Gene regulation in embryonic development

Berg, P.R. van den

Gene regulation in embryonic development

Patrick R. van den Berg

Gene regulation in embryonic development

Gene regulation in embryonic development

PROEFSCHRIFT

ter verkrijging van
de graad van doctor aan de Universiteit Leiden,
op gezag van rector magnificus prof.dr.ir. H. Bijl,
volgens besluit van het College voor Promoties
te verdedigen op 19 mei 2021
klokke 13:45 uur

door

Patrick Robert van den Berg

geboren te Amsterdam, Nederland
in 1990

Promotor:	Prof. dr. T. Schmidt
Co-promotor:	Dr. S. Semrau
Promotiecommissie:	Dr. V. V. Orlova (Leids Universitair Medisch Centrum)
	Dr. J. van Zon (AMOLF)
	Prof. dr. E.R. Eliel
	Prof. dr. ir. S. J. T. van Noort
	Prof. dr. M. E. Drukker

Cover (front): "Not a Gene" - A matrix of DNA letters (A,C,G,T,N) together forming a pixelized representation of a DNA molecule.
Cover (back): "Not a Pipe" - Maggritte's famous image of a pipe in the same style as the image on the front.

Casimir PhD Series, Delft-Leiden, 2021-9
ISBN 978-90-8593-475-2
An electronic version of this thesis can be found at
https://openaccess.leidenuniv.nl

This work was supported by the Netherlands Organisation for Scientific Research (NWO/OCW), as part of the Frontiers of Nanoscience (NanoFront) program.
Data analysis was carried out on the Dutch national e-infrastructure with the support of SURF Foundation.

“But man is not made for defeat,” he said. “A man can be destroyed but not defeated.”

– Ernest Hemingway, *The Old Man and the Sea*

Voor mijn ouders.

CONTENTS

INTRODUCTION

Our bodies are unbelievably complex systems. An important contributor to this complexity is the enormous diversity of cell types. Each cell type has its own properties and functions, such as the transfer of information in the brain or the transport of oxygen in the bloodstream. Despite all this diversity, each cell in our body has the same origin: a single fertilized egg cell. Developmental biology is the science of life's journey from this single cell, to trillions of cells with hundreds of different cell types.

But what even defines a cell type? The exact answer to this is not trivial and can be approached at different levels. Functional specialization or location in the body can both be used to define a cell type. What a cell can do (which is largely determined by its protein composition) and where a cell is located (when and where it was formed during development) is all determined by the genome of that single fertilized egg cell.

To discover what parts of the genome, i.e. which genes, are active when and where in a developing embryo, is thus an important task of developmental biology. Various biochemical processes, collectively termed *gene regulatory mechanisms*, determine the amount of transcription, translation and protein degradation, resulting in precisely tuned protein abundances. These processes are at the center of defining a cell type, since they ultimately determine its protein composition. In this thesis we will study embryonic development with gene regulation as a common thread. Here we will introduce the different types of gene regulatory mechanisms discussed in this thesis, as well as some of the state-of-the-art experimental techniques we employ to study these mechanisms.

DNA methylation

The human genome comprises 46 DNA molecules, totaling two meters in length, which are packaged inside the cell's nucleus. How tightly different parts of the DNA are packed determines, at least in part, which genes are active. In turn, the gene activity profile is strongly correlated with the cell's type. Which parts of the DNA are accessible or inaccessible is not encoded in the DNA sequence itself but determined by chemical modifications of the DNA or proteins that are associated with it. The field of *epigenetics* studies how this meta-information is modified by the cell and how it determines when a gene is active. One layer of this meta-information is stored in the modifications of a specific DNA base. A methyl-group can be added and removed from cytosine and this triggers a whole cascade of other processes influencing DNA packing. Knowing where throughout the genome the DNA is methylated, its *methylome*, thus reveals vital information about how genes are regulated.

Transcriptional and translational regulation

The instructions on how to build proteins are stored in the genome, but these instructions are not read out directly. In between DNA and protein, messenger RNA (mRNA) acts as a carrier of these instructions. The process of copying DNA into mRNA is called transcription and reading mRNA's instructions for protein is called translation. Both of these are heavily regulated processes.

Transcription is the process of copying DNA sequences to the equivalent of mRNA. In contrast to DNA there can be thousands of copies of mRNA in the cell simultaneously. The amount of mRNA in the cell will largely determine the amount of the corresponding protein in the cell so controlling transcription is vitally important. There are many actors that bind DNA that are responsible for either promoting or repressing transcription. These molecules are called transcription factors and at any given transcription site there is typically a complex set of these present that fine-tune transcription levels. Conversely, each transcription factor can bind multiple locations in the genome. In a famous example, there is a set of transcription factors that, when introduced into the cell, can reprogram a mature cell type back into a progenitor cell. Another important role in transcriptional regulation is that of the enhancer. These regions of DNA are typically positioned far away from the transcription site but can nonetheless control the transcription level.

Translation is the process of converting genes encoded in mRNA into proteins. Akin to transcriptional regulation, translation can be regulated in multiple ways. Most regulatory mechanisms involve the binding of proteins or other RNAs to specific sequences in the mRNA. One mechanism studied in this thesis involves micro-RNAs (miRs). miRs are short pieces of RNA that bind to complementary sequences in their target mRNAs. They are known to either signal *slicer* proteins to degrade the mRNA or to simply block the translation machinery. When it comes to important cellular processes like differentiation, big changes in translational regulation is much less prevalent than in transcriptional regulation. However, in particular cases translational regulation plays a decisive role and should not be overlooked.

Omics

To measure is to know, therefore science is forever pushing the limits of measurements. This thesis makes use of several high-throughput methods to measure the molecular profiles of cells. In order to discern what is being measured we use a set of suffixes. For instance we measure the gen*ome* with gen*omics* and the prote*ome* with prote*omics*. *-ome* relates to all the information the type of molecule holds while *-omics* describes the measurement of this information. In this thesis we will deal with the genome (the DNA sequence), the methylome (where the DNA is methylated), the transcriptome (the composition of mRNAs), the mirnome (the composition of miRs) and the proteome (the composition of proteins). All of these measurements have been made possible by recent scientific and commercial advances.

In particular DNA sequencing methods have recently driven transcriptomics and genomics. Mass spectrometry is one of the main tools to measure the proteome. Most of these -omics tools measure many cells to get accurate averages, but a recent trend is to measure single cells. Single cell -omics allows us to study cell populations without averaging out important variability between cells.

Thesis outline

In this thesis we explore the ways cells exert control over their states, particularly in the context of development. We mirror the central dogma of molecular biology in the order of the chapters: from DNA to RNA to protein.

In **Chapter 1** we look at the dynamics of DNA methylation in mouse embryonic stem cells (ESCs). It has been previously found that there are regions in the genome that are differentially methylated depending on the cell type. Bulk DNA sequencing based methods can only provide a static snapshot of average methylation levels. In particular, bulk methods obscure heterogeneity within the cell population, which might have important functional consequences. Here we used a locus-specific fluorescent reporter for DNA methylation to study two super enhancers sites. Our study reveals that their methylation level is highly dynamic and correlated with gene expression. Moreover, methylation changes occur independently on the two alleles (i.e. the mother's and father's copy of the super enhancers). In summary, Chapter 1 demonstrates a new and unique tool for studying enhancer DNA methylation in heterogeneous populations of cells.

In **Chapter 2** we look at how the human kidney develops. Most research on kidney development has been done in mice, even though there are many crucial differences between the species. We were one of the first to measure human fetal kidneys with a new state-of-the-art technique: single-cell RNA-sequencing. In this dataset we identified 22 distinct kidney cell types at five developmental ages. One of our key findings is that there are several subclasses of nephron progenitors, which give rise to the basic functional unit of the kidney. In summary, Chapter 2 creates a better picture of how the human kidney develops and how it compares to the mouse kidney.

In **Chapter 3** we look at the discordance between mRNA and protein abundance to discover mechanisms of translational regulation. mRNA changes are often used as surrogate for protein changes in the cell. However, protein expression over time does not directly correspond to mRNA expression due to finite rates of protein synthesis and degradation and a resulting delay. For example, mRNA expression may spike upwards in a matter of minutes, but protein expression may take a day to catch up, owing to a slow protein synthesis rate. To better understand these differences we measured both mRNA and protein in a mouse ESCs differentiation experiment. By modeling protein expression as a birth-death process we discern which gene is entirely determined by mRNA abundance and which is the target of post-transcriptional regulation. We further integrate a small RNA-sequencing dataset into

our model to identify miRs that may be responsible for some of this regulation. In summary, Chapter 3 determines the important differences between mRNA and protein dynamics and uses these differences to identify cases of post-transcriptional regulation.

In **Chapter 4** follows up on some of the results of the preceding chapter. The regulation that miRs exert on mRNAs adds a layer of complexity to gene regulation that is often ignored. This is in part due to the enormous number of possible miR-gene interactions and the ambiguity of the binding sites. In the previous chapter we used our birth-death model to predict several miR-gene interactions in mouse ESC differentiation. Here, we set out to validate some of these interactions by introducing mimics and inhibitors of these miRs into ESCs. We created reporter cell lines of miR activity to accurately select the timing and dosage of those reagents. We found that four out of six candidate miRs down-regulated their predicted target. In summary, Chapter 4 validates the birth-death model as a tool for finding miR-gene interactions.

1 Dynamic Enhancer DNA Methylation as Basis for Transcriptional and Cellular Heterogeneity of ESCs

This chapter is based on:
Yuelin Song, Patrick R van den Berg, Styliani Markoulaki, Frank Soldner, Alessandra Dall'Agnese, Jonathan E Henninger, Jesse Drotar, Nicholas Rosenau, Malkiel A Cohen, Richard A Young, Stefan Semrau, Yonatan Stelzer, Rudolf Jaenisch. "Dynamic Enhancer DNA Methylation as Basis for Transcriptional and Cellular Heterogeneity of ESCs". In: *Molecular cell* 0.0 (2019), 905–920.e6. DOI: 10.1016/j.molcel.2019.06.045

Chapter preface

The following chapter is a near verbatim reproduction of Song et al. [1] (supplementary figures, however, are not reproduced and for these I refer to in the original article). This chapter demonstrates that DNA methylation is a highly dynamic process in steady-state mouse embryonic stem cells (ESCs). Furthermore, is it shown that DNA methylation of super-enhancer loci correlates strongly with gene expression, which makes the observed fluctuations highly relevant for ESC biology. I contributed significantly to making the connection between DNA methylation and gene expression.

To measure DNA methylation of specific loci in live cells, fluorescent reporter lines were created. By using F1 hybrid cell lines, it was possible to create independent reporters (with distinguishable fluorophores) for the paternal and maternal allele, respectively. After careful characterization of the methylation dynamics, we set out to study the effect on gene expression. To that end, cell populations sorted on the reporter signals were profiled by RNA-seq. My task was to analyze the RNA-seq data. The specific challenge was to design an analysis pipeline that distinguishes reads from the two maternal and paternal allele, which differ in only a few nucleotides. Moreover I assisted by reanalysing single cell WGBS (scWGBS) data and validating the expression levels of *Sox2* in these sorted populations using single-molecule FISH (smFISH).

Patrick van den Berg

Abstract

Variable levels of DNA methylation have been reported at tissue-specific differential methylation regions (DMRs) overlapping enhancers, including super-enhancers (SEs) associated with key cell identity genes, but the mechanisms responsible for this intriguing behavior are not well understood. We used allele-specific reporters at the endogenous *Sox2* and *Mir290* SEs in embryonic stem cells and found that the allelic DNA methylation state is dynamically switching, resulting in cell-to-cell heterogeneity. Dynamic DNA methylation is driven by the balance between DNA methyltransferase and transcription factor binding on one side and co-regulated with the Mediator complex recruitment and H3K27ac level changes at regulatory elements on the other side. DNA methylation at the *Sox2* and the *Mir290* SEs is independently regulated and has distinct consequences on the cellular differentiation state. Dynamic allele-specific DNA methylation at the two SEs was also seen at different stages in preimplantation embryos, revealing that methylation heterogeneity occurs *in vivo*.

1.1 Introduction

Tissue-specific differential methylation regions (T-DMRs) have been found to strongly associate with low CpG density and inter-genic enhancers [2, 3, 4, 5, 6], and the vast majority of cell-type specific DNA methylation changes occur at distal regulatory elements [7, 8]. whole-genome bisulfite sequencing (WGBS) data indicate a low but detectable level of DNA methylation at T-DMRs overlapping active enhancers [9, 10, 11, 12, 13, 14, 8]. Recent single cell WGBS (scWGBS) data from mouse emrbyonic stem cells (ESCs) and the early mouse embryo suggest that the variable low-to-intermediate DNA methylation levels found at enhancer regions in bulk-cell measurements are largely due to averaging signals across cells with heterogeneous methylation states [15, 16, 17, 18, 19, 20]. However, due to the static snapshot view of sequencing-based methods, it has been difficult to define the basis, regulation, and functional impact of DNA methylation heterogeneity on gene expression and cellular states.

The hierarchy and casual relationship between the regulation of enhancer DNA methylation, active enhancer histone marks, transcription factor (TF) binding, and *cis*-regulated transcription has been challenging to define due to the epigenetic heterogeneity among cells [21, 13, 22]. While genome-wide epigenetic profiling provided insights into the relationship between DNA methylation, histone marks, and TFs and coactivators binding [13, 23, 24], these approaches, even at the single-cell level, did not allow resolving fast dynamics of individual epigenetic processes in heterogeneous tissues and cell populations. Thus, currently there is no clear understanding of the basis, regulation, and functional consequences of DNA methylation heterogeneity.

Our recently developed Reporter of Genome Methylation (RGM) allows tracing of locus-specific DNA methylation based on the on-and-off of a fluorescent signal in single cells in real time, and has been shown to faithfully reflect the endogenous DNA methylation states at multiple genomic loci [25, 26]. This system allows for robustly tracking locus-specific DNA

methylation at enhancer regions and for functionally dissecting the hierarchy of epigenetic events that regulate enhancer activity and cellular states, overcoming the challenges faced by bulk measurements or sequence-based methods. We utilized this system at two pluripotency SEs, *Sox2* and *Mir290* SE, in ESCs. Both SEs overlap with ESC-specific DMRs, which display consistently low levels of methylation, indicating potential heterogeneity [27, 28, 19, 29, 8]. We targeted RGMs to both alleles of the two SEs in F1 129xCasteneous (129xCAST) hybrid ESCs allowing to visualize allele-specific DNA methylation changes. We observed highly dynamic switching between different methylation states on individual alleles resulting in cell-to-cell heterogeneity and were able to distinguish the DNA methylation pathways driving these changes. The RGM system enables isolation of rare and transient populations exclusively based on their locus-specific methylation states, which allowed defining the relationship between dynamic SE DNA methylation changes, the Mediator complex condensation, histone H3K27 acetylation, TF binding, *cis*-regulated target gene expression, and changes in cellular states. Finally, transgenic methylation reporter mice for both SEs revealed the previously underappreciated epigenetic heterogeneity and dynamics of the pluripotent cells in cleavage embryos, recapitulating and extending the observations in ESCs.

1.2 Results

1.2.1 DNA Methylation at the *Sox2* and *Mir290* SEs Is Heterogeneous at the Allelic Level

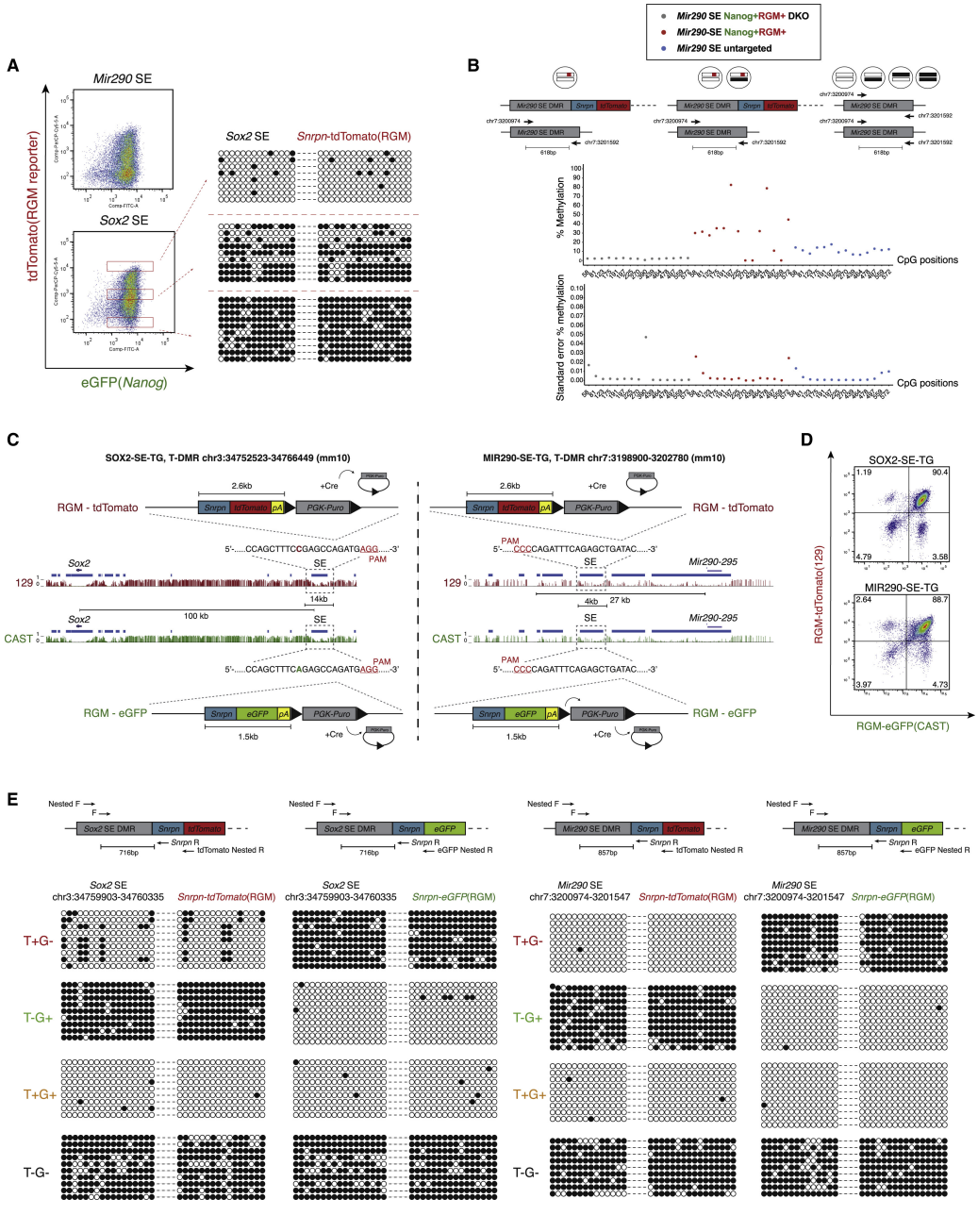
Sox2 and *Mir290* SEs reside on chromosome 3 and 7, respectively. Both SEs overlap with T-DMRs, which are hypo-methylated in ESCs but become *de novo* methylated upon differentiation [25]. The T-DMR of the *Sox2* SE is located about 100 kb upstream of the *Sox2* gene, whereas the *Mir290* SE, consisting of hypo-methylated DMR constituents interspersed by small hyper-methylated regions, is proximal to the *Mir290-295* cluster (Song et al. [1] Figure S1A). WGBS of ESCs indicates that the *Sox2* and *Mir290* SE DMRs have overall DNA methylation levels higher than that of hypo-methylated promoters of highly expressed genes in ESCs, such as *Gapdh* and *Oct4*, but lower than that of imprinting control regions or retroelements, which are monoallelically and hyper-methylated, respectively (Song et al. [1] Figure S1B) [27, 28, 29, 8]. This low-to-intermediate level of methylation at both SEs in bulk cell WGBS suggests that they are hypermethylated in a small population of cells. Re-analysis of published scWGBS data [20] revealed that the T-DMRs of both SEs belong to the 5% regions with the most variable DNA methylation level compared to other regions of chromosome 7 or chromosome 3 (Song et al. [1] Figure S1C), further supporting the presence of rare cells with hypermethylated SE DMRs.

Consistent with published scWGBS studies reporting heterogeneity in the wild-type genome [16, 17, 18, 30, 31, 20], we previously observed methylation heterogeneity in ESCs with the endogenous *Nanog* tagged with eGFP and RGM-tdTomato reporter inserted mono-

allelically into the *Sox2* or *Mir290* SE DMRs [25]. The heterogeneity at these two specific loci was manifested by the bi-modal distribution of RGM activity in Nanog positive ($Nanog^+$) pluripotent cells as seen in FACS (Figure 1A). Sorting cells based on florescence intensity, followed by bisulfate PCR (BS-PCR) and sequencing, validated that RGM methylation strictly correlates with the endogenous methylation in both regions (Figure 1A). Analyzing the *Sox2* SE revealed that hyper-methylation occurred on both the targeted and the untargeted alleles in the pluripotent ESC population ($Nanog^+$), indicating that rare allelic methylation exists among cells (Song et al. [1] Figure S1D). The rare methylated alleles were also detected at the *Mir290* SE by high-throughput sequencing of BS-PCR amplicons from the wild-type allele. Figure 1B shows that, comparing to *Dnmt3a/b* double-knockout cells (described later in Song et al. [1] Figure S3A), we found methylation at the *Mir290* SE in non-manipulated wild-type ESCs as well as on the untargeted allele in the $Nanog^+RGM^+$ ESCs. These results indicate that SE DNA methylation heterogeneity is created by allele-specific hypermethylation in rare ESC populations independent of RGM targeting. To track DNA methylation heterogeneity on each allele, we targeted the *Mir290* and the *Sox2* SE independently in 129xCastaneus F1 hybrid ESCs with allele-specific RGM reporters and generated two cell lines, *Sox2*-$129^{\text{SE-RGM-tdTomato}}$/*Sox2*-$\text{CAST}^{\text{SE-RGM-eGFP}}$ (abbreviated below as SOX2-SE-TG) and *Mir290*-$129^{\text{SE-RGM-tdTomato}}$/*Mir290*-$\text{CAST}^{\text{SE-RGM-eGFP}}$ (abbreviated below as MIR290-SE-TG) (Figure 1C and Song et al. [1] Figure S1E) allowing to visualize the SE locus-specific DNA methylation state at allelic and single-cell resolution. These cell lines also enabled dissection of allelic functional output of SE methylation states by distinguishing the two alleles based on the abundance of 129 or CAST allele-specific single nucleotide polymorphisms (SNPs) at both the DNA and the mRNA level.

The initial FACS analysis detected a small fraction of single-positive (T^+G^-, T^-G^+) as well

Figure 1 *(following page)*. DNA Methylation at the *Sox2* and *Mir290* SEs Is Heterogeneous at the Allelic Level. (A) Left, DNA methylation heterogeneity at both the *Sox2* and the *Mir290* SE in v6.5-*Nanog*-eGFP ESC where the RGM-tdTomato reporter was mono-allelically targeted. Right, BS-PCR followed by sequencing of the *Sox2* SE in different populations of the bimodal distribution. (B) Average methylation percentage and standard errors were quantified from high-throughput sequencing of BS-PCR amplicons of the *Mir290* SE wild-type alleles in *Dnmt3a/b* double-knockout ESCs, in Nanog+RGM+ ESCs and in untargeted wild-type ESCs. BS-PCRs were amplified allele-specifically as illustrated from potential epigenetic states indicated above. Standard error was estimated assuming number of methylated counts as a binomial random variable. (C) Targeting strategy for generating SOX2-SE-TG and *Mir290-SE-TG* ESCs using CRISPR/Cas9 and targeting vectors. Methylation tracks from (Stadler et al., 2011) were used as the genome reference with blue bars highlighting the DMRs of the two SEs. Red tracks, 129 allele; green tracks, CAST allele. (D) FACS analysis of CASTx129 F1 ESC clones targeted with allele-specific RGMs at either the *Mir290* or the *Sox2* SE. (E) Allele-specific BS-PCR of the SEs with RGM (Snprn-tdTomato or Snprn-eGFP) in single PCR amplicons followed by Sanger sequencing in sorted cells from both SOX2-SE-TG and *Mir290-SE-TG* See also Song et al. [1] Figure S1.


A
Mir290 SE
Sox2 SE
tdTomato(RGM reporter)
eGFP(Nanog)
Sox2 SE
Snrpn-tdTomato(RGM)
B
Mir290 SE Nanog+RGM+ DKO
Mir290-SE Nanog+RGM+
Mir290 SE untargeted
% Methylation
CpG positions
Standard error % methylation
CpG positions
C
SOX2-SE-TG, T-DMR chr3:34752523-34766449 (mm10)
MIR290-SE-TG, T-DMR chr7:3198900-3202780 (mm10)
RGM - tdTomato
2.6kb
+Cre
Snrpn
tdTomato
pA
PGK-Puro
5'-.....CCAGCTTTCCGAGCCAGATGAGG.....-3'
PAM
Sox2
SE
129
CAST
14kb
100 kb
Mir290-295
4kb
27 kb
RGM - eGFP
eGFP
1.5kb
D
SOX2-SE-TG
MIR290-SE-TG
RGM-tdTomato(129)
RGM-eGFP(CAST)
1.19
90.4
4.79
3.58
2.64
88.7
3.97
4.73
E
Nested F
F
Sox2 SE DMR
Snrpn
tdTomato
eGFP
Snrpn R
716bp
857bp
tdTomato Nested R
eGFP Nested R
Sox2 SE
chr3:34759903-34760335
Snrpn-tdTomato(RGM)
Snrpn-eGFP(RGM)
Mir290 SE
chr7:3200974-3201547
T+G-
T-G+
T+G+
T-G-

as double-negative (T^-G^-) cells in both cell lines, though the majority of cells were double-positive (T^+G^+) (Figure 1D), consistent with the heterogeneity reported in scWGBS data by others (Song et al. [1] Figure S1C) and in our BS-PCR analysis on both targeted and wild-type alleles (Figure 1A, 1B, and Song et al. [1] Figure S1D). To confirm that the RGM reporter activity faithfully reflected the allele-specific endogenous DNA methylation state, we sorted the four populations and performed allele-specific BS-PCR followed by Sanger sequencing of the DMRs upstream of the reporters. Figure 1E shows that the reporter activities on both alleles were consistent with the DNA methylation levels of the genomic SE regions and the inserted RGMs in all sorted populations. Quantitative pyro-sequencing further confirmed that T^+G^+ and T^-G^- populations represent two extreme methylation states of the intrinsic epigenetic heterogeneity at both SEs (Song et al. [1] Figure S1F). As expected, both unmethylated alleles in sorted T^+G^+ cells from both cell lines gained methylation synchronously upon retinoic acid (RA)-induced differentiation. This confirms that the RGM-targeted SEs undergo the predicted methylation changes when exiting pluripotency (Song et al. [1] Figure S1G).

1.2.2 Dynamic Allele-Specific SE DNA Methylation Is Regulated by *De Novo* Methylation and Passive Demethylation during Cell Proliferation

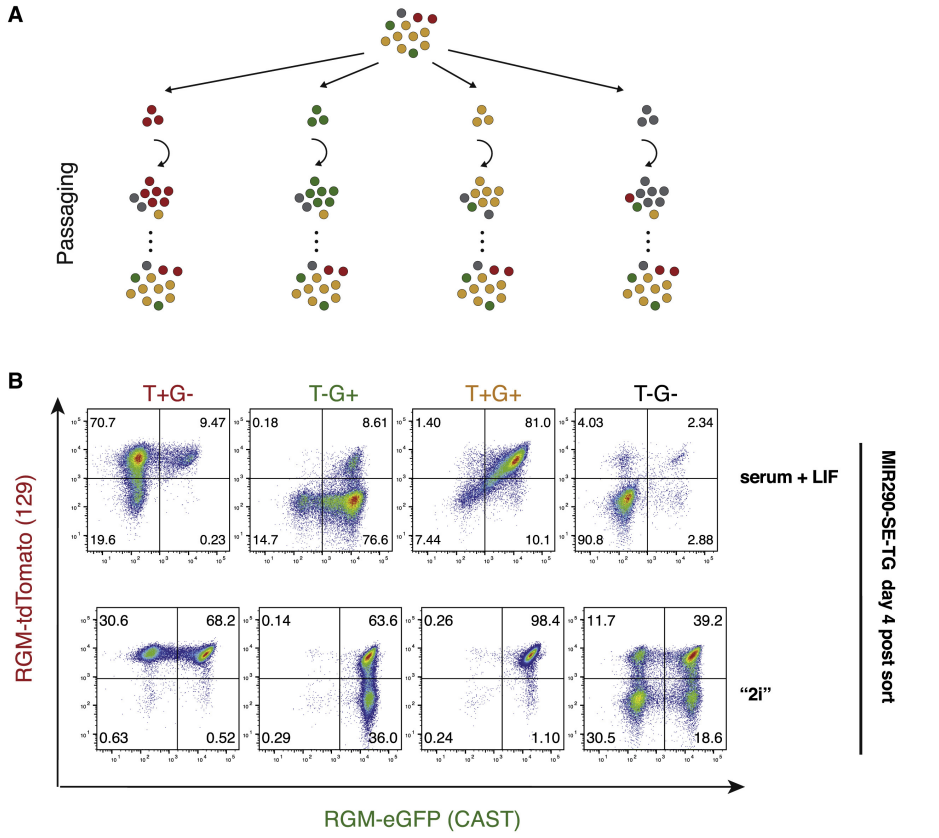
To gain insights into the origin of DNA methylation heterogeneity, we FACS sorted equal numbers of the four populations from both reporter cell lines and monitored the RGM activity upon passaging in serum + LIF medium (Figure 2A). Figure 2B (serum + LIF) and Song et al. [1] Figure S2A show that the SE DNA methylation states in the four sorted populations were not stable but highly dynamic with each allele independently switching the RGM on-and-off over the course of only a few days. This indicates that the observed SE DNA methylation heterogeneity is a result of fast dynamic and reversible switching of allelic DNA methylation states. When sorted cells were passaged and cultured in "2i" (GSKi and MAPKi) medium, the kinetics of the transitions between different methylation states was significantly altered with slowed *de novo* methylation for both SEs and an initial acceleration of demethylation at the *Mir290* SE (Figure 2B and 2C; Song et al. [1] Figure S2B). Demethylation of T^-G^- population of SOX2-SE-TG in "2i", however, is slower over the long term than that in serum + LIF, possibly due to impaired cell division as shown in the later part of this article. The observed DNA methylation difference between "2i" and serum + LIF is consistent with the extensive global demethylation induced in "2i" by downregulation of *de novo* and maintenance methyltransferases [32, 33, 34, 35, 36].

Demethylation in "2i" suggests that changes in DNA methyltransferase (DNMT) activities modulate the observed dynamics. To determine the main *de novo* methyltransferase driver for SE methylation, we compared RGM activities in *Dnmt3a* or *Dnmt3b* single-knockout and *Dnmt3a/3b* double-knockout (DKO) cells (Song et al. [1] Figure S3A). Although the number of RGM negative cells was reduced in *Dnmt3a* or *Dnmt3b* single-knockout cells, cells with

methylated SEs were eliminated only in the absence of both *de novo* methyltransferases in DKO cells preventing any *de novo* methylation (Figure 3A and Song et al. [1] Figure S3B). The hypomethylation of both SEs was further confirmed by pyro-sequencing in DKO ESCs as well as in cells induced to differentiate by RA (Song et al. [1] Figure S3C). These results suggest that both DNMT3A and DNMT3B have redundant functions and independently contribute to *de novo* methylation of SE DMRs.

DNA demethylation can occur either passively in rapidly dividing cells, caused by inhibition of DNMT1 or by active removal of the methyl group mediated by Tet enzymes and base excision repair (BER) pathways [37]. To assess whether demethylation of the SEs involved active or passive mechanisms, we analyzed whether DNA demethylation would be affected in cells upon delaying cell-cycle progression using thymidine block. In all three populations carrying at least one methylated allele, the kinetics of demethylation upon thymidine block was significantly decreased upon 3 days in culture (Figure 3B and 3C). This suggests that cell proliferation-driven passive demethylation is responsible for SE demethylation. To confirm this observation genetically, we transfected $129^{\text{SE-RGM-tdTomato}}$ $T^{-}G^{+}$ cells with Cas9 and sgRNAs against genes encoding the maintenance enzymes DNMT1/UHRF1, which upon down-regulation would lead to genome-wide passive dilution of methylation. In addition, we used sgRNAs against enzymes implicated in mediating active demethylation (*Tets*/*Tdg*/*Aid*). Figure 3D shows the predicted outcomes of $129^{\text{SE-RGM-tdTomato}}$ allele demethylation (changes of the fraction of $T^{+}G^{+}$ cells) after disruption of these genes. When *Dnmt1* or *Uhrf1* were disrupted, the $129^{\text{SE-RGM-tdTomato}}$ allele became demethylated in a substantial fraction of cells (Figure 3E). In contrast, transduction of sgRNAs against *Tet* enzymes, *Aid*, or *Tdg* had no substantial effect indicating that active demethylation is not significantly involved in SE demethylation. To confirm that the lack of methylation changes upon disruption of *Tets*, *Aid*, or *Tdg* was not due to inefficient Cas9-sgRNA transfection, we further compared the demethylation kinetics of the $129^{\text{SE-RGM-tdTomato}}$ allele in single clones harboring homozygous *Tdg* and *Aid* frameshift mutations (Song et al. [1] Figure S3D) with that of wild-type cells and observed no difference (Song et al. [1] Figure S3E). In addition, DNA methylation levels, as quantified by pyro-sequencing, did not reveal a significant difference among *Tet1*, *2*, and *3* single-knockout, *Tet1*, *2* double-knockout, *Tet1*, *2*, *3* triple-knockout ESCs, and the isogenic wild-type cells [38, 39, 40] (Song et al. [1] Figure S3C). Given the rapid proliferation of ESCs, our data are consistent with the notion that locus-specific DNA methylation at both SEs is subjected to intrinsically dynamic changes at the allelic level in each cell due to un-

Figure 2 *(following page).* SE DNA Methylation Heterogeneity Is Created by Dynamic Switching of Methylation States. (A) Experiment setup for monitoring SE DNA methylation dynamics. Yellow cells: T+G+; gray cells: T-G-; red cells: T+G-; green cells: T-G+. (B) FACS analyses on the dynamics of T+G-, T-G+, T+G+, and T-G- populations 4 days post-sorting for both *MIR290-SE-TG* in serum + LIF or "2i" medium. (C) Quantifications of the dynamics of 4 sorted populations from *MIR290-SE-TG* in percentages change over time when cultured in the serum + LIF or the "2i" medium after sorting. See also Song et al. [1] Figure S2.

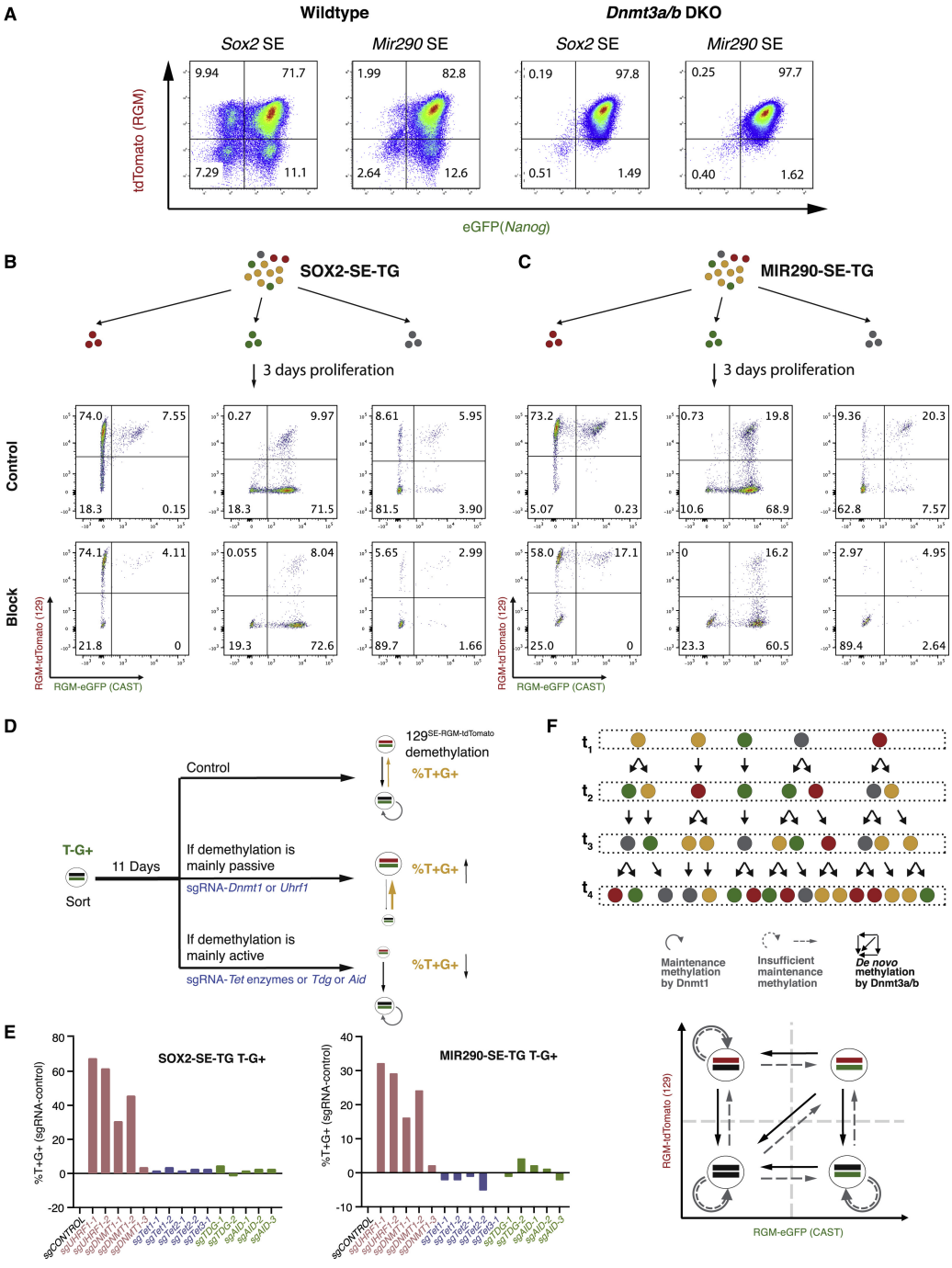

A
Passaging
B
T+G-
T-G+
T+G+
T-G-
70.7
9.47
19.6
0.23
0.18
8.61
14.7
76.6
1.40
81.0
7.44
10.1
4.03
2.34
90.8
2.88
serum + LIF
30.6
68.2
0.63
0.52
0.14
63.6
0.29
36.0
0.26
98.4
0.24
1.10
11.7
39.2
30.5
18.6
"2i"
MIR290-SE-TG day 4 post sort
RGM-tdTomato (129)
RGM-eGFP (CAST)
C
T-G+
T+G-
T-G-
T+G+
Percentage
100
50
0
Day 0
Day 4
Day 6
Day 8
Day 13
Day 11
serum + LIF
"2i"
MIR290-SE-TG

synchronized cell division and passive DNA demethylation, which leads to heterogeneous SE methylation at a snapshot sampling time (t1, …, t4, Figure 3F, top). The steady-state of such dynamic heterogeneity reflects a balance between *de novo* methylation dependent on both DNMT3A and DNMT3B and passive demethylation during rapid cell proliferation (Figure 3F, bottom).

1.2.3 TF Binding at SE Promotes Demethylation and Inhibits *De Novo* Methylation

To explore additional regulators of SE DNA methylation dynamics besides DNMTs activities and cell division, we investigated the impact of TF binding on the transition between DNA methylation states. Some TFs can serve as readers of DNA methylation or inducing changes to DNA methylation states upon binding to target sequences [41, 42, 43, 22]. The *Sox2* SE harbors multiple enrichment sites for the master TFs OCT4 and NANOG in ESCs [44] (Figure 4A, top). We deleted enrichment sites for the two TFs (peak 1 for NANOG and 2 for both NANOG and OCT4) at the *Sox2* SE DMR on either the $129^{\text{SE-RGM-tdTomato}}$ or the $\text{CAST}^{\text{SE-RGM-eGFP}}$ allele using sgRNAs against allele-specific SNPs (Figure 4A, bottom) and generated ESC clones harboring allele-specific peak deletions (ΔPeak 1-CAST, ΔPeak 2-CAST, and ΔPeak 2-129 clones; Song et al. [1] Figure S4A). We sorted the T^-G^- and T^+G^+ populations from these clones and monitored the re-establishment of allelic heterogeneity across deletion genotypes (Figure 4B). The fraction of T^+G^- or T^-G^+ cells transitioning from T^+G^+ or T^-G^- cells were quantified as allelic *de novo* methylation rates or demethylation rates, respectively (Figure 4C). We found that both the $129^{\text{SE-RGM-tdTomato}}$ and the $\text{CAST}^{\text{SE-RGM-eGFP}}$ allele exhibited a faster *de novo* methylation rate after deletion of its TF enrichment sites as compared to the intact wild-type allele (Figure 4D, top), indicating higher susceptibility to *de novo* methylation upon loss of TF binding. Similarly, the allele that had its TF enrichment site deleted showed a slower demethylation rate than the wild-type allele, indicating less resistance to maintenance methylation upon loss of TF binding (Figure 4D, bottom). To confirm that the observed RGM activity changes correspond to changes in DNA methylation, we performed

Figure 3 *(following page)*. The Dynamics of SE DNA Methylation Is Driven by De Novo Methylation and Passive Demethylation during Cell Proliferation. (A) Elimination of the population with methylated SEs in Dnmt3a and Dnmt3b DKO v6.5-*Nanog*-eGFP ESCs with the RGM-tdTomato reporter targeted mono-allelically at either the *Sox2* or the *Mir290* SE. (B and C) Demethylation of sorted T+G-, T-G+, and T-G- cells from (B) SOX2-SE-TG and (C) *MIR290-SE-TG* cells with and without thymidine block. (D) Expected changes in the T+G+ cell percentage for each demethylation mechanism upon CRISPR/Cas9-mediated gene disruptions. Changes in the percentage of T+G+ cells indicate the rate of demethylation on the 129SE-RGM-tdTomato allele. (E) Relative changes in the T+G+ percentage upon transfecting sgRNAs against enzymes involved in DNA demethylation, as compared to cells transfected with the same vector without sgRNA (sgControl). (F) A model for the origin of locus-specific DNA methylation heterogeneity. See also Song et al. [1] Figure S3.

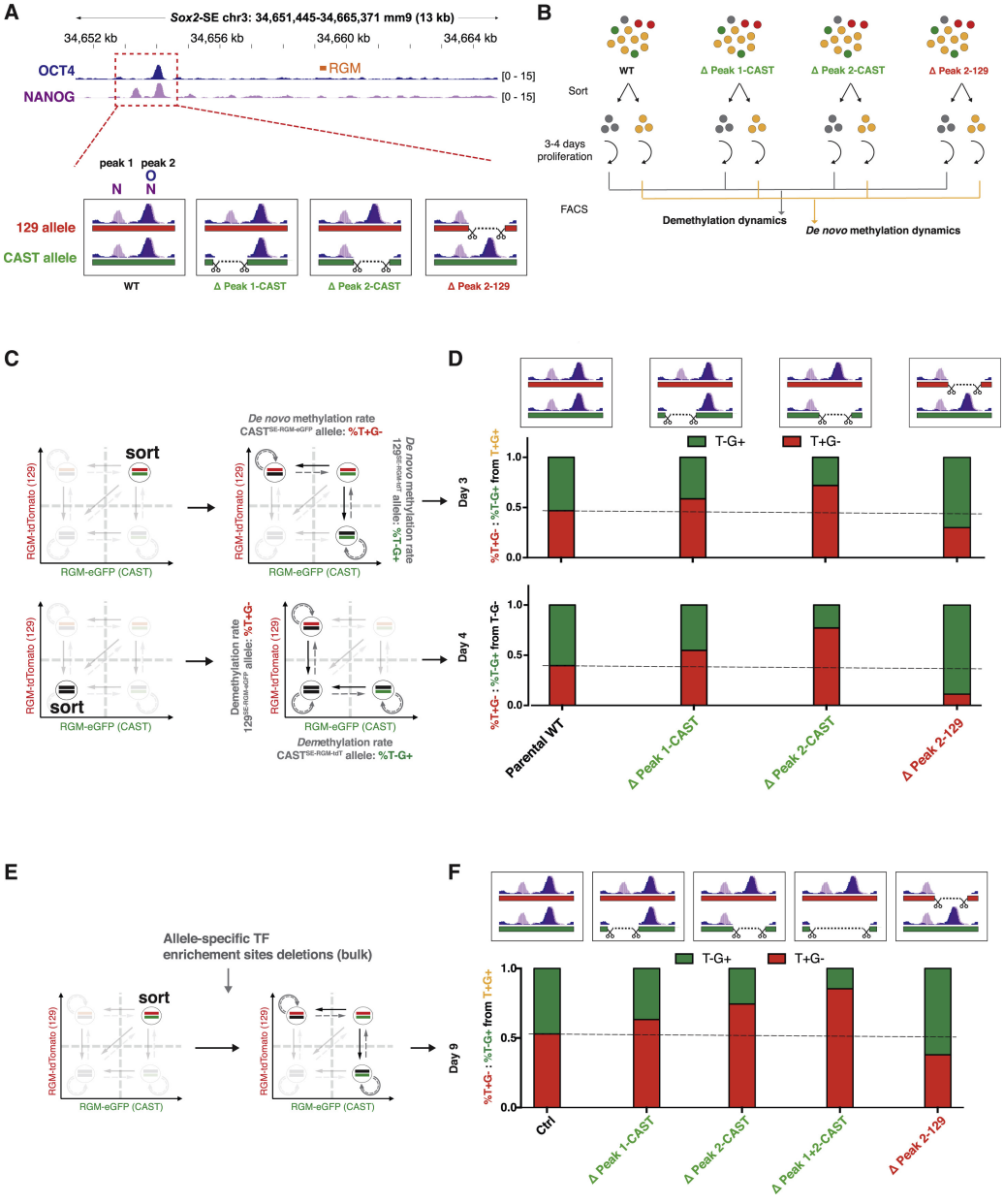

A
Wildtype
Dnmt3a/b DKO
Sox2 SE
Mir290 SE
Sox2 SE
Mir290 SE
9.94
71.7
7.29
11.1
1.99
82.8
2.64
12.6
0.19
97.8
0.51
1.49
0.25
97.7
0.40
1.62
tdTomato (RGM)
eGFP(Nanog)
B
SOX2-SE-TG
3 days proliferation
C
MIR290-SE-TG
3 days proliferation
Control
Block
RGM-tdTomato (129)
RGM-eGFP (CAST)
D
129SE-RGM-tdTomato demethylation
%T+G+
Control
T-G+
11 Days
Sort
If demethylation is mainly passive
sgRNA-Dnmt1 or Uhrf1
If demethylation is mainly active
sgRNA-Tet enzymes or Tdg or Aid
E
SOX2-SE-TG T-G+
MIR290-SE-TG T-G+
%T+G+ (sgRNA-control)
sgCONTROL
F
t1
t2
t3
t4
Maintenance methylation by Dnmt1
Insufficient maintenance methylation
De novo methylation by Dnmt3a/b
RGM-tdTomato (129)
RGM-eGFP (CAST)

BS-PCR followed by Sanger sequencing on sorted cells from ΔPeak 1-CAST and ΔPeak 2-129 clones. This analysis confirmed that the methylation status of the endogenous SE region was consistent with that of the *Snrpn* promoter as well as RGM activities at allelic resolution after genetic manipulation (Song et al. [1] Figure S4B). The TF binding effect on methylation dynamics was seen not only in cloned cells but also in sorted T^+G^+ cell population transfected with allele-specific sgRNAs against TF enrichment sites (Figure 4E). Consistent with the single-cell clone analyses, the allele with TF enrichment site deletion showed a faster *de novo* methylation rates than the wild-type allele that was not targeted by the sgRNAs (Figure 4F).

1.2.4 DNA Methylation Decreases MED1 Association with SE, Enhancer-Promoter H3K27ac, and *in cis* Transcription of the Target Genes

We investigated whether the rapid changes in SE DNA methylation would dynamically affect target gene transcription. Promoter DNA methylation has long been associated with stable silencing of gene expression [45, 46, 47, 48]; in comparison, enhancer methylation's role in transcription is less well characterized. The Mediator complex has been shown to be dynamically involved in phase-separated condensates concentrating at SEs for transcription of key cell-identity genes [49]. Since SE DNA methylation is dynamically changing, we investigated whether different allelic methylation states affect association of MED1 condensates with the *Mir290* SE. We performed DNA FISH at the *Mir290* SE locus and MED1 immunostaining on sorted cell populations. Figure 5A and Song et al. [1] Figure S5A show that MED1 was not enriched at the methylated *Mir290* SE as T^-G^- cell populations did not have DNA FISH foci that overlapped with MED1 enrichment as compared to cells in which at least one *Mir290* SE was

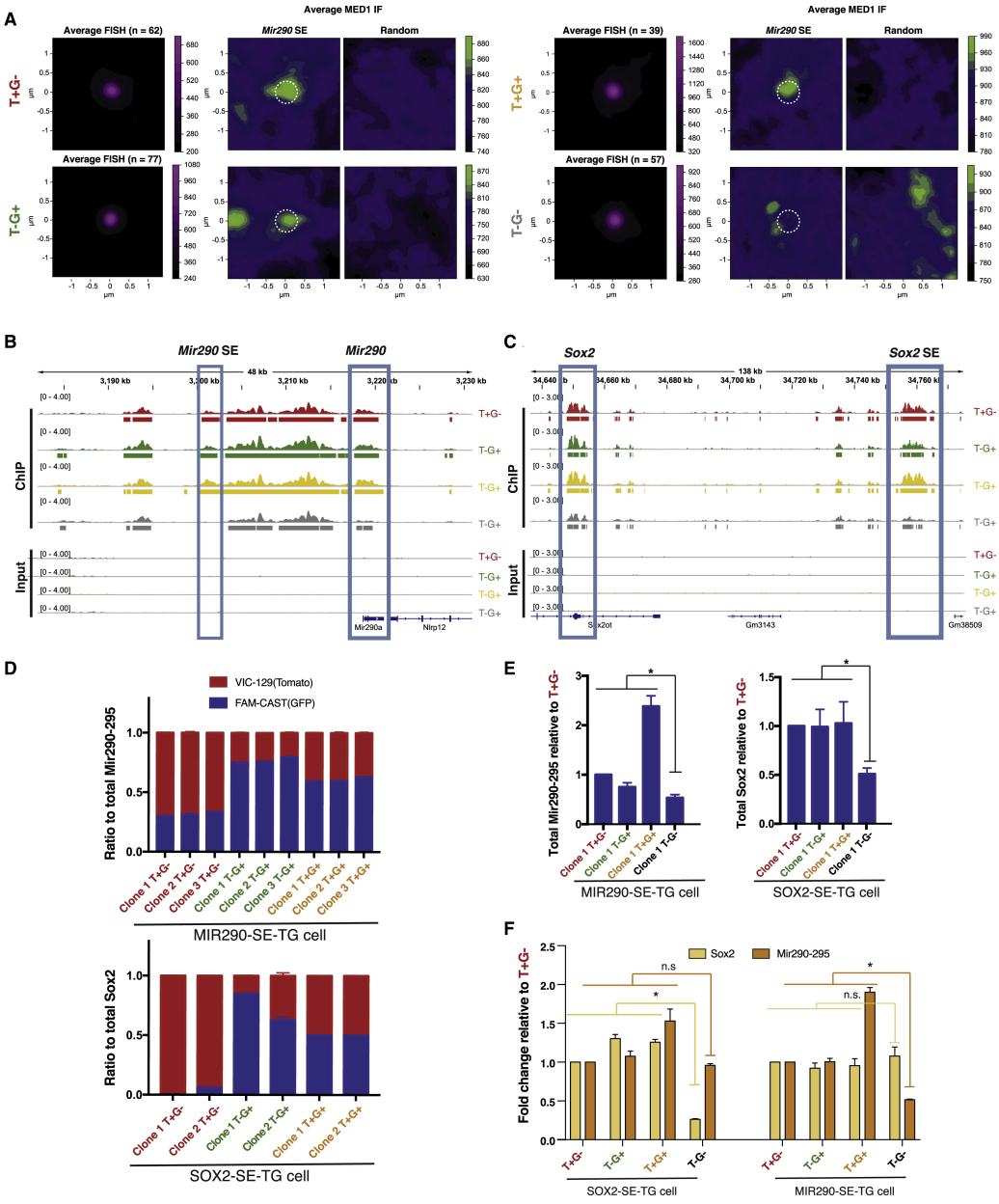
Figure 4 *(following page).* TF Binding at SEs Promotes Demethylation and Inhibits De Novo Methylation. (A) Top, schematic representation of TF enrichment sites (based on the ChIP-seq data of NANOG [pink track, peak 1 and 2] and OCT4 [blue track, peak 2]. ENCODE: ENCSR779CZG and ENCSR392DGA) relative to the RGM targeted site (orange). Bottom, allele-specific deletions of individual peaks after overlapping NANOG (N) and OCT4 (O) ChIP tracks. Red: $129^{\text{SE-RGM-tdTomato}}$ allele, green: $\text{CAST}^{\text{SE-RGM-eGFP}}$ allele. Scissors illustrate sgRNA targeting sites. ChIP-seq value is presented as fold-change-over-control. (B) Experimental setup using cells with different allelic TF enrichment site deletions in assessing the impacts of TF binding on SE methylation dynamics. (C) Top, T+G+ cells were sorted from the genotyped single-cell clones with allelic TF enrichment site deletions. Bottom, T-G0 cells were sorted from the same clones. (D) Quantification of allele-specific de novo methylation rates (top panels, T+G- or T-G+ cells derived from T+G+ cells) and demethylation rates (bottom panels, T+G- or T-G+ cells derived from T-G- cells) of the respective ESC clones compared to that of an unmodified parental wild-type clone (dotted line level). (E) Bulk T+G+ cells were sorted from the SOX2-SE-TG cell line and transfected with allele-specific sgRNA pairs to delete TF enrichment sites or with empty vectors. (F) Quantification of allele-specific de novo methylation rates of the bulk cells transfected with different sgRNAs. See also Song et al. [1] Figure S4.


A
Sox2-SE chr3: 34,651,445-34,665,371 mm9 (13 kb)
34,652 kb
34,656 kb
34,660 kb
34,664 kb
OCT4
NANOG
RGM
[0 - 15]
[0 - 15]
peak 1 peak 2
N O N
129 allele
CAST allele
WT
Δ Peak 1-CAST
Δ Peak 2-CAST
Δ Peak 2-129
B
WT
Δ Peak 1-CAST
Δ Peak 2-CAST
Δ Peak 2-129
Sort
3-4 days proliferation
FACS
Demethylation dynamics
De novo methylation dynamics
C
sort
RGM-tdTomato (129)
RGM-eGFP (CAST)
De novo methylation rate
CAST SE-RGM-eGFP allele: %T+G-
129 SE-RGM-tdT allele: %T+G+
Demethylation rate
129 SE-RGM-eGFP allele: %T+G-
Demethylation rate
CAST SE-RGM-tdT allele: %T-G+
D
Day 3
Day 4
T-G+
T+G-
%T+G- : %T-G+ from T+G-
%T+G- : %T-G+ from T-G-
1.0
0.5
0.0
Parental WT
Δ Peak 1-CAST
Δ Peak 2-CAST
Δ Peak 2-129
E
Allele-specific TF enrichement sites deletions (bulk)
sort
RGM-tdTomato (129)
RGM-eGFP (CAST)
F
Day 9
T-G+
T+G-
%T+G- : %T-G+ from T+G-
1.0
0.5
0.0
Ctrl
Δ Peak 1-CAST
Δ Peak 2-CAST
Δ Peak 1+2-CAST
Δ Peak 2-129

unmethylated. Since the Mediator complex interacts with both the SE and the promoter [50], a loss of MED1 enrichment upon SE DNA methylation may affect promoter activity as well. We therefore performed H3K27ac chromatin immunoprecipitation sequencing (ChIP-seq) as a proxy epigenetic mark defining active enhancers and promoters on four sorted populations from both reporter cell lines. H3K27ac was significantly reduced at both methylated SE regions, as measured by total (Figure 5B and 5C, *Sox2* SE and *Mir290* SE boxes; Song et al. [1] Figure S5B, enhancer panels) as well as allele-specific H3K27ac enrichment (Song et al. [1] Figure S5C, enhancer panels). As expected, a decrease in H3K27ac was also observed at promoters residing on the same chromosome with the methylated SE (Figure 5B and 5C, *Sox2* and *Mir290* boxes, and Song et al. [1] Figures S5B and S5C, promoter panels) but not at adjacent regions (Song et al. [1] Figure S5B, adjacent regions panels). This demonstrates that SE methylation affects the promoter H3K27ac level, likely through a loss of enhancer-promoter communication.

To test whether synchronized H3K27ac changes upon transient DNA methylation at enhancers and promoters affects *in cis* target gene expression, we performed allele-specific qRT-PCR on the four sorted cell populations from both reporter cell lines. As shown in Figure 5D, methylation of either allele of the SEs resulted in decreased target gene expression on the same chromosome. However, the *Sox2* SE and the *Mir290* SE have different effects on the total expression level of their respective target genes. The suppressive effect of transient DNA methylation was independent and additive when either *Mir290* SE allele was methylated (Figure 5E, left). In contrast, total *Sox2* expression only significantly decreased when both *Sox2* SE alleles were methylated (Figure 5E, right), and in single-positive cells only single-

Figure 5 *(following page).* DNA Methylation Decreases MED1 Association at SE, Enhancer-Promoter H3K27ac and *in cis* Transcription of the Target Genes. (A) Averaged DNA FISH (Magenta, *Mir290* SE) and co-immunofluorescence staining (Green, MED1) signal in the nuclei of *MIR290-SE-TG* cells sorted based on allelic methylation states. Random spots were selected in the same image away from the DNA FISH spots. (B) Peak calling from H3K27ac ChIP-seq of 4 sorted populations from *MIR290-SE-TG*. *Mir290* SE and *Mir290-295* cluster are boxed in blue. Peak values are normalized using RPKM (reads per million) with a 10-bp bin size. (C) Peak calling from H3K27ac ChIP-seq of 4 sorted populations from SOX2-SE-TG. *Sox2* SE and *Sox2* gene are boxed in blue. Peak values are normalized using RPKM (reads per million) with a 10-bp bin size. (D) Allele-specific expression of *Mir290-295* pri-miRNA (top) and *Sox2* mRNA (bottom) in 3 sorted populations, with VIC-TaqMan probe detecting the 129SE-RGM-tdTomato allele, and FAM-TaqMan probe detecting the $CAST^{SE\text{-}RGM\text{-}eGFP}$ allele in both SE cases. Independently targeted clones for each SE were used as biological replica. Data are represented as mean ± SD. (E) Fold change of total *Mir290-295* pri-miRNA (left) and total *Sox2* mRNA (right) from the 4 sorted populations normalized to that of the T+G- population. Independently targeted clones for each SE were used as biological replica. Data are represented as mean ± SD. (F) Quantification of *Mir290-295* expression on sorted SOX2-SE-TG cells compare to *Sox2* expression (left) and quantification of *Sox2* expression on sorted *MIR290-SE-TG* cells compare to *Mir290-295* expression (right). Data are represented as mean ± SD. See also Song et al. [1] Figure S5.

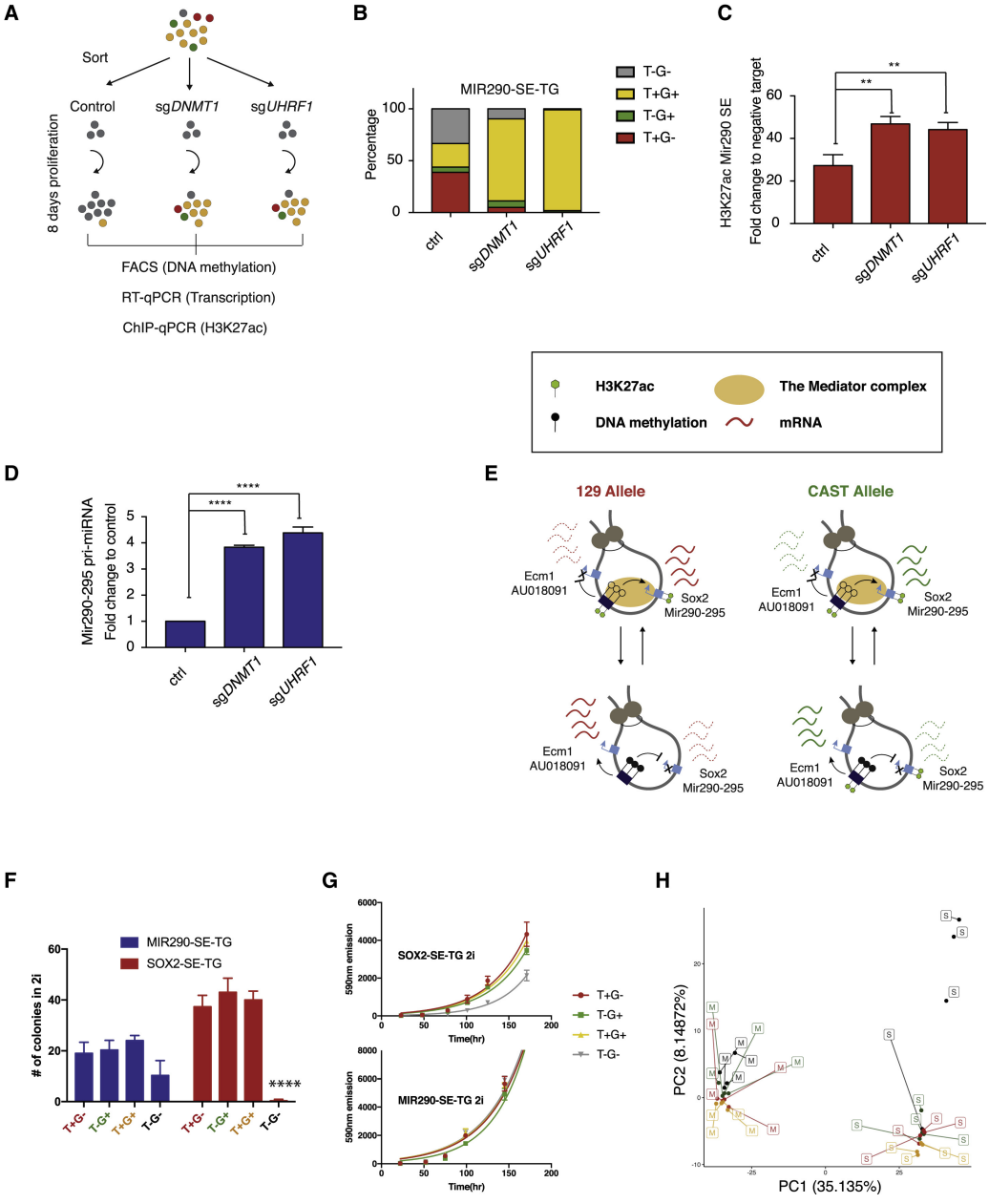

A
Average MED1 IF
Average FISH (n = 62)
Mir290 SE
Random
Average FISH (n = 39)
Average FISH (n = 77)
Average FISH (n = 57)
T+G-
T+G+
T-G+
T-G-
B
Mir290 SE
Mir290
ChIP
Input
Mir290a
Nlrp12
C
Sox2
Sox2 SE
Sox2ot
Gm3143
Gm38509
D
VIC-129(Tomato)
FAM-CAST(GFP)
Ratio to total Mir290-295
MIR290-SE-TG cell
Ratio to total Sox2
SOX2-SE-TG cell
E
Total Mir290-295 relative to T+G-
Total Sox2 relative to T+G-
MIR290-SE-TG cell
SOX2-SE-TG cell
F
Sox2
Mir290-295
Fold change relative to T+G-
n.s
SOX2-SE-TG cell
MIR290-SE-TG cell

molecule RNA FISH (smFISH) could detect a slight decrease of *Sox2* transcripts (Song et al. [1] Figures S5D and S5E), indicating a compensating mechanism on total *Sox2* transcripts when one SE allele is methylated. Notably, DNA methylation at two SEs exclusively anti-correlated with their respective *in cis* target genes, and little difference is seen in *Mir290-295* expression if cells were sorted based on the methylation state at the *Sox2* SE locus and vice versa (Figure 5F). This indicates that the DNA methylation state of the two SEs switches independently of each other.

To determine whether SE methylation has a causal role in suppressing enhancer-promoter H3K27ac and transcription, we transfected Cas9-sgRNAs targeting *Dnmt1* and *Uhrf1* and removed DNA methylation in sorted T^-G^- MIR290-SE-TG cells to induce rapid passive demethylation (Figure 6A). Figure 6B shows that cells deficient for *Dnmt1* or *Uhrf1* displayed significantly faster demethylation resulting in a higher proportion of T^+G^+ cells as compared to the control. Both acetylation of H3K27 at the SE (Figure 6C) and *Mir290-295* expression (Figure 6D) were significantly increased upon *Dnmt1*/*Uhrf1* disruptions, as measured by ChIP-qPCR and qRT-PCR from the same cultures, respectively. This suggests that change in DNA methylation directly regulates SE function and transcription *in cis*.

Since correlating abundance in RNA allele-specific SNPs with allele-specific RGM activities allows distinguishing direct targets regulated *in cis* by the SE methylation status versus expression changes caused by secondary effects, we searched additional genomic targets on the same chromosomes that are directly regulated by SE methylation by allele-specific RNA sequencing (RNA-seq) analysis on sorted populations. We quantified allele-specific expression of genes on chromosome 3 (for MIR290-SE-TG) and chromosome 7 (for SOX2-SE-TG) in single positive cells and calculated the ratio between expressions from the allele with an unmethylated SE over that of the other allele with a methylated SE. We plotted this ratio of each gene calculated in T^-G^+ cells as the x-axis value and the ratio calculated in T^+G^- cells as the y-axis value (Song et al. [1] Figure S6A). As expected, *Mir290-295* and *Sox2* both appeared in the

Figure 6 ***(following page).*** **DNA Methylation Directly Suppresses SE Activity and Affects ESC State.** (A) Experimental setup for assessing the causal role of SE DNA methylation suppresses H3K27ac. FACS (DNA methylation), RT-qPCR (*Mir290-295*), and ChIP-qPCR (H3K27ac) were co-assessed from the same pool of cells from each sample. (B) Loss of DNA methylation in *MIR290-SE-TG* T-G- cells 8 days post-Dnmt1 and Uhrf1 sgRNA transfection as compared to controls. (C) H3K27ac ChIP-qPCR at the *Mir290* SE from the experimental groups in (B), respectively. Data are represented as mean ± SD. (D) *Mir290-295* pri-miRNA level from the experimental groups in (B). Data are represented as mean ± SD. (E) Summary of the dynamic regulation and functional impact of allelic SE methylation. (F) Colony formation assays in "2i" starting from 100 sorted cells. Data are represented as mean ± SD. (G) Growth curves measured by AlamarBlue Cell Viability Reagent. Data are represented as mean ± SD. (H) Principal-component analysis of the top 5% highly variable genes from different populations of SOX2-SE-TG (Labeled as S. red: T+G-, green, T-G+, black: T-G-, yellow: T+G+) and *MIR290-SE-TG* (labeled as M; color code same as S). See also Song et al. [1] Figures S5-S7.


A
Sort
Control
sgDNMT1
sgUHRF1
8 days proliferation
FACS (DNA methylation)
RT-qPCR (Transcription)
ChIP-qPCR (H3K27ac)
B
MIR290-SE-TG
Percentage
T-G-
T+G+
T-G+
T+G-
ctrl
sgDNMT1
sgUHRF1
C
H3K27ac Mir290 SE
Fold change to negative target
ctrl
sgDNMT1
sgUHRF1
H3K27ac
The Mediator complex
DNA methylation
mRNA
D
Mir290-295 pri-miRNA
Fold change to control
ctrl
sgDNMT1
sgUHRF1
E
129 Allele
CAST Allele
Ecm1
AU018091
Sox2
Mir290-295
F
MIR290-SE-TG
SOX2-SE-TG
# of colonies in 2i
T+G-
T-G+
T+G+
T-G-
G
SOX2-SE-TG 2i
MIR290-SE-TG 2i
590nm emission
Time(hr)
H
PC2 (8.14872%)
PC1 (35.135%)

upper right corner as they were *in cis* directly suppressed by allelic SE methylation. Surprisingly, two antisense transcripts relative to *Sox2* and *Mir290-295*, *Ecm1* and *AU018091*, respectively, were oppositely regulated by allele-specific *Sox2* or *Mir290* SE methylation: SE hypermethylation strongly correlated with upregulations of both anti-sense transcripts, whereas SE hypo-methylation correlated with inhibition (Song et al. [1] Figures S6B and S6C). This result shows that direct transcriptional targets of SE methylation are highly specific with possibly opposite effects on some *cis*-regulated genes. Though the detailed mechanism of such regulation remains to be elucidated, *Ecm1* was upregulated in *Sox2* SE deletion cells [51].

Our results suggest that DNA methylation at both SEs fluctuates independently and dynamically, altering Mediator complex condensates at the SE and allelic H3K27ac at enhancers and promoters *in cis* and ultimately leading to heterogeneous allelic transcription of the target genes (Figure 6E).

1.2.5 *Sox2* and *Mir290* SE Methylation Heterogeneities Have Different Biological Impacts on ESC State

Culture in "2i" medium has been shown to only allow naïve pluripotent cells to proliferate [52]. Long-term culture of MIR290-SE-TG and SOX2-SE-TG cells in "2i" after passaging from serum + LIF media, though favoring T^+G^+ population decreased but did not abolish heterogeneity completely (Song et al. [1] Figure S6D). The persistence of all four populations in both reporter cell lines indicates that DNA methylation at both SEs have different degrees of heterogeneity in different culture conditions. Both *Sox2* and *Mir290-295* are highly expressed in ESCs [53, 44, 54, 55], raising the possibility that allelic transcriptional heterogeneity caused by SE methylation heterogeneity may lead to co-existing heterogeneous cellular states of ESCs. In "2i" media, SOX2-SE-TG T^-G^- cells exhibited significantly impaired colony-forming ability and proliferation (Figure 6F and 6G). However, under the same condition, the heterogeneous DNA methylation at the *Mir290* SE did not lead to any obvious changes of ESCs, despite the slight colony formation disadvantage of MIR290-SE-TG T^-G^- cells (Figure 6F and 6G). We further explored the functional differences among populations *in vivo* by injecting sorted cells to form teratomas. Surprisingly, despite the significant growth disadvantage of SOX2-SE-TG T^-G^- population, they were able to contribute to all three germ layers in teratoma formation assays with no obvious contribution bias towards any germ layer compared to SOX2-SE-TG T^+G^+, MIR290-SE-TG T^+G^+, and T^-G^- cells (Song et al. [1] Figure S6E). This indicates that ESCs with biallelic methylation at the *Sox2* SE are still pluripotent. However, when examined at the molecular level, these cells were distinct from other populations in principal-component analysis on difference in the 5% most highly variably expressed genes (Figure 6H) and 17,000 uniquely distinct H3K27ac enrichment peaks in ChIP-seq (Song et al. [1] Figure S6F). GO analysis on RNA-seq revealed that the SOX2-SE-TG T^-G^- population preferentially expressed genes in differentiation-related pathways as compared to the MIR290-SE-TG T^-G^- population (Song et al. [1] Figure S7A). The epigenetic and transcriptional differences of SOX2-SE-TG T^-G^- cells indicate that these cells downregulate *Sox2* expression and

are prone to differentiate but not as yet committed to a certain fate. Our results are consistent with the notion that pluripotent ESC are heterogeneous as reflected by the dynamic allelic DNA methylation of key pluripotency SE.

1.2.6 DNA Methylation Is Dynamic at Both SEs in Blastocysts while Exhibiting Spatial-Temporal Differences in Pre-implantation Embryos

In vivo, both *Sox2* and *Mir290-295* are expressed in preimplantation embryos. As reported previously *Sox2* expression increases between the morula and the blastocyst stage [56] and *Mir290-295* expression significantly upregulates at the 4-cell stage [57]. To investigate changes in DNA methylation of the two SEs at single-cell and allelic resolution, we generated transgenic mice homozygous for the $129^{\text{SE-RGM-tdTomato}}$ allele or the $\text{CAST}^{\text{SE-RGM-eGFP}}$ allele and obtained 2–4 cell embryos carrying one $129^{\text{SE-RGM-tdTomato}}$ allele and one $\text{CAST}^{\text{SE-RGM-eGFP}}$ allele by mating animals homozygous for RGM-eGFP or RGM-tdTomato (Figure 7A). The two SEs gained allelic DNA methylation heterogeneity at different times: reporter activity became apparent as early as the 4-cell stage for the *Mir290* SE but only at the morula stage for the *Sox2* SE (Figure 7B). At the blastocyst stage, *Sox* expression was restricted to the inner cell mass (ICM), whereas the *Mir290-295* displayed broad expression in both ICM and trophectoderm (TE) [52, 58, 59]. Heterogeneous SE DNA methylation was consistent with the established spatial expression pattern of the two genes in blastocysts (Figure 7C). We further investigated whether the observed methylation heterogeneity was due to dynamic allelic methylation state switching *in vivo*. We sorted the four populations from SOX2-SE-TG and MIR290-SE-TG ESCs, injected each population into 8-cell stage wild-type CD1-IGS host embryos, and cultured embryos for 2 days to monitor *de novo* methylation or demethylation at single-cell resolution (Song et al. [1] Figure S7B). A long-term membrane bound dye (Cy5) was used to track the injected cells (Figure 7D; Song et al. [1] Figure S7C). Figure 7D (SOX2-SE-TG cells) and Song et al. [1] Figure S7C (MIR290-SE-TG cells) show that, at the blastocyst stage, injected T^-G^- cells demethylated the SE as they turned on the RGMs on either or both alleles and became single positive or T^+G^+ cells (T^-G^- columns, white arrows). Demethylation also was observed in injected single-positive cells as the originally methylated allele at injection became unmethylated and cells became T^+G^+ (T^+G^- and T^-G^+ columns, white arrows). Similarly, dynamic *de novo* methylation was observed *in vivo*, as injected T^+G^+ cells shut down RGM activities on either or both alleles (T^+G^+ columns, yellow arrows) and single-positive cells became T^-G^- cells (T^-G^+ and T^-G^+ columns, yellow arrows).

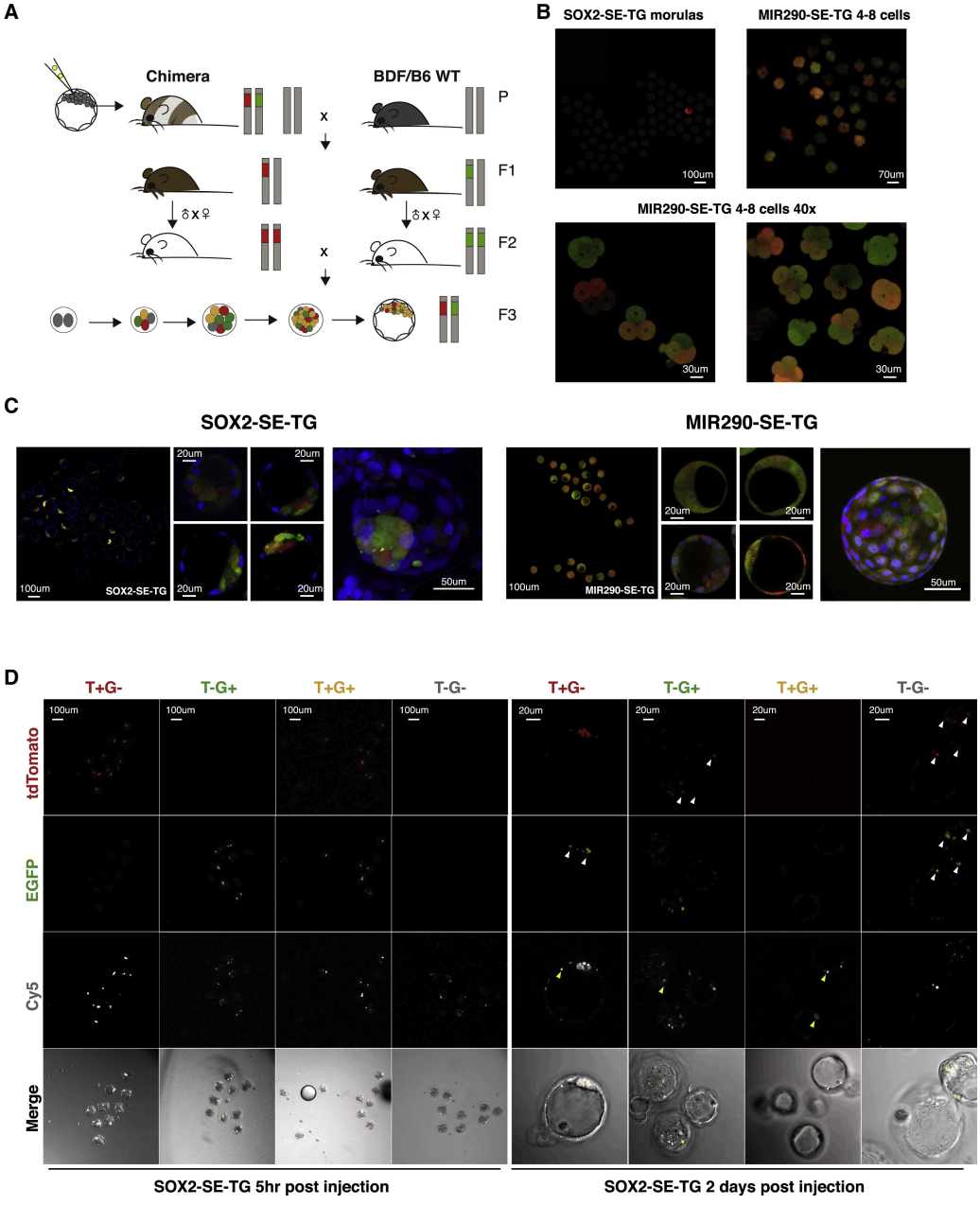
In summary, our data indicate that dynamic DNA methylation exists at active SEs in early preimplantation embryos creating locus-specific epigenetic heterogeneity, recapitulating and extending our observations in ESCs *in vitro*.

1.3 Discussion

The importance of DNA methylation regulation at *cis*-regulatory elements is increasingly recognized as many developmental- and disease-associated DMRs overlap with these regions [60, 61, 62]. Locus-specific DNA methylation heterogeneity across cells has been shown by recent scWGBS as a potential explanation for the variable low-to-intermediate levels of methylation at active enhancers in bulk measurements. The present work was based on an experimental paradigm that overcomes some of the limitations of single-cell sequencing approaches using an allele-specific reporter system. This allowed us to address questions that were not resolved by previously used sequencing-based methods. (1) Our study shows that in ESCs the methylation state of the two alleles of the *Sox2* and *Mir290* SEs change dynamically and independently of each other. (2) We demonstrate that the dynamic change of SE DNA methylation is driven by the balance between three DNMTs and cell proliferation, with TF binding promoting the hypomethylated state. (3) We show that DNA methylation dynamically regulates target genes *in cis* and inhibits formation of Mediator complex condensates at the SE as well as enhancer-promoter H3K27 acetylation. (4) Allelic variation of SE DNA methylation, reflecting the epigenetic heterogeneity of ESCs, can originate from cells of different transcriptional landscapes and proliferative potentials as for the *Sox2* SE or of developmentally identical states as for the *Mir290* SE. (5) Finally, we show that dynamic DNA methylation is not only seen in cultured ESCs but also in preimplantation embryos.

Allele-specific RGM reporters targeted to the endogenous *Sox2* and the *Mir290* SEs allowed us to trace DNA methylation both *in vitro* and *in vivo*. Detailed analyses showed that the low levels of DNA methylation of the *Sox2* and the *Mir290* SE are due to the presence of a small fraction of cells with hypermethylated SE alleles. The methylation heterogeneity in these cells results from highly dynamic and reversible switching between allelic DNA methylation states. Because the RGM reporter allowed isolation of cells with defined allele-specific SE DNA methylation states, we were able to demonstrate that dynamic changes in SE

Figure 7 *(following page).* DNA Methylation Is Dynamic at Both SEs in Blastocysts while Exhibiting Spatial-Temporal Differences in Pre-implantation Embryos. (A) Mating scheme for generating SOX2-SE-TG and MIR290-SE-TG mice and heterozygous preimplantation embryos genetically carry 129SE-RGM-tdTomato and CASTSE-RGM-eGFP at the Sox2 or the Mir290 SE for imaging analyses. (B) Live 4-8 cell (MIR290-SE-TG) and morula stage (SOX2-SE-TG) embryos. (C) Live E3.5-E4.5 blastocysts of SOX2-SE-TG and MIR290-SE-TG in 10X low magnification, 40X high magnification, and 3D projections (left to right in each group). Red: tdTomato, green: eGFP, blue: Hoechst 33342. (D) Tracking Sox2 SE DNA methylation dynamics in vivo. Columns are sorted and injected populations and rows are different imaging channels. Red: RGM-tdTomato; Green: RGM-eGFP; Cy5: Qtracker 705 was used to label and track injected cells. White arrows indicate demethylation, and yellow arrows indicate de novo methylation, at 2 days post-injection compare to 5 h post-injection. Channels were adjusted for brightness and contrast for optimal visibility. See also Song et al. [1] Figure S7.


A
Chimera
BDF/B6 WT
P
F1
F2
F3
♂x♀
B
SOX2-SE-TG morulas
MIR290-SE-TG 4-8 cells
MIR290-SE-TG 4-8 cells 40x
100um
70um
30um
C
SOX2-SE-TG
MIR290-SE-TG
20um
50um
D
T+G-
T-G+
T+G+
T-G-
tdTomato
EGFP
Cy5
Merge
SOX2-SE-TG 5hr post injection
SOX2-SE-TG 2 days post injection

DNA methylation are tightly anti-correlated *in cis* with enhancer-promoter H3K27ac levels. This is likely due to disruption of enhancer-promoter interactions consistent with the Mediator complex condensates showing decreased association at the methylated *Mir290* SE. The Mediator complex and its unit MED1 have been shown previously to form condensates with liquid-like properties, which allows dynamic interactions with TFs and the transcription apparatus [63, 49]. Our study shows that DNA methylation can affect these transcriptional condensates. Given the dynamic state switching of allelic SE DNA methylation as well as the dynamic nature of MED1 condensate formation, it is highly likely that one process mediates the other. We also show that SE DNA methylation can have opposing effects on transcription of different genes located on the same chromosome: the direct target genes *Sox2* and *Mir290-295* were repressed, while the antisense genes *Ecm1* and *AU018091* were activated by SE methylation. By removing DNA methylation at the *Mir290 SE* through *Dnmt1*/*Uhrf1* deletion, we showed that changes in SE DNA methylation is a dynamic process actively regulating its transcriptional activity. By enabling sorting for a particular epigenetic state and combined with allelic expression analyses, we demonstrate that dynamic DNA methylation serves as an epigenetic basis for allelic heterogeneity in gene expression and that dynamic DNA methylation at SEs is a likely mechanism for dynamic random monoallelic transcription seen in mammalian cells [64, 65]. However, it warrants further exploration to establish the causal link between allelic epigenetic and transcriptional heterogeneity *in vivo*.

While *Sox2* and *Mir290* SE methylation affect target gene expression similarly, we detected some differences on cellular growth and differentiation. Cells with biallelically methylated *Sox2* SE revealed impaired growth and upregulation of differentiation-related pathways (Figure 6F and 6G; Song et al. [1] Figure S7A). In contrast, *Mir290* SE methylation had little effects on cell state. We identified additional differences of how DNA methylation suppresses activity of the two SE. *Mir290-295* expression was independently suppressed by methylation at either *Mir290* SE DMR allele consistent with the observation that individual DMR constituents have independent activities [66]. In contrast, monoallelic *Sox2* SE methylation did not significantly affect the overall *Sox2* expression, suggesting additional regulatory mechanisms.

The experimental platform described here allows rapid tracing and isolating rare cell populations based on their transient methylation signatures at specific loci and thus can provide mechanistic insights into the nature of enhancer DNA methylation in heterogeneous cell populations both *in vivo* and *in vitro* in real time, which is difficult in sequencing-based approaches. Furthermore, this system enables manipulation of different molecular components to define interactions and hierarchies between layers of epigenetic regulation in dynamic systems with rapid changes. Our study provides a path towards the mechanistic understanding of dynamic T-DMR regulation in heterogeneous tissues and complex biological processes, such as development and diseases [67, 68].

1.4 STAR Methods

1.4.1 Key Resources Table

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Rabbit poly-clonal anti-H3K27ac	Abcam	Cat#ab4729; RRID:http://antibodyregistry.org/AB_2118291
Rabbit poly-clonal anti-MED1	Abcam	Cat#ab64965; RRID:http://antibodyregistry.org/AB_1142301
Chemicals, Peptides, and Recombinant Proteins		
Puromycin	Sigma Aldrich	Cat#P7255
Thymidine	Sigma Aldrich	Cat#T1895
Retinoic acid	Sigma Aldrich	Cat#R2625
LIF recombinant protein	House-made	N/A
MAPK inhibitor PD0325901	Stemgent	Cat#04-0006-10
GSK-3β inhibitor CHIR99021	Stemgent	Cat#04-0004-10
Critical Commercial Assays		
Zymoclean Gel DNA Recovery Kit	Zymo Research	Cat#D4002
X-tremeGENE 9 DNA Transfection Reagent	Sigma Aldrich	Cat#6365809001
Xfect ESC Transfection Reagent	Clontech	Cat#631320
AlamarBlue Cell Viability Reagent	Bio-Rad	BUF012A
NEBNext®Ultra™DNA Library Prep Kit for Illumina	NEB	Cat#E7370S
NEBNext®Multiplex Oligos for Illumina®	NEB	Cat#E7335S
KAPA mRNA HyperPrep Kit	Roche	Cat# 08098115702
TrueSeq Stranded PolyA prep	Illumina	Cat# 20020595
Accel-NGS 2S PCR-Free Library Kit (96 rxns)	Swift Biosciences	Cat#20096
Qtracker™705 Cell Labeling Kit	Thermo Fisher	Cat# Q25061MP
TaqMan Assay, primer information see table S3	Sigma	N/A
Direct-zol RNA Miniprep	Zymo Research	Cat#R2050
SuperScript III First-Strand Synthesis SuperMix	Life Technologies	Cat#18080400
Fast SYBR Green Master Mix	Life Technologies	Cat#4385618
ProlongGold	Life Technologie	Cat#P36930
Prime-It II Random Primer Labeling Kit	Agilent Technologies	Cat#300385

Continued on next page

Continued from previous page

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited Data		
Raw and processed data	This paper	GEO: GSE132416 (subseries: GEO: GSE132376, geo: GSE132404, and geo: GSE132414)
Flowcytometry data, Sanger sequencing trace files, raw images, gels, algorithm scripts	This paper	https://doi.org/10.17632/6vbc6htfnf.1
Experimental Models: Cell Lines		
Mouse: SOX2-SE-TG ESCs	This paper	N/A
Mouse: MIR290-SE-TG ESCs	This paper	N/A
Mouse: Sox2-SE-NanogRGM ESCs	[25]	N/A
Mouse: miR290-SE-NanogRGM ESCs	[25]	N/A
Mouse: SOX2-SE-TG Δpeak 1-CAST ESCs	This paper	N/A
Mouse: SOX2-SE-TG Δpeak 2-CAST ESCs	This paper	N/A
Mouse: SOX2-SE-TG Δpeak 2-129 ESCs	This paper	N/A
Tet1-/- #19	[38]	N/A
Tet1-/- #34	[38]	N/A
Tet2 -/- KO	[39]	N/A
Tet1-/- Tet2-/- #26	[39]	N/A
Tet1-/- Tet2-/- #51	[39]	N/A
Tet1-/- Tet2-/- Tet3 -/- #26	[40]	N/A
Tet1-/- Tet2-/- Tet3 -/- #29	[40]	N/A
Experimental Models: Organisms/Strains		
Mouse: SOX2-SE-TT:	This paper	N/A
Mouse: SOX2-SE-GG	This paper	N/A
Mouse: MIR290-SE-TT: CAST/EiJ /129/BDF1/C57BL/6	This paper	N/A
Mouse: MIR290-SE-GG: CAST/EiJ /129/BDF1/C57BL/6	This paper	N/A
Mouse: NSG NOD.Cg-PrkdcscidIl2rgtm1Wjl/SzJ	Jackson Laboratory	005557
Mouse: CD1®IGS	Charles River	022
Oligonucleotides		
sgRNA for targeting and knockout, see table S1	This paper	N/A
Primers for bisulfite PCR and pyro-sequencing, see Table S2	This paper	N/A
Primers for TaqMan Assay and qRT-PCR, see table S3	This paper	N/A
Primers for ChIP-qPCR, see table S4		N/A

Continued on next page

Continued from previous page

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Primers for reporter ESC line, KO cell line, and mouse genotyping, see table S5	This paper	N/A
Stellaris®DesignReady FISH Probes, Sox2	LGC Biosearch Technologies	Cat# VSMF-3075-5-BS
DNA FISH probe for miR290 SE	[49]	N/A
Recombinant DNA		
miR290-SE-RGM-tdTomato targeting vector	[25]	N/A
miR290-SE-RGM-eGFP targeting vector	This paper	N/A
Sox2-SE-RGM-tdTomato targeting vector	[25]	N/A
Sox2-SE-RGM-eGFP targeting vector	This paper	N/A
pTurbo-Cre	GenBank	AF334827
px330-BFP-sgRNA	[25]	N/A
Software and Algorithms		
MACS2 (ChIP-seq algorithms)	[69]	https://github.com/taoliu/MACS/wiki
Samtools	[70]	http://samtools.sourceforge.net/
BWA	[71]	http://bio-bwa.sourceforge.net/
deepTools 3.0.2	[72]	http://deeptools.readthedocs.io/en/
STAR (v.2.5.3.a)	[73]	https://github.com/alexdobin/STAR
DESeq2 (v1.18.1)	[74]	https://bioconductor.org/packages/release/bioc/html/DESeq2.html
SNPsplit (v0.3.2)	Babraham Bioinformatics	https://www.bioinformatics.babraham.ac.uk/projects/SNPsplit/
RSEM (v1.2.31)	N/A	https://deweylab.github.io/RSEM/
PANTHER	[75, 76]	http://pantherdb.org/
Image J	[77])	https://imagej.net/
FlowJo	N/A	https://www.flowjo.com/
PyroMark Q48 Autoprep	QIAGEN	http://www.qiagen.com/us/
Python scripts	This paper; Mendeley Data	https://doi.org/10.17632/6vbc6htfnf.1
MATLAB scripts	[78]	N/A
IGV	Broad Institute	https://software.broadinstitute.org/software/igv
Bismark v0.21.0	Babraham Bioinformatics	https://www.bioinformatics.babraham.ac.uk/projects/bismark

1.4.2 Lead Contact and Materials Availability

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Rudolf Jaenisch (jaenisch@wi.mit.edu).

1.4.3 Experimental Model and Subject Details

ESC cell lines

ESC cell culture and proliferation assays All cells were cultured at 37°C with 5% CO_2. 129xCAST or v6.5 mouse male ESCs were cultured on irradiated mouse embryonic fibroblasts (MEFs) with standard ESCs medium: (500 ml) DMEM supplemented with 10% FBS (Hyclone), 10 mg recombinant leukemia inhibitory factor (LIF), 0.1 mM beta-mercaptoethanol (Sigma-Aldrich), penicillin/streptomycin, 1 mM L-glutamine, and 1% nonessential amino acids (all from Invitrogen). For experiments in 2i culture conditions, ESCs were cultured on gelatin-coated plates with N2B27 + 2i + LIF medium containing: (500 ml), 240 ml DMEM/F12 (Invitrogen; 11320), 240 ml Neurobasal media (Invitrogen; 21103), 5 ml N2 supplement (Invitrogen; 17502048), 10 ml B27 supplement (Invitrogen; 17504044), 10 mg recombinant LIF, 0.1 mM beta-mercaptoethanol (Sigma-Aldrich), penicillin/streptomycin, 1 mM L-glutamine, and 1% nonessential amino acids (all from Invitrogen), 50 mg/ml BSA (Sigma), PD0325901 (Stemgent, 1 mM), and CHIR99021 (Stemgent, 3 mM). For measuring cell proliferation, AlamarBlue Cell Viability Reagent (Bio-Rad, BUF012A) was added to cell culture and incubated at 37°C with 5% CO_2 and emission at 590nm was monitored every 50hrs. At each sampling time point, relative changes in cell numbers were compared to 0hr after sorting.

Generating biallelically targeted reporter cell lines To generate SOX2-SE-TG and MIR290-SE-TG reporter cell lines, targeting vectors (Mir290-SE-RGM-tdTomato, Mir290-SE-RGM-eGFP, Sox2-SE-RGM-tdTomato, Sox2-SE-RGM-eGFP), and CRISPR/Cas9 were transfected into ESCs using Xfect ESC Transfection Reagent (Clontech, Cat#631320), according to the provider's protocol. Forty-eight hours following transfection, cells were selected for puromycin resistance (Sigma Aldrich, Cat#P7255) and plated on MEF feeder plates. Single colonies were further analyzed for proper and single integration by Southern blot and Junction PCR analysis. PGK-Puromycin resistance cassette were looped out by overexpression of Cre recombinase (pTurbo-Cre, GenBank accession number AF334827) and followed by Southern blot validation.

ESCs with CRISPR-Cas9-mediated deletions Tet-enzyme single-, double- and triple knockouts were generated and described previously [38, 39, 40]. sgRNA sequences are cloned into *px330-BFP* vector under U6 promoter. *px330-BFP-sgRNA* vectors were transfected into pre-plated ESC cells using Xfect ESC Transfection Reagent, according to the provider's protocol. For analysis in populations, cells were sorted for BFP 48 hours post-transfection and cultured on MEF feeder plates. For single clone analysis, cells were genotyped using Southern blot or TA cloning of PCR products of CRISPR targeting site from each allele followed by sequencing. For *Dnmt3a* and *Dnmt3b*, *Aid* and *Tdg* single knockouts, single clones with frame-shifting indels were selected for further analysis; for TF binding

site deletions, single clones have allele-specific entire peak site deletions were selected for further analysis. For TF enrichment site deletion experiments, sgRNA pairs for generating deletion are transfected as following: Δpeak 1-CAST: sgTFBS-*Sox2*-SE-1(CAST) and sgTFBS-*Sox2*-SE-2(CAST); Δpeak 2-CAST: sgTFBS-*Sox2*-SE-2(CAST) and sgTFBS-*Sox2*-SE-3(CAST); Δpeak 2-129: sgTFBS-*Sox2*-SE-2(129) and sgTFBS-*Sox2*-SE-3(Both); Δ Peak 1+2-CAST: sgTFBS-*Sox2*-SE-1(CAST) and sgTFBS-*Sox2*-SE-3(CAST). All sgRNA sequences are listed in Song et al. [1] Table S1.

Animals

Blastocyst injections and generation of reporter mice Blastocyst injections were performed using (C57BL/6xDBA) B6D2F1 (Charles River) or CD1 (Charles River) host embryos. In brief, 6-7-week old B6D2F1 females were hormone primed by an intraperitoneal (i.p.) injection of pregnant mare serum gonadotropin (PMS, EMD Millipore) followed 46 hr later by an injection of human chorionic gonadotropin (hCG, VWR). Embryos were harvested at the morula stage and cultured in a CO2 incubator overnight. To obtain tetraploid (4n) blastocysts, electrofusion was performed at approximately 44–47 h post hCG using a BEX LF-301 cell fusion device (Protech International Inc., Boerne, TX). On the day of the injection, groups of embryos were placed in drops of M2 medium using a 16-um diameter injection pipet (CytoSpring). Approximately ten cells were injected into the blastocoel cavity of each embryo using a Piezo micromanipulator (Prime Tech). Approximately 20 blastocysts were subsequently transferred to each recipient female; the day of injection was considered as 2.5 days postcoitum (DPC). Male chimera mice were mated to C57BL/6 females and the ones that gave birth to agouti pups (F1) have germ-line transmitted CASTX129 ESC. Mice were handled in accordance with institutional guidelines and approved by the Committee on Animal Care (CAC) and Department of Comparative Medicine (DCM) of Massachusetts Institute of Technology.

Mouse mating scheme and genotyping All mouse F1 mice heterozygous for either the SE-RGM-tdTomato (abbreviated as SOX2-SE-T0 or MIR290-SE-T0) or the SE-RGM-eGFP allele (abbreviated as SOX2-SE-G0 or MIR290-SE-G0) were obtained by mating germ-line transmitted chimeras to C57BL/6 females. F2 mice homozygous for either SE-RGM-tdTomato (abbreviated as SOX2-SE-TT or MIR290-SE-TT) or the SE-RGM-eGFP allele (abbreviated as SOX2-SE-GG or MIR290-SE-GG) were generated by inbreeding (SOX2-SE-T0 x SOX2-SE-T0, SOX2-SE-G0 x SOX2-SE-G0, MIR290-SE-T0 x MIR290-SE-T0, MIR290-SE-G0 x MIR290-SE-G0). Mice are genotyped by PCR the 5'junction of the SE RGM: SOX2-SE-F (or MIR290-SE-F) with tdTomato-R for the RGM-tdTomato allele, SOX2-SE-F(or SOX2-SE-F) with eGFP-R for the RGM-eGFP allele, and SOX2-SE-F (or MIR290-SE-F) with SOX2-SE-R (or MIR290-SE-R) for the wild-type allele (Song et al. [1] Table S5).

Confocal imaging of live pre-implantation embryos 2-cell embryos were obtained from mating SOX2-SE-TT or MIR290-SE-TT females hormone primed step-wise with PMS and hCG to SOX2-SE-GG or MIR290-SE-GG males, respectively, or the opposite mating strategy (SOX2-SE-GG or MIR290-SE-GG females to SOX2-SE-TT or MIR290-SE-TT males, respectively). 2-cell embryos were flushed out from the oval-duct by M2 media with BSA (CytoSpring # m2113) 48hrs post mating. The embryos were then cultured in 25-50μl KSOM media droplets (CytoSpring # KO102) covered by mineral oil in a 37°C 5% CO_2 incubator. Embryos will become blastocysts at E3.5. For monitoring methylation dynamics *in vivo* , ESCs were cultured in serum + LIF, pre-plated and sorted based on RGM activity before injection. 2-3 cells were injected into 8-cell stage CD1 host embryos and cultured in M2 media with BSA at 37°C in 5% CO_2. Images were taken by a Zeiss LSM 710

Laser Scanning Confocal microscope. Images were taken using either 10x or 40x water lenses and saved in LSM format. Channels for eGFP (excitation 488nm), tdTomato (excitation 594nm), Cy5 (excitation 633nm), and Hoechst 33342 (excitation 405nm) were merged into image composites.

Teratoma formation assays and H&E staining 0.5-1 million sorted ESCs in serum + LIF media were 1:1 mixed with Matrigel and injected subcutaneously into the femur on both sides of the NSG mice. Tumors were taken when reaching 1cm in diameter and mice euthanized. Mice were handled in accordance with institutional guidelines and approved by the Committee on Animal Care (CAC) and Department of Comparative Medicine (DCM) of Massachusetts Institute of Technology. Tissues were dissected and fixed in 10% formalin overnight. Tissues were embedded in paraffin, sectioned, and stained for H&E.

1.4.4 Method Details

Southern blots

Genomic DNA (10–15 mg) was digested with appropriate restriction enzymes overnight. Subsequently, genomic DNA was separated on a 0.8% agarose gel, transferred to a nylon membrane (Amersham) and hybridized with ^{32}P probe labeled by Prime-It II Random Primer Labeling Kit (Agilent Technologies, Cat#300385).

Flow cytometry

To assess the proportion of eGFP and tdTomato in the established reporter cell lines, a single-cell suspension was filtered and assessed on the BD Aria or FACSCanto II. Compensation was achieved by using cells with either tdTomato or eGFP fluorescence. Fsc files were analyzed by FlowJo.

Bisulfite conversion-PCR (BS-PCR) and pyro-sequencing

Bisulfite conversion of genomic DNA, nested PCR, and sequencing was established as described previously [25]. Pyro-seq of all bisulfite converted genomic DNA samples were performed with PyroMark Q48 Autoprep (QIAGEN) according to the manufacturer's instructions. Primers used for BS-PCR and pyro-sequencing are listed in Song et al. [1] Table S2.

Retinoic acid differentiation

ESCs carrying the reporter for both *Mir290* and *Sox2* SE regions were sorted for *Nanog*-eGFP positive and RGM-tdTomato positive and plated on gelatin-coated plates in ESC medium (+LIF). The next day, cells were washed with PBS, re-suspended in basal N2B27 medium (2i medium without LIF, insulin, and the two inhibitors), and supplemented with 0.25μM retinoic acid (RA, Sigma Aldrich, Cat#R2625-50MG). Medium was replaced every other day.

Double thymidine block

10-20k cells/per well were plated onto 12-well plates after sorting with media containing 2.5mM thymidine for 12hrs. Blocking was released by washing twice with PBS and culturing in serum + LIF mouse ES media for 9hrs. Cells were then again blocked with 2.5mM thymidine for 14hrs and FACS analyses were done 6hrs post release.

1.4.5 Quantification and Statistical Analysis

qRT-PCR and TaqMan assays

Total mRNA was extracted from ESCs using Direct-zol RNA Miniprep (Zymo Research, Cat#R2050) after pre-plating for elimination of MEF feeders, treated with DNase A defined amount of mRNA reverse-transcribed into cDNA using SuperScript III First-Strand Synthesis SuperMix (Life Technologies, Cat#18080400) using random hexamers. Total expression of transcripts were quantified by qRT-PCR using Fast SYBR Green Master Mix (Life Technologies), and allele-specific transcripts are quantified by TaqMan Assay customized probes (Sigma, Song et al. [1] Table S3) targeting *Sox2* and *Mir290-295* pri-mRNA SNPs. Tukey's multiple comparison (****P<0.0001, **P<0.01, *P<0.05). Both qRT-PCR and TaqMan assays used at least 2 independently targeted clones as biological replica. The probes and context sequences are listed in Song et al. [1] Table S3.

ChIP-qPCR

ChIP was done on 2-5 million cells of each same-culture-sorted population from both reporter cell lines as described previously [79], 2μg of anti-H3K27ac antibody (abcam ab4729) was used for precipitation. Eluted DNA was quantified using real-time qPCR with Fast SYBR

Green Master Mix. Each ChIP-qPCR was repeated 3 times. Enrichment was calculated using as percentage of input. Statistical differences between samples are calculated with two-way ANOVA (alpha = 0.05), followed by Tukey's multiple comparison (****P<0.0001, **P<0.01). Primers used for detecting positive and negative control sequences, and SE targets are listed in Song et al. [1] Table S4.

H3K27ac ChIP-Seq and analysis

ChIP samples of 4 same-culture-sorted populations from SOX2-SE-TG and MIR290-SE-TG, respectively, were validated for positive and negative targets using qPCR. Libraries of Input-ChIP pairs were prepared with Accel-NGS 2S PCR-Free Library Kit (Cat#20096) and sequenced using Illumina HiSeq 2500. Raw reads were aligned to the reference genome mm10 using BWA using default parameters. Peak calling was done using MACS2. Peak intensities at SE, promoter and in-between regions are quantified and compared using bamCompare – deepTools 3.0.2 with FPKM from 10bp genomic bins of each sample. SNPs specific to 129 or CAST genomes at SE and promoters were counted from mapped raw reads and SNPs covered by more than 3 reads are accepted for quantification. Coordinates for analysis (mm10): *Mir290*-SE: chr7:3198900-3202780, *Mir290*-promoter: chr7:3215340-3221110; *Sox2*-SE: chr3:34752523-34766449, *Sox2*-promoter chr3:34649995-34652460.

RNA-Seq and analysis

For each reporter cell line, 2 independently targeted clones are independently sorted twice, generating 2 biological replica x 2 experimental replica = 4 replica in total. Stranded mRNA libraries were prepared using KAPA HyperPrep (SOX2-SE-TG) and TrueSeq Stranded PolyA prep (MIR290-SE-TG). mRNA libraries were sequenced on Illumina HiSeq 2500. Allele-specific RNA expression was quantified with a custom pipeline. In short, raw fastq files are aligned to a consensus genome using STAR (v.2.5.3.a). The reference transcriptome includes the *Mir290-295* pri-miRNA or *Sox2* and the RGMs on pseudo-chromosomes. After alignment SNPsplit (v0.3.2) splits the reads into four files based on single nucleotide variations (SNV). The reads were either allele specific (for CAST or S129), unassigned (if there are no SNVs present) or conflicting (if the SNVs in the read are from both alleles). The split read files were quantified using RSEM (v1.2.31) separately for each sample. Raw counts were then normalized to library using DESeq2 (v1.18.1) for each split. To obtain sample-level quantifications raw counts were summed over the splits before normalization. Differential expression analysis (DEA) was performed using DESeq2 (v1.18.1) at the level of samples. Samples were corrected for genetic clone and batch effect. GO analyses were performed using PANTHER. All expressed genes in the respective cell lines were used as the reference backgrounds. All P-values were controlled for false discovery rate (Benjamin-Hochberg procedure).

RNA smFISH and image analyses

Cells were fixed for 15 min with 4% PFA at room temperature and subsequently permeabilized in 70% EtOH overnight. Custom designed smFISH probes for Sox2 labeled with Quasar 670 (Stellaris®DesignReady FISH Probes, Cat# VSMF-3075-5-BS) were incubated with the samples for 16 hours at 30°C in hybridization buffer (100 mg/mL dextran sulfate, 25% formamide, 2X SSC, 1 mg/mL E.coli tRNA, 1 mM vanadyl ribonucleoside complex, 0.25 mg/mL BSA). Samples were washed twice for 30 min at 30°C with wash buffer (25% formamide, 2X SSC) containing DAPI (1 μg/mL, Sigma D9542). All solutions were prepared with RNAse-free water. Finally, the sections were mounted using ProlongGold (Life Technologies, P36930) and imaged two days later. Mounted samples were imaged on a Nikon Ti-Eclipse epifluorescence microscope equipped with an Andor iXON Ultra 888 EMCCD camera, using a 100X /1.45 Plan Apo Lambda oil objective (Nikon) and dedicated, custom-made fluorescence filter sets (Nikon). z-stacks with a distance of 0.3 μm between planes were collected. The number of Sox2 (mRNA) signals per cell was quantified using home-made MATLAB scripts.

DNA FISH, Med1 IF and average image analyses

DNA FISH of the *Mir290* SE and IF of MED1 were done as previously described [49]. For analysis of RNA/DNA FISH with immunofluorescence, custom Python scripts were written to process and analyze 3D image data gathered in FISH and IF channels. Nuclear stains were blurred with a Gaussian filter (sigma = 2.0), maximally projected in the z plane, and clustered into 2 clusters (nuclei and background) by K-means. FISH foci were either manually called with ImageJ or automatically called using the scipy ndimage package. For automatic detection, an intensity threshold (mean + 3* standard deviation) was applied to the FISH channel. The ndimage find_objects function was then used to call contiguous FISH foci in 3D. These FISH foci were then filtered by various criteria, including size (minimum 100 voxels), circularity of a max z-projection (circularity = 4* areaperimeter2 ; 0.7), and being present in a nucleus (determined by nuclear mask described above). For manual calling, FISH foci were identified in maximum z-projections of the FISH channel, and the x and y coordinates were used as reference points to guide the automatic detection described above. The FISH foci were then centered in a 3D-box (length size $i = 3.0$ μm). The IF signal centered at FISH foci for each FISH and IF pair are then combined and an average intensity projection is calculated, providing averaged data for IF signal intensity within a i x i square centered at FISH foci. As a control, this same process was carried out for IF signal centered at an equal number of randomly selected nuclear positions. These average intensity projections were then used to generate 2D contour maps of the signal intensity. Contour plots are generated using the matplotlib python package. For the contour plots, the intensity-color ranges presented were customized across a linear range of colors ($n! = 15$). For the FISH channel, black to magenta was used. For the IF channel, we used chroma.js (an online color generator) to generate colors across 15 bins, with the key transition colors chosen as black, blueviolet, medium-blue,

lime. This was done to ensure that the reader's eye could more readily detect the contrast in signal. The generated colormap was employed to 15 evenly spaced intensity bins for all IF plots. The averaged IF centered at FISH or at randomly selected nuclear locations are plotted using the same color scale, set to include the minimum and maximum signal from each plot.

High-throughput sequencing of bisulfite PCR

PCR amplicons were sonicated using Covarius into 150-200bp range. NEBNext®Ultra™DNA Library Prep Kit for Illumina and NEBNext®Multiplex Oligos for Illumina®were used to construct libraries according to manufacturer's protocol. Single barcoded library was prepared from sonicated bisulfite PCR amplicon fragments of the *Mir290* SE wildtype-allele using NEBNext®Ultra™DNA Library Prep Kit for Illumina (NEB #E7370S) and NEBNext®Multiplex Oligos for Illumina®(Index Primers Set 1, NEB #E7335S). Libraries were sequenced with 40bp single reads, adapter trimmed, aligned and analyzed with Bismark v0.21.0 (bismark -nondirectional). CpGs with >1000 coverage were counted to generate average percentage of methylation. Methylation percentage and its standard error were estimated as described in [20], and number of methylated counts was assumed to be a binomial random variable.

1.4.6 Data and Code Availability

Description: https://doi.org/10.17632/6vbc6htfnf.1

The raw confocal, gel and film images and original fsc files have been deposited at Mendeley Data

Description: Raw and processed high-throughput sequencing data have been deposited at NCBI Gene Expression Omnibus under ID code GEO: GSE132416 (subseries: GEO: GSE132376 for H3K27ac ChIP-seq, GEO: GSE132404 for BS-seq, and GEO: GSE132414 for RNA-seq).

1.5 Acknowledgments

We thank George Bell, Prathapan Thiru, and Bingbing Yuan for their help in ChIP-seq analysis and BS-PCR sequencing analysis; Ruth Flannery and Dina Rooney for their help with animal husbandry, injections of the ESCs, and harvesting pre-implantation embryos; and Dongdong Fu for sectioning and processing of teratoma samples. We would like to thank Tom Volkert, Sumeet Gupta, Kevin Truong, Amanda Chilaka, and Jennifer Love of the Whitehead Genome Technology Core for their help in ChIP-seq; Wendy Salmon of the W.M. Keck Microscopy Facility for help with confocal microscopy; Glenn Paradis, Patti Wisniewski, Patrick Autissier, Michael Jennings, Michele Griffin, Mervelina Saturno-Condon, Hanna Aharonov, and Eleanor Kincaid of the Whitehead Institute and MIT flow cytometry facilities for their

help with cell sorting. We thank Dr. Roderick Bronson and Kathleen Cormier at the KI Swanson Biotechnology Center Histology Core for teratoma sample consultation. We thank Raaji Alagappan, Tenzin Lungjangwa, and Carrie Garrett-Engele for their technical support. We thank Alicia V. Zamudio from Young Lab, Jian Shu, Shawn Liu, Haiting Ma, Emile Wogram, and all of the members of the Jaenisch lab for helpful discussions. Y.S. was supported by HFSP long-term fellowship, ISF grant no. 1610/18 and is the incumbent of the Louis and Ida Rich Career Development Chair. R.J. was supported by NIH grants HD 045022, 1U19AI131135, 5R01MH104610, and 1R01GM114864.

1.5.1 Author Contributions

Y. Song, Y. Stelzer, and R.J. conceived the project. Y. Stelzer and R.J. designed and supervised the experiments, S.S., R.A.Y., and R.J. acquired funding for this study. Y. Song conducted experiments, interpreted results, and wrote the manuscript with input from all authors. S.M., J.D., and N.R. conducted blastocyst injections. S.S. and P.R.v.d.B. performed re-analysis of the published scWGBS data, RNA-seq analysis, and smFISH. A.D. and J.E.H. assisted with DNA FISH, IF, and quantitative image analyses. F.S. assisted with cloning, targeting, and designing of the CRISPR knockout experiments and contributed instrumentally to the writing of the manuscript. M.A.C. assisted with teratoma injection.

1.5.2 Declaration of Interests

R.J. is a cofounder of Fate Therapeutics, Fulcrum Therapeutics, and Omega Therapeutics and an advisor to Dewpoint Therapeutics. R.A.Y. is a founder and shareholder of Syros Pharmaceuticals, Camp4 Therapeutics, Omega Therapeutics, and Dewpoint Therapeutics.

Acronyms

129xCAST 129xCasteneous

DKO double-knockout

DMR differential methylation region

DNMT DNA methyltransferase

ESC emrbyonic stem cell

PCR polymerase chain reaction

RGM Reporter of Genome Methylation

RNA-seq RNA sequencing

scWGBS single cell WGBS

SE super-enhancer

smFISH single-molecule RNA FISH

SNP single nucleotide polymorphism

T-DMR tissue-specific differential methylation region

TF transcription factor

WGBS whole-genome bisulfite sequencing

1.6 References

[1] Yuelin Song et al. “Dynamic Enhancer DNA Methylation as Basis for Transcriptional and Cellular Heterogeneity of ESCs”. In: *Molecular cell* 0.0 (2019), 905–920.e6. DOI: 10.1016/j.molcel.2019.06.045.

[2] Kenneth C Ehrlich et al. “DNA Hypomethylation in Intragenic and Intergenic Enhancer Chromatin of Muscle-Specific Genes Usually Correlates with their Expression.” In: *The Yale Journal of Biology and Medicine* 89.4 (2016), pp. 441–455.

[3] Thomas Fleischer et al. “DNA methylation at enhancers identifies distinct breast cancer lineages”. In: *Nature Communications* 8.1 (2017), pp. 1–14. DOI: 10.1038/s41467-017-00510-x.

[4] Benedetta Izzi et al. “Allele-specific DNA methylation reinforces PEAR1 enhancer activity”. In: *Blood* 128.7 (2016), pp. 1003–1012. DOI: 10.1182/blood-2015-11-682153.

[5] Peter A Jones. “Functions of DNA methylation: islands, start sites, gene bodies and beyond”. In: *Nature Publishing Group* 13.7 (2012), pp. 484–492. DOI: 10.1038/nrg3230.

[6] Lorenzo Rinaldi et al. “Dnmt3a and Dnmt3b Associate with Enhancers to Regulate Human Epidermal Stem Cell Homeostasis”. In: *Cell Stem Cell* 19.4 (2016), pp. 491–501. DOI: 10.1016/j.stem.2016.06.020.

[7] Chongyuan Luo et al. “Dynamic DNA methylation: In the right place at the right time”. In: *Science* 361.6409 (2018), pp. 1336–1340. DOI: 10.1126/science.aat6806.

[8] Michael B Stadler et al. “DNA-binding factors shape the mouse methylome at distal regulatory regions”. In: *Nature* 480.7378 (2011), pp. 490–495. DOI: 10.1038/nature10716.

[9] GiNell Elliott et al. “Intermediate DNA methylation is a conserved signature of genome regulation”. In: *Nature Communications* 6.1 (2015), pp. 1–10. DOI: 10.1038/ncomms7363.

[10] Holger Heyn et al. “Epigenomic analysis detects aberrant super-enhancer DNA methylation in human cancer”. In: *Genome Biology* 17.1 (2016), p. 1198. DOI: 10.1186/s13059-016-0879-2.

[11] Gary C Hon et al. “Epigenetic memory at embryonic enhancers identified in DNA methylation maps from adult mouse tissues”. In: *Nature Genetics* 45.10 (2013), pp. 1198–1206. DOI: 10.1038/ng.2746.

[12] Ruiwei Jiang et al. “Discordance of DNA Methylation Variance Between two Accessible Human Tissues”. In: *Scientific Reports* 5.1 (2015), pp. 1–8. DOI: 10.1038/srep08257.

[13] Andrew D King et al. “Reversible Regulation of Promoter and Enhancer Histone Landscape by DNA Methylation in Mouse Embryonic Stem Cells”. In: *Cell Reports* 17.1 (2016), pp. 289–302. DOI: 10.1016/j.celrep.2016.08.083.

[14] Austin Y Shull et al. *DNA Hypomethylation within B-Cell Enhancers and Super Enhancers Reveal a Dependency on Immune and Metabolic Mechanisms in Chronic Lymphocytic Leukemia*. Tech. rep. 2016.

[15] Lih Feng Cheow et al. "Multiplexed locus-specific analysis of DNA methylation in single cells". In: *Nature Protocols* 10.4 (2015), pp. 619–631. DOI: 10.1038/nprot.2015.041.

[16] H Guo et al. "Single-cell methylome landscapes of mouse embryonic stem cells and early embryos analyzed using reduced representation bisulfite sequencing". In: *Genome Research* 23.12 (2013), pp. 2126–2135. DOI: 10.1101/gr.161679.113.

[17] Hongshan Guo et al. "Profiling DNA methylome landscapes of mammalian cells with single-cell reduced-representation bisulfite sequencing". In: *Nature Protocols* 10.5 (2015), pp. 645–659. DOI: 10.1038/nprot.2015.039.

[18] Fan Guo et al. "Single-cell multi-omics sequencing of mouse early embryos and embryonic stem cells". In: *Cell Research* 27.8 (2017), pp. 967–988. DOI: 10.1038/cr.2017.82.

[19] Steffen Rulands et al. "Genome-Scale Oscillations in DNA Methylation during Exit from Pluripotency". In: *Cell Systems* 7.1 (2018), 63–76.e12. DOI: 10.1016/j.cels.2018.06.012.

[20] Sébastien A Smallwood et al. "Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity". In: *Nature Methods* 11.8 (2014), pp. 817–820. DOI: 10.1038/nmeth.3035.

[21] B Jin et al. "DNA Methylation: Superior or Subordinate in the Epigenetic Hierarchy?" In: *Genes & Cancer* 2.6 (2011), pp. 607–617. DOI: 10.1177/1947601910393957.

[22] Heng Zhu et al. "Transcription factors as readers and effectors of DNA methylation". In: *Nature Publishing Group* 17.9 (2016), pp. 551–565. DOI: 10.1038/nrg.2016.83.

[23] Anshul Kundaje et al. "Integrative analysis of 111 reference human epigenomes". In: *Nature* 518.7539 (2015), pp. 317–330. DOI: 10.1038/nature14248.

[24] Stephen Wilson et al. "A network of epigenomic and transcriptional cooperation encompassing an epigenomic master regulator in cancer". In: *npj Systems Biology and Applications* 4.1 (2018), pp. 1–10. DOI: 10.1038/s41540-018-0061-4.

[25] Yonatan Stelzer et al. "Tracing Dynamic Changes of DNA Methylation at Single-Cell Resolution". In: *CELL* 163.1 (2015), pp. 218–229. DOI: 10.1016/j.cell.2015.08.046.

[26] Yonatan Stelzer et al. "Parent-of-Origin DNA Methylation Dynamics during Mouse Development". In: *CellReports* 16.12 (2016), pp. 3167–3180. DOI: 10.1016/j.celrep.2016.08.066.

[27] Hisato Kobayashi et al. "Contribution of Intragenic DNA Methylation in Mouse Gametic DNA Methylomes to Establish Oocyte-Specific Heritable Marks". In: *PLOS Genetics* 8.1 (2012), e1002440. DOI: 10.1371/journal.pgen.1002440.

[28] Danny Leung et al. “Regulation of DNA methylation turnover at LTR retrotransposons and imprinted loci by the histone methyltransferase Setdb1”. In: *Proceedings of the National Academy of Sciences* 111.18 (2014), pp. 6690–6695. DOI: 10.1073/pnas.1322273111.

[29] Stefanie Seisenberger et al. “The Dynamics of Genome-wide DNA Methylation Reprogramming in Mouse Primordial Germ Cells”. In: *Molecular cell* 48.6 (2012), pp. 849–862. DOI: 10.1016/j.molcel.2012.11.001.

[30] Yidan Hu et al. “An estimated method of urban PM2.5 concentration distribution for a mobile sensing system”. In: *Pervasive and Mobile Computing* 25 (2016), pp. 88–103. DOI: 10.1016/j.pmcj.2015.06.004.

[31] Zakary S Singer et al. “Dynamic Heterogeneity and DNA Methylation in Embryonic Stem Cells”. In: *Molecular cell* 55.2 (2014), pp. 319–331. DOI: 10.1016/j.molcel.2014.06.029.

[32] Jiho Choi et al. “Prolonged Mek1/2 suppression impairs the developmental potential of embryonic stem cells”. In: *Nature* 548.7666 (2017), pp. 219–223. DOI: 10.1038/nature23274.

[33] Harry G Leitch et al. “Naive pluripotency is associated with global DNA hypomethylation”. In: *Nature Structural & Molecular Biology* 20.3 (2013), pp. 311–316. DOI: 10.1038/nsmb.2510.

[34] Ye-Ji Sim et al. “2i Maintains a Naive Ground State in ESCs through Two Distinct Epigenetic Mechanisms”. In: *Stem Cell Reports* 8.5 (2017), pp. 1312–1328. DOI: 10.1016/j.stemcr.2017.04.001.

[35] Ferdinand von Meyenn et al. “Impairment of DNA Methylation Maintenance Is the Main Cause of Global Demethylation in Naive Embryonic Stem Cells”. In: *Molecular cell* 62.6 (2016), p. 983. DOI: 10.1016/j.molcel.2016.06.005.

[36] Masaki Yagi et al. “Derivation of ground-state female ES cells maintaining gamete-derived DNA methylation”. In: *Nature* 548.7666 (2017), pp. 224–227. DOI: 10.1038/nature23286.

[37] Xiaoji Wu et al. “TET-mediated active DNA demethylation: mechanism, function and beyond”. In: *Nature Reviews Genetics* 18.9 (2017), pp. 517–534. DOI: 10.1038/nrg.2017.33.

[38] Meelad M Dawlaty et al. “Tet1 Is Dispensable for Maintaining Pluripotency and Its Loss Is Compatible with Embryonic and Postnatal Development”. In: *Cell Stem Cell* 9.2 (2011), pp. 166–175. DOI: 10.1016/j.stem.2011.07.010.

[39] Meelad M Dawlaty et al. “Combined Deficiency of Tet1 and Tet2 Causes Epigenetic Abnormalities but Is Compatible with Postnatal Development”. In: *Developmental Cell* 24.3 (2013), pp. 310–323. DOI: 10.1016/j.devcel.2012.12.015.

[40] Meelad M Dawlaty et al. “Loss of Tet Enzymes Compromises Proper Differentiation of Embryonic Stem Cells”. In: *Developmental Cell* 29.1 (2014), pp. 102–111. DOI: 10.1016/j.devcel.2014.03.003.

[41] Angelika Feldmann et al. “Transcription Factor Occupancy Can Mediate Active Turnover of DNA Methylation at Regulatory Regions”. In: *PLOS Genetics* 9.12 (2013), e1003994. DOI: 10.1371/journal.pgen.1003994.

[42] Matthew T Maurano et al. “Role of DNA Methylation in Modulating Transcription Factor Occupancy”. In: *Cell Reports* 12.7 (2015), pp. 1184–1195. DOI: 40_10.1016/j.celrep.2015.07.024.

[43] Yinong Yin et al. “Recent advances in oxide thermoelectric materials and modules”. In: *Vacuum* 146 (2017), pp. 356–374. DOI: 10.1016/j.vacuum.2017.04.015.

[44] Denes Hnisz et al. “Super-Enhancers in the Control of Cell Identity and Disease”. In: *CELL* 155.4 (2013), pp. 934–947. DOI: 10.1016/j.cell.2013.09.053.

[45] Aimée M Deaton et al. “CpG islands and the regulation of transcription.” In: *Genes & Development* 25.10 (2011), pp. 1010–1022. DOI: 10.1101/gad.2037511.

[46] Yuval Dor et al. “Principles of DNA methylation and their implications for biology and medicine”. In: *The Lancet* 392.10149 (2018), pp. 777–786. DOI: 10.1016/S0140-6736(18)31268-6.

[47] Dirk Schübeler. “Function and information content of DNA methylation”. In: *Nature* 517.7534 (2015), pp. 321–326. DOI: 10.1038/nature14192.

[48] Zachary D Smith et al. “DNA methylation: roles in mammalian development”. In: *Nature Publishing Group* 14.3 (2013), pp. 204–220. DOI: 10.1038/nrg3354.

[49] Benjamin R Sabari et al. “Coactivator condensation at super-enhancers links phase separation and gene control”. In: *Science* 361.6400 (2018), eaar2555. DOI: 10.1126/science.aar3958.

[50] Warren A Whyte et al. “Master Transcription Factors and Mediator Establish Super-Enhancers at Key Cell Identity Genes”. In: *CELL* 153.2 (2013), pp. 307–319. DOI: 10.1016/j.cell.2013.03.035.

[51] James A Gagnon et al. “Efficient Mutagenesis by Cas9 Protein-Mediated Oligonucleotide Insertion and Large-Scale Assessment of Single-Guide RNAs”. In: *PLOS ONE* 9.5 (2014), e98186. DOI: 10.1371/journal.pone.0098186.

[52] Jennifer Nichols et al. “Naive and Primed Pluripotent States”. In: *Cell Stem Cell* 4.6 (2009), pp. 487–492. DOI: 10.1016/j.stem.2009.05.015.

[53] J M Calabrese et al. “RNA sequence analysis defines Dicer’s role in mouse embryonic stem cells”. In: *Proceedings of the National Academy of Sciences* 104.46 (2007), pp. 18097–18102. DOI: 10.1073/pnas.0709193104.

[54] Rudolf Jaenisch et al. “Stem Cells, the Molecular Circuitry of Pluripotency and Nuclear Reprogramming”. In: *CELL* 132.4 (2008), pp. 567–582. DOI: 10.1016/j.cell.2008.01.015.

[55] Matt Thomson et al. “Pluripotency Factors in Embryonic Stem Cells Regulate Differentiation into Germ Layers”. In: *CELL* 145.6 (2011), pp. 875–889. DOI: 10.1016/j.cell.2011.05.017.

[56] TK Mistri et al. “Dynamic changes in Sox2 spatio-temporal expression promote the second cell fate decision through Fgf4/ Fgfr2 signaling in preimplantation mouse embryos”. In: *The Biochemical journal* 475.6 (2018), pp. 1075–1089. DOI: 10.1042/BCJ20170418.

[57] Lea A Medeiros et al. “Mir-290–295 deficiency in mice results in partially penetrant embryonic lethality and germ cell defects”. In: *Proceedings of the National Academy of Sciences* 108.34 (2011), pp. 14163–14168. DOI: 10.1073/pnas.1111241108.

[58] Alireza Paikari et al. “The eutheria-specific miR-290 cluster modulates placental growth and maternal-fetal transport”. In: *Development* 144.20 (2017), pp. 3731–3743. DOI: 10.1242/dev.151654.

[59] Eryn Wicklow et al. “HIPPO Pathway Members Restrict SOX2 to the Inner Cell Mass Where It Promotes ICM Fates in the Mouse Blastocyst”. In: *PLOS Genetics* 10.10 (2014), e1004618. DOI: 10.1371/journal.pgen.1004618.

[60] Matthew D Schultz et al. “Human body epigenome maps reveal noncanonical DNA methylation variation”. In: *Nature* 523.7559 (2015), pp. 212–216. DOI: 10.1038/nature14465.

[61] Christoph Weigel et al. “Epigenetic regulation of diacylglycerol kinase alpha promotes radiation-induced fibrosis”. In: *Nature Communications* 7.1 (2016), pp. 1–12. DOI: 10.1038/ncomms10893.

[62] Michael J Ziller et al. “Charting a dynamic DNA methylation landscape of the human genome”. In: *Nature* 500.7463 (2013), pp. 477–481. DOI: 10.1038/nature12433.

[63] Won-Ki Cho et al. “Mediator and RNA polymerase II clusters associate in transcription-dependent condensates”. In: *Science* 361.6400 (2018), pp. 412–415. DOI: 10.1126/science.aar4199.

[64] Qiaolin Deng et al. “Single-Cell RNA-Seq Reveals Dynamic, Random Monoallelic Gene Expression in Mammalian Cells”. In: *Science* 343.6167 (2014), pp. 193–196. DOI: 10.1126/science.1245316.

[65] Björn Reinius et al. “Random monoallelic expression of autosomal genes: stochastic transcription and allele-level regulation”. In: *Nature Publishing Group* 16.11 (2015), pp. 653–664. DOI: 10.1038/nrg3888.

[66] Hiroshi I Suzuki et al. “Super-Enhancer-Mediated RNA Processing Revealed by Integrative MicroRNA Network Analysis”. In: *CELL* 168.6 (2017), 1000–1014.e15. DOI: 10.1016/j.cell.2017.02.015.

[67] Holger Heyn et al. “DNA methylation profiling in the clinic: applications and challenges”. In: *Nature Publishing Group* 13.10 (2012), pp. 679–692. DOI: 10.1038/nrg3270.

[68] Keith D Robertson. “DNA methylation and human disease”. In: *Nature Publishing Group* 6.8 (2005), pp. 597–610. DOI: 10.1038/nrg1655.

[69] Yong Zhang et al. “Model-based Analysis of ChIP-Seq (MACS)”. In: *Genome Biology* 9.9 (2008), pp. 1–9. DOI: 10.1186/gb-2008-9-9-r137.

[70] H Li et al. “The Sequence Alignment/Map format and SAMtools”. In: *Bioinformatics* 25.16 (2009), pp. 2078–2079. DOI: 10.1093/bioinformatics/btp352.

[71] H Li et al. “Fast and accurate short read alignment with Burrows-Wheeler transform”. In: *Bioinformatics* 25.14 (2009), pp. 1754–1760. DOI: 10.1093/bioinformatics/btp324.

[72] Fidel Ramírez et al. “deepTools: a flexible platform for exploring deep-sequencing data”. In: *Nucleic Acids Research* 42.W1 (2014), W187–W191. DOI: 10.1093/nar/gku365.

[73] Alexander Dobin et al. “STAR: ultrafast universal RNA-seq aligner”. In: *Bioinformatics* 29.1 (2012), pp. 15–21. DOI: 10.1093/bioinformatics/bts635.

[74] Michael I Love et al. “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2”. In: *Genome Biology* 15.12 (2014), p. 31. DOI: 10.1186/s13059-014-0550-8.

[75] H Mi. “The PANTHER database of protein families, subfamilies, functions and pathways”. In: *Nucleic Acids Research* 33.Database issue (2004), pp. D284–D288. DOI: 10.1093/nar/gki078.

[76] Paul D Thomas et al. “PANTHER: a library of protein families and subfamilies indexed by function.” In: *Genome Research* 13.9 (2003), pp. 2129–2141. DOI: 10.1101/gr.772403.

[77] Caroline A Schneider et al. “NIH Image to ImageJ: 25 years of image analysis”. In: *Nature Methods* 9.7 (2012), pp. 671–675. DOI: 10.1038/nmeth.2089.

[78] Arjun Raj et al. “Imaging individual mRNA molecules using multiple singly labeled probes”. In: *Nature Methods* 5.10 (2008), pp. 877–879. DOI: 10.1038/nmeth.1253.

[79] Bluma J Lesch et al. “A set of genes critical to development is epigenetically poised in mouse germ cells from fetal stages through completion of meiosis”. In: *Proceedings of the National Academy of Sciences* 110.40 (2013), pp. 16061–16066. DOI: 10.1073/pnas.1315204110.

2 SINGLE-CELL TRANSCRIPTOMICS REVEALS GENE EXPRESSION DYNAMICS OF HUMAN FETAL KIDNEY DEVELOPMENT

THIS CHAPTER IS BASED ON:
Mazène Hochane, Patrick R van den Berg, Xueying Fan, Noémie Bérenger-Currias, Esmée Adegeest, Monika Bialecka, Maaike Nieveen, Maarten Menschaart, Susana M Chuva de Sousa Lopes, Stefan Semrau. "Single-cell transcriptomics reveals gene expression dynamics of human fetal kidney development". In: *PLOS Biology* 17.2 (Feb. 2019), e3000152. DOI: 10.1371/journal.pbio.3000152

Abstract

Regenerative medicine offers an exciting avenue to potential cures of kidney disease. However, a detailed knowledge of the structure and embryonic development of the kidney is crucial, both for stimulating regeneration in the body and for growing healthy kidney tissue in a dish. Most of such knowledge has been obtained from mice, whose development differs in crucial ways from that of humans. We therefore studied the composition of human fetal kidney tissue from five developmental ages, using a technique that can measure gene expression in individual cells. Our measurements revealed 22 distinguishable cell types, some of which we localized in the tissue by fluorescence microscopy. We found several subpopulations of nephron progenitor-cells that give rise to the nephron, the functional unit of the kidney. Our study also focused on the development of podocytes, a cell type that is crucial for the filtration function of the kidney, and our results might inform attempts to recreate these cells in a dish. We hope that our dataset, made conveniently accessible through a web application, will help scientists develop new regenerative medicine approaches to kidney disease.

2.1 Introduction

Mammalian kidney development initiates in the intermediate mesoderm through crosstalk between the metanephric mesenchyme (MM) and the ureteric bud (UB). The UB originates from the nephric duct, invades the MM, and starts to subdivide progressively into multiple ramifications. The UB tip cells, which make the first contact with the MM, become enveloped by an assembly of mesenchymal cells, the cap mesenchyme (CM) (Fig 1A-B). The CM contains nephron progenitor cells (NPCs), which give rise to the whole nephron epithelium through tightly regulated morphogenic transformations [2]. Self-renewal of (mouse) NPCs is governed by key transcription factors, such as *Six2* and *Meox1*, which mark the nephrogenic zone of the kidney [3]. Signaling between UB tip cells and NPCs regulates the balance between self-renewal and differentiation of the NPCs [4]. In humans, about 1 million nephrons are produced before the NPC population is irrevocably exhausted a few weeks before birth [5]. During nephrogenesis, NPCs undergo mesenchymal-epithelial transition and differentiate into a succession of intermediate structures: the pretubular aggregate (PTA), renal vesicle (RV), comma-shaped body (CSB) and s-shaped body (SSB). Then, via the capillary loop stage, mature and functional glomerular and tubular structures are eventually formed. In contrast to the nephron epithelium, the collecting duct system originates from the UB. GDNF/RET signaling between the UB and CM critically regulates proliferation of UB tip cells and branching morphogenesis of the UB [6]. Stromal cells-such as interstitial cells (ICs), mesangial cells, juxtaglomerular cells, smooth muscle cells, fibroblasts, and pericytes-derive from a common interstitial progenitor [7, 8]. Finally, vascular endothelial cells and the highly specified glomerular endothelium originate from the MM [9], and leukocytes and erythrocytes enter with the blood stream. The current understanding of mammalian kidney development is largely based on mouse studies, although it is clear that human and mouse kidneys are morphologically different. Three recent landmark studies have revealed, in great detail, a significant divergence between mouse and human renal embryogenesis in terms of morphology as well as gene expression [10, 11, 12]. These studies underline that the prevailing lack of data on human kidney development severely hinders the detailed understanding of human kidney development and possible developmental origins of kidney disease.

In the study described here, we used scRNA-seq to study gene expression dynamics in human fetal kidney development. Analysis of a fetal kidney from week 16 (w16) of gestation revealed 22 cell types, which we identified by known marker genes. Pseudotime analysis clarified their temporal relationship. We further defined specifically expressed cell type marker genes of which many have not been implied in kidney development. Comparison to four additional samples (from w9, w11, w13, and w18) suggested that most cell types have a constant expression pattern, with the notable exception of podocytes. To highlight two ways in which our dataset can be interrogated, we then explored the nephrogenic niche and the development of podocytes. Gene expression differences between four NPC clusters were related to spatial heterogeneity by immunostaining and single-molecule FISH (smFISH). Expression of the disease-associated gene *UNCX* was localized to NPCs and their early derivatives. Fi-

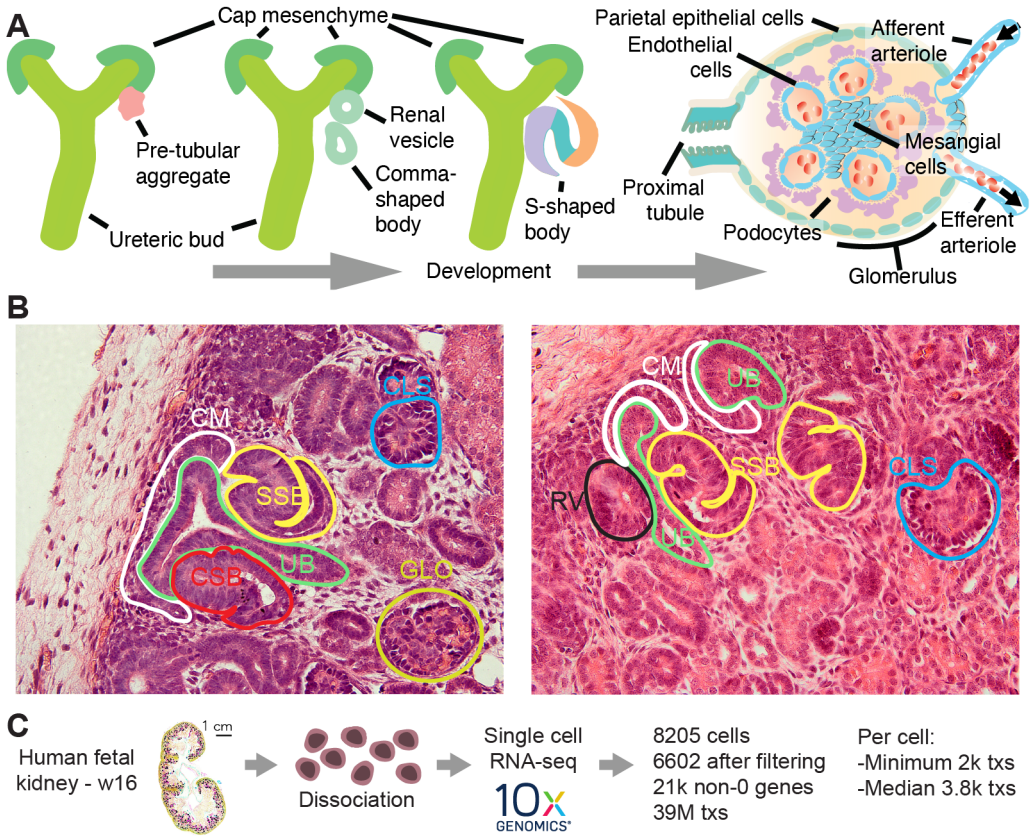


Figure 1. Overview of kidney development and the experimental design. (A) Schematic of kidney epithelium development. (B) Several morphologically distinct stages of nephrogenesis are highlighted by colored lines in images of human fetal kidney sections stained with HE. (C) Overview of the scRNA-seq experiment. Ureteric bud (UB), cap mesenchyme (CM), pretubular aggregate (PTA), renal vesicle (RV) (Stage I), comma-shaped body (CSB) (Stage II), s-shaped body (SSB) (Stage II), capillary-loop stage (CLS) (Stage III), glomerulus (GLO) (Stage IV)

nally, we focused on podocyte development, which proceeds via a distinct precursor state. By immunostaining and smFISH, we localized these precursors in situ and confirmed the disease-associated gene *OLFM3* as a marker.

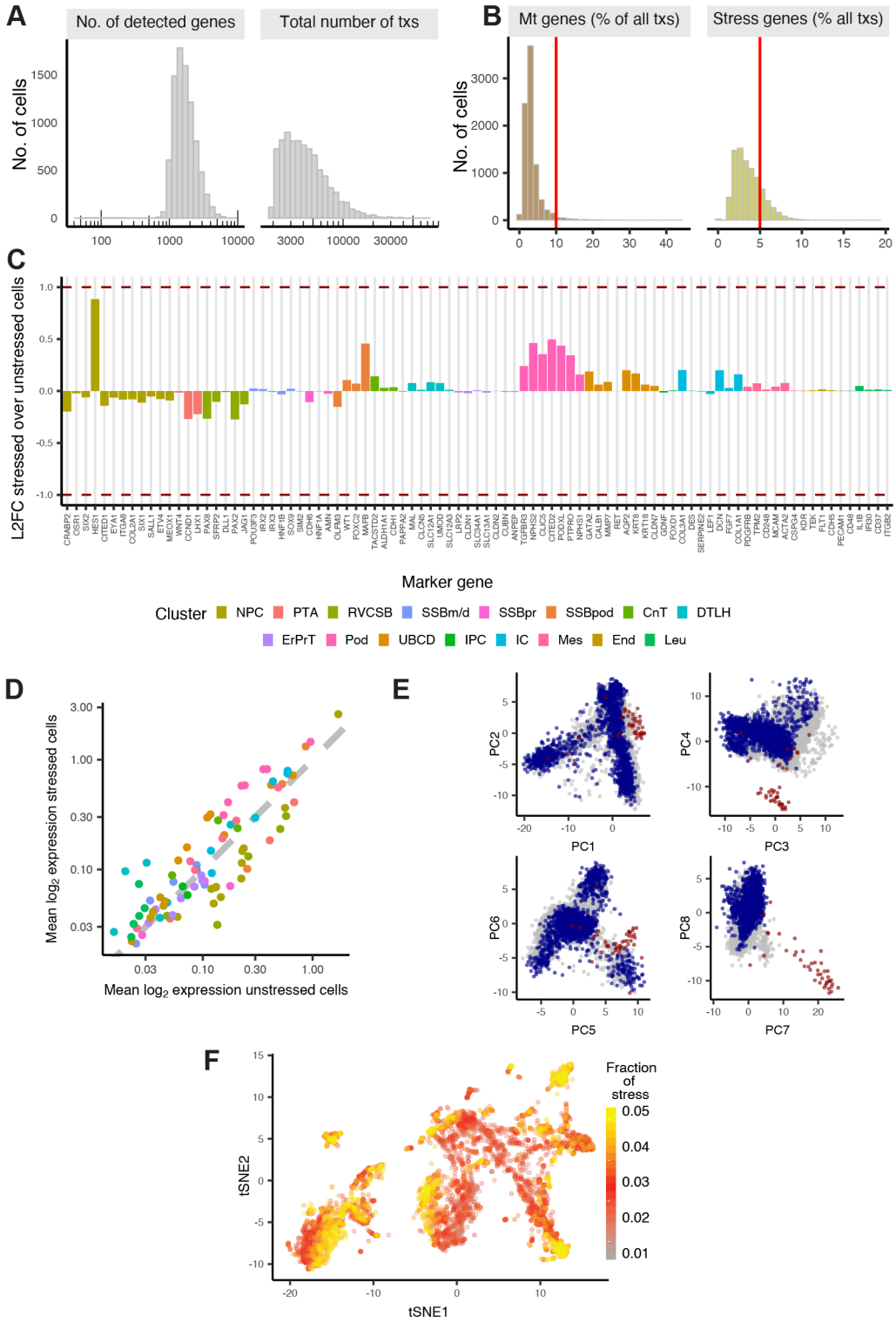
2.2 Results

2.2.1 Clustering and identification of cell types

We performed single-cell transcriptomics on a human fetal kidney from w16 of gestation, equivalent to 14 weeks of development (Fig 1C). After data pruning and stringent removal of

cells affected by stress (Fig 2, Materials and methods), 6,602 cells were retained for further analysis. Clusters of cells were identified by hierarchical clustering after k-nearest neighbors (knn) smoothing [13]. We assigned cell types (see Cell types) to these clusters by expression of marker genes from the literature on mouse kidney development. The studies that linked the genes of this literature set to particular cell types are referenced in Table 2.1. After merging similar clusters (Fig 4, Fig 5, Materials and methods), we obtained 22 cell types (Hochane et al. [1] S4 Fig) and visualized the single-cell transcriptomes in a two-dimensional t-distributed stochastic neighbor embedding (tSNE) map [14] (Fig 3).

Figure 2 ***(following page).*** **Removing stressed cells did not bias the scRNA-seq results in the w16 sample.** (A) Number of detected genes and total number of transcripts per cell. tx = transcript (B) Relative expression of mitochondrial and stress marker genes per cell. Red line indicates the threshold used to define stressed cells. See Materials and methods for the list of mitochondrial genes and Hochane et al. [1] S2 Table for the list of stress markers. (C and D) $\log_2$ fold change (L2FC) and scatter plot of the literature set genes (Table 2.1). Red dashed lines indicate fold-change of 0.5 and 2. (E) PCs one to eight of the top 5% most HVGs for all cells. Blue and red points indicate stressed cells and red blood cells, respectively. Expression values in C-E are normalized to library size and log-transformed with a pseudocount of 1. (F) Fraction of stress markers in the 6,602 remaining cells. tSNE map corresponds to Fig 3.


A
No. of detected genes
Total number of txs
No. of cells
B
Mt genes (% of all txs)
Stress genes (% all txs)
No. of cells
C
L2FC stressed over unstressed cells
Marker gene
Cluster NPC PTA RVCSB SSBm/d SSBpr SSBpod CnT DTLH ErPrT Pod UBCD IPC IC Mes End Leu
D
Mean log2 expression stressed cells
Mean log2 expression unstressed cells
E
PC1
PC2
PC3
PC4
PC5
PC6
PC7
PC8
F
tSNE1
tSNE2
Fraction of stress
0.05
0.04
0.03
0.02
0.01

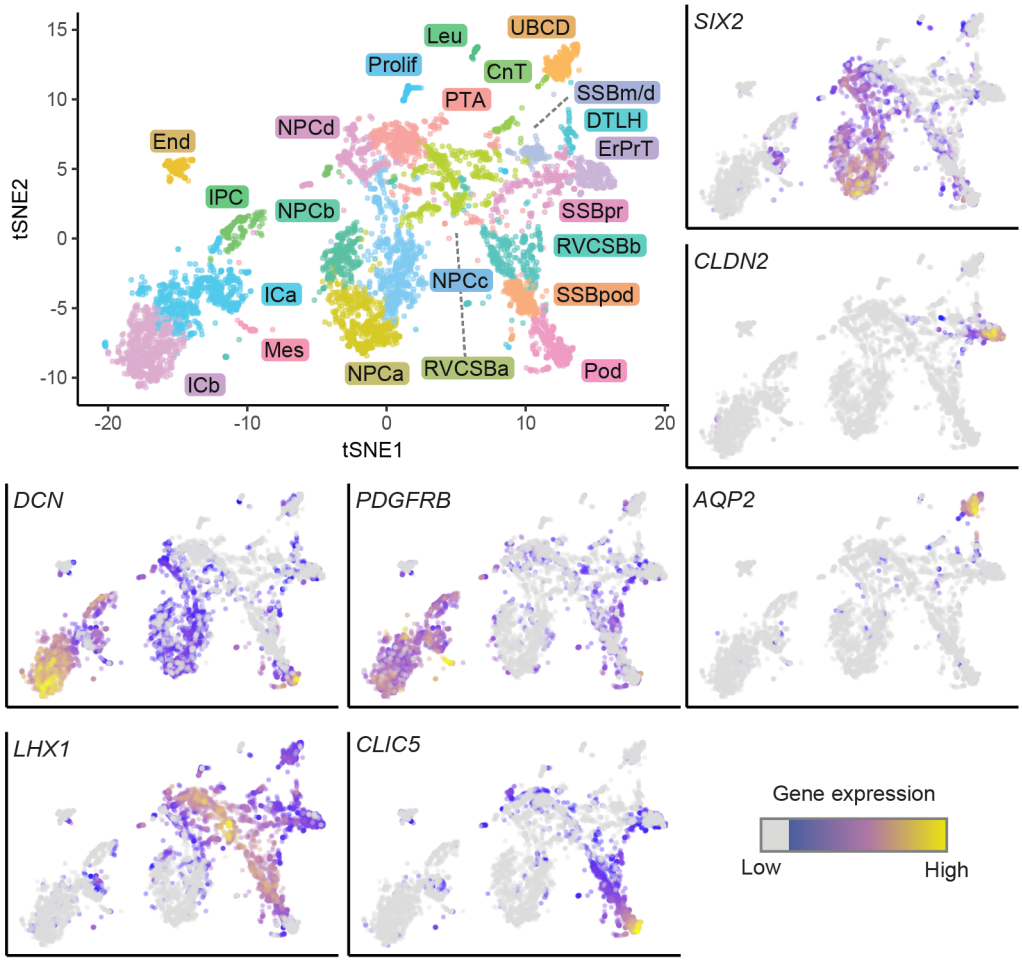
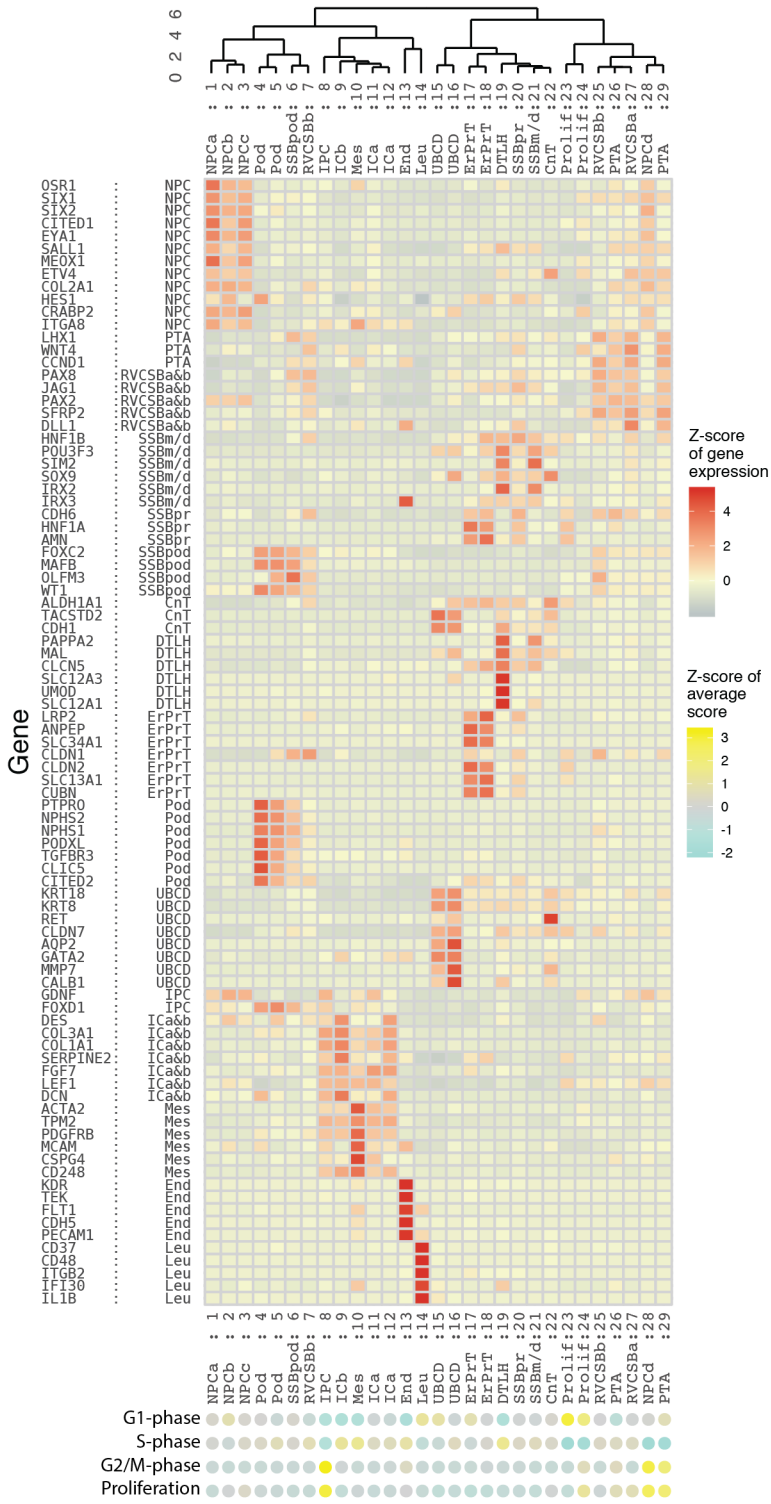
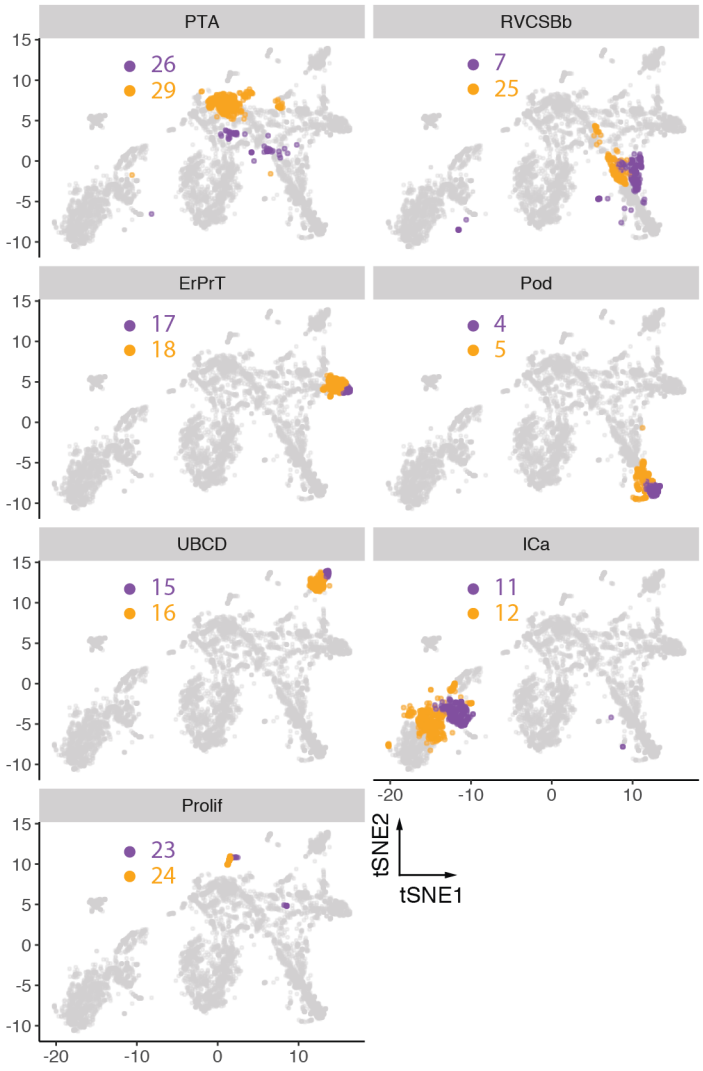


Figure 3. Single-cell transcriptomics identified 22 unique cell types in the human fetal kidney. Top left: 2D tSNE map of 6,602 human fetal kidney cells. Colors and labels indicate the assigned cell type. (Other panels) tSNE maps indicating expression of *SIX2*, *LHX1*, *CLDN2*, *CLIC5*, *DCN*, *PDGFRB*, and *AQP2*. Expression is indicated by color; expression values of 1 are plotted in gray.

Figure 4 *(following page)*. Adjacent clusters were merged based on similarity in literature set gene expression. Heat map of literature set gene expression. Expression was Freeman-Tukey (FT) transformed, averaged over all cells in the 29 clusters found by hierarchical clustering (indicated by the dendrogram on top of the heat map) and standardized gene-wise. Cluster average cell cycle scores, calculated by Cyclone [15] as well as average expression of proliferation markers [16], are indicated by colored circles below each cluster (Z-score of the mean score or mean expression).

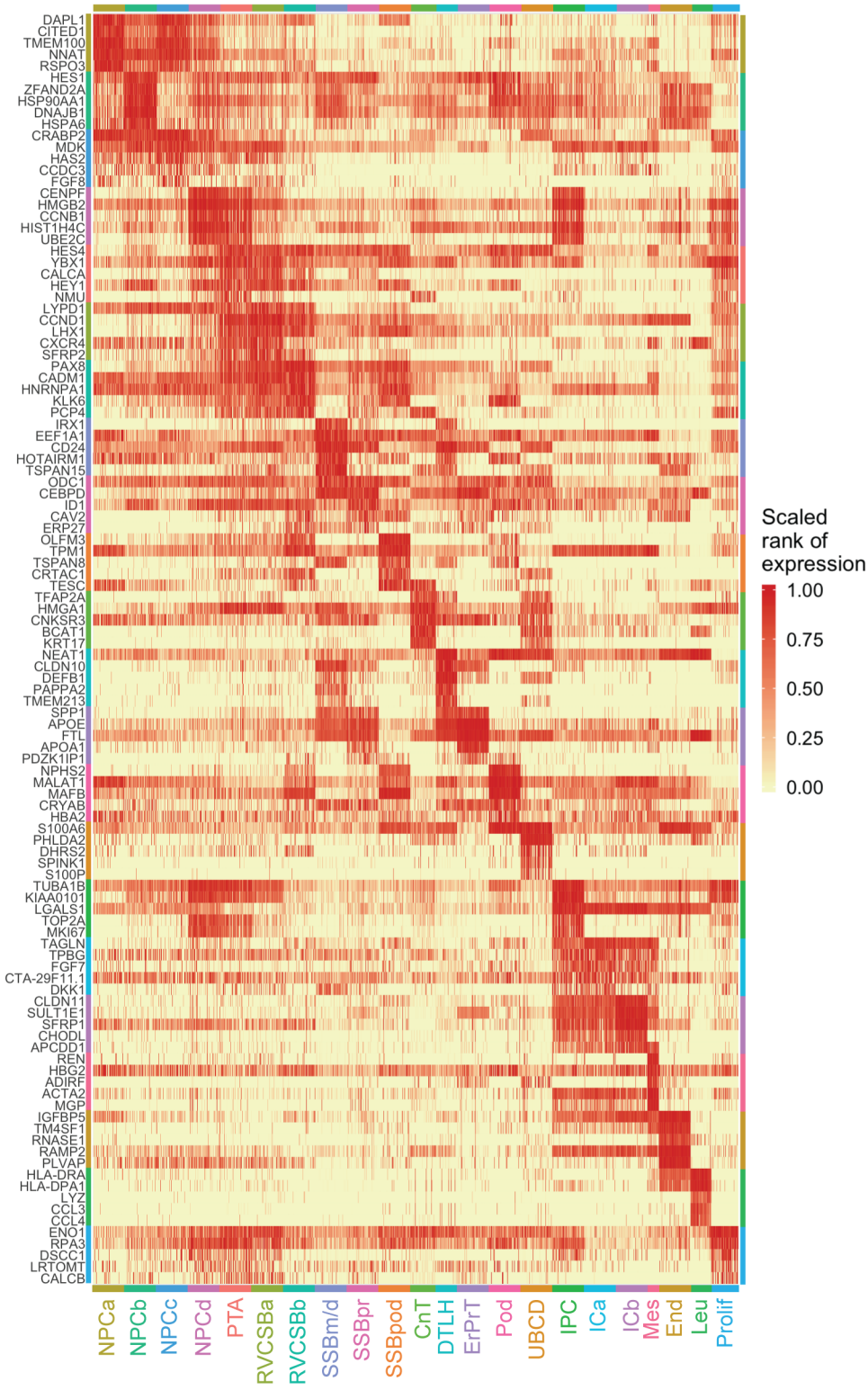

Gene
Z-score of gene expression
4
2
0
Z-score of average score
3
2
1
0
-1
-2
G1-phase
S-phase
G2/M-phase
Proliferation



Cell type	Clusters	No of cells
NPCa	1	522
NPCb	2	296
NPCc	3	596
NPCd	28	280
PTA	26&29	56&403
RVCSBa	27	367
RVCSBb	7&25	181&131
SSBm/d	21	151
SSBpr	20	213
SSBpod	6	257
CnT	22	80
DTLH	19	67
ErPrT	17&18	46&292
Pod	4&5	224&206
UBCD	15&16	55&262
IPC	8	177
ICa	11&12	272&381
ICb	9	677
Mes	10	37
End	13	223
Leu	14	65
Prolif	23&24	16&69

Figure 5. Merged hierarchical clusters were similar on the tSNE map. tSNE maps highlighting the clusters that were merged to give the cell types indicated in the titles of each map. (Inset right) Table listing the numbers of cells in each of the 29 original clusters. tSNE map corresponds to Fig 3.

Figure 6 *(following page)*. HVGs adequately described all cell clusters. Heat map of 2,034 randomly chosen cells (maximum 100 per cluster) and the five most HVGs with a minimum mean expression of 0.01 excluding stress markers (Hochane et al. [1] S2 Table) and ribosomal genes. Genes were assigned to clusters based on highest mean expression within that cluster. Values shown are the ranks of nonzero cells (cells with no expression receive rank 0) divided by the highest rank per gene.


DAPL1
CITED1
TMEM100
NNAT
RSPO3
HES1
ZFAND2A
HSP90AA1
DNAJB1
HSPA6
CRABP2
MDK
HAS2
CCDC3
FGF8
CENPF
HMGB2
CCNB1
HIST1H4C
UBE2C
HES4
YBX1
CALCA
HEY1
NMU
LYPD1
CCND1
LHX1
CXCR4
SFRP2
PAX8
CADM1
HNRNPA1
KLK6
PCP4
IRX1
EEF1A1
CD24
HOTAIRM1
TSPAN15
ODC1
CEBPD
ID1
CAV2
ERP27
OLFM3
TPM1
TSPAN8
CRTAC1
TESC
TFAP2A
HMGA1
CNKSR3
BCAT1
KRT17
NEAT1
CLDN10
DEFB1
PAPPA2
TMEM213
SPP1
APOE
FTL
APOA1
PDZK1IP1
NPHS2
MALAT1
MAFB
CRYAB
HBA2
S100A6
PHLDA2
DHRS2
SPINK1
S100P
TUBA1B
KIAA0101
LGALS1
TOP2A
MKI67
TAGLN
TPBG
FGF7
CTA-29F11.1
DKK1
CLDN11
SULT1E1
SFRP1
CHODL
APCDD1
REN
HBG2
ADIRF
ACTA2
MGP
IGFBP5
TM4SF1
RNASE1
RAMP2
PLVAP
HLA-DRA
HLA-DPA1
LYZ
CCL3
CCL4
ENO1
RPA3
DSCC1
LRTOMT
CALCB
NPCa
NPCb
NPCc
NPCd
PTA
RVCSBa
RVCSBb
SSBm/d
SSBpr
SSBpod
CnT
DTLH
ErPrT
Pod
UBCD
IPC
ICa
ICb
Mes
End
Leu
Prolif
Scaled rank of expression
1.00
0.75
0.50
0.25
0.00

The mean expression levels of the literature set genes showed clear differences between cell types (Fig 7). NPCs, which were distributed over four distinct clusters (NPCa-d), were marked by the established markers *SIX2* (Fig 3), *CITED1*, *MEOX1*, and *EYA1*. Expression of these progenitor markers was highest in NPCa, which we hence considered *bona fide* self-renewing NPCs. NPCb showed lower levels of *CITED1* and *SALL1* and higher levels of *GDNF* and *HES1* compared to the other NPC clusters. *HES1*, a transcription factor downstream of Notch signaling, is important for further renal cell differentiation. Compared to the other NPC clusters, NPCc showed higher expression of *CRABP2*, which is related to retinoic acid (RA) signaling [17]. NPCd exhibited low *OSR1*, *CITED1*, and *MEOX1* expression and increased levels of *LEF1*, a known indicator of NPC induction towards differentiation. Compared to the other NPC subtypes, NPCd was also marked by a larger fraction of cells in G2/M-phase (Fig 8A-B) and a higher expression of proliferation markers (Fig 8C), which indicated faster proliferation. We will discuss the relationship between the various NPC clusters in more detail below (see Heterogeneity in the nephrogenic niche).

Nephrogenesis continues with the creation of PTA cells, which in turn develop into the RV and CSB cells. In our data, PTA cells were identified based on high expression of *LHX1* (Fig 3), *JAG1*, *WNT4*, and *CCND1*. Because RV and CSB are mainly distinguishable by morphology, cells belonging to these two structures were grouped in our analysis (RVCSB). RVCSBs were marked by the same genes as PTA cells but they appeared to proliferate less (Fig 8). Furthermore, they expressed markers reflecting more advanced regional patterning, which allowed us to discriminate between two subtypes (a and b). RVCSBa had a higher expression of genes that were recently associated with the distal RV (*SFRP2*, *DLL1*, *LHX1*), whereas RVCSBb expressed genes that indicate the proximal RV (*CDH6*, *FOXC2*, *MAFB*, *CLDN1*, *WT1*).

The next step in development is the formation of the SSB. In our dataset, this structure was represented by three clusters, named according to the part of tubule and glomerular epithelium they are known to give rise to SSBpr, proximal tubule; SSB medial/distal (SSBm/d); SSB podocyte precursor cell (SSBpod), podocytes. SSBpr were identified in our data by markers of the early proximal tubule (ErPrT) (such as *NF1A*, *CDH6*, *AMN*), as well as low expression of SLC3A1, *LRP2*, and *SLC13A1*, which are known to be found in more mature proximal tubule cells. Therefore, SSBpr were likely precursors of the ErPrT cells, which expressed higher levels of early proximal markers together with CLDN2 (Fig 3), *ANPEP*, and *SLC34A1*. Another cluster accounted for the precursor cells of the loop of Henle (LOH) and the distal tubule in the SSB (SSBm/d). This cluster could be identified by the presence of *IRX1*, *IRX2*, *SIM2*, *SOX9*, *POU3F3*, and *HNF1B*, together with low expression of *PAPPA2* and *MAL* and the absence of *CDH6* and *HNF1A*. Cell types that are known to develop from the SSBm/d were found together in one cluster (DTLH). This cluster showed high expression of the distal markers *MAL*, *CLCN5*, *SLC12A3*, and *POU3F3*, which are specific to the distal tubule, as well as *SLC12A1*, *PAPPA2*, and *UMOD*, which are found in the LOH. Finally, cells that likely gave rise to podocytes, SSBpod, clustered separately. These cells showed high expression of *MAFB* and *FOXC2*, both transcription factors necessary for the development of podocyte identity, and low lev-

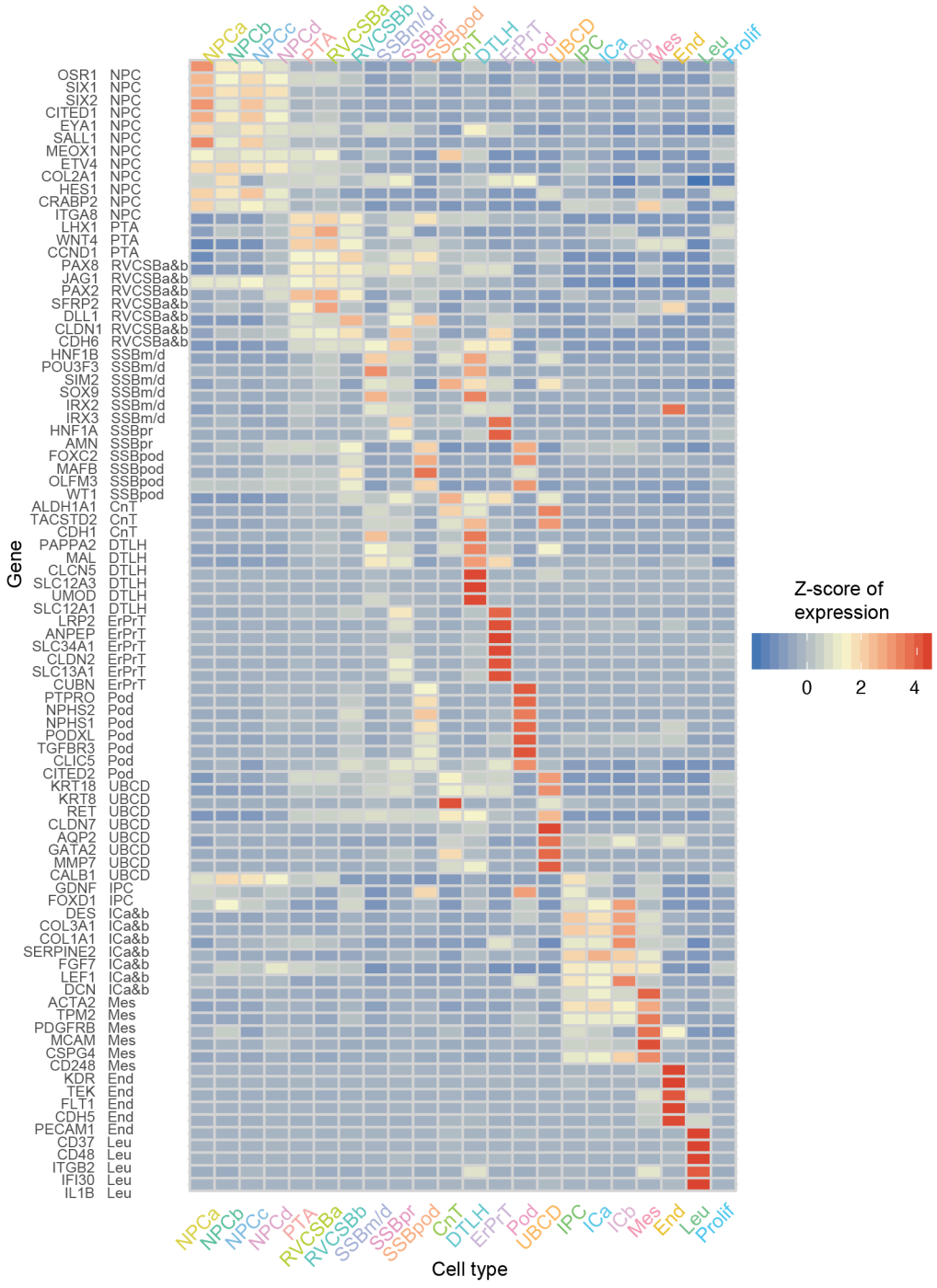


Figure 7. Known markers elucidated the cell types corresponding to each cluster. Heatmap of literature set gene expression in the 22 identified cell types. Expression was FT transformed, averaged over all cells in a cluster and standardized gene-wise.

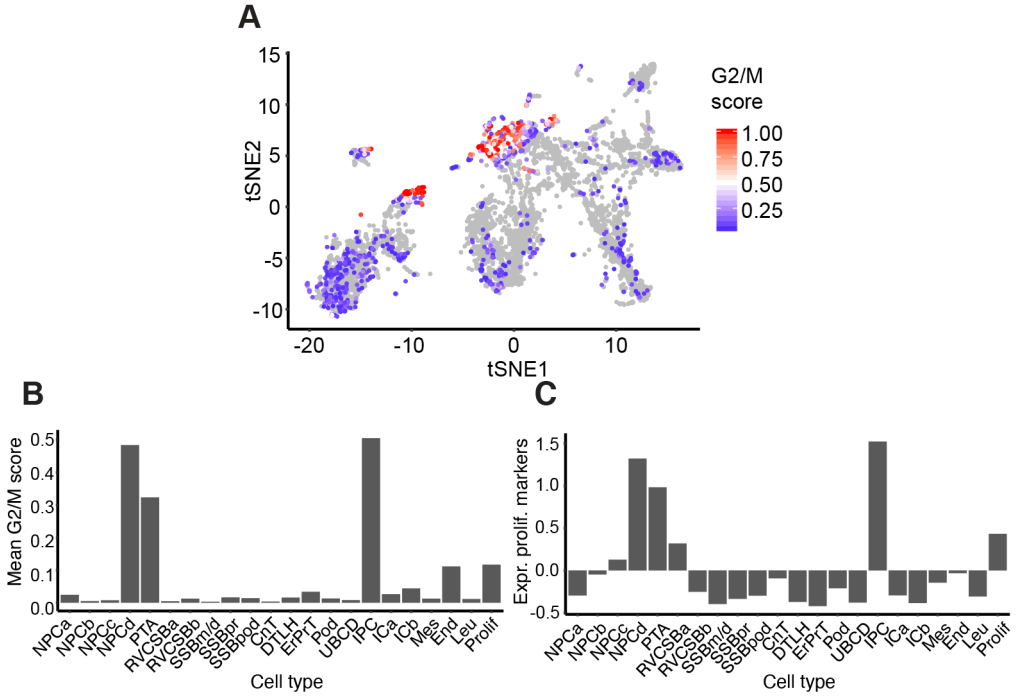


Figure 8. Proliferation states varied per cell type. (A) tSNE map of all cells with color indicating the G2/M score (calculated by the Cyclone tool [15]. This score reflects the likelihood that a cell is in G2/M phase. (B) G2/M scores from panel A averaged over the cells in each cell type. (C) Mean expression of proliferation markers [16] (Z-scores) per cell type.

els of the mature podocytes markers *CLIC5* (Fig 3), *PTPRO*, *NPHS1*, and *NPHS2*. This cluster also showed the highest expression of *OLFM3*, previously identified as a specific marker of podocyte precursors residing in the visceral part of the proximal segment of the SSB. In contrast to SSBpod, podocytes showed higher expression of mature podocyte markers and lower levels of *MAFB*. Differences between SSBpod and podocytes will be studied in more detail below (see Podocyte development). Because of the high similarity in gene expression between SSB and capillary loop stage, we could not exclude that the SSB clusters also contained cells from the capillary loop stage.

Cells of the connecting tubule (CnT), which connects the distal tubule to the collecting duct, could also be identified in the data. They shared markers with the collecting duct (such as *ALDH1A1*, *TACSTD2*, and *CDH1*), distal tubule (*SOX9*, *POU3F3*), and UB (*RET*, *KRT8*, *KRT18*, *MMP7*). Cells of the ureteric bud/collecting duct (UBCD) were strongly marked by well-known genes like *AQP2* (Fig 3), *CALB1*, *KRT8*, *KRT18*, *RET*, and *GATA2*, found in the collecting duct as well as the stalk and tip of the UB.

The developing nephrons are surrounded by interstitial tissue, a separate lineage that originates in interstitial progenitor cells (IPCs). We identified IPCs by coexpression of *FOXD1*

and *GDNF*. These cells also expressed lower levels of markers known to be found in more mature cells like *PDGFRA* for ICs or *PDGFRB* (Fig 3) and ACTA2 for mesangial cells. We identified two subtypes of IC (a and b), which were similar in their marker gene profile. Compared to IPCs, they lacked *FOXD1* and expressed less (ICa) or no (ICb) GDNF. ICa showed high levels of *FGF7*, which has been localized to the renal fibroblasts or stroma surrounding the ureter and the collecting system. ICa also showed high levels of *TPM2* and *ACTA2*, markers of smooth muscle-like cells. ICb, on the other hand, expressed genes like DCN (Fig 3), *DES*, *SERPINE2*, and *COL3A1*, which are known to mark cortical stromal cells. Endothelial cells were identified by markers such as *KDR* and *TEK*, whereas leukocytes showed many specifically expressed genes, such as *CD37* or *CD48*. Finally, one cluster of cells (proliferating cells) had a higher expression of proliferation markers compared to most other cell types (Fig 8C) but lacked discernible cell type markers.

2.2.2 Developmental flow

The literature-based analysis of the found clusters seemed to suggest that cells cluster by developmental progression (e.g., NPCs versus PTA cells), as well as location (e.g., RVCSBa, distal, versus RVCSBb, proximal) Because the interpretation of clusters is sometimes based on genes that are expressed in multiple developmental stages, we wanted to retrieve the developmental flow with an independent method We used *Monocle 2* [18] to learn a graph that represents the developmental hierarchy of the cell types from the PTA on (Fig 9A) Subsequently, cells were placed on a pseudotime scale rooted in the PTA (Fig 9B-C) This analysis showed that PTA cells were followed by RVCSBa and RVCSBb, the SSB clusters, and finally the clusters identified as more mature types (DTLH, ErPrT, podocytes) Therefore, the clustering was strongly driven by developmental progression RVCSBa cells were distributed over a fairly broad period of pseudotime and already occurred before branch point 1, which separates proximal from distal cell fates (Fig 9A) This might indicate that some of these cells preceded the RVCSBb, whereas others were primed to develop into distal fates RVCSBb cells, however, only appeared after branch point 1, which confirmed that they were likely progenitors of proximal cell fates On three separate branches, SSBm/d preceded DTLH, SSBpr preceded ErPrT, and SSBpod preceded podocytes, which confirmed the identity of the SSB clusters The temporal relationship of the NPC subtypes will be discussed in detail below (see Heterogeneity in the nephrogenic niche)

2.2.3 Comparison with existing single-cell transcriptomics data

To further confirm the interpretation of the cell clusters, we wanted to compare our data with an existing single-cell transcriptomics study of a w17 fetal kidney by Lindström et al. [20]. To that end, we first corrected for batch effects, using a method based on matching mutual nearest neighbors in the two datasets [21]. After correction, the two datasets showed a large degree of overlap (Fig 10, Fig 11). This allowed us to use the cell types found by Lind-

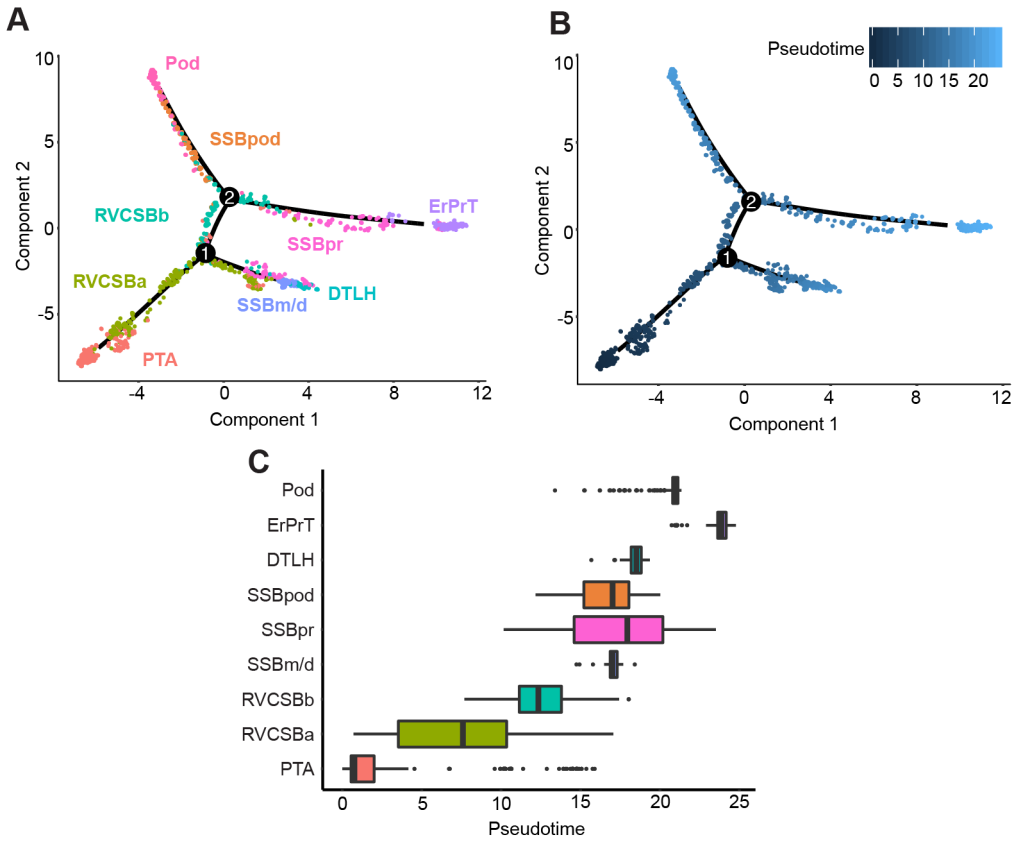


Figure 9. Pseudotime analysis clarified the developmental relationship of the cell clusters. (A) Two-dimensional embedding (with the *DDRTree* algorithm [19] of all w16 kidney cells, calculated by *Monocle 2* [18]. The graph learned by the algorithm is shown as a black line. Colors and labels indicate cell types. (B) Same embedding as in panel A. Color indicates pseudotime calculated by *Monocle 2*. (C) Box plots of cell type distribution over pseudotime.

ström et al. [20] to classify the cell clusters found here, using a knn approach (see Materials and methods). NPCa-c were also classified as NPC by Lindström et al. [20], whereas NPCb were considered *primed NPC*, which supports the notion that NPCb were primed to differentiate. The NPCd cluster was classified as *proliferating cells* This classification is in agreement with our observation that NPCd seemed to proliferate more than other NPC subtypes (Fig 8). Because NPCd cells expressed low levels of NPC markers (such as *SIX2* and *CITED1*), these cells were likely in a transition state between NPCs and PTA cells. Although the majority of PTA cells identified here were considered *PTA/RV I* by Lindström et al. [20], RVCSBa cells were spread over multiple cell types. This spread was likely due to the fact that transitory cell types are transcriptionally similar, and their clustering is therefore less robust. Nevertheless, the *PTA/RV II* cluster received most of the RVCSBa cells. RVCSBb cells were called *podocyte*

precursors in the Lindström et al. [20], whereas SSBpod as well as podocytes were classified as *podocytes*. In our dataset, RVCSBb directly preceded SSBpod (Fig 9A), so they could indeed be considered podocyte progenitors. Below, we will show that SSBpod did form a cell state separate from podocytes and should not be grouped with them (see Podocyte development). In agreement with our analysis, the majority of SSBpr were classified as *proximal precursor* or *proximal tubule*, and all ErPrT were considered *proximal tubule* by Lindström et al. [20]. CnT and DTLH were both classified as *distal/LOH precursor*. The fact that two cell types in the study by Lindström et al. [20] (*podocytes* and *distal/LOH precursor*) were split in multiple subclusters in our study likely reflects differences in sample preparation. Whereas Lindström et al. [20] preferentially released single cells from the nephrogenic niche, here, the whole kidney was used. Consequently, the Lindström et al. [20] dataset has a finer resolution of NPCs and early, proliferating cell types, whereas our dataset allowed us to resolve more mature cell types. The two datasets therefore complement each other.

2.2.4 Marker identification

To confirm the inferred cell types and also identify novel markers, we pursued two complementary strategies. First, we determined a set of marker genes based on their usefulness as classifiers for individual cell types: for each gene, the performance of a binary classifier was evaluated by the area under the ROC (AUROC) and combined with expression level filtering (see Materials and methods). This resulted in 88 marker genes (See Fig 12, marker set, Hochane et al. [1] S3 Table). Only 11 of these markers overlapped with the 89 genes in the literature set (See Fig 13C). To our knowledge, many of the remaining markers had not been associated with kidney development in previous studies. As an independent approach, we used the KeyGenes algorithm [22] to identify classifier genes among the 500 most HVGs, using two-thirds of all cells as a training set. Based on the classifier genes determined by KeyGenes, we next predicted the cell types of the remaining one-third of the cells (test set). Cell types could be predicted with an average certainty (id score) of 0.59; 24% of the cells in the test set obtained an id score higher than 0.8. Of the 95 classifier genes (See Fig 12A, KeyGenes set, Hochane et al. [1] S3 Table), 24 were the same as in the marker set, and 14 were common with the literature set (See Fig 13B).

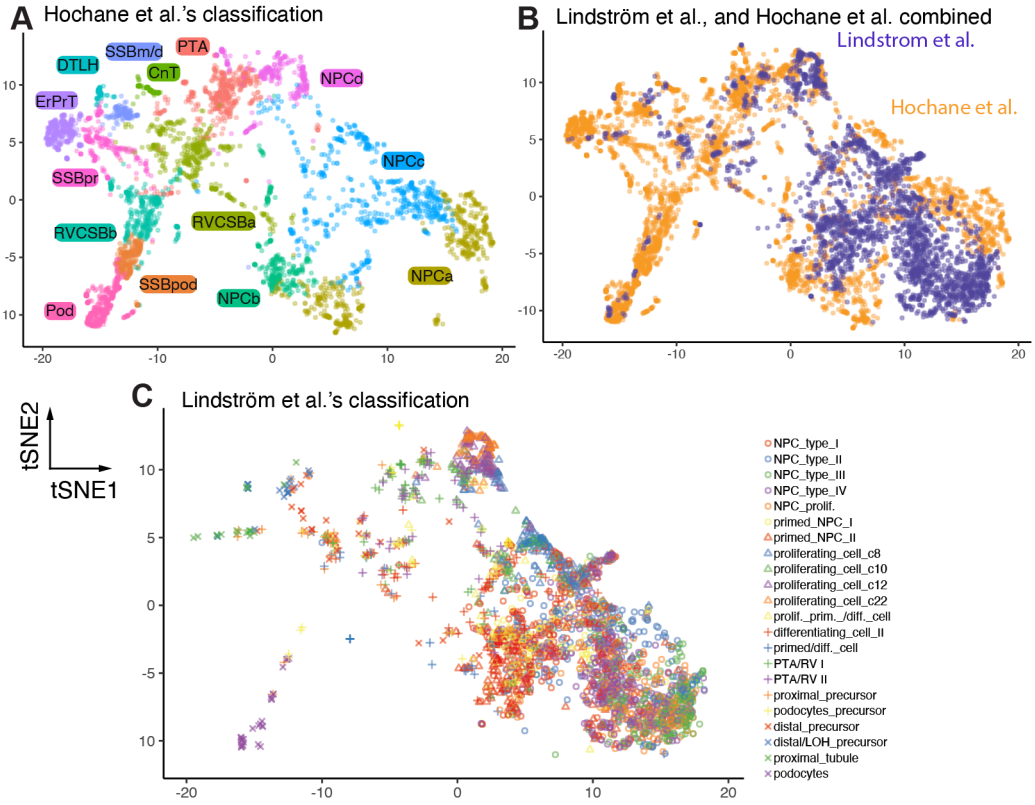


Figure 10. Comparison with an existing single-cell transcriptomics dataset showed congruent expression profiles. Two-dimensional tSNE maps comparing the data presented here with the data from Lindström et al. [20] both restricted to the nephrogenic niche by their own classification. The map was calculated using both data sets after batch correction [21]. (A) Only cells measured in this study are shown. Color and labels indicate the classification developed in this study. (B) Same tSNE map as above. Color indicates the data set. (C) Same tSNE map as (B). Only cells measured by Lindström et al. are shown. Color and labels indicate the classification by Lindström et al. See Fig 11.

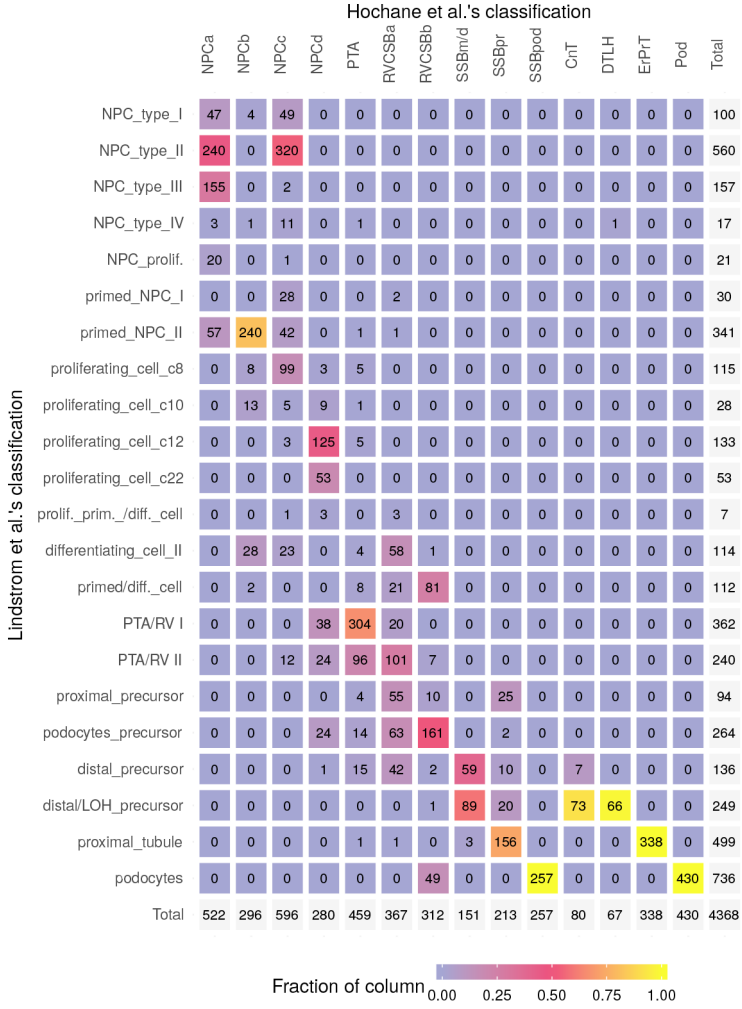


Figure 11. Comparison with an existing single-cell transcriptomics dataset showed differences in cell type distribution. Confusion matrix relating the cells measured in this study to the classification by Lindström et al. [20]. After batch correction, cells measured here were mapped on the cells in the Lindström et al. data set using a nearest neighbors-based approach (see Materials and methods). See Fig 10.

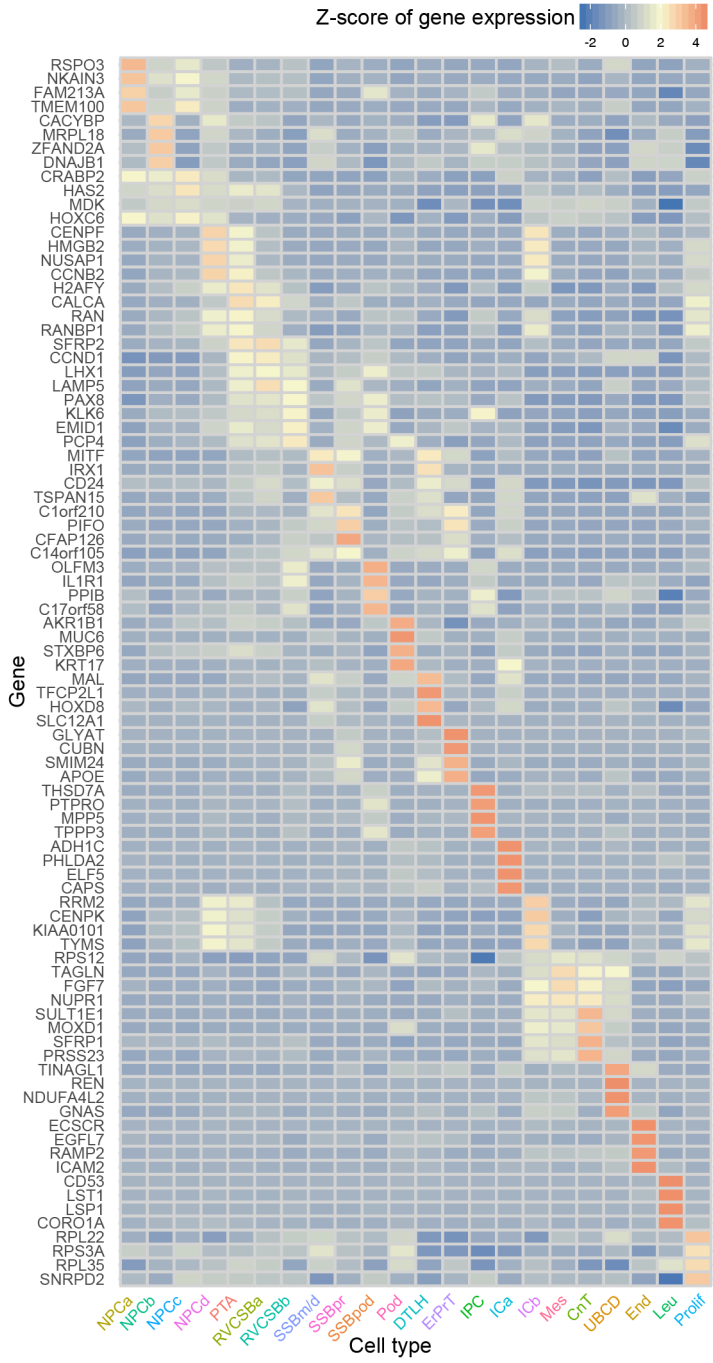


Figure 12. An ROC-based method identified novel marker genes. Expression heat map of the 88 genes identified by a method that evaluates the ROC for each gene (marker set, Hochane et al. [1] S3 Table). Expression was FT transformed, averaged over all cells in a cluster, and standardized gene-wise. See Fig 13.

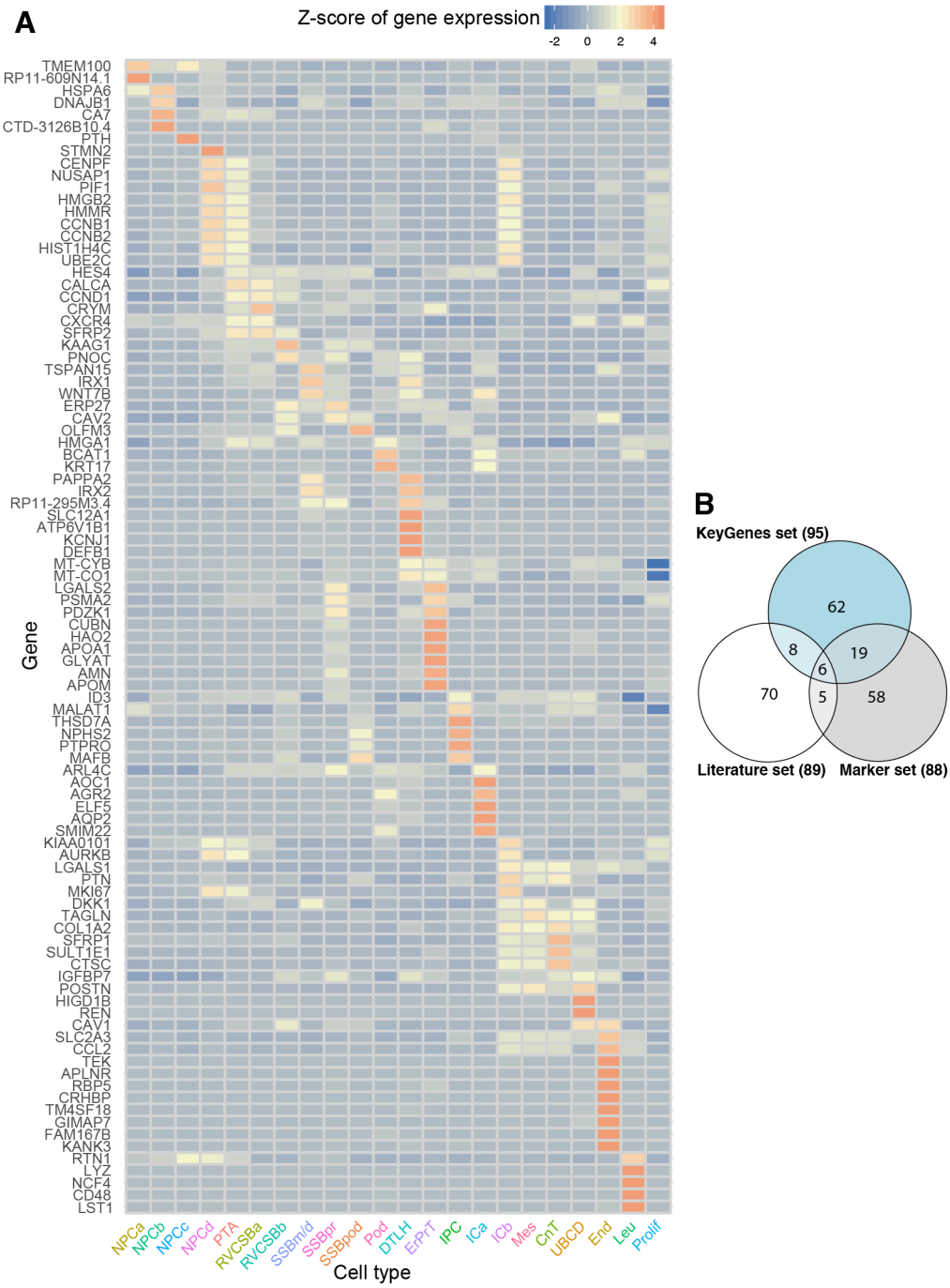


Figure 13. The KeyGenes method identified novel marker genes. (A) Expression heat map of the 95 genes identified by the KeyGenes algorithm (KeyGenes set, Hochane et al. [1] S3 Table). Expression was FT transformed, averaged over all cells in a cluster, and standardized gene-wise. (B) Euler diagram of the literature set, marker set, and KeyGenes set (Hochane et al. [1] S3 Table). See Fig 12.

Because the interpretation of the found cell clusters was largely based on markers identified in mouse development, we were wondering whether the new markers identified here were informative for the classification of cell types in the mouse kidney. Using a scRNA-seq measurement of cells from a whole P1 mouse kidney [23], we plotted the expression of the newly identified marker genes in single cells (Hochane et al. [1] S7 Fig). In many cases, markers that were found to label a particular cell type in the human fetal kidney were coexpressed in the same subset of mouse cells. A few markers, however, were either ubiquitously expressed or almost completely absent. This might be due to interspecies differences.

2.2.5 Comparison of different developmental ages

By establishing the identity of cell clusters at w16, we obtained a snapshot of cell type diversity in the fetal kidney. To explore whether the identified expression patterns change dynamically throughout development, we analyzed four additional samples from different developmental ages (w9, w11, w13, and w18), which together contained 11,359 usable cells. Using, again, batch correction based on mutual nearest neighbors [21], we visualized all samples in a common tSNE map (Fig 15, Fig 14). Overall, gene expression in the different samples was largely overlapping for the majority of cell types. For example, proximal tubules cells (ErPrT) appeared at the same positions in the tSNE map in all samples (Fig 15B). The position of podocytes shifted systematically across different ages, which corresponds to a continuing change in expression pattern (Fig 15C). This observation might suggest that podocytes further matured in terms of their expression pattern after being specified.

Differential expression analysis of podocytes of different ages revealed 109 differentially expressed genes (fold change > 2 in any comparison, FDR < 0.05, Hochane et al. [1] S4 Table). Functional annotation analysis of these genes showed significant enrichment of two gene ontology (GO) terms: *proteinaceous extracellular matrix* (adjusted p-value $= 1.9 \cdot 10^{-3}$, including *SPON2*, *BGN*, *COL1A2*, and *CTGF*) and *extracellular exosomes* (adjusted p-value $= 1.4 \cdot 10^{-3}$, including *NPNT*, *S100A10*, *ANXA1*, and *EPCAM*). Some of the differentially expressed genes have been shown to be important for kidney development. For example, *NPNT* and *DCN* showed increasing expression from w11 to w18. Knockout of the extracellular matrix protein NPNT in mice decreases the invasion of the UB and causes agenesis or hypoplasia [24]. *NPNT* was further shown to be expressed in the glomerular basement membrane and to be necessary for podocyte adhesion in mice [25]. Ablation of this gene in mice causes podocyte effacement. As in the case of *NPNT*, *DCN* has been reported to be part of the glomerular basement membrane proteins [26]. This gene appeared strongly up-regulated in podocytes between w11 and w13 or w18 (fold changes of 3.25 and 4.6, respectively). The increase of *NPNT* and *DCN* expression over time in our data set could reflect an increase in adhesion between podocytes and glomerular basement membrane. Podocytes further showed significant differential expressions of genes related to stress, like *HSPA1A* and *HSPA1B* or *NFKB* genes (*NFKB2*, *NFKBIA*, and *REL*), with the highest levels at w18. This might suggest that dissociation-related stress increases with age for podocytes, maybe related to stronger

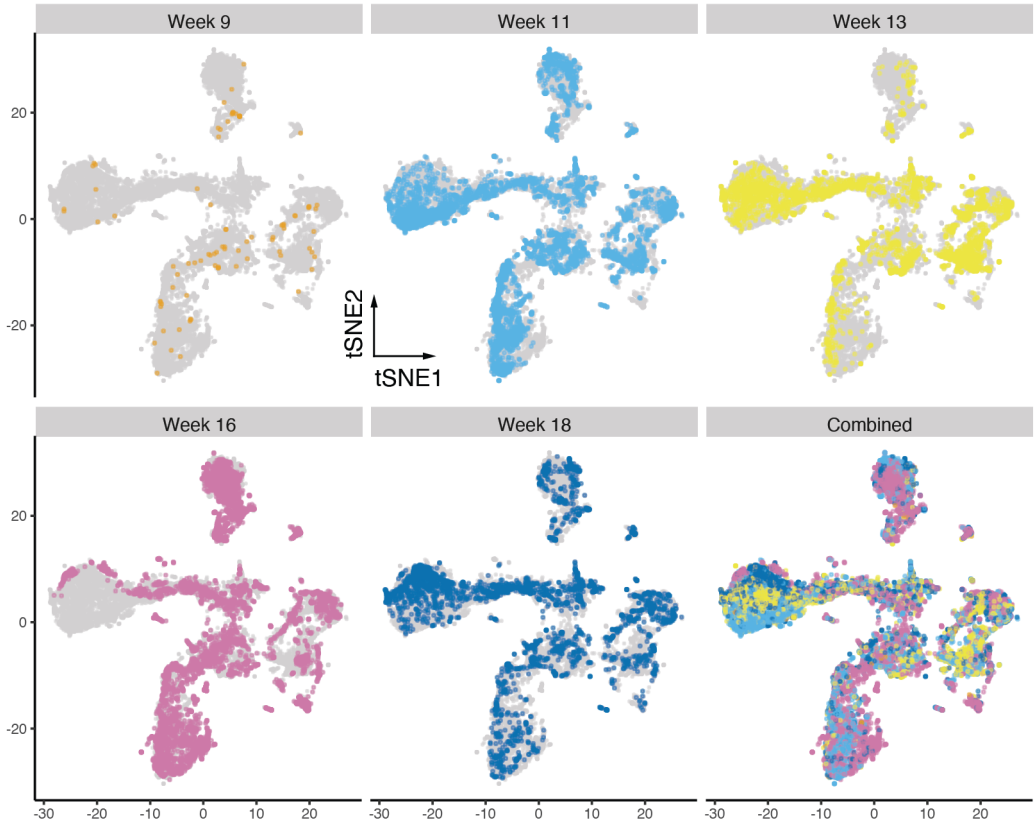


Figure 14. Samples of different developmental ages had a similar cell type diversity. tSNE map calculated for all five samples (w9, w11, w13, w16, w18) combined after batch correction [21]. Developmental age is indicated by color.

adhesion of the cells, or that stress-related genes have another, physiological role in development.

We would like to emphasize that the observed gene expression changes with age should be considered with caution because they might be related to the differences in genotype between the samples. A much larger number of samples would be necessary to rule out such interindividual differences as a cause.

Having established the identity of the cell clusters, we next wanted to demonstrate how the dataset can be used to explore different aspects of kidney development. We specifically focused on the nephrogenic niche, which showed pronounced heterogeneity, and the development of podocytes, which progressed via a distinct, intermediate cell state (SSBpod).

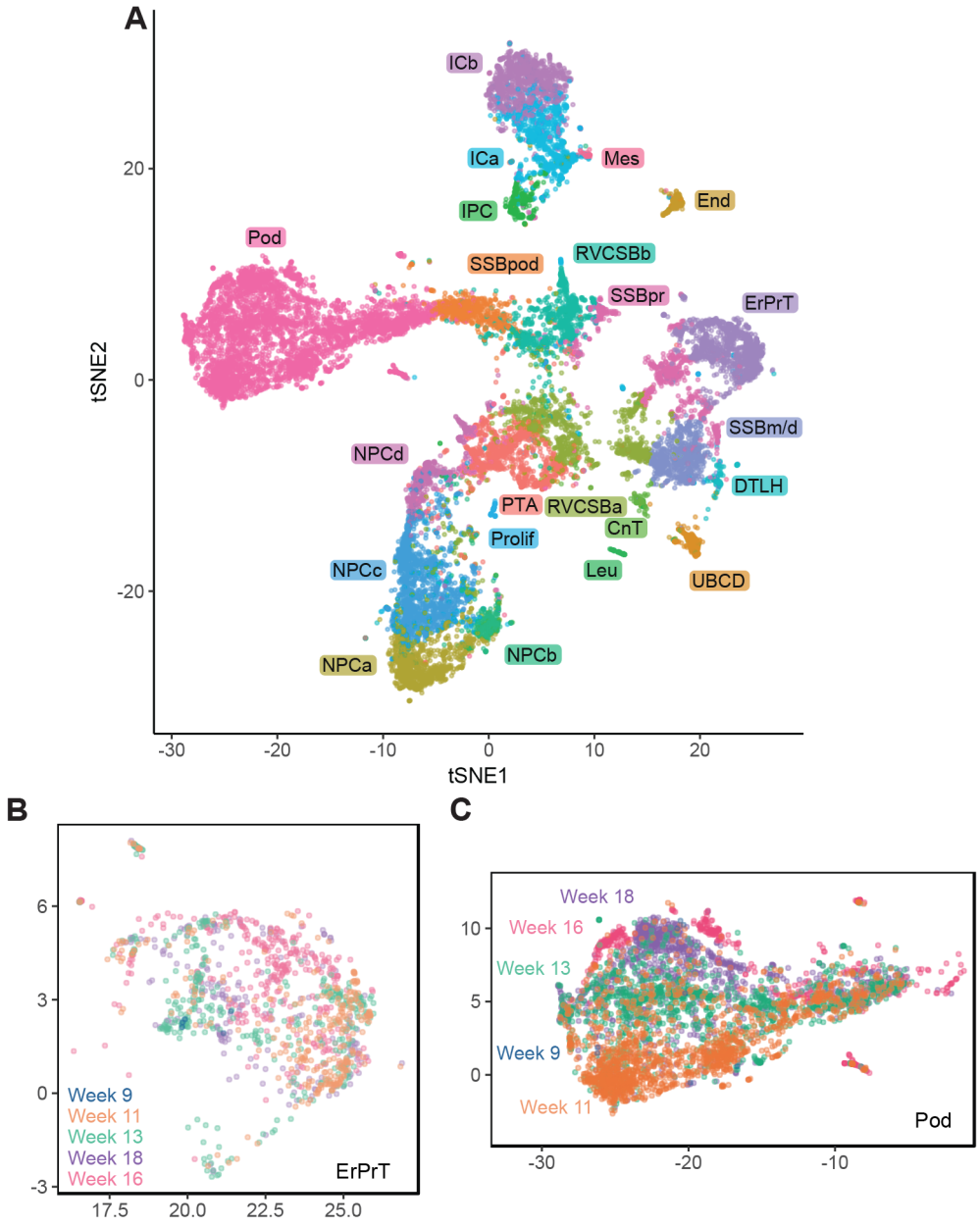


Figure 15. Comparison of different developmental ages suggested continued expression changes in podocytes. (A) tSNE map combining all five samples (w9, w11, w13, w16, w18). Samples were corrected for batch effects by matching mutual nearest neighbors [21]. Cells in the w9, w11, w13, and w18 samples were classified by comparing to the w16 sample using a knn-based approach (see Materials and methods). (B) tSNE map of all ages restricted to ErPrT. Labels and colors indicate ages. Six outlier cells were omitted from this plot to improve visualization. (C) tSNE map of all ages restricted to podocytes. Labels and colors indicate ages.

2.2.6 Heterogeneity in the nephrogenic niche

The formation of the nephron epithelium starts with the NPCs that differentiate and form the PTA, RV, and CSB. Studies in the mouse suggest that cells in the NPC compartment are not biased towards a particular lineage and patterning is first detectable in the PTA [27]. Nevertheless, the w16 scRNA-seq indicated the presence of several nephron progenitor subpopulations, NPCa-d. To clarify the temporal relationship of these clusters, we employed *Monocle 2* again to arrange them together with the PTA cells on a pseudotime scale (Fig 16A). NPCa clearly preceded NPCb and c, which seemed to appear around the same pseudotime. NPCd cells followed NPCb and c and preceded PTA. This analysis suggested that NPCa are the *bona fide* NPCs and give rise to NPCb and c. NPCd, which were likely more proliferative than the other NPCs (Fig 8), seemed to be an intermediate state between (slowly cycling) NPCa-c and the PTA.

To localize the NPC clusters in the tissue, we made use of the fact that they expressed various levels of *CITED1* and *SIX2* (Fig 7): although NPCa and NPCc exhibited roughly similar levels of these markers, NPCb and NPCd had lost *CITED1* almost completely, while retaining some *SIX2* expression. In an immunostaining of a w15 kidney, CITED1 and SIX2 appeared overlapping in a subset of cells (Fig 16C). Quantification of the fluorescence signal (see Materials and methods) revealed clear differences between their expression patterns. Although SIX2 expression was approximately constant throughout the CM, CITED1 expression decreased, relative to SIX2, with increasing (radial) distance from the UB (Fig 16D-E). A marked drop of CITED1 was visible between 10 and 20 μm from the UB, which approximately corresponds to the first layer of cells. To exclude that the observed difference between SIX2 and CITED1 expression was due to the different fluorophores on the secondary antibodies, we repeated the experiment with swapped fluorophores. This measurement produced a very similar expression gradient (Fig 16F). To exclude that the observed effect was influenced by PTA found in the CM towards the stalk of the UB, the analysis was repeated, taking only the 20% of CM cells closest to the edge of the cortex into account. A similar expression gradient was observed (Fig 16H). This result implies the existence of a CITED1 low/SIX2 high subpopulation of cells, which are not in contact with the UB. Secondly, we observed that CITED1 decreased relative to SIX2 towards the interface with the PTA and the stalk of the UB (Fig 16D-E). A similar observation was made when the experiment was repeated with swapped fluorophores (Fig 16G). Taken together, these results suggested that NPCa and NPCc were located closer to the surface of the UB and closer to the tip of the UB compared to the other NPC subtypes. Additionally, we also observed differences in subcellular localization of CITED1 protein within the CITED1 high compartment. Although for the majority of cells CITED1 was found in the cytoplasm, in several cells it was concentrated in the nucleus (right inset in Fig 16C). In contrast, SIX2 was always found restricted to the nucleus (left inset in Fig 16C). This observation might indicate that CITED1 was only active in a small population of cells, which would constitute another layer of cell-cell heterogeneity.

In addition to the observed heterogeneity in *CITED1* and *SIX2*, differences between the

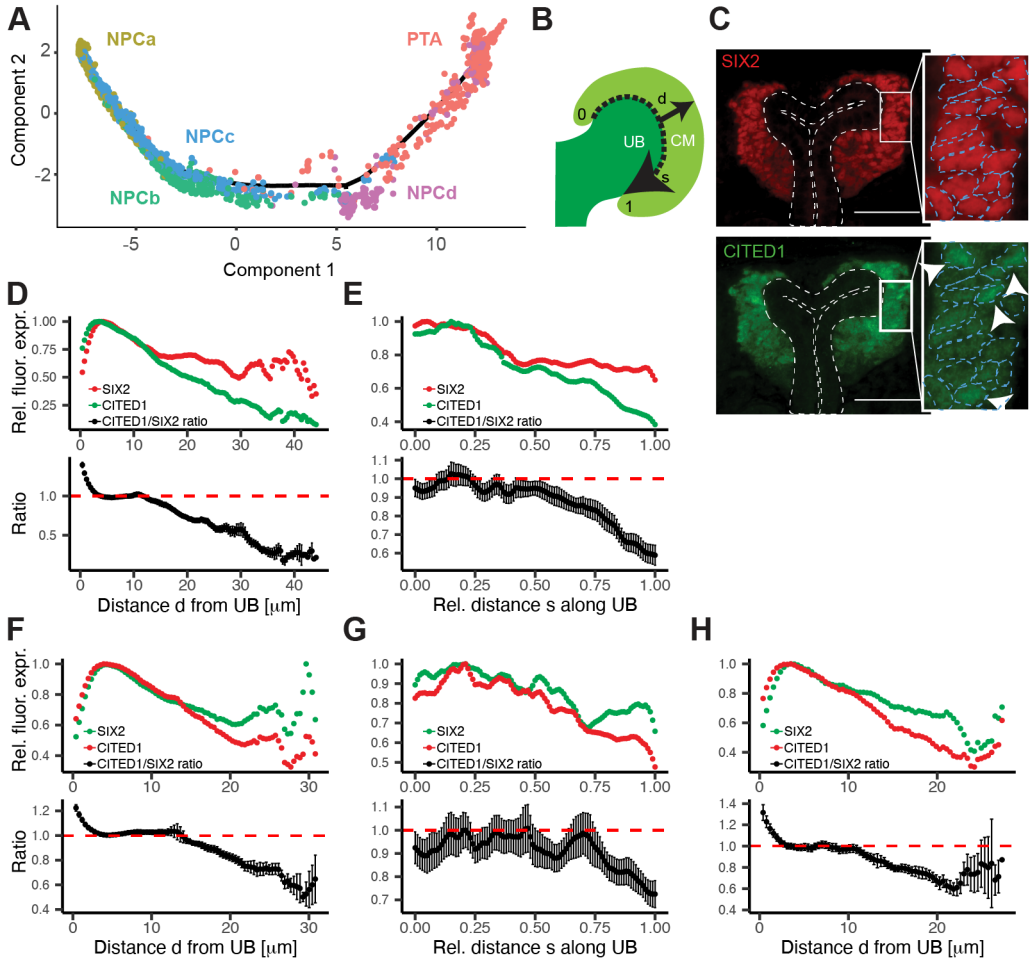


Figure 16. The nephrogenic niche exhibited a complex spatial organization. (A) Pseudotime analysis of the nephrogenic niche and the PTA. Two-dimensional *DDRTree* [19] embedding and the learned graph (shown as a black line) were calculated with *Monocle 2* [18]. Labels and colors indicate cell types. (B) Schematic sketch of the CM indicating the distance d from the UB to the edge of the CM (solid arrow) and the relative distance s along the UB (dashed arrow), in which 0 and 1 represent the top and bottom of the CM, respectively. (C) Representative image of SIX2 and CITED1 immunostaining in a w15 human fetal kidney. Dashed lines in the insets indicate the outline of the nuclei, based on DAPI signal. Arrows in the inset point to cells in which CITED1 is concentrated in the nucleus. Scale bar = 50μm. (D and E) Quantification of SIX2 and CITED1 immunostaining with respect to the distance d from UB or distance s along the UB; see panel A. Error bars indicate the SEM calculated over all evaluated profiles ($n = 24$). (F and G) Same as D and E, but the fluorophores on the secondary antibodies were swapped. ($n = 19$). (H) Quantification of SIX2 and CITED1 immunostaining with respect to the distance d from UB in which only cells with a relative distance s (along the UB) < 0.2 were taken into account. Error bars indicate the SEM calculated over all evaluated profiles ($n = 19$).

NPC clusters could also be gleaned from the set of novel markers (marker set, Hochane et al. [1] S3 Table).

TMEM100 and *RSPO3* specifically marked NPCa. *RSPO3* is an activator of the canonical WNT signaling pathway [28], suggesting a role of WNT either in NPC self-renewal or UB branching morphogenesis. Notably, all markers of NPCb (*CACYBP*, *MRPL18*, *ZFAND2A*, *DNAJB1*) were related to the stress response in some form. The markers of NPCc (*CRABP2*, *HAS2*, *MDK*, *HOXC6*), which were also expressed in the other NPC types, are all either targets of RA or binding it [17, 29, 30]. *MDK* has been shown to be expressed in the CM of the developing rat kidney, and its neutralization reduced the number of formed nephrons in vitro [31, 32]. Finally, the NPCd markers *CENPF*, *HMGB2*, *CCNB2*, and *NUSAP1* all have a role in cell cycle regulation or proliferation [16]. *HMGB2* was recently implied in the activation of quiescent adult neural stem cells [33].

The observation that markers of NPCb were related to the stress response seemed to suggest that this cluster was created as an artifact of cell dissociation [34], despite our best efforts to remove stressed cells (see Materials and methods). On the other hand, the vast majority of NPCb cells were classified as *primed NPC II* in the Lindström et al. [20] dataset (Fig 10, Fig 11). The fact that NPCb cells were only detected in the w16 and w18 kidneys is consistent with single-cell dissociation becoming increasingly difficult with fetal age, or alternatively, with a progenitor cell aging phenomenon. To explore the differences between NPCb and the other NPC clusters further, we immunostained HSPA1A and NR4A1, both known stress-response genes, in w15 kidney sections (Fig 17). *HSPA1A* was identified as a marker of NPCb (Fig 18A, Hochane et al. [1] S3 Table), whereas *NR4A1* was expressed in multiple NPC clusters but highest in NPCb (Fig 18A). Furthermore, *NR4A1* was also identified in the study by Adam et al. [23] to be up-regulated in response to elevated temperatures during enzymatic dissociation. HSPA1A and NR4A1 were both observable in the nephrogenic niche at the level of the stalk of the UB and at the transition to the PTA or RV. Additionally, we studied the expression of *EGR1*, another stress-related gene that marked NPCb, with smFISH (Fig 18B). *EGR1* was mainly found toward the stalk of the UB and in a few cells around the tip of the UB, whereas *SIX2* and *CITED1* transcripts were visible throughout the CM (Fig 19A). Because results obtained in fixed tissue sections are not confounded by dissociation-related artifacts, these immunostainings and smFISH measurements supported the existence of NPCb cells in the fetal kidney.

The fourth NPC cluster, NPCd, was clearly distinguished by proliferation markers (Fig 8). To locate NPCd in the tissue, we immunostained the cell cycle regulator and NPCd marker *CKS2* (Fig 18C). *CKS2* signal could be observed around the stalk of the UB and in RV (Fig 17). This result supported the interpretation that NPCd were a proliferating transitory state between NPCa-c and PTA.

Given the crucial role of the nephrogenic niche in the development of nephrons, it is likely that misexpression or mutation of genes that are specifically expressed in NPC affect kidney function. Mining a database of genome-wide association studies (GWAS) revealed that genes

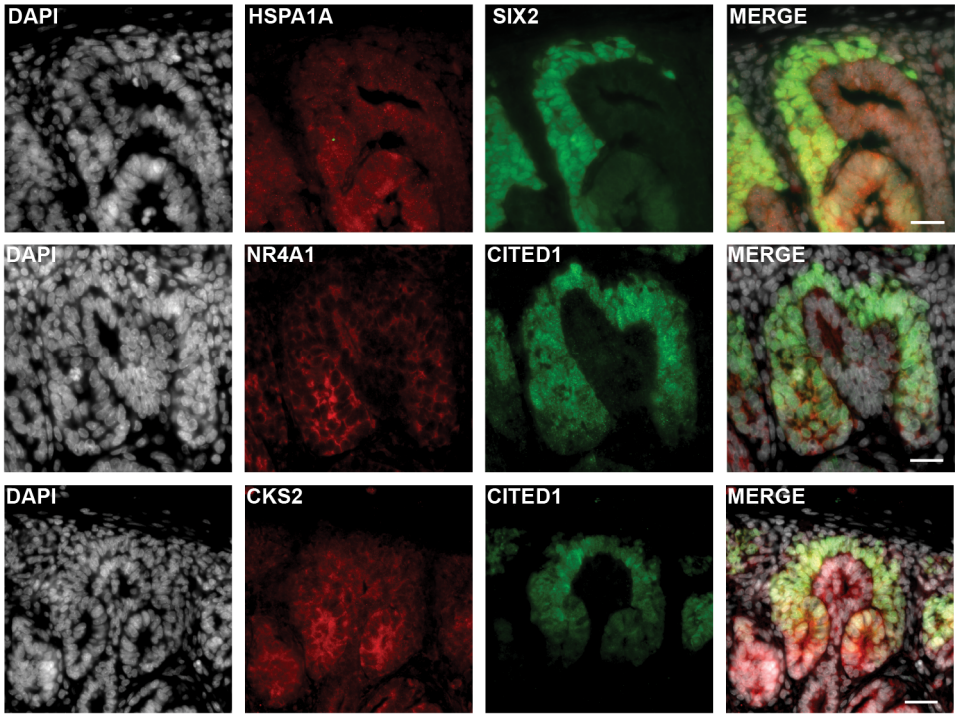


Figure 17. Expression variability in the nephrogenic niche. (Representative image of HSPA1A, NR4A1, and CKS2 immunostaining in a w15 human fetal kidney. Scale bar = 20μm.

that were differentially expressed in NPCs were significantly enriched for association with kidney disease (p-value = $1.7 \cdot 10^{-3}$, one-sided Fisher's exact test). No enrichment was found for lung diseases (p-value = 0.21), one-sided Fisher's exact test) (see Materials and methods, Fig 20A, Hochane et al. [1] S4 Table). Unsurprisingly, several of the disease-associated genes are known regulators of kidney development, such as *SALL1*, *SOX11*, and *HAS2* [35, 36, 37, 38]. The other identified genes had not been previously associated with kidney development. For example, *DDX1*, which was differentially expressed in NPCs as well as SSBpod, is an RNA helicase that promotes microRNA maturation [39]. *UNCX*, which was broadly expressed in all NPC clusters, is a homeobox transcription factor involved in somitogenesis and neurogenesis [40] and has also been found to be up-regulated in the induced mouse nephrogenic mesenchyme in culture [41]. It was recently associated with renal-function-related traits [42] as well as glomerular filtration rate [43, 44, 45]. In our data, the expression profile of *UNCX* was similar to that of *CITED1* (Fig 18D). Immunostaining of UNCX confirmed the scRNA-seq results and showed expression of UNCX in the nephrogenic zone, as marked by CITED1 (Fig 19C-D). These findings suggested *UNCX* as a novel potential regulator of early nephrogenesis.

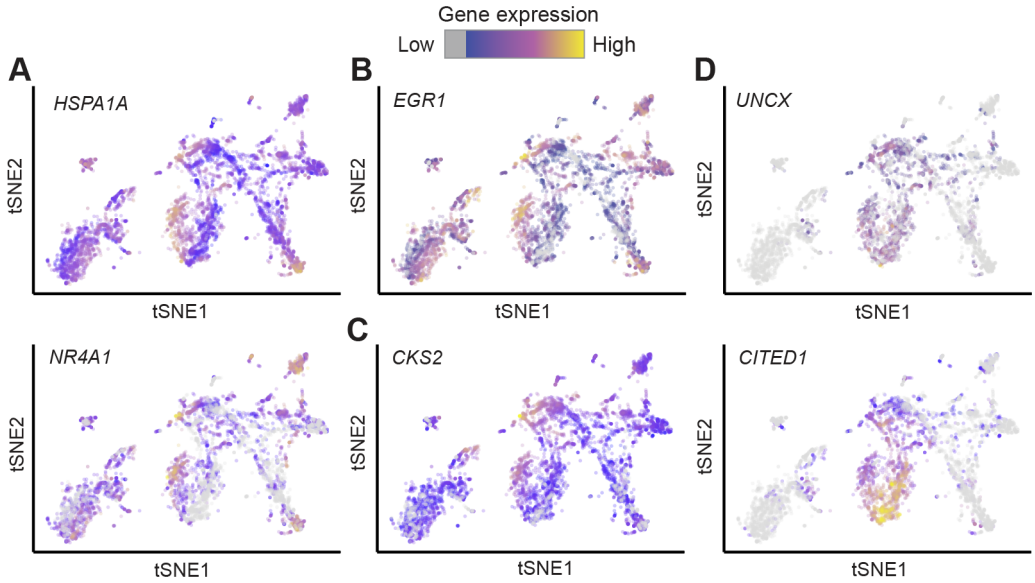


Figure 18. Markers of heterogeneity in the nephrogenic niche. (A [top], A [bottom], B, C, D [top] and D [bottom]) tSNE maps showing expression of *HSPA1A*, *NR4A1*, *EGR1*, *CKS2*, *CITED1* and *UNCX* respectively. Expression is indicated by color; expression values of 1 are plotted in gray.

2.2.7 Podocyte development

Another cell type of high relevance for kidney function is the podocyte. This cell type is critical for filtration and is implied in several forms of kidney disease [25]. podocyte (Pod)ocytes wrap around the glomerular basement membrane (capillary bed) inside Bowman's capsule (Fig 21A). Clustering (Fig 3) and pseudotime analysis (Fig 9) of the w16 kidney dataset had indicated that development into podocytes occurs via a distinct intermediate state that we dubbed SSBpod here. This cell state was likely related or even identical to previously discovered podocyte precursors [20, 46]. In the Lindström et al. [20] dataset, SSBpod and podocytes were both classified as *podocytes*, and the RVCSB were considered *podocyte precursors* (Fig 11). To show that SSBpod cells were indeed a localizable cell state distinct from podocytes, we further investigated their expression pattern (Fig 21B), focusing on known literature markers (literature set) and the marker set (Hochane et al. [1] S3 Table). Compared to RVCSB, SSBpod showed higher expression of *MAFB* and *FOXC2*, which are necessary for the determination of podocyte identity [47, 48]. On the other hand, compared to podocyte cells, they exhibited lower expression of genes typically associated with more mature podocytes, like *CLIC5*, *PODXL*, and *PTPRO*. Filtration function-related genes like *NPHS1*, *NPHS2*, and *PTPRO* were expressed at intermediate levels in SSBpod compared to RVCSBb, where they were absent, and podocytes, where they are highly expressed. A similar pattern could be

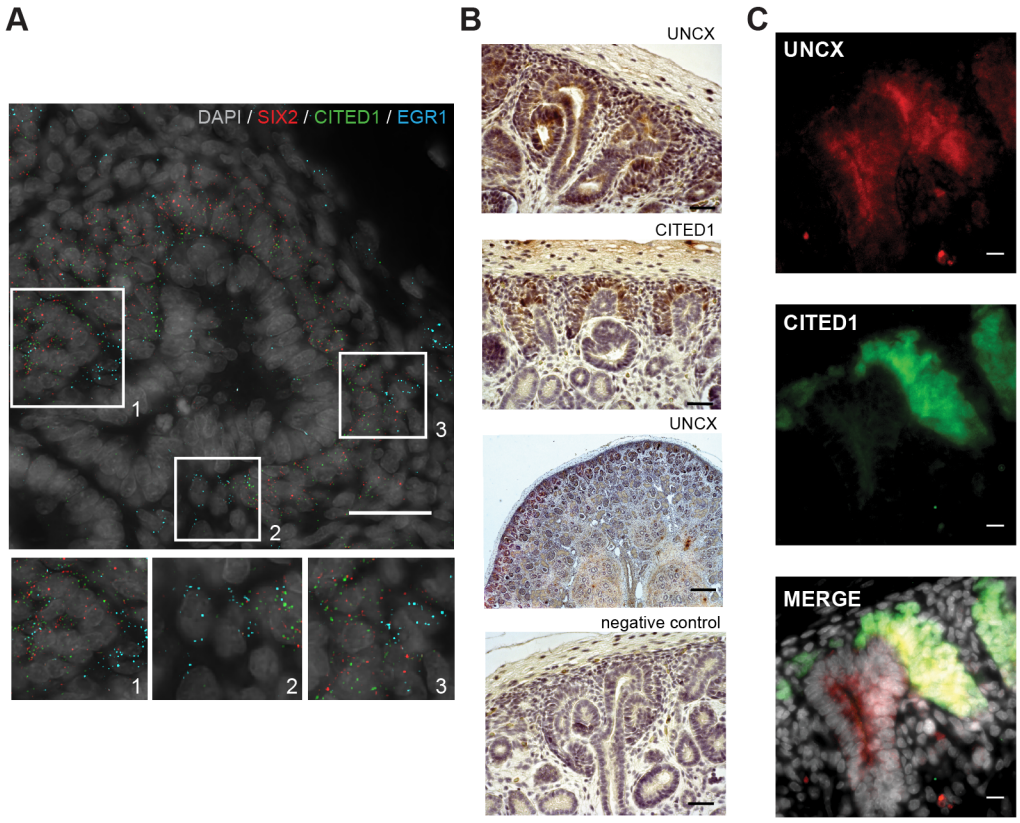


Figure 19. In situ expression of NPC markers. (A)) smFISH of *SIX2*, *CITED1*, and *EGR1*. The three insets at the bottom correspond to the three areas marked by solid boxes in the main image. Scale bar = 25μm. (B) Representative image of UNCX and CITED1 immunostaining. Arrowheads indicate the presence of immunostaining signal. Scale bar = 100μm. (C) Immunostaining of CITED1 and UNCX. Scale bar = 10μm.

observed for genes associated with podocyte polarization or structural organization as well as pedicel growth and patterning. Finally, podocytes showed the expression of genes that negatively regulate the cell cycle and support long term survival, consistent with their post-mitotic nature [49]. In contrast, SSBpod specifically expressed *ORC4*, which has a function in DNA replication. However, proliferation markers were lowly expressed in both SSBpod and podocytes (Fig 8C), which suggested low proliferative potential in both cell types. In contrast to NPCs, association with kidney disease was not significantly enriched among genes differentially expressed in SSBpod (p-value = 0.1, one-sided Fisher's exact test). One of the disease-associated genes was *OLFM3*, which has been associated with glomerular filtration rate (Fig 20B) [50]. *OLFM3*, a secreted glycoprotein, has a known function in brain and retina development [51] and has been identified as a marker for podocyte precursors in two independent studies [20, 46]. In our dataset, it was specifically expressed in SSBpod (Fig 21C)

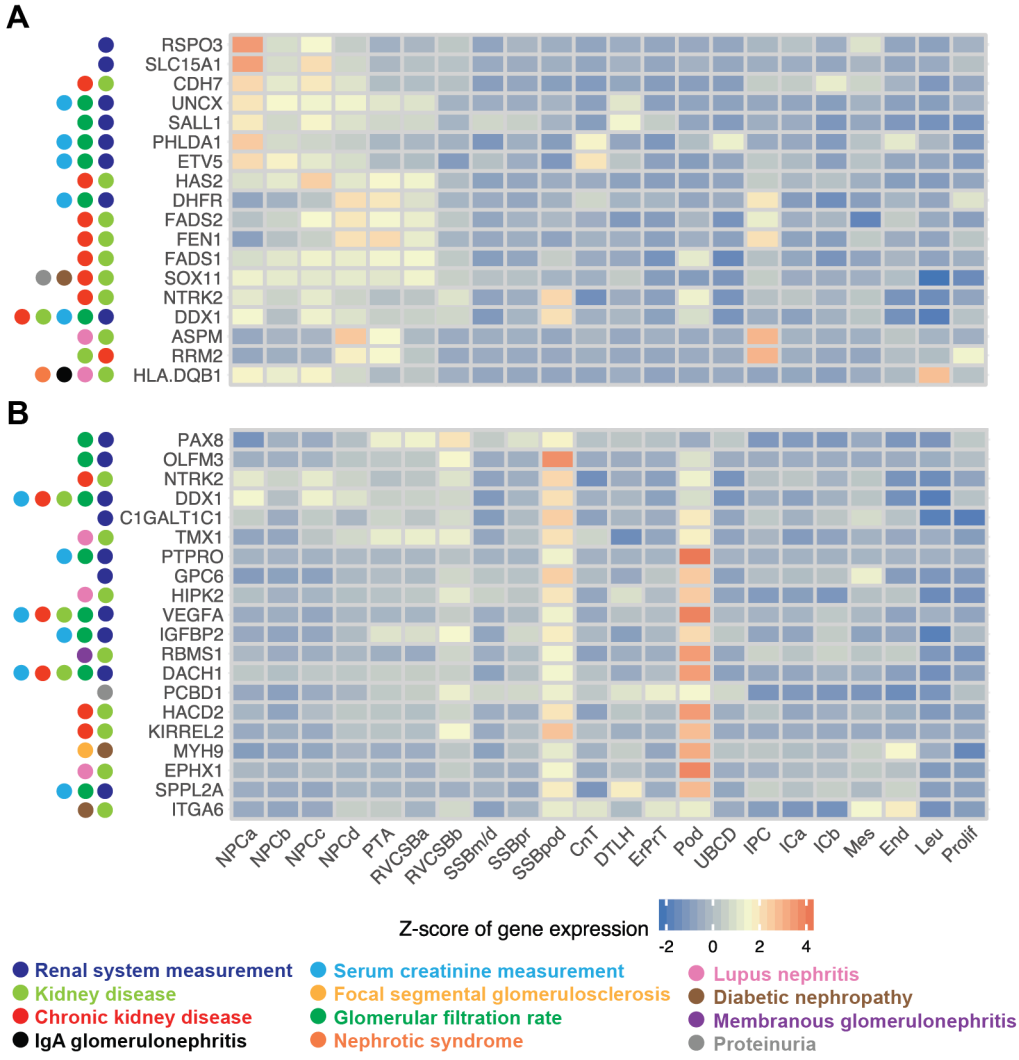


Figure 20. Disease-associated genes were specifically expressed in transient cell types. Expression of genes from GWAS traits related to kidney disease. Disease phenotypes associated with these genes are indicated by color; genes were filtered for high expression in cluster(s) of interest relative to all other cell types. Expression was FT transformed, averaged over all cells in a cluster, and standardized gene-wise. (A) Disease-associated genes expressed in early nephron progenitor states (NPC to PTA). (B) Disease-associated genes expressed in SSBpod.

and was a marker for this cell type in the marker set and KeyGenes set (Hochane et al. [1] S3 Table).

In order to localize SSBpod, podocytes and mesangial cells in situ, we immunostained w15 kidney sections with antibodies for MAFB, PODXL, and ACTA2 (Fig 21D). As expected, PODXL and MAFB were found in podocytes at the capillary loop stage and in more mature glomeruli. MAFB staining extended to the proximal segment of the SSB, which indicated that SSBpod may be part of this structure. To locate the SSBpod cells more precisely, we performed smFISH on *CLIC5* and *MAFB*, expressed both in podocytes and SSBpod (Fig 22A). We observed a subpopulation of *MAFB+/CLIC5-* cells outside the glomeruli, which we identified as the SSBpod. These cells could be found predominantly in the visceral part of the proximal segment of the SSB but also at the capillary loop stage. This result supported the notion that SSBpod were transient cells that preceded (mature) podocytes. Having localized the SSBpod, we next wanted to confirm *OLFM3* as a marker of this cell type. smFISH of *OLFM3*, *MAFB*, and *CLIC5* showed *OLFM3* to be coexpressed with *MAFB* but absent in cells that were positive for *CLIC5* (Fig 22A), a marker that persists in podocytes in the adult kidney. Quantification of the density of smFISH signals (Fig 22B) showed that *OLFM3* was absent in glomeruli but could be detected in the subpopulation we identified as SSBpod (*MAFB+/CLIC5-*). In summary, these results supported *OLFM3* as a robust marker of podocyte precursors.

Finally, we were wondering whether our dataset would also allow us to identify candidate mechanisms that drive development from SSBpod to podocytes. Differential expression analysis revealed 228 genes that had a significant, bigger than 2-fold changes between SSBpod and podocytes (Fig 23, Hochane et al. [1] S4 Table). Among these we found factors belonging to multiple signaling pathways, such as *FGF1*, *VEGFA*, *HES1*, and *EGF1*. *Vegfa* and *Fgf1* are known to have a homeostatic function in podocytes [52, 53, 54], whereas *Hes1*, a target of the Notch signaling pathway, seems to be necessary for the synthesis of extracellular matrix proteins in these cells [55]. Binding sites for the transcription factor AP-1 were strongly enriched in this set of genes (145 out of 228 genes, adjusted p-value $= 1.3 \cdot 10^{-5}$). *AP-1* would therefore be an interesting target for perturbation studies in mouse models.

All in all, the results presented here complement other, recent, single-cell transcriptomics studies of the fetal kidney. We demonstrated how the data can be interrogated to find expression patterns that will improve our understanding of human kidney development.

2.3 Discussion

2.3.1 The nephrogenic niche is heterogeneous

Building an organ during development requires the careful balance between two fundamental processes—growth and the creation of structure. In many organs, these two functions are reconciled by self-renewing progenitor cells that can be induced to differentiate. In kidney development, NPCs give rise to the epithelium of the nephron, the functional unit of the kid-

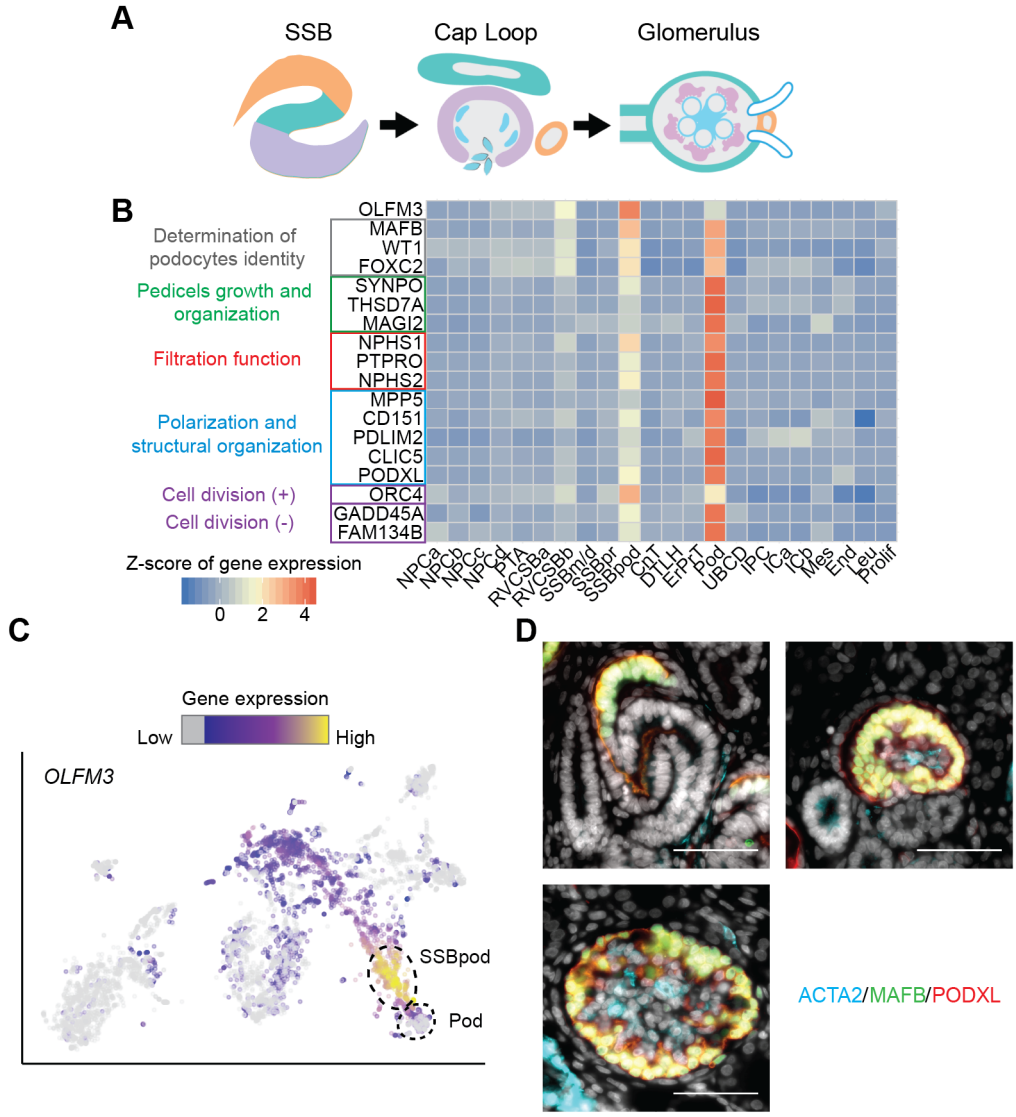


Figure 21. Pod develop via a precursor state. (A) Schematic of development from the SSB to the glomerulus. Regional patterning is shown in color—proximal (purple), medial (green), distal (orange). (B) Expression heat map of the subset of marker set genes that are markers for SSBpod and Pod. Expression was FT transformed, averaged over all cells in a cluster and standardized gene-wise. (C) Two-dimensional tSNE map showing the expression of *OLFM3*. Expression is indicated by color; expression values of 1 are plotted in gray. (D) Typical images of MAFB, PODXL, and ACTA2 immunostaining in SSB and glomeruli. Scale bar = 50μm.

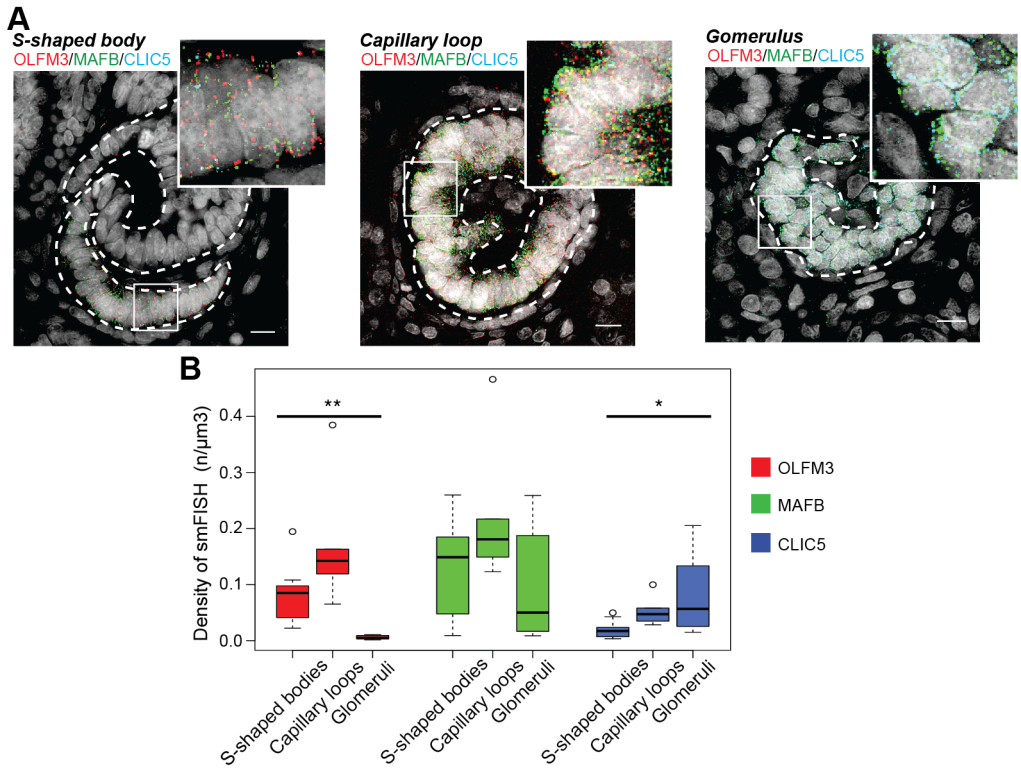


Figure 22. SSBpod is localized in the visceral proximal SSB. (A) Representative images of smFISH of *OLFM3*, *MAFB*, and *CLIC5* in SSBpod and Pod. w15 female kidney. Scale bar = 10µm. (B) Box plots of smFISH signal densities in SSB ($n = 10$), capillary loop ($n = 4$), and glomeruli ($n = 8$), for *OLFM3*, *MAFB* and *CLIC5* (* adjusted $p < 0.05$, ** adjusted $p < 0.0005$).

ney. To balance growth with patterning, self-renewal and differentiation of NPCs have to be tightly controlled. It is well established that the niche of the NPC plays an important role in this control, but the precise mechanisms are not well understood. In particular, it is not clear how the position and movement of NPCs in the niche might impact the induction towards differentiation.

Heterogeneity in the nephrogenic niche was brought to light first by Mugford et al. [27] in 2009 and has been confirmed by multiple recent studies [11, 20, 46, 56, 57]. Mugford et al. [27] used in situ hybridization to study the localization of transcriptional regulators in E15.5 mouse kidney. Three distinct compartments were defined in the CM—inner capping mesenchyme (which lies closest to the cleft of the UB), outer capping mesenchyme (at the tip of the UB), and induced mesenchyme (at the level of the stalk of the UB). Although all compartments express *Six2*, only inner and outer capping mesenchyme express *Cited1*. The induced CM was distinguished by Wnt pathway activity, as evidenced by *Wnt4* expression. Several recent scRNA-seq studies confirmed heterogeneity in the nephrogenic niche. Brunskill et al.

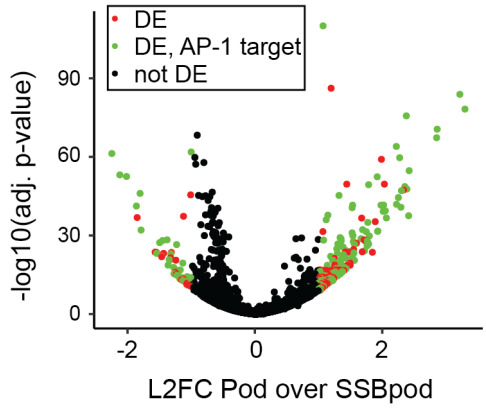


Figure 23. AP-1 targets are enriched in the DEG of Pod versus SSBpod. Volcano plot of differential gene expression between SSBpod and Pod. L2FC Pod over SSBpod versus $-\log_{10}$(adjusted p-value). Genes with an adjusted $p < 0.05$ and L2FC > 1 were considered significant (colored data points). Genes with an AP-1 binding site are shown in red.

[56] studied an E12.5 mouse kidney and found two subpopulations in the CM, which they classified as uninduced (*Six2* positive, *Cited1* positive) and induced (*Six2* positive, *Cited1* negative). Among the hundreds of genes that were differentially expressed between these two populations, they found genes related to the Wnt signaling pathway as well as protein vesicular trafficking and degradation. Wang et al. [57] also found two subclusters in the CM of the human fetal kidney. They interpreted one subcluster as the self-renewing compartment due to higher expression of markers for cell division. The other subcluster, which showed gene expression related to Notch signaling (*HES1*, *HEY1*), was considered induced. Two studies by [11, 20] also explored NPC heterogeneity. The first study [11] identified four NPC clusters (self-renewing, primed, differentiating, and proliferating), whereas the second [20] revealed four clusters of NPCs (I-IV), two clusters of primed NPCs (I-II), as well as several clusters of proliferating cells.

In the dataset presented here, we identified four clusters of NPCs. Among these, NPCa are most likely the self-renewing compartment. In agreement with the studies by Lindström et al. [20] [11, 20], they expressed the highest levels of *CITED1* and *TMEM100* compared with the other NPCs. Furthermore, they preceded all other NPC clusters in pseudotime analysis. NPCb showed expression of several genes that modulate Notch, BMP, and TGF-β pathway activity, as well as low levels of *LEF1*, which has been shown to indicate induction towards differentiation [11, 20]. The classification of NPCb as *primed NPC* by comparison to the Lindström et al. [20] dataset supported the interpretation of NPCb as a state distinct from NPCa that is primed to differentiate. The fact that we detected NPCb only at w16 and w18 leads us to speculate that NPCb could be the result of continuous changes in the nephrogenic niche over the course of development. The third NPC cluster, NPCc, appeared together with NPCb

in pseudotime and was distinguished from the other NPCs by higher expression of genes involved in or regulated by RA signaling. The RA binding protein CRABP2 has been identified as an NPC marker in other reports [11, 57]. We speculate that NPCc are the result of spatially varying concentrations of RA, which is produced in the cortical interstitium [58]. Finally, NPCd appeared between NPCb-c and PTA in pseudotime and were clearly distinguished from the other NPC clusters by increased proliferation, at least as far as that can be inferred from gene expression data. In agreement with our analysis, NPCd were classified as *proliferating cells* by comparison with the Lindström et al. [20] dataset. NPCd cells also lowly expressed markers of induction towards differentiation (such as *LEF1*, *LHX1*, *WNT4*), which indicates a transitory state between induced and/or primed NPCs and PTA. The suggested developmental flow from NPCa via NPCb-c to NPCd was supported by a gradual decrease of *OSR1*, which is a well-known marker of the early CM.

By in situ detection of CITED1, SIX2, and other genes, we also explored the spatial localization of the different NPC clusters. NPCa seemed to reside closest to the tip of the UB, the induced and/or primed NPCb and c were situated closer to the stalk, and NPCd were closest to the PTA. This finding is consistent with the recent report of NPCs streaming from their niche at the UB tip towards the UB branch point to form the PTA and RV [20]. On this path, the cells gradually lose the NPC transcriptional program, and differentiation is induced. In the mouse, trajectories of NPCs also seem to have a large stochastic component: NPCs repeatedly detach from the UB and attach again and also shuttle back and forth between the *uninduced region* at the UB tip and the *committed region* around the stalk of the UB [59]. This observation could indicate that varying expression levels of genes such as *CITED1* occur as a consequence of cell migration and are not necessarily functionally relevant. Indeed, *Cited1* knockout has no adverse effects on kidney development in the mouse [60]. Taken together, our results support a model in which (self-renewing) NPCa reside at the tip of the UB, probably in close proximity or even in contact with the UB. Movement away from the UB tip, toward the stalk, is accompanied by decreased *CITED1* expression and transformation to the (induced and/or primed) NPCb-c states. Arrival at the stalk of the UB is characterized by the NPCd expression state, increased proliferation, and eventually transformation to the PTA. It is conceivable that cells sometimes visit the different NPC states in reverse order, which would reconcile this model with the observed high, multidirectional motility of NPCs [59].

2.3.2 Proximal-distal patterning

In the prevailing model of mammalian kidney development, self-renewing NPCs are not prepatterned to develop into a certain lineage. When the developing nephron first displays signs of proximal-distal patterning is an important, outstanding question. Mugford et al. [27] have found evidence that the PTA, which succeeds the NPC, is already polarized. A recent study by Lindström et al. [20] proposed an intriguing mechanism that couples temporal and spatial cues: although NPCs that are recruited to the PTA early develop into distal cell types, NPCs that are integrated later contribute to the proximal compartment. In our study,

we identified the PTA by known marker genes (*CCND1*, *LHX1*, and *WNT4*) and high proliferation. We were unable to detect any substructure within the PTA, which might be due to the limited resolution of our scRNA-seq method. RV/CSB, the next developmental stage, however, was split in two clusters (a and b). Pseudotime analysis suggested that RVCSBa was a heterogeneous cluster comprising early RV cells (which appeared before RVCSBb) and the distal segment of the RVCSB. This observation is consistent with the time-dependent recruitment model by Lindström et al. [20] in the sense that in that model, distal specification precedes proximal patterning.

2.3.3 Fetal podocytes may have varying degree of maturation

Single-cell transcriptomics studies of various organs have brought to light many new, intermediate cell states. This has provoked the question of whether we should consider gene expression in complex tissues as a continuum rather than a collection of distinct expression profiles. In developmental systems, it is certainly useful to think about gene expression change as a continuous process. Nevertheless, there are clearly distinct intermediate cell states even within linear developmental paths. In our study, we observed that RVCSBb gave rise to podocytes via an intermediate state, the SSBpod, which directly preceded the podocytes in pseudotime. Specific expression of *OLFM3* made it likely that this cluster is identical to previously identified podocyte precursors, which were marked by this gene [20, 46]. Menon et al. [46] defined a cluster of *immature* or *early* podocytes, characterized by high *OLFM3* and low *MAFB* expression. In that study, podocytes showed increased *MAFB* expression but loss of *OLFM3*. Lindström et al. [20] located *OLFM3* positive cells to the proximal part of the SSB. In our study, we confirmed all of these observations: *OLFM3* was localized to the visceral part of the proximal segment of the SSB, and *OLFM3* negative podocytes showed higher expression of mature podocyte markers compared to the *OLFM3* positive SSBpod. Functional annotation analysis of genes that were differentially expressed between SSBpod and podocytes revealed enrichment of a binding site for AP-1. This transcription factor has been found to be important for the development of the skin [61], neural precursor cells [62], and the heart valve [63] in mice. A role of *AP-1* in kidney development has not been described yet, and further research is needed to elucidate its potential function. The analysis of the kidneys from different gestational ages showed high similarity of cell types across different ages with the exception of podocytes. These displayed a systematic change in expression pattern, which might indicate the continued maturation of podocytes over time. This observation is in agreement with a study by Brunskill et al. [64] in the mouse, which compared embryonic (E13.5 and E15.5) with adult podocytes (defined as *Mafb* positive cells). That study found hundreds of genes that were differentially expressed between embryonic and adult podocytes. Furthermore, targeted experiments are needed to demonstrate the possible maturation of podocytes in human kidney development.

In summary, we have leveraged a combination of single-cell transcriptomics and in situ imaging to study the intricate structure of the developing human kidney. The transcrip-

tomics data, accessible via a web application http://www.semraulab.com/kidney, will be a valuable starting point for discovering gene regulatory mechanisms or finding new disease mechanisms.

2.4 Materials and methods

2.4.1 Ethics statement

The collection and use of human material in this study was approved by the Medical Ethics Committee from the Leiden University Medical Center (P08.087). The gestational age was determined by ultrasonography, and the tissue was obtained by vacuum aspiration from women undergoing elective abortion. The material from six embryos (w9, male; w11, male; w13, female; w15, female; w16, male; and w18, female) was donated with written informed consent. Questions about the human material should be directed to S. M. Chuva de Sousa Lopes (Lopes@lumc.nl).

2.4.2 Experimental methods

Single-cell dissociation of human fetal kidney

One human embryo of w16 (male) was isolated and the kidney dissected in cold saline solution (0.9% NaCl, Versylene Fresenius). For sex genotyping, polymerase chain reaction (PCR) for AMELX/Y was used as previously described [65]. The obtained kidney was decapsulated and kept on ice in dissociation buffer (DPBS + Penicillin 100U/mL + Streptomycin 0.1mg/mL; all from Life Technologies) before cutting it into 1-2mm pieces. The pieces were washed three times with washing solution (Advanced DMEM F12 supplemented with ITS commercial solution (Insulin-Transferrin-Selenium; Thermofisher), Glutamax, Penicillin 100U/mL, and Streptomycin 0.1mg/mL) with brief centrifugation (160g) in order to remove as many red blood cells as possible. The washed kidney tissue was then incubated with digestion solution (Trypsin/EDTA solution 0.25% and Collagenase-II 280U/ml) and incubated overnight at 4°C. The next day, the digestion solution was removed, and the kidney was rinsed with washing solution and incubated with washing solution for 30min at 37°C with agitation. Subsequently, the sample was sequentially passed through sterile cell strainers of 100, 70, and 40μm pore size with the help of washing solution. The cells were then centrifuged and counted, and viability was measured to be 78% (trypan blue assay) before proceeding with scRNA-seq library preparation. Four additional human fetal kidneys (w9, male; w11, male; w13, female; and w18, female) were dissected as described above, but, additionally, live cells were purified by FACS before library preparation [66].

scRNA-seq library preparation and sequencing

scRNA-seq libraries were prepared using the Chromium Single Cell 3' Reagent Kit, Version 2 Chemistry (10X Genomics) according to the manufacturer's protocol. Libraries were sequenced on a NextSeq500 in Mid Output mode using a version 2, 150-cycle kit (Illumina).

Immunostaining

A paraffin-embedded w15 human kidney (female) was sectioned (5μm) using a RM2255 microtome (Leica Microsystems GmbH) and mounted on StarFrost slides (Waldemar Knittel).

For immunofluorescence, sections were deparaffinized and rehydrated by standard procedures, starting with xylene (twice for 20min), followed by ethanol with sequential dilution and ending with distilled water, all at room temperature (RT). Antigen retrieval was performed by a double treatment of 10min in a microwave (97°C) with 0.01M sodium citrate buffer (pH 6.0). The sample was then allowed to cool down, rinsed three times with PBS, and blocked for 1h at RT in blocking buffer (1% BSA, 0.05% Tween-20 in PBS). Subsequently, sections were incubated overnight with the following antibodies diluted in blocking buffer—rabbit anti-SIX2 (1:100, 11562-1-AP;Proteintech), mouse anti-CITED1 (1:500, H00004435-M03; Novus Biologicals), mouse anti-MAFB (1:200, LS-C336952; LifeSpan Biosciences), rabbit anti-ACTA2 (1:200, ab5694; Abcam), goat anti-PODXL (1:200, AF1658; R&D Systems), and rabbit anti-UNCX (1:10, PA5-69485; Thermo Fisher Scientific), rabbit anti-CKS2 (HPA003424, 1:100; Sigma Aldrich), rabbit anti-NURR77 (NR4A1) (ab13851, 1:50; Abcam Biochemicals), and mouse anti-HSP70 (HSPA1A) (ab2787, 1:50; Abcam Biochemicals). The secondary antibodies were diluted in blocking buffer and applied at RT for 1h followed by nuclear counterstaining with 4′,6-diamidino-2-phenylindole (DAPI; Life Technologies). The secondary antibodies used were Alexa Fluor 647 donkey anti-rabbit (1:500, A-31573; Life Technologies), Alexa Fluor 594 donkey anti-mouse (1:500, A-21203; Life Technologies), and Alexa Fluor 555 donkey anti-goat (1:500, A32727; Life Technologies). The sections were then mounted using ProLong Gold (Life Technologies).

For immunohistochemistry, sections were deparaffinized and blocked as above. After overnight incubation with primary antibodies rabbit anti-UNCX (1:10, PA5-69485; Thermo Fisher Scientific) and mouse anti-CITED1 (1:500, H00004435-M03; Novus Biologicals) in blocking buffer, 0.3% H_2O_2 was used to quench endogenous peroxidase activity for 20min. Next, the sections were incubated with biotin-labeled goat anti-rabbit IgG (1:200, BA-1000; Vector Laboratories) diluted in normal goat serum (1:66, S-1000; Vector Laboratories) or biotin-labeled horse anti-mouse (1:200, BA-2000; Vector Laboratories) diluted in normal horse serum (1:66, S-2000; Vector Laboratories) for 40min. Sections were then treated for 40min with avidin-biotin-peroxidase complex (VECTASTAIN Elite ABC HRP Kit, #PK-6100; Vector Laboratories) following the manufacturer's instructions, followed by DAB (D5637; Sigma-Aldrich) and hematoxylin (1043020025; Merck) and were mounted with Entellan (1079610100; Merck).

Single-molecule FISH

Paraffin embedded sections from the w15 human fetal kidney (female) used for immunostaining were also used for smFISH experiments. Paraffin was removed by immersion in xylene twice for 10min at RT. The sections were then rehydrated by sequential immersion in ethanol solutions—100% (2x, 10min), 85% (2x, 5min), and 70% (2x, 3min). Subsequently, sections were permeabilized in 70% ethanol for 5h before incubation with proteinase-K (P4850; Sigma Aldrich) for 15min at 37°C (23μg/mL in TE buffer at pH = 8) and a wash in RNAse-free water (3x, 5min). smFISH was performed as described previously [67]. Briefly, custom designed smFISH probes (BioCat,Hochane et al. [1] S5 Table), labeled with Quasar 570, CAL FLuor Red 610, or Quasar 670, were incubated with the samples for 16h at 30°C in hybridization buffer (100 mg/mL dextran sulfate, 25% formamide, 2X SSC, 1mg/mL E.coli tRNA, 1mM vanadyl ribonucleoside complex, 0.25mg/mL BSA). Samples were washed twice for 30min at 30°C with wash buffer (25% formamide, 2x SSC) containing DAPI (1μg/mL, D9542; Sigma). All solutions were prepared with RNAse-free water. Finally, the sections were mounted using ProlongGold (P36930; Life Technologies) and imaged the next day.

Imaging

Immunostained and smFISH-treated kidney sections were imaged on a Nikon Ti-Eclipse epifluorescence microscope equipped with an Andor iXON Ultra 888 EMCCD camera, using a 100x /1.45 Plan Apo Lambda oil objective (Nikon) and dedicated, custom-made fluorescence filter sets (Nikon). To cover large areas of the sectioned kidney, images of multiple adjacent areas were taken and combined using the tiling feature of the NIS Elements software (Nikon). For imaging of smFISH signals, z-stacks were collected with distances of 0.3-0.5μm between planes in four fluorescence channels (GFP, Quasar 570, CAL FLuor Red 610, Quasar 670).

2.4.3 Quantification and statistical analysis

scRNA-seq data pruning and normalization

Single-cell expression for the w16 sample was quantified using unique molecular identifiers (UMIs) by 10X Genomics' *Cell Ranger* software. After removing cells with less than 2,000 transcripts per cell, 8,503 cells were retained for further analysis. On average, 1,789 genes were detected per cell and a median of 4,805 transcripts per cell (Fig 2A). Given the recent report that dissociation can have a significant influence on the single-cell transcriptome [34] and that the kidney is notoriously difficult to dissociate, special attention was paid to dissociation-related artifacts. The amount of 1,859 cells with signs of stress were removed from the dataset (Fig 2B). These cells had more than 10% of their expression come from mitochondrial genes (*MT-ND1*, *MT-ND2*, *MT-CO1*, *MT-CO2*, *MT-ATP8*, *MT-ATP6*, *MT-CO3*, *MT-ND3*, *MT-ND4L*, *MT-ND4*, *MT-ND5*, *MT-ND6*, *MT-CYB*) or more than 5% from stress markers. Stress markers were defined as those genes that were significantly up-regulated upon prolonged enzy-

matic incubation of mouse kidney tissue in the study by Adam et al. [23] (Hochane et al. [1] S2 Table, Fig 2B). Mouse genes from this list were converted to human genes using biomart [68]. Genes of the literature set (Table 2.1) only showed small differences between stressed and nonstressed cells (Fig 2C-D), and stressed cells did not form a separate cluster (Fig 2E-F). Therefore, removing stressed cells did not reduce the cell type diversity in the sample. Additionally, 42 cells had more than 1% of their expression coming from *HBB*, *HBA1*, and *HBA2* and were therefore classified as red blood cells and discarded from any further analysis (Fig 2E). Sporadic expression of hemoglobin genes in other cells was likely due to red blood cells that burst before isolation. The same filtering approach was applied to the samples from the other developmental ages as well as the data from Lindström et al. [20] [20]. Raw UMI counts were smoothened by k-nearest neighbors smoothing version 2.1 [13]. This procedure reduces technical noise by sharing information between transcriptionally similar cells, which likely belong to the same cell type. Briefly, the expression profiles of each cell and its knn were summed (k = 10; distance metric: Euclidean distance of the first 10 PCs with a dither of 0.05). The resulting smoothened count matrix had a higher total count than the original and was therefore scaled back to the original matrix by a global factor. Expression was normalized by the method developed by Lun and colleages [69] (as implemented in the scran (version 1.10.1)R package using the functions quickCluster and computeSumFactors). Normalized gene expression was FT transformed in further analyses unless stated otherwise.

Reduction of dimensionality

Variability of gene expression was calculated using the improvedCV2-function from the scran R package. Intercell distances were calculated using the 5% most HVGs excluding stress markers [23] (Hochane et al. [1] S2 Table) and ribosomal genes (obtained from the HGNC website) without any filter for minimum mean expression. For maps of individual samples, we used (1–Pearson correlation) as distance measure. For maps of combined samples, we used Euclidean distance in the MNN-corrected PC space. All tSNE maps used a perplexity setting of 500. For the DDRTree embedding used with pseudotime analysis, see Pseudotime analysis.

Clustering

Hierarchical cluster analysis was performed using Ward linkage and the same intercell distances as for the reduction of dimensionality. The dendrogram of this clustering was cut at height 0.6 to yield 29 clusters of cells (Fig 4). The cut off was chosen such that the number of resulting clusters was comparable to the number of cell types expected from the literature on mouse development [70] and other scRNA-seq studies of the human fetal kidney [11, 20, 46, 56, 57]. We estimated the number of cell types to be around 20 but created slightly more as a starting point to allow for the discovery of new cell types. On the other hand, we did not want to use a much higher number to avoid overclustering (i.e., creating many clusters that

are merely driven by noise, which would then have to be merged manually). The presence of known markers of the different cell types in the kidney (literature set, Table 2.1) was then used to identify cell types (Fig 4). Based on this analysis, some adjacent clusters (clusters 4 and 5, 11 and 12, 15 and 16, and 17 and 18) showed very similar expression of known marker genes of podocytes, ICs, UB, and collecting duct and proximal tubule cells, respectively. In addition, the aforementioned clusters were in close proximity, both in the clustering dendrogram as well as in tSNE space (Fig 4). Consequently, these clusters were merged. For example, clusters 4 and 5 had similar expression of genes known to be expressed in mature podocytes (*NPHS2*, *PTPRO*, *PODXL*) compared to cluster 6, which showed very weak expression of these genes and had distinctive expression of *OLFM3*, which has been shown to be specifically expressed in podocytes precursors [20, 46]. Furthermore, we also merged clusters 7 and 25, which were more distant in the dendrogram of the hierarchical clustering but had very similar literature marker profiles (e.g., *LHX1*, *WNT4*, *CCND1*, *JAG1*, *PAX2*, and *PAX8*) and appeared in close proximity in tSNE space. Finally, clusters 26 and 29 were also merged. Cluster 26 was a heterogeneous cluster of only 56 cells that were spread in tSNE space between multiple other clusters. This cluster was closest to PTA (cluster 29) in terms of literature marker expression (*WNT4*, *LHX1*, and *CCND1*) and differed from it with respect to proliferative state, which may account for the heterogeneous distribution.

Combining different datasets

To compare cells from multiple scRNA-seq datasets, we used the fastMNN function [21] implemented in scran (version 1.10.1) on the first 50 PCs of the 5% HVGs without stress markers or ribosomal genes. We used a knn approach to infer the cell types of unclassified cells from already classified cells. For each unclassified cell, the 20 nearest neighbors in batch-corrected PC space (Euclidean distance) were determined. The most common cell type among these neighbors was then assigned to the unclassified cell. For the comparison with the dataset from Lindström et al. [20], we restricted our dataset to the nephrogenic niche. The cluster identities for the Lindström et al. [20] dataset were kindly provided to us by the group of Andrew D. Smith.

Cell cycle and proliferation

Cell cycle scores were calculated using the Cyclone tool [15] from the scran (version 1.10.1) R package. A list of proliferation markers was adopted from a publication by Whitfield et al. [16].

Pseudotime analysis

We used the *Monocle 2* algorithm [18] to perform embedding and pseudotime analyses on the 2,594 cells of the nephron epithelium, starting from the PTA (cells classified as PTA, RVCSBa, RVCSBb, SSBm/d, SSBpr, SSBpod, DTLH, ErPrT, or podocytes), and separately on the 2,153

cells of the nephrogenic niche (NPC) and the PTA. The 5% HVGs (without stress or ribosomal genes) were used as input to the algorithm. We used the reduceDimension function (max_components = 3 for Fig 9; max_components = 2 for Fig 16A) to run the DDRTree algorithm [19]. The root of the graph learned by DDRTree was placed on the branch that starts with the PTA to obtain the pseudotime shown in Fig 9B.

Marker genes and KeyGenes

For each gene, the cluster of interest (COI) was defined as the cluster that had the highest mean expression of the gene. Then, a binary classifier based on an expression threshold was defined: cells with expression above that threshold were considered to be part of the COI. We systematically varied this threshold to create a ROC based on the cells' true cluster identities. The AUROC was then used to determine the usefulness of this gene as a marker (rather than the specificity or sensitivity at a specific threshold). Genes that had an AUROC exceeding 0.8 were detected in at least 80% of the cells in the COI, had a minimum mean expression of 1.5 in the COI, and those for which maximally 25% of the cells outside the COI had significant expression, were defined as marker set candidates (Hochane et al. [1] S3 Table). Significant expression was defined here as an expression level higher than the 25th percentile of expression in the COI. Subsequently, the top four candidate marker genes per cluster, as ranked by the AUROC, resulted in a final set of 88 marker genes (marker set, Hochane et al. [1] S3 Table).

To apply the KeyGenes prediction algorithm [22], two-thirds of the cells were assigned to the training set and one-third to the test set. A multinomial logistic regression model was trained on the training set with LASSO shrinkage using the 500 most HVGs, filtered for stress markers and ribosomal genes. The shrinkage parameter was determined by 20-fold cross validation. To apply the KeyGenes method to single cells, each cell was treated as a sample, and cross validation was used to control for overfitting. The model obtained a list of 95 classifier genes with nonzero weights (KeyGenes set, Hochane et al. [1] S3 Table). Thereafter, the cells in the test set were assigned to the cell type with the highest identity (id) score; 84% of the cells in the test set were classified correctly (16% test error). On average, the id score was 0.59, and 24% of the cells in the test set obtained an id score higher than 0.8.

Differential expression analysis

For all differential expression analyses, we used EdgeR (version 3.24.0) [71] on raw counts. Normalization and dispersion estimates were calculated by calcNormFactors and estimateDisp, respectively. We modeled gene expression with a negative binomial generalized linear model with glmQLFit. Besides the conditions to be compared, a detection rate for each gene was added to the design matrix. The detection rate is defined as the fraction of cells with nonzero expression. In the comparison of different ages, we excluded the w9 and w16 samples. The w9 sample contained only a few cells, which results in high uncertainty for average gene expression levels. The w16 sample was created separately from the other samples.

Therefore, to avoid batch effects, which are not corrected for in the differential expression analysis, we therefore also excluded the w16 sample.

GWAS analysis

The NHGRI-EBI GWAS catalog was used to retrieve genes associated with traits related to kidney diseases. Specifically, we selected the following kidney traits: *kidney stone*, *kidney disease*, *rapid kidney function decline*, *chronic kidney disease*, *kidney amyloid deposition measurement*, *acute kidney injury*, *type 1 diabetes nephropathy*, *nephrolithiasis*, *diabetic nephropathy*, *proteinuria*, *GFR change measurement*, *renal cell carcinoma*, *serum creatinine measurement*, *cystatin C measurement*, *type 2 diabetes nephropathy*, *immunosuppressive agent*, *tacrolimus measurement*, *focal segmental glomerulosclerosis*, *nephrotic syndrome*, *membranous glomerulonephritis*, *lupus nephritis*, *IgA glomerulonephritis*, *renal system measurement*, and *Wegener's granulomatosis*. This selection resulted in a list of 560 genes (Hochane et al. [1] S4 Table, Kidney GWAS genes). As a negative control, we also obtained a list of 1,508 genes associated with traits related to lung diseases (Hochane et al. [1] S4 Table, Lung GWAS genes) in which we selected the following traits: lung adenocarcinoma, lung carcinoma, interstitial lung disease, squamous cell lung carcinoma, lung disease severity measurement, family history of lung cancer, non-small cell lung carcinoma, diffusing capacity of the lung for carbon monoxide, pulmonary function measurement, vital capacity, emphysema, idiopathic pulmonary fibrosis, chronic bronchitis, chronic obstructive pulmonary disease, pneumonia, and asthma. We performed a one-sided Fisher's exact test to determine whether the genes in the GWAS lists were significantly enriched in the genes that were differentially expressed in our clusters of interest.

Multiple hypothesis testing

In all cases in which significance is reported, p-values were adjusted for multiple hypothesis testing using the Benjamini-Hochberg method.

Functional annotation enrichment

To look for enrichment of gene ontology (GO) terms or transcription factor binding sites we use the DAVID Functional Annotation tool [72], version 6.8 https://david.ncifcrf.gov/ with all genes in the human genome as background gene set. For enrichment of transcription factor binding sites, we used the *UCSC_TFBS* category.

Image analysis

smFISH image stacks were processed with 3D deconvolution and background correction (rolling ball, radius: 3 pixels = 0.39μm), using the NIS Elements software. Subsequently,

maximum projection was used to create a 2D representation of the image stack. The resulting smFISH images were analyzed with homemade MATLAB scripts. First, autofluorescent background was removed by subtracting the appropriately scaled signal of the GFP channel from each of the other channels. Then a region of interest (ROI) containing the structure of interest was defined manually, and candidate smFISH signals were detected by binarizing the image using a global threshold. Connected components were then counted as smFISH signals if they fulfilled two criteria: their average intensity was bigger than the third quartile of individual pixel intensities and they had an area of three pixels or bigger. The density of smFISH signals in the ROI was calculated as the number of retained connected components divided by the area of the ROI.

Images of immunostaining were pre-processed by background subtraction (rolling ball, radius: 100 pixels = 13μm) using ImageJ [73]. Quantification of the immunostaining signal was carried out using homemade MATLAB scripts. First, the CM region was segmented manually. Then, cross-sections of the CM, roughly perpendicular to the outline of the UB, were drawn by hand, approximately 30 pixels apart. For the starting point of each cross-section, the contour length s along the UB starting from the top of the CM (close to the edge of the cortex) was determined (see Fig 16B). The distance s was expressed relative to the total contour length (from top to bottom of the UB). The distance from the UB along the cross-sections was defined to be the distance d (see Fig 16B). Fluorescence intensities were then averaged over lines of 30 pixels length perpendicular to the drawn cross-section. The resulting intensity profiles (which depend on d and s) were then averaged over multiple images and either s or d to get average intensity profiles depending only on d or s. Normalization to the maximum intensity of each profile resulted in the intensity profiles reported in Fig 16D-E. Division of the CITED1 intensity profile by the SIX2 intensity profile gave the ratio plotted in Fig 16D-E. Accuracy, indicated by error bars in the plots, was quantified as the standard error of the mean calculated over all evaluated profiles.

Data availability

The scRNA-seq data have been deposited in the GEO database under accession number GSE114530. An interactive web application accompanying this paper, which provides convenient access to the data, can be found here: http://www.semraulab.com/kidney.

Acknowledgments

We are thankful to Gynaikon Clinic in Rotterdam for their efforts in collecting and providing the fetal material; Susan Kloet and Emile Meijer from the Leiden Genome Technology Center for cell encapsulation, library preparation, single-cell sequencing, primary data mapping, and quality control; Vanessa Torrens-Juaneda and Ioannis Moustakas for primary and secondary data analysis and discussions; GenomseScan for technical support; and the group of Andrew D. Smith for providing the cluster identities of the Lindström et al. dataset.

2.4.4 Author contributions

Conceptualization: Mazène Hochane, Susana M. Chuva de Sousa Lopes, Stefan Semrau.

Data curation: Mazène Hochane, Patrick R. van den Berg.

Formal analysis: Patrick R. van den Berg.

Funding acquisition: Susana M. Chuva de Sousa Lopes, Stefan Semrau.

Investigation: Mazène Hochane, Patrick R. van den Berg, Xueying Fan, Noémie Bérenger-Currias, Esmée Adegeest, Monika Bialecka, Maaike Nieveen.

Project administration: Mazène Hochane, Susana M. Chuva de Sousa Lopes, Stefan Semrau.

Resources: Xueying Fan, Monika Bialecka, Maaike Nieveen, Susana M. Chuva de Sousa Lopes..

Software: Patrick R. van den Berg, Maarten Menschaart.

Supervision: Susana M. Chuva de Sousa Lopes, Stefan Semrau.

Visualization: Patrick R. van den Berg, Noémie Bérenger-Currias, Esmée Adegeest, Maarten Menschaart.

Writing - original draft: Mazène Hochane, Patrick R. van den Berg, Noémie Bérenger-Currias, Esmée Adegeest, Stefan Semrau.

Writing - review & editing: Mazène Hochane, Patrick R. van den Berg, Xueying Fan, Noémie Bérenger-Currias, Esmée Adegeest, Monika Bialecka, Maaike Nieveen, Maarten Menschaart, Susana M. Chuva de Sousa Lopes, Stefan Semrau.

2.5 Supplementary information

Abbreviation	Cell type	Gene	Reference
NPCs	nephron progenitor cells	OSR1	[35]
		SIX1	[74]
		SIX2	[75, 76]
		CITED1	[77, 78]
		EYA1	[79]
		SALL1	[80, 60]
		MEOX1	[3]
		GDNF	[81]
		ETV4	[82]
		COL2A1	[83]
		HES1	[84, 78]
		CRABP2	[20]
		LEF1	[20]
		ITGA8	[46]
PTA	pretubular aggregate	LHX1	[20]
		WNT4	[85]
		CCND1	[86]
RVCSB a and b	renal vesicle/comma-shaped body	LHX1	[87]
		PAX8	[85]
		JAG1	[88]
		PAX2	[41]
		WNT4	[89]
		SFRP2	[90]
		DLL1	[4]
SSBm/d	SSB medial/distal	HNF1B	[91]
		POU3F3	[92]
		SIM2	[93]
		SOX9	[20]
		IRX2	[46]
		IRX3	[46]
SSBpr	SSB proximal precursor cell	CDH6	[94, 95]
		HNF1A	[96, 97]
		AMN	[93]
SSBpod	SSB podocyte precursor cell	FOXC2	[98]
		MAFB	[48]
		OLFM3	[20]
		WT1	[99]
Pod	podocytes	PTPRO	[100]
		NPHS2	[101]
		NPHS1	[102]
		PODXL	[103]
		TGFBR3	[103]

Continued on next page

Continued from previous page

Abbreviation	Cell type	Gene	Reference
		CLIC5	[104]
		CITED2	[98]
DTLH	distal tubule/loop of Henle	PAPPA2	[105]
		MAL	[106]
		CLCN5	[107]
		SLC12A3	[108]
		UMOD	[20]
		SLC12A1	[109]
ErPrT	early proximal tubule	LRP2	[110]
		ANPEP	[111]
		SLC34A1	[112]
		CLDN1	[113]
		CLDN2	[114]
		SLC13A1	[115]
		CUBN	[116]
CnT	connecting tubule	ALDH1A1	[20]
		TACSTD2	[20]
		CDH1	[117]
IPC	interstitial progenitor cell	GDNF	[118]
		FOXD1	[119]
ICs a and b	interstitial cells a and b	DES	[120]
		COL3A1	[121]
		COL1A1	[122]
		SERPINE2	[123]
		FGF7	[41, 124]
		LEF1	[125]
		DCN	[126]
Mes	mesangial cells	ACTA2	[123]
		TPM2	[127]
		PDGFRB	[128]
		MCAM	[123]
		CSPG4	[123]
		CD248	[129]
UBCD	ureteric bud/collecting duct	KRT18	[130]
		KRT8	[130]
		RET	[131]
		CLDN7	[132]
		AQP2	[133]
		GATA2	[133]
		MMP7	[134]
		CALB1	[117]
End	endothelial cell	KDR	[135]
		TEK	[136]
		FLT1	[135]
		CDH5	[137]

Continued on next page

Continued from previous page			
Abbreviation	Cell type	Gene	Reference
		PECAM1	[138]
Leu	leukocytes	CD37	[139]
		CD48	[140]
		ITGB2	[141]
		IFI30	[142]
		IL1B	[143]

Table 2.1. Literature marker set and references.

Acronyms

AUROC area under the ROC

COI cluster of interest

FDR false discovery rate

FISH fluorescence in situ hybridization

GDNF glial cell-derived neurotrophic factor

GO gene ontology

GWAS genome-wide association studies

knn k-nearest neighbors

L2FC $\log_2$ fold change

PCR polymerase chain reaction

RA retinoic acid

RET ret proto-oncogene

ROC receiver operating characteristic

ROI region of interest

RT room temperature

smFISH single-molecule FISH

tSNE t-distributed stochastic neighbor embedding

UMI unique molecular identifier

w16 week 16

Cell types

CM cap mesenchyme

CnT connecting tubule

CSB comma-shaped body

DTLH distal tubule/loop of Henle

End endothelial cell

ErPrT early proximal tubule

IC interstitial cell

IPC interstitial progenitor cell

Leu leukocytes

LOH loop of Henle

Mes mesangial cells

MM metanephric mesenchyme

NPC nephron progenitor cell

Pod podocyte

PTA pretubular aggregate

RV renal vesicle

RVCSB renal vesicle/comma-shaped body

SSB s-shaped body

SSBm/d SSB medial/distal

SSBpod SSB podocyte precursor cell

SSBpr SSB proximal precursor cell

UB ureteric bud

UBCD ureteric bud/collecting duct

2.6 References

[1] Mazène Hochane et al. “Single-cell transcriptomics reveals gene expression dynamics of human fetal kidney development”. In: *PLOS Biology* 17.2 (Feb. 2019), e3000152. DOI: 10.1371/journal.pbio.3000152.

[2] Frank Costantini et al. “Patterning a Complex Organ: Branching Morphogenesis and Nephron Segmentation in Kidney Development”. In: *Developmental Cell* 18.5 (2010), pp. 698–712. DOI: 10.1016/j.devcel.2010.04.008.

[3] Alexander N. Combes et al. “Haploinsufficiency for the Six2 gene increases nephron progenitor proliferation promoting branching and nephron number”. In: *Kidney International* 93.3 (2018), pp. 589–598. DOI: 10.1016/j.kint.2017.09.015.

[4] Melissa H. Little et al. “Mammalian kidney development: principles, progress, and projections”. In: *Cold Spring Harb Perspect Biol* 4.5 (2012), a008300–a008300. DOI: 10.1101/cshperspect.a008300.

[5] John F. Bertram et al. “Why and how we determine nephron number”. In: *Pediatr. Nephrol.* 29.4 (2014), pp. 575–580. DOI: 10.1007/s00467-013-2600-y.

[6] Frank Costantini. “Genetic controls and cellular behaviors in branching morphogenesis of the renal collecting system”. In: *Wiley Interdiscip Rev Dev Biol* 1.5 (2012), pp. 693–713. DOI: 10.1002/wdev.52.

[7] Akio Kobayashi et al. “Identification of a multipotent self-renewing stromal progenitor population during mammalian kidney organogenesis”. In: *Stem Cell Reports* 3.4 (2014), pp. 650–662. DOI: 10.1016/j.stemcr.2014.08.008.

[8] Maria Luisa S. Sequeira-Lopez et al. “The earliest metanephric arteriolar progenitors and their role in kidney vascular development”. In: *Am. J. Physiol. Regul. Integr. Comp. Physiol.* 308.2 (2015), R138–149. DOI: 10.1152/ajpregu.00428.2014.

[9] B. Robert et al. “Evidence that embryonic kidney cells expressing flk-1 are intrinsic, vasculogenic angioblasts”. In: *Am. J. Physiol.* 271.3 Pt 2 (1996), F744–753. DOI: 10.1152/ajprenal.1996.271.3.F744.

[10] Nils O. Lindström et al. “Conserved and Divergent Features of Human and Mouse Kidney Organogenesis”. In: *JASN* 29.3 (2018), pp. 785–805. DOI: 10.1681/ASN.2017080887.

[11] Nils O. Lindström et al. “Conserved and Divergent Features of Mesenchymal Progenitor Cell Types within the Cortical Nephrogenic Niche of the Human and Mouse Kidney”. In: *JASN* 29.3 (2018), pp. 806–824. DOI: 10.1681/ASN.2017080890.

[12] Nils O. Lindström et al. “Conserved and Divergent Molecular and Anatomic Features of Human and Mouse Nephron Patterning”. In: *JASN* 29.3 (2018), pp. 825–840. DOI: 10.1681/ASN.2017091036.

[13] Florian Wagner et al. “K-nearest neighbor smoothing for high-throughput single-cell RNA-Seq data”. In: *bioRxiv* (2018), p. 217737. DOI: 10.1101/217737.

[14] Laurens van der Maaten et al. “Visualizing Data using t-SNE”. In: *Journal of Machine Learning Research* 9.86 (2008), pp. 2579–2605.

[15] Antonio Scialdone et al. “Computational assignment of cell-cycle stage from single-cell transcriptome data”. In: *Methods* 85 (2015), pp. 54–61. DOI: 10.1016/j.ymeth.2015.06.021.

[16] Michael L. Whitfield et al. “Common markers of proliferation”. In: *Nat. Rev. Cancer* 6.2 (2006), pp. 99–106. DOI: 10.1038/nrc1802.

[17] Joseph L. Napoli. “Cellular retinoid binding-proteins, CRBP, CRABP, FABP5: Effects on retinoid metabolism, function and related diseases”. In: *Pharmacol. Ther.* 173 (2017), pp. 19–33. DOI: 10.1016/j.pharmthera.2017.01.004.

[18] Xiaojie Qiu et al. “Reversed graph embedding resolves complex single-cell trajectories”. In: *Nat. Methods* 14.10 (2017), pp. 979–982. DOI: 10.1038/nmeth.4402.

[19] Qi Mao et al. “Dimensionality Reduction Via Graph Structure Learning”. In: 2015, pp. 765–774. DOI: 10.1145/2783258.2783309.

[20] Nils O Lindström et al. “Progressive Recruitment of Mesenchymal Progenitors Reveals a Time-Dependent Process of Cell Fate Acquisition in Mouse and Human Nephrogenesis”. In: *Developmental Cell* 45.5 (2018), 651–660.e4. DOI: 10.1016/j.devcel.2018.05.010.

[21] Laleh Haghverdi et al. “Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors”. In: *Nat. Biotechnol.* 36.5 (2018), pp. 421–427. DOI: 10.1038/nbt.4091.

[22] Matthias S. Roost et al. “KeyGenes, a Tool to Probe Tissue Differentiation Using a Human Fetal Transcriptional Atlas”. In: *Stem Cell Reports* 4.6 (2015), pp. 1112–1124. DOI: 10.1016/j.stemcr.2015.05.002.

[23] Mike Adam et al. “Psychrophilic proteases dramatically reduce single-cell RNA-seq artifacts: a molecular atlas of kidney development”. In: *Development* 144.19 (2017), pp. 3625–3632. DOI: 10.1242/dev.151142.

[24] J M Linton et al. “The ECM protein nephronectin promotes kidney development via integrin 8 1-mediated stimulation of Gdnf expression”. In: *Development* 134.13 (2007), pp. 2501–2509. DOI: 10.1242/dev.005033.

[25] Janina Müller-Deile et al. “Podocytes from the diagnostic and therapeutic point of view”. In: *Pflugers Arch.* 469.7-8 (2017), pp. 1007–1015. DOI: 10.1007/s00424-017-1993-z.

[26] Liliana Schaefer et al. “Small proteoglycans of normal adult human kidney: Distinct expression patterns of decorin, biglycan, fibromodulin, and lumican”. In: *Kidney International* 58.4 (2000), pp. 1557–1568. DOI: 10.1046/j.1523-1755.2000.00317.x.

[27] Joshua W. Mugford et al. “High-resolution gene expression analysis of the developing mouse kidney defines novel cellular compartments within the nephron progenitor population”. In: *Dev. Biol.* 333.2 (2009), pp. 312–323. DOI: 10.1016/j.ydbio.2009.06.043.

[28] Kyung-Ah Kim et al. “R-Spondin Family Members Regulate the Wnt Pathway by a Common Mechanism”. In: *MBoC* 19.6 (2008), pp. 2588–2596. DOI: 10.1091/mbc.e08-02-0187.

[29] C. Pedraza et al. “A retinoic acid-responsive element in human midkine gene”. In: *J. Biochem.* 117.4 (1995), pp. 845–849. DOI: 10.1093/oxfordjournals.jbchem.a124785.

[30] Katri M. Makkonen et al. “Regulation of the hyaluronan synthase 2 gene by convergence in cyclic AMP response element-binding protein and retinoid acid receptor signaling”. In: *J. Biol. Chem.* 284.27 (2009), pp. 18270–18281. DOI: 10.1074/jbc.M109.012492.

[31] José Vilar et al. “Midkine is involved in kidney development and in its regulation by retinoids”. In: *J. Am. Soc. Nephrol.* 13.3 (2002), pp. 668–676.

[32] Waichi Sato et al. “Midkine expression in the course of nephrogenesis and its role in ischaemic reperfusion injury”. In: *Nephrol. Dial. Transplant.* 17 Suppl 9 (2002), pp. 52–54. DOI: 10.1093/ndt/17.suppl_9.52.

[33] Ayaka Kimura et al. “HMGB2 expression is associated with transition from a quiescent to an activated state of adult neural stem cells”. In: *Dev. Dyn.* 247.1 (2018), pp. 229–238. DOI: 10.1002/dvdy.24559.

[34] Susanne C. van den Brink et al. “Single-cell sequencing reveals dissociation-induced gene expression in tissue subpopulations”. In: *Nature Methods* 14.10 (2017), pp. 935–936. DOI: 10.1038/nmeth.4437.

[35] Joshua W. Mugford et al. “Osr1 expression demarcates a multi-potent population of intermediate mesoderm that undergoes progressive restriction to an Osr1-dependent nephron progenitor compartment within the mammalian kidney”. In: *Dev. Biol.* 324.1 (2008), pp. 88–98. DOI: 10.1016/j.ydbio.2008.09.010.

[36] Yasmine Neirijnck et al. “Sox11 gene disruption causes congenital anomalies of the kidney and urinary tract (CAKUT)”. In: *Kidney Int.* 93.5 (2018), pp. 1142–1153. DOI: 10.1016/j.kint.2017.11.026.

[37] Priscilla Soulié et al. “Spatially restricted hyaluronan production by Has2 drives epithelial tubulogenesis in vitro”. In: *Am. J. Physiol., Cell Physiol.* 307.8 (2014), pp. C745–759. DOI: 10.1152/ajpcell.00047.2014.

[38] Tomoko Ohmori et al. “Sall1 in renal stromal progenitors non-cell autonomously restricts the excessive expansion of nephron progenitors”. In: *Sci Rep* 5.1 (2015), p. 15676. DOI: 10.1038/srep15676.

[39] Cecil Han et al. “The RNA-binding protein DDX1 promotes primary microRNA maturation and inhibits ovarian tumor progression”. In: *Cell Rep* 8.5 (2014), pp. 1447–1460. DOI: 10.1016/j.celrep.2014.07.058.

[40] Neeraja Sammeta et al. “Uncx regulates proliferation of neural progenitor cells and neuronal survival in the olfactory epithelium”. In: *Mol. Cell. Neurosci.* 45.4 (2010), pp. 398–407. DOI: 10.1016/j.mcn.2010.07.013.

[41] Minoru Takasato et al. “Identification of kidney mesenchymal genes by a combination of microarray analysis and Sall1-GFP knockin mice”. In: *Mech. Dev.* 121.6 (2004), pp. 547–557. DOI: 10.1016/j.mod.2004.04.007.

[42] Yukinori Okada et al. “Meta-analysis identifies multiple loci associated with kidney function-related traits in east Asian populations”. In: *Nat. Genet.* 44.8 (2012), pp. 904–909. DOI: 10.1038/ng.2352.

[43] Mathias Gorski et al. “1000 Genomes-based meta-analysis identifies 10 novel loci for kidney function”. In: *Sci Rep* 7 (2017), p. 45040. DOI: 10.1038/srep45040.

[44] Anubha Mahajan et al. “Trans-ethnic Fine Mapping Highlights Kidney-Function Genes Linked to Salt Sensitivity”. In: *Am. J. Hum. Genet.* 99.3 (2016), pp. 636–646. DOI: 10.1016/j.ajhg.2016.07.012.

[45] Cristian Pattaro et al. “Genetic associations at 53 loci highlight cell types and biological pathways relevant for kidney function”. In: *Nat Commun* 7 (2016), p. 10023. DOI: 10.1038/ncomms10023.

[46] Rajasree Menon et al. “Single-cell analysis of progenitor cell dynamics and lineage specification in the human fetal kidney”. In: *Development* 145.16 (2018). DOI: 10.1242/dev.164038.

[47] Masaru Motojima et al. “Foxc1 and Foxc2 are necessary to maintain glomerular podocytes”. In: *Exp. Cell Res.* 352.2 (2017), pp. 265–272. DOI: 10.1016/j.yexcr.2017.02.016.

[48] Virginia Sadl et al. “The mouse Kreisler (Krml1/MafB) segmentation gene is required for differentiation of glomerular visceral epithelial cells”. In: *Dev. Biol.* 249.1 (2002), pp. 16–29. DOI: 10.1006/dbio.2002.0751.

[49] Kevin V. Lemley. “Mechanical challenges to the glomerulus and podocyte loss: evolution of a paradigm”. In: *Pflugers Arch.* 469.7-8 (2017), pp. 959–963. DOI: 10.1007/s00424-017-2012-0.

[50] S. Nanayakkara et al. “An integrative study of the genetic, social and environmental determinants of chronic kidney disease characterized by tubulointerstitial damages in the North Central Region of Sri Lanka.” In: *J Occup Health* 56.1 (2014), pp. 28–38. DOI: 10.1539/joh.13-0172-OA.

[51] T. A. Moreno et al. “The secreted glycoprotein Noelin-1 promotes neurogenesis in Xenopus”. In: *Dev. Biol.* 240.2 (2001), pp. 340–360. DOI: 10.1006/dbio.2001.0472.

[52] Diangeng Li et al. “Mesenchymal stem cells protect podocytes from apoptosis induced by high glucose via secretion of epithelial growth factor”. In: *Stem Cell Res Ther* 4.5 (2013), p. 103. DOI: 10.1186/scrt314.

[53] Guang Liang et al. “Fibroblast growth factor 1 ameliorates diabetic nephropathy by an anti-inflammatory mechanism”. In: *Kidney Int.* 93.1 (2018), pp. 95–109. DOI: 10.1016/j.kint.2017.05.013.

[54] Christina S. Bartlett et al. “Vascular Growth Factors and Glomerular Disease”. In: *Annu. Rev. Physiol.* 78 (2016), pp. 437–461. DOI: 10.1146/annurev-physiol-021115-105412.

[55] Min Yao et al. “The Notch pathway mediates the angiotensin II-induced synthesis of extracellular matrix components in podocytes”. In: *Int. J. Mol. Med.* 36.1 (2015), pp. 294–300. DOI: 10.3892/ijmm.2015.2193.

[56] Eric W. Brunskill et al. “Single cell dissection of early kidney development: multilineage priming”. In: *Development* 141.15 (2014), pp. 3093–3101. DOI: 10.1242/dev.110601.

[57] Ping Wang et al. “Dissecting the Global Dynamic Molecular Profiles of Human Fetal Kidney Development by Single-Cell RNA Sequencing”. In: *Cell Rep* 24.13 (2018), 3554–3567.e3. DOI: 10.1016/j.celrep.2018.08.056.

[58] Carolina Rosselot et al. “Non-cell-autonomous retinoid signaling is crucial for renal development”. In: *Development* 137.2 (2010), pp. 283–292. DOI: 10.1242/dev.040287.

[59] Alexander N. Combes et al. “Cap mesenchyme cell swarming during kidney development is influenced by attraction, repulsion, and adhesion to the ureteric tip”. In: *Dev. Biol.* 418.2 (2016), pp. 297–306. DOI: 10.1016/j.ydbio.2016.06.028.

[60] Scott Boyle et al. “Cited1 and Cited2 are differentially expressed in the developing kidney but are not required for nephrogenesis”. In: *Dev. Dyn.* 236.8 (2007), pp. 2321–2330. DOI: 10.1002/dvdy.21242.

[61] Christina A. Young et al. “Embryonic AP1 Transcription Factor Deficiency Causes a Collodion Baby-Like Phenotype”. In: *J. Invest. Dermatol.* 137.9 (2017), pp. 1868–1877. DOI: 10.1016/j.jid.2017.04.032.

[62] Fumiaki Kawashima et al. “c-jun is differentially expressed in embryonic and adult neural precursor cells”. In: *Histochem. Cell Biol.* 147.6 (2017), pp. 721–731. DOI: 10.1007/s00418-016-1536-2.

[63] Victoria C. Garside et al. “SOX9 modulates the expression of key transcription factors required for heart valve development”. In: *Development* 142.24 (2015), pp. 4340–4350. DOI: 10.1242/dev.125252.

[64] Eric W. Brunskill et al. “Defining the molecular character of the developing and adult kidney podocyte”. In: *PLoS ONE* 6.9 (2011), e24640. DOI: 10.1371/journal.pone.0024640.

[65] A Heeren et al. “Development of the follicular basement membrane during human gametogenesis and early folliculogenesis”. In: *BMC Developmental Biology* 15.1 (2015), p. 4. DOI: 10.1186/s12861-015-0054-0.

[66] Ábel Vértesy et al. “Parental haplotype-specific single-cell transcriptomics reveal incomplete epigenetic reprogramming in human female germ cells”. In: *Nat Commun* 9.1 (2018), p. 1873. DOI: 10.1038/s41467-018-04215-7.

[67] Stefan Semrau et al. “FuseFISH: robust detection of transcribed gene fusions in single cells”. In: *Cell Rep* 6.1 (2014), pp. 18–23. DOI: 10.1016/j.celrep.2013.12.002.

[68] Steffen Durinck et al. “Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt”. In: *Nat Protoc* 4.8 (2009), pp. 1184–1191. DOI: 10.1038/nprot.2009.97.

[69] Aaron T. L. Lun et al. “Pooling across cells to normalize single-cell RNA sequencing data with many zero counts”. In: *Genome Biol.* 17 (2016), p. 75. DOI: 10.1186/s13059-016-0947-7.

[70] Andrew P. McMahon. “Development of the Mammalian Kidney”. In: *Curr. Top. Dev. Biol.* 117 (2016), pp. 31–64. DOI: 10.1016/bs.ctdb.2015.10.010.

[71] Mark D. Robinson et al. “edgeR: a Bioconductor package for differential expression analysis of digital gene expression data”. In: *Bioinformatics* 26.1 (2010), pp. 139–140. DOI: 10.1093/bioinformatics/btp616.

[72] Da Wei Huang et al. “Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources”. In: *Nature Protocols* 4.1 (2009), pp. 44–57. DOI: 10.1038/nprot.2008.211.

[73] Caroline A. Schneider et al. “NIH Image to ImageJ: 25 years of image analysis”. In: *Nat. Methods* 9.7 (2012), pp. 671–675. DOI: 10.1038/nmeth.2089.

[74] Pin-Xian Xu et al. “Six1 is required for the early organogenesis of mammalian kidney”. In: *Development* 130.14 (2003), pp. 3085–3094. DOI: 10.1242/dev.00536.

[75] Michael Marcotte et al. “Gene regulatory network of renal primordium development”. In: *Pediatr Nephrol* 29.4 (2014), pp. 637–644. DOI: 10.1007/s00467-013-2635-0.

[76] Akio Kobayashi et al. “Six2 defines and regulates a multipotent self-renewing nephron progenitor population throughout mammalian kidney development”. In: *Cell Stem Cell* 3.2 (2008), pp. 169–181. DOI: 10.1016/j.stem.2008.05.020.

[77] Scott Boyle et al. “Fate mapping using Cited1-CreERT2 mice demonstrates that the cap mesenchyme contains self-renewing progenitor cells and gives rise exclusively to nephronic epithelia”. In: *Dev. Biol.* 313.1 (2008), pp. 234–245. DOI: 10.1016/j.ydbio.2007.10.014.

[78] Sayoko Fujimura et al. “Notch2 activation in the embryonic kidney depletes nephron progenitors”. In: *J. Am. Soc. Nephrol.* 21.5 (2010), pp. 803–810. DOI: 10.1681/ASN.2009040353.

[79] P. X. Xu et al. “Eya1-deficient mice lack ears and kidneys and show abnormal apoptosis of organ primordia”. In: *Nat. Genet.* 23.1 (1999), pp. 113–117. DOI: 10.1038/12722.

[80] Ryuichi Nishinakamura et al. “Murine homolog of SALL1 is essential for ureteric bud invasion in kidney development”. In: *Development* 128.16 (2001), pp. 3105–3115.

[81] Alexander N. Combes et al. “Cell-cell interactions driving kidney morphogenesis”. In: *Curr. Top. Dev. Biol.* 112 (2015), pp. 467–508. DOI: 10.1016/bs.ctdb.2014.12.002.

[82] Benson C. Lu et al. “Etv4 and Etv5 are required downstream of GDNF and Ret for kidney branching morphogenesis”. In: *Nat. Genet.* 41.12 (2009), pp. 1295–1302. DOI: 10.1038/ng.476.

[83] Masahito Miyamoto et al. “In-depth proteomic profiling of the normal human kidney glomerulus using two-dimensional protein prefractionation in combination with liquid chromatography-tandem mass spectrometry”. In: *J. Proteome Res.* 6.9 (2007), pp. 3680–3690. DOI: 10.1021/pr070203n.

[84] Sophie Jarriault et al. “Signalling downstream of activated mammalian Notch”. In: *Nature* 377.6547 (1995), pp. 355–358. DOI: 10.1038/377355a0.

[85] K. Stark et al. “Epithelial transformation of metanephric mesenchyme in the developing kidney regulated by Wnt-4”. In: *Nature* 372.6507 (1994), pp. 679–683. DOI: 10.1038/372679a0.

[86] Thomas F. Gallegos et al. “A Protein Kinase A and Wnt-dependent network regulating an intermediate stage in epithelial tubulogenesis during kidney development”. In: *Dev Biol* 364.1 (2012), pp. 11–21. DOI: 10.1016/j.ydbio.2012.01.014.

[87] Kimmo Halt et al. “Coordination of kidney organogenesis by Wnt signaling”. In: *Pediatr Nephrol* 29.4 (2014), pp. 737–744. DOI: 10.1007/s00467-013-2733-z.

[88] Madhulika Sharma et al. “Coexpression of Cux-1 and Notch signaling pathway components during kidney development”. In: *Dev. Dyn.* 231.4 (2004), pp. 828–838. DOI: 10.1002/dvdy.20175.

[89] S. Steven Potter et al. “Laser capture-microarray analysis of Lim1 mutant kidney development”. In: *Genesis* 45.7 (2007), pp. 432–439. DOI: 10.1002/dvg.20309.

[90] Cornelia Leimeister et al. “Developmental expression patterns of mouse sFRP genes encoding members of the secreted frizzled related protein family”. In: *Mechanisms of Development* 75.1 (1998), pp. 29–42. DOI: 10.1016/S0925-4773(98)00072-0.

[91] Claire Heliot et al. “HNF1B controls proximal-intermediate nephron segment identity in vertebrates by regulating Notch signalling components and Irx1/2”. In: *Development* 140.4 (2013), pp. 873–885. DOI: 10.1242/dev.086538.

[92] Alexandra Rieger et al. “Missense Mutation of POU Domain Class 3 Transcription Factor 3 in Pou3f3L423P Mice Causes Reduced Nephron Number and Impaired Development of the Thick Ascending Limb of the Loop of Henle”. In: *PLOS ONE* 11.7 (2016), e0158977. DOI: 10.1371/journal.pone.0158977.

[93] Eric W. Brunskill et al. “Atlas of Gene Expression in the Developing Kidney at Microanatomic Resolution”. In: *Dev Cell* 15.5 (2008), pp. 781–791. DOI: 10.1016/j.devcel.2008.09.007.

[94] Steven P. Mah et al. “Kidney Development in Cadherin-6 Mutants: Delayed Mesenchyme-to-Epithelial Conversion and Loss of Nephrons”. In: *Developmental Biology* 223.1 (2000), pp. 38–53. DOI: 10.1006/dbio.2000.9738.

[95] Liwei Jia et al. “Distinct roles of cadherin-6 and E-cadherin in tubulogenesis and lumen formation”. In: *Mol Biol Cell* 22.12 (2011), pp. 2031–2041. DOI: 10.1091/mbc.E11-01-0038.

[96] Filippo Massa et al. “Hepatocyte nuclear factor 1β controls nephron tubular development”. In: *Development* 140.4 (2013), pp. 886–896. DOI: 10.1242/dev.086546.

[97] M. Pontoglio et al. “Hepatocyte nuclear factor 1 inactivation results in hepatic dysfunction, phenylketonuria, and renal Fanconi syndrome”. In: *Cell* 84.4 (1996), pp. 575–585. DOI: 10.1016/s0092-8674(00)81033-8.

[98] Minoru Takemoto et al. “Large-scale identification of genes implicated in kidney glomerulus development and function”. In: *EMBO J.* 25.5 (2006), pp. 1160–1174. DOI: 10.1038/sj.emboj.7601014.

[99] A. J. Buckler et al. “Isolation, characterization, and expression of the murine Wilms’ tumor gene (WT1) during kidney development”. In: *Mol. Cell. Biol.* 11.3 (1991), pp. 1707–1712. DOI: 10.1128/mcb.11.3.1707.

[100] Pedro J. Beltran et al. “Expression of PTPRO during mouse development suggests involvement in axonogenesis and differentiation of NT-3 and NGF-dependent neurons”. In: *J. Comp. Neurol.* 456.4 (2003), pp. 384–395. DOI: 10.1002/cne.10532.

[101] Natalya V. Kaverina et al. “Partial podocyte replenishment in experimental FSGS derives from nonpodocyte sources”. In: *Am. J. Physiol. Renal Physiol.* 310.11 (2016), F1397–1413. DOI: 10.1152/ajprenal.00369.2015.

[102] Ellen F. Carney. “Podocytes: ShcA regulates nephrin turnover”. In: *Nat Rev Nephrol* 13.12 (2017), p. 722. DOI: 10.1038/nrneph.2017.153.

[103] C. Schell et al. “Glomerular development–shaping the multi-cellular filtration unit”. In: *Semin. Cell Dev. Biol.* 36 (2014), pp. 39–49. DOI: 10.1016/j.semcdb.2014.07.016.

[104] Brian A. Pierchala et al. “Proteomic analysis of the slit diaphragm complex: CLIC5 is a protein critical for podocyte morphology and function”. In: *Kidney International* 78.9 (2010), pp. 868–882. DOI: 10.1038/ki.2010.212.

[105] Allen W. Cowley et al. “Pappa2 is linked to salt-sensitive hypertension in Dahl S rats”. In: *Physiol Genomics* 48.1 (2016), pp. 62–72. DOI: 10.1152/physiolgenomics.00097.2015.

[106] Monica Carmosino et al. “MAL/VIP17, a New Player in the Regulation of NKCC2 in the Kidney”. In: *Mol. Biol. Cell* 21.22 (2010), pp. 3985–3997. DOI: 10.1091/mbc.E10-05-0456.

[107] J. Christopher Hennings et al. “The ClC-K2 Chloride Channel Is Critical for Salt Handling in the Distal Nephron”. In: *J Am Soc Nephrol* 28.1 (2017), pp. 209–217. DOI: 10.1681/ASN.2016010085.

[108] Catherina A. Cuevas et al. “Potassium Sensing by Renal Distal Tubules Requires Kir4.1”. In: *J. Am. Soc. Nephrol.* 28.6 (2017), pp. 1814–1825. DOI: 10.1681/ASN.2016090935.

[109] Hayo Castrop et al. “Physiology and pathophysiology of the renal Na-K-2Cl cotransporter (NKCC2)”. In: *Am. J. Physiol. Renal Physiol.* 307.9 (2014), F991–F1002. DOI: 10.1152/ajprenal.00432.2014.

[110] Christopher P. Larsen et al. “LDL Receptor-Related Protein 2 (Megalin) as a Target Antigen in Human Kidney Anti-Brush Border Antibody Disease”. In: *J. Am. Soc. Nephrol.* 29.2 (2017), pp. 644–653. DOI: 10.1681/ASN.2017060664.

[111] Kumar Kotlo et al. “Aminopeptidase N reduces basolateral Na+ -K+ -ATPase in proximal tubule cells”. In: *Am. J. Physiol. Renal Physiol.* 293.4 (2007), F1047–1053. DOI: 10.1152/ajprenal.00074.2007.

[112] Daniel Caballero et al. “Intraperitoneal pyrophosphate treatment reduces renal calcifications in Npt2a null mice”. In: *PLoS ONE* 12.7 (2017), e0180098. DOI: 10.1371/journal.pone.0180098.

[113] Yumiko Kiuchi-Saishin et al. “Differential Expression Patterns of Claudins, Tight Junction Membrane Proteins, in Mouse Nephron Segments”. In: *JASN* 13.4 (2002), pp. 875–886.

[114] Takamoto Ohse et al. “Establishment of Conditionally Immortalized Mouse Glomerular Parietal Epithelial Cells in Culture”. In: *J Am Soc Nephrol* 19.10 (2008), pp. 1879–1890. DOI: 10.1681/ASN.2007101087.

[115] Daniel Markovich. “Na+-sulfate cotransporter SLC13A1”. In: *Pflugers Arch - Eur J Physiol* 466.1 (2014), pp. 131–137. DOI: 10.1007/s00424-013-1388-8.

[116] Erik Ilsø Christensen et al. “Receptor-mediated endocytosis in renal proximal tubule”. In: *Pflugers Arch.* 458.6 (2009), pp. 1039–1048. DOI: 10.1007/s00424-009-0685-8.

[117] Kylie Georgas et al. “Analysis of early nephron patterning reveals a role for distal RV proliferation in fusion to the ureteric tip via a cap mesenchyme-derived connecting segment”. In: *Dev. Biol.* 332.2 (2009), pp. 273–286. DOI: 10.1016/j.ydbio.2009.05.578.

[118] Bliss Magella et al. “Cross-platform single cell analysis of kidney development shows stromal cells express Gdnf”. In: *Dev. Biol.* 434.1 (2018), pp. 36–47. DOI: 10.1016/j.ydbio.2017.11.006.

[119] Benjamin D. Humphreys et al. “Fate tracing reveals the pericyte and not epithelial origin of myofibroblasts in kidney fibrosis”. In: *Am. J. Pathol.* 176.1 (2010), pp. 85–97. DOI: 10.2353/ajpath.2010.090517.

[120] Scott C. Boyle et al. “Notch signaling is required for the formation of mesangial cells from a stromal mesenchyme precursor during kidney development”. In: *Development* 141.2 (2014), pp. 346–354. DOI: 10.1242/dev.100271.

[121] Lino Muñoz Cuellar et al. “Identification and localization of novel genes preferentially expressed in human kidney glomerulus”. In: *Nephrology (Carlton)* 14.1 (2009), pp. 94–104. DOI: 10.1111/j.1440-1797.2008.01009.x.

[122] Prashant S. Patole et al. “Toll-like receptor-4: renal cells and bone marrow cells signal for neutrophil recruitment during pyelonephritis”. In: *Kidney Int.* 68.6 (2005), pp. 2582–2587. DOI: 10.1111/j.1523-1755.2005.00729.x.

[123] Yuqiu Lu et al. “Single-cell RNA-sequence analysis of mouse glomerular mesangial cells uncovers mesangial cell essential genes”. In: *Kidney Int.* 92.2 (2017), pp. 504–513. DOI: 10.1016/j.kint.2017.01.016.

[124] J. Qiao et al. “FGF-7 modulates ureteric bud growth and nephron number in the developing kidney”. In: *Development* 126.3 (1999), pp. 547–554.

[125] Naoki Nakagawa et al. “Dicer1 activity in the stromal compartment regulates nephron differentiation and vascular patterning during mammalian kidney organogenesis”. In: *Kidney Int.* 87.6 (2015), pp. 1125–1140. DOI: 10.1038/ki.2014.406.

[126] Jennifer L. Fetting et al. “FOXD1 promotes nephron progenitor differentiation by repressing decorin in the embryonic kidney”. In: *Development* 141.1 (2014), pp. 17–27. DOI: 10.1242/dev.089078.

[127] Konstantinos Stamatakis et al. “Identification of novel protein targets for modification by 15-deoxy-Delta12,14-prostaglandin J2 in mesangial cells reveals multiple interactions with the cytoskeleton”. In: *J. Am. Soc. Nephrol.* 17.1 (2006), pp. 89–98. DOI: 10.1681/ASN.2005030329.

[128] Taizo Nakagawa et al. “Roles of PDGF receptor-beta in the structure and function of postnatal kidney glomerulus”. In: *Nephrol. Dial. Transplant.* 26.2 (2011), pp. 458–468. DOI: 10.1093/ndt/gfq468.

[129] Stuart W. Smith et al. “CD248+ stromal cells are associated with progressive chronic kidney disease”. In: *Kidney Int.* 80.2 (2011), pp. 199–207. DOI: 10.1038/ki.2011.103.

[130] Sonja Djudjaj et al. “Keratins are novel markers of renal epithelial cell injury”. In: *Kidney Int.* 89.4 (2016), pp. 792–808. DOI: 10.1016/j.kint.2015.10.015.

[131] Frank Costantini. “GDNF/Ret signaling and renal branching morphogenesis: From mesenchymal signals to epithelial cell behaviors”. In: *Organogenesis* 6.4 (2010), pp. 252–262. DOI: 10.4161/org.6.4.12680.

[132] Halim Khairallah et al. “Claudin-7, -16, and -19 during mouse kidney development”. In: *Tissue Barriers* 2.4 (2014), e964547. DOI: 10.4161/21688362.2014.964547.

[133] Lei Yu et al. “GATA2 Regulates Body Water Homeostasis through Maintaining Aquaporin 2 Expression in Renal Collecting Ducts”. In: *Mol Cell Biol* 34.11 (2014), pp. 1929–1941. DOI: 10.1128/MCB.01659-13.

[134] John K. McGuire et al. “Matrilysin (MMP-7) inhibition of BMP-7 induced renal tubular branching morphogenesis suggests a role in the pathogenesis of human renal dysplasia”. In: *J. Histochem. Cytochem.* 60.3 (2012), pp. 243–253. DOI: 10.1369/0022155411435152.

[135] M. Shibuya. “Structure and dual function of vascular endothelial growth factor receptor-1 (Flt-1)”. In: *Int. J. Biochem. Cell Biol.* 33.4 (2001), pp. 409–420. DOI: 10.1016/s1357-2725(01)00026-7.

[136] S. Davis et al. “Isolation of angiopoietin-1, a ligand for the TIE2 receptor, by secretion-trap expression cloning”. In: *Cell* 87.7 (1996), pp. 1161–1169. DOI: 10.1016/s0092-8674(00)81812-7.

[137] P. Carmeliet et al. “Targeted deficiency or cytosolic truncation of the VE-cadherin gene in mice impairs VEGF-mediated endothelial survival and angiogenesis”. In: *Cell* 98.2 (1999), pp. 147–157. DOI: 10.1016/s0092-8674(00)81010-7.

[138] Qi Ren et al. “Platelet endothelial cell adhesion molecule-1 (PECAM1) plays a critical role in the maintenance of human vascular endothelial barrier function”. In: *Cell Biochem. Funct.* 33.8 (2015), pp. 560–565. DOI: 10.1002/cbf.3155.

[139] R. Schwartz-Albiez et al. “The B cell-associated CD37 antigen (gp40-52). Structure and subcellular expression of an extensively glycosylated glycoprotein”. In: *J. Immunol.* 140.3 (1988), pp. 905–914.

[140] S. J. Davis et al. “The structure and ligand interactions of CD2: implications for T-cell function”. In: *Immunol. Today* 17.4 (1996), pp. 177–187. DOI: 10.1016/0167-5699(96)80617-7.

[141] Sam W. Moore et al. “The ITGB2 immunomodulatory gene (CD18), enterocolitis, and Hirschsprung’s disease”. In: *J. Pediatr. Surg.* 43.8 (2008), pp. 1439–1444. DOI: 10.1016/j.jpedsurg.2007.12.057.

[142] Priya Srinivasan et al. “Signal transducer and activator of transcription 1 negatively regulates constitutive gamma interferon-inducible lysosomal thiol reductase expression”. In: *Immunology* 132.2 (2011), pp. 209–216. DOI: 10.1111/j.1365-2567.2010.03355.x.

[143] øyvind Salvesen et al. “Activation of innate immune genes in caprine blood leukocytes after systemic endotoxin challenge”. In: *BMC Vet Res* 12.1 (2016), p. 241. DOI: 10.1186/s12917-016-0870-x.

3 KINETIC MODELING OF MULTI-OMICS DATA REVEALS MICRORNA-MEDIATED TRANSLATIONAL REGULATION IN STEM CELL DIFFERENTIATION

THIS CHAPTER IS BASED ON:
Patrick van den Berg, Noémie Bérenger-Currias, Marleen Feliksik, Esmée Adegeest, Mazène Hochane, Maria Mircea, Bogdan Budnik, Nikolai Slavov, Stefan Semrau. "Kinetic modeling of multi-omics data reveals microRNA-mediated translational regulation in stem cell differentiation". In: *Unpublished* (2020)[1]

Abstract

Stem cell differentiation is a highly dynamical process involving intricate gene regulatory mechanisms at multiple levels. A lack of detailed understanding of these mechanisms makes it challenging to improve existing differentiation protocols, which are gleaned from in vivo development and are typically slow and inefficient. The large majority of existing studies on differentiation has focused on transcriptional regulation, while the extent and mechanisms of translational regulation are much less explored. Here, we present a time-resolved, multi-omics study of retinoic-acid driven differentiation of mouse embryonic stem cells, comprising mass spectrometry, mRNA-sequencing of cytoplasmic and nuclear fractions, as well as micro-RNA sequencing. We develop a hierarchical kinetic rate model that allows us to integrate these datasets and explore the factors that determine protein levels. While the cytoplasmic-to-nuclear ratio of mRNA only has a minor effect, our model reveals microRNAs that have a significant influence on the translation of their putative targets. Multi-omics factor analysis finally identifies the major biological factors involved in the differentiation process. All in all, our study shows how a refined kinetic model, in conjunction with stringent model selection, can be used to discover regulatory mechanisms in a high-throughput manner, without the need for perturbations.

[1] S.S. and N.S. conceived the project. S.S. acquired funding. N.S. and B.B. supervised and performed the proteomics experiments. S.S., P.v.d.B, N.N. and M.F. performed all other experiments with support from E.A. and M.H.. P.v.d.B. analyzed, interpreted and modeled the data with assistance from M.M.. P.v.d.B. and S.S. wrote the manuscript.

3.1 Introduction

Much of the medical potential of pluripotent stem cells is due to their ability to differentiate into all tissue types of the adult body [2]. While tremendous progress has been made in guiding cells through successive lineage decisions, the gene regulatory mechanisms underlying these decisions remain largely unknown. This gap in knowledge hampers the streamlining and acceleration of differentiation protocols. A large body of work has focused on transcriptional regulation, charting transcriptome changes during differentiation, most recently down to the single-cell level [3, 4, 5, 6, 7]. Gene regulation occurring at the level of translation is much less explored. Most transcriptomics studies make the implicit assumption that mRNA levels are a good proxy for protein levels. It has been shown that in steady state, roughly 40% of protein variability across the proteome, can be explained by differences in mRNA abundance ([8]). Models of the steady-state protein to mRNA ratio (PTR) can explain up to two-thirds of the variability when taking transcript sequence features -such as coding sequence length or amino acid frequencies- into account [9]. In highly dynamical systems, such as differentiating stem cells, protein abundance is typically modeled with differential equations. These models are different from steady-state models in that they cannot explain absolute protein levels, but they can be used to infer kinetic rates for protein synthesis and degradation [10, 11, 12]. Here, we show that such models can also be used to reveal regulatory mechanisms during stem cell differentiation in an unbiased, high-throughput manner. We collected a multi-omics dataset of retinoic acid (RA) driven differentiation of mouse embryonic stem cells. Samples taken over a period of 96h were subjected to: mass spectrometry, bulk RNA-sequencing of nuclear and cytoplasmic fractions, as well as small RNA sequencing to quantify micro-RNA (miR) abundance. To model protein dynamics we refined a birth-death model by considering explicitly the cytoplasmic-to-nuclear ratio of mRNA abundance and the influence of certain technical artifacts related to mass spectrometry. By modeling the influence of miRs on protein synthesis, we identified several miR that likely have a significant influence on protein regulation. Finally, we used multi-omics factor analysis (MOFA) to reveal the overall relevance of translational regulation for in vitro differentiation.

3.2 Results

3.2.1 Pervasive discordance between RNA and protein in retinoic acid driven mESC differentiation

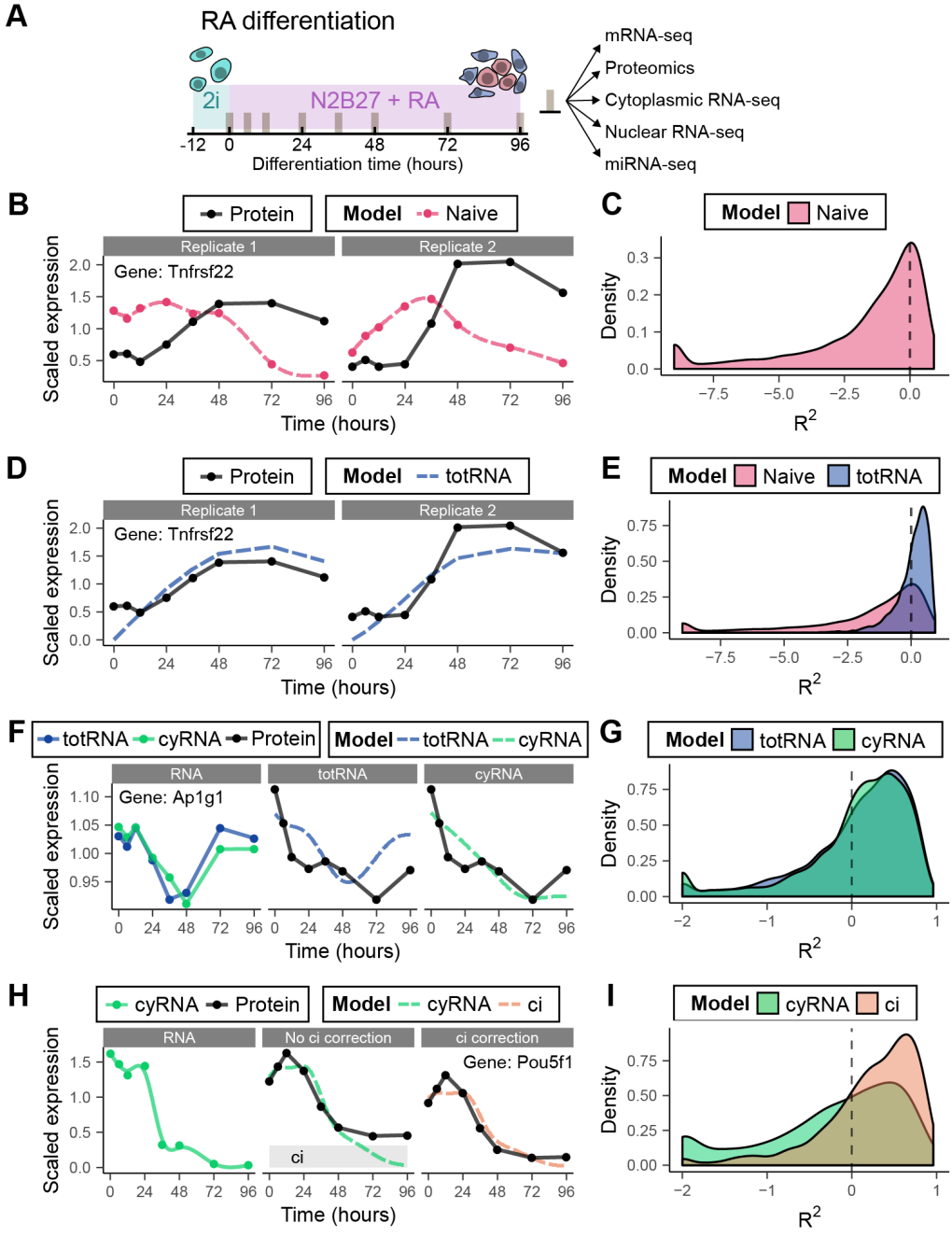
We used RA differentiation of mouse emrbyonic stem cells (ESCs) as a generic model for in vitro differentiation. Previously, we characterized this differentiation assay in detail at the transcriptional level by single-cell RNA-seq [3] In particular, we have shown that within 96 h of RA exposure, mouse ESCs bifurcate into an extraembryonic endoderm-like and an ectoderm-like cell type (XEN and ECT respectively). Here we collected samples during an

RA differentiation time course (Fig 1A). For each time point we quantified total poly(A) RNA by RNA-seq and protein expression by tandem mass tag (TMT) labeling followed by tandem mass spectrometry (MS/MS). In total, we obtained both RNA and protein expression of 6271 of genes (Fig S1A-E) for 8 time points in duplicate. After correction for batch effects due to different runs and sequencing methods (Fig 1H), we achieved highly similar results for the two biological replicates. To investigate, in how far protein expression can be predicted from RNA expression, we started with the simplest conceivable model (termed *naive* here), which assumes that protein expression is identical to RNA scaled with a constant factor. This model is justified if RNA expression changes slowly on the time scale of protein degradation, resulting in a quasi-steady state. Consequently, the PTR would be approximately constant over the time course. To test this model, we scaled both protein and RNA to their respective means, which should result in a constant PTR of 1, if the naive model is valid. We observed that for a large fraction of genes the naive model is inaccurate, resulting in a low coefficient of determination (R^2) and low correlation coefficient (Fig 1C, Fig 2A). For particular genes we could even observe significant anti-correlation between RNA and protein (Fig 1B). This result shows that the assumptions of the naive model are likely wrong for the majority of genes and a more sophisticated model is necessary to explain the relationship between RNA and protein.

3.2.2 Protein turnover model explains RNA-protein discordance for most genes

To relax the assumption that expression is in steady state, we next considered a kinetic model that implements a birth-death process for protein turnover (Eq 3.1). Similar models have been used previously to describe protein turnover during the stress response in yeast [10], as well as embryonic development of *Xenopus* [11] and *Drosophila* [12]. The birth-death model assumes constant rates for protein synthesis (k_s) and degradation (k_d). All processes related to protein production (translation, initiation, elongation, etc) are lumped into k_s, while k_d represents all processes leading to a reduction in protein levels (dilution due to cell division, active degradation, etc.). We do not consider simpler, degenerate models (without k_s and/or

Figure 1 *(following page)*. Protein turnover models outperform the naive model in predicting protein temporal profiles. (A) Schematic overview of RA differentiation time course and subsequent omics measurements. (B) Example fit of the naive model. The naive model is a smoothing spline fit of RNA scaled to match the mean protein expression. (C) R^2 distribution of the naive model. (D) Example fit of the totRNA model. (E) R^2 distribution of the totRNA model model. (F) Example fit of the totRNA and fullRNA model, replicate 1. (G) R^2 distribution of the cyRNA model. (H) Example fit of the ci model, replicate 1. The height of the grey bar indicates the fitted ci parameter. (I) R^2 distributions of the ci model. Only genes that are improved by the ci model are shown. The full distribution of genes is shown in Fig S2E. Some genes with extremely low R^2 values are set to the minimum value of the plot for clarity. Corresponding Pearson's r distributions are plotted in Fig S2.


A
RA differentiation
2i
N2B27 + RA
-12
0
24
48
72
96
Differentiation time (hours)
mRNA-seq
Proteomics
Cytoplasmic RNA-seq
Nuclear RNA-seq
miRNA-seq
B
Protein
Model
Naive
Replicate 1
Replicate 2
Gene: Tnfrsf22
Scaled expression
Time (hours)
C
Model
Naive
Density
R2
D
Protein
Model
totRNA
Replicate 1
Replicate 2
Gene: Tnfrsf22
Scaled expression
Time (hours)
E
Model
Naive
totRNA
Density
R2
F
totRNA
cyRNA
Protein
Model
totRNA
cyRNA
RNA
totRNA
cyRNA
Gene: Ap1g1
Scaled expression
Time (hours)
G
Model
totRNA
cyRNA
Density
R2
H
cyRNA
Protein
Model
cyRNA
ci
RNA
No ci correction
ci correction
Gene: Pou5f1
ci
Scaled expression
Time (hours)
I
Model
cyRNA
ci
Density
R2

k_d [11]), because these models are not biologically meaningful in our opinion. It seems reasonable to assume that synthesis and degradation always occur to some degree. To reduce the influence of uninformative small fluctuations, we applied a smoothing spline to the expression data prior to inferring model parameters by non-linear least-squares fitting. Compared to the naive model R^2 and correlation improved markedly (Fig 1DE, Fig 2B), which might be expected given the increase in model flexibility. To correct for a difference in the number of fit parameters and thus compare model performance fairly, we used the Bayesian information criterion (BIC) (see Methods). According to the BIC, 3551 out of 4580 genes were better fit by the kinetic model. These genes are thus likely out of steady state for the duration of the experiment as a result of the differentiation cue.

$$\begin{aligned} P^g(t) &= k_s^g \cdot R^g(t) - k_d^g \cdot P^g(t) \\ k_s^g &\geq 0, k_d^g \geq 0 \end{aligned} \tag{3.1}$$

In summary, these results showed that a simple birth-death model outperforms the naive model of protein turnover.

Despite the overall improvement observed with the kinetic model, many genes were still not properly fit. We would like to interpret the remaining discrepancies as signs of biologically interesting, dynamic regulation. To be able to do so, we had to exclude technical limitations of our measurements as possible explanations. We first considered the subcellular localization of mRNA. In our first experiment we measured total RNA, whereas only cytoplasmic mRNA is available for translation. Nuclear retention of mRNA was found to reduce variability in cytoplasmic mRNA concentration and thereby protein synthesis. Moreover, specific genes are retained in the nucleus as a form of translational regulation [13, 14, 15]. To measure the cytoplasmic mRNA fraction of each gene, we repeated the differentiation experiment in triplicate and separated cell lysates into a nuclear and cytoplasmic fraction before performing RNA-seq. To obtain a global scaling factor between cytoplasmic and nuclear expression, we regressed totRNA reads, measured previously, on nuclear RNA (nuRNA) and cyRNA reads across all genes (see methods) . Then the cytoplasmic fraction C was calculated for each gene and each time point. To our surprise, C did not vary substantially between genes (Mean= 0.817, Std=0.0161, subset of 3,563 genes without any missing values) (Fig S1F). In addition, C also did not fluctuate much in time for individual genes (Fig S1G). Despite the low variability of C, we incorporated this parameter into our model (Eq 3.2). As expected, adding C brought overall only a very subtle improvement (Fig 1G), although for individual cases, the improvement can be quite significant (Fig 1F). We opted to fit further models including the cytoplasmic fraction due to the overall slightly better performance.

$$\begin{aligned} P^g(t) &= k_s^g \cdot C^g(t) \cdot R^g(t) - k_d^g \cdot P^g(t) \\ 0 &\leq C^g(t) \leq 1 \end{aligned} \tag{3.2}$$

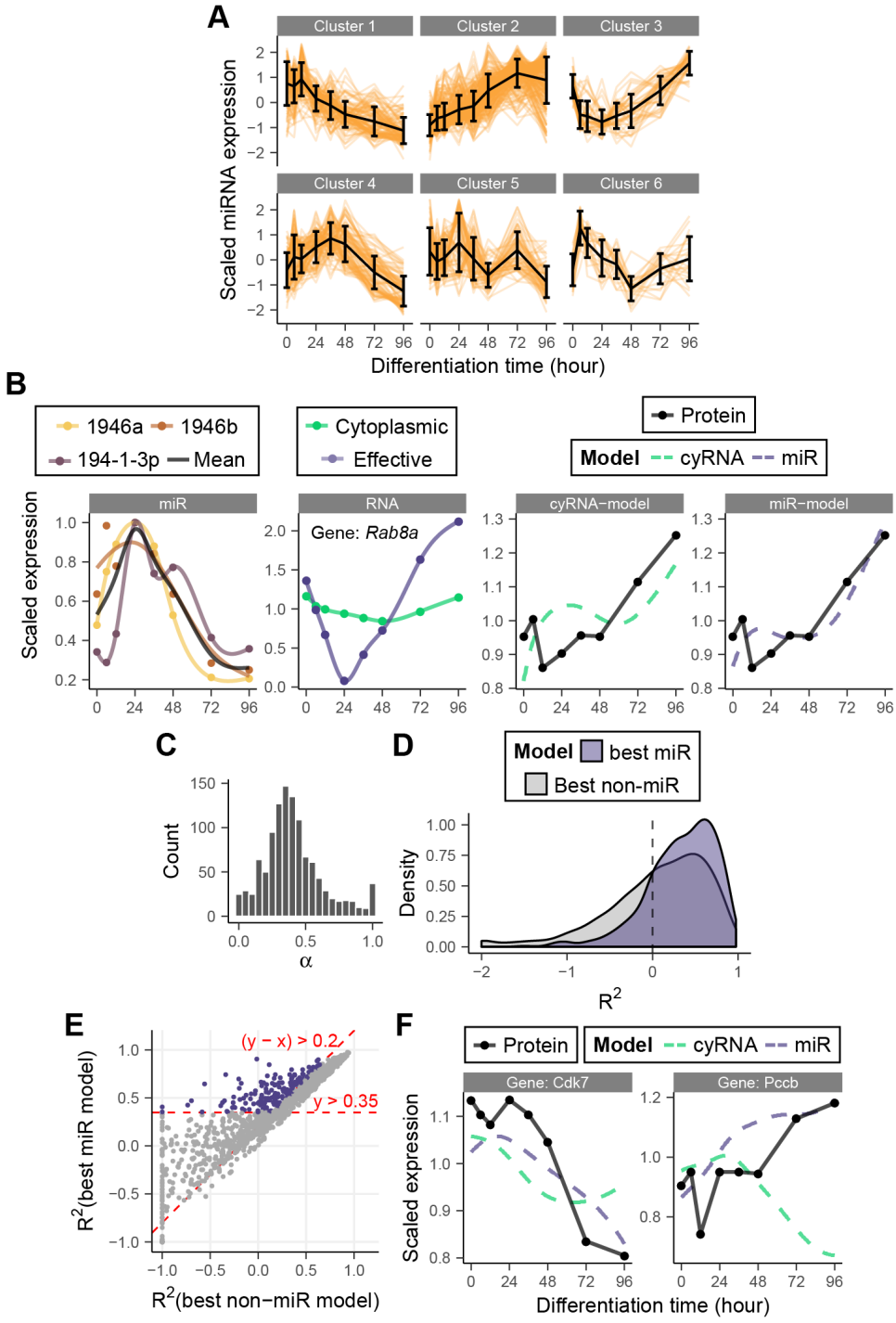
Another important technical limitation was inherent to the proteomics method we employed. TMT-based proteomics suffers from coisolation interference, a process in which two peptides are co-isolated in the second MS step. The contaminating peptide interferes with the quantification of the peptide of interest. To model this contamination, we added an additional parameter (ci) to the model (Eq 3.3), which we assume to be constant for all TMT tags (i.e. time points). Effectively, including this parameter allows protein expression to have a bigger dynamic range, which can improve the fit for certain genes significantly (Fig 1H-I, Fig 1DE). Judged by the BIC, 598 genes were fit better including ci. All in all, this result shows that it is very important to consider co-isolation interference.

$$\begin{aligned} P^g(t) + ci^g = k_s^g \cdot C^g(t) \cdot R^g(t) - k_d^g \cdot (P^g(t) + ci^g) \\ 0 \leq ci^g \leq \min\{P^g(t)\} \end{aligned} \tag{3.3}$$

3.2.3 Including miRs improves model performance and identifies miR-gene interactions

Having ruled out major systematic errors, we were in a position to explore biological explanations for the remaining discordance between RNA and protein. We chose to explore miRs since they are known to play an important role in gene regulation during differentiation. Specifically, we wanted to study the influence of miRNAs on protein translation initiation. In order to study the role of miRs in our system we repeated the differentiation assay and measured the miRnome by small RNA-seq in quadruplicate. We quantified around 1000 mature miRs per time point (Fig 1ACE). To identify possible miR interactions, we used the list of predicted targets provided by TargetScanMouse [16]. We further limited the number of miRs per mRNA by filtering the miR-gene interactions leniently using the context score (Fig 1D). In the end we retained 4527 genes with 560 unique mature miRs and 45,882 potential interactions

Figure 2 *(following page).* The addition of miRs further improves the protein turnover model for a subset of genes and reveals novel candidate miR-gene interactions. (A) Expression profiles of 560 miRs in six clusters. (B) Example fit of miR model for the gene *Rab8a*, replicate 1. First panel: expression of the assigned miRs of a single cluster. Colored lines are individual smoothing spline fits. Second panel: Cytoplasmic RNA expression and the effective RNA concentration available for translation (see Materials and methods). Solid lines represent smoothing splines. Third/fourth panel: cyRNA and miR model fits. (C) Distribution of inferred α for genes that benefit from miR model. (D) R^2 distribution of the miR model and the next best model (either naive, totRNA, cyRNA or ci). Only genes that benefit from the miR model are shown. Some genes with extremely low R^2 values are set to the minimum value of the plot for clarity. Corresponding Pearson's r distributions are shown in Fig S2F. (E) R^2 distributions of (D) compared in a scatter plot. Colored dots are defined by the cutoffs indicated in red and represent a subset of genes with a miR-gene interaction of higher confidence. Some genes with extremely low R^2 values are set to the minimum value of the plot for clarity. (F) miR model fits of two genes (*Cdk7, Pccb*) from the subset highlighted in (E).


A
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
Cluster 6
Scaled miRNA expression
Differentiation time (hour)
B
1946a
1946b
194-1-3p
Mean
Cytoplasmic
Effective
Protein
Model
cyRNA
miR
miR
RNA
cyRNA−model
miR−model
Scaled expression
Gene: Rab8a
C
Count
α
D
Model
best miR
Best non-miR
Density
R2
E
(y − x) > 0.2
y > 0.35
R2(best miR model)
R2(best non−miR model)
F
Protein
Model
cyRNA
miR
Gene: Cdk7
Gene: Pccb
Scaled expression
Differentiation time (hour)

(Fig 1DE).

If multiple miRs with similar temporal profiles target the same gene, we considered them to be indistinguishable. Therefore we globally clustered miRs into six clusters with similar temporal profiles (Fig 2A) and averaged miRs targeting the same gene per cluster (Fig 2B). To keep the model simple we assumed that the inhibitory effect of miRs on protein translation grows linearly with miR abundance (Eq 3.4). We fit one of the six miR clusters at a time and identified the improvement in model performance for each cluster and each gene.

$$\begin{aligned} P_m^g(t) = k_s^g \cdot (1 - \alpha_m^g \cdot M_m^g) \cdot C^g(t) \cdot R^g(t) - k_d^g \cdot P^g(t) \\ 0 < \alpha_m^g \leq 1 \end{aligned} \tag{3.4}$$

Including miRs greatly improved the fits for some genes, especially when there is a transient discordance between RNA and protein expression (Fig 2B). Typically, the "effective" mRNA abundance (cytoplasmic mRNA corrected for miR effects) was more dynamic than nominal mRNA abundance. For many of the genes that benefit from the addition of miRs, their influence is typically large. For these genes, 50% of translation is blocked on average at peak miR expression (Fig 2C). Overall, the addition of miRs significantly improved the coefficient of determination for a quarter of the genes (Fig 2D).

To use the model for identifying novel miR-gene interactions, we ranked the genes by the quality of the model fit and model performance improvement compared to the simpler models without miRs (Fig 2E, Suppl Table 2). Among this list of candidate genes we selected seven genes and their putative miRs (Fig 2BF, Fig S3), whose interaction we intend to validate in a follow-up study. (*Rab8a*, *Cdk7*, *Pccb*, *Acad8*, *Mfge8*, *Eif4h* and *Srgap2*).

3.2.4 The best protein turnover model explains 45% of total protein variance

While each model refinement introduced above improved model performance overall, each discussed model was optimal for a subset of genes, judged by the BIC (Fig 3A). In about 16% of cases the naive mode was optimal, meaning that for these genes none of the protein turnover models improve prediction by a significant amount. 25% and 26% of genes are best predicted with the kinetic model without or with considering mRNA localization, respectively. So for 51% of genes, protein expression is out of steady state, but explainable by a simple model with fixed synthesis and degradation rates. For 8% of genes the model including co-isolation interference was optimal. The increased relative dynamic range due to subtracting a constant increased the fit for these genes significantly. Finally, 25% of genes were fit optimally with one of the miR clusters, meaning that translational regulation plays a significant role for these genes. All things considered, 84% of genes were insufficiently described by RNA alone, leading to a very significant lack of variance explained (Fig 3BC). Therefore, it seems in general not advisable to consider mRNA abundance a good proxy for protein levels in a highly dynamical setting.

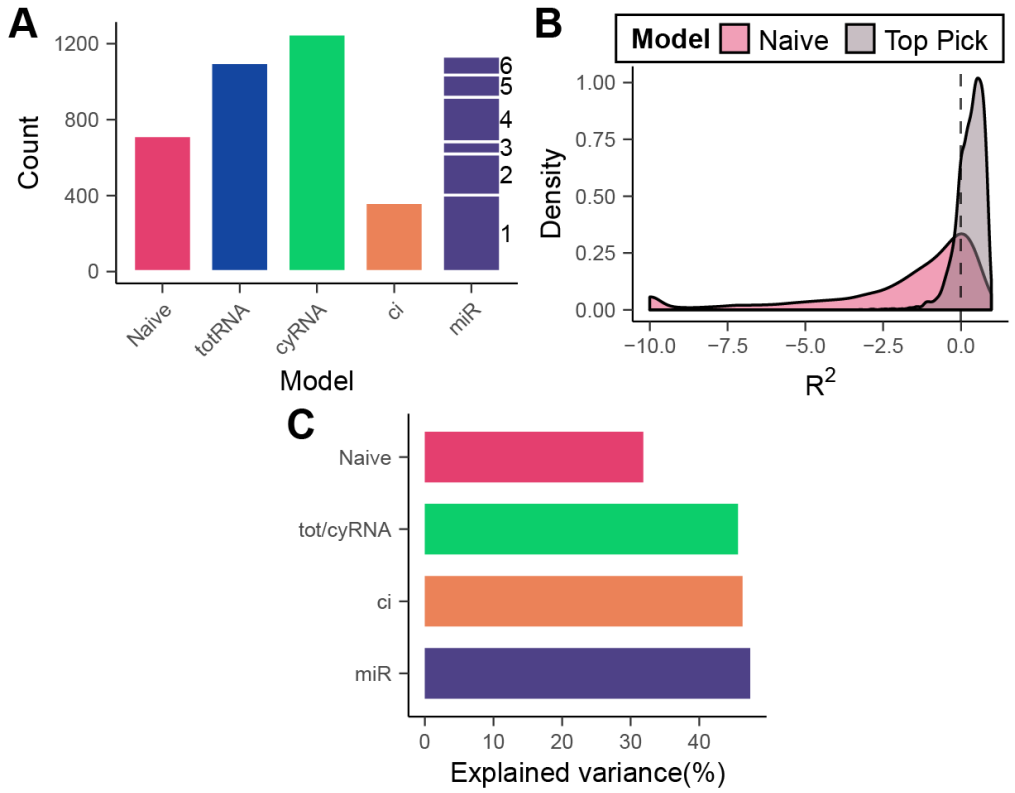


Figure 3. Selecting the optimal model on a gene-by-gene basis increases the total explained variance of protein expression from 30% to 50%. (A) Assignment of the optimal model for each gene based on BIC. The number next to the miR bar indicates the miR cluster giving the best fit. (B) R^2 distribution of the optimal fits from (A) and their naive model counterpart. Some genes with extremely low R^2 values are set to the minimum value of the plot for clarity. (C) Median percentage of protein variance explained by each model, picking among the best models progressively. Fits with negative R^2 were ignored.

3.2.5 Multi-omics factor analysis reveals global factors driving translational regulation in mESC differentiation

The above analyses focused on how individual protein turnover is regulated. In order to describe translational regulation in our system in a more comprehensive manner we performed MOFA. MOFA is an extension of factor analysis which integrates omics data from different sources, like the genome, transcriptome or metabolome. MOFA extracts low dimensional, latent factors that represent the biological processes underlying the variation observed in the data. In order to identify factors involved in translation only, we used two datasets: PTR and miR abundance (Fig 4A). MOFA is typically run directly on gene expression, but here we used the PTR, because we were most interested in explaining how post-transcriptional

regulation can drive fluctuations away from steady state. From this analysis we obtained six driving factors of translational regulation (Fig 4B).

The six factors had distinct temporal trends (Fig 4C), which we used to interpret their biological identity. Since MOFA sorts the factors in order of variance explained and the first three factors had the simplest trends we restricted our interpretation to these three factors (Fig 4D). The genes contributing most to factor 1 are enriched for gene ontology (GO) terms related to development, differentiation, cell cycle and nucleotide synthesis (Fig 4E). Thus, we interpret this factor as the main differentiation program which includes changes in metabolism. This interpretation is supported by the miR that contribute most to factor 1 (Table 3.1): Several of these miRs are known to be involved in differentiation. The let-7 family in particular is known to play an important role in embryonic stem cell pluripotency and self-renewal. Factor 2 is enriched for GO terms in morphogenesis, cell adhesion and signaling. It appears to be the factor driving the specification of the cell type as a response to the external signals. Top miRs of this factor are involved in multiple differentiation paths (osteocytes, adipocytes, trophectoderm, neurons, see Table 3.1). Finally, Factor 3 appears to be related to epithelial-mesenchymal transition (EMT) and mesenchymal-epithelial transition (MET) as many of the top miRs are involved in EMT various types of cancers. Moreover, one of the few enriched GO terms is "epithelial tube morphogenesis". All in all, this analysis shows that meaningful biological factors can be revealed by considering protein synthesis and its regulation.

Factor	Rank	miR symbol	Short description
1	$\bigtriangledown$ 1	miR-1843a-5p	Differentially expressed in traumatic brain injury [17].
1	$\bigtriangledown$ 2	miR-27b-5p	Clustered with miR-23b. Upregulated in *Smad4* knockout cardiomyocytes, involved in cardiac hypertrophy [18]. Induces EMT in gastric cancer [19].
1	$\bigtriangledown$ 3	miR-23b-5p	Clustered with 27b. Regulates osteoclast differentiation [20]. Attenuates glucose-mediated EMT in diabetic nephropathy [21].
1	$\bigtriangledown$ 4	let-7d-3p	let-7 family is involved in pluripotency and self-renewal, and is differentially expressed between different ESC states [22]. 7d: inhibitions of leads to EMT in idiopathic pulmonary fibrosis [23].
1	$\bigtriangledown$ 5	let-7f-5p	
1	$\bigtriangleup$ 1	miR-34c-5p	Downregulation of this miR promotes EMT in breast cancer.[24]
1	$\bigtriangleup$ 2	miR-34c-3p	
1	$\bigtriangleup$ 3	miR-10a-5p	Critical for smooth muscle cell differentiation from mESC [25].
1	$\bigtriangleup$ 4	miR-9-3p	Involved in neurogenesis [26]. Suppressor of EMT in nasopharyngeal carcinoma. [27]
1	$\bigtriangleup$ 5	miR-34b-5p	See miR-34c.
2	$\bigtriangledown$ 1	miR-7b-5p	Represses self-renewal [28]. Reverses EMT in breast cancer through *STAT3* [29].

Continued on next page

Continued from previous page

Factor	Rank	miR symbol	Short description
2	▽ 2	miR-3058-3p	
2	▽ 3	miR-195a-3p	Inhibits adipocyte differentiation [30]. Inhibits EMT in Prostate cancer through *FGF2* [31]. Inhibits EMT in colorectal cancer through *NOTCH2* [32].
2	▽ 4	miR-3095-5p	
2	▽ 5	miR-187-3p	Inhibits osteogenic differentiation [33]. Inhibits EMT in hepatocellular carcinoma [34].
2	△ 1	miR-297c-3p	
2	△ 2	miR-297a-3p	Involved in trophectoderm specification in mouse [35].
2	△ 3	miR-297b-3p	
2	△ 4	miR-466f-3p	In cluster with each other. Inhibits *NeuroD1*, which is required for neuron differentiation [36].
2	△ 5	miR-669f-3p	
3	▽ 1	miR-770-3p	Inhibits EMT in non-small cell lung cancer [37].
3	▽ 2	miR-760-3p	Inhibits EMT in breast cancer [38].
3	▽ 3	miR-1306-5p	Involved in hepatocellular carcinoma, regulates *Snail*-mediated metastasis [39].
3	▽ 4	miR-301b-5p	Promotes proliferation, mobility and EMT in bladder cancer by targeting *EGR1* [40].
3	▽ 5	miR-369-5p	The -3p variant targets *Sox4* [41].
3	△ 1	miR-452-3p	Inhibits EMT in hepatocellular carcinoma through TGF-β1 [42].
3	△ 2	miR-340-5p	Targets *Bcl-w* and *Sox2* and inhibitions of miR promotes cancer progression [43].
3	△ 3	miR-186-3p	Affects EMT through Cdc42 in lung cancer [44].
3	△ 4	miR-700-5p	
3	△ 5	miR-106a-5p	Downregulates Twist1 which causes EMT [45].

Table 3.1. Top 5 negative and positive miRs for the first three MOFA factors.

3.3 Discussion

Widespread discordance between steady-state protein and mRNA levels has been observed in several mammalian systems [46, 47, 48]. Importantly, low correlation does not immediately imply a significant role of gene-specific regulation, as technical noise tends to reduce the observed correlation and conventional correction schemes typically ignore the effect of systematic, correlated errors [49] Edfors et al. [48] showed recently that the PTR for a specific gene is constant across several tissues [48]. While the PTR might allow the prediction of absolute protein levels, it is unable to capture relative changes over time or relative differences

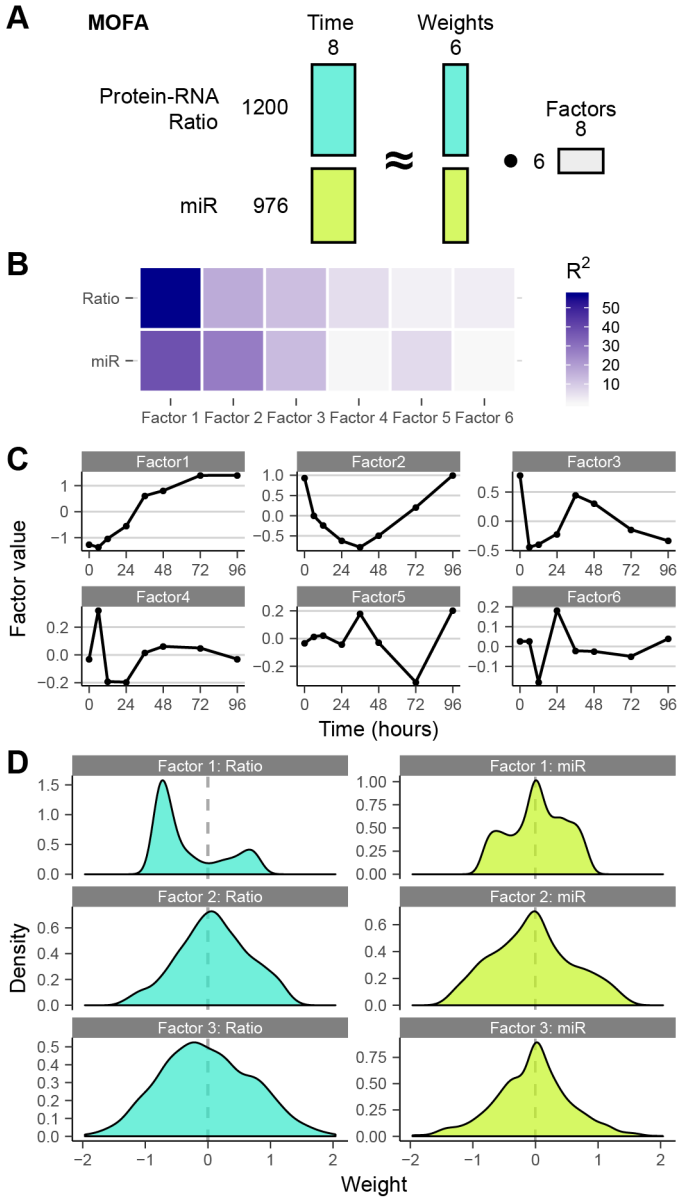


Figure 4. Multi-omics factor analysis uncovers underlying factors that drive translational regulation in mESC differentiation. **(Continued on next page)**

Figure 4. Multi-omics factor analysis uncovers underlying factors that drive translational regulation in mESC differentiation. (Continued from last page) (A) Schematic of the matrices used in MOFA analysis. (B) Variance explained (R^2) by each of the six factors. (C) Temporal profile of each factor. (D) Distributions of the factor weights for the first three factors. (E) GO term over and under representation by the top (+) and bottom (-) 120 genes of the first three factors. The top 10 enriched GO terms (with p<0.1) is shown for each factor and direction. Test performed is elim with KS statistic, see Materials and methods).

between tissues [50].

To identify dynamic translational regulation during stem cell differentiation, we therefore collected a time-resolved gene expression dataset and found overall low correlation between mRNA and protein abundance across time. Such low correlation has been observed recently in several other systems, in particular: *Xenopus* development [11], *C. elegans* development [51], macrophage differentiation [52] and mouse ESC differentiation [53]. While the lack of strong correlation is typically interpreted as a sign of (post) translational regulation [51, 53], theoretical work showed that a simple delay between mRNA and protein production can lead to a reduction in gene-wise correlation [54, 55] A simple model with constant kinetic rates explained the protein dynamics of a third of all genes during stress response in yeast [10] and of 75% of all genes in *Xenopus* development [11]. In our system, 3551 out of 4580 genes were explained better by this model, compared to a naive model which assumes a constant PTR.

To explain the remaining discordance we explored the cytoplasmic-to-nuclear ratio of mRNA abundance, but did not find a strong effect. On the other hand, including a parameter modeling co-isolation interference markedly improved the fit for some genes. We therefore posit that co-isolation interference should be included in any kinetic model when TMT labeling is used for multiplexing the mass spectrometry measurements.

miRs have been identified as a key regulator of stem cell pluripotency and differentiation [56]. For example, members of the the let-7 and miR-290 families have been implied as drivers for differentiation of ESCs as well as in the maintenance of pluripotency [57, 22, 56, 58] To find putative targets of miRs, various computational methods, typically based on sequence complementarity and conservation, have been developed [16, 59, 60]. These methods predict hundreds of thousands of interactions, among which are likely many false positives. The gold standard for validation, the luciferase assay, is time consuming, which means that the majority of potential interactions have not been verified. To our knowledge, there is currently no high-throughput experimental method to identify miR-mediated translational regulation in a genome- and miRnome-wide manner. We believe that our modeling approach is able to reduce the number of potential interactions to a much smaller set, which can be easily validated by conventional methods.

In addition to the possibility to infer regulatory interactions between different molecular players, multi-omics data sets are also useful to identify major driving factors of biological processes in development and disease. A living cell is typically considered to be a highly complex dynamical system that defies many traditional modeling approaches due to the large amount of unobserved or indeterminable parameters. There is, however, the hope that many biological processes in fact occur on low-dimensional manifolds within the high-dimensional space needed to describe the state of a cell. Multi-omics measurements will allow us to ascertain if there are in fact such manifolds, which would significantly simplify a complete quantitative understanding of biological dynamics. Our study indicates the presence of at least 3 factors that co-regulate miR and protein abundance during differentiation. To unravel how co-regulation is achieved molecularly and how the factors can be perturbed

to in vitro differentiation are fascinating challenges for the future.

3.4 Materials and methods

3.4.1 Cell culture

E14 mouse embryonic stem cells were cultured as previously described [3]. Briefly, cells were grown in modified 2i medium [61]: DMEM/F12 (Life technologies) supplemented with 0.5x N2 supplement, 0.5x B27 supplement, 4mM L- glutamine (Gibco), 20 μg/ml human insulin (Sigma-Aldrich), 1x 100U/ml penicillin/streptomycin (Gibco), 1x MEM Non-Essential Amino Acids (Gibco), 7 μl 2-Mercaptoethanol (Sigma-Aldrich), 1 μM MEK inhibitor (PD0325901, Stemgent), 3 μM GSK3 inhibitor (CHIR99021, Stemgent), 1000 U/ml mouse LIF (ESGRO). Cells were passaged every other day with Accutase (Life technologies) and replated on gelatin coated tissue culture plates (Cellstar, Greiner bio-one).

3.4.2 Retinoic acid differentiation and sample collection

Retinoic acid induced differentiation was carried out exactly as described before [3]. Prior to differentiation cells were grown in 2i medium for at least 2 passages. Cells were seeded at 2.5e5 per 10 cm dish and grown over night (12 h). Cells were then washed twice with PBS and differentiated in basal N2B27 medium (2i medium without the inhibitors, LIF and the additional insulin) supplemented with 0.25 μM all-trans retinoic acid (RA, Sigma-Aldrich). Spent medium was exchanged with fresh medium after 48 h. To collect samples, cells were dissociated with Accutase and spun down. Full RNA and cytoplastmic/nuclear RNA were always immediately extracted (RNeasy, Qiagen and SurePrep, Fisher Scientific, resp.) and the purified RNA was stored at -80C until RNA-sequencing was performed. For proteomics and miR-sequencing, pellets were flash frozen in liquid nitrogen and stored at -80C until further processing.

3.4.3 RNA and miR sequencing

The libraries for RNA sequencing were prepared under standard conditions using Illumina's TruSeq stranded mRNA sample preparation kit. The stranded single end libraries were sequenced using Illumina HiSeq at 40bp with an average read depth of 40 million reads per sample. Paired-end libraries for RNA sequencing were sequenced on an Illumina NextSeq 500 at 150bp per strand at a read depth of 10 million reads per sample. miRs were extracted from frozen pellets using miRNeasy (Qiagen) kit. Libraries for small RNA sequencing (miR sequencing) were prepared using NEBNext Small RNA Library Prep Set for Illumina (New England Biolabs) and were sequenced on an Illumina NovaSeq 600 at 150bp paired-end with a range of 4 to 15 million samples. Specification table for the sequencing strategy is available upon request. All sequencing data is available through GEO.

3.4.4 Mass spectrometry

Pelleted cells were lysed in 400 µl RIPA buffer, except for the sorted cells. Volumes of cell lysate corresponding to 100 µg protein per sample were digested with trypsin using a modified FASP protocol [62]. Subsequently each sample was labeled with TMT 10-plex, 6-plex or 11-plex reagent (Thermo Fisher)) according to the manufacturer's protocol. All labeled samples were combined into a set-sample. Which labels were assigned to each sample is specified in the specification table, which is available upon request. The labeled set-sample was fractionated by electrostatic repulsion-hydrophilic interaction chromatography (ERLIC) run on an HPLC 1200 Agilent system using PolyWAX LP column (200x2.1 mm, 5 µM, 30nm, PolyLC Inc, Columbia, MD) and a fraction collector (Agilent Technologies, Santa Clara, CA). Set-samples were fractionated into a total of 40 ERLIC fractions. Each ERLIC fraction was subsequently further separated by online nano-LC and submitted for tandem mass spectrometry analysis to both LTQ OrbitrapElite or Q exactive high field (HF). One third of each fraction was injected from an auto-sampler into the trapping column (75 um column ID, 5 cm length packed with 5 um beads with 20 nm pores, from Michrom Bioresources, Inc.) and washed for 15 min;; the sample was eluted to analytic column with a gradient from 2 to 32 % of buffer B (0.1% formic acid in ACN) over 180 min gradient and fed into LTQ OrbitrapElite or Q exactive HF. The instruments were set to run in TOP 20 MS/MS mode method with dynamic exclusion. After MS1 scan in Orbitrap with 60K resolving power, each ion was submitted to an HCD MS/MS with 60K resolving power and to CID MS/MS scan subsequently. All quantification data were derived from HCD spectra.

3.4.5 RNA-seq processing

We used genome assembly mm10 release 93 from Ensembl. First an RSEM (v1.3.1) reference was created with default settings. Next we performed adapter and quality trimming with Trimmomatic (v0.38). Finally all reads were aligned with RSEM with STAR (v2.6.1a) with the option enabled for stranded libraries. Expected counts from RSEM were used as input for DESeq2 (v1.26) to obtain regularized $\log_2$ counts with stabilized variance to make comparisons between samples more reliable. From these values we obtained regularized counts and used these for all further analyses and as input for batch correction.

3.4.6 Proteomics processing

Peptide search was performed on peptides identified in full RNA seq data to increase specificity of the protein quantification with MaxQuant. Proteins were quantified from the peptide measurements in the evidence.txtoutputs. Reversed peptides and contaminants were removed. Each column in the file was then normalized to the mean. Some peptides for some samples were quantified multiple times, due to multiple mass-spectrometry runs or multiple tmt tags in the same sample mix. These values were averaged. Multiple peptides assigned

to a single ensembl gene ID were also averaged to obtain normalized protein expression that we used for batch correction.

3.4.7 Batch correction

We observed global expression differences in protein and rna expression depending on the seeding day which we decided to correct using the RemoveBatchEffect function from limma (v3.42.2). We applied it on the protein and totalRNA datasets separately with three different levels of batches, one each for the first replicate, the second replicate and for two samples we used to replace failed samples of the first replicate (Specification table available upon request). The resulting batch-corrected values were used as input for most further analysis (Fig S1H).

3.4.8 miR-seq processing

For alignment of miR-seq data we used the same genome release as above with miRnome release 22.1 from miRBase using the mature miR sequences. To prepare the reads we performed adapter and quality trimming with Trimmomatic and obtained a consensus forward sequence using both the forward and reverse read and PEAR (v0.9.6). We next ran bowtie-prepare from bowtie (v1.0.0-1). Finally we quantified each sample with the mapper.pl and quantifier.pl scripts from mirdeep2 (v2.0.1.2). The obtained counts were processed the same way as the RNA-seq data, but separately.

3.4.9 miR-gene interactions

We looked for putative miR-gene interaction using TargetScanMouse release 7.1. We filtered the "miR family" table for expressed miRs and expressed RNAs. We next did a lenient filter on the miR-gene interactions; we filtered out all interactions with a cumweightscore lower than -0.3. Finally, we wanted to keep only miRs with high dynamics over the time course and high reproducibility. To achieve this we calculated for each miR the Coefficient of variation (CV) across the mean miR expression of each time point and the mean of the CV's across the biological replicates. We fit a gaussian mixture model to these to values using mclust (v5.4.6) where each distribution has an equal diagonal shape, but with varying volumes ("VEE" modelNames option). We filtered the miR-gene interactions list for miRs from cluster 1 because they fit our high variance and high reproducibility criteria (Fig S1I). The final putative list of miR-gene interactions comprised 560 miRs and was used in the miR clustering and model fit (see below).

3.4.10 miR clustering

To cluster miRs into sets of similar temporal profiles, miR expression of the miRs in the interaction table was first averaged per time point. A miR to miR distance matrix was created with

1–Pearson correlation on $\log_2$-transformed values. This matrix was then used to perform hierarchical clustering with complete linkage (base R) and the resulting dendrogram was cut into 6 clusters.

3.4.11 C-fraction calculation

To obtain a per gene C-fraction we needed to map how cytoplasmic and nuclear sequencing reads relate to each other. We took total, cytoplasmic and nuclear RNA and removed genes that had any raw count lower than 10. We then took the top 500 genes with the lowest variance and, under the assumption that for these genes that C-fraction was the most stable, fit the following linear model: $R_{tot} = \beta_c \cdot R_c + \beta_n \cdot R_n$, where R_{tot} is total RNA, R_c is cytoplasmic RNA and R_n is nuclear RNA. T For each we applied the regularized $\log_2$ counts. The fit beta parameters map cytoplasmic and nuclear values to total RNA values and were 0.815 and 0.183 respectively. Then for each gene g we calculated the C-fraction: $C^g = \frac{\beta_c \cdot R_c^g}{\beta_c \cdot R_c^g + \beta_n \cdot R_n^g}$

3.4.12 Rate model fitting

We fit several rate models for every gene in our clean set of genes (Fig S1E). We first scaled totRNA and protein expression by dividing each by its replicate mean. We next with a smoothing spline (smooth.spline function, base R) to the totRNA data for each replicate with 7 degrees of freedom (DF). We opted for manually setting the DF over letting the function determine it because we observed that what seems like noise at the RNA level is sometimes replicated at the protein level. smooth.spline would sometimes oversimplify the dynamics of RNA and this would lead to bad fits at the protein level if the protein has more dynamics than the resulting spline. Perhaps counterintuitively, we deemed the fit more conservative if we forced high dynamics at the RNA level at the cost of introducing some noise. We let smooth.spline determine the DF for all other smoothing spline fits using leave-one-out-cross-validation. We fit smoothing splines to the C-fraction and multiplied this with the smooth totRNA to get smooth cyRNA. miRs that were assigned to each gene were first averaged over replicates and then divided by the miRs maximum value. Smooth splines were fit to each miR and then the smooth miRs for each miR cluster were averaged. We solved the differential equation using deSolve (v1.28), given a rate model, parameters, totRNA or cyRNA and a miR cluster. Parameters put into the solvers were $\log_2(k_{prod})$, $\log_2(k_{div})$, P_0, ci and α, depending on the differential equation being solved. $k_{prod} = k_s \cdot k_d$ and $k_{div} = k_s/k_d$, which are perpendicular to the original values on purpose because the optimization algorithms sometimes had difficulty finding an optimal fit, because covarying k_s and k_d may result in a very similar fit. We $\log_2$ transformed these values to give the optimization function more control over fitting it. P_0 is the protein concentration at t=0h, we opted for adding this as a model parameter instead of setting it to the observed concentration at t=0h and losing that value for fitting. ci is set to be between 0 and the minimum

observed protein expression, because the consolation interference cannot exceed what was actually observed. α is between 0 and 1 and describes the miR clusters' influence on translation rate. Since both α and the miR clusters' expression cannot exceed 1, only at peak miR cluster expression can translation be completely turned off in the model. To find the optimal parameters we fit using the optim function (base R). Sum of squared residuals (SSR) were minimized using the "L-BFGS-B" method of the optim function. However for both Eq 3.2 and 3.3 we first minimized: $SSR + 10 \cdot \log_2(k_{div}) \cdot 2$. Due to our scaling of RNA and protein $\log_2(k_{div})$ is expected to be close to 0 so we penalized divergence first to get a decent estimate for $\log_2(k_{prod})$ first, since we observed that otherwise the fits sometimes had extreme values. The resulting parameters were used as starting values for the unpenalized fits. We used BIC to compare models with different numbers of parameters: $BIC = k \cdot \ln(n) - 2 \cdot SLL$, where k is the number of parameters, n is the number of samples, and SLL the sum of log-likelihood. k is 0 for the naive model, 3 for Eq 3.1 and 4 for Eq 3.2,3.3. n=8, the number of time points. The error of the fits was assumed to be normally distributed in order to calculate the LSL. When comparing models, the model with the lower BIC was considered superior.

3.4.13 MOFA analysis

We performed MOFA analysis to identify factors that drive translation. In contrast to what MOFA is intended for we applied it on the PTR ratios, to only look at the shifts out of steady state and not the actual values. Since we did not have matching protein-rna-miR samples we averaged ratios and miR to each time point. We filtered miRs for a minimum raw count of 10 and obtained 976 miRs. We then selected the top 1200 genes with the highest PTR ratio CV to obtain a similar number as recommended by the authors. MOFA was run in R with the default settings (MOFA2, v1.0).

3.4.14 GO term enrichment

We performed GO term enrichment analysis on the genes with extreme weights in the MOFA analysis. The top 120 positive and negative genes of each factor were used as ordered input for the topGO package (v2.38.1) We used a minimum term size of 20, the "ks" statistic and the "elim" algorithm. The elim algorithm takes the neighboring GO terms into account when calculating p-values. As a result, these p-values are not independent and therefore the authors do not recommend correcting for multiple hypotheses [63].

3.5 Supplementary information

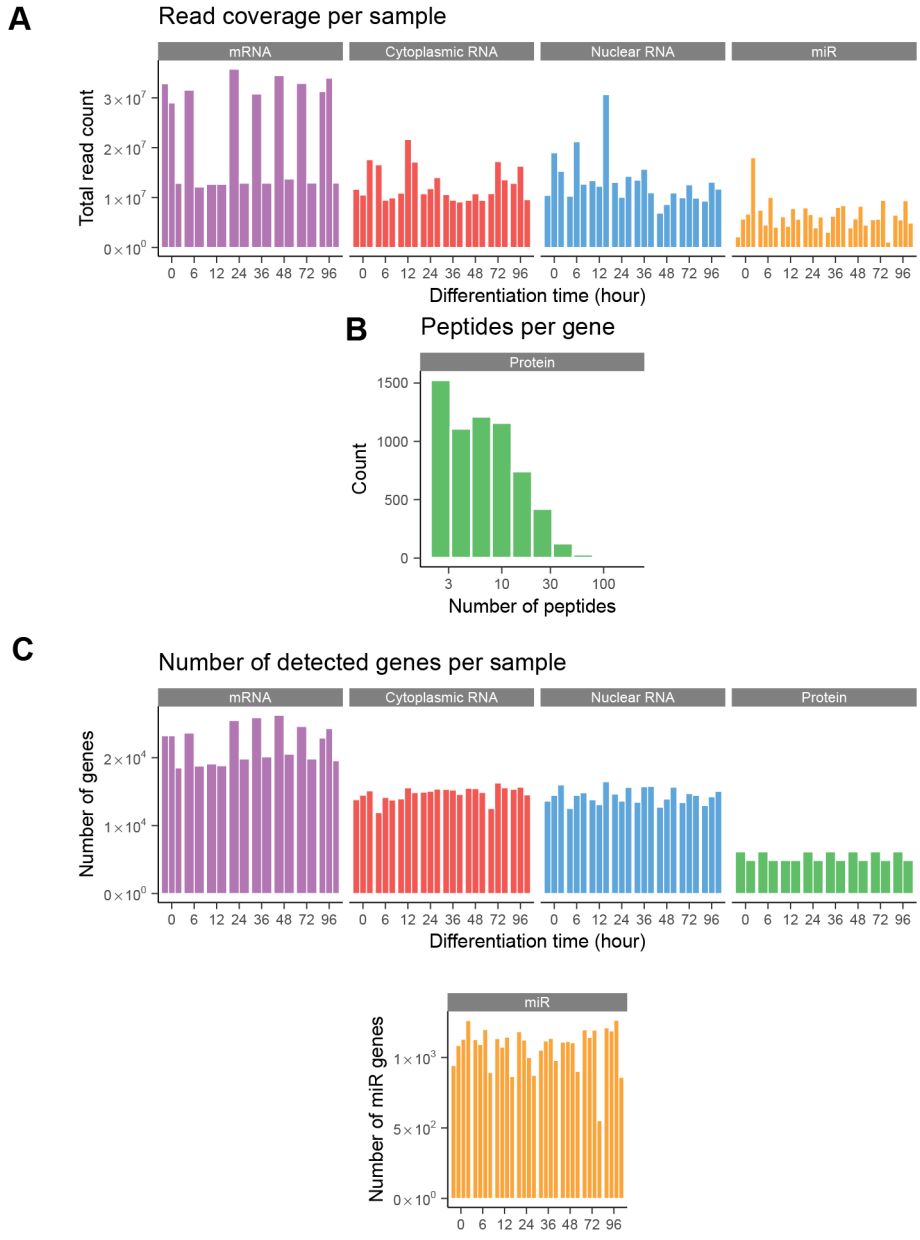


Figure S1. Quality control of full, cytoplasmic, nuclear RNA-seq, miRNA sequencing and proteomics. **(Continued on next page)**

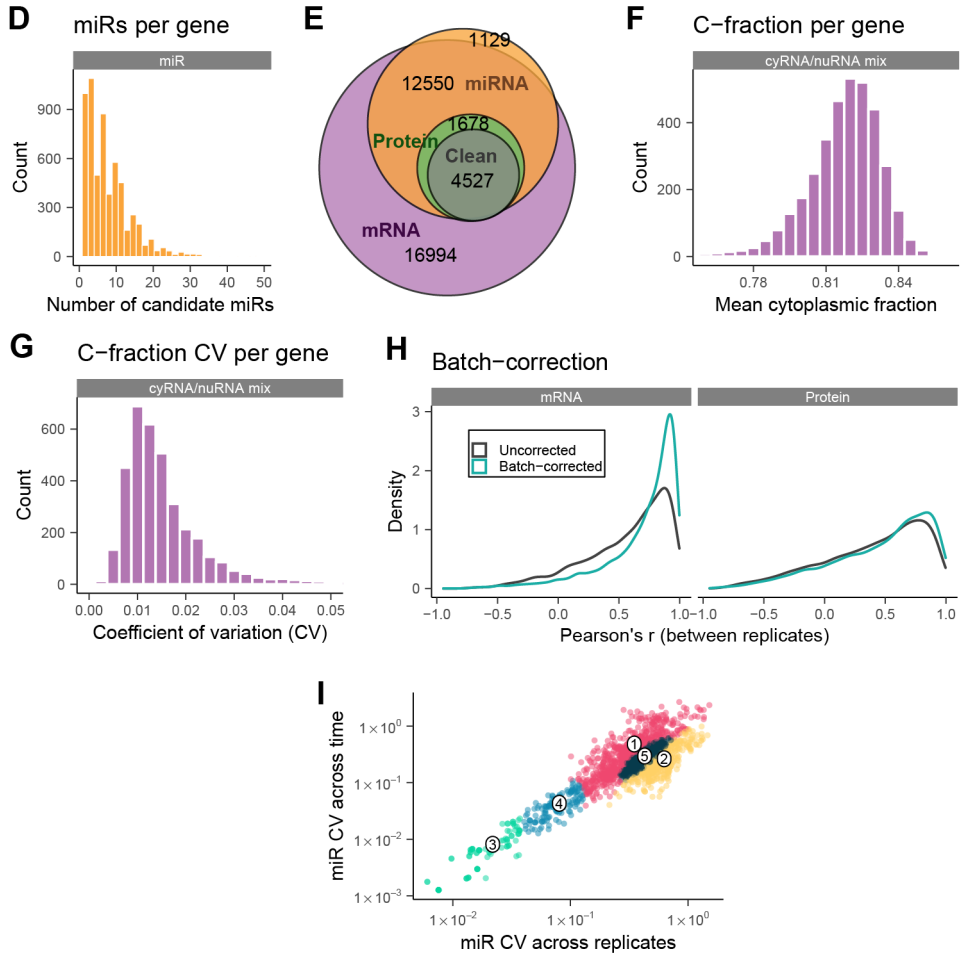


Figure S1. Quality control of full, cytoplasmic, nuclear RNA-seq, miRNA sequencing and proteomics. (Continued from last page) (A) Total number of reads for all sequencing samples. (B) Distribution of the number of peptides used for the quantification of each protein. (C) Number of detected genes or miRs in each sample. Individual replicates are plotted as separate bars in (A, C). (D) Distribution of miR-gene interactions per gene. (E) Euler diagram of all gene sets. The "miRNA" set indicates genes with predicted miR interaction and the "clean" set is a subset of genes without missing values in either RNA or protein. 53 genes are in the set: RNA&Protein&Clean (no miR-gene interactions), 13 genes are in the set: RNA&Protein (no miR-gene interactions, and some genes have missing values). (F) Distribution of the mean cytoplasmic fraction (C-fraction) per gene. (G) Coefficient of variation of C-fraction per gene. (H) Distribution of RNA-RNA and protein-protein Pearson's r with and without batch correction. (I) Gaussian mixture model based clustering of miRs to select a cluster with low noise and high variance (cluster 1), see Materials and methods.

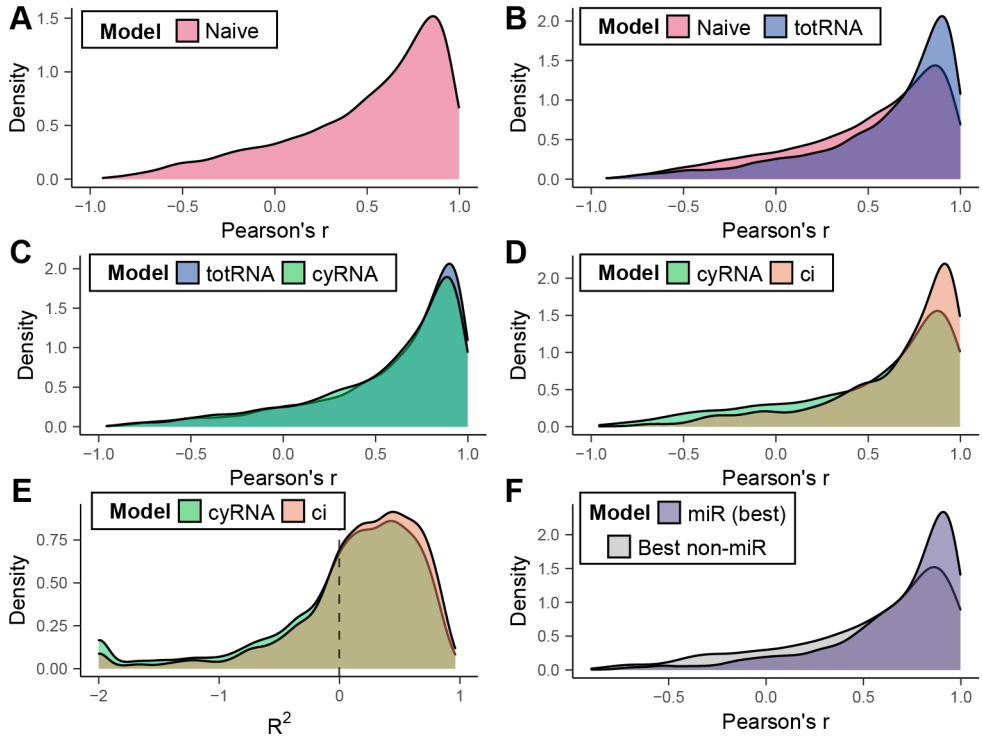
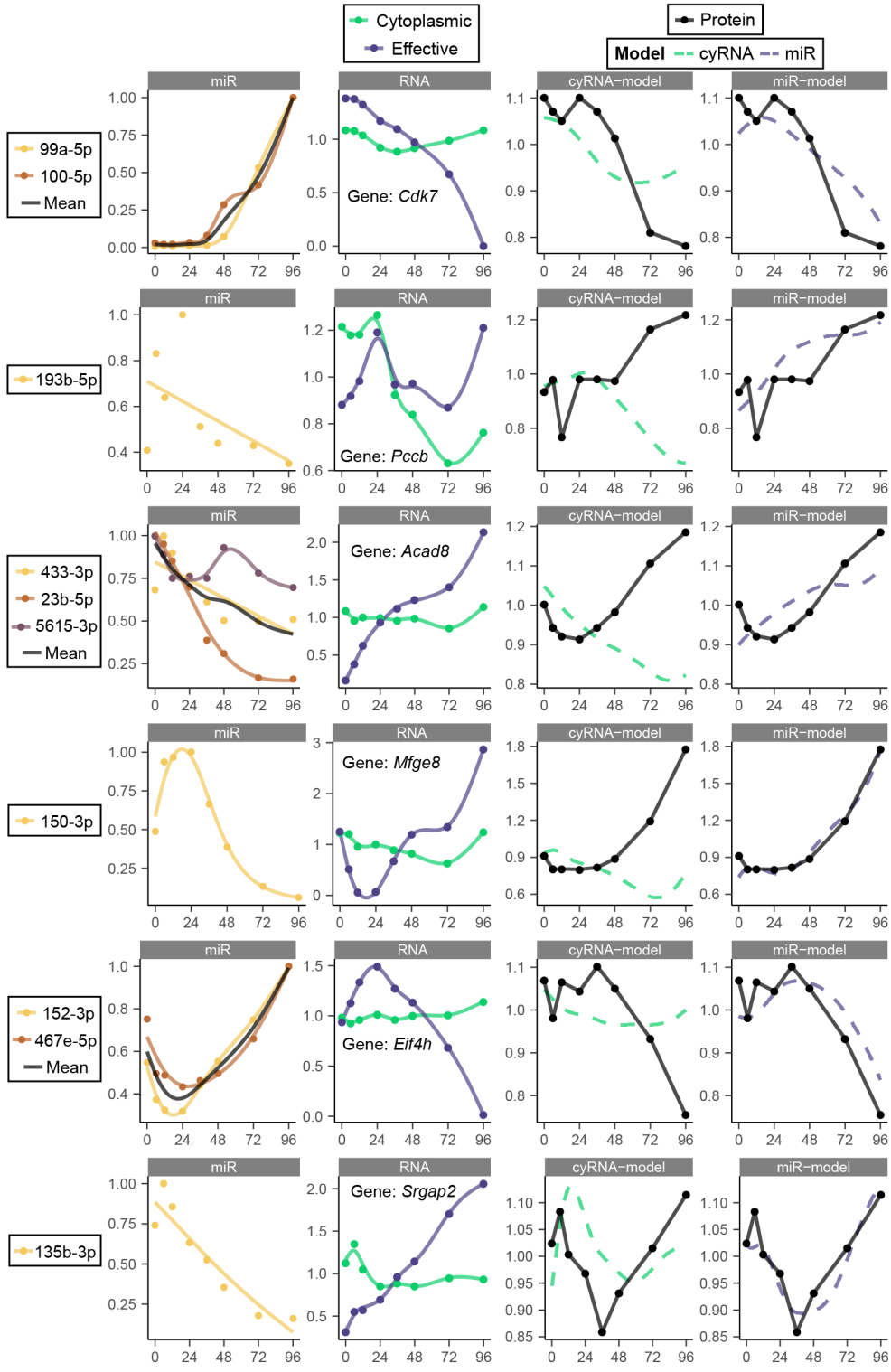


Figure S2. Model performance comparison with Pearson's r. (A-D) Pearson's r distribution of various protein models. Corresponding R^2 distributions are shown in Fig 1. (E) R^2 distribution of cyRNA and ci model for all genes. The R^2 distribution of the subset of genes that are best fit by the ci model is shown in Fig 1I. (F) Pearson's r distribution of the miR model and the next best model (either naive, totRNA, cyRNA or ci). Only genes that are best fit by the miR model are shown. Corresponding R^2 distributions are plotted in Fig 2D.

Figure S3 *(following page).* Six candidate miR-gene interactions. Example fit of the miR model for genes *Cdk7*, *Pccb*, *Acad8*, *Mfge8*, *Eif4h* and *Srgap2* (rows). First column: expression of the assigned miRs of a single cluster. Colored lines are individual smoothing spline fits. Second column: Cytoplasmic RNA expression and the effective RNA concentration available for translation (see Materials and methods). Solid lines represent smoothing splines. Third/fourth column: cyRNA and miR model fits.


Cytoplasmic
Effective
Protein
Model cyRNA miR
miR
RNA
cyRNA−model
miR−model
99a-5p
100-5p
Mean
Gene: Cdk7
193b-5p
Gene: Pccb
433-3p
23b-5p
5615-3p
Mean
Gene: Acad8
150-3p
Gene: Mfge8
152-3p
467e-5p
Mean
Gene: Eif4h
135b-3p
Gene: Srgap2

Acronyms

BIC Bayesian information criterion

CV Coefficient of variation

DF degrees of freedom

EMT epithelial-mesenchymal transition

ERLIC electrostatic repulsion-hydrophilic interaction chromatography

ESC emrbyonic stem cell

GO gene ontology

HF high field

MET mesenchymal-epithelial transition

miR micro-RNA

MOFA multi-omics factor analysis

MS/MS tandem mass spectrometry

nuRNA nuclear RNA

PTR protein to mRNA ratio

RA retinoic acid

TMT tandem mass tag

3.6 References

[1] Patrick van den Berg et al. “Kinetic modeling of multi-omics data reveals microRNA-mediated translational regulation in stem cell differentiation”. In: *Unpublished* (2020).

[2] F Soldner et al. “Medicine. iPSC disease modeling”. In: *Science* 338.6111 (2012), pp. 1155–1156. DOI: 10.1126/science.1227682.

[3] Stefan Semrau et al. *Dynamics of lineage commitment revealed by single-cell transcriptomics of differentiating embryonic stem cells*. Tech. rep. 1. 2016, pp. 1–16. DOI: 10.1038/s41467-017-01076-4.

[4] Kyle M Loh et al. “Mapping the Pairwise Choices Leading from Pluripotency to Human Bone, Heart, and Other Mesoderm Cell Types”. In: *CELL* 166.2 (2016), pp. 451–467. DOI: 10.1016/j.cell.2016.06.011.

[5] Allon M Klein et al. “Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells”. In: *CELL* 161.5 (2015), pp. 1187–1201. DOI: 10.1016/j.cell.2015.04.044.

[6] Anna S E Cuomo et al. “Single-cell RNA-sequencing of differentiating iPS cells reveals dynamic genetic effects on gene expression”. In: *Nature Communications* 11.1 (2020), pp. 1–14. DOI: 10.1038/s41467-020-14457-z.

[7] Pavithra Kumar et al. “Understanding development and stem cells using single cell-based analyses of gene expression”. In: *Development (Cambridge, England)* 144.1 (2017), pp. 17–32. DOI: 10.1242/dev.133058.

[8] Christine Vogel et al. “Insights into the regulation of protein abundance from proteomic and transcriptomic analyses”. In: *Nature Publishing Group* 13.4 (2012), pp. 227–232. DOI: 10.1038/nrg3185.

[9] C Vogel et al. “Sequence signatures and mRNA concentration can explain two-thirds of protein abundance variation in a human cell line”. In: *Mol Syst Biol* 6.1 (2010), p. 400. DOI: 10.1038/msb.2010.59.

[10] Konstantine Tchourine et al. “One third of dynamic protein expression profiles can be predicted by a simple rate equation”. In: *Molecular BioSystems* 10.11 (2014), pp. 2850–2862. DOI: 10.1039/C4MB00358F.

[11] L Peshkin et al. “On the Relationship of Protein and mRNA Dynamics in Vertebrate Embryonic Development”. In: *Developmental Cell* 35.3 (), pp. 383–394. DOI: 10.1016/j.devcel.2015.10.010.

[12] Kolja Becker et al. “Quantifying post-transcriptional regulation in the development of Drosophila melanogaster”. In: *Nature Communications* 9.1 (2018), p. 4970. DOI: 10.1038/s41467-018-07455-9.

[13] Kannanganattu V Prasanth et al. “Regulating Gene Expression through RNA Nuclear Retention”. In: *CELL* 123.2 (2005), pp. 249–263. DOI: 10.1016/j.cell.2005.08.033.

[14] Carole Iampietro et al. “Developmentally Regulated Elimination of Damaged Nuclei Involves a Chk2-Dependent Mechanism of mRNA Nuclear Retention”. In: *Developmental Cell* 29.4 (2014), pp. 468–481. DOI: 10.1016/j.devcel.2014.03.025.

[15] Antoine Graindorge et al. “Sex-lethal promotes nuclear retention of msl2 mRNA via interactions with the STAR protein HOW.” In: *Genes & Development* 27.12 (2013), pp. 1421–1433. DOI: 10.1101/gad.214999.113.

[16] Vikram Agarwal et al. “Predicting effective microRNA target sites in mammalian mRNAs”. In: *eLife* 4 (2015). DOI: 10.7554/eLife.05005.

[17] Niketa A Patel et al. “Long noncoding RNA MALAT1 in exosomes drives regenerative function and modulates inflammation-linked networks following traumatic brain injury”. In: *Journal of Neuroinflammation* 15.1 (2018), pp. 1–23. DOI: 10.1186/s12974-018-1240-3.

[18] Jian Wang et al. “Cardiomyocyte overexpression of miR-27b induces cardiac hypertrophy and dysfunction in mice”. In: *Cell Research* 22.3 (2012), pp. 516–527. DOI: 10.1038/cr.2011.132.

[19] Ziping Zhang et al. “miR-27 promotes human gastric cancer cell metastasis by inducing epithelial-to-mesenchymal transition”. In: *Cancer Genetics* 204.9 (2011), pp. 486–491. DOI: 10.1016/j.cancergen.2011.07.004.

[20] J Chai et al. “miR-23b-3p regulates differentiation of osteoclasts by targeting PTEN via the PI3k/AKT pathway”. In: *Archives of Medical Science* (2019). DOI: 10.5114/aoms.2019.87520.

[21] Jing Zhang et al. “MiR-23b-3p induces the proliferation and metastasis of esophageal squamous cell carcinomas cells through the inhibition of EBF3”. In: *Acta Biochimica et Biophysica Sinica* 50.6 (2018), pp. 605–614. DOI: 10.1093/abbs/gmy049.

[22] Roshan M Kumar et al. “Deconstructing transcriptional heterogeneity in pluripotent stem cells”. In: *Nature* 516.7529 (2014), pp. 56–61. DOI: 10.1038/nature13920.

[23] Haihai Liang et al. “miR-26a suppresses EMT by disrupting the Lin28B/let-7d axis: potential cross-talks among miRNAs in IPF”. In: *Journal of Molecular Medicine* 94.6 (2016), pp. 655–665. DOI: 10.1007/s00109-016-1381-8.

[24] Fengyan Yu et al. “MicroRNA 34c gene down-regulation via DNA methylation promotes self-renewal and epithelial-mesenchymal transition in breast tumor-initiating cells.” In: *Journal of Biological Chemistry* 287.1 (2012), pp. 465–473. DOI: 10.1074/jbc.M111.280768.

[25] Huarong Huang et al. “miR-10a contributes to retinoid acid-induced smooth muscle cell differentiation.” In: *Journal of Biological Chemistry* 285.13 (2010), pp. 9383–9389. DOI: 10.1074/jbc.M109.095612.

[26] Marion Coolen et al. “miR-9: a versatile regulator of neurogenesis”. In: *Frontiers in Cellular Neuroscience* 7 (2013), p. 220. DOI: 10.3389/fncel.2013.00220.

[27] Yu Ding et al. “Elevation of MiR-9–3p suppresses the epithelial-mesenchymal transition of nasopharyngeal carcinoma cells via down-regulating FN1, ITGB1 and ITGAV”. In: *Cancer Biology & Therapy* 18.6 (2017), pp. 414–424. DOI: 10.1080/15384047.2017.1323585.

[28] Qimin Wang et al. “Sevoflurane represses the self-renewal ability by regulating miR-7a,7b/Klf4 signalling pathway in mouse embryonic stem cells”. In: *Cell Proliferation* 49.5 (2016), pp. 609–617. DOI: 10.1111/cpr.12283.

[29] Hongyi Zhang et al. “MiR-7, Inhibited Indirectly by LincRNA HOTAIR, Directly Inhibits SETDB1 and Reverses the EMT of Breast Cancer Stem Cells by Downregulating the STAT3 Pathway”. In: *STEM CELLS* 32.11 (2014), pp. 2858–2868. DOI: 10.1002/stem.1795.

[30] Ui Jeong Yun et al. “miR-195a Inhibits Adipocyte Differentiation by Targeting the Preadipogenic Determinator Zfp423”. In: *Journal of Cellular Biochemistry* 116.11 (2015), pp. 2589–2597. DOI: 10.1002/jcb.25204.

[31] Chunhui Liu et al. “miR-195 Inhibits EMT by Targeting FGF2 in Prostate Cancer Cells”. In: *PLOS ONE* 10.12 (2015), e0144073. DOI: 10.1371/journal.pone.0144073.

[32] Xiaobin Lin et al. “miR-195-5p/NOTCH2-mediated EMT modulates IL-4 secretion in colorectal cancer to affect M2-like TAM polarization”. In: *Journal of Hematology & Oncology* 12.1 (2019), pp. 1–14. DOI: 10.1186/s13045-019-0708-7.

[33] Aihua Xu et al. “Inhibiting effect of microRNA-187-3p on osteogenic differentiation of osteoblast precursor cells by suppressing cannabinoid receptor type 2”. In: *Differentiation* 109 (2019), pp. 9–15. DOI: 10.1016/j.diff.2019.07.002.

[34] Changwei Dou et al. “miR-187-3p inhibits the metastasis and epithelial–mesenchymal transition of hepatocellular carcinoma by targeting S100A4”. In: *Cancer Letters* 381.2 (2016), pp. 380–390. DOI: 10.1016/j.canlet.2016.08.011.

[35] Srinivas R Viswanathan et al. “microRNA Expression during Trophectoderm Specification”. In: *PLOS ONE* 4.7 (2009), e6143. DOI: 10.1371/journal.pone.0006143.

[36] Jui-Tung Liu et al. “Arsenic Induces Members of the mmu-miR-466-669 Cluster Which Reduces NeuroD1 Expression”. In: *Toxicological Sciences* 162.1 (2017), pp. 64–78. DOI: 10.1093/toxsci/kfx241.

[37] Zhang Z et al. “MiR-770 inhibits tumorigenesis and EMT by targeting JMJD6 and regulating WNT/β-catenin pathway in non-small cell lung cancer”. In: *Life Sciences* 188 (2017), pp. 163–171. DOI: 10.1016/j.lfs.2017.09.002.

[38] S-H Hu et al. “miR-760 mediates chemoresistance through inhibition of epithelial mesenchymal transition in breast cancer cells.” In: *European review for medical and pharmacological sciences* 20.23 (2016), pp. 5002–5008.

[39] Zhi-Jiang He et al. “miR-1306–3p targets FBXL5 to promote metastasis of hepatocellular carcinoma through suppressing snail degradation”. In: *Biochemical and Biophysical Research Communications* 504.4 (2018), pp. 820–826. DOI: 10.1016/j.bbrc.2018.09.059.

[40] Lei Yan et al. “MiR-301b promotes the proliferation, mobility, and epithelial-to-mesenchymal transition of bladder cancer cells by targeting EGR1”. In: *Biochemistry and Cell Biology* 95.5 (2017), pp. 571–577. DOI: 10.1139/bcb-2016-0232.

[41] Ana Rita Lourenço et al. “SOX4: Joining the Master Regulators of Epithelial-to-Mesenchymal Transition?” In: *Trends in Cancer* 3.8 (2017), pp. 571–582. DOI: 10.1016/j.trecan.2017.06.002.

[42] Tong Zhang et al. “Downregulation of miR-542-3p promotes cancer metastasis through activating TGF-β/Smad signaling in hepatocellular carcinoma”. In: *OncoTargets and Therapy* 11 (2018), pp. 1929–1939. DOI: 10.2147/OTT.S154416.

[43] Sanghwa Kim et al. “miR-340-5p Suppresses Aggressiveness in Glioblastoma Multiforme by Targeting Bcl-w and Sox2”. In: *Molecular Therapy - Nucleic Acids* 17 (2019), pp. 245–255. DOI: 10.1016/j.omtn.2019.05.022.

[44] Ying Dong et al. “MiR-186 Inhibited Migration of NSCLC via Targeting cdc42 and Effecting EMT Process”. In: *Molecules and Cells* 40.3 (2017), pp. 195–201. DOI: 10.14348/molcells.2017.2291.

[45] Rui Wang et al. “The PDGF-D/miR-106a/Twist1 pathway orchestrates epithelial-mesenchymal transition in gemcitabine resistance hepatoma cells”. In: *Oncotarget* 6.9 (2015), pp. 7000–7010. DOI: 10.18632/oncotarget.3193.

[46] B Schwanhäusser et al. “Global quantification of mammalian gene expression control”. In: *Nature* 473.7347 (2011), pp. 337–342. DOI: 10.1038/nature10098.

[47] Mathias Wilhelm et al. “Mass-spectrometry-based draft of the human proteome”. In: *Nature* 509.7502 (2014), pp. 582–587. DOI: 10.1038/nature13319.

[48] Fredrik Edfors et al. “Gene-specific correlation of RNA and protein levels in human cells and tissues”. In: *Molecular Systems Biology* 12.10 (2016), pp. 883–2. DOI: 10.15252/msb.20167144.

[49] G Csárdi et al. “Accounting for Experimental Noise Reveals That mRNA Levels, Amplified by Post-Transcriptional Processes, Largely Determine Steady-State Protein Levels in Yeast”. In: *PLoS Genet* 11.5 (2015), e1005206. DOI: 10.1371/journal.pgen.1005206.

[50] Alexander Franks et al. “Post-transcriptional regulation across human tissues”. In: *PLoS Computational Biology* 13.5 (2017), e1005535. DOI: 10.1371/journal.pcbi.1005535.

[51] Dominic Grün et al. “Conservation of mRNA and Protein Expression during Development of C. elegans”. In: *CellReports* 6.3 (2014), pp. 565–577. DOI: 10.1016/j.celrep.2014.01.001.

[52] Anders R Kristensen et al. “Protein synthesis rate is the predominant regulator of protein expression during differentiation”. In: *Molecular Systems Biology* 9 (2013), p. 689. DOI: 10.1038/msb.2013.47.

[53] Rong Lu et al. “Systems-level dynamic analyses of fate change in murine embryonic stem cells”. In: *Nature* 462.7271 (2009), pp. 358–362. DOI: 10.1038/nature08575.

[54] Tomáš Gedeon et al. “Delayed Protein Synthesis Reduces the Correlation between mRNA and Protein Fluctuations”. In: *Biophysical Journal* 103.3 (2012), pp. 377–385. DOI: 10.1016/j.bpj.2012.06.025.

[55] B Munsky et al. “From analog to digital models of gene regulation”. In: *Phys. Biol* 12.4 (2015), p. 045004. DOI: 10.1088/1478-3975/12/4/045004.

[56] Nelly Rahkonen et al. “Mature Let-7 miRNAs fine tune expression of LIN28B in pluripotent human embryonic stem cells”. In: *Stem Cell Research* 17.3 (2016), pp. 498–503. DOI: 10.1016/j.scr.2016.09.025.

[57] Lin He Meng Amy Li. “microRNAs as novel regulators of stem cell pluripotency and somatic cell reprogramming”. In: *BioEssays : news and reviews in molecular, cellular and developmental biology* 34.8 (2012), pp. 670–680. DOI: 10.1002/bies.201200019.

[58] Zsuzsanna Lichner et al. “The miR-290-295 cluster promotes pluripotency maintenance by regulating cell cycle phase distribution in mouse embryonic stem cells”. In: *Differentiation* 81.1 (2011), pp. 11–24. DOI: 10.1016/j.diff.2010.08.002.

[59] Harsh Dweep et al. “miRWalk – Database: Prediction of possible miRNA binding sites by “walking” the genes of three genomes”. In: *Journal of Biomedical Informatics* 44.5 (2011), pp. 839–847. DOI: 10.1016/j.jbi.2011.05.002.

[60] S Griffiths-Jones et al. “miRBase: tools for microRNA genomics”. In: *Nucleic Acids Research* 36.Database (2007), pp. D154–D158. DOI: 10.1093/nar/gkm952.

[61] Qi-Long Ying et al. “The ground state of embryonic stem cell self-renewal”. In: *Nature* 453.7194 (2008), pp. 519–523. DOI: 10.1038/nature06968.

[62] Jacek R Wisniewski et al. “Universal sample preparation method for proteome analysis”. In: *Nature Methods* 6.5 (2009), pp. 359–362. DOI: 10.1038/nmeth.1322.

[63] Adrian Alexa et al. *topGO: Enrichment Analysis for Gene Ontology*. 2019.

4 RNA-SEQUENCING VALIDATES MICRORNA-GENE PREDICTIONS FROM TRANSLATION MODEL

Abstract

Micro-RNAs (miRs) are post-transcriptional regulators of gene expression, which play important roles in development and cancer. Current methods to discover the effects of miRs on particular genes are either purely computational, and prone to produce many false positive, or cumbersome and low-throughput biochemical assays. In Chapter 3 we modeled the kinetics of protein. In this way we reduced the large number of sequence-based predictions to a few highly likely candidates that can be validated experimentally. Here, we validate the effect of six miRs on their predicted three target genes. We optimize assays for introducing miRs mimics or miR inhibitors in mouse embryonic stem cells using two fluorescent reporter cell lines that indicate the activity of specific miR . We performed RNA-sequencing on mouse emrbyonic stem cells (ESCs) transfected with mimics of six miRs and found in four cases that the predicted target gene is differentially expressed to a comparable extent as known targets of the respective miRs. These results corroborate the use of the kinetic model from Chapter 3 as a tool to identify novel miR-gene interactions.

4.1 Introduction

Mature miRs are single-stranded RNAs of circa 22 nucleotides that mediate RNA interference in eukaryotes. miRs typically pair to the 3' untranslated region (3'-UTR) of an mRNA and silence gene expression either through RNA-induced silencing complex (RISC)-mediated cleavage, destabilization of the poly-A tail, or blocking the translational machinery. In animals, the binding region in miRs is short and mostly not 100% complementary to the mRNA binding site. This, combined with the fact that there are around 24,000 predicted miRs and 140,000 predicted mRNA transcripts makes predicting miR-gene a difficult task. One popular database of predicted miR-gene interactions is TargetScan [1]. TargetScan reports a score which reflects the likelihood of miR binding, however neither its accuracy nor its precision is well characterized. One obvious reason an interaction might not exist is the lack of co-occurrence in cells. In Chapter 3 of this thesis we pre-selected dozens of mir-gene interactions using a very lenient TargetScan score threshold and whittled down this selection using our protein turnover rate model. By definition of our model, these miRs and genes are expressed simultaneously. In this chapter we will investigate three genes for which we predicted interactions: *Acad8*, *Cdk7* and *Eif4h*.

Acad8 is part of the acyl-CoA dehydrogenase family of enzymes and is not known to play a significant role in embryonic development or differentiation. However, mutations in *Acad8* are known to cause the rare genetic disease Isobutyryl-CoA dehydrogenase deficiency. *Acad8*'s predicted regulator miR-433-3p, on the other hand, is known to target genes involved in development: the transcription factor *CREB* [2], the *WNT* regulator *DKK1* [3] and the *Egfr* binding adaptor protein *GRB2* [4]. The second predicted regulator of *Acad8* is miR-23b-5p, which is known for its role in cancer via targets like *EBF3* [5], *FOXC1* [6] and *Hmgb2* [7]. We decided not to investigate the third miR, miR-5615-3p, due to its low expression levels.

The second gene is the cell cycle gene *Cdk7* which is part of the cyclin-dependent protein kinase family. *Cdk7* is essential during the very early stages of development [8, 9]. The two miRs that are predicted to regulate *Cdk7* are mir-99a-3p and mir-100-3p, which are very similar in sequence. Both are known to regulate *MTOR*/*Mtor* [10, 11], a serine-threonine kinase that is essential for growth and proliferation [12]. Other targets of mir-99a-3p and mir-100-3p include *NOX4* [13], *CDC25A* [14] and *Hoxa1* [15], and *RASGRP3* [16] and *IGF1R* [17] respectively.

The third gene with candidate interactions is *Eif4h*, a translation initiation factor. This gene is part of the machinery that recruits ribosomes to mRNA. Williams-Beuren syndrome is a rare genetic defect that results from a deletion of *Eif4h* [18]. We predicted regulation of the miRs miR-152-3p and miR-467e-3p, where, to the best of our knowledge, the latter has no experimentally validated targets. Some of miR-152-3p's known targets are the two cell cycle genes *CDKN1B* [19] and *CDK8* [20], the pluripotency inducing *KLF4* [21], and the DNA methyltransferase *Dnmt1* [22].

In this chapter we will demonstrate an miR mimic and inhibitor assay, which we optimized using two fluorescent reporter cell lines for miR activity [23]. By RNA-sequencing of

mouse ESCs transfected with miR mimics we validate four out of six predicted miR-gene interactions.

4.2 Results

In Chapter 3 we identified *Acad8*, *Cdk7* and *Eif4h* to be translationally regulated by miRs Fig 1a. TargetScan predicts one or two binding sites per miR of which only one is conserved between species Fig 1b. Based on the cumulative weighted context++ score (CWC++S) for these genes alone, these interactions do not stand out among the hundreds of predicted interaction each of these genes have with the exception for *Acad8* with miR-433-3p [1].

In order to investigate the effect of these miRs on mouse ESCs we set up a transfection assay with miR mimics and inhibitors Fig 2A. miR mimics (Pre-miR miRNA Precursors, Thermo Fisher Scientific) are double stranded RNAs designed to be processed by the cell into mature miRs that are identical to a mature miR. Their small size facilitates transfection, making them a potent tool to simulate miR overexpression in the cell. In contrast, miR inhibitors (miRCURY LNA miRNA Power Inhibitor, Qiagen) block miRs by complementary binding to the mature miR. The modified LNA bases have a higher binding affinity compared to RNA and therefore inhibit and degrade endogenous miR effectively. In order to evaluate the effectiveness of the transfections we created two fluorescent reporter cell lines Fig 2B. These cell lines have bi-directional CAG promoters with highly correlated transcription of two fluorescent proteins: mCherry and citrine. The citrine transcript has additionally been cloned with miR binding sites at its 3'-end, resulting in reduced expression of citrine relative to mCherry if the respective miR is present. We created reporter cell lines for a miR that is undetected in our system (mir-590-3p) and one that is highly expressed (miR292a-5p) in order to evaluate the mimic and inhibitors respectively. Flow cytometry measurements of the mimic transfection revealed a high percentage of positively transfected and regulated cells after 24h Fig 2C. Although the effect increased slightly over time, we picked 24h as the ideal time point in order to limit the amount of secondary effects of the mimic Fig 2D. Transfection of the inhibitor was slightly less effective even at higher doses Fig 2E. For the miR inhibitor transfection we selected 48h transfection with 2X the suggested dose to be ideal Fig 2F.

Having set up effective dose and timings for our transfection assays, we next set out to validate the predicted miR targets. Although our kinetic model was set up to predict regulation at the level of translation, we used mRNA abundance as a readout, reasoning that mRNA levels are likely affected as well. After 24h of exposure to the miR mimics, cell samples were collected and purified mRNA was subjected to RNA-seq Fig 3A. Differential expression analysis revealed significant downregulation of all three predicted target genes by at least one of the proposed miRs Fig 3B. *Acad8* is downregulated by miR-433-3p (P=1.7e–3), *Cdk7* is downregulated by miR-99a-5p (P=0.014) and *Eif4h* is downregulated by both miR-152-3p (P=0.015) and miR-467e-3p (P=1.21e–19). The observed downregulation supported the hypothesis that miR-gene binding takes place in these four cases.

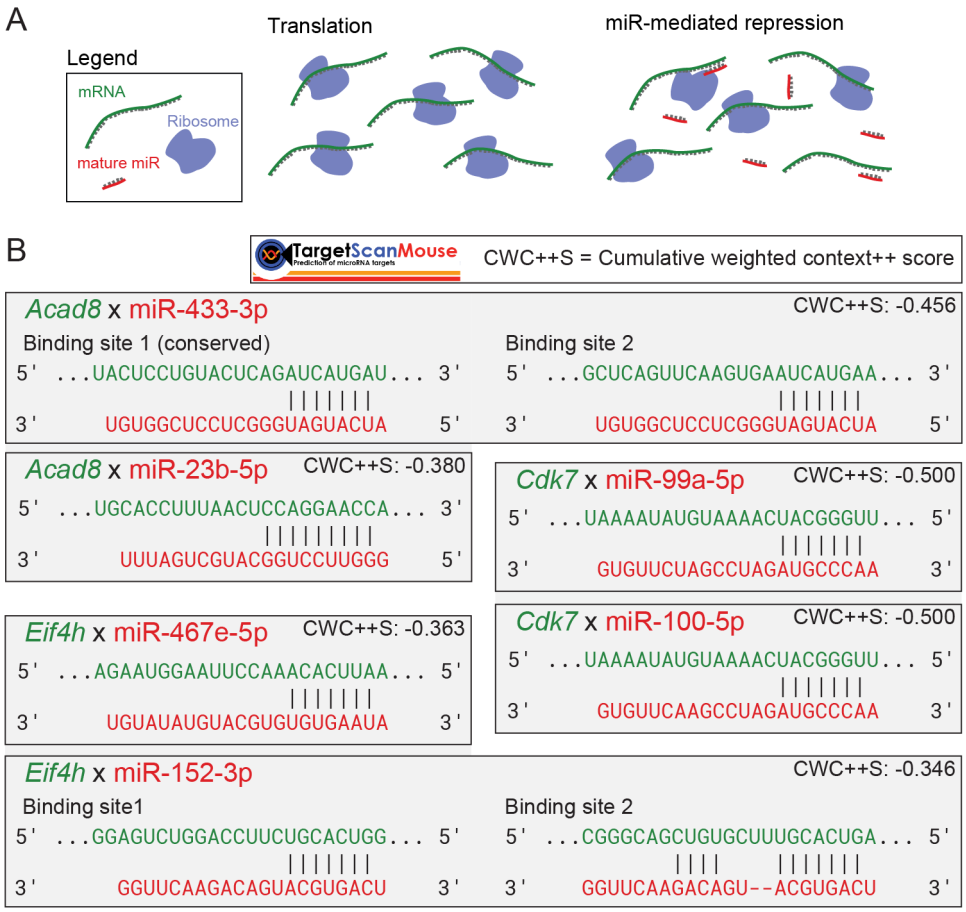


Figure 1. Proposed miR-gene interactions from chapter 3 are not scored highly by TargetScan. (A) Cartoon of miR-mediated translational regulation, miR-mediated mRNA degradation is not shown. (B) miR-gene interaction predictions from TargetScan 7.2 with their CWC++S [1]. Only one of eight predicted binding sites is preserved across species. Vertical lines indicate binding bases.

In order to assess the level of downregulation of our target genes, we compared these genes to some known targets in the context of the full transcriptome Fig 3C. The genes *Dkk1*, *Ebf3*, *Foxc1*, *Nox4*, *Hoxa1* and *Rasgrp3* were undetected in our system. Of the rest of the genes with known interactions most appear to be downregulated. Fixing the false discovery rate (FDR) at 0.1% only miR-100-5p–*Cdk7*, miR-152-3p–*Dnm1* and our own proposed miR–3p–*Eif4h*, are identified as interactions. Importantly, the lack of measured differential mRNA expression does not exclude binding of the mIRs and regulation of translation. These data show that our proposed significant miR-gene interactions, miR-433-3p–*Acad8*, miR-99a-5p–*Cdk7*, miR-152-3p–*Eif4h* and miR-467e-3p–*Eif4h*, lead to a similar extent of downregulation

as found in known targets.

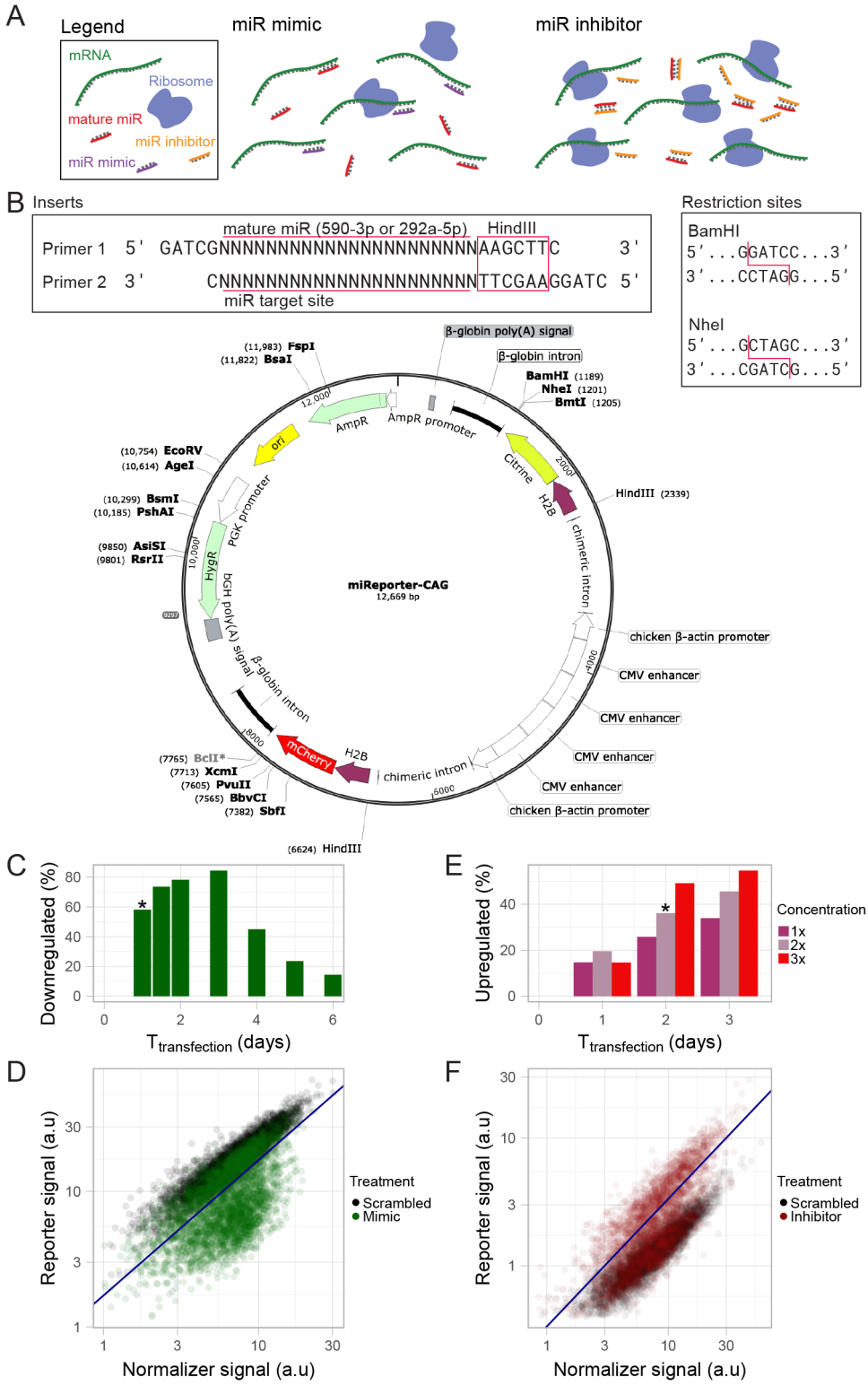
4.3 Discussion

Due to an enormous amount of possible combinations of miRs and genes it is a formidable challenge to identify true interactions. The kinetic model discussed in Chapter 3 whittles down potential interactions to a much smaller set of interactions that may have impact on translational regulation. We created reporter cell lines for activity of a lowly expressed and a highly expressed miR. Using these cell lines, we set up transfection assays for miR mimics and miR inhibitors that can enhance or annul a miRs effect on protein expression. We performed transcriptomic measurements of the mimic assay for six proposed miRs. These data showed that at least four out of six miRs bind and downregulate their predicted target mRNA, a surprisingly high fraction considering the search space. Moreover, *Eif4h* is, to the best of our knowledge, the first confirmed target for miR-467e-3p.

In this study we decided not to investigate the miRs using the dual luciferase assay. Although this is considered the gold standard of validating miR-gene interactions it only reports on the interaction between a specific miR-gene pair and does therefore not scale well to many possible pairs. By RNA-seq we can simultaneously observe the effects on the whole transcriptome. This can potentially reveal biological relevance of miRs in addition to providing controls in the form of known targets. On the other hand, some of the observed differential expression might be a secondary effect of direct miR targets. However, we start out with a clear hypothesis about the interaction from a very different source. In a follow up study we will complement the mimic experiments with the corresponding inhibitor assays. We hypothesize that quenching the miRs of miR-433-3p–*Acad8*, miR-99a-5p–*Cdk7*, miR-152-3p–*Eif4h* and miR-467e-3p–*Eif4h*, will result in higher expression of the respective mRNAs.

miR-433-3p is downregulated upon retinoic acid (RA) differentiation meaning any tar-

Figure 2 *(following page)*. Dose and timing for miR mimic and inhibitors transfection experiments can be obtained using fluorescent reporters of miR activity. (A) Cartoon of miR mimic and miR inhibitor translational regulation. (B) miReporter plasmid, inserts and digestion sites (BamHI and NheI). The insert overhangs are compatible with BamHI and NheI, but block redigestion. See Methods for full cloning strategy. (C) Inhibition of the miR-590-3p reporter transcript by the miR-590-3p mimic for seven time points as measured by flow cytometry. The asterisk indicates the optimal transfection timing shown in D (1d). (D) Fluorescence signal of miR-590-3p reporter for miR-590-3p mimic or scrambled control at optimal transfection conditions. Blue line indicates 1st percentile of reporter/normalizer ratio of the scrambled control. (E) Relief of inhibition on the miR-292a-5p reporter transcript by the miR-292a-5p inhibitor for three time points at three transfection concentrations as measured by flow cytometry. The asterisk indicates the optimal transfection timing shown in F (2days, 2X). (F) Fluorescence signal of miR-292a-5p reporter for miR-292a-5p inhibitor or scrambled control at optimal transfection conditions. Blue line indicates 99th percentile of reporter/normalizer ratio of the scrambled control.


A
Legend
mRNA
Ribosome
mature miR
miR inhibitor
miR mimic
miR mimic
miR inhibitor
B
Inserts
mature miR (590-3p or 292a-5p)
HindIII
Primer 1 5' GATCGNNNNNNNNNNNNNNNNNNNNNNNAAGCTTC 3'
Primer 2 3' CNNNNNNNNNNNNNNNNNNNNNNNTTCGAAGGATC 5'
miR target site
Restriction sites
BamHI
5'...GGATCC...3'
3'...CCTAGG...5'
NheI
5'...GCTAGC...3'
3'...CGATCG...5'
miReporter-CAG
12,669 bp
β-globin poly(A) signal
β-globin intron
BamHI (1189)
NheI (1201)
BmtI (1205)
HindIII (2339)
chicken β-actin promoter
CMV enhancer
CMV enhancer
CMV enhancer
CMV enhancer
chicken β-actin promoter
(6624) HindIII
(7765) BclI*
(7713) XcmI
(7605) PvuII
(7565) BbvCI
(7382) SbfI
mCherry
H2B
chimeric intron
β-globin intron
bGH poly(A) signal
HygR
PGK promoter
ori
AmpR
AmpR promoter
Citrine
H2B
chimeric intron
(11,983) FspI
(11,822) BsaI
(10,754) EcoRV
(10,614) AgeI
(10,299) BsmI
(10,185) PshAI
(9850) AsiSI
(9801) RsrII
C
Downregulated (%)
Ttransfection (days)
D
Reporter signal (a.u)
Normalizer signal (a.u)
Treatment
Scrambled
Mimic
E
Upregulated (%)
Ttransfection (days)
Concentration
1x
2x
3x
F
Reporter signal (a.u)
Normalizer signal (a.u)
Treatment
Scrambled
Inhibitor

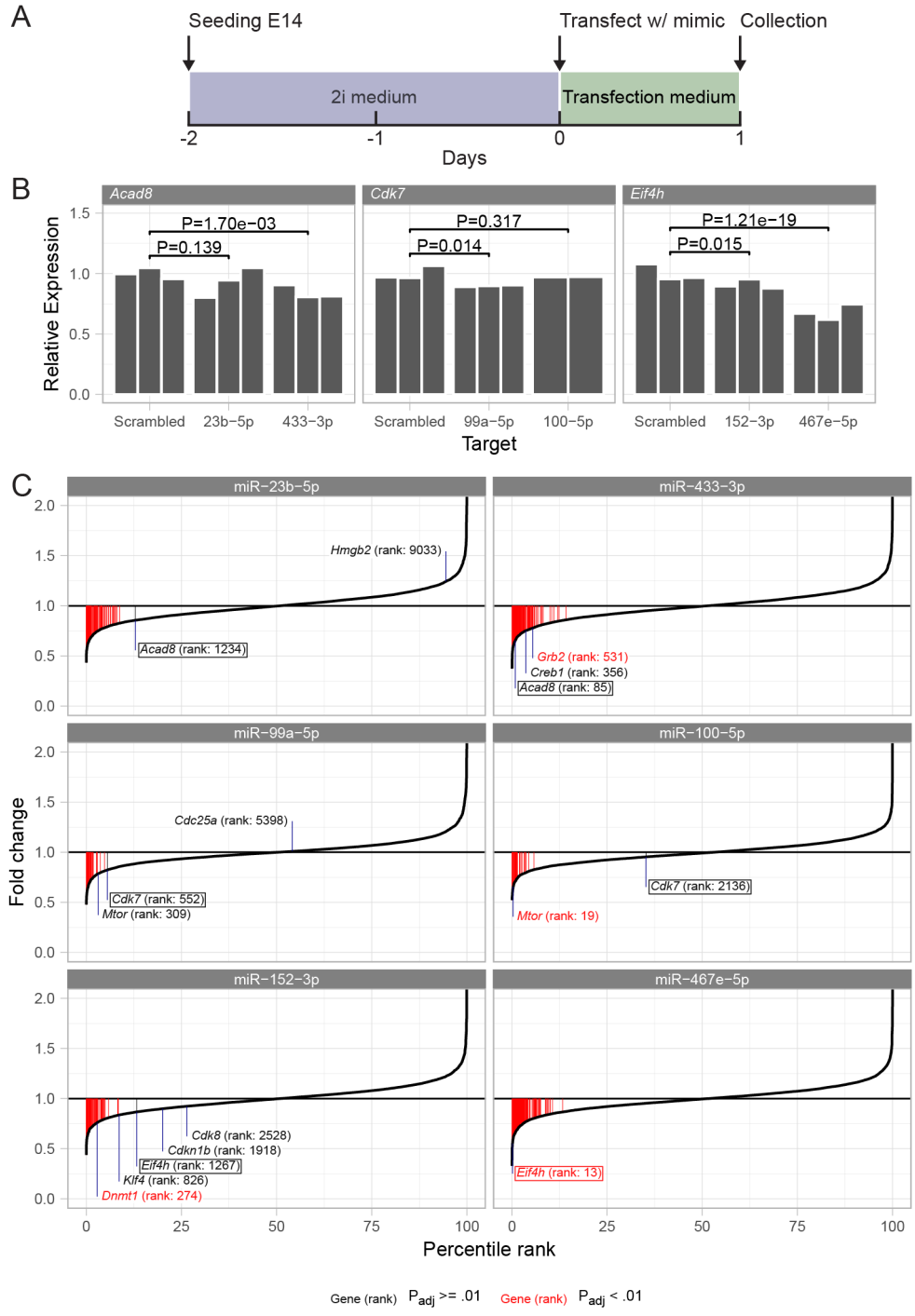
get's translational efficiency increases. Two of its known targets were present in our mouse ESCs, *Grb2* (downregulated FDR = 0.10%) and *Creb1* (downregulated FDR = 5.8%). *Grb2* is a known suppressor of *Nanog* so it may aid the exit of pluripotency [24]. Furthermore, *Creb* may be involved with the differentiation towards ectoderm or primitive endoderm, as it is involved in the differentiation towards many cell types like endothelial cells, brown adipocytes and osteoclasts [25, 26, 27]. Co-regulation of our discovered gene *Acad8* could suggest that miR-433-3p increases mitochondrial activity and simultaneously, reinforces the exit from pluripotency and differentiation.

In Chapter 3 we have observed that miR-99-5p expression increases in RA differentiation. Therefore we expect an increase in downregulation of *Ckd7* and *Mtor* as the cells differentiate. By targeting both of these genes miR-99-5p may simultaneously regulate the cell cycle, proliferation and cell growth. *Mtor*$^{-/-}$ mice embryos have shown to be embryonic lethal with a disorganized visceral endoderm at E5.5 [28]. Visceral endoderm is a derivative of primitive endoderm, which is one of the two cell types emerging in RA differentiation. Perhaps *Mtor*, and by extension *Ckd7*, are involved with the specification of the cell types induced by RA. *Ckd7*, despite being a cell cycle gene, is known to be involved in many differentiation processes.

The observation that the translational regulator *Eif4h* is itself translationally regulated demonstrates the complexity of translation regulation. One notable, known target of miR-152-3p is *Klf4*, which is part of the pluripotency gene regulatory network [29] and one of the four factors that were originally used to induce pluripotency in adult fibroblast cells [30]. In chapter 3 we observed that miR-152-3p, as well miR-467e-5p, increase over time around the exit from pluripotency during RA differentiation (Chapter 3 Fig S3). Perhaps miR-152-3p co-regulates the exit from pluripotency (by down-regulating *Klf4*, while simultaneously tweaking translation by downregulating *Eif4h*. Interestingly, according to our data, miR-467e-5p is more likely to downregulate *Klf4* expression than miR-152-3p (FDR = 1.4% vs 3.2%), but the validity of this interaction would require further investigation.

Using miRs the cell has many possible ways to subtly change its state. Unfortunately, many of the interactions between miRs and their target genes remain unconfirmed. The methods presented here, will serve as tools to discover bona fide interactions in a more scalable way than before. Future studies will have to reveal the functional relevance of these miRs for embryonic stem cell differentiation.

Figure 3 *(following page)*. Four out of six miR mimic transfections downregulate their proposed mRNA targets. (A) Schematic of the transfection protocol. (B) Expression levels (regularized counts scaled to scrambled control) of the proposed targets *Acad8*, *Cdk7* and *Eif4h* after miR mimic transfection and scrambled control. P-value shown is for an uncorrected one-sided test (see Methods). (C) Expression fold changes relative to scrambled control of six miR mimic transfections. Text annotations are known or proposed targets. The boxed genes are our proposed targets. Red lines indicate significantly downregulated genes. P_{adj} = Benjamini-Hochberg corrected p-value.


A
Seeding E14
Transfect w/ mimic
Collection
2i medium
Transfection medium
-2
-1
0
1
Days
B
Acad8
Cdk7
Eif4h
Relative Expression
P=1.70e−03
P=0.139
P=0.317
P=0.014
P=1.21e−19
P=0.015
Scrambled
23b−5p
433−3p
99a−5p
100−5p
152−3p
467e−5p
Target
C
miR−23b−5p
miR−433−3p
miR−99a−5p
miR−100−5p
miR−152−3p
miR−467e−5p
Fold change
Hmgb2 (rank: 9033)
Acad8 (rank: 1234)
Grb2 (rank: 531)
Creb1 (rank: 356)
Acad8 (rank: 85)
Cdc25a (rank: 5398)
Cdk7 (rank: 552)
Mtor (rank: 309)
Cdk7 (rank: 2136)
Mtor (rank: 19)
Cdk8 (rank: 2528)
Cdkn1b (rank: 1918)
Eif4h (rank: 1267)
Klf4 (rank: 826)
Dnmt1 (rank: 274)
Eif4h (rank: 13)
Percentile rank
Gene (rank) Padj >= .01
Gene (rank) Padj < .01

4.4 Materials and Methods

4.4.1 Cell culture

E14 mouse embryonic stem cells were maintained as described in Chapter 3 of this thesis. During transfections, cells were temporarily cultured in serum+LIF medium (10% ES certified FBS, 1X non-essential amino acids, 0.1mM β-mercaptoethanol, 1X pen/strep, 2mM L-glutamine, 10,000U/ml mLIF, mLIF from Merck, rest from Thermo Fisher Scientific). Furthermore, miR reporter cell line clone selection took place on homegrown mouse embryonic fibroblast feeders.

4.4.2 Cloning

The miReporter backbone (AddGene, Plasmid #82478) was transformed into DH5a competent cells (Cat. 18265017, Thermo Fisher Scientific) as per manufacturer's instructions. Then, transformed cells were expanded and harvested for miReporter backbone by miniprep (Qiaprep, Qiagen). A set of two oligos were synthesized for each of the two reporter cell lines: miR-590-3p-fwd: 5'-GATCG TAATTTTATGTATAAGCTAGT AAGCTTC-3', miR-590-3p-rev: 5'-CTAGGAAGCTT ACTAGCTTATACATAAAATTA C-3', miR-292a-5p-fwd: 5'-GATCG ACTCAAACTGGGGGCTCTTTTG AAGCTTC-3', miR-292a-5p-rev: 5'-CTAGGAAGCTT CAAAAGAGCCCCCAGTTTGAGT C-3' (Integrated DNA Technologies, see Fig 2b). Pairs of oligos were annealed and phosphorylated in a thermocycler: 30m at 37°C, 5m at 95°C, for 12 cycles (1μM fwd oligo, 1μM rev oligo, 1X T4 buffer, 1U/μl T4 Polynucleotide Kinase; buffer and enzyme from New England Biolabs). Next, backbone digestion and ligation was performed in one step in a thermocycler, which was facilitated by the ligated inserts destroying the restriction sites for the enzymes (See Fig 2b): 5m at 37°C, 5m at 23°C, for 12 cycles (1:2500 dilution of phosphorylated oligo duplex, 2.5ng/μl backbone, 5% v/v DTT, 0.15U/μl BamHI, 0.5U/μl NheI, 1U/μl T4 ligase, 1X restriction buffer; T4 from New England Biolabs, rest from Thermo Fisher Scientific). Plasmids were then amplified in the same manner as the backbone: Plasmids were transformed into DH5a cells, which were then expanded and used for midiprep extraction (Plasmid Midi, Qiagen).

4.4.3 miReporter cell lines creation

miReporter-590-3p and miReporter-292a-5p plasmids were transfected into ESCs with lipofectamine 3000 (Thermo Fisher Scientific) as per manufacturer's instruction. Briefly, 125μl DMEM (Sigma-Aldrich) was mixed with 5μl lipofectamine 3000 and briefly vortexed. Separately 125μl was mixed with 5μl p3000 reagent and 5μg of plasmid and also briefly vortexed. Both mixtures were combined and incubated at room temperature for 5 minutes to create DNA-lipid complexes. 2i medium was removed from pre-seeded ESCs at a confluency of about 70-90% and replaced with serum+LIF medium. DNA-lipid complexes were added to the medium for 24 h. Medium was then aspirated, cells washed twice with PBS, and cells were

left to grow for two days in 2i. Transfected cells were selected for by hygromycin (100 μg/ml in 2i) for three days. Single clones were selected differently for the two miReporter cell lines. Double-positive, single cells of the miReporter-590-3p cell line were sorted by Fluorescence-activated cell sorting in 96-well feeder-coated plates and expanded thereafter. miReporter-292a-5p cells however, were strongly diluted upon passage and single colonies with double-positive cells were picked by hand using a benchtop microscope and a 200μl pipette. Double positive colonies were left to expand on feeders in 48-well plates. Clones were grown for at least two passages to ascertain the stability of the transfection. The reporter activity was confirmed by flow cytometry.

4.4.4 Mimic and inhibitor transfection

ESCs and both miReporter cell lines were transfected with miR mimic and inhibitors in an identical fashion. Cells were seeded 48 h prior to transfection in 12-well plates. DNA-lipid complexes (Lipofectamine RNAiMax, Thermo Fisher Scientific) were prepared at the ratios recommended by the manufacturer but siRNA was replaced with either miR mimic or miR inhibitor (see Table 4.1). We considered 1X concentration of mimic/inhibitor to be 100nM in the DNA-lipid mixture. We used 37.5μl of DNA-lipid mixture per well. After 24h transfection cells were either harvested or washed with PBS and left to grow in 2i.

miR-target	Our name	Company name	Company	Catalog number
None	mimic scrambled	Pre-miR™miRNA Precursor Negative Control #1	Thermo Fisher Scientific	AM17110
miR-590-3p	miR-590-3p mimic	Pre-miR miRNA Precursor	Thermo Fisher Scientific	AM17100-PM20392
miR-23b-5p	miR-23b-5p mimic	Pre-miR miRNA Precursor	Thermo Fisher Scientific	AM17100-PM15655
miR-433-3p	miR-433-3p mimic	Pre-miR miRNA Precursor	Thermo Fisher Scientific	AM17100-PM10774
miR-99a-5p	miR-99a-5p mimic	Pre-miR miRNA Precursor	Thermo Fisher Scientific	AM17100-PM10719
miR-100-5p	miR-100-5p mimic	Pre-miR miRNA Precursor	Thermo Fisher Scientific	AM17100-PM10188
miR-152-3p	miR-152-3p mimic	Pre-miR miRNA Precursor	Thermo Fisher Scientific	AM17100-PM12269
miR-467e-5p	miR-467e-5p mimic	Pre-miR miRNA Precursor	Thermo Fisher Scientific	AM17100-PM12611
None	scrambled inhibitor	miRCURY LNA miRNA Inhibitor Control: Negative control A	Qiagen	YI00199006-DDA

Continued on next page

Continued from previous page

miR-target	Our name	Company name	Company	Catalog number
miR-292a-5p	miR-292a-5p inhibitor	miRCURY LNA miRNA Power Inhibitors	Qiagen	YI04101165-DDA

Table 4.1. Overview of miR mimic and inhibitors.

4.4.5 Flow cytometry

Transfected cells were harvested for flow cytometry by washing with PBS and detachment using Accutase (Sigma-Aldrich). Detached cells were washed and resuspended in 2i. Cells were fixed in 4%Formaldehyde in medium (Cat. 43368, Alfa Aesar) for 15 min at room temperature. Cells were then centrifuged and the supernatant was removed. Cells were resuspended in 1% BSA (Cat. A2153, Sigma-Aldrich) in PBS and stored at 4°C until the measurement.

Fixed cells were measured on a BD LSRFortessa X-20. Forward and side scatter was measured as well as Citrine fluorescence (488nm laser, 530/30nm emission filter) and mCherry fluorescence (561nm laser, 610/20 emission filter). Live cell gating and Citrine/mCherry positive selection was achieved using custom R scripts (FlowCore v1.52.1 [31]). To determine relative down- or upregulation of citrine expression we calculated the ratio mCherry/citrine for each cell in the mimic/inhibitor assays and scrambled controls. Mimic cells with a ratio lower than the 1^{st} percentile of scrambled control ratios were deemed transfected with successful citrine inhibition. Inhibitor cells with a ratio higher than the 99^{th} percentile of scrambled control ratios were deemed transfected with successful miR inhibition.

4.4.6 RNA-sequencing

RNA-sequencing was performed as described in Chapter 3 of this thesis with a minimum of 10 million raw reads per sample.

4.4.7 RNA-sequencing analysis

RNA-sequencing data was preprocessed as described in Chapter 3 of this thesis. Sample 3 of the miR-100p-5p mimic was excluded due to low complexity. Only genes with expression counts larger than 20 were kept and were subsequently processed using DEseq2 (v1.26.0, [32]). Regularized counts are defined as the 2-base exponent of the *rlog*-values. DESeq2 was also used to identify differentially expressed genes and obtain $\log_2$ fold-changes. 14 genes were differentially expressed between the scrambled control and no treatment control: *Tjp2*, *Serp1*, *Rap1b*, *Cdk2ap1*, *Ssr2*, *B230219d22rik*, *Cyld*, *Fam98b*, *Tpm3*, *Uggt1*, *Snx6*, *Rfc4*, *Tcf7l1* and *Adam10*. This list of genes was excluded in other comparisons.

Acknowledgements

Stefan Semrau and Patrick van den Berg conceived the project. S.S. acquired funding. P.v.d.B., Noémie Bérenger-Currias and Marleen Feliksik performed the experiments. P.v.d.B analyzed, interpreted and modeled the data. P.v.d.B. and S.S. wrote the manuscript.

Acronyms

CWC++S cumulative weighted context++ score

ESC emrbyonic stem cell

FDR false discovery rate

miR micro-RNA

RA retinoic acid

RISC RNA-induced silencing complex

4.5 References

[1] Vikram Agarwal et al. “Predicting effective microRNA target sites in mammalian mRNAs”. In: *eLife* 4 (2015). DOI: 10.7554/eLife.05005.

[2] Shupeng Sun et al. “MiR-433-3p suppresses cell growth and enhances chemosensitivity by targeting CREB in human glioma”. In: *Oncotarget* 8.3 (2016), pp. 5057–5068. DOI: 10.18632/oncotarget.13789.

[3] Xiaolin Tang et al. “MicroRNA-433-3p promotes osteoblast differentiation through targeting DKK1 expression”. In: *PLOS ONE* 12.6 (2017), e0179860. DOI: 10.1371/journal.pone.0179860.

[4] Qizhong Shi et al. “MiR-433-3p Inhibits Proliferation and Invasion of Esophageal Squamous Cell Carcinoma by Targeting GRB2.” In: *Cellular physiology and biochemistry : international journal of experimental cellular physiology, biochemistry, and pharmacology* 46.5 (2018), pp. 2187–2196. DOI: 10.1159/000489548.

[5] Jing Zhang et al. “MiR-23b-3p induces the proliferation and metastasis of esophageal squamous cell carcinomas cells through the inhibition of EBF3”. In: *Acta Biochimica et Biophysica Sinica* 50.6 (2018), pp. 605–614. DOI: 10.1093/abbs/gmy049.

[6] Guo dong Hu et al. “Long noncoding RNA CCAT2 functions as a competitive endogenous RNA to regulate FOXC1 expression by sponging miR-23b-5p in lung adenocarcinoma”. In: *Journal of Cellular Biochemistry* 120.5 (2018), pp. 7998–8007. DOI: 10.1002/jcb.28077.

[7] Diafara Boureima Oumarou et al. “Involvement of microRNA-23b-5p in the promotion of cardiac hypertrophy and dysfunction via the HMGB2 signaling pathway”. In: *Biomedicine & Pharmacotherapy* 116 (2019), p. 108977. DOI: 10.1016/j.biopha.2019.108977.

[8] Miguel Ganuza et al. “Genetic inactivation of Cdk7 leads to cell cycle arrest and induces premature aging due to adult stem cell exhaustion”. In: *The EMBO Journal* 31.11 (2012), pp. 2498–2510. DOI: 10.1038/emboj.2012.94.

[9] Shetal A Patel et al. “Functional analysis of the Cdk7.cyclin H.Mat1 complex in mouse embryonic stem cells and embryos.” In: *Journal of Biological Chemistry* 285.20 (2010), pp. 15587–15598. DOI: 10.1074/jbc.M109.081687.

[10] Junhong Cai et al. “MiR-100-5p, miR-199a-3p and miR-199b-5p induce autophagic death of endometrial carcinoma cell through targeting mTOR.” In: *International Journal of Clinical and Experimental Pathology* 10.9 (2017), pp. 9262–9272.

[11] Xiaoyang Ye et al. “MicroRNAs 99b-5p/100-5p Regulated by Endoplasmic Reticulum Stress are Involved in Abeta-Induced Pathologies”. In: *Frontiers in Aging Neuroscience* 7 (2015), p. 210. DOI: 10.3389/fnagi.2015.00210.

[12] Mirei Murakami et al. “mTOR Is Essential for Growth and Proliferation in Early Mouse Embryos and Embryonic Stem Cells”. In: *Molecular and Cellular Biology* 24.15 (2004), pp. 6710–6718. DOI: 10.1128/MCB.24.15.6710-6718.2004.

[13] Y Shi et al. “MiR-99a-5p regulates proliferation, migration and invasion abilities of human oral carcinoma cells by targeting NOX4”. In: *Neoplasma* 64.05 (2017), pp. 666–673. DOI: 10.4149/neo_2017_503.

[14] Hongzhen Qin et al. “MicroRNA-99a-5p suppresses breast cancer progression and cell-cycle pathway through downregulating CDC25A”. In: *Journal of Cellular Physiology* 234.4 (2019), pp. 3526–3537. DOI: 10.1002/jcp.26906.

[15] Faman Xiao et al. “Downregulation of HOXA1 gene affects small cell lung cancer cell survival and chemoresistance under the regulation of miR-100”. In: *European Journal of Cancer* 50.8 (2014), pp. 1541–1554. DOI: 10.1016/j.ejca.2014.01.024.

[16] Qian Peng et al. “FOXA1 Suppresses the Growth, Migration, and Invasion of Nasopharyngeal Carcinoma Cells through Repressing miR-100-5p and miR-125b-5p”. In: *Journal of Cancer* 11.9 (2020), pp. 2485–2495. DOI: 10.7150/jca.40709.

[17] Hongliang Zhang et al. “miR-100-5p Inhibits Malignant Behavior of Chordoma Cells by Targeting IGF1R”. In: *Cancer Management and Research* 12 (2020), pp. 4129–4137. DOI: 10.2147/CMAR.S252185.

[18] Rossella De Cegli et al. “A transcriptomic study of Williams-Beuren syndrome associated genes in mouse embryonic stem cells”. In: *Scientific Data* 6.1 (2019), pp. 1–5. DOI: 10.1038/s41597-019-0281-5.

[19] L Wang et al. “MiR-152-3p promotes the development of chronic myeloid leukemia by inhibiting p27.” In: *European review for medical and pharmacological sciences* 22.24 (2018), pp. 8789–8796. DOI: 10.26355/eurrev_201812_16646.

[20] Tao Yin et al. “miR-152-3p Modulates hepatic carcinogenesis by targeting cyclin-dependent kinase 8”. In: *Pathology - Research and Practice* 215.6 (2019), p. 152406. DOI: 10.1016/j.prp.2019.03.034.

[21] Feng Feng et al. “miR-148-3p and miR-152-3p synergistically regulate prostate cancer progression via repressing KLF4”. In: *Journal of Cellular Biochemistry* 120.10 (2019), pp. 17228–17239. DOI: 10.1002/jcb.28984.

[22] Jin Sun et al. “Regulation of human glioma cell apoptosis and invasion by miR-152-3p through targeting DNMT1 and regulating NF2”. In: *Journal of Experimental & Clinical Cancer Research* 36.1 (2017), p. 1061. DOI: 10.1186/s13046-017-0567-4.

[23] Hanna L Sladitschek et al. “Bidirectional Promoter Engineering for Single Cell MicroRNA Sensors in Embryonic Stem Cells”. In: *PLOS ONE* 11.5 (2016), e0155177. DOI: 10.1371/journal.pone.0155177.

[24] Takashi Hamazaki et al. “The Grb2/Mek Pathway Represses Nanog in Murine Embryonic Stem Cells”. In: *Molecular and Cellular Biology* 26.20 (2006), pp. 7539–7549. DOI: 10.1128/MCB.00508-06.

[25] Kohei Yamamizu et al. “PKA/CREB Signaling Triggers Initiation of Endothelial and Hematopoietic Cell Differentiation via Etv2 Induction”. In: *STEM CELLS* 30.4 (2012), pp. 687–696. DOI: 10.1002/stem.1041.

[26] Aaron M Cypess et al. “Insulin/IGF-I Regulation of Necdin and Brown Adipocyte Differentiation Via CREB- and FoxO1-Associated Pathways”. In: *Endocrinology* 152.10 (2011), pp. 3680–3689. DOI: 10.1210/en.2011-1229.

[27] Kojiro Sato et al. “Regulation of osteoclast differentiation and function by the CaMK-CREB pathway”. In: *Nature Medicine* 12.12 (2006), pp. 1410–1416. DOI: 10.1038/nm1515.

[28] Yann-Gaël Gangloff et al. “Disruption of the Mouse mTOR Gene Leads to Early Postimplantation Lethality and Prohibits Embryonic Stem Cell Development”. In: *Molecular and Cellular Biology* 24.21 (2004), pp. 9508–9516. DOI: 10.1128/MCB.24.21.9508-9516.2004.

[29] Mo Li et al. “Deconstructing the pluripotency gene regulatory network”. In: *Nature Cell Biology* 20.4 (2018), pp. 382–392. DOI: 10.1038/s41556-018-0067-6.

[30] Kazutoshi Takahashi et al. “Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors”. In: *CELL* 126.4 (2006), pp. 663–676. DOI: 10.1016/j.cell.2006.07.024.

[31] B Ellis et al. *flowCore: flowCore: Basic structures for flow cytometry data*. 2019.

[32] Michael I. Love et al. “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2”. In: *Genome Biology* 15.12 (2014), p. 31. DOI: 10.1186/s13059-014-0550-8.

SUMMARY

The cells in our body are entrusted with countless different tasks. For instance, some specialized cells form a physical barrier to the outside, some generate the hormone insulin and some produce acid in our stomachs. To achieve these diverse tasks we carry a diverse set of genes, i.e. the functional units of information we inherited from our ancestors. Genes are stored in the DNA molecules in our cells' nucleus, which is collectively called our genome. However, just storing this information is not enough; it needs to be converted into different types of molecules that can perform the cell's tasks. The central dogma of molecular biology describes the flow of genetic information into these other molecules:

1. Genes in the genome can be *replicated*, creating an exact replica of the genome. This happens when cells divide into two cells, that each have to contain the genome.

2. Genes in the genome can be *transcribed*, creating transcripts (RNA molecules). This happens when certain proteins are needed in the cell. All transcripts of the cell are, collectively, called the transcriptome.

3. Genes in the transcriptome can be *translated*, creating proteins. These molecules are responsible for most of the work in the cell. All proteins of the cell are, collectively, called the proteome.

The central dogma of molecular biology is an excellent model to keep in mind when thinking about how genetic information flows from DNA to RNA to protein. However, it fails to describe *which* information flows at a given time and at which *rates*. That is where *gene regulation* comes into play. Gene regulation is a set of biochemical processes the cell employs that ensures a finely tuned proteome, which is crucial for its function. Gene regulation is also one of the common threads among the chapters of this thesis along with mammalian development. Each chapter of this thesis covers a different type of gene regulation (at the different levels of the central dogma, if you will).

In the **Introduction** we introduce the important concept of cell types, and how this subject closely ties into gene regulation and development (another common thread of this thesis). We then give a layman's explanation of several concepts that are needed to understand the remainder of the thesis. We introduce DNA methylation as a form of epigenetic regulation (epigenetics: inheritable alterations of the genome that do not change its DNA sequence). We next explain transcriptional regulation, the control mechanisms that decide which genes are transcribed at which levels, and translational regulation, which controls the rate at which translation take place. And finally we discuss *omics*, a relatively recent collection of techniques that are vital for measuring the above phenomena and are heavily featured in this thesis.

In **Chapter 1** we look at how epigenetics influences cell identity and transcription in mouse embryonic stem cells (mESCs). We investigate a set of super-enhancers (sets of closely-clustered enhancers, which are regions of the genome that do not code for genes but can promote transcription elsewhere). Super-enhancers are often associated with cell identity and therefore highly relevant in developmental systems. Moreover, it was found that in some cell types the DNA of super-enhancers is methylated at varying levels between individual cells. However, the mechanisms causing this heterogeneous methylation were not well understood. We created methylation reporter cell lines for the *Sox2* and miR-290 super-enhancers to uncover these mechanisms. These cell lines have fluorescent proteins of two distinct colors that are turned on and off depending on the methylation state of each of the alleles (the maternal and paternal copies of the super-enhancers). We show that the methylation states at these super-enhancers are not only heterogeneous but also highly dynamic, because they switch from on to off and vice versa over a matter of days. We also show that the methylation state of the super-enhancers influences the transcription of genes *in cis* (meaning the methylation state has a local effect on the same chromosome). Finally, we observe that this dynamic methylation is not an *in vitro* artifact but also occurs in pre-implantation embryos.

In **Chapter 2** we use the transcriptomes of cells as key indicators of their cell type. We performed single-cell transcriptomics on human fetal kidneys at different stages in development. Kidneys consist out of about 1 million nephrons, functional units that operate independently. The development of these nephrons is asynchronous, meaning that we can observe multiple stages of nephron development simultaneously in the fetal kidney. In the transcriptomics dataset we identify 22 different cell types, ranging from precursor cells to fully differentiated cells. Some of these cell types were novel and more nuanced subclassifications of previously known cell types like the *nephron progenitor cells*. We also observe that for the *podocyte* the transcriptome continues to change over the course of development, even though the cell type is already established. This kind of more detailed information on the kidney cell types can potentially open up avenues for the development of cures for kidney diseases in the form of regenerative medicine.

In **Chapter 3** we investigate translation and degradation rates in mESCs. Measuring RNA concentrations is often more convenient than measuring protein concentrations and because of the hierarchical relationship between the two, changes in RNA are often assumed to occur in protein as well. If there is a sudden change in RNA concentration of a gene then the protein concentration should reflect this. How responsive a gene's protein concentration is to its RNA concentration differs per gene and changes dynamically. Therefore, in a system undergoing pervasive changes, e.g. during differentiation, one cannot simply measure RNA concentrations and assume that the corresponding proteins follow a similar pattern. In this chapter we model protein synthesis and degradation rates using RNA and protein data of differentiating mESCs. This allows us to measure the extent of discordance between RNA and protein. More interestingly, the models were useful for identifying when our prior

assumption of constant turnover rates does *not* hold, hinting that these genes may be post-transcriptionally regulated (i.e. regulated at the protein synthesis or degradation steps). Using these models we identify new cases of translational regulation by microRNA (miR), small RNA molecules that block gene-encoding RNAs. Interactions between mRNA and miRs are notoriously hard to identify.

In **Chapter 4** we put the models of the previous chapter to the test. We set out to verify some of the miR-gene interactions we had identified. To probe the interactions we *transfected* the cells with miR mimics and miR inhibitors (to transfect: to introduce nucleic acid molecules or proteins into a cell). To optimize the timings and concentrations of these transfections we first created cell lines that report when these transfections take place by means of fluorescent proteins. We then transfected six different miR mimics, performed transcriptomics and show that these miRs negatively influence the predicted genes' concentrations. These results reinforce the use of the models from Chapter 3 as a means of identifying cases of translational regulation by miR.

SAMENVATTING

We vertrouwen er op dat de cellen in ons lichaam talloze taken verrichten. Zo zijn er bijvoorbeeld cellen die een fysieke barrière vormen met de buiten wereld, cellen die insuline produceren, en cellen die het zuur in onze maag uitscheiden. Om aan deze taken te voldoen heeft de evolutie ervoor gezorgd dat wij veel verschillende genen hebben, de functionele eenheden van informatie die we hebben geërfd van ons voorouders. Genen zijn opgeslagen in de DNA moleculen in de kern van onze cellen, en samen worden ze ons genoom genoemd. Echter, het louter opslaan van informatie is niet voldoende, het dient ook vertaald te worden naar verschillende soorten moleculen die de taken van de cel daadwerkelijk kunnen uitvoeren. De overdracht van genetisch informatie van het DNA naar deze moleculen vormt het centrale dogma van de moleculaire biologie:

1. Genen in het genoom kunnen worden *gerepliceerd*, waarmee een identieke kopie wordt gemaakt. Dit gebeurt als cellen zich delen en elke cel een kopie van het genoom nodig heeft.

2. Genen in het genoom kunnen worden *getranscribeerd*, wat transcripten produceert (RNA moleculen). Dit gebeurt wanneer bepaalde eiwitten nodig zijn in de cel. Alle transcripten in de cel worden samen het transcriptoom genoemd.

3. Genen in het transcriptoom kunnen worden *getransleerd*, wat eiwitten produceert. Deze moleculen zijn verantwoordelijk voor het merendeel van de arbeid die de cel verricht. Alle eiwitten in de cel worden samen het proteoom genoemd.

Het centrale dogma van de moleculaire biologie is een uitstekend model dat in staat is om te beschrijven hoe genetische informatie wordt overgedragen van DNA naar eiwit. Echter, het omschrijft niet *welke* informatie wordt overgedragen en hoe *snel*. Dat is waar *gen regulatie* een rol speelt. Gen regulatie is een verzameling biochemische processen die de cel inzet om er voor te zorgen dat het fijn afgesteld proteoom heeft, wat cruciaal is voor het functioneren van de cel. Gen regulatie is ook een van de rode draden tussen de hoofdstukken van dit proefschrift. Elke hoofdstuk in dit proefschrift betreft een ander type gen regulatie (grofweg langs de lijnen van het centrale dogma).

In de **Introductie** introduceren we het concept *cell type* en beschrijven we hoe nauw verwant dit onderwerp is aan gen regulatie en ontwikkeling (nog een rode draad van dit proefschrift). We beschrijven dan in lekentaal een aantal concepten die nodig zijn om de rest van het proefschrift te begrijpen. We beschrijven DNA methylatie als form van epigenetische regulatie (epigenetica: overdraagbare veranderingen in het genoom die de sequentie zelf niet aantasten). Daarna duiken we kort in de transcriptie regulatie, de controle mechanismen

die beslissen welke genen op welk moment worden getranscribeerd, en translatie regulatie, wat bepaalt met welke snelheid translatie plaatsvindt. Tot slot beschrijven we *omics*, een verzameling van relatief nieuwe technieken die cruciaal zijn voor het meten van de eerder genoemde processen en die veel worden toegepast in dit proefschrift.

In **Hoofdstuk 1** bekijken we hoe epigenetica de cel identiteit en transcriptie van *mouse embryonic stem cells* (mESCs) beïnvloedt. We onderzoeken een verzameling van super-enhancers (groepen van nabijgelegen enhancers, wat regio's zijn in het genoom die niet coderen voor genen maar betrokken zijn bij de transcriptie van genen elders). Super-enhancers worden vaak geassocieerd met cell identiteit en zijn daarom zeer relevant voor ontwikkelingssystemen. Daarnaast blijkt uit recente studies dat de mate van DNA methylatie varieert van cel tot cel. De mechanismen die zorgen voor deze heterogene methylatie zijn echter nog niet goed onderzocht. Om deze mechanismen te achterhalen creëerden we methylatie reporter cellijnen voor de *Sox2* and miR-290 super-enhancers. Deze cellijnen hebben twee kleuren fluorescente eiwitten, die aan of uit zijn afhankelijk van de methylatie-graad van elk van de twee allelen (de kopieën van de super-enhancers van moeders- dan wel vaderszijde). We tonen aan dat de methylatie-graad van de super-enhancers niet alleen heterogeen is maar ook zeer dynamisch, omdat ze schakelen tussen aan en uit, en vice versa, in een tijdsbestek van enkele dagen. We tonen ook aan dat de methylatie-graad de transcriptie beïnvloed van de genen *in cis* (wat betekent dat de methylatie-graad een lokaal effect heeft op hetzelfde chromosoom). Tot slot nemen we waar dat deze dynamische methylatie niet slechts een *in vitro* artefact is omdat het ook plaatsvindt in pre-implantatie embryo's.

In **Hoofdstuk 2** gebruiken we de transcriptomen van cellen als indicator van celtype. We deden *single-cell transcriptomics* op menselijke foetale nieren in verschillende stadia van ontwikkeling. Nieren bestaan uit ongeveer 1 miljoen nefronen, de functionele eenheden die onafhankelijk kunnen opereren. De ontwikkeling van deze nefronen is asynchroon, wat betekent dat meerdere stadia van nefron-ontwikkeling tegelijk zichtbaar zijn in de foetale nier. In de transcriptomics dataset identificeren wij 22 verschillende celtypes, variërend van voorloper cellen tot volledig gedifferentieerde cellen. Sommige van deze cellen waren nieuwe, meer genuanceerde, subclassificaties van vooraf bekende cellen zoals de *nephron progenitor cells*. We zien ook dat voor de *podocyte* het transcriptoom verder blijft veranderen gedurende ontwikkeling, ondanks dat het celtype wel al vast staat. Dit soort gedetailleerde informatie over nier celtypes kan mogelijk het pad effenen naar de ontwikkeling van medicijnen voor nierziektes in de vorm van regeneratieve therapieën.

In **Hoofdstuk 3** onderzoeken we de translatie en degradatie snelheden in mESCs. Het meten van RNA is dikwijls gemakkelijker dan het meten van eiwit, en vanwege de hiërarchische relatie tussen de twee wordt er vaak van uit gegaan dat veranderingen in RNA ook plaatsvinden in het eiwit. Een plotselinge verandering van RNA concentratie zou zich dan ook moeten uiten in het bijbehorende eiwit. Hoe snel een gen's eiwit concentratie reageert op de RNA concentratie verschilt per gen en verandert dynamisch. Zodoende kan men in een

systeem dat veel veranderingen doorgaat, bijvoorbeeld gedurende differentiatie, niet simpelweg RNA meten en aannemen dat de bijpassende eiwitten gelijke patronen vertonen. In dit hoofdstuk modelleren we eiwit synthese en degradatie snelheden aan de hand van RNA en eiwit data van differentiërende mESCs. Dit stelt ons in staat om de verschillen tussen RNA en eiwit in kaart te brengen. Nog interessanter aan de modellen is dat ze de gevallen identificeren waarbij onze aanname van constante omzetsnelheden *niet* kloppen, wat suggereert dat deze genen post-transcriptioneel worden gereguleerd (gereguleerd in de eiwit synthese of degradatie stappen). Met deze modellen identificeren we nieuwe gevallen van translatie regulatie via microRNA (miR), kleine RNA moleculen die gen-coderende RNAs blokkeren. Interacties tussen mRNAs en miRs zijn doorgaans zeer moeilijk te identificeren.

In **Hoofdstuk 4** worden de modellen van het voorgaande hoofdstuk op de proef gesteld. We wilden enkele van de gevonden miR-gen interacties verifiëren. Om de interacties te testen *transfecteerden* we de cellen met *miR mimics* en *miR inhibitors* (transfecteren: het introduceren van nucleïnezuren of eiwitten in de cel). Om de timing en concentraties van de transfecties te optimaliseren creëerden we cellijnen die transfectie kunnen rapporteren middels fluorescerende eiwitten. Vervolgens transfecteerden we zes verschillende miR mimics, deden we transcriptomics en tonen we aan dat de miRs de concentratie van de voorspelde genen negatief beïnvloeden. Deze resultaten ondersteunen de toepassing van modellen als middel voor het ontdekken van translatie regulatie via een miR.

CURRICULUM VITAE

I was born on March 30 1990 at the *VU medical center* in Amsterdam. I went to *Laar & Berg* high school in Laren (NH) during which time which I took part in a bilingual study program called *Middle Years Programme* (MYP). I received my MYP diploma in 2006 and VWO diploma in 2008. I obtained a Bachelor of Science in Biomedical sciences in 2012 at the *University of Amsterdam* after an internship at the group of Prof. dr. Joost Teixera De Mattos. There I studied the adaptation of cyanobacteria (photosynthetically-capable bacteria) for use as a "fuel factory", with waste products and sunlight as input. Desiring a more analytical continuation of my studies I chose to pursue a Master of Science in Bioinformatics and Systems biology at *VU Amsterdam* as my next step. During this Master's I took part in an international exchange program called *CanSys* (a portmanteau of cancer and systems biology). As a result, I spent three months at the *Université du Luxembourg* (Luxembourg) for an internship in the activation of a receptor involved in gastrointestinal stromal tumors. I then spent 11 months in Buffalo (NY, USA) studying cancer at the *State University of New York at Buffalo*. This included oncology courses and an internship in the group of Dr. Moray Campbell at the *Roswell Park Cancer Institute*. During this internship I created an analytical pipeline for the integration of multiple sources of publicly available data in the context of a nuclear receptor that is involved in multiple types of cancer. I received the Master of Science degree in Natural sciences (interdisciplinary) from the SUNY at Buffalo in 2014 at the end of the CanSys program as well as a Master of Science in Bioinformatics and Systems Biology from the VU Amsterdam.

In 2015 I joined the group of Dr. Stefan Semrau at *Leiden University* as his very first PhD candidate. My position was funded by the research program *Frontiers of Nanoscience* (NanoFront), a consortium of researchers from the fields of quantum nanoscience, bionanoscience and nanotechnology. During my time as a PhD candidate, I worked on a variety of subjects including gene regulation at the levels of the epigenome, genome, transcriptome and proteome. I also worked on developing a new technique and I set up the processing pipelines for transcriptomics data. I attended the workshop *RNA-seq data analysis* by *BioSB*, and the *EMBO* practical course *Single cell omics* in Heidelberg. I presented at various conferences in the Netherlands, USA, Germany and France.

Presently, I am working as a data scientist at the *Rijksinstituut voor volksgezondheid en milieu* as a data scientist dealing with COVID-19 data.

LIST OF PUBLICATIONS

[1] Mark D Long, Patrick R van den Berg, James L Russell, Prashant K Singh, Sebastiano Battaglia, and Moray J Campbell. "Integrative genomic analysis in K562 chronic myelogenous leukemia cells reveals that proximal NCOR1 binding positively regulates genes that govern erythroid differentiation and Imatinib sensitivity." In: *Nucleic Acids Research* 43.15 (2015), pp. 7330–7348. DOI: 10.1093/nar/gkv642.

[2] Patrick R van den Berg, Bogdan Budnik, Nikolai Slavov, and Stefan Semrau. "Dynamic post-transcriptional regulation during embryonic stem cell differentiation". In: *bioRxiv* (2017), p. 123497. DOI: 10.1101/123497.

[3] Prashant K Singh, Patrick R van den Berg, Mark D Long, Angie Vreugdenhil, Laurie Grieshober, Heather M Ochs-Balcom, Jianmin Wang, Sylvie Delcambre, Sami Heikkinen, Carsten Carlberg, Moray J Campbell, and Lara E Sucheston-Campbell. "Integration of VDR genome wide binding and GWAS genetic variation data reveals co-occurrence of VDR and NF-κB binding that is linked to immune phenotypes". In: *BMC genomics* 18.1 (2017), p. 132. DOI: 10.1186/s12864-017-3481-4.

[4] Tobias C Messemaker, Selina M van Leeuwen, Patrick R van den Berg, Anke E J t Jong, Robert-Jan Palstra, Rob C Hoeben, Stefan Semrau, and Harald M M Mikkers. "Allele-specific repression of Sox2 through the long non-coding RNA Sox2ot". In: *Scientific Reports* 8.1 (2018), p. 386. DOI: 10.1038/s41598-017-18649-4.

[5] Mazène Hochane, Patrick R van den Berg, Xueying Fan, Noémie Bérenger-Currias, Esmée Adegeest, Monika Bialecka, Maaike Nieveen, Maarten Menschaart, Susana M Chuva de Sousa Lopes, and Stefan Semrau. "Single-cell transcriptomics reveals gene expression dynamics of human fetal kidney development". In: *PLOS Biology* 17.2 (Feb. 2019), e3000152. DOI: 10.1371/journal.pbio.3000152.

[6] Mark D Long, Prashant K Singh, James R Russell, Gerard Llimos, Spencer Rosario, Abbas Rizvi, Patrick R van den Berg, Jason Kirk, Lara E Sucheston-Campbell, Dominic J Smiraglia, and Moray J Campbell. "The miR-96 and RARγ signaling axis governs androgen signaling and prostate cancer progression". In: *Oncogene* 38.3 (2019), pp. 421–444. DOI: 10.1038/s41388-018-0450-6.

[7] Yuelin Song, Patrick R van den Berg, Styliani Markoulaki, Frank Soldner, Alessandra Dall'Agnese, Jonathan E Henninger, Jesse Drotar, Nicholas Rosenau, Malkiel A Cohen, Richard A Young, Stefan Semrau, Yonatan Stelzer, and Rudolf Jaenisch. "Dynamic Enhancer DNA Methylation as Basis for Transcriptional and Cellular Heterogeneity of ESCs". In: *Molecular cell* 0.0 (2019), 905–920.e6. DOI: 10.1016/j.molcel.2019.06.045.

[8] Esmée Adegeest, Noémie Bérenger-Currias, Patrick R van den Berg, Marleen Feliksik, Mazène Hochane, Maria Mircea, and Stefan Semrau. *Scrum for Science blogpost*. 2020.

ACKNOWLEDGEMENTS

A tremendous number of people have helped me throughout the journey of obtaining my PhD. I will attempt to express the extent of my gratitude to these people in the coming section.

First and foremost, I would like to to thank Stefan Semrau. Stefan, you have been more than just a mentor throughout my PhD; your enthusiasm and wisdom has been crucial to my progress and I am extremely grateful to be your graduate student. Other members of our group have all been directly involved in my work, and I deeply appreciate your support in all its forms over the past years: Kate Sokolova, Noémie Bérenger-Currias, Marleen Feliksik, Mazène Hochane, Esmée Adegeest, Maria Mircea, Maarten Menschaart and Thomas Pool.

The Physics of Life (FvL) group would not be the same without its leaders and staff, so many thanks to Thomas Schmidt, John van Noort, Doris Heinrich and Yvonne Kerkhof for providing the foundation of an amazing working environment. I would also like to thank the other members of FvL; you gave me all kinds of support, you made my time at Leiden University more sociable and you set important examples of what it means to be a scientist: Lena Beletkaia, Nelli Bossert, Thomas Brouwer, Sara Carozza, Stefano Coppola, Julia Eckert, Jeremy Ernst, Klaas Hermans, Olga Iendeltseva, Babette de Jong, Artur Kaczmarcyk, Veer Keizer, Gert-jan Kuijntjes, Kirsten Martens, Maria Mytiliniou, Lionel Ndamba, Sylvie Olthuis, Chi Pham, Wim Pomp, Ivo Severins, Redmar Vlieg and Joeri Wondergem. Honorable mentions outside of FvL that left a positive mark on me include: Martin Caldarola, Aquiles Carattino, Thomas Jollans and Marco Tompitak.

The whole is greater than the sum of its parts and this holds especially true for science. Many thanks to all the people who collaborated with me, with special mention to Nikolai Slavov, Harald Mikkers, Susana Chuva de Sousa Lopes, Yuelin Song and Peter van Veelen.

Finally, I would like to thank my family and my friends outside of Leiden. Your encouragement has proven invaluable.

Casimir PhD Series 2021-9
ISBN 978-90-8593-475-2