif information to the TF-target pairs (Figure 5.35 B) revealed that most TF-target clusters represented a certain motif cluster or combination of motifs. TF-target cluster 7 consisted of motif cluster 5, and motif cluster 5 was only found in one other TF outside of TF-target cluster 7. Similarly, motif cluster 2 was the most prevalent in TF-target clusters 1, 3, and 6, which were the clusters with the most connections between TFs and target genes.

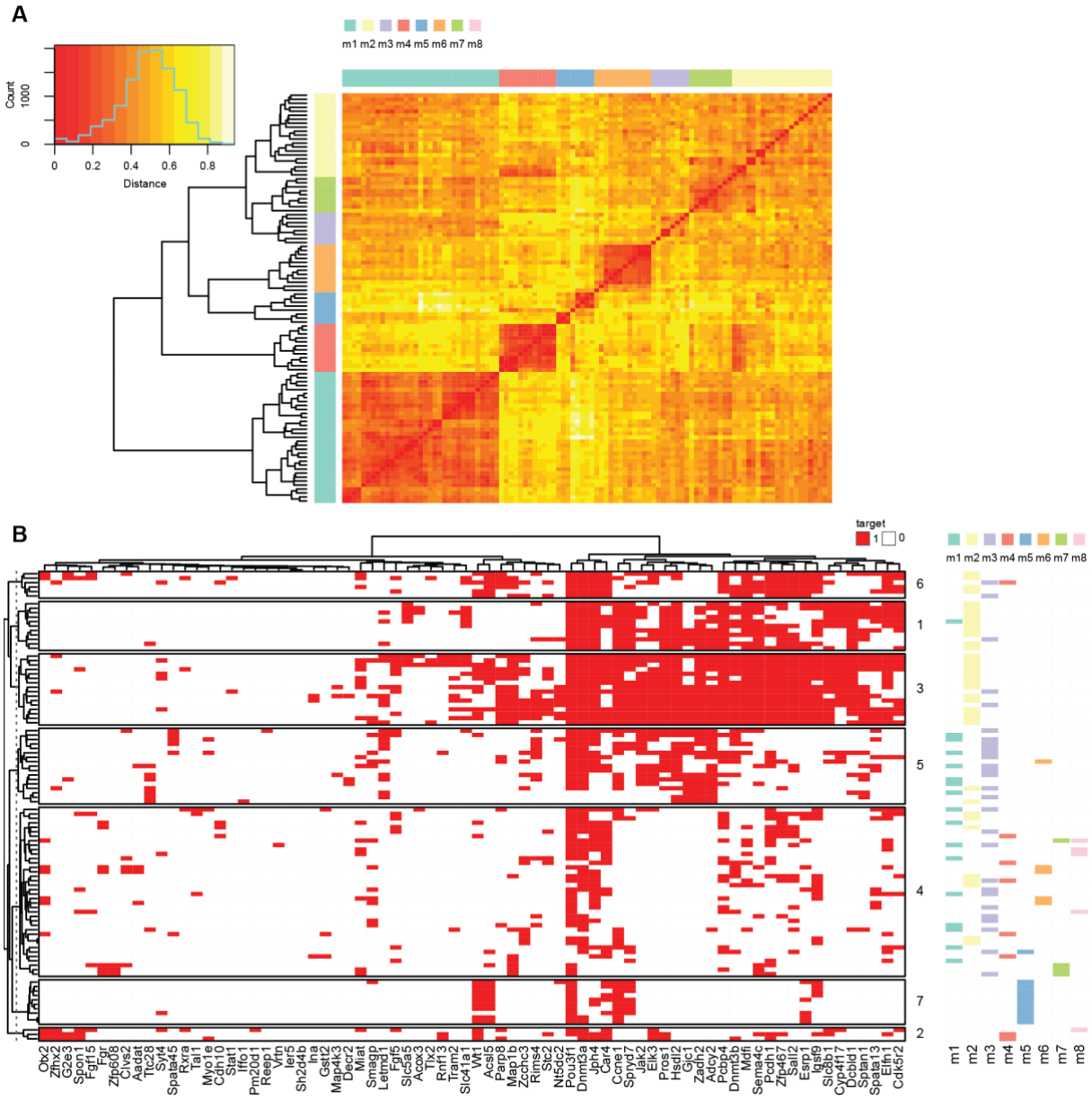

Figure 5.35: Heatmaps assessing the effect of TF motifs on the TF-target interactions.

A) Heatmap of motif-motif similarity for motifs of TFs remaining in the analysis. Color represents distance from red (low distance) to yellow (high distance). The annotation on the side and top shows clusters based on hierarchical clustering. B) Figure 5.34 with added motif clusters to the right of the heatmap.

Due to the overlap of TF-target and motif clusters, we next identified TFs most likely to represent the corresponding combination of TF-target and motif clusters. To decide on possible TFs representing the different combinations, expression values in 2i and change of expression of the corresponding TFs in the 32 hours differentiation time course were used. Here, we assumed that a more abundant TF would substantially affect the cells more than a less abundant TF. Thus, we decided to choose highly expressed TFs to remove redundancy. Additionally, the pattern of change of expression could also hint at the role of the TF in differentiation. After selecting a subset of TFs based on their expression in 2i and the WT differentiation changes, 37 TFs that were connected to 68 target genes from the extended formative genes were kept (Figure 5.36 A). When the TFs were grouped by the

motifs connected to each TF (Figure 5.36 B), there were 14 clusters of motif combinations. The biggest group was motif cluster 2, followed by motif cluster 3. Motif cluster 2 included three outlier TFs with PLAGL1, TFDP1, and TFDP2 that showed the most different TF-target patterns from the rest of the TFs in this motif cluster. The only bigger group of TFs represented by a mixture of motif groups was a combination of motif clusters 1 and 3. This group consisted of the TFs RFX2, E2F2, and ZIC3. Interestingly, ZIC2 did not share the motif from motif cluster 1 but still overlapped with the TF-target profile of ZIC3. All potential interactions that were present in ZIC3 were also identified in ZIC2. However, in line with an additional binding motif assigned to ZIC3, it also had additional TF-target interactions compared to ZIC2.

Even though the TFs have already been selected for expression values, some can be grouped further. Aside from the example of ZIC3 and ZIC2 already mentioned above, other TFs were also further grouped based on TF-target and motif clusters (Figure 5.36). SOX2, SOX3, and SOX4 clustered together in the TF-target clusters, and all have a motif connected from cluster 7. MYC, MYCN, and MAX were also connected to the same motif cluster and showed similar TF-target interactions. The same was applied to TFDP1 and TFDP2; DNMT1 and NRF1; SP1 and ZFP341; ZFP281, ZFP38, and MAZ; as well as ZFP740 and SALL4.

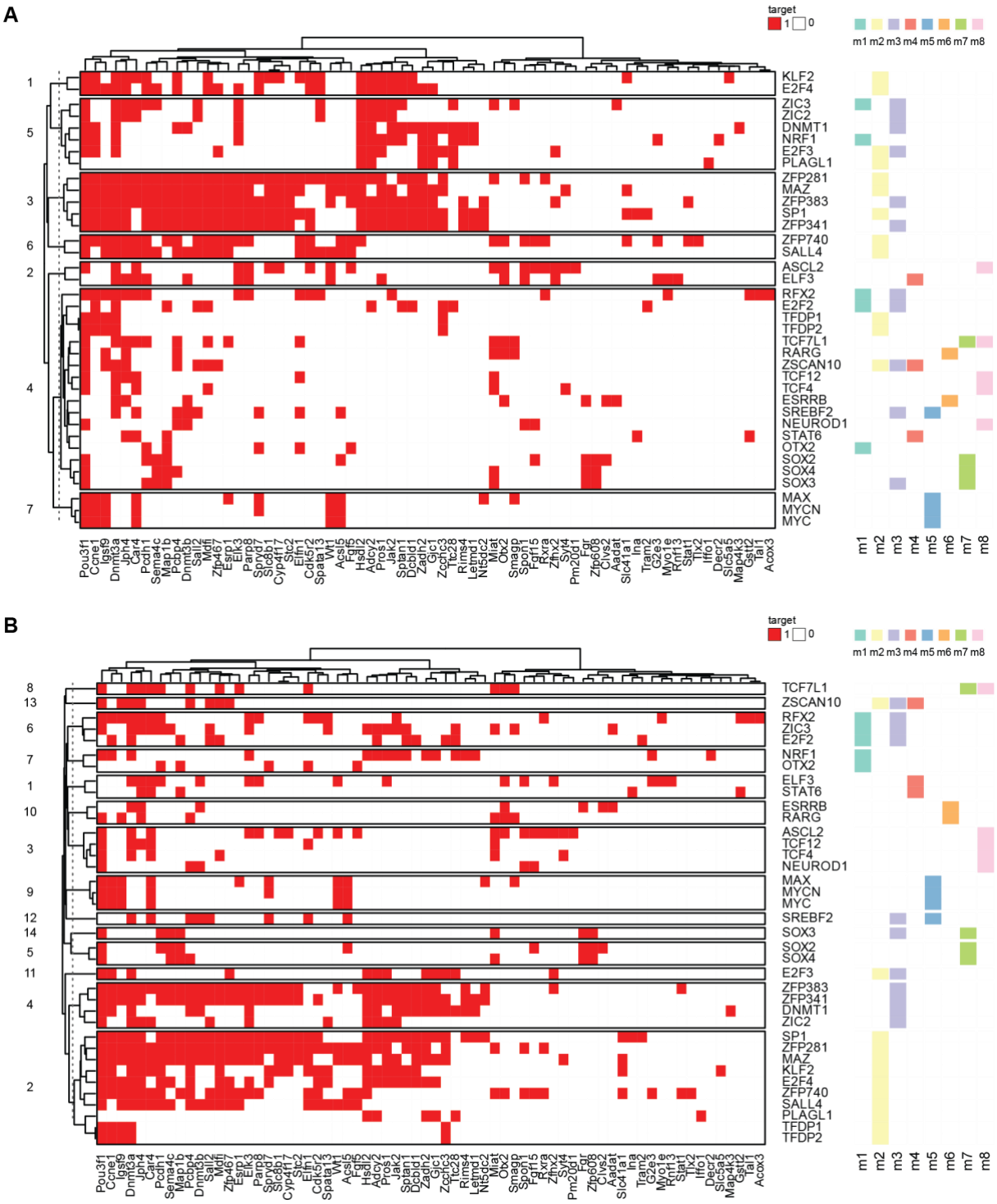

Figure 5.36: Heatmaps of TF to target mapping from Figure 5.34 only show TFs that were selected to represent TF-target and motif cluster combinations.

A) Rows of the heatmap are clustered by TF target mapping. B) Rows of the heatmap are clustered based on motif clusters.

Using the potential TF-target interactions between selected TFs and the extended formative genes, we built a model of potential upstream regulators of the formative network (Figure 5.37). In addition to potential upstream regulators identified through the analysis of ATAC-seq data, known influences on the naïve and formative TF network were integrated into this model. The influence of LIF and 2i medium and how

these affect the naïve network are well established (Figure 5.37 A). The potential upstream regulators proposed here are a group of TFs to investigate further processes regulating the establishment of the formative TF network (Figure 5.37 B). One pathway known to regulate the establishment of the formative state is FGF/ERK signaling, which is also subject to one of the inhibitors. However, in the previous section (5.2.8), we found that even under FGF/ERK inhibition, parts of the formative network have already adapted to differentiation signals. In addition to the potential regulatory effect on formative extended genes, we also investigated TFs with the most potential to affect the formative markers. ZFP281, ZFP383, MAZ, ASCL2, ZFP740, and SALL4 all potentially regulated four of the five formative markers, and ELF3, ZFP341, SP1, DNMT1, TCF7L1, and ESRRB potentially regulated three of the five formative markers.

The potential upstream regulators of the formative network include candidates that are well-studied in their interactions with the naïve network and importance for processes involved in differentiation. ESRRB, for example, is one of the potential upstream regulators of the formative markers and extended formative genes. While it is established that ESRRB regulates the naive network[210,271,275,276] it has only recently been shown that ESRRB also regulates the formative network[240]. TCF7L1, also known as TCF3, is a regulator of ESRRB[210] and a known pathway member of canonical WNT signaling[204]. Genes of the TCF/LEF family are known repressors[277], and TCF7L1 is known to repress expression of *Nanog*[278]. The role of TCF7L1 as a repressor of the naïve network and its influence on the balance between self-renewal and differentiation is well decribed[212,279]. ZIC2 and ZIC3, together with OTX2, are candidate TFs important in the redistribution of OCT4 binding in the naïve to primed differentiation[280]. Depletion of *Zic3* additionally highlights the importance of ZIC3 in the naïve to primed transition[281]. TFDP1 has recently been identified as a modulator of global chromatin accessability[282], which aligns with the reorganization of chromatin accessibility in the naïve to primed transition[138,140,146,147]. Depletion of *Tfdp1* also led to a gain in efficiency in the induction of PSCs [283].

SALL4 and ZFP281 are candidates for upstream regulators of the formative network that were identified to potentially target multiple formative markers. SALL4 was found to be important in the early differentiation of mESCs and identified as a regulator of *Oct4*[284,285]. In combination with other factors, such as ESRRB and NANOG, SALL4 has also been used to induce pluripotency[286]. ZFP281 is a zinc finger TF well known to affect pluripotency through direct activation or repression of its target genes[287–292]. The TF has been identified as an important interaction partner for the core pluripotency factors NANOG, OCT4, and SOX2[287]. It was also shown to be important for establishing and maintaining primed pluripotency[291] and the interconversion of naïve and primed pluripotency[293]. Additionally, a recent study identified ZFP281 as a transcriptional activator of the formative genes *Dnmt3a* and *Dnmt3b*[294].

Taken together, across different approaches and datasets, we saw the adaptation of the formative network to differentiation independent of the naïve network. Additionally, parts of the formative network already exhibited adaptation to differentiation while cells were still exposed to FGF/ERK inhibition. This led to the

assumption that other factors must be crucial in establishing the formative network. The ATAC-seq data and TF motifs analysis in this chapter were used to propose a set of potential upstream regulators of the formative network. Some of these candidate TFs are well studied in the context of their interaction with the core pluripotency network or the naïve network. While the interaction of ZFP281 and the formative network has been shown recently[294], the role of the other candidates on the formative network remains to be elucidated.

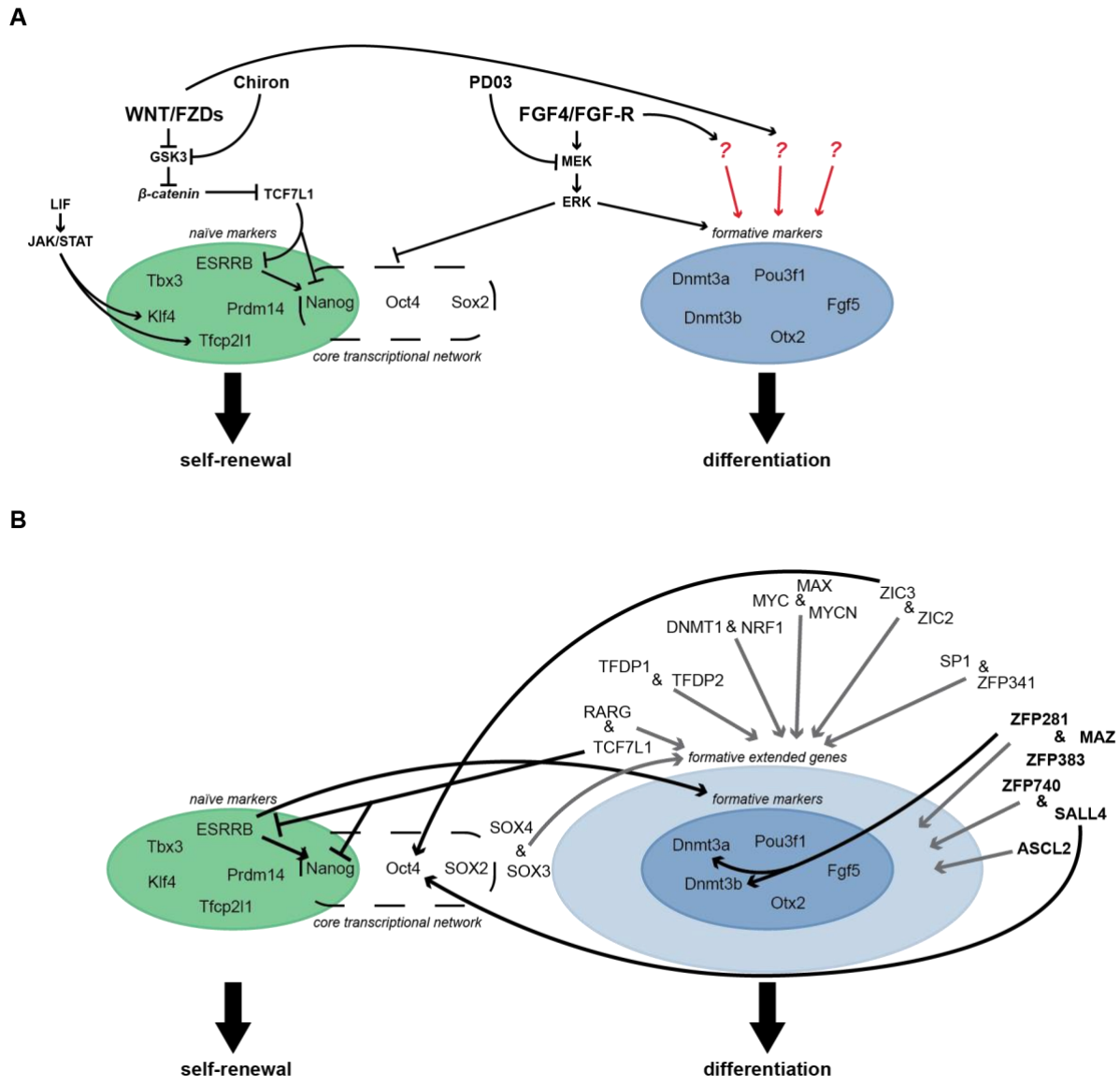

Figure 5.37: Regulation of the naïve and formative networks.

Naïve and formative core networks and corresponding genes are shown in green and blue circles, respectively. The dashed rectangle includes three factors described as core transcriptional networks in the KEGG pathway (annotation of mmu04550). Pathways and inhibitors regulating the naïve network are shown on the left. Edges with arrowheads indicate induction of the target gene, and edges with an orthogonal line indicate repression of the target gene. Edges in black are known interactions from literature. Grey edges indicate potential regulators of the formative network.

6 Discussion

In this thesis, we investigated the mechanisms that break the circuit of self-renewal of naïve ESCs and initiate differentiation to the formative state. Our results show that the transcriptional adaptation of genes from the formative network is independent of genes of the naïve network. This independence is consistent in two different approaches to infer regulatory dependencies and different underlying data. The first analysis indicating possible independence of the formative network from the naïve network is the timing of 73 KO cell lines based on both networks. This is further supported by systematically testing independence and possible dependence between gene pairs using only a subset of the KOs. Examination of the adaptation of the formative network using independent external scRNA-seq data further supports the conclusions retrieved from our data. Based on our findings, we propose 22 different TFs, including ZFP281, TCF7L1, ZIC3, and ESRRB, as potential upstream regulators of the formative network. These TFs might play an important role in breaking the circuit of self-renewal and promoting differentiation.

6.1 Wildtype differentiation

Analysis of the 32-hour differentiation time course revealed that most transcriptional change in the naïve to formative transition is initiated with the onset of differentiation. Many genes that were induced or repressed at the onset of differentiation were exposed to a reversal of the initial transcriptional change after 6 hours of differentiation. However, we demonstrated that the process of changing the medium itself only had minimal impact on the observed transcriptional change. Additionally, quantification of differentiation delays of the 73 KO cell lines was the first analysis revealing that transcriptional adjustment of the formative network might be independent from the naïve network. Here, the dense 32-hour differentiation time course was used as a reference for naïve to formative transition in WT.

For the 32-hour time course, we decided to use denser sampling instead of measuring replicates at fewer time points. Thus, the resulting setup included measurements every 2 hours without replicates, except for the 2i, 2i LIF, and 32 hours measurements, which were done in duplicates. This was done assuming that close neighboring time points contain information about their neighbors and thus allow for an estimation of noise without replicates. Even though there is no considerable body of literature concerning this issue, a study that tested this assumption on simulated and collected data concluded that denser time points are more beneficial than more replicates[295].

The denser time points were also beneficial for applying GPR to the time course data. GPR was used to retrieve a function representing the changes in expression that accounts for noise in the data. The noise is assumed to be approximately the same for all time points[296,297]. Additionally, the correlation model used for GPR here assumes higher correlation for close-by time points[249]. These two assumptions were used to retrieve smoother output functions from the GPR. Even though the GPR led

to smoother functions while maintaining the trends of the expression changes, fast expression changes only captured in one or two time points can be lost. This was beneficial in cases such as *Tcfp2l1* (Figure 5.2), where the 6-hour time point seems to be an outlier and is not represented in the outcome function. One potential downside of GPR is that it could miss a rapid biological burst in expression with fast degradation. However, in this scenario the burst in translation and degradation of the mRNA would have to take place in a window of a maximum of 4 hours. Considering that the mean half-life time of mRNA in the naïve to formative transition was found to be approximately 2 hours[298], this scenario might occur in some genes.

To answer the question of when genes are regulated in response to the onset of differentiation, we clustered genes based on kinetics in the 32-hour differentiation time course. Additionally, we identified specific time points at which genes were regulated. The clustered kinetics and the specific regulation points help gain an overview of how genes of interest behave in the naïve to formative transition. While the points of induction and repression indicate how a gene behaves at specific time points, the unsupervised clustering reveals more general trends. In addition to providing an overview of the regulation of genes, the analysis of induction and repression time points in the WT differentiation revealed that most regulation takes place early in the naïve to formative transition.

Most genes are regulated at 0 hours; however, approximately 20% of genes show a reversal of the initial direction around 6 hours after changing the medium. The immediate regulation at 0 hours with reversal at 6 hours opened the question of whether this was an artifact of the medium change from N2B27 + 2i medium to N2B27 medium. Previous research has shown that transcriptional adjustment to heat shock differs depending on whether fresh or conditioned medium was used[299]. There were no previous studies on removing the conditioned medium and adding fresh medium in the setup of naïve mESCs. Therefore, we conducted additional experiments to separate the effect of changing the medium from the differentiation-associated effects. The effects of changing the medium could either be caused by stress as cells are without medium for a short time when replacing the medium or caused by replacing the conditioned medium with fresh medium. It is important to state that the differentiation-associated effects cannot be disentangled using our experimental setup, as removing the two inhibitors, Chiron and PD, leads to differentiation. Thus, the change of medium composition coincides with the onset of differentiation. Differential expression analysis of cells cultivated in the fresh 2i medium 4 hours and 8 hours after replacing the medium showed only 33 and 44 DEGs, respectively. However, these numbers were drastically increased when removing both inhibitors from the medium. This shows that the transcriptional reversal we observed at 6 hours was neither an artifact of the stress of removing and adding medium, nor an artifact of replacing the conditioned medium with fresh medium. In contrast to findings of the study on the heat shock reaction[299], we demonstrated that the process of changing medium alone has minimal impact on the transcriptional level in the naïve ground state.

6.2 Quantification of differentiation delay

In order to quantify the differentiation delay of the 73 KO cell lines, we mapped the KOs on the 32-hour WT differentiation time-axis. This was done for the naïve marker genes, the formative marker genes, and the DEGs between 24 hours of differentiation and 2i in the WT. So far, we only applied the quantification of differentiation delay to those three groups of genes. However, the same approach can be applied to further gene sets to gain insights into their differentiation delay over specific KOs or all KOs. This can be useful to answer questions about the effect of certain KOs on critical pathways or to compare delays of different pathways as a consequence of the KOs.

The gene set used to quantify the differentiation delay will heavily impact how representative the timing of the KOs is. This mostly depends on two factors: First, how well the corresponding genes correlate with the differentiation status. As the differentiation axis in our experiments is unidirectional, genes with monotonic kinetics correlate well with differentiation. The first point is directly connected to the kinetics of the corresponding genes. The genes most suitable for timing the KOs behave monotonically over the time course, such as naïve or formative markers. Genes that might cause problems would be genes that plateau, thus only adding precise information to the timing in the time frame where expression does not plateau. Genes that peak in the middle of the 32-hour time course might have multiple time points with the same change towards the naïve ground state. This results in multiple time points with an equally good fit, and timings of other genes must be incorporated to resolve these cases. Second, how many genes are in the group, and what percentage of the genes correlates well with the differentiation status of the cells. The non-monotonic kinetics described above can be compensated when monotonic genes are used to resolve cases of plateaus or peaks. If enough monotonic genes are part of a gene set, it should still be sufficient to resolve those cases, and it should be possible to use the gene set to quantify the differentiation delay.

While the potential weakness of the quantification of KO delay has to be considered for each gene set, the timing based on naïve and formative markers is less susceptible to those weaknesses as all genes from these groups are monotonically down or upregulated, respectively. While the DEGs have a similar order in rankings of delays compared to those based on the naïve marker genes, the timings appear noisier. This hints at added noise caused by non-monotonic genes included in the analysis.

Overall, the timings based on the naïve networks align with known effects and observations of the cells. The KO cell line of *Csnk1a1*, for example, appears undifferentiated when it comes to the time axis. This agrees with the lack of cell shape changes observed in the first 24 hours of cell culture under differentiation-permitting conditions[251]. Other KOs that showed a substantial delay of differentiation were *Tcf7l1* and *Zfp281*. Both TFs are known regulators of the naïve network and interact with the core pluripotency network[210,278,279,287,288,290]. TCF7Ll1, also known as TCF3, is an inhibitor of expression of *Esrrb*[210] and Nanog[278]; thus, the strong delay of

the naïve network under depletion of *Tcf7l1* can be linked to the lack of downregulation, at least of parts of the naïve network. The strong delay of both the naive and the formative network under KO of *Zfp281* also aligns with the known interactions of the TF with parts of both networks[287,288,294]. Additionally, the KO of *Myc* was included as a control as it is known to accelerate differentiation when depleted[251,300]. This is consistent with being either the most progressed KO (naïve network and DEGs) or in the top three most progressed KOs (formative network).

One notable case of KOs of the same complex with diverging behavior are *Jarid2*, *Suz12*, and *Eed* as members of the polycomb repressive complex 2 (PRC2). While the three KOs are relatively similar in their timings based on the naïve network, ranging from 14 to 17.75 hours, there is a uncoupling when considering timings based on the formative network. Here, the *Suz12* and *Eed* KOs show an acceleration of 8 and 1.75 hours, respectively. The recruiter of PRC2, *Jarid2*, leads to a delay of 13.75 hours when it is depleted. The different effects on the formative network of the PRC2 KOs might be interesting to follow up on to find the cause of this disconnect.

Aside from allowing quantification of the KOs, the quantification of differentiation delay analysis provided the first hint at a disconnect between the transcriptional adaptation to differentiation for the naïve and formative networks. The timings of the KOs based on the three different gene sets (Figure 5.9, Figure 5.10) highlight a stronger delay of the naïve marker genes compared to the formative markers and DEGs. Additionally, the order of strengths of delays was relatively consistent between using naïve markers and DEGs. However, the order of KO delays between using naïve markers and formative markers was more inconsistent. This shows a disconnect between the naïve and formative networks in quantifying differentiation delays of the KO cell lines. Additionally, the more consistent order of strengths of delays between naïve markers and all DEGs suggests that the downregulation of the naïve network might function as a gatekeeper to many genes in their adaptation to differentiation.

6.3 Identification of gene regulatory dependencies

The disconnect between the naïve and formative network revealed by the timing of KOs motivated the analysis of potential gene regulatory dependencies. Here, we compared the log2FCs at 24 hours vs 2i of each KO to the corresponding log2FCs in the WT. This resulted in the distribution of WT completion per gene, and the differences in WT completion for a gene pair were used to calculate the dependencies. Interestingly, this approach was sensitive to the phenotypic strength of the KOs used in the analysis. We showed that KOs with weak or no differentiation delay phenotype were more likely to contradict the overall direction of the resulting dependency. Removing those KOs from the analysis led to higher consistency of results and more dependencies with significant differences to zero. The latter point is more outstanding, as excluding KOs from the analysis decreases statistical power from the test. However, with the exclusion of the KOs, we also seemingly removed noise from the data used for the approach that compensates for the loss of statistical

power. The correlation between the phenotypic strength of the KOs included in the analysis and the sensitivity of the analysis shows that the dependencies capture the biology behind the differentiation of the KOS.

The analysis of gene regulatory dependencies is based on the percentage of WT completion in each KO. Thus, we compared the log2FCs 24 hours vs. 2i for each KO against the log2FC 24 hours vs 2i of the WT. This has no effect if the KO does not already influence the transcription of a given gene in 2i. However, if a KO influences the expression level of the gene in 2i, it will also distort the WT completion of that KO. This problem might occur in a KO-specific manner, and it is not clear how the gene should adjust under differentiation-permitting conditions. It can be assumed that a particular change of expression has to happen under differentiation to reflect what happens in the WT, as we did in ours. The alternative would be to assume that a particular expression level has to be reached to reflect what happens in the WT. However, this would mean that KOs can reflect partial WT completion or negative WT completion in 2i. In both approaches, assumptions have to be made in some way to define what behavior is expected to reflect the adaptation that the cells excel in the WT. Here, we assumed that a specific change of expression has to happen for the gene to contribute or adapt to the naïve to formative transition. Additionally, the approach assumes monotonic behavior between 2i and 24 hours. While the thresholds on WT change (abs(log2FC 24 hours vs. 2i) >= 0.5 & padj <= 0.05) disregard genes that show initial regulation with a reversal to the initial expression values at 24 hours (such as genes from clusters 5, 6, 7, and 8 Figure 5.3), it does not guarantee that 24 hours is the most extreme change observed for all 1203 genes.

To test the resulting dependencies on independent data, we analyzed external sc data of WT cells differentiating from naïve to primed pluripotency. The sc approach to calculate gene regulatory dependencies in the naïve to formative transition was the limiting factor when overlapping results of the KO and sc-based approaches. This is due to the strict filtering criteria applied to the genes to be included in the analysis. Here, we required monotonic behavior in the sc data as the approach was based on ranking cells from early to late expression. While this criterion has to be kept on the gene level, it could be less restrictive when applying a similar approach to groups of genes instead. However, as we implemented the approach for single genes, we had to keep the restrictive filter to guarantee a clear transition from early to late expression behavior. Additionally, we considered cells with dropouts as cells with early or late expression levels depending on whether the gene was upregulated or downregulated respectively. While it is impossible to distinguish whether a count of zero has a technical or a biological reason for a single cell[301], we assume that for many cells at least some of that information can be obtained. Thus, we calculated the maximum rank difference that occurs at the time points in which high expression is expected. Here, we know that the gene should show increased expression and dropouts are more likely to be technical than biological compared to time points where low expression is expected. Calculating the maximum rank difference of cells with zeros from the time points with expected high expression thus delivers an estimation of rank difference caused by technical zeros. This difference in ranks

defines the buffer zone we implemented to avoid overinterpretation of rank differences more likely to be driven by dropouts.

The gene-wise dependencies are consistent between approaches and recapitulate known dependencies from pathways important in differentiation, for example WNT signaling. However, while the independence of one gene from another gene is strongly supported by the approaches we use, the second gene's dependence must be interpreted carefully. A dependency observed in our analysis always consists of one independent gene and a second gene, possibly depending on the independent gene. This possible dependence can be caused by a third gene that impacts the independent and the dependent gene or only the dependent gene. In this case, the dependent gene causally depends on the third gene, and the independent gene has no direct effect on the dependent gene. Even though we can add confidence to dependencies by requiring consistency between KO-based and sc-based results, this does not imply causal dependence. Thus, how dependency is defined in this thesis should not be confused with causal dependence as this would require more evidence. The other factor that must be considered is that the dependencies from the analysis in this thesis are restricted to transcriptional adaptation to the naïve to formative transition.

There are cases of known gene regulatory dependencies that do not agree with those calculated here, for example the independence of formative markers from *Esrrb*. That can have multiple reasons. First, the dependence is derived from the function of proteins or does depend on epigenetic modifications. In such a case, feedback mechanisms would allow a downstream gene to influence the transcription of a gene upstream in the pathway. Additionally, the ability to influence the transcription of the other gene can depend on epigenetic marks and thus differ between different states. This also leads to the second reason that the calculated dependencies might disagree with known dependencies. The dependencies in this thesis are in the context of the naïve to formative pluripotency transition. Thus, dependencies can differ at early or later stages of development as the transition from naïve to formative pluripotency includes heavy rewiring both transcriptionally and epigenetically. An example that can explain the differences between known dependencies and dependencies identified in this thesis is *Esrrb*. In the case of Esrrb, it was shown that it regulates both formative and naïve markers[240]. Which network is regulated by ESRRB, however is influenced by epigenetic modifications and TFs ESRRB can interact with between the different states. Furthermore, the study showed that these modifications change in the naïve to formative transition, thus changing the targets ESRRB regulates. At first, it seems to contradict the downstream position of *Esrrb* in our analysis. This suggests that the transcriptional adaptation of the other markers is mostly independent of the transcriptional adaptation of *Esrrb*. However, this case shows that the regulatory function of ESRRB is important for those other markers even though their transcriptional adaptation is independent of transcriptional changes of *Esrrb*.

This work established two methods to infer dependencies between genes from bulk RNA-seq data or scRNA-seq data. The resulting dependencies recapitulate known

dependencies and agree with structures of pathways critical for differentiation for example demonstrated by WNT signaling genes. Additionally, we saw that the two approaches are more likely to agree than disagree despite the different underlying data structures and laboratories in which the data was generated. Even though the degree of agreement depends on the restrictiveness of the scRNA-seq approach (Figure 5.27), we already observed agreement with the KO-based dependencies under unrestrictive implementation of the approach (only 150 cells outside of buffer zone). Additionally, gene-pairs including at least one naïve or formative marker showed even higher consistency between approaches. This likely results from the monotonic kinetics of expression change of those genes in the naïve to formative transition.

6.4 Independent adjustment of formative markers

As mentioned before, our dependency analysis indicated that the transcriptional adaptation of the formative network is independent of the transcriptional adaptation of the naïve network. This was consistent between the KO-based analysis and the sc-based analysis. Additionally, the consistency between both approaches was higher for pairs of genes that included at least one naïve or formative marker than the overall consistency (Figure 5.27). Thus, the dependencies between naïve and formative markers were not influenced much when removing dependencies that contradicted between approaches (Figure 5.18, Figure 5.31 A). The general trend between the two networks is still observed when only those dependencies that can be measured in both approaches are considered (Figure 5.31 B). However, the dependencies are much sparser due to the sc approach's lack of measurements for some gene pairs. This could be solved by integrating additional sc data to strengthen our analysis. While integrating additional sc data would remove some dependencies from the layer without contradictions, it would add even more confidence in this layer. At the same time, the most restrictive layer could be extended through additional data integration by including consistent dependencies between the KO approach and at least one sc approach.

In addition to the independent adjustment of the formative markers from the naïve markers, we also observed an independent adjustment of other genes specific to formative pluripotency from genes specific to naïve pluripotency (Figure 5.20 B). Even though some formative markers (*Dnmt3b*, *Pou3f1*, and *Fgf5*) and naïve markers (*Prdm14* and *Nanog*) were included in the corresponding gene sets, the observation holds without considering the marker genes. The three clusters of more independently adjusting genes from our analysis predominantly included genes corresponding to the later stages of differentiation investigated in this study[241]. The later stages of differentiation were represented by the preimplantation postimplantation EPI-specific genes and postimplantation EPI-specific genes. Similarly, the three clusters of more dependently adjusting genes from our analysis include predominantly ICM preimplantation EPI-specific genes and preimplantation EPI-specific genes. The ICM preimplantation EPI and preimplantation EPI

correspond to the state of early naïve pluripotency we investigate in this work. An exception to the trend mentioned above is *Sox2*, which is important in both naïve and formative pluripotency and not regulated in the naïve to formative transition[144,188,189]. Thus, the observation that the network needed in the later formative pluripotency state adapts independently from the earlier naïve pluripotency state can be extended to other genes describing those two states.

The last layer of confirmation for the independent transcriptional adjustment of the formative network from the naïve network came from additional data from experiments performed in the Leeb lab (Figure 5.33). The unpublished data comes from two experiments independent from the KO experiments we used to identify the gene dependencies. One experiment included quant-seq data of a double KO. This dKO cell line is unable to differentiate even after long-term passaging in basal medium. This contrasts the example of *Csnk1a1*, which first had a non-differentiation phenotype but started differentiation after further passaging[251]. In addition to the dKO data, we analyzed further data from the medium change experiment. The samples included come from the change from 2i to N2B27 and 2i to N2B27 supplemented with only Chiron or only PD. The independent adaptation of the formative genes, or at least some of the formative genes, was particularly striking 8 hours after the change to a PD-containing N2B27 medium. Here, the formative network has already adjusted to differentiation despite the ongoing inhibition of FGF/ERK signaling, which promotes the transition from naïve to formative pluripotency[132,273,274]. Additionally, the medium change to PD and Chiron confirms that the two inhibitors act through the impairment of different mechanisms instead of reinforced impairment of the same mechanism. While the changes of both single inhibitors correlate well with the changes in the inhibitor-free medium, both conditions with only one inhibitor correlate weakly (Figure 5.32). Both conditions share parts of the transcriptional adaptation with the condition without inhibitors but only share little of the transcriptional adaptation present in the other inhibitor. If the inhibitors just reinforced the effect of the other inhibitor and act on the same mechanisms, the correlation to each other should be higher than the correlation to the inhibitor-free medium.

The independent transcriptional adaptation of the formative network was also captured in work from Neagu and colleagues[132]. Here, a model representing the rosette state was established through simultaneous inhibition of both WNT and FGF/ERK (MEK) signaling. This system was defined by upregulation of the formative markers, while naïve markers showed expression behavior corresponding to the naïve state. From this rosette state, the cells either return to the naïve state or progress to the primed state, depending on the signals the cells are exposed to. While the *in vitro* rosette state, much like 2i, is an artificial cell state achieved through a combination of inhibitors, it provides a model with an already adjusted formative network without concomitant adjustment of the naïve network.

Aside from the analysis of the independence of the formative network from the naïve network, the different layers of the dependency network can be used as a helpful resource for other questions in the future. The dependencies can be a practical starting point to investigate open debates in early differentiation. The need for

sequential activation of WNT and FGF/ERK signaling described in the rosette state *in vitro* model[132] could further be investigated. A recent publication by Jayaram and colleagues opened the debate about which role the cell cycle plays in the exit from naïve pluripotency[272]. In contrast to many publications in the field[302–307], often performed in human ESCs, it is suggested that cell cycle phases cannot explain the asynchrony in exit from naïve pluripotency. Even though the sc data used in our analysis is from this study, the dependency network can still be a helpful resource to investigate cell cycle genes and other genes to identify candidates potentially influencing the asynchrony.

The results discussed in the previous paragraphs supported the initial observation that the formative network adapts independently from the naïve network. This opens the question of which mechanisms or factors control the formative network and initiate the shift from self-renewal to differentiation. The change from medium containing the two inhibitors to medium without the inhibitors is the cause of the most apparent change in mechanisms playing a role in this shift. Changes in WNT signaling and FGF/ERK signaling due to the lack of inhibitors in the medium promote differentiation. Upon differentiation and removal of Chiron, canonical WNT signaling is downregulated[208], and TCF7L1 can repress core pluripotency genes[210,278]. Simultaneously, the upregulation of FGF/ERK signaling in the absence of PD promotes differentiation even further[219,220]. The adaptation of parts of the formative network in the PD-only medium especially suggests that there are FGF/ERK independent mechanisms or factors influencing the breakage of the self-renewal program.

6.5 Potential regulators upstream of formative network

The analysis of potential TFs targeting the extended set of formative genes allowed us to propose a model of potential upstream regulators of the formative markers. The formative markers were extended parallel to the definition of naïve associated genes in previous work[251] using multiple regression analysis. While the approach helps to reveal potentially important regulators for the naïve to formative transition, the analysis has some limitations due to assumptions that had to be made. First, we used motif matches in open chromatin of available ATAC-seq data[257]. On the one hand, this allowed us to identify more potential TFs as upstream regulators as the approach is unbiased in comparison to using ChIP-seq data. On the other hand, the TF sites we identified indicate potential binding in open chromatin and do not represent measured binding in those areas. Second, the algorithm we used to assign target genes to the TFs assumes that only the closest gene to a binding motif is a target of the corresponding TF. While it is safest to assume that a TF acts close to the potential binding sites, the TF can still affect multiple genes with TSSs in proximity, which our approach would miss. However, as we used binding motifs in open chromatin of independent ATAC-seq data instead of ChIP-seq signal, we decided to be stricter in the TF to target assignment.

The resulting network includes TFs with known roles in differentiation and TFs previously not linked to the naïve to formative transition. While the regulatory role of some of the candidates has already been studied in the context of the naïve network, only few interactions with the formative network have been identified so far. The TF ZIC3 was found to be crucial for redistribution of OCT4 binding and its importance in the naïve to formative transition[280,281]. The role of TCF7L1 as a regulator of expression of *Esrrb*[210] and *Nanog*[278] is well established and leveraged by the inhibitor Chiron. However, the potential regulatory role of TCF7L*1* on the formative network suggested by our analysis would need to be investigated further. The recent identification of TFDP1 as a modulator of global chromatin accessibility in SCs[282] aligns with its potential importance in the naïve to formative transition due to the global reorganization of chromatin methylation during this transition[138,140,146,147]. A recent study on ZFP281 revealed that aside from the well-established role in pluripotency and its interaction with core pluripotency genes[287–291] ZFP281 also regulates the formative genes *Dnmt3a* and *Dnmt3b* directly[294]. Additionally, ESRRB was identified as a regulator of the formative network[240].

Even though there is a large body of literature on the proposed candidate TFs and their interactions with the naïve and core pluripotency network, little is known about their impact on the formative network. The only exceptions are ZFP281 and ESRRB, for which the regulatory role on the formative network has recently been found. Investigation of the potential effect of these candidate TFs on the formative network can help to further explain the mechanisms that transmit the changes in WNT and FGF/ERK signaling to the formative network, thus shifting the program from self-renewal to differentiation.

6.6 Limitations

In previous sections, we highlighted the limitations of different aspects of the approaches developed in this thesis and the assumptions made during different steps of the analysis. Aside from potential limitations of specific steps of the analysis, the limitations of this work come in the direct transferability to the differentiation of the mouse embryo *in vivo*.

First, this work only investigates the transcriptional adaptation of mESCs in the transition from naïve to formative pluripotency. Epigenetic changes and proteome changes are not considered in the analysis yet. Thus, the gene dependencies from this work are limited to the transcriptional adaptation of those genes to the differentiation signals. Integration of changes of proteins and epigenetic marks must be performed to fully understand the naïve to formative transition. The example of *Esrrb*[240] nicely highlights that the different layers of mRNA, protein, and epigenetic modifications all influence the function of the gene and have to be considered in combination.

Second, the experiments are done *in vitro*, and the comparability to the *in vivo* changes at this stage has to be assured to draw direct conclusions for the mouse

embryo. This comparison was already performed for the changes of DEGs and naïve associated genes in the WT and KOs[251]. Here, comparing *in vitro* changes in the mESCs and *in vivo* changes in mouse[142,308] and macaque[309] revealed high transcriptional similarity *in vivo* and *in vitro*, especially for the naïve associated genes. Thus, the dependencies identified in this work are based on data sharing high similarity to the *in vivo* pre- to postimplantation differentiation.

Even though the dependencies from our analysis provide first insights into transcriptional dependencies between genes, additional information on the proteome and epigenetic regulation is required to provide an extensive picture for the onset of differentiation. Additionally, the dependencies need further experimental support to be considered causal dependencies. However, the data and approaches used in this work strongly support the independent transcriptional adaptation of one gene from other genes in the naïve to formative transition.

6.7 Outlook

This work focused on the identification of processes and mechanisms that cause naïve mESCs to break the circuit of self-renewal and initiate differentiation, ultimately enabling establishment of the body plan of the embryo. While the results from this thesis elucidated gene regulatory dependencies in the naïve to formative transition, additional experiments are needed to further support our findings and investigate the transferability beyond the transcriptional adaptation.

First, the effect of the medium change on the proteome of differentiating mESCs can further be explored. We observed that genes tend to be regulated almost in immediately after the medium change on the transcriptional level. We showed that this was not caused by removing and replacing the conditioned medium with new medium. Whether the proteome is equally insensitive to the process of replacing conditioned medium with new medium remains to be tested. Additionally, it would be interesting to investigate the concept of translation on demand[310] on this data. Here, it could be tested if there are proteins poised for adaptation to differentiation independent of transcriptional changes and thus allow mESCs to react faster to differentiation signals.

Experimental validation of the proposed upstream regulators will be required to identify those TFs that regulate the formative markers. Here, the interference of those TFs and the corresponding downstream effects must be performed experimentally and analyzed. The analysis will investigate the effect of the interference on both the formative and naïve networks. While the interference of TFs would lead to impairments of upregulation of parts of the formative network to be considered an upstream regulator, it is unclear what effects to expect on the naïve network. Here, impairment of the downregulation of the naïve network or parts of it is the most likely outcome. The result of such experiments will help minimize the proposed model from this thesis and yield new insights into important factors in the naïve to formative transition of mESCs. The identified factors could extend the current knowledge of

early pluripotency states and provide a comprehensive understanding of the early differentiation similar to the understanding of the naïve state. Some of the proposed candidates, such as ZIC3, TCF7L1, and ZFP281, might provide a more comprehensive understanding of how changes of WNT and FGF/ERK signaling transfer to the formative network and kick-off differentiation.

The findings from this thesis help to draw a more comprehensive picture of the pre- to post-implantation differentiation of mESCs on the transcriptional level. Here, we provided an overview of transcriptional regulation and potential gene regulatory dependencies during this transition. However, these insights have to be extended to proteome and epigenetic changes for a more ample understanding. Specifically, the changes on the proteome are essential to translate the findings on the transcriptional adaptations to the functional level of proteins. Extending the time course with proteomic data will allow us to further investigate findings from this work and their effect on the corresponding proteins. First, a general comparison of changes in proteins and transcripts in the WT time course would identify genes mainly driven by transcription and translation and genes where other processes cause a disconnect of changes in the transcripts and the proteins. The dependencies of genes with similar changes of transcripts and proteins could mostly be extended to the level of proteins. However, the dependency of genes with disconnected transcript and protein changes would need further investigation.

6.8 Conclusions

In this work, we aimed to investigate the mechanisms that promote the differentiation of naïve ESCs and break the self-renewal program. This shift in programs initiates processes that lead to the establishment of the embryos' body plan and, ultimately, the differentiation into different cell types, enabling maintenance and regeneration of the adult organism.

We developed a method to quantify the delays of differentiation of 73 different KO cell lines to study the effects of the KOs on the networks supporting either self-renewal or differentiation. Here, a dense 32-hour WT differentiation time course was used as reference to quantify the delay of 73 different KO cell lines. Additionally, we developed two methods to infer gene regulatory dependencies in the naïve to formative transition to investigate which genes might drive this transition. The first approach to identify gene regulatory dependencies was based on the 73 KO cell lines and leveraged the variability introduced by the different KOs. The quality of the results improved when KOs with weak or no differentiation delay phenotypes were removed from the analysis. The second approach for identification of the gene regulatory dependencies was used on external scRNA-seq data and made use of the variability of the WT scs. The resulting regulatory dependencies of both approaches showed high concordance despite using different data structures and data from different laboratories.

The gene regulatory dependencies revealed the independent transcriptional adaptation of the formative network from the naïve network in the early differentiation of mESCs. Analysis of the transcriptional adaptation in medium containing PD showed that the formative network had already adapted to differentiation despite the inhibition of FGF/ERK. Thus, there must be FGF/ERK independent mechanisms that shift the program of the cells from self-renewal to differentiation and initiate upregulation of parts of the formative network. Using ATAC-seq data and motif information of TFs, we suggest a set of potential upstream regulators of the formative network that might play a role in the upregulation of the formative network and initiate the shift from self-renewal to differentiation. Taken together, we have shown that the transcriptional adaptation of the formative network is independent of the naïve network and proposed upstream regulators of the formative network as potential drivers of the naïve to formative transition.

7 Availability of code and data

The KO and 32-hour time course datasets are available on the Gene Expression Omnibus database under:

https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE145653

The medium change and dKO datasets will be made available to the scientific community upon publication of the corresponding results. The sc dataset provided by the group of Christa Brücker was published in:

Jayaram, S., Romeike, M. & Buecker, C. The asynchrony in the exit from naive pluripotency cannot be explained by differences in the cell cycle phase. bioRxiv 2023.09.15.557731 (2023) doi:10.1101/2023.09.15.557731

All R-scripts used in the analyses of this thesis are accessible under:

https://github.com/ftitztei/Gene-Regulatory-Dependencies-in-mESCs

8 Appendix

Supplementary Tables

KO gene	timing naïve marker	delay naïve marker	timing formative marker	delay formative marker	timing DEGs	delay DEGs
Csnk1a1	4.25	19.75	-0.75	24.75	-0.5	24.5
Zfp423	9.75	14.25	15.25	8.75	15.5	8.5
Rbpj	10	14	14.75	9.25	15.75	8.25
Fgfr1	11.75	12.25	20.5	3.5	16.5	7.5
Etv5	12.25	11.75	18.5	5.5	15.25	8.75
Smg5	12.25	11.75	8.25	15.75	16	8
Zfp281	12.25	11.75	14.25	9.75	16.5	7.5
Tcf7l1	12.5	11.5	18.5	5.5	17.5	6.5
Tsc2	12.5	11.5	13.25	10.75	16.25	7.75
Jarid2	14	10	10.25	13.75	15.25	8.75
Trp53	15	9	19.25	4.75	18.5	5.5
Smg6	15.25	8.75	11.75	12.25	17.25	6.75
Ptpn11	15.5	8.5	22.25	1.75	19	5
Nmt1	15.75	8.25	17.75	6.25	20.75	3.25
Trim71	16	8	18.75	5.25	21.5	2.5
Fryl	16.5	7.5	21	3	20.75	3.25
Raf1	16.75	7.25	21.75	2.25	19.75	4.25
Kdm6a	17	7	18.25	5.75	20.75	3.25
Suz12	17	7	32	-8	24	0
Mbd3	17.5	6.5	23.25	0.75	20.75	3.25
Pten	17.5	6.5	16	8	21	3
Cdk8	17.75	6.25	20.25	3.75	21.5	2.5
Eed	17.75	6.25	25.75	-1.75	23.5	0.5
Ncor1	18.75	5.25	21	3	23.5	0.5
Nmnat2	18.75	5.25	18.75	5.25	22.5	1.5
L3mbtl3	19	5	19.75	4.25	22.25	1.75
Usp7	19	5	23	1	22.5	1.5
Gabbr1	19.25	4.75	18.75	5.25	22	2
Apc	19.5	4.5	20.25	3.75	22.75	1.25
Mapk1	19.5	4.5	25	-1	20.25	3.75
Mta3	19.75	4.25	23	1	21.75	2.25
Smg7	19.75	4.25	23	1	21.25	2.75
Sulf1	19.75	4.25	21	3	22.75	1.25
Mthfd1l	20	4	22.5	1.5	23.25	0.75
Leo1	20.25	3.75	18.75	5.25	21	3
Pecam1	20.25	3.75	23.75	0.25	22.75	1.25

Zfp532	20.25	3.75	23	1	23	1
Jmjd1c	20.5	3.5	21.75	2.25	22	2
Lmna	20.5	3.5	18.5	5.5	21.5	2.5
Usp9x	20.5	3.5	20.75	3.25	23.75	0.25
Ctbp2	20.75	3.25	23.75	0.25	21	3
Ski	20.75	3.25	21.5	2.5	22.25	1.75
Myo10	21	3	22.25	1.75	23.5	0.5
Aff3	21.25	2.75	16.75	7.25	22.25	1.75
Alg13	21.25	2.75	21.5	2.5	21.75	2.25
Fbxw7	21.25	2.75	23	1	22.75	1.25
Arid5b	21.5	2.5	27.75	-3.75	23.5	0.5
Bcor	21.5	2.5	23.25	0.75	23.25	0.75
Nsd1	21.75	2.25	24	0	23.5	0.5
Dym	22.25	1.75	24.5	-0.5	23.75	0.25
Gtf2ird1	22.25	1.75	23	1	23.75	0.25
Atg13	22.5	1.5	21	3	23.75	0.25
Macf1	22.5	1.5	21	3	24.25	-0.25
Pitpnc1	22.5	1.5	23.25	0.75	23.75	0.25
Prkci	22.5	1.5	22	2	23.5	0.5
Rbpms	22.5	1.5	22.5	1.5	23	1
Cabin1	23	1	25	-1	23	1
Foxp1	23	1	20.25	3.75	22.75	1.25
Arih2	23.25	0.75	20.75	3.25	24.5	-0.5
Fgfr2	23.5	0.5	21.75	2.25	24.25	-0.25
Igf2bp1	23.5	0.5	22.5	1.5	24.25	-0.25
Irak3	23.75	0.25	23.75	0.25	24.25	-0.25
Etl4	24.25	-0.25	22.5	1.5	23.75	0.25
Tet1	24.25	-0.25	30.25	-6.25	24.25	-0.25
Hprt	24.5	-0.5	25.25	-1.25	24	0
Pum1	24.5	-0.5	27.75	-3.75	24.5	-0.5
Ssr2	24.5	-0.5	20.5	3.5	24.25	-0.25
Nes	25.5	-1.5	27.5	-3.5	25	-1
Dido1	25.75	-1.75	27.25	-3.25	25	-1
Rps6ka1	26	-2	17.25	6.75	24.5	-0.5
Hnrnph1	26.25	-2.25	24	0	25.5	-1.5
Msi2	26.5	-2.5	27.25	-3.25	24.25	-0.25
Myc	27.5	-3.5	29.25	-5.25	26.5	-2.5

Supplementary Table 1: timings and delays of all KO genes (rows) for naïve markers (first two columns), formative markers (columns 3 and 4), and DEGs (columns 5 and 6).

Cluster	Genes
1	Adhfe1, Aox1, Scn3a, Hist2h2be, Snap91, Ramp3, Cisd3, Jdp2, Lgmn, Syne3, Mok, Tmem117, Ephx1, Ak1, Trim44, Ckmt1, Mxd4, Micall2, Rtkn, Ptpn6, Snx20, Htr1b, Anks1b, Lif, Trim7, Hist3h2a, Foxj1, Aoah, 1500009C09Rik, Vdr, Tap1, Ptprn, Rbm44, Rbbp5, Fkbp7, Ndufaf5, Ggh, Fbxo6, Lrpap1, Cd9, Dnaaf3, 1700008O03Rik, Ap3b2, Fbxl12, Abca7, Atp8b3, Xkr6, Piwil2, Egfl6, Gm973, Rcsd1, Wfdc2, Pianp, Ldhb, Nrg4, Sec14l2, Camk2b, Mpp3, Dnaic2, Ocln, Fzd3, Ftx, Tdrd5, Vim, B3galt1, Nbea, Plch1, Cxxc4, Lin28a, Sh2d5, Pde3b, Itgam, Col6a5, Lin28b, Fam49a, Fam84a, Nrn1, Prr7, Has2, Usp25, Flywch1, H2-K1, Satb1, Tdrd1, Suv39h1, Sox3, Tspan6, Aox3, Epha4, Lyn, Tinagl1, Myo18b, Ffar3, Cemip, Nkx6-3, Mst1, Adgrg6, Ttll6, Bhmt, D830044D21Rik, Fam25c, Sox21, Qpct, Pqlc1, Nrp2, Prrx2, Hey1, Ppp2r2c, Ttll3, St8sia1, Ido2, Tmem231, Anxa2, Tekt1, Itgb3, Itpk1, Camk2a, Il13ra1, Soat1, Car14, Gatsl2, Calb2, Rassf9, Btbd17, Gm2762, Sh3tc2, Cnih2, Cd6, H19, Gm8994, Ptpn18, Klf7, Tfcp2l1, Gli2, Cd55, Rab29, Elf3, Nr5a2, Frmd4a, Proser2, Arrdc1, Gpsm1, Fam129b, Ptgs1, Baz2b, Galnt3, Accs, Pard6b, Zfp704, Chmp4c, Setd7, Trpc4, Mme, B3galnt1, Il6ra, Dram2, Cnn3, Dapp1, Plekhf2, Fkbp15, Rasef, Tesk2, Mfsd2a, Zc3h12a, Alpl, Garem2, Clock, Ptpn13, Ulk1, Fzd10, Pdgfa, Flnc, Fam131b, Adamts9, Spred3, Pcsk6, Mfge8, Pde8a, Tead1, Mylpf, Prss8, Hapln4, D230025D16Rik, Plcg2, Zbtb44, Fam81a, Tex264, Golga4, Cnksr3, Arid5b, Lss, Dcn, Ksr1, Plekhh1, Ism2, Id4, Serinc5, Thbs4, Hexb, Kat6b, Slc15a1, Prlr, Cthrc1, Naprt, Tmem184b, Mapk12, Kif21a, Pphln1, Tango2, Fstl1, Tmem8, Stk38, B4galt6, Epb41l4a, Jakmip2, Afap1l1, Cpt1a, Clcf1, Tmem151a, Dtx4, Pip5k1b, Erlin1, Vwa2, Pim2, Ndp, Abcd1, Acsl4, Lamc2, R3hdml, Dlgap3, Chrna9, Samd9l, Tead4, Ano1, Cd109, Xaf1, Mep1b, Klhl34, Tfap2c, Uox, Col9a2, Hspb8, Zc3hav1, Zmat4, Platr11, Kcns3, Cdhr1, A4galt, Bfsp1, D8Ertd82e, Irgm2, Rhbdl3, Tmem98, Igfbp5, Lad1, Tor3a, Lypd6b, Mir670hg, Srrm4, Rimbp2, Stard13, Magel2, Adm, Jade2, Id2, Bcl11b, Fgf9, Tnfsf11, Zcrb1, Prob1, Gna14, Pak3, Lonrf2, Klf2, Myh13, Hoxb13, Axin2, Tdh, Sall3, Hecw2, Fam171b, Themis2, Prtg, Tnfrsf19, Greb1l, Egr1, Lrrn2, Pdgfc, Aldh1b1, Ptpn3, Tnnt1, Shank1, Pknox2, Tmcc3, Rassf3, Hlf, Fkbp10, Uts2r, Ephx2, Slc1a3, Laptm4b, Lipg, Tfe3, Mir363, 2810459M11Rik, Rxrg, Rtn4rl2, Prom2, Cxcl12, Atp8b1, Gfra1, Exoc6b, Apela, Trim67, Pcolce2, Nefl, Mb21d2, Cited1, Sat1, Ass1, Casq2, Ddx58, Sh3gl2, Plekhg5, Trh, Ttyh1, Abhd2, Zfp710, Adgrg5, Slc29a3, Urgcp, Rnf135, Svil, Tmem88b, Gria1, Epn3, Slc22a3, Tubb4a
2	Slc7a11, Tmem144, Col11a1, Anxa3, Dio3, Emp2, Lama3, Ablim1, Spats2l, Pid1, Slc45a3, D1Pas1, Galnt12, Adamts7, Col6a4, Lgals9, Baiap2, Itm2a, Pax8, Dpp7, Uap1l1, Stmn3, Ggcx, Pdia5, Gpsm3, Slc22a7, Treml2, Dstyk, Slc30a10, Tcf15, S100a1, Rimkla, Pik3cd, Lgi2, Reep1, Sema4f, Rhcg, Nlrp14, Gas6, Ndrg4, Vstm5, 2810417H13Rik, Ccnjl, Tbc1d9b, Slc47a1, Ccdc92b, Tex19.1, Six1, Smc1b, Col2a1, Scn8a, Epb41l3, Ttc39c, Ppargc1b, Tex11, Vwa3b, Bend5, Crmp1, Fras1, Zcwpw1, Ccne1, Atp10a,

	Lhpp, Cdkn1c, Fgf15, Slc5a4b, Stc2, Serpinf1, Fzd2, Dnmt3a, C130071C03Rik, Btla, Plcl2, Syt4, Ctxn3, Ms4a14, Foxr2, Catip, Grid2, Nt5e, Phlda1, Tbc1d8, AA986860, C130074G19Rik, Mgarp, Tmem8b, Adamts8, Sept8, Lrrc75a, Dusp14, Lama1, Mppe1, Fam189a2, Ccdc73, Tet3, Tuba3a, Nlrp4c, Fndc1, Rhox5, Usp26, Gm364, 4930591A17Rik, 4933427D06Rik, Gm18336, Tram2, She, Hsdl2, Grik3, Tnip2, Fut1, Ubap1l, Hebp2, Tet1, Slc36a1, Slc35d2, Decr2, Mdfi, Zadh2, Jak2, Hells, Acsl5, Mob3b, Wnt5b, Fam53b, Msl1, Cables1, Ifit1, Zfp36l3, Hes3, Otop1, Ccno, C1qtnf2, AI427809, Eomes, Pcdh10, Edn2, Gm43293, Plxna4, Csrnp1, Atcay, Cacna1g, Cd96, Rbbp8nl, Cdx2, Nkx1-2, Fam107b, Patl2, B2m, Yes1, Trpm1, Grid1, Atp6v1c2, Gmnc
3	Pkhd1, Car8, Laptm5, Pramef12, Wscd2, Cadm4, Hs3st3b1, Serpinb6c, Cd34, Fam78a, Dnaic1, Hmgcll1, Unc45b, Lefty2, Lefty1, Fblim1, Acacb, Cabp1, 2900076A07Rik, Pkp3, Mical1, Pcsk4, H2-M3, Pitpnm1, Gm21992, Lrrfip1, Nat8l, Pcolce, Plekhg4, Ctrl, Tmod2, Sp6, Socs3, Nkd2, Fermt1, Sort1, Rnf207, Sox1, Elovl4, Abhd14b, Slc16a13, Mal2, Foxh1, Hspa1a, Tro, Cxcr4, Rusc2, Vax2, Adrb3, Slco2a1, Irak3, Mapt, Tmem63c, Btn2a2, Pnpla3, Anxa1, Mecom, Sh3tc1, Oas1a, Sycp2l, Slc1a1, Jazf1, Slc4a5, Cend1, Gm26945, Prdm14, Aff3, Kif1a, Slco4c1, Lypd6, Epb41l1, Sox2, Kirrel, Syt11, Sv2a, Vangl1, Mov10, Klf4, Lpar1, Zyg11a, Tbx3, A730049H05Rik, Nanog, Gprc5a, Lsr, Triml1, Gabarapl2, Izumo1r, Fam214a, Mras, 1110002J07Rik, Cobl, Trim25, Ngfr, Calcoco2, Cpsf4l, Nfkbia, Esrrb, Nid1, Rreb1, Dbn1, Kctd6, Nid2, Asb8, St6gal1, Jam2, Notch4, Rfx2, Clip4, Rnf125, Sema6a, Tbc1d25, Ebp, Porcn, Gabra3, Slc6a8, Stard8, Gsta3, Nkain4, Kifc3, Slc38a8, Slc6a15, Ndrg1, Ildr1, Pla1a, Slc6a7, Myof, Lax1, Mybpc3, Dnajc6, Zfp9, Thrsp, Adgrg1, Gadd45b, Osm, Tcl1, Serpinb9b, Gm807, Speg, Unc5a, Arl4c, Pogk, Zeb2, Kcnj3, Hivep2, Rnf112, Vasn, Tead3, Fndc3c1, Ankrd33b, Trim46, Tcea3, Nfic, Crip2, Dapk1, Glrx, Lpp, Epha1, Rfx4, Dpys, G0s2, Slc17a9, Kazn, Smarcd3, Rbm47, Mdfic, Mfsd12, Adap2, Jup, Ubxn2a, Dmtn, Klf8, Acot9
4	Dock10, Lemd1, Grid2ip, Nell1, Mfsd7c, Ahnak2, Gch1, Krt79, Cnga3, Neurod1, Sirpa, Sdc4, Lxn, Vgf, Alox5, Bbc3, Hs3st4, Pou4f2, Dgkb, Nkx2-9, Slc4a3, Ptgis, Rnf13, Lmna, Acox3, Scarb2, Pon2, Atp1a3, Zfp607b, Ryr1, Mctp2, Cpeb1, Themis, Ebf1, Pomc, Uhrf2, Sfxn4, Pcsk1n, Col20a1, Fut9, Cadps2, Tex101, Fam169b, Pls1, Plcd1, Rhbdf1, Slc43a2, Tunar, Actn2, Gpr137b, Mylip, Dok2, Parp10, Syngap1, Nlrc4, Crocc2, Kif21b, 5730559C18Rik, Cadm3, Syt13, Snap25, Insm1, Sox18, Car3, AI504432, Gbp2b, Map7d1, Iffo2, Fgd5, Pou2f2, Tmem145, Pcbp4, Usp44, Mrc2, Trib2, Acot2, Sp8, 2810429I04Rik, Vcan, Nt5dc2, Matn2, Gramd3, Pten, Ccdc160, Ptpn22, Slc13a5, Ly6a, Lix1, Stat1, Itgav, Ltk, Hdc, S100a13, S100a10, Selenbp1, Tmem64, Fabp3, Tlx2, Scube2, Adcy2, Crtac1, Pnck, Ier5, Astn1, Pdpn, Fgf5, Hip1, Peg3, Pde2a, Bst2, Calml4, Rnf217, Gstt2, Dlk1, Hus1b, Smarca1, Hs6st2, Iqca, Brsk2, 4930502E18Rik, Jph1, Sema4c, Mpzl1, Susd4, Rxra, Stom, Rab25, Aqp3, Tal1, Id3, Gabra4, Kit, Slc8b1, Adap1, Slc2a3, Gpm6a, Aadat, Slc5a5, Sorl1, Tmem136, Myo1e,

	Nhsl1, Dcbld1, Dnajc12, Adcy1, Gjc1, Ccdc40, Parp8, Cyp4f17, Cyb561a3, Rab33a, Ar, Itga9, Bmp4, Gjb3, Smyd1, Mycl, Dusp4, Ina, Bdnf, Myl9, Bnipl, Lef1, En2, Plcxd1, Miat, Elfn1, Podxl, Zfp467, Sez6l2, Mt3, Sox11, Rab15, Samd4, Pdzrn4, St8sia3, Ms4a4d, Irs4, Gm20362, Ntn1, Csrnp3, Fjx1, Stra8, Halr1, Cacng6, Plcxd2, Plekhh2, Cidea, Synpo2, Foxd3, Calcr, Tril, 1700027J19Rik, L3mbtl3, Clvs2, P3h4, Arl4d, Spata13, Parvg, Smagp, Rnf165, Aadac, Hoxb1, Cdk6, Jakmip1, Plbd1, Ifi35, Dnah8, Cask, Etnk2, Nrarp, Phf19, Ccdc141, Klf15, Il17rc, Nanos3, Robo4, Dixdc1, Ptrf, Il10, Slc26a5, Iqsec3, Enpp2, Prss22
5	Lamb3, Clca3b, Slc30a2, Tmem82, Pramel1, Chd5, Sfrp1, Hap1, Tmem179, Pcsk1, Sgk3, Pm20d1, Nfatc2, Abcb1b, Pqlc3, Foxd1, Pcdh8, Cpped1, Pros1, Cyyr1, Fhl1, Tmem81, Urm1, Ttll9, Atp8b2, Ciart, Rhoc, Usp48, Trmt44, Clstn3, Apoe, Hexa, Slc25a20, Als2cl, Cd63, Cnp, Lrrc45, Sptb, Ctsl, Dzip1, Gnpda1, Renbp, Slc41a1, Casq1, Psd4, Capn3, Emilin3, Tcfl5, Wdr86, Rilpl2, M1ap, Ceacam20, Man2a2, Lpl, Dzip1l, Myl7, Fbxo32, Parp9, Slco5a1, Cdk5r2, Pbx1, Vash2, Bmf, Sox12, Zcchc3, Tmem125, Zbtb8a, Vwa1, Gm38393, Snrpn, Efnb2, Ncan, Unc13a, Rasd2, Kcnk1, Cnn1, Soga3, Bcl6b, Eno3, Grb7, Otub2, Sox4, Tfap2a, Gadd45g, Map1b, Grhl2, Ccdc116, Dll1, Epcam, Gng3, Myrf, Ifit2, Nhs, Msc, Gm31108, Bpifb5, S100a11, Epb41l4b, Ncmap, Rab11fip5, Crxos, Tdrp, Cmtm7, Susd2, Elk3, Ckb, Naga, Scube3, Dsg2, Pdgfrb, Ahnak, Spata45, Crabp2, Esrp1, Ror1, Fgr, Bcl11a, Car4, Elmo1, Cdh10, Slc23a1, Apln, Peg12, Pam, Igsf9, Sptan1, Gpr107, Dab2ip, Ly75, Ext2, Wt1, Mavs, Dnmt3b, Rims4, Glmp, Ptgfrn, Artn, 2610528J11Rik, Pou3f1, Tmem54, Rcan3, Ttc28, Amz1, Creb3l2, Iffo1, Zfp59, Idh2, Spon1, Nr2f6, Sall1, Prss35, Shisa5, Fam26e, Prep, Tcn2, Smtnl2, Cep112, Ppm1a, Vrtn, Aldh5a1, Bhmt2, Dmgdh, Ap3b1, Adk, Sh2d4b, Sall2, Jph4, Spryd7, Mtmr12, Trio, Arfgap3, Tcte2, Tbc1d24, Memo1, Map4k3, Pcdh1, Plgrkt, Stag2, Lrrc34, Lrrc31, Eya1, Pla2g4e, Bmp7, Ddah1, Cttnbp2, Slc16a3, Nptxr, Trp53cor1, Mkx, Gucy2c, Fam57b, Sdr42e1, Shisa6, Vstm4, Fam84b, Psors1c2, Zfp608, Ephb6, Slc18b1, Hoxa1, Mobp, Rhobtb1, Nrcam, Msx2, Zfhx2, Letmd1, 2510009E07Rik, Crybg3, Pcdh19, Dnajb3, Thbs1, Gm12688, Cilp2, Unc5b, G2e3, Arhgef28, Zswim6, Rasgrp2, Mtap7d3, Kynu, Efhd1, Rasgrp1, Shf, Grsf1, Pglyrp1, Tacc2, Ror2, Ctnnd2, Cdkn1a, Catsperd, Ccbe1, Tmem51, Spsb1, Cdc42ep4, Sept9, Ehhadh, Sh3pxd2a
6	Fcgrt, Sept1, Coro2a, Fndc5, Asphd2, Tspan33, Smo, Nxph3, Rps6kl1, Sh3bp5, Prrc1, Blnk, Plac8, Hdgfrp3, Tmem158, Apol6, Pramef17, Gm13128, Gm26829, Alox12e, Efhb, Upk1a, Nfatc4, Nefm, Foxb1, Prdm1, Lrrc3b, Mreg, Kyat3, Hspb1, Zfp296, Triml2, Zfp42, Tppp3, Ephb1, Nodal, Gdf11, Pik3ip1, Tmem106a, Kcnk5, Phf11d, Klf5, Lifr, Trps1, Hbegf, Smad7, Cep55, 1700057H21Rik, Spp1, Slc6a1, Klk1, Cd37, Notum, Tgm1, Apobec2, Adgre5, Khdc3, 4930444M15Rik, Faxc, Rflnb, Phactr1, Spin2c, Tgm3, Grhl3, Rsph4a, Sox15, Prkca, Mcc, Ppl, Hck, Slitrk5, Nmnat2, Ceacam1, Cdyl2, Spire2, Pfkp, Fgf17, Rorc

Supplementary Table 2: Clusters of genes taken from NMF driven clustering approach.

Cut-off	Ratio (agreement/disagreement)	Number of dependencies in agreement
100	1.51	11092
250	1.67	9947
400	1.81	8325
450	1.87	7887
500	1.94	7469
550	2.02	7155
600	2.11	6893
650	2.20	6642
700	2.26	6281
750	2.35	6011
1000	2.65	4483
1250	2.76	2911
1500	2.41	1199
2000	1.32	308

Supplementary Table 3: resulting agreement/disagreement ratios and total number of remaining dependencies under different cut-offs.

9 List of Figures

10 Bibliography

1. Fuchs, E. & Segre, J. A. Stem Cells: A New Lease on Life. *Cell* **100**, 143–155 (2000).
2. Hall, P. A. & Watt, F. M. Stem cells: the generation and maintenance of cellular diversity. *Development* **106**, 619–633 (1989).
3. Deneke, V. E. & Pauli, A. The Fertilization Enigma: How Sperm and Egg Fuse. *Annu Rev Cell Dev Biol* **37**, 391–414 (2021).
4. Bianchi, E. & Wright, G. J. Sperm meets egg: the genetics of mammalian fertilization. *Annu. Rev. Genet.* **50**, 93–111 (2016).
5. Clift, D. & Schuh, M. Restarting life: fertilization and the transition from meiosis to mitosis. *Nat. Rev. Mol. Cell Biol.* **14**, 549–562 (2013).
6. Johnson, M. H. & McConnell, J. M. L. Lineage allocation and cell polarity during mouse embryogenesis. *Semin Cell Dev Biol* **15**, 583–597 (2004).
7. Tam, P. P. L. & Loebel, D. A. F. Gene function in mouse embryogenesis: Get set for gastrulation. *Nat Rev Genet* **8**, 368–381 (2007).
8. Greaves, M. F. Stem cell origins of leukaemia and curability. *Br J Cancer* **67**, 413–423 (1993).
9. Jackson, G. H. *et al.* Philadelphia positive acute leukaemia with minor breakpoint cluster rearrangement may be a stem cell disease. *Br J Haematol* **81**, 77–80 (1992).
10. Nishigaki, H. *et al.* Prevalence and Growth Characteristics of Malignant Stem Cells in B-Lineage Acute Lymphoblastic Leukemia. *Blood* **89**, 3735–3744 (1997).
11. Fialkow, P. J., Denman, A. M., Jacobson, R. J. & Lowenthal, M. N. Chronic myelocytic leukemia. Origin of some lymphocytes from leukemic stem cells. *J Clin Invest* **62**, 815–823 (1978).
12. Singh, S. K. *et al.* Identification of human brain tumour initiating cells. *Nature* **432**, 396–401 (2004).
13. Hemmati, H. D. *et al.* Cancerous stem cells can arise from pediatric brain tumors. *Proc Natl Acad Sci U S A* **100**, 15178–15183 (2003).
14. Galli, R. *et al.* Isolation and characterization of tumorigenic, stem-like neural precursors from human glioblastoma. *Cancer Res* **64**, 7011–7021 (2004).
15. Wang, Y. *et al.* Expression of Mutant p53 Proteins Implicates a Lineage Relationship between Neural Stem Cells and Malignant Astrocytic Glioma in a Murine Model. *Cancer Cell* **15**, 514–526 (2009).
16. Mimeault, M. & Batra, S. K. Concise review: recent advances on the significance of stem cells in tissue regeneration and cancer therapies. *Stem Cells* **24**, 2319–2345 (2006).
17. Mimeault, M., Hauke, R. & Batra, S. K. Stem cells: a revolution in therapeutics-recent advances in stem cell biology and their therapeutic applications in regenerative medicine and cancer therapies. *Clin Pharmacol Ther* **82**, 252–264 (2007).

18. Mahla, R. S. Stem Cells Applications in Regenerative Medicine and Disease Therapeutics. *Int J Cell Biol* **2016**, (2016).

19. O'Brien, T. & Barry, F. P. Stem Cell Therapy and Regenerative Medicine. *Mayo Clin Proc* **84**, 859 (2009).

20. Chien, K. R. Regenerative medicine and human models of human disease. *Nature 2008 453:7193* **453**, 302–305 (2008).

21. Aversa, F. *et al.* Treatment of High-Risk Acute Leukemia with T-Cell–Depleted Stem Cells from Related Donors with One Fully Mismatched HLA Haplotype. *N Engl J Med* **339**, 1186–1193 (1998).

22. Goldman, J. M. & Melo, J. V. Chronic Myeloid Leukemia — Advances in Biology and New Approaches to Treatment. *N Engl J Med* **349**, 1451–1464 (2003).

23. Copelan, E. A. Hematopoietic Stem-Cell Transplantation. *N Engl J Med* **354**, 1813–1826 (2006).

24. Pellegrini, G. *et al.* Biological parameters determining the clinical outcome of autologous cultures of limbal stem cells. *Regenerative Med* **8**, 553–567 (2013).

25. Rama, P. *et al.* Autologous fibrin-cultured limbal stem cells permanently restore the corneal surface of patients with total limbal stem cell deficiency. *Transplantation* **72**, 1478–1485 (2001).

26. Tsai, R. J.-F., Li, L.-M. & Chen, J.-K. Reconstruction of damaged corneas by transplantation of autologous limbal epithelial cells. *N Engl J Med* **343**, 86–93 (2000).

27. Pellegrini, G. *et al.* Location and clonal analysis of stem cells and their differentiated progeny in the human ocular surface. *J Cell Biol* **145**, 769–782 (1999).

28. Rodriguez, A. M. *et al.* Transplantation of a multipotent cell population from human adipose tissue induces dystrophin expression in the immunocompetent mdx mouse. *J Exp Med* **201**, 1397 (2005).

29. Zhang, Y. V., Cheong, J., Ciapurin, N., McDermitt, D. J. & Tumbar, T. Distinct self-renewal and differentiation phases in the niche of infrequently dividing hair follicle stem cells. *Cell Stem Cell* **5**, 267–278 (2009).

30. Loeffler, M. & Roeder, I. Tissue stem cells: definition, plasticity, heterogeneity, self-organization and models--a conceptual approach. *Cells Tissues Organs* **171**, 8–26 (2002).

31. Marshman, E., Booth, C. & Potten, C. S. The intestinal epithelial stem cell. *Bioessays* **24**, 91–98 (2002).

32. Clayton, E. *et al.* A single type of progenitor cell maintains normal epidermis. *Nature* **446**, 185–189 (2007).

33. Ho, A. D. Kinetics and symmetry of divisions of hematopoietic stem cells. *Exp Hematol* **33**, 1–8 (2005).

34. Zhong, W. & Chia, W. Neurogenesis and asymmetric cell division. *Curr Opin Neurobiol* **18**, 4–11 (2008).

35. Fuchs, E., Tumbar, T. & Guasch, G. Socializing with the neighbors: Stem cells and their niche. *Cell* **116**, 769–778 (2004).

36. Knoblich, J. A. Mechanisms of asymmetric stem cell division. *Cell* **132**, 583–597 (2008).

37. Friedenstein, A. J. *et al.* Precursors for fibroblasts in different populations of hematopoietic cells as detected by the in vitro colony assay method. *Exp Hematol* **2**, 83–92 (1974).

38. O'Connor, M. D. *et al.* Alkaline phosphatase-positive colony formation is a sensitive, specific, and quantitative indicator of undifferentiated human embryonic stem cells. *Stem Cells* **26**, 1109–1116 (2008).

39. Sarugaser, R., Hanoun, L., Keating, A., Stanford, W. L. & Davies, J. E. Human mesenchymal stem cells self-renew and differentiate according to a deterministic hierarchy. *PLoS One* **4**, (2009).

40. Passegué, E., Jamieson, C. H. M., Ailles, L. E. & Weissman, I. L. Normal and leukemic hematopoiesis: are leukemias a stem cell disorder or a reacquisition of stem cell characteristics? *Proc Natl Acad Sci U S A* **100 Suppl 1**, 11842–11849 (2003).

41. Jordan, C. T. & Guzman, M. L. Mechanisms controlling pathogenesis and survival of leukemic stem cells. *Oncogene* **23**, 7178–7187 (2004).

42. Bonnet, D. & Dick, J. E. Human acute myeloid leukemia is organized as a hierarchy that originates from a primitive hematopoietic cell. *Nat Med* **3**, 730–737 (1997).

43. Lapidot, T. *et al.* A cell initiating human acute myeloid leukaemia after transplantation into SCID mice. *Nature* **367**, 645–648 (1994).

44. Evans, M. J. & Kaufman, M. H. Establishment in culture of pluripotential cells from mouse embryos. *Nature* vol. 292 154–156 Preprint at https://doi.org/10.1038/292154a0 (1981).

45. Martin, G. R. Isolation of a pluripotent cell line from early mouse embryos cultured in medium conditioned by teratocarcinoma stem cells. *Proc Natl Acad Sci U S A* **78**, 7634–7638 (1981).

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

47. Huch, M. & Koo, B. K. Modeling mouse and human development using organoid cultures. *Development* **142**, 3113–3125 (2015).

48. Lancaster, M. A. & Knoblich, J. A. Organogenesis in a dish: Modeling development and disease using organoid technologies. *Science (1979)* **345**, (2014).

49. Lancaster, M. A. *et al.* Cerebral organoids model human brain development and microcephaly. *Nature 2013 501:7467* **501**, 373–379 (2013).

50. Zhao, Z. *et al.* Organoids. *Nature Reviews Methods Primers 2022 2:1* **2**, 1–21 (2022).

51. Luo, C. *et al.* Cerebral Organoids Recapitulate Epigenomic Signatures of the Human Fetal Brain. *Cell Rep* **17**, 3369 (2016).

52. Lancaster, M. A. *et al.* Guided self-organization and cortical plate formation in human brain organoids. *Nature Biotechnology 2017 35:7* **35**, 659–666 (2017).

53. Zhou, T. *et al.* High-Content Screening in hPSC-Neural Progenitors Identifies Drug Candidates that Inhibit Zika Virus Infection in Fetal-like Organoids and Adult Brain. *Cell Stem Cell* **21**, 274-283.e5 (2017).

54. Takahashi, T. Organoids for Drug Discovery and Personalized Medicine. *Annual Review of Pharmacology and Toxicology* **59**, 447–462 (2019).

55. Miranda, C. C., Fernandes, T. G., Diogo, M. M. & Cabral, J. M. S. Towards Multi-Organoid Systems for Drug Screening Applications. *Bioengineering* **5**, (2018).

56. Vandana, J. J., Manrique, C., Lacko, L. A. & Chen, S. Human pluripotent-stem-cell-derived organoids for drug discovery and evaluation. *Cell Stem Cell* **30**, 571–591 (2023).

57. Lau, K. Y. C., Amadei, G. & Zernicka-Goetz, M. Assembly of complete mouse embryo models from embryonic and induced stem cell types in vitro. *Nature Protocols 2023 18:12* **18**, 3662–3689 (2023).

58. Sozen, B. *et al.* Self-assembly of embryonic and two extra-embryonic stem cell types into gastrulating embryo-like structures. *Nat Cell Biol* **20**, 979–989 (2018).

59. Lau, K. Y. C. *et al.* Mouse embryo model derived exclusively from embryonic stem cells undergoes neurulation and heart development. *Cell Stem Cell* **29**, 1445-1458.e8 (2022).

60. Ocampo, A. *et al.* In Vivo Amelioration of Age-Associated Hallmarks by Partial Reprogramming. *Cell* **167**, 1719-1733.e12 (2016).

61. Yang, J. H. *et al.* Chemically induced reprogramming to reverse cellular aging. *Aging (Albany NY)* **15**, 5966 (2023).

62. Gill, D. *et al.* Multi-omic rejuvenation of human cells by maturation phase transient reprogramming. *Elife* **11**, (2022).

63. Mazid, M. A. *et al.* Rolling back human pluripotent stem cells to an eight-cell embryo-like stage. *Nature* **605**, 315–324 (2022).

64. Yang, M. *et al.* Chemical-induced chromatin remodeling reprograms mouse ESCs to totipotent-like stem cells. *Cell Stem Cell* **29**, 400–418 (2022).

65. Hu, Y. *et al.* Induction of mouse totipotent stem cells by a defined chemical cocktail. *Nature* **617**, 792–797 (2022).

66. Solter, D. From teratocarcinomas to embryonic stem cells and beyond: a history of embryonic stem cell research. *Nat. Rev. Genet.* **7**, 319–327 (2006).

67. Posfai, E. *et al.* Evaluating totipotency using criteria of increasing stringency. *Nature Cell Biology 2021 23:1* **23**, 49–60 (2021).

68. López-Otín, C., Blasco, M. A., Partridge, L., Serrano, M. & Kroemer, G. The Hallmarks of Aging. *Cell* **153**, 1194–1217 (2013).

69. Wassarman, P. M. & Litscher, E. S. Mammalian fertilization: the egg's multifunctional zona pellucida. *Int. J. Dev. Biol.* **52**, 665–676 (2008).

70. Wassarman, P. M. & Mortillo, S. Structure of the Mouse Egg Extracellular Coat, the Zona Pellucida. *Int Rev Cytol* **130**, 85–110 (1991).

71. Hartmann, J. F., Gwatkin, R. B. & Hutchison, C. F. Early Contact Interactions between Mammalian Gametes In Vitro: Evidence That the Vitellus Influences Adherence between Sperm and Zona Pellucida. *Proceedings of the National Academy of Sciences* **69**, 2767–2769 (1972).

72. Hertwig, O. W. A. Beiträge zur Kenntniss der Bildung, Befruchtung und Theilung des thierischen Eies. in *Morphologisches Jahrbuch 1* 347–434 (1876).

73. Trebichalská, Z. & Holubcová, Z. Perfect date-the review of current research into molecular bases of mammalian fertilization. *J. Assist. Reprod. Genet.* **37**, 243–256 (2020).

74. Kojima, Y., Tam, O. H. & Tam, P. P. L. Timing of developmental events in the early mouse embryo. *Semin Cell Dev Biol* **34**, 65–75 (2014).

75. Ziomek, C. A. & Johnson, M. H. Cell surface interaction induces polarization of mouse 8-cell blastomeres at compaction. *Cell* **21**, 935–942 (1980).

76. Fleming, T. P., Javed, Q. & Hay, M. Epithelial differentiation and intercellular junction formation in the mouse early embryo. *Journal of Neuroscience* **13**, 105–112 (1993).

77. Ducibella, T. & Anderson, E. Cell shape and membrane changes in the eight-cell mouse embryo: Prerequisites for morphogenesis of the blastocyst. *Dev Biol* **47**, 45–58 (1975).

78. Fierro-González, J. C., White, M. D., Silva, J. C. & Plachta, N. Cadherin-dependent filopodia control preimplantation embryo compaction. *Nat Cell Biol* **15**, 1424–1433 (2013).

79. Aiken, C. E. M., Swoboda, P. P. L., Skepper, J. N. & Johnson, M. H. The direct measurement of embryogenic volume and nucleo-cytoplasmic ratio during mouse pre-implantation development. *Reproduction* **128**, 527–535 (2004).

80. Motosugi, N., Bauer, T., Polanski, Z., Solter, D. & Hiiragi, T. Polarity of the mouse embryo is established at blastocyst and is not prepatterned. *Genes Dev* **19**, 1081–1092 (2005).

81. Chazaud, C., Yamanaka, Y., Pawson, T. & Rossant, J. Early Lineage Segregation between Epiblast and Primitive Endoderm in Mouse Blastocysts through the Grb2-MAPK Pathway. *Dev Cell* **10**, 615–624 (2006).

82. Krupa, M. *et al.* Allocation of inner cells to epiblast vs primitive endoderm in the mouse embryo is biased but not determined by the round of asymmetric divisions (8→16- and 16→32-cells). *Dev Biol* **385**, 136–148 (2014).

83. Morris, S. A. *et al.* Origin and formation of the first two distinct cell types of the inner cell mass in the mouse embryo. *Proc Natl Acad Sci U S A* **107**, 6364–6369 (2010).

84. Handyside, A. H. & Hunter, S. Cell division and death in the mouse blastocyst before implantation. *Roux's Archives of Developmental Biology* **195**, 519–526 (1986).

85. Dey, S. K. *et al.* Molecular Cues to Implantation. *Endocr Rev* **25**, 341–373 (2004).

86. Huisman, G. J., Fauser, B. C. J. M., Eijkemans, M. J. C. & Pieters, M. H. E. C. Implantation rates after in vitro fertilization and transfer of a maximum of two

embryos that have undergone three to five days of culture. *Fertil Steril* **73**, 117–122 (2000).

87. Wang, H. & Dey, S. K. Roadmap to embryo implantation: clues from mouse models. *Nat. Rev. Genet.* **7**, 185–199 (2006).

88. Cohen, J. *et al.* Impairment of the hatching process following IVF in the human and improvement of implantation by assisting hatching using micromanipulation. *Hum. Reprod.* **5**, 7–13 (1990).

89. Meilhac, S. M. *et al.* Active cell movements coupled to positional induction are involved in lineage segregation in the mouse blastocyst. *Dev Biol* **331**, 210–221 (2009).

90. Plusa, B., Piliszek, A., Frankenberg, S., Artus, J. & Hadjantonakis, A. K. Distinct sequential cell behaviours direct primitive endoderm formation in the mouse blastocyst. *Development* **135**, 3081–3091 (2008).

91. Bedzhov, I. & Zernicka-Goetz, M. Self-Organizing Properties of Mouse Pluripotent Cells Initiate Morphogenesis upon Implantation. *Cell* **156**, 1032–1044 (2014).

92. Brennan, J. *et al.* Nodal signalling in the epiblast patterns the early mouse embryo. *Nature 2001 411:6840* **411**, 965–969 (2001).

93. Snow, M. H. L. Gastrulation in the mouse: growth and regionalization of the epiblast. *J Embryol Exp Morphol* **Vol. 42**, 293–303 (1977).

94. Smith, R. & McLaren, A. Factors affecting the time of formation of the mouse blastocoele. *Development* **41**, 79–92 (1977).

95. Rands, G. F. Size regulation in the mouse embryo. II. The development of half embryos. *J Embryol Exp Morphol* **Vol. 98**, 209–217 (1986).

96. Power, M. A. & Tam, P. P. L. Onset of gastrulation, morphogenesis and somitogenesis in mouse embryos displaying compensatory growth. *Anat Embryol (Berl)* **187**, 493–504 (1993).

97. Schmidt, W. The amniotic fluid compartment: the fetal habitat. *Adv Anat Embryol Cell Biol* **127**, 1–100 (1992).

98. Nonaka, S. *et al.* Randomization of Left–Right Asymmetry due to Loss of Nodal Cilia Generating Leftward Flow of Extraembryonic Fluid in Mice Lacking KIF3B Motor Protein. *Cell* **95**, 829–837 (1998).

99. Okada, Y. *et al.* Abnormal Nodal Flow Precedes Situs Inversus in iv and inv mice. *Mol Cell* **4**, 459–468 (1999).

100. Hogan, B., Costantini, F. & Lacy, E. *Manipulating the Mouse Embryo: A Laboratory Manual*. (Cold spring harbor laboratory Cold Spring Harbor, NY, 1986).

101. Tam, P. P. L. & Tan, S. S. The somitogenetic potential of cells in the primitive streak and the tail bud of the organogenesis-stage mouse embryo. *Development* **115**, 703–715 (1992).

102. Tam, P. P. L. & Beddington, R. S. P. The Metameric Organization of the Presomitic Mesoderm and Somite Specification in the Mouse Embryo. *Somites in Developing Embryos* 17–36 (1986) doi:10.1007/978-1-4899-2013-3_2.

103. Bindu A, H. & B, S. Potency of Various Types of Stem Cells and their Transplantation. *J Stem Cell Res Ther* **01**, (2011).

104. Clarke, D. & Frisén, J. Differentiation potential of adult stem cells. *Curr Opin Genet Dev* **11**, 575–580 (2001).

105. Morrison, S. J., Shah, N. M. & Anderson, D. J. Regulatory Mechanisms in Stem Cell Biology. *Cell* **88**, 287–298 (1997).

106. Lawson, K. A., Meneses, J. J. & Pedersen, R. A. Clonal analysis of epiblast fate during germ layer formation in the mouse embryo. *Development* **113**, 891–911 (1991).

107. Dekoninck, S. & Blanpain, C. Stem cell dynamics, migration and plasticity during wound healing. *Nat Cell Biol* **21**, 18 (2019).

108. Kolios, G. & Moodley, Y. Introduction to Stem Cells and Regenerative Medicine. *Respiration* **85**, 3–10 (2012).

109. Posfai, E. *et al.* Position- and hippo signaling-dependent plasticity during lineage segregation in the early mouse embryo. *Elife* **6**, (2017).

110. Tarkowski, A. K., Suwińska, A., Czolowska, R. & Ozdzezski, W. Individual blastomeres of 16- and 32-cell mouse embryos are able to develop into foetuses and mice. *Dev. Biol.* **348**, 190–198 (2010).

111. Rossant, J. Investigation of the determinative state of the mouse inner cell massII. The fate of isolated inner cell masses transferred to the oviduct. *Development* **33**, 991–1001 (1975).

112. Condic, M. L. Totipotency: What It Is and What It Is Not. *Stem Cells Dev* **23**, 796–812 (2013).

113. Denker, H. W. Early human development: new data raise important embryological and ethical questions relevant for stem cell research. *Naturwissenschaften* **91**, 1–21 (2004).

114. Tarkowski, A. K. Experiments on the Development of Isolated Blastomeres of Mouse Eggs. *Nature* **184**, 1286–1287 (1959).

115. Genet, M. & Torres-Padilla, M. E. The molecular and cellular features of 2-cell-like cells: a reference guide. *Development* **147**, (2020).

116. Macfarlan, T. S. *et al.* Embryonic stem cell potency fluctuates with endogenous retrovirus activity. *Nature* **487**, 57–63 (2012).

117. Rossant, J. Postimplantation development of blastomeres isolated from 4-and 8-cell mouse eggs. *J Embryol Exp Morphol* **36**, 283–290 (1976).

118. Nicholas, J. S. & Hall, B. V. Experiments on developing rats. II. The development of isolated blastomeres and fused eggs. *Journal of Experimental Zoology* **90**, 441–459 (1942).

119. Johnson, W. H., Loskutoff, N. M., Plante, Y. & Betteridge, K. J. Production of four identical calves by the separation of blastomeres from an in vitro derived four-cell embryo. *Vet Rec* **137**, 15–16 (1995).

120. Willadsen, S. M. & Polge, C. Attempts to produce monozygotic quadruplets in cattle by blastomere separation. *Vet Rec* **108**, 211–213 (1981).

121. Nichols, J. & Smith, A. Naive and Primed Pluripotent States. *Stem Cell* **4**, 487–492 (2009).

122. Boroviak, T. & Nichols, J. The birth of embryonic pluripotency. *Philosophical Transactions of the Royal Society B: Biological Sciences* **369**, (2014).

123. Rossant, J. & Tam, P. P. L. New Insights into Early Human Development: Lessons for Stem Cell Derivation and Differentiation. *Cell Stem Cell* **20**, 18–28 (2017).

124. Smith, A. Formative pluripotency: The executive phase in a developmental continuum. *Development (Cambridge)* **144**, 365–373 (2017).

125. De Los Angeles, A. *et al.* Hallmarks of pluripotency. *Nature* **525**, 469–478 (2015).

126. Shahbazi, M. N. *et al.* Pluripotent state transitions coordinate morphogenesis in mouse and human embryos. *Nature* **552**, 239–243 (2017).

127. Wu, J. & Izpisua Belmonte, J. C. Dynamic Pluripotent Stem Cell States and Their Applications. *Cell Stem Cell* **17**, 509–525 (2015).

128. Hackett, J. A. & Azim Surani, M. Regulatory principles of pluripotency: From the ground state up. *Cell Stem Cell* **15**, 416–430 (2014).

129. Weinberger, L., Ayyash, M., Novershtern, N. & Hanna, J. H. Dynamic stem cell states: Naive to primed pluripotency in rodents and humans. *Nat Rev Mol Cell Biol* **17**, 155–169 (2016).

130. Ying, Q. *et al.* The ground state of embryonic stem cell self-renewal. *Nature* **453**, 519–523 (2008).

131. Marks, H. *et al.* The transcriptional and epigenomic foundations of ground state pluripotency. *Cell* **149**, 590–604 (2012).

132. Neagu, A. *et al.* In vitro capture and characterization of embryonic rosette-stage pluripotency between naive and primed states. *Nat Cell Biol* **22**, 534–545 (2020).

133. Furlan, G., Huyghe, A., Combémorel, N. & Lavial, F. Molecular versatility during pluripotency progression. *Nature Communications 2023 14:1* **14**, 1–13 (2023).

134. Boroviak, T., Loos, R., Bertone, P., Smith, A. & Nichols, J. The ability of inner-cell-mass cells to self-renew as embryonic stem cells is acquired following epiblast specification. *Nat Cell Biol* **16**, 513–525 (2014).

135. Brook, F. A. & Gardner, R. L. The origin and efficient derivation of embryonic stem cells in the mouse. *Proc Natl Acad Sci U S A* **94**, 5709–5712 (1997).

136. Nichols, J. *et al.* Validated germline-competent embryonic stem cell lines from nonobese diabetic mice. *Nature Medicine 2009 15:7* **15**, 814–818 (2009).

137. Kiyonari, H., Kaneko, M., Abe, S. I. & Aizawa, S. Three inhibitors of FGF receptor, ERK, and GSK3 establishes germline-competent embryonic stem cells of C57BL/6N mouse strain with high efficiency and stability. *Genesis* **48**, 317–327 (2010).

138. Leitch, H. G. *et al.* Naive pluripotency is associated with global DNA hypomethylation. *Nat Struct Mol Biol* **20**, 311–316 (2013).

139. Ficz, G. *et al.* FGF signaling inhibition in ESCs drives rapid genome-wide demethylation to the epigenetic ground state of pluripotency. *Cell Stem Cell* **13**, 351–359 (2013).

140. Habibi, E. *et al.* Whole-genome bisulfite sequencing of two distinct interconvertible DNA methylomes of mouse embryonic stem cells. *Cell Stem Cell* **13**, 360–369 (2013).

141. Yagi, M. *et al.* Derivation of ground-state female ES cells maintaining gamete-derived DNA methylation. *Nature* **548**, 224–227 (2017).

142. Mohammed, H. *et al.* Single-Cell Landscape of Transcriptional Heterogeneity and Cell Fate Decisions during Mouse Early Gastrulation. *Cell Rep* **20**, 1215–1228 (2017).

143. Kalkan, T. *et al.* Tracking the embryonic stem cell transition from ground state pluripotency. *Development* **144**, 1221 (2017).

144. Hayashi, K., Ohta, H., Kurimoto, K., Aramaki, S. & Saitou, M. Reconstitution of the mouse germ cell specification pathway in culture by pluripotent stem cells. *Cell* **146**, 519–532 (2011).

145. Ohinata, Y. *et al.* A Signaling Principle for the Specification of the Germ Cell Lineage in Mice. *Cell* **137**, 571–584 (2009).

146. Hackett, J. A. *et al.* Synergistic mechanisms of DNA demethylation during transition to ground-state pluripotency. *Stem Cell Reports* **1**, 518–531 (2013).

147. Hayashi, K., Lopes, S. M. C. de S., Tang, F. & Surani, M. A. Dynamic Equilibrium and Heterogeneity of Mouse Pluripotent Stem Cells with Distinct Functional and Epigenetic States. *Cell Stem Cell* **3**, 391–401 (2008).

148. Williams, L. H., Kalantry, S., Starmer, J. & Magnuson, T. Transcription precedes loss of Xist coating and depletion of H3K27me3 during X-chromosome reprogramming in the mouse inner cell mass. *Development* **138**, 2049–2057 (2011).

149. Rossant, J. Stem Cells and Early Lineage Development. *Cell* **132**, 527–531 (2008).

150. Jaing, T. H. Umbilical cord blood: A trustworthy source of multipotent stem cells for regenerative medicine. *Cell Transplant* **23**, 493–496 (2014).

151. Bacakova, L. *et al.* Stem cells: their source, potency and use in regenerative therapies with focus on adipose-derived stem cells – a review. *Biotechnol Adv* **36**, 1111–1126 (2018).

152. Avilion, A. A. *et al.* Multipotent cell lineages in early mouse development depend on SOX2 function. *Genes Dev* **17**, 126 (2003).

153. Fehling, H. J. *et al.* Tracking mesoderm induction and its specification to the hemangioblast during embryonic stem cell differentiation. *Development* **130**, 4217–4227 (2003).

154. Buckingham, M., Meilhac, S. & Zaffran, S. Building the mammalian heart from two sources of myocardial cells. *Nat Rev Genet* **6**, 826–835 (2005).

155. Ferretti, E. & Hadjantonakis, A. K. Mesoderm specification and diversification: from single cells to emergent tissues. *Curr Opin Cell Biol* **61**, 110 (2019).

156. Zorn, A. M. & Wells, J. M. Vertebrate Endoderm Development and Organ Formation. *Annual Review of Cell and Developmental Biology* **25**, 221–251 (2009).

157. Zaret, K. S. Genetic programming of liver and pancreas progenitors: lessons for stem-cell differentiation. *Nature Reviews Genetics 2008 9:5* **9**, 329–340 (2008).

158. Cardoso, W. V. & Lü, J. Regulation of early lung morphogenesis: questions, facts and controversies. *Development* **133**, 1611–1624 (2006).

159. Sengel, P. Pattern formation in skin development. *Int J Dev Biol* **34**, 33–50 (1990).

160. Stiles, J. & Jernigan, T. L. The Basics of Brain Development. *Neuropsychol Rev* **20**, 327 (2010).

161. Morgani, S., Nichols, J. & Hadjantonakis, A. K. The many faces of Pluripotency: in vitro adaptations of a continuum of in vivo states. *BMC Developmental Biology 2017 17:1* **17**, 1–20 (2017).

162. Kim, Y. S. *et al.* Deciphering epiblast lumenogenesis reveals proamniotic cavity control of embryo growth and patterning. *Sci Adv* **7**, (2021).

163. Kinder, S. J. *et al.* The orderly allocation of mesodermal cells to the extraembryonic structures and the anteroposterior axis during gastrulation of the mouse embryo. *Development* **126**, 4691–4701 (1999).

164. Lawson, K. A. & Pedersen, R. A. Clonal Analysis of Cell Fate During Gastrulation and Early Neurulation in the Mouse. in *Ciba Foundation Symposium 165 - Postimplantation Development in the Mouse* vol. 165 3–26 (John Wiley & Sons, Ltd, 2007).

165. Quinlan, G. A., Williams, E. A., Tan, S. S. & Tam, P. P. L. Neuroectodermal fate of epiblast cells in the distal region of the mouse egg cylinder: implication for body plan organization during early embryogenesis. *Development* **121**, 87–98 (1995).

166. Tam, P. P. L. & Beddington, R. S. P. The formation of mesodermal tissues in the mouse embryo during gastrulation and early organogenesis. *Development* **99**, 109–126 (1987).

167. Martello, G. & Smith, A. The Nature of Embryonic Stem Cells. *Annu Rev Cell Dev Biol* **30**, 647–675 (2014).

168. Ying, Q. L., Nichols, J., Chambers, I. & Smith, A. BMP induction of Id proteins suppresses differentiation and sustains embryonic stem cell self-renewal in collaboration with STAT3. *Cell* **115**, 281–292 (2003).

169. Nichols, J., Evans, E. P. & Smith, A. G. Establishment of germ-line-competent embryonic stem (ES) cells using differentiation inhibiting activity. *Development* **110**, 1341–1348 (1990).

170. Williams, R. L. *et al.* Myeloid leukaemia inhibitory factor maintains the developmental potential of embryonic stem cells. *Nature* **336**, 684–687 (1988).

171. Smith, A. *et al.* Inhibition of pluripotential embryonic stem cell differentiation by purified polypeptides. *Nature* **336**, (1988).

172. Gaspar-Maia, A., Alajem, A., Meshorer, E. & Ramalho-Santos, M. Open chromatin in pluripotency and reprogramming. *Nature Reviews Molecular Cell Biology 2011 12:1* **12**, 36–47 (2010).

173. Ahmed, K. *et al.* Global chromatin architecture reflects pluripotency and lineage commitment in the early mouse embryo. *PLoS One* **5**, (2010).

174. Efroni, S. *et al.* Global transcription in pluripotent embryonic stem cells. *Cell Stem Cell* **2**, 437–447 (2008).

175. Aoto, T., Saitoh, N., Ichimura, T., Niwa, H. & Nakao, M. Nuclear and chromatin reorganization in the MHC-Oct3/4 locus at developmental phases of embryonic stem cell differentiation. *Dev Biol* **298**, 354–367 (2006).

176. Chen, G. *et al.* Single-cell analyses of X Chromosome inactivation dynamics and pluripotency during differentiation. *Genome Res* **26**, 1342–1354 (2016).

177. Yeo, J. C. & Ng, H. H. The transcriptional regulation of pluripotency. *Cell Research 2013 23:1* **23**, 20–32 (2012).

178. Nichols, J. & Smith, A. Pluripotency in the Embryo and in Culture. *Cold Spring Harb Perspect Biol* (2012).

179. Luo, J. *et al.* Placental abnormalities in mouse embryos lacking the orphan nuclear receptor ERR-β. *Nature 1997 388:6644* **388**, 778–782 (1997).

180. Davenport, T. G., Jerome-Majewska, L. A. & Papaioannou, V. E. Mammary gland, limb and yolk sac defects in mice lacking Tbx3,the gene mutated in human ulnar mammary syndrome. *Development* **130**, 2263–2273 (2003).

181. Dahl, J. A., Reiner, A. H., Klungland, A., Wakayama, T. & Collas, P. Histone H3 Lysine 27 Methylation Asymmetry on Developmentally-Regulated Promoters Distinguish the First Two Lineages in Mouse Preimplantation Embryos. *PLoS One* **5**, e9150 (2010).

182. Rugg-Gunn, P. J., Cox, B. J., Ralston, A. & Rossant, J. Distinct histone modifications in stem cell lines and tissue lineages from the early mouse embryo. *Proc Natl Acad Sci U S A* **107**, 10783–10790 (2010).

183. Alder, O. *et al.* Ring1B and Suv39h1 delineate distinct chromatin states at bivalent genes during early mouse lineage commitment. *Development* **137**, 2483–2492 (2010).

184. Tang, F. *et al.* Tracing the Derivation of Embryonic Stem Cells from the Inner Cell Mass by Single-Cell RNA-Seq Analysis. *Cell Stem Cell* **6**, 468–478 (2010).

185. Martin Gonzalez, J. *et al.* Embryonic Stem Cell Culture Conditions Support Distinct States Associated with Different Developmental Stages and Potency. *Stem Cell Reports* **7**, 177–191 (2016).

186. Brons, I. G. M. *et al.* Derivation of pluripotent epiblast stem cells from mammalian embryos. *Nature* **448**, 191–195 (2007).

187. Osorno, R. *et al.* The developmental dismantling of pluripotency is reversed by ectopic Oct4 expression. *Development* **139**, 2288–2298 (2012).

188. Tesar, P. J. *et al.* New cell lines from mouse epiblast share defining features with human embryonic stem cells. *Nature 2007 448:7150* **448**, 196–199 (2007).

189. Kojima, Y. *et al.* The Transcriptional and Functional Properties of Mouse Epiblast Stem Cells Resemble the Anterior Primitive Streak. *Cell Stem Cell* **14**, 107–120 (2014).

190. Tosolini, M. & Jouneau, A. From naive to primed pluripotency: In vitro conversion of mouse embryonic stem cells in epiblast stem cells. *Methods in Molecular Biology* **1341**, 209–216 (2015).

191. Guo, G. *et al.* Klf4 reverts developmentally programmed restriction of ground state pluripotency. *Development* **136**, 1063–1069 (2009).

192. Yu, L. *et al.* Derivation of Intermediate Pluripotent Stem Cells Amenable to Primordial Germ Cell Specification. *Cell Stem Cell* **28**, 550-567.e512 (2021).

193. Wang, X. *et al.* Formative pluripotent stem cells show features of epiblast cells poised for gastrulation. *Cell Research 2021 31:5* **31**, 526–541 (2021).

194. Smith, A. G. *et al.* Inhibition of pluripotential embryonic stem cell differentiation by purified polypeptides. *Nature* **336**, 688–690 (1988).

195. Niwa, H., Burdon, T., Chambers, I. & Smith, A. Self-renewal of pluripotent embryonic stem cells is mediated via activation of STAT3. *Genes Dev* **12**, 2048–2060 (1998).

196. Yoshida, K. *et al.* Maintenance of the pluripotential phenotype of embryonic stem cells through direct activation of gp130 signalling pathways. *Mech Dev* **45**, 163–171 (1994).

197. Nichols, J., Chambers, I., Taga, T. & Smith, A. Physiological rationale for responsiveness of mouse embryonic stem cells to gp130 cytokines. *Development* **128**, 2333–2339 (2001).

198. Nichols, J. *et al.* Complementary tissue-specific expression of LIF and LIF-receptor mRNAs in early mouse embryogenesis. *Mech Dev* **57**, 123–131 (1996).

199. Nakashima, K. *et al.* Developmental requirement of gp130 signaling in neuronal survival and astrocyte differentiation. *J Neurosci* **19**, 5429–5434 (1999).

200. Ware, C. B. *et al.* Targeted disruption of the low-affinity leukemia inhibitory factor receptor gene causes placental, skeletal, neural and metabolic defects and results in perinatal death. *Development* **121**, 1283–1299 (1995).

201. Li, M., Sendtner, M. & Smith, A. Essential function of LIF receptor in motor neurons. *Nature* **378**, 724–727 (1995).

202. Do, D. V. *et al.* A genetic and developmental pathway from STAT3 to the OCT4-NANOG circuit is essential for maintenance of ICM lineages in vivo. *Genes Dev* **27**, 1378–1390 (2013).

203. Doble, B. & Woodgett, J. R. GSK-3 : tricks of the trade for a multi-tasking kinase. *J Cell Sci.* **116**, 1175–1186 (2003).

204. Clevers, H. Wnt/β-Catenin Signaling in Development and Disease. *Cell* **127**, 469–480 (2006).

205. Bain, J. *et al.* The selectivity of protein kinase inhibitors: a further update. *Biochem J* **408**, 297–315 (2007).

206. Murray, J. T. *et al.* Exploitation of KESTREL to identify NDRG family members as physiological substrates for SGK1 and GSK3. *Biochem J* **384**, 477–488 (2004).

207. ten Berge, D. *et al.* Embryonic stem cells require Wnt proteins to prevent differentiation to epiblast stem cells. *Nature Cell Biology 2011 13:9* **13**, 1070–1075 (2011).

208. Endoh, M. & Niwa, H. Stepwise pluripotency transitions in mouse stem cells. *EMBO Rep* **23**, (2022).

209. Gordon, M. D. & Nusse, R. Wnt Signaling: Multiple Pathways, Multiple Receptors, and Multiple Transcription Factors. *Journal of Biological Chemistry* **281**, 22429–22433 (2006).

210. Martello, G., Sugimoto, T., Diamanti, E., Joshi, A. & Hannah, R. Esrrb Is a Pivotal Target of the Gsk3 / Tcf3 Axis Regulating Embryonic Stem Cell Self-Renewal. *Cell Stem Cell* **11**, 491–504 (2012).

211. Kelly, K. F. *et al.* β-Catenin Enhances Oct-4 Activity and Reinforces Pluripotency through a TCF-Independent Mechanism. *Cell Stem Cell* **8**, 214–227 (2011).

212. Yi, F. *et al.* Opposing effects of Tcf3 and Tcf1 control Wnt stimulation of embryonic stem cell self-renewal. *Nature Cell Biology 2011 13:7* **13**, 762–770 (2011).

213. Thisse, B. & Thisse, C. Functions and regulations of fibroblast growth factor signaling during embryonic development. *Dev Biol* **287**, 390–402 (2005).

214. Roux, P. P. & Blenis, J. ERK and p38 MAPK-Activated Protein Kinases: a Family of Protein Kinases with Diverse Biological Functions. *Microbiology and Molecular Biology Reviews* **68**, 320 (2004).

215. Kunath, T. *et al.* FGF stimulation of the Erk1 / 2 signalling cascade triggers transition of pluripotent embryonic stem cells from self-renewal to lineage commitment. *Development* **134**, 2895–2902 (2007).

216. Stavridis, M. P., Simon Lunn, J., Collins, B. J. & Storey, K. G. A discrete period of FGF-induced Erk1/2 signalling is required for vertebrate neural specification. *Development* **134**, 2889–2894 (2007).

217. Saiz, N., Grabarek, J. B., Sabherwal, N., Papalopulu, N. & Plusa, B. Atypical protein kinase C couples cell sorting with primitive endoderm maturation in the mouse blastocyst. *Development (Cambridge)* **140**, 4311–4322 (2013).

218. Nichols, J., Silva, J., Roode, M. & Smith, A. Suppression of Erk signalling promotes ground state pluripotency in the mouse embryo. *Development* **136**, 3215–3222 (2009).

219. Yeo, J. C. *et al.* Klf2 Is an Essential Factor that Sustains Ground State Pluripotency. *Cell Stem Cell* **14**, 864–872 (2014).

220. Tee, W. W., Shen, S. S., Oksuz, O., Narendra, V. & Reinberg, D. Erk1/2 Activity Promotes Chromatin Features and RNAPII Phosphorylation at Developmental Promoters in Mouse ESCs. *Cell* **156**, 678–690 (2014).

221. Hébert, J. M., Rosenquist, T., Götz, J. & Martin, G. R. FGF5 as a regulator of the hair growth cycle: evidence from targeted and spontaneous mutations. *Cell* **78**, 1017–1025 (1994).

222. Niswander, L. & Martin, G. R. Fgf-4 expression during gastrulation, myogenesis, limb and tooth development in the mouse. *Development* **114**, 755–768 (1992).

223. Rappolee, D. A., Basilico, C., Patel, Y. & Werb, Z. Expression and function of FGF-4 in peri-implantation development in mouse embryos. *Development* **120**, 2259–2269 (1994).

224. Feldman, B., Poueymirou, W., Papaioannou, V. E., DeChiara, T. M. & Goldfarb, M. Requirement of FGF-4 for postimplantation mouse development. *Science* **267**, 246–249 (1995).

225. Goldin, S. N. & Papaioannou, V. E. Paracrine action of FGF4 during periimplantation development maintains trophectoderm and primitive endoderm. *Genesis* **36**, 40–47 (2003).

226. Yamanaka, Y., Lanner, F. & Rossant, J. FGF signal-dependent segregation of primitive endoderm and epiblast in the mouse blastocyst. *Development* **137**, 715–724 (2010).

227. Molotkov, A., Mazot, P., Brewer, J. R., Cinalli, R. M. & Soriano, P. Distinct Requirements for Fgfr1 and Fgfr2 in Primitive Endoderm Development and Exit from Pluripotency. *Dev Cell* **41**, 511 (2017).

228. Wang, J. *et al.* A protein interaction network for pluripotency of embryonic stem cells. *Nature 2006 444:7117* **444**, 364–368 (2006).

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

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

231. Maherali, N. *et al.* Directly reprogrammed fibroblasts show global epigenetic remodeling and widespread tissue contribution. *Cell Stem Cell* **1**, 55–70 (2007).

232. Wernig, M. *et al.* In vitro reprogramming of fibroblasts into a pluripotent ES-cell-like state. *Nature* **448**, 318–324 (2007).

233. Takahashi, K. *et al.* Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell* **131**, 861–872 (2007).

234. Okita, K., Ichisaka, T. & Yamanaka, S. Generation of germline-competent induced pluripotent stem cells. *Nature* **448**, 313–317 (2007).

235. Hall, J. *et al.* Oct4 and LIF/Stat3 additively induce Krüppel factors to sustain embryonic stem cell self-renewal. *Cell Stem Cell* **5**, 597–609 (2009).

236. Xie, L. *et al.* A dynamic interplay of enhancer elements regulates Klf4 expression in naïve pluripotency. *Genes Dev* **31**, 1795–1808 (2017).

237. Yoon, H. S. & Yang, V. W. Requirement of Krüppel-like Factor 4 in Preventing Entry into Mitosis following DNA Damage. *Journal of Biological Chemistry* **279**, 5035–5041 (2004).

238. Chen, X. *et al.* Krüppel-like Factor 4 (Gut-enriched Krüppel-like Factor) Inhibits Cell Proliferation by Blocking G1/S Progression of the Cell Cycle. *Journal of Biological Chemistry* **276**, 30423–30428 (2001).

239. Zhang, P., Andrianakos, R., Yang, Y., Liu, C. & Lu, W. Kruppel-like factor 4 (Klf4) prevents embryonic stem (ES) cell differentiation by regulating Nanog gene expression. *J Biol Chem* **285**, 9180–9189 (2010).

240. Carbognin, E. *et al.* Esrrb guides naive pluripotent cells through the formative transcriptional programme. *Nat Cell Biol* **25**, 643–657 (2023).

241. Boroviak, T. *et al.* Lineage-Specific Profiling Delineates the Emergence and Progression of Naive Pluripotency in Mammalian Embryogenesis. *Dev Cell* **35**, 366–382 (2015).

242. Yang, P. *et al.* Multi-omic Profiling Reveals Dynamics of the Phased Progression of Pluripotency. *Cell Syst* **8**, 427-445.e10 (2019).

243. Kalkan, T. & Smith, A. Mapping the route from naive pluripotency to lineage specification. *Phil. Trans. R. Soc. B* **369**, (2014).

244. Niwa, H. The principles that govern transcription factor network functions in stem cells. *Development* **145**, dev157420 (2018).

245. Okano, M., Bell, D. W., Haber, D. A. & Li, E. DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell* **99**, 247–257 (1999).

246. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

247. Zhu, A., Ibrahim, J. G. & Love, M. I. Heavy-tailed prior distributions for sequence count data: removing the noise and preserving large differences. *Bioinformatics* **35**, 2084–2092 (2019).

248. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 1–21 (2014).

249. Gramacy, R. B. tgp: An R Package for Bayesian Nonstationary, Semiparametric Nonlinear Regression and Design by Treed Gaussian Process Models. *J Stat Softw* **19**, 1–46 (2007).

250. Krueger, F. Babraham Bioinformatics - Trim Galore! http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/.

251. Lackner, A. *et al.* Cooperative genetic networks drive embryonic stem cell transition from naïve to formative pluripotency. *EMBO J* **40**, e105776 (2021).

252. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)* **57**, 289–300 (1995).

253. Gaujoux, R. & Seoighe, C. A flexible R package for nonnegative matrix factorization. *BMC Bioinformatics* **11**, 1–9 (2010).

254. Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573-3587.e29 (2021).

255. Van Der Maaten, L. & Hinton, G. Visualizing Data using t-SNE. *Journal of Machine Learning Research* **9**, 2579–2605 (2008).

256. Macnair, W., Gupta, R. & Claassen, M. psupertime: supervised pseudotime analysis for time-series single-cell RNA-seq data. *Bioinformatics* **38**, i290–i298 (2022).

257. Kinoshita, M. *et al.* Capture of Mouse and Human Stem Cells with Features of Formative Pluripotency. *Cell Stem Cell* **28**, 453-471.e8 (2021).

258. Neph, S. *et al.* BEDOPS: high-performance genomic feature operations. *Bioinformatics* **28**, 1919–1920 (2012).

259. Weirauch, M. T. *et al.* Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* **158**, 1431–1443 (2014).

260. Cuellar-Partida, G. *et al.* Gene expression Epigenetic priors for identifying active transcription factor binding sites. *Bioinformatics* **28**, 56–62 (2012).

261. Sikora-Wohlfeld, W., Ackermann, M., Christodoulou, E. G., Singaravelu, K. & Beyer, A. Assessing Computational Methods for Transcription Factor Target Gene Identification Based on ChIP-seq Data. *PLoS Comput Biol* **9**, e1003342 (2013).

262. Tremblay, B. & Nystrom, S. Universalmotif: Import, modify, and export motifs with R. *R package Version* **1**, (2020).

263. Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32**, 2847–2849 (2016).

264. Warnes, G. R. *et al.* gplots: Various R programming tools for plotting data. *R package version* **2**, 1 (2009).

265. Wei, T. *et al.* Package 'corrplot'. *Statistician* **56**, e24 (2017).

266. Conway, J. R., Lex, A. & Gehlenborg, N. UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics* **33**, 2938–2940 (2017).

267. Wickham, H. *Ggplot2: Elegant Graphics for Data Analysis*. *Use R!* (Springer International Publishing, Cham, Switzerland, 2016).

268. Kanehisa, M., Furumichi, M., Sato, Y., Kawashima, M. & Ishiguro-Watanabe, M. KEGG for taxonomy-based analysis of pathways and genomes. *Nucleic Acids Res* **51**, D587–D592 (2023).

269. Kanehisa, M. Toward understanding the origin and evolution of cellular organisms. *Protein Sci* **28**, 1947–1951 (2019).

270. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* **28**, 27–30 (2000).

271. Festuccia, N. *et al.* Esrrb Is a Direct Nanog Target Gene that Can Substitute for Nanog Function in Pluripotent Cells. *Cell Stem Cell90* **11**, 477–490 (2012).

272. Jayaram, S., Romeike, M. & Buecker, C. The asynchrony in the exit from naive pluripotency cannot be explained by differences in the cell cycle phase. *bioRxiv* 2023.09.15.557731 (2023) doi:10.1101/2023.09.15.557731.

273. Mulas, C. *et al.* ERK signalling orchestrates metachronous transition from naïve to formative pluripotency. *bioRxiv* 2023.07.20.549835 (2023) doi:10.1101/2023.07.20.549835.

274. Kalkan, uzer *et al.* Complementary Activity of ETV5, RBPJ, and TCF3 Drives Formative Transition from Naive Pluripotency. *Cell Stem Cell* **24**, 785–801 (2019).

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

276. Zhang, X., Zhang, J., Wang, T., Esteban, M. A. & Pei, D. Esrrb activates Oct4 transcription and sustains self-renewal and pluripotency in embryonic stem cells. *J Biol Chem* **283**, 35825–35833 (2008).

277. Liu, F., van den Broek, O., Destrée, O. & Hoppler, S. Distinct roles for Xenopus Tcf/Lef genes in mediating specific responses to Wnt/β-catenin signalling in mesoderm development. *Development* **132**, 5375–5385 (2005).

278. Pereira, L., Yi, F. & Merrill, B. J. Repression of Nanog gene transcription by Tcf3 limits embryonic stem cell self-renewal. *Mol Cell Biol* **26**, 7479–7491 (2006).

279. Cole, M. F., Johnstone, S. E., Newman, J. J., Kagey, M. H. & Young, R. A. Tcf3 is an integral component of the core regulatory circuitry of embryonic stem cells. *Genes Dev* **22**, 746–755 (2008).

280. Buecker, C. *et al.* Reorganization of enhancer patterns in transition from naïve to primed pluripotency. *Cell Stem Cell* **14**, 838 (2014).

281. Yang, S. H. *et al.* ZIC3 Controls the Transition from Naive to Primed Pluripotency. *Cell Rep* **27**, 3215-3227.e6 (2019).

282. Ishii, S. *et al.* Genome-wide ATAC-see screening identifies TFDP1 as a modulator of global chromatin accessibility. *Nature Genetics 2024* 1–10 (2024) doi:10.1038/s41588-024-01658-1.

283. Yang, C. S., Chang, K. Y. & Rana, T. M. Genome-wide functional analysis reveals factors needed at the transition steps of induced reprogramming. *Cell Rep* **8**, 327–337 (2014).

284. Zhang, J. *et al.* Sall4 modulates embryonic stem cell pluripotency and early embryonic development by the transcriptional regulation of Pou5f1. *Nature Cell Biology 2006 8:10* **8**, 1114–1123 (2006).

285. Miller, A. *et al.* Sall4 controls differentiation of pluripotent cells independently of the Nucleosome Remodelling and Deacetylation (NuRD) complex. *Development* **143**, 3074 (2016).

286. Wang, B. *et al.* Induction of Pluripotent Stem Cells from Mouse Embryonic Fibroblasts by Jdp2-Jhdm1b-Mkk6-Glis1-Nanog-Essrb-Sall4. *Cell Rep* **27**, 3473-3485.e5 (2019).

287. Wang, Z.-X. *et al.* The transcription factor Zfp281 controls embryonic stem cell pluripotency by direct activation and repression of target genes. *Stem Cells* **26**, 2791–2799 (2008).

288. Fidalgo, M. *et al.* Zfp281 mediates Nanog autorepression through recruitment of the NuRD complex and inhibits somatic cell reprogramming. *Proceedings of the National Academy of Sciences* **109**, 16202–16207 (2012).

289. Huang, X. *et al.* Zfp281 is essential for mouse epiblast maturation through transcriptional and epigenetic control of Nodal signaling. *Elife* **6**, (2017).

290. Fidalgo, M. *et al.* Zfp281 functions as a transcriptional repressor for pluripotency of mouse embryonic stem cells. *Stem Cells* **29**, 1705–1716 (2011).

291. Fidalgo, M. *et al.* Zfp281 Coordinates Opposing Functions of Tet1 and Tet2 in Pluripotent States. *Cell Stem Cell* **19**, 355–369 (2016).

292. Betschinger, J. *et al.* Exit from pluripotency is gated by intracellular redistribution of the bHLH transcription factor Tfe3. *Cell* **153**, 335–347 (2013).

293. Mayer, D. *et al.* Zfp281 orchestrates interconversion of pluripotent states by engaging Ehmt1 and Zic2. *EMBO J* **39**, (2020).

294. Huang, X. *et al.* ZFP281 controls transcriptional and epigenetic changes promoting mouse pluripotent state transitions via DNMT3 and TET1. *Dev Cell* (2024) doi:10.1016/J.DEVCEL.2023.12.018.

295. Sefer, E., Kleyman, M. & Bar-Joseph, Z. Tradeoffs between dense and replicate sampling strategies for high throughput time series experiments. *Cell Syst* **3**, 35 (2016).

296. Rasmussen, C. E. Gaussian Processes in Machine Learning. in *Advanced Lectures on Machine Learning: ML Summer Schools 2003, Canberra, Australia, February 2 - 14, 2003, Tübingen, Germany, August 4 - 16, 2003, Revised Lectures* (ed. Bousquet Olivier and von Luxburg, U. and R. G.) 63–71 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2004). doi:10.1007/978-3-540-28650-9_4.

297. Rasmussen, C. E. & Williams, C. K. I. Gaussian Processes for Machine Learning. *the MIT Press* (2006).

298. Huth, M. *et al.* NMD is required for timely cell fate transitions by fine-tuning gene expression and regulating translation. *Genes Dev* **34**, 348–367 (2022).

299. Mahat, D. B. & Lis, J. T. Use of conditioned media is critical for studies of regulation in response to rapid heat shock. *Cell Stress Chaperones* **22**, 155 (2017).

300. Smith, K. N., Singh, A. M. & Dalton, S. Myc Represses Primitive Endoderm Differentiation in Pluripotent Stem Cells. *Cell Stem Cell* **7**, 343–354 (2010).

301. Hicks, S. C., Townes, F. W., Teng, M. & Irizarry, R. A. Missing data and technical variability in single-cell RNA-sequencing experiments. *Biostatistics* **19**, 562–578 (2018).

302. Pauklin, S. & Vallier, L. The Cell-Cycle State of Stem Cells Determines Cell Fate Propensity. *Cell* **155**, 135–147 (2013).

303. Waisman, A. *et al.* Cell cycle dynamics of mouse embryonic stem cells in the ground state and during transition to formative pluripotency. *Scientific Reports 2019 9:1* **9**, 1–10 (2019).

304. Chaigne, A. *et al.* Abscission Couples Cell Division to Embryonic Stem Cell Fate Chaigne et al. *Dev Cell* **55**, 195-208.e5 (2020).

305. Singh, A. M. *et al.* Cell-cycle control of developmentally regulated transcription factors accounts for heterogeneity in human pluripotent cells. *Stem Cell Reports* **1**, 532–544 (2013).

306. Strawbridge, S. E., Blanchard, G. B., Smith, A., Kugler, H. & Martello, G. Embryonic stem cells commit to differentiation by symmetric divisions following a variable lag period. *bioRxiv* 2020.06.17.157578 (2020) doi:10.1101/2020.06.17.157578.

307. Waisman, A. *et al.* Inhibition of Cell Division and DNA Replication Impair Mouse-Naïve Pluripotency Exit. *J Mol Biol* **429**, 2802–2815 (2017).

308. Deng, Q., Ramsköld, D., Reinius, B. & Sandberg, R. Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science* **343**, 193–196 (2014).

309. Nakamura, T. *et al.* A developmental coordinate of pluripotency among mice, monkeys and humans. *Nature* **537**, 57–62 (2016).

310. Beyer, A., Hollunder, J., Nasheuer, H. P. & Wilhelm, T. Post-transcriptional expression regulation in the yeast Saccharomyces cerevisiae on a genomic scale. *Mol Cell Proteomics* **3**, 1083–1092 (2004).