

Universiteit
Leiden
The Netherlands

Unravelling cell fate decisions through single cell methods and mathematical models

Mircea, M.

Citation

Mircea, M. (2022, December 20). *Unravelling cell fate decisions through single cell methods and mathematical models*. *Casimir PhD Series*. Retrieved from https://hdl.handle.net/1887/3505763

Version:	Publisher's Version
License:	Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden
Downloaded from:	https://hdl.handle.net/1887/3505763

Note: To cite this publication please use the final published version (if applicable).

Unravelling cell fate decisions through single cell methods and mathematical models

- Maria Mircea -

Unravelling cell fate decisions through single cell methods and mathematical models

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 dinsdag 20 december 2022
klokke 15:00 uur

door

Maria Mircea

geboren te Boekarest, Roemenië
in 1993

Promotores:	Dr. D. Garlaschelli Prof. dr. T. Schmidt
Co-promotor:	Dr. S. Semrau
Promotiecommissie:	Dr. A. Mahfouz (Leids Universitair Medisch Centrum) Dr. V. V. Orlova (Leids Universitair Medisch Centrum) Prof. dr. J. Aarts Prof. dr. R. M. H. Merks Prof. dr. ir. S. J. T. van Noort

© 2022 by M. Mircea. All rights reserved.

Cover (front): The path of a single stem cell during cellular differentiation located on a potential landscape on top of a gene regulatory network.
Cover (back): Molecules involved in cellular differentiation.

Casimir PhD Series, Delft-Leiden 2022-30
ISBN 978-90-8593-540-7

An electronic version of this dissertation is available at
https://openaccess.leidenuniv.nl

The first principle is that you must not fool yourself – and you are the easiest person to fool.

Richard Feynman

Contents

1 How a cell decides its own fate: A single-cell view of molecular mechanisms and dynamics of cell type specification 1
1.1 Introduction . . . 2
1.2 Molecular embodiment of a cell type . . . 3
1.3 Molecular profiling . . . 5
1.4 Dynamics of differentiation . . . 8
1.5 Measurement of differentiation dynamics . . . 9
1.6 Data analysis . . . 12
1.7 Conceptual framework . . . 14
1.8 Perspectives . . . 15
1.9 Outline of Thesis . . . 16

2 Phiclust: A clusterability measure for single-cell transcriptomics reveals phenotypic subpopulations 29
2.1 Background . . . 30
2.2 Results . . . 33
2.2.1 Phiclust is derived from first principles and does not have free parameters . . . 33
2.2.2 Phiclust is a proxy for clusterability . . . 39
2.2.3 Confounder regression removes unwanted variability . . . 40
2.2.4 Phiclust has high sensitivity for the detection of sub-structure . . . 42
2.2.5 Genes responsible for the detected substructure can be identified . 45
2.2.6 Application of phiclust to a BMNC data set drives the discovery of biologically meaningful sub-clusters . . . 46
2.2.7 Phiclust reveals subpopulations in a fetal human kidney data set that can be confirmed experimentally . . . 50
2.3 Discussion . . . 54
2.4 Conclusion . . . 55
2.5 Methods . . . 56
2.5.1 Preprocessing . . . 56
2.5.2 Phiclust . . . 56
2.5.3 Clustering . . . 58
2.5.4 Variance driving genes . . . 59
2.5.5 Data sets . . . 60
2.5.6 Single cell data analysis . . . 61
2.5.7 Staining . . . 61
2.5.8 Data availability . . . 62
2.6 Supplementary Note 1 . . . 63
2.7 Supplementary Note 2 . . . 75

3 ETV2 upregulation marks the specification of early cardiomyocytes and endothelial cells during co-differentiation 81
3.1 Introduction . . . 82
3.2 Results . . . 84
3.2.1 ETV2 is upregulated after a bifurcation into CMs and ECs . . . 84
3.2.2 Generation and characterization of an ETV2^{mCherry} fluorescent hiPSC reporter line . . . 87
3.2.3 The ETV2^{mCherry} fluorescent reporter allows for the purification of differentiating cells with lineage-specific expression profiles. . . . 89
3.2.4 ETV2+ cells contain lineage-restricted subpopulations . . . 91
3.3 Discussion . . . 93
3.4 Conclusion . . . 94
3.5 Materials and Methods . . . 95
3.5.1 Experimental Methods . . . 95
3.5.2 Bulk RNA sequencing and analysis . . . 96
3.5.3 Single-cell RNA sequencing and analysis . . . 97
3.5.4 Data availability . . . 99

4 Tissue microenvironment partially removes signatures of developmental origin in a 3D in vitro model of cardiac endothelial cell differentiation 105
4.1 Introduction . . . 106
4.2 Results . . . 108
4.2.1 Derivation of endothelial cells from different mesodermal origins . 108
4.2.2 Transcriptomic profiling of CMECs and PMECs . . . 110
4.2.3 Characterization of CMEC and PMEC differentiation by single-cell RNA-seq . . . 111
4.2.4 hiPSC-ECs acquired organ-specific signatures in a cardiac microenvironment . . . 114
4.2.5 Extraction of organ-specific signatures of human fetal heart ECs from a published scRNA-seq dataset . . . 116
4.2.6 Both CMECs and PMECs acquired intra-myocardial identity in MT culture . . . 117
4.3 Discussion . . . 119
4.4 Materials and Methods . . . 121
4.4.1 Experimental Methods . . . 121
4.4.2 Bulk RNA sequencing (RNA-seq) and analysis . . . 122
4.4.3 Single-cell RNA sequencing and analysis . . . 123
4.4.4 Data availability . . . 125

5 A gastruloid model of the interaction between embryonic and extra-embryonic cell types 131
5.1 Introduction . . . 132
5.2 Results . . . 133
5.2.1 XEN cells induce neuroepithelial structures in XEN enhanced gastruloids. . . . 133

5.2.2 Neuroepithelial cells in XEGs are heterogeneous and show further specification . . . 135
5.2.3 Signaling perturbation experiments and further differentiation support the neuroepithelial character . . . 137
5.2.4 Single-cell RNA-seq reveals the transcriptional profiles of XEG cells 140
5.2.5 Most XEN cells become VE-like in XEGs . . . 145
5.2.6 XEN cells guide symmetry breaking by local inhibition of WNT signaling . . . 146
5.3 Discussion . . . 152
5.4 Methods . . . 154
5.4.1 Experimental Methods . . . 154
5.4.2 Computational methods . . . 158
5.4.3 Data availability . . . 162

6 Gene regulatory network inference with physics informed neural networks 169
6.1 Introduction . . . 170
6.2 Results . . . 172
6.2.1 Cell communication drives bifurcations in a GRN model of differentiation . . . 172
6.2.2 Feedforward NN regression is unsuitable for GRN parameter inference . . . 175
6.2.3 Physics informed neural networks can infer GRN parameters from partial and noisy data. . . . 176
6.2.4 PINNs can infer GRN parameters from snapshot data in the absence of cell communication. . . . 181
6.3 Discussion . . . 182
6.4 Materials and Methods . . . 185
6.4.1 Inference of a GRN with cell communication from trajectories . . . 185
6.4.2 Feed-forward neural network . . . 186
6.4.3 Physics Informed Neural Network . . . 187
6.4.4 Inference of a GRN without cell communication from snapshot data 188

Summary 195

Samenvatting 201

Zusammenfassung 207

List of Publications 213

Curriculum Vitæ 215

1

How a cell decides its own fate: A single-cell view of molecular mechanisms and dynamics of cell type specification

On its path from a fertilized egg to one of the many cell types in a multicellular organism, a cell turns the blank canvas of its early embryonic state into a molecular profile fine-tuned to achieve a vital organismal function. This remarkable transformation emerges from the interplay between dynamically changing external signals, the cell's internal, variable state and a tremendously complex molecular machinery we are only beginning to understand. Recently developed single-cell omics techniques have started to provide an unprecedented, comprehensive view of the molecular changes during cell type specification and promise to reveal the underlying gene regulatory mechanism. The exponentially increasing amount of quantitative molecular data being created at the moment is slated to inform predictive, mathematical models. Such models can suggest novel ways to manipulate cell types experimentally, which has important biomedical applications. This review is meant to give the reader a starting point to participate in this exciting phase of molecular developmental biology. We first introduce some of the principal molecular players involved in cell type specification and discuss the important organizing ability of biomolecular condensates, which has been discovered recently. We then review some of the most important single-cell omics methods and relevant findings they produced. We devote special attention to the dynamics of the molecular changes and discuss methods to measure them, most importantly lineage tracing. Finally, we introduce a conceptual framework that connects all molecular agents in a mathematical model and helps us make sense of the experimental data.

This chapter is based on **Maria Mircea**, Stefan Semrau; How a cell decides its own fate: a single-cell view of molecular mechanisms and dynamics of cell-type specification. Biochem Soc Trans 17 December 2021; 49 (6): 2509–2525. doi: https://doi.org/10.1042/BST20210135 [1]

1

1.1 Introduction

Single-cell omics technologies:
Experimental methods to measure the entire genome, epigenome, transcriptome, proteome etc. of a cell in high-throughput.
Pluripotency:
Ability of a cell to give rise to multiple cell types.
Histone:
Proteins that are crucial for the organization of DNA in the nucleus. DNA is tightly wound around nucleosome core particles which consist of 8 histones.

What is a cell type, anyway? Traditionally, cell types have been defined by their function within an organism: Neurons process and transmit information, macrophages remove harmful microorganisms and podocytes are crucial for blood filtration in the kidney. As function can be difficult to ascertain, especially for subtle variants of cell types, cell morphology and the presence of certain marker genes are often used as proxies [2, 3]. With the advent of single-cell **omics technologies**, cell types have increasingly come to be identified with their molecular profiles. While most cell types persist over long periods of time, often the entire life span of an adult organism, cells are found in short-lived, transient states such as different phases of the cell cycle, different metabolic states or multiple forms of stress response. Here, we are only concerned with the specification of cell types, which occurs during embryonic development or regeneration of adult tissues. During development, **pluripotent** embryonic cells differentiate into progenitors with diminishing developmental potential, and eventually fully specified cell types [2, 4–6]. In adult tissue, long-lived adult stem cells give rise to multiple types of descendants. These processes are collectively termed differentiation. Differentiation involves changes in gene expression (i.e., messenger RNA and protein levels), which are accompanied and guided by epigenetic changes. Broadly, the epigenetic profile of a cell encompasses any heritable molecular mark, with the exceptions of changes in the DNA sequence [7, 8]. Two of the most important epigenetic marks are DNA methylation and **histone** modifications. These marks are tightly linked to the accessibility of the DNA and thus influence the expression of specific genes. A comprehensive introduction to epigenetics can be found in [9].

Here, we will review a few of the many cell-autonomous molecular mechanism that make differentiation a reproducible process and ensure the long-term stability of cell types. Importantly, cells do not develop in isolation. Their communication with neighboring cells via chemical and mechanical signaling is an integral part of embryonic development and tissue regeneration, which we will not discuss here (for recent reviews see [10, 11]). Equally important, but also outside the scope of this review, is the role of molecular noise, which can drive cell type decisions but must also be controlled to ensure the stability of the fully differentiated state (for recent reviews see [12, 13]). In this review, we will first introduce some of the most important molecular players, which can be used to define a cell type, and discuss omics techniques that can measure molecular profiles comprehensively in single cells. We will then focus on the dynamics of differentiation and novel methods that allow the inference of the developmental lineage tree. Finally, we will discuss challenges arising in the analysis of data sets comprising multiple modalities and a conceptual framework that enables a quantitative understanding of differentiation (Figure 1.1).

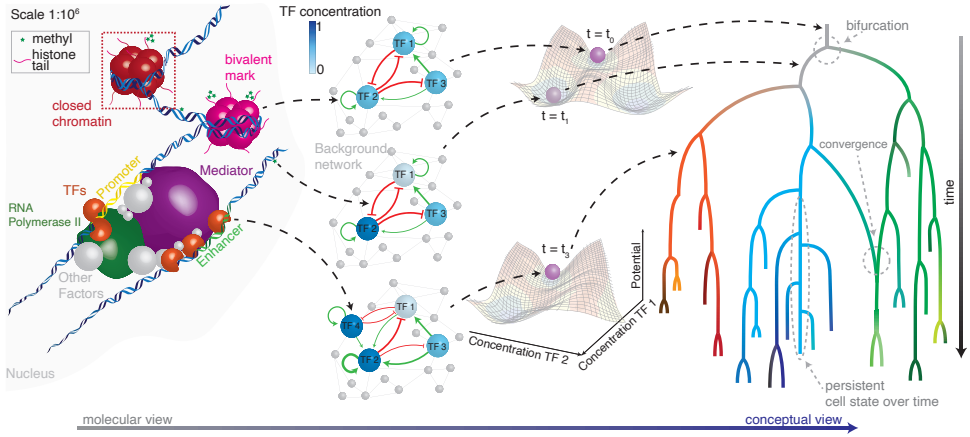


Figure 1.1: Conceptual framework for cell types and differentiation dynamics. Left: TFs bind to enhancer and promoter regions depending on the chromatin state. All molecules are drawn at a scale of 1:106. Center left: A simplified gene regulatory network consisting of TFs (nodes of the network graph) and their interactions (edges of the network graph). Center right: A gene regulatory network can be modeled by a dynamical system, which can be represented by a potential energy landscape. The position in the landscape is determined by TF abundance. The shape of the landscape depends on the TF interactions. Cells follow the path of steepest descent to stable states, which correspond to cell types. Right: The collection of trajectories in the potential landscape form a cell type decision tree, which highlights the hierarchical nature of differentiation.

1.2 Molecular embodiment of a cell type

Most, if not all, cell type decisions involve specific **transcription factors** (TFs) [2, 3, 5, 6, 14, 15]. These DNA binding proteins control a gene's transcription level by binding to **cis-regulatory elements** (CREs) in the DNA. Enhancers, CREs that can be found at large distances from the regulated gene, play a particularly important role for cell type determination. Enhancers work in concert and physically interact with promoters, another type of CRE that is usually found near the regulated gene. TF binding of CREs not only depends on the presence of specific DNA sequence motifs but is also strongly modulated by the configuration of the **chromatin** (the complex of DNA, **nucleosomes** and other associated proteins, see Figure 1, left). With the exception of so-called **pioneer factors** [16], TFs only bind accessible, nucleosome-free DNA [17]. Chromatin configuration and TF binding are affected by chemical modifications of histones (the components of a nucleosome), as well as the DNA [17]. Different histone modifications, or marks, are associated with different functions, broadly categorized as active and repressive, and the effect of DNA modifications strongly depends on the genomic context [18, 19]. For example, DNA methylation at the enhancer regions of the pluripotency gene Sox2 results in silencing of its expression in embryonic stem cells [19]. Importantly, the interaction between TFs and chromatin configuration is reciprocal: TFs recruit enzymes that locally

Transcription factor:
A protein that binds to specific DNA sequences and regulates transcription.
Cis-regulatory elements:
Sequences of non-coding DNA which regulate the transcription of genes.
Chromatin:
The complex of nucleosomes, DNA and other associated proteins.
Nucleosome:
Smallest unit of DNA organization. Consists of DNA wound around 8 histones.
Pioneer factor:
A transcription factor that can bind to nucleosome-bound DNA.

change the molecular make-up of the chromatin [17]. Both histone marks and DNA methylation are heritable molecular marks, as they are copied to the newly synthesized DNA during cell division [8, 9, 20]. They can therefore function as long-term memory of a cell's molecular profile and hence cell type. The pattern of chromatin accessibility and epigenetic marks can thus be used to identify a cell type and reveal relevant CREs [17, 21]. Importantly, cell type specification cannot be understood by studying individual TFs or epigenetic features in isolation. Cell types rather emerge from the complex interactions of several TFs. The presence of particular subsets of TFs has therefore been used to define a periodic table of cell types [5]. Together with their target genes, TFs form **gene regulatory networks** that establish and maintain cell identity [2], see Figure 1, middle. Regulatory interactions between TFs, in particular negative feedback loops, are crucial for the stability of molecular states. Due to the presence of fluctuations in the environment as well as the internal state of the cell, robustness is an important requirement for regulatory networks. At the same time, they need to be dynamic and react appropriately to external signaling inputs [2]. Mutual repression of TFs is one mechanism by which multiple, alternative cell types can be created. A prominent example is the interaction between the TFs GATA6 and NANOG, which governs the lineage decision between two of the earliest cell types in the mammalian embryo [22–24]. The conceptual framework discussed in the final section of this review explains how various stable cell types and unidirectional differentiation dynamics emerge from gene regulatory networks, see Figure 1.1, right.

Regulatory network:
A system of interacting molecules that regulate each other's gene expression as well as a set of target genes.
Master transcription factor / master regulator:
A transcription factor that effects the transcription of multiple downstream genes and is essential for cell type specification.
Super-enhancer:
A group of multiple enhancers in close proximity characterized by high levels of Mediator complex, which strongly drives gene expression of its target genes.
Mediator:
A multiprotein complex that coactivates transcription by interacting with TFs and RNA polymerase II.
Biomolecular condensates:
Droplets of a condensed liquid phase formed in cells by homotypic, multivalent interactions (i.e., interactions between identical molecules that involve multiple binding sites). One example are membrane-less organelles.
Liquid-liquid phase separation:
De-mixing of a homogeneous liquid into two distinct liquid phases.

Despite the fact that TFs always work in concert, some have a particularly large impact on lineage decisions: Overexpression of certain TFs can revert a differentiated cell back to a pluripotent state (reprogramming) or convert one cell type into another (transdifferentiation) [25]. The remarkable power of these TFs, termed **master TFs** or **master regulators**, can be rationalized by their DNA binding patterns. Master TFs have been shown to bind clusters of enhancers, or **super-enhancers**, which drive high levels of key cell type-specific genes [26, 27], see Figure 1.2. Super-enhancers owe their special role to a high density of co-localized **Mediator** complex [26, 27], a protein complex that links TF binding to the recruitment of the transcription machinery and therefore gene expression. A well-studied example of master TFs that bind cell-type specific super-enhancers are regulators of the pluripotent state in embryonic stem cells: NANOG, SOX2 and OCT4 [27]. TFs, CREs, epigenetic marks and enzymes that modify chromatin state are just a small, albeit important, subset of the many molecular species that are involved in cell type specification. It has long been unclear, how all of these mobile molecules, some of which are freely diffusing in the nucleus, can interact in an efficient manner. Recently, **biomolecular condensates**, which form through **liquid-liquid phase separation** (LLPS) [28–30], have been suggested as a possible answer to this question. Biomolecular condensates form

Paraspeckle:
A biomolecular condensate that forms in the presence of the long non-coding RNA NEAT1 and several RNA binding proteins.
Intrinsically disordered regions:
Segments of a protein that do not form a stable three-dimensional structure.

according to well-known thermodynamic principles as a result of multivalent, homotypic interactions between molecules [31]. The high concentration of several molecular species in the condensed phase leads to increased interaction rates [30]. Examples of biomolecular condensates are the well-known membrane-less organelles, such as the nucleolus or Cajal bodies, as well as **paraspeckles** and many more [28, 29, 32]. It has been found that **intrinsically disordered regions** (IDRs) of proteins can lead to the multivalent interactions that can cause condensates to form [33–35]. Interestingly, MED1, a member of the Mediator complex, and BRD4, a coactivator of transcription, have large IDRs and form condensates at super-enhancers [33], see Figure 1.2. Thus, phase-separated condensates likely concentrate components of the transcription apparatus and thereby ensure robust transcription of key cell type-specific genes. Additionally, the large size of the Mediator cluster at super-enhancers enables the contact with multiple promoter sites [36]. Therefore, biomolecular condensates are likely of crucial importance for establishing a cell type.

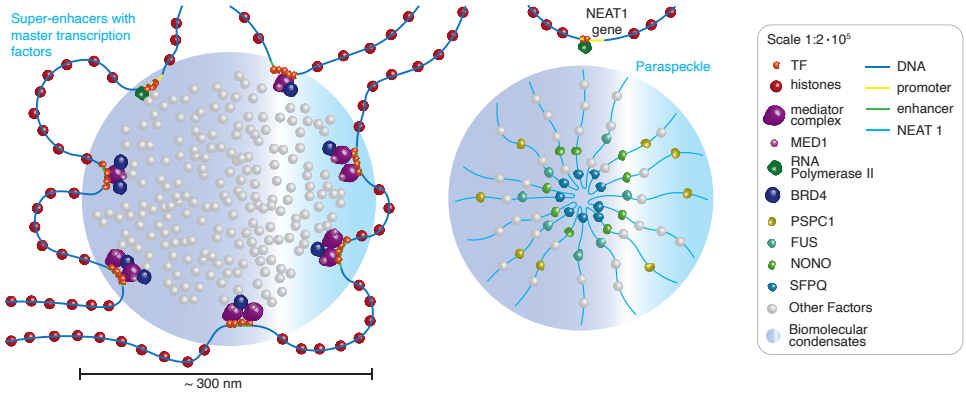


Figure 1.2: Master transcription factors and super-enhancers play a major role in guiding cell type specification and are compartmentalized by biomolecular condensates. Left: super-enhancer with bound master transcription factors in a large biomolecular condensate. Right: A paraspeckle that forms in the presence of the lncRNA NEAT1. All molecules are drawn at a scale of $1 : 2 \cdot 10^5$.

1.3 Molecular profiling

In recent years, omics technologies have emerged that measure one or multiple molecular species comprehensively in single cells (see Box 3 and Figure 1.3 for a selection of common methods). These technologies can reveal cell type-specific molecular profiles in high throughput. With single-cell RNA-sequencing (scRNA-seq) the transcriptomes of individual cells can be obtained [2, 5, 6], which enables the identification of new cell types and cell states in complex tissues [37]. Multiple large consortia are currently generating transcriptional atlases of entire organisms (reviewed in [38]). The human cell atlas [39]

and Tabula Muris [40] are two prominent examples. Notwithstanding the great value of scRNA-seq measurements, gene expression should ideally be measured at the protein level. Numerous regulatory mechanisms at the translational and post-translational level make mRNA abundance just a proxy for protein abundance. Protein measurements have indeed revealed phenotypic features that could not be discerned with scRNA-seq alone [41, 42]. As

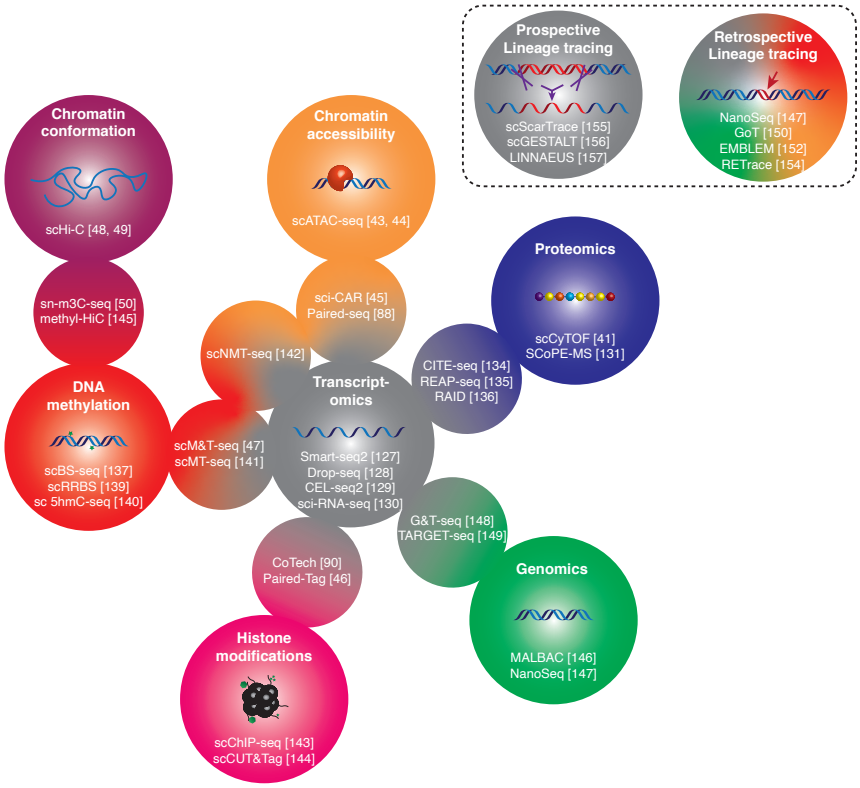


Figure 1.3: A selection of single-cell omics and multi-omics techniques useful for studying cell type specification. Colors indicate the measurement of different molecular species in a cell (green: DNA sequence, grey: RNA abundance, blue: protein abundance , orange: chromatin accessibility, red: DNA methylation, pink: histone modifications, violet: chromatin conformation). Circles with color gradients contain techniques that measure several types of molecules at the same time. Dashed lines envelop techniques that aim at understanding similar concepts.

mentioned before, chromatin state is an important factor in gene regulation. Knowledge of the chromatin landscape can therefore improve the identification of cell types [43]. There is a growing variety of single-cell methods that measure chromatin features. For example, by using scATAC-seq [44, 45], which reveals accessible chromatin regions in single cells, it is possible to identify cell type-specific regulatory elements and candidate master TFs. By combining scATAC-seq and scRNA-seq, open chromatin regions can be associated with active transcription, which improves the identification of TFs and target genes compared

to pure RNA measurements. It was also shown that considering the chromatin state of distal CREs significantly increased the power to predict cell-type specific gene expression, compared to using promoter chromatin state alone [46].

Box 1: Common single-cell omics and multi-omics techniques

Transcriptomics: The currently most prevalent single-cell omics method is single-cell RNA sequencing, which measures RNA abundance. Different experimental implementations of this method include: Smart-seq2 [47], Drop-seq [48], CEL-seq2 [49] and Sci-RNA-seq [50].

Proteomics: It is not yet feasible to measure every protein in a single cell. Antibody-based methods, for example scCyTOF [42], can measure hundreds of proteins, but cannot easily be scaled to the whole proteome and rely on the existence of highly specific antibodies. Mass spectrometry-based proteomics methods, which recently became available, might soon produce high-quality proteomes of single cells. One example is SCoPE-MS [51], which detects around 1000 proteins per cell. Improvements to this method have been recently made in SCoPE2 [52] and another method [53]. By sequencing of DNA-tagged antibodies, the quantification of hundreds of proteins together with the transcriptome in single cells is possible with CITE-seq [54] and REAP-seq [55]. RAID [136] uses RNA-tagged antibodies for the same purpose.

Epigenomics: To gain insights into chromatin accessibility, one of the most prominent techniques is scATAC-seq [44, 45], which uses transposons to barcode accessible DNA. scATAC-seq can be performed simultaneously with scRNA-seq, which was implemented, for example, by sci-CAR [46] and Paired-seq [56]. DNA methylation is measured by scBS-seq [57, 58] and scRRBS [59]. Single-cell 5hmC-seq measures DNA hydroxymethylation [60]. Joint measurements of DNA methylation and the transcriptome is possible with, for example, scM&T-seq [61] and scMT-seq [62]. A method to measure all three molecular profiles (DNA methylation, transcriptome and chromatin accessibility) is scNMT-seq [63]. Histone modifications can be measured, for example, with scChIP-seq [64] and scCUT&Tag [65]. New methods to measure the transcriptome jointly with histone modifications are CoTech [66] and PairedTag [67]. It is now also possible to study the chromatin conformation in every single cell with scHi-C [68, 69]. Recent methods have allowed the capture of both chromatin conformation and DNA methylation (sn-m3C-seq [70] and methyl-HiC [71]).

Genomics: The DNA sequence of single cells can be measured by methods such as MALBAC [72] and NanoSeq [73]. NanoSeq which has been designed to detect even small somatic mutations in single DNA molecules. Measurement methods that combine DNA sequencing with transcriptomics are, for example, G&T-seq [74] and TARGET-seq [75].

Similarly, a simultaneous measurement of histone modifications and transcriptome showed that active enhancers are epigenetically more variable across cell types than promoter regions [67]. A related finding resulted from the simultaneous measurement of DNA methylation and transcriptome (scM& T-seq [61]). The authors confirmed that promoter DNA methylation in mouse ESCs is typically correlated with reduced gene expression. By contrast, DNA methylation of distal enhancers is more often correlated with increased gene expression, compared to promoters. Since active transcription typically requires the physical proximity of enhancers and promoters, knowledge of chromatin organization can be helpful to understand cell type decision making.
scHi-C is a high-throughput method to reveal chromatin interactions throughout the genome in single cells [68, 69]. In combination with DNA methylation measurements, cell type specific chromatin conformations can be obtained [70], which might help to clarify the role of biomolecular condensates [76, 77]. In a recent study, a new variant of Hi-C [77] was used to determine the stability of chromatin interactions, which were revealed to vary substantially between organelles. Approaches to measure the spatial distribution of transcripts [78, 79] and proteins [80] with sub-cellular resolution might lead to an even better understanding of cellular compartmentalization through biomolecular condensates. Multi-omics single-cell methods, like those presented here, promise to enable an improved mechanistic understanding of cell type-specific gene regulation [81]. A more comprehensive discussion of these methods can be found in [82].

1.4 Dynamics of differentiation

Chromatin remodeler:
Protein complexe that catalyze molecular changes of the chromosome, such as nucleosome removal.
Bivalent domain:
Chromatin domain that carries both activating and repressing histone marks.
RNA Polymerase II:
A multiprotein complex that transcribes DNA into messenger RNA.

During differentiation, the molecular profile of a cell is remodeled substantially. TFs are, unsurprisingly, important drivers of this transformation. As the majority of TFs binds to accessible chromatin regions, differentiation is accompanied by pervasive changes in chromatin accessibility [13, 16, 17]. One underlying mechanism involves pioneer TFs, which bind to nucleosome-associated DNA and create an open chromatin state [16, 17]. These TFs can explore nucleosomal DNA through non-specific and transient binding, which in turn allows partial opening of the chromatin and other, non-pioneering factors to bind [16]. This mechanism has been recently validated, for example, for the pioneer factor PAX7 [83]. Another mechanism is passive competition of TFs for DNA binding during short periods of local chromatin opening, which increases and stabilizes with higher TF concentrations [17, 84].
Chromatin state is also influenced by **chromatin remodelers** that are recruited by TFs [16, 18, 85] and bind to different histone marks [85, 86]. This is one mechanism by which epigenetic marks strongly impact chromatin accessibility. Importantly, activating and repressing histone marks can also occur simultaneously, on the same nucleosome. These **bivalent domains** play a particularly important role in cell type decisions [87] and are more abundant in embryonic stem cells (ESCs) than in adult tissues. A prominent example is the combination of H3K4me3 (Trimethylation of histone H3 on lysine 4) which is associated with active transcription and H3K27me3 (Trimethylation of histone H3 on lysine

27), which causes chromatin compaction and is thus a repressive mark. It has been shown that bivalent domains are positioned at key TF genes that are important for development [87–89]. Enhancers and promoters with bivalent marks are thought to be in a poised state, that can be quickly resolved to either activation or repression. This effect can be mediated by multiprotein complexes composed of polycomb group (PcG) proteins. These proteins cause gene silencing by, for example, catalyzing methylation of H3K27 [90, 91]. In ESCs, the occupancy of PcG proteins at bivalent histone marks can change during differentiation, which results in altered gene expression [89, 92]. Poised enhancers have been found to be necessary, for example, for the differentiation into specific neural cell types [93]. The examples mentioned here are only few of many epigenetic mechanisms that drive dynamic chromatin remodeling during stem cell differentiation (reviewed in [94]).

Epigenetic marks are not homogeneously distributed in the nucleus, but rather need to be localized at important regulatory sites, which might be promoted by biomolecular condensates. The formation of biomolecular condensates during differentiation has been linked to different long non-coding RNAs (lncRNA) and RNA-binding proteins (RBPs) [32, 95]. For example, the lncRNA DIGIT forms biomolecular condensates together with the RNA binding protein BRD3, which contains an IDR [95]. BRD3 is recruited to sites of the activating histone mark H3K18ac (Acetylation of histone H3 on lysine 18). Paraspeckles are another important class of biomolecular condensates defined by the presence of the lncRNA NEAT1, which recruits several RBPs [96], see Figure 1.2. These condensates can, for example, influence transcriptional regulation via associated RBPs [96–98]. NEAT1 has been found to physically interact with EZH2, a PcG protein, which is involved in catalyzing histone methylation [99]. Interestingly, paraspeckles were found to be involved in slowing down the differentiation process and their number changes dynamically during differentiation to several lineages [32, 97]. Another example of dynamic transcriptional regulation through biomolecular condensates is the association of RNA polymerase II with Mediator condensates (see Figure 1.2). It has been found that, upon phosphorylation, **RNA polymerase II** transitions from condensates involved in transcription initiation to condensates involved in RNA splicing at genes associated with super-enhancers [35].

1.5 Measurement of differentiation dynamics

In simple organisms the entire lineage tree can be assembled using a microscope [100]. In larger organisms that becomes unfeasible and scRNA-seq data of developing tissues has been used instead to infer lineage relationships [101]. Due to asynchrony in embryonic development or regeneration of adult tissues, a single scRNA-seq measurement can capture cells in different stages of differentiation [43, 102] and developmental order, or pseudotime, can be inferred by computational methods (reviewed in [103], see also the section on data analysis below). If the developmental process is sufficiently accessible for repeated sampling, scRNA-seq measurements at several time points can be used to resolve developmental dynamics [14, 104–107]. This approach improves the temporal resolution and revealed that cells with different lineage histories can converge to globally similar cells [108]. However, combining multiple data sets to infer the correct developmental trajectory is challenging.

Lineage reconstruction has also been performed based on protein measurements in single

cells at different time points. In a recent study [108], 27 proteins, of which 16 were TFs, were measured over a time course of 22 days during hematopoietic differentiation. This study showed that, at the protein level, cell type decisions are accompanied by gradual changes in lineage specific TFs, as no abrupt switches in TF levels were observed.

Box 2: Lineage tracing

There are two, conceptionally distinct approaches to lineage tracing: retrospective and prospective. In retrospective lineage tracing, lineage relationships are inferred from naturally occurring somatic mutations. These mutations can be traced using DNA sequencing methods [73]. In a recent study, such mutations were linked with scRNA-seq data to investigate clonal relationships and cell types in human [109]. Mitochondrial DNA has a 10 fold higher mutation rate than nuclear DNA [110, 111], which makes it a good candidate for retrospective lineage tracing. Interestingly, these mutations can be tracked with ATAC-seq measurements because mitochondrial DNA is accessible [111]. DNA methylation also undergoes stochastic changes during cell division known as epimutations, which allows tracking of lineage histories through measurements of DNA methylation [110, 112]. Coupling genomics to DNA methylation measurements allows both lineage tracing and the study of cell type specific methylation patterns [113]. However, naturally occurring mutations are rare, which requires highly accurate and sensitive measurement techniques and computational methods. In prospective lineage tracing, heritable markers are introduced that are read out at a later time point. The most recently developed dynamic lineage tracing methods insert 'scars' into the DNA at random or pre-determined locations, resulting in a large variety of different markers [114, 115]. In some cases, these markers, or barcodes, are also transcribed, so that scRNA-seq is able to capture transcriptomes and lineage information simultaneously. Different omics technologies have been used in the context of lineage tracing (see Box 1 for a list of omics techniques). For retrospective lineage tracing, NanoSeq [73] has been used to track even small somatic mutations and GoT [109] linked transcriptomics to genotyping. scATAC-seq has been used to track mutations in mitochondrial DNA [111] and scRRBS has been used together with DNA sequencing to track DNA mutations together with DNA methylation [113]. Examples of prospective lineage tracing techniques that use transcriptomics measurements are scScarTrace [116], scGESTALT [117] and LINNAEUS [118].

To reveal the gene regulatory programs that cause gene expression changes, chromatin conformation measurements during development can be used. Bulk methods have been used extensively to measure epigenetic changes and chromatin accessibility of cell populations [119], which produced many important insights. However, cell-to-cell variability and rare cell populations can only be distinguished with single-cell methods. Therefore, time-resolved single cell chromatin accessibility measurements can be very informative [120], in particular in combination with transcriptomics [46, 56, 121, 122]. One study found a class of genes with a high number of putative enhancers whose chromatin accessibility is predictive of gene expression [122]. These genes are enriched in TFs that regulate cell type-specific gene expression. These findings suggest participation in super-enhancers and

a central role in cell-type specification. Additionally, it was observed that the expression of TFs precedes the accessibility of their target sites, which might indicate a causal role of TFs in chromatin remodeling, possibly through additional epigenetic mechanisms [121]. Another interesting case is the simultaneous measurement of chromatin accessibility, DNA methylation and transcriptome (scNMT [123]) at several timepoints in mouse development. The authors were able to study the dynamic changes of all three profiles in time and confirmed ectoderm, one of the three embryonic germ layers, as the default developmental pathway. Specific histone marks have also been measured during differentiation and development. For example, the co-occurrence of H3K4me3 and H3K27me3 (bivalent mark) was measured in mouse ESCs together with scRNA-seq. The authors calculated a bivalency score along an RNA based pseudotime trajectory and were able to classify genes by trends in bivalency dynamics [66]. A similar method found a significant overlap between H3K27ac (Acetylation of histone H3 on lysine 27) and H3K27me3 in the adult mouse brain at CREs related to forebrain development [67].

An entirely different approach to study developmental dynamics is used in lineage tracing techniques [114, 115, 124] (see Box 2, which aim to find the correct phylogenetic tree [125, 126] from pluripotent cells to fully specified cell types. Lineage tracing methods have produced a large number of valuable insights. A recent study used lineage tracing to reveal early biases towards particular cell types [127] that are not resolved with transcriptomics: Transcriptionally similar cells were found to be committed to particular cell types prior to the divergence of their transcriptional profiles [114, 128]. Importantly, such cells can easily be mistaken for multipotent progenitors. Coupling lineage tracing with epigenomics or proteomics measurements might help to avoid some of these biases and pinpoint the correct sequence of transcriptional and epigenetic changes during development. Lineage tracing experiments also seem to indicate that cell fate decisions occur in a more continuous manner rather than abruptly, as previously believed [129]. Finally, lineage tracing made it possible to observe the convergence of differentiation trajectories from distinct developmental origins [128].

1

1.6 Data analysis

Many single-cell methods involve advanced data analysis (see Box 3 and Figure 1.4 for a selection of computational methods). In scRNA-seq data, cell types can in principle be identified by clustering similar transcriptomes [130] and the underlying gene regulatory networks can be inferred [131–133]. However, both cell type identification and network inference are improved by integrating multiple omics data sets [41, 134]. Integration methods typically aim to extract variations common to all measured modalities [135–137]. That is even possible if molecular species are not measured simultaneously in the same cell, as shown, for example, for DNA methylation and transcriptome measurements [138]. Trajec-

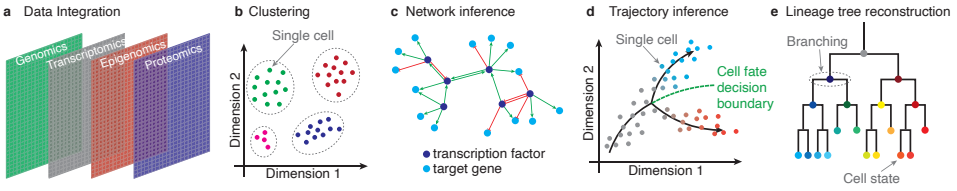


Figure 1.4: Common elements of single-cell omics data analysis. a) Each rectangle represents a data matrix from an omics technology with cells in columns and features in rows. Data Integration methods can be used to combine these data sets. b) Each circle represents a cell projected into a two-dimensional space. Cells that are closer have more similar molecular profiles. Different colors indicate different clusters of cells, which have been determined through a clustering algorithm. Clusters are usually identified with cell types. c) Inferred gene regulatory network. Circles show transcription factors and their target genes. Edges correspond to the interactions between them. d) Each circle represents a cell in a two-dimensional space. Developmental dynamics, indicated by a black line, are revealed with trajectory inference methods. The cell fate decision boundary separates molecular profiles that develop into different cell types. e) Each node represents a cell type. The leaves of the tree (lowest nodes) are the observed final cell types, whereas the other cell types are transient. Lines indicate the lineage relationships.

tory inference algorithms seek to reconstruct gene expression dynamics from scRNA-seq measurements of developing tissues. Many of these methods use similarity of transcriptomes to estimate temporal proximity, which comes with many challenges and limitations [102, 127, 139, 140]: For example, the starting point of a differentiation trajectory has to be provided by the user, because most methods cannot infer directionality. One exception is RNA velocity [141, 142], which exploits RNA splicing dynamics to infer gene expression dynamics and directionality. Time-resolved measurements can be analyzed with optimal transport theory to infer probabilities for the transitions between the observed cell types [14, 106].

Prospective lineage tracing presents a completely different set of challenges for data analysis. The increase in data complexity caused by randomly inserted barcodes, necessitates the development of novel algorithms to infer the underlying phylogenetic tree [127, 143], which captures the hierarchy and relationship of cells during differentiation. As barcoding is often limited to a short period of time, it becomes difficult to infer the lineage tree beyond the point where barcoding has stopped. However, a new method [127] leverages covariances between barcodes to transcend this limitation. An interesting concept in this regard is phylodynamics [125], which studies how the cell type distribution changes over

time, given an observed lineage tree. For example, it has been shown that a model with constant cell division rate, can result in a skewed lineage tree that appears like earlier generations were dividing more rapidly [125].
The algorithms mentioned here are just a small selection of the many tools that have been developed specifically to deal with the challenges arising in single-cell methods. We refer the reader to [144–146] for a much more comprehensive overview.

Box 3: A selection of computational methods for the analysis of single-cell omics data

Clustering methods have been used extensively for transcriptomics data (reviewed in [99]), where they partition cells based on the similarity of their transcriptomes. Clustering is now also applied to a combination of different omics data sets (reviewed in [147–149]). Clusters can be first obtained separately for each modality and then combined, or the different data sets are integrated prior to clustering. Popular examples of integration methods are WNN [41], totalVI [137], MOFA+ [135] and LIGER [138].

Inference of gene regulatory networks has been frequently performed using scRNA-seq data [150]. Examples of existing algorithms are GENIE3 [131], SCENIC [132], SINCERITIES [133] and Scribe [151]. A new method, CellOracle [134], allows the identification of gene regulatory networks from a combination of scRNA-seq and scATAC-seq data. Symphony [152] provides multi-omics clustering as well as gene regulatory network inference. Importantly, these methods often rely on the proper identification of TFs and CREs.

Inference of differentiation trajectories was first introduced for transcriptomics data. These methods make use of the asynchrony during differentiation and order cells by developmental progress (pseudotime). Examples of trajectory inference methods are PAGA [153], DPT [154], Monocle3 [155], FateID [156] and Palantir [157] (reviewed in [103]). A pseudotime method that makes use of spliced and unspliced RNA is RNA velocity [141, 142], which has also been expanded to include protein dynamics [158]. In order to combine several transcriptomics data sets and recreate the differentiation trajectory, optimal transport theory has been applied [14, 106]. A new, interesting method is MATCHER [159], which infers pseudotime based on multi-omics assays.

Reconstruction of lineage trees is the goal of dynamic lineage tracing techniques, where barcodes are introduced randomly during a short period of time. Classic reconstruction methods, like neighbor joining [160] are not robust enough for this purpose. Several studies therefore designed custom made methods [118, 161] and additionally, a new, more robust inference method has been proposed, Cassiopeia [143]. Building on the neighbor joining algorithm, CLiNC [127] tries to discover inconsistencies within the phylogenic tree.

1

1.7 Conceptual framework

Coupled differential equations: Differential equations describe the temporal evolution of a system. They are coupled, if variables appear in several equations. Such equations can have multiple stable solutions, which do not evolve in time, unless perturbed.
Critical point: A point in parameter space where the number or stability of solutions to a dynamical system change abruptly.
Intrinsically disordered regions: Segments of a protein that do not form a stable three-dimensional structure.
Stable states: A solution of a dynamical system that is a local minimum of the corresponding potential landscape.

Even with appropriate data analysis algorithms in place, we still need a conceptual framework for the quantitative understanding of cell types and their formation. The challenge is to reveal, how gene regulatory networks with certain topologies give rise to the observed cell types and molecular dynamics during differentiation. Dynamical systems theory has been used extensively to model gene regulatory networks quantitatively. In this framework, cell types can be understood as **stable states** in a system of **coupled differential equations** [162]. Number, position and robustness of these stable states all depend on parameters that reflect the interactions between TFs and other members of the regulatory network. These parameters can be difficult to infer from experiments, except for (unrealistically) small networks. Nevertheless, dynamical systems describe key properties of the differentiation process. They explain how the interactions between several TFs jointly give rise to cell types that are robust up to a certain level of perturbation [163]. They also explain how a change in TF interactions causes cell types to destabilize [164]. Finally, unstable, intermediate cell states can be found, depending on the parameters of the system [165, 166].

A dynamical systems model can be represented by a potential energy landscape, where a cell follows the path of steepest decent into locally stable states, that correspond to cell types [166–168], see Figure 1.1, middle right. This potential energy landscape is closely related to Waddington's epigenetic landscape [169], a pioneering metaphor that abstracted from molecular details to conceptualize embryonic development. Importantly, the shape of Waddington's landscape is constant in time and a location in the landscape corresponds to the complete molecular profile of a cell. By contrast, most dynamical systems models identify the state of a cell by its transcriptome or even just the expression levels of the TFs in a gene regulatory network (see Figure 1.1, middle left). The shape of the potential landscape is then defined by the gene regulatory network, most importantly the interactions between TFs and their target genes [162]. Changes in the epigenetic state and other gene regulatory molecules can modulate the strength of those interactions (i.e. the parameters of the dynamic system) and thereby cause different stable and unstable states to appear or disappear [166, 167]. Defining the gene expression profile as the state of the cell and modeling the epigenetic profile as parameters of the gene regulatory network has certain conceptual advantages. For example, at **critical points**, which have been studied extensively by catastrophe theory, small changes of the parameters can cause large changes in the stable states of a dynamical system [162, 166]. Lineage decisions might thus be driven by dynamic epigenetic changes around critical points. Importantly, Waddington's landscape implies a strict hierarchy of differentiation, leading from multipotent to more and more specified, unipotent states (see Figure 1.1, right).

Despite its many advantages, the landscape model also has clear drawbacks, including its inability to describe periodic trajectories, e.g. caused by the cell cycle [102, 167]. Therefore,

many other ways to conceptualize differentiation have been devised. For example, the spin glass, a model that originated in physics, describes a system of interacting particles that can have stable low energy states corresponding to different cell types [162, 170]. It accommodates different strengths of interactions between TFs, can describe symmetry breaking events and is scalable to larger numbers of TFs. However, it is often simplified by the usage of binary TF expression (on/off) and symmetric interactions for mathematical tractability. The concepts discussed here are just a few examples of the many models that are currently being developed. More comprehensive overviews can be found in [162, 171].

1.8 Perspectives

- To discover the molecular underpinnings of cell types and their formation is of fundamental interest in developmental and stem cell biology. It is equally important for the understanding of diseases such as cancer, where cell types lose their stability and are transformed to malignant states.
- New single-cell measurement techniques have given us unprecedented insights into the interactions and dynamics of the relevant molecular agents. In the current paradigm, transcription factors, regulatory DNA elements and other classes of molecules form a regulatory network from which cell types emerge.
- In the future, lineage tracing and other quantitative methods will be leveraged to reveal the complete lineage tree and infer a predictive mathematical model of the underlying gene regulatory network. Such a model would allow us to manipulate cell types at will, which has numerous medical applications.

Author Contributions
S.S. and M.M. wrote the manuscript. M.M. created the figures.

Competing Interests
The authors declare no competing interests.

Funding
M.M and S.S. were supported by the Netherlands Organisation for Scientific Research (NWO/OCW, www.nwo.nl), as part of the Frontiers of Nanoscience (NanoFront) program. We acknowledge funding by an NWO/OCW Vidi grant (016.Vidi.189.007) for S.S.

1

1.9 Outline of Thesis

Despite being the object of intense study, embryonic development has been difficult to model for several reasons. One primary reason is that complex tissues can comprise many cell types, of which we probably only know a subset. Thus, we first focus on discovering cell types with single-cell RNA sequencing (scRNA-seq), which has proven very successful. Throughout this thesis, we will use scRNA-seq to uncover different cell types and transcriptional changes during cellular differentiation.
Additionally, many signaling processes and morphogenic events co-occur in a developing tissue, making it hard to isolate the cell's individual contributions. For this purpose, we look at stem cell-derived in vitro systems, in which a small number of specific cell types can be studied in isolation. In this thesis, we mainly focus on stem-cell-derived endothelial cells, which line the inside of vessels, and gastruloids which mimic the process of gastrulation. We use scRNA-seq to investigate changes in gene expression patterns of master transcription factors and their target genes. Cellular differentiation, particularly, depends on extensive communication between cells and must lead to the formation of non-trivial spatial patterns in a robust and reproducible way. For this purpose, we analyze different in vitro model systems showing that cellular communication causes morphogenic and transcriptional changes in the developing cells. Specifically, we show in one chapter that cellular communication can overrule the developmental origin and in another that it can cause the formation of epithelial structures.
Lastly, we want to use measurements of developmental processes to reveal the underlying regulatory mechanisms. To that end, we use a neural network to infer the parameters of a model for gene regulation and cellular communication.

In **Chapter 2** we develop a new method to assess clusterability in single-cell transcriptomics data. Single-cell transcriptomics data has revolutionized biology by its ability to find new cell phenotypes through clustering. However, all clustering methods have adjustable parameters, making it challenging to find the correct number of clusters. There was no principled method to decide whether a cluster of cells contains meaningful sub-populations that can be further resolved. We developed phiclust (ϕ_{clust}) a clusterability measure derived from random matrix theory and low-rank perturbation theory that can identify the presence of non-random substructures. We showed that, by using this method, we could identify previously overlooked subtypes.

In **Chapter 3** we analyze the co-differentiation of endothelial cells and cardiomyocytes from human-induced pluripotent stem cells (hiPSCs). Both cell types are building blocks of the heart: Cardiomyocytes generate contractile forces, and endothelial cells line the inside of blood vessels. During embryonic development, these two cell types arise from a common hematoendothelial lineage. A known master regulator of this specification in mice is ETV2. We used scRNA-seq and a reporter cell line to uncover the role of ETV2 in the differentiation from human iPSCs. We showed that a transient expression of an ETV2-high state initiates the specification of endothelial cells. Functional cardiomyocytes can arise from cells that do not express ETV2 or, surprisingly, have sub-threshold expression.

In **Chapter 4** we investigate the importance of cellular communication compared to developmental origin in a 3D model of endothelial cell differentiation from hiPSCs. Endothelial cells appear across different organs where, on the one hand, their common function is to create a barrier between a liquid, such as blood, and the surrounding tissue. On the other hand, each vessel has specific requirements for each environment and liquid. We lacked an understanding of how endothelial cells acquire their organ-specific function. We investigated the hypothesis that developmental origin plays a role in the specification. For this purpose, endothelial cells were differentiated from paraxial or cardiac mesoderm. We showed that upon integration into a 3D microtissue, including cardiomyocytes and fibroblasts, the transcriptomic signature of the developmental origin is partially removed. This finding suggests that environmental cues might be more critical in function specification than developmental origin.

In **Chapter 5** we characterize the effect of cellular communication between embryonic and extra-embryonic cell types in an enhanced gastruloid system. In particular, we combined in vitro mouse embryonic stem cells (mESCs), used in regular gastruloid protocols, with extraembryonic-endoderm (XEN) cells. With this addition, we observed the formation of a neural epithelium absent in gastruloids derived only from mESCs. We characterized the neural epithelia with scRNA-seq, imaging, and differentiation protocols and observed similarities with the formation of a neural tube with dorsal characteristics. The XEN cells influence morphogenic changes in mESCs, and we also showed that mESCs induce differentiation in the XEN cells. We analyzed candidates of signaling pathways between these cell types and found that local inhibition of WNT signaling is one of the processes.

In **Chapter 6** we explore the use of physics-informed neural networks (PINNs) for the inference of gene regulatory networks (GRNs). GRNs regulate, among many other things, the robust differentiation of a progenitor into specified cell types. Existing methods to infer GRNs mostly use correlation or other similarity-based metrics which limits the predictive power of the resulting model. In order to incorporate prior biological knowledge and thereby make the inference of mechanistic relationships feasible, we chose to use PINNs. We analyze two relevant experimental scenarios in detail: In the first scenario single-cell trajectories are available and cells communicate with each other. In the other scenario, cells do not communicate and the provided data is only a snapshot, which corresponds to a destructive single-cell RNA-sequencing measurement. In both cases, a PINN is used to infer the strengths of gene interactions. We show the benefits of using PINNs compared to regular feedforward NNs in this context and determine the performance level of PINNs for different experimental designs. This analysis will serve as the starting point for exploring the great potential of PINNs for GRN inference.

1

References

[1] M. Mircea and S. Semrau. How a cell decides its own fate: a single-cell view of molecular mechanisms and dynamics of cell-type specification. *Biochemical Society Transactions*, dec 2021.

[2] S. A. Morris. The evolving concept of cell identity in the single cell era. *Development (Cambridge)*, 146(12), jun 2019.

[3] D. Arendt et al. The origin and evolution of cell types, dec 2016.

[4] A. F. Savulescu et al. Pinpointing Cell Identity in Time and Space. *Frontiers in Molecular Biosciences*, 7:209, aug 2020.

[5] B. Xia and I. Yanai. A periodic table of cell types. *Development (Cambridge)*, 146(12), jun 2019.

[6] N. Almeida et al. Employing core regulatory circuits to define cell identity. *The EMBO Journal*, 40(10):e106785, may 2021.

[7] C. Dupont, D. R. Armant, and C. A. Brenner. Epigenetics: Definition, Mechanisms and Clinical Perspective. *Seminars in reproductive medicine*, 27(5):351, sep 2009.

[8] C. D. Allis and T. Jenuwein. The molecular hallmarks of epigenetic control, aug 2016.

[9] R. Paro, U. Grossniklaus, R. Santoro, and A. Wutz. *Introduction to Epigenetics*. Learning Materials in Biosciences. Springer International Publishing, Cham, 2021.

[10] E. Armingol, A. Officer, O. Harismendy, and N. E. Lewis. Deciphering cell–cell interactions and communication from gene expression, feb 2021.

[11] A. A. Almet, Z. Cang, S. Jin, and Q. Nie. The landscape of cell–cell communication through single-cell transcriptomics. *Current Opinion in Systems Biology*, 26:12–23, jun 2021.

[12] N. Eling, M. D. Morgan, and J. C. Marioni. Challenges in measuring and understanding biological noise. *Nature Reviews Genetics*, 20(9):536–548, may 2019.

[13] A. Guillemin and M. P. Stumpf. Noise and the molecular processes underlying cell fate decision-making. *Physical Biology*, 18(1):11002, jan 2021.

[14] M. Mittnenzweig et al. A single-embryo, single-cell time-resolved model for mouse gastrulation. *Cell*, 184(0):1–18, apr 2021.

[15] R. Stadhouders, G. J. Filion, and T. Graf. Transcription factors and 3D genome conformation in cell-fate decisions, may 2019.

[16] K. S. Zaret. Pioneer Transcription Factors Initiating Gene Network Changes, nov 2020.

[17] S. L. Klemm, Z. Shipony, and W. J. Greenleaf. Chromatin accessibility and the regulatory epigenome, apr 2019.

[18] Y. Stelzer et al. Tracing Dynamic Changes of DNA Methylation at Single-Cell Resolution. *Cell*, 163(1):218–229, sep 2015.

[19] D. Song, D. Yang, C. A. Powell, and X. Wang. Cell–cell communication: old mystery and new opportunity, apr 2019.

[20] A. D. Riggs, R. A. Martienssen, and V. E. Russo. *Epigenetic Mechanisms of Gene Regulation.* Cold Spring Harbor Laboratory Press, 1996.

[21] C. H. Ludwig and L. Bintu. Mapping chromatin modifications at the single cell level. *Development (Cambridge)*, 146(12), jun 2019.

[22] Z. Cang et al. A multiscale model via single-cell transcriptomics reveals robust patterning mechanisms during early mammalian embryo development. *PLOS Computational Biology*, 17(3):e1008571, mar 2021.

[23] S. Bessonnard et al. Gata6, Nanog and Erk signaling control cell fate in the inner cell mass through a tristable regulatory network. *Development (Cambridge)*, 141(19):3637–3648, oct 2014.

[24] N. Schrode, N. Saiz, S. D. Talia, and A.-K. Hadjantonakis. GATA6 Levels Modulate Primitive Endoderm Cell Fate Choice and Timing in the Mouse Blastocyst. *Developmental Cell*, 29(4):454–467, may 2014.

[25] T. Graf and T. Enver. Forcing cells to change lineages, dec 2009.

[26] D. Hnisz et al. Super-Enhancers in the Control of Cell Identity and Disease. *Cell*, 155(4):934, nov 2013.

[27] W. A. Whyte et al. Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell*, 153(2):307–319, apr 2013.

[28] S. V. Razin and A. A. Gavrilov. The Role of Liquid–Liquid Phase Separation in the Compartmentalization of Cell Nucleus and Spatial Genome Organization, jun 2020.

[29] A. S. Lyon, W. B. Peeples, and M. K. Rosen. A framework for understanding the functions of biomolecular condensates across scales, mar 2021.

[30] S. F. Banani, H. O. Lee, A. A. Hyman, and M. K. Rosen. Biomolecular condensates: Organizers of cellular biochemistry, may 2017.

[31] E. Gomes and J. Shorter. The molecular language of membraneless organelles, may 2019.

[32] M. Grosch, S. Ittermann, D. Shaposhnikov, and M. Drukker. Chromatin-Associated Membraneless Organelles in Regulation of Cellular Differentiation, dec 2020.

[33] B. R. Sabari et al. Coactivator condensation at super-enhancers links phase separation and gene control. *Science*, 361(6400):eaar3958, jul 2018.

[34] M. Esposito et al. TGF-β-induced DACT1 biomolecular condensates repress Wnt signalling to promote bone metastasis. *Nature Cell Biology*, 23(3):257–267, mar 2021.

[35] Y. E. Guo et al. Pol II phosphorylation regulates a switch between transcriptional and splicing condensates. *Nature*, 572(7770):543–548, aug 2019.

[36] W. K. Cho et al. Mediator and RNA polymerase II clusters associate in transcription-dependent condensates. *Science*, 361(6400):412–415, jul 2018.

[37] M. Hochane et al. Single-cell transcriptomics reveals gene expression dynamics of human fetal kidney development. *PLOS Biology*, 17(2):e3000152, feb 2019.

[38] Y. Ando, A. T. J. Kwon, and J. W. Shin. An era of single-cell genomics consortia, sep 2020.

[39] A. Regev et al. The human cell atlas. *eLife*, 6, dec 2017.

[40] N. Schaum et al. Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature*, 562(7727):367–372, oct 2018.

[41] Y. Hao et al. Integrated analysis of multimodal single-cell data. *Cell*, 0(0), may 2021.

[42] S. C. Bendall et al. Single-Cell Mass Cytometry of Differential Immune and Drug Responses Across a Human Hematopoietic Continuum. *Science*, 332(6030):687–696, may 2011.

[43] A. Tanay and A. Regev. Scaling single-cell genomics from phenomenology to mechanism, jan 2017.

[44] J. D. Buenrostro et al. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature*, 523(7561):486–490, jul 2015.

[45] D. A. Cusanovich et al. Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science*, 348(6237):910–914, may 2015.

[46] J. Cao et al. Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science*, 361(6409):1380–1385, sep 2018.

[47] S. Picelli et al. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nature Methods*, 10(11):1096–1098, sep 2013.

[48] E. Z. Macosko et al. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell*, 161(5):1202–1214, may 2015.

[49] T. Hashimshony et al. CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. *Genome Biology*, 17(1):1–7, apr 2016.

[50] J. Cao et al. Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science*, 357(6352):661–667, aug 2017.

[51] B. Budnik, E. Levy, G. Harmange, and N. Slavov. SCoPE-MS: mass spectrometry of single mammalian cells quantifies proteome heterogeneity during cell differentiation. *Genome Biology*, 19(1):1–12, oct 2018.

[52] H. Specht et al. Single-cell proteomic and transcriptomic analysis of macrophage heterogeneity using SCoPE2. *Genome Biology*, 22(1):1–27, jan 2021.

[53] E. M. Schoof et al. Quantitative single-cell proteomics as a tool to characterize cellular hierarchies. *Nature Communications*, 12(1):1–15, jun 2021.

[54] M. Stoeckius et al. Simultaneous epitope and transcriptome measurement in single cells. *Nature Methods*, 14(9):865–868, aug 2017.

[55] V. M. Peterson et al. Multiplexed quantification of proteins and transcripts in single cells. *Nature Biotechnology*, 35(10):936–939, aug 2017.

[56] C. Zhu et al. An ultra high-throughput method for single-cell joint analysis of open chromatin and transcriptome. *Nature Structural and Molecular Biology*, 26(11):1063–1070, nov 2019.

[57] S. A. Smallwood et al. Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nature Methods*, 11(8):817–820, jul 2014.

[58] M. Farlik, N. C. Sheffield, J. Klughammer, and C. Bock Correspondence. Single-Cell DNA Methylome Sequencing and Bioinformatic Inference of Epigenomic Cell-State Dynamics. *CellReports*, 10:1386–1397, 2015.

[59] H. Guo et al. Single-cell methylome landscapes of mouse embryonic stem cells and early embryos analyzed using reduced representation bisulfite sequencing. *Genome Research*, 23(12):2126–2135, dec 2013.

[60] D. Mooijman et al. Single-cell 5hmC sequencing reveals chromosome-wide cell-to-cell variability and enables lineage reconstruction. *Nature Biotechnology*, 34(8):852–856, jan 2016.

[61] C. Angermueller et al. Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nature Methods*, 13(3):229–232, feb 2016.

[62] Y. Hu et al. Simultaneous profiling of transcriptome and DNA methylome from a single cell. *Genome Biology*, 17(1):1–11, may 2016.

[63] S. J. Clark et al. scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells. *Nature Communications*, 9(1):1–9, feb 2018.

[64] A. Rotem et al. Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. *Nature Biotechnology*, 33(11):1165–1172, nov 2015.

[65] H. S. Kaya-Okur et al. CUT&Tag for efficient epigenomic profiling of small samples and single cells. *Nature Communications*, 10(1):1–10, dec 2019.

[66] H. Xiong et al. Single-cell joint detection of chromatin occupancy and transcriptome enables higher-dimensional epigenomic reconstructions. *Nature Methods*, pages 1–9, may 2021.

[67] C. Zhu et al. Joint profiling of histone modifications and transcriptome in single cells from mouse brain. *Nature Methods*, 18(3):283–292, mar 2021.

[68] V. Ramani et al. Massively multiplex single-cell Hi-C. *Nature Methods*, 14(3):263–266, jan 2017.

[69] T. Nagano et al. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature*, 502(7469):59–64, sep 2013.

[70] D.-S. Lee et al. Simultaneous profiling of 3D genome structure and DNA methylation in single human cells. *Nature Methods*, 16(10):999–1006, sep 2019.

[71] G. Li et al. Joint profiling of DNA methylation and chromatin architecture in single cells. *Nature Methods*, 16(10):991–993, aug 2019.

[72] C. Zong, S. Lu, A. R. Chapman, and X. S. Xie. Genome-Wide Detection of Single-Nucleotide and Copy-Number Variations of a Single Human Cell. *Science*, 338(6114):1622–1626, dec 2012.

[73] F. Abascal et al. Somatic mutation landscapes at single-molecule resolution. *Nature*, 593(7859):405–410, may 2021.

[74] I. C. Macaulay et al. G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. *Nature Methods*, 12(6):519–522, apr 2015.

[75] A. Rodriguez-Meira et al. Unravelling Intratumoral Heterogeneity through High-Sensitivity Single-Cell Mutational Analysis and Parallel RNA Sequencing. *Molecular Cell*, 73(6):1292–1305.e8, mar 2019.

[76] S. V. Ulianov et al. Suppression of liquid–liquid phase separation by 1,6-hexanediol partially compromises the 3D genome organization in living cells. *Nucleic Acids Research*, 49(18):10524–10541, oct 2021.

[77] H. Belaghzal et al. Liquid chromatin Hi-C characterizes compartment-dependent chromatin interaction dynamics. *Nature Genetics*, 53(3):367–378, feb 2021.

[78] K. Holler et al. Spatio-temporal mRNA tracking in the early zebrafish embryo. *Nature Communications*, 12(1):3358, dec 2021.

[79] C.-H. L. Eng et al. Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+. *Nature*, 568(7751):235–239, mar 2019.

[80] E. Lundberg and G. H. Borner. Spatial proteomics: a powerful discovery tool for cell biology, may 2019.

[81] E. Shema, B. E. Bernstein, and J. D. Buenrostro. Single-cell and single-molecule epigenomics to uncover genome regulation at unprecedented resolution, jan 2019.

[82] J. Lee, D. Y. Hyeon, and D. Hwang. Single-cell multiomics: technologies and data analysis methods, sep 2020.

[83] A. Mayran et al. Pioneer factor Pax7 deploys a stable enhancer repertoire for specification of cell fate. *Nature Genetics*, 50(2):259–269, jan 2018.

[84] S. R. Joseph et al. Competition between histone and transcription factor binding regulates the onset of transcription in zebrafish embryos. *eLife*, 6, apr 2017.

[85] C. R. Clapier and B. R. Cairns. The Biology of Chromatin Remodeling Complexes. *Annual Review Biochem*, 78:273–304, jun 2009.

[86] M. Tyagi, N. Imam, K. Verma, and A. K. Patel. Chromatin remodelers: We are the drivers!! *Nucleus*, 7(4):388, jul 2016.

[87] V. Azuara et al. Chromatin signatures of pluripotent cell lines. *Nature Cell Biology*, 8(5):532–538, mar 2006.

[88] B. E. Bernstein et al. A Bivalent Chromatin Structure Marks Key Developmental Genes in Embryonic Stem Cells. *Cell*, 125(2):315–326, apr 2006.

[89] E. Blanco et al. The Bivalent Genome: Characterization, Structure, and Regulation. *Trends in Genetics*, 36(2):118–131, feb 2020.

[90] R. Margueron et al. Role of the polycomb protein EED in the propagation of repressive histone marks. *Nature*, 461(7265):762–767, sep 2009.

[91] L. Di Croce and K. Helin. Transcriptional regulation by Polycomb group proteins. *Nature Structural and Molecular Biology*, 20(10):1147–1155, oct 2013.

[92] S. Kundu et al. Polycomb Repressive Complex 1 Generates Discrete Compacted Domains that Change during Differentiation. *Molecular Cell*, 65(3):432–446.e5, feb 2017.

[93] S. Cruz-Molina et al. PRC2 Facilitates the Regulatory Topology Required for Poised Enhancer Function during Pluripotent Stem Cell Differentiation. *Cell Stem Cell*, 20(5):689–705.e9, may 2017.

[94] Y. Atlasi and H. G. Stunnenberg. The interplay of epigenetic marks during stem cell differentiation and development. *Nature Reviews Genetics*, 18(11):643–658, aug 2017.

[95] K. Daneshvar et al. lncRNA DIGIT and BRD3 protein form phase-separated condensates to regulate endoderm differentiation. *Nature Cell Biology*, 22(10):1211–1222, oct 2020.

[96] M. Modic et al. Cross-Regulation between TDP-43 and Paraspeckles Promotes Pluripotency-Differentiation Transition. *Molecular Cell*, 74(5):951–965.e13, jun 2019.

[97] M. Grosch et al. Nucleus size and DNA accessibility are linked to the regulation of paraspeckle formation in cellular differentiation. *BMC Biology*, 18(1):1–19, apr 2020.

[98] G. J. Knott, C. S. Bond, and A. H. Fox. The DBHS proteins SFPQ, NONO and PSPC1: a multipurpose molecular scaffold. *Nucleic Acids Research*, 44(9):3989–4004, may 2016.

[99] S. Wang et al. Long noncoding RNA Neat1 modulates myogenesis by recruiting Ezh2. *Cell Death and Disease*, 10(7):1–15, jun 2019.

[100] J. E. Sulston and H. R. Horvitz. Post-embryonic cell lineages of the nematode, Caenorhabditis elegans. *Developmental Biology*, 56(1):110–156, mar 1977.

[101] Sagar and D. Grün. Deciphering Cell Fate Decision by Integrated Single-Cell Sequencing Analysis. *Annual Review of Biomedical Data Science*, 3(1):1–22, jul 2020.

[102] C. Weinreb et al. Fundamental limits on dynamic inference from single-cell snapshots. *Proceedings of the National Academy of Sciences of the United States of America*, 115(10):E2467–E2476, mar 2018.

[103] W. Saelens, R. Cannoodt, H. Todorov, and Y. Saeys. A comparison of single-cell trajectory inference methods. *Nature Biotechnology*, 37(5):547–554, may 2019.

[104] B. Pijuan-Sala et al. A single-cell molecular map of mouse gastrulation and early organogenesis. *Nature*, 566(7745):490–495, feb 2019.

[105] J. Delile et al. Single cell transcriptomics reveals spatial and temporal dynamics of gene expression in the developing mouse spinal cord. *Development*, 146(12):dev173807, jun 2019.

[106] G. Schiebinger et al. Optimal-Transport Analysis of Single-Cell Gene Expression Identifies Developmental Trajectories in Reprogramming. *Cell*, 176(4):928–943.e22, feb 2019.

[107] S. Nowotschin et al. The emergent landscape of the mouse gut endoderm at single-cell resolution. *Nature*, 569(7756):361–367, apr 2019.

[108] C. G. Palii et al. Single-Cell Proteomics Reveal that Quantitative Changes in Co-expressed Lineage-Specific Transcription Factors Determine Cell Fate. *Cell Stem Cell*, 24(5):812–820.e5, may 2019.

[109] A. S. Nam et al. Somatic mutations and cell identity linked by Genotyping of Transcriptomes. *Nature*, 571(7765):355–360, jul 2019.

[110] A. Abyzov and F. M. Vaccarino. Cell Lineage Tracing and Cellular Diversity in Humans. *Annual Review ofGenomics and Human Genetics*, 21:101–116, sep 2020.

[111] J. Xu et al. Single-cell lineage tracing by endogenous mutations enriched in transposase accessible mitochondrial DNA. *eLife*, 8, 2019.

[112] F. Gaiti et al. Epigenetic evolution and lineage histories of chronic lymphocytic leukaemia. *Nature*, 569(7757):576–580, may 2019.

[113] C. J.-Y. Wei and K. Zhang. RETrace: simultaneous retrospective lineage tracing and methylation profiling of single cells. *Genome Research*, 30(4):gr.255851.119, mar 2020.

[114] D. E. Wagner and A. M. Klein. Lineage tracing meets single-cell omics: opportunities and challenges, jul 2020.

[115] C. S. Baron and A. van Oudenaarden. Unravelling cellular relationships during development and regeneration using genetic lineage tracing, dec 2019.

[116] A. Alemany et al. Whole-organism clone tracing using single-cell sequencing. *Nature*, 556(7699):108–112, apr 2018.

[117] B. Raj, J. A. Gagnon, and A. F. Schier. Large-scale reconstruction of cell lineages using single-cell readout of transcriptomes and CRISPR–Cas9 barcodes by scGESTALT. *Nature Protocols*, 13(11):2685–2713, nov 2018.

[118] B. Spanjaard et al. Simultaneous lineage tracing and cell-type identification using CrIsPr-Cas9-induced genetic scars. *Nature Biotechnology*, 36(5):469–473, jun 2018.

[119] B. Carter and K. Zhao. The epigenetic basis of cellular heterogeneity. *Nature Reviews Genetics*, 22(4):235–250, nov 2020.

[120] S. Preissl et al. Single-nucleus analysis of accessible chromatin in developing mouse forebrain reveals cell-type-specific transcriptional regulation. *Nature Neuroscience*, 21(3):432–439, mar 2018.

[121] G. Jia et al. Single cell RNA-seq and ATAC-seq analysis of cardiac progenitor cell transition states and lineage settlement. *Nature Communications*, 9(1):1–17, dec 2018.

[122] A. E. Trevino et al. Chromatin and gene-regulatory dynamics of the developing human cerebral cortex at single-cell resolution, dec 2020.

[123] R. Argelaguet et al. Multi-omics profiling of mouse gastrulation at single-cell resolution. *Nature*, 576(7787):487–491, dec 2019.

[124] A. McKenna and J. A. Gagnon. Recording development with single cell dynamic lineage tracing. *Development (Cambridge)*, 146(12), jun 2019.

[125] T. Stadler, O. G. Pybus, and M. P. Stumpf. Phylodynamics for cell biologists, jan 2021.

[126] C. Weinreb and A. M. Klein. Lineage reconstruction from clonal correlations. *Proceedings of the National Academy of Sciences of the United States of America*, 117(29):17041–17048, jul 2020.

[127] C. Weinreb, A. Rodriguez-Fraticelli, F. D. Camargo, and A. M. Klein. Lineage tracing on transcriptional landscapes links state to fate during differentiation. *Science*, 367(6479), feb 2020.

[128] F. Wagner, Y. Yan, and I. Yanai. K-nearest neighbor smoothing for high-throughput single-cell RNA-Seq data. *bioRxiv*, page 217737, 2018.

[129] A. E. Rodriguez-Fraticelli et al. Clonal analysis of lineage fate in native haematopoiesis. *Nature*, 553(7687):212–216, jan 2018.

[130] M. Krzak et al. Benchmark and Parameter Sensitivity Analysis of Single-Cell RNA Sequencing Clustering Methods. *Frontiers in Genetics*, 10:1253, dec 2019.

[131] V. A. Huynh-Thu, A. Irrthum, L. Wehenkel, and P. Geurts. Inferring regulatory networks from expression data using tree-based methods. *PLoS ONE*, 5(9):12776, 2010.

[132] S. Aibar et al. SCENIC: Single-cell regulatory network inference and clustering. *Nature Methods*, 14(11):1083–1086, oct 2017.

[133] N. Papili Gao, S. M. Ud-Dean, O. Gandrillon, and R. Gunawan. SINCERITIES: Inferring gene regulatory networks from time-stamped single cell transcriptional expression profiles. *Bioinformatics*, 34(2):258–266, jan 2018.

[134] K. Kamimoto, C. M. Hoffmann, and S. A. Morris. CellOracle: Dissecting cell identity via network inference and in silico gene perturbation, feb 2020.

[135] R. Argelaguet et al. MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biology*, 21(1):1–17, may 2020.

[136] A. Butler et al. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology*, 36(5):411–420, apr 2018.

[137] A. Gayoso et al. Joint probabilistic modeling of single-cell multi-omic data with totalVI. *Nature Methods*, 18(3):272–282, feb 2021.

[138] J. D. Welch et al. Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain Cell Identity. *Cell*, 177(7):1873–1887.e17, jun 2019.

[139] S. Tritschler et al. Concepts and limitations for learning developmental trajectories from single cell genomics, jun 2019.

[140] C. Weinreb, S. Wolock, and A. M. Klein. SPRING: A kinetic interface for visualizing high dimensional single-cell expression data. *Bioinformatics*, 34(7):1246–1248, 2018.

[141] V. Bergen et al. Generalizing RNA velocity to transient cell states through dynamical modeling. *Nature Biotechnology*, 38(12):1408–1414, dec 2020.

[142] G. La Manno et al. RNA velocity of single cells. *Nature*, 560(7719):494–498, aug 2018.

[143] M. G. Jones et al. Inference of single-cell phylogenies from lineage tracing data using Cassiopeia. *Genome Biology*, 21(1):1–27, apr 2020.

[144] H. Nguyen et al. A comprehensive survey of regulatory network inference methods using single cell RNA sequencing data. *Briefings in Bioinformatics*, 22(3):1–15, may 2021.

[145] T. S. Andrews, V. Y. Kiselev, D. McCarthy, and M. Hemberg. Tutorial: guidelines for the computational analysis of single-cell RNA sequencing data. *Nature Protocols*, 16(1):1–9, dec 2020.

[146] T. Stuart and R. Satija. Integrative single-cell analysis. *Nature Reviews Genetics*, 20(5):257–272, jan 2019.

[147] N. Rappoport and R. Shamir. Multi-omic and multi-view clustering algorithms: review and cancer benchmark. *Nucleic Acids Research*, 46(20):10546, nov 2018.

[148] M. Colomé-Tatché and F. J. Theis. Statistical single cell multi-omics integration. *Current Opinion in Systems Biology*, 7:54–59, feb 2018.

[149] M. Efremova and S. A. Teichmann. Computational methods for single-cell omics across modalities. *Nature Methods*, 17(1):14–17, jan 2020.

[150] A. Pratapa et al. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nature Methods*, 17(2):147–154, feb 2020.

[151] X. Qiu et al. Inferring Causal Gene Regulatory Networks from Coupled Single-Cell Expression Dynamics Using Scribe. *Cell Systems*, 10(3):265–274.e11, mar 2020.

[152] C. Burdziak, E. Azizi, S. Prabhakaran, and D. Pe'er. A Nonparametric Multi-view Model for Estimating Cell Type-Specific Gene Regulatory Networks. *arXiv*, feb 2019.

[153] F. A. Wolf et al. PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biology*, 20(1):1–9, mar 2019.

[154] L. Haghverdi et al. Diffusion pseudotime robustly reconstructs lineage branching. *Nature Methods 2016 13:10*, 13(10):845–848, aug 2016.

[155] J. Cao et al. The single-cell transcriptional landscape of mammalian organogenesis. *Nature*, 566(7745):496–502, feb 2019.

[156] J. S. Herman, Sagar, and D. Grün. FateID infers cell fate bias in multipotent progenitors from single-cell RNA-seq data. *Nature Methods*, 15(5):379–386, apr 2018.

[157] M. Setty et al. Characterization of cell fate probabilities in single-cell data with Palantir. *Nature Biotechnology*, 37(4):451–460, mar 2019.

[158] G. Gorin, V. Svensson, and L. Pachter. Protein velocity and acceleration from single-cell multiomics experiments. *Genome Biology*, 21(1):39, feb 2020.

[159] J. D. Welch, A. J. Hartemink, and J. F. Prins. MATCHER: manifold alignment reveals correspondence between single cell transcriptome and epigenome dynamics. *Genome Biology*, 18(1):1–19, jul 2017.

[160] N. Saitou and M. Nei. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4(4):406–425, jul 1987.

[161] M. M. Chan et al. Molecular recording of mammalian embryogenesis. *Nature*, 570(7759):77–82, jun 2019.

[162] A. E. Teschendorff and A. P. Feinberg. Statistical mechanics meets single-cell biology. *Nature reviews. Genetics*, pages 1–18, apr 2021.

[163] C. Furusawa and K. Kaneko. A Dynamical-Systems View of Stem Cell Biology. *Science*, 338(6104):215–217, oct 2012.

[164] J. J. Tyson and B. Novak. A Dynamical Paradigm for Molecular Cell Biology. *Trends in Cell Biology*, 30(7):504–515, jul 2020.

[165] A. L. MacLean, T. Hong, and Q. Nie. Exploring intermediate cell states through the lens of single cells, jun 2018.

[166] A. Guillemin and M. P. H. Stumpf. Non-equilibrium statistical physics, transitory epigenetic landscapes, and cell fate decision dynamics. *Mathematical Biosciences and Engineering*, 17(6):7916–7930, nov 2020.

[167] P. Greulich, R. Smith, and B. D. MacArthur. The physics of cell fate. In *Phenotypic Switching*, pages 189–206. Elsevier, jan 2020.

[168] L. Xu and J. Wang. Quantifying Waddington landscapes, paths, and kinetics of cell fate decision making of differentiation/development. In *Phenotypic Switching*, pages 157–187. Elsevier, jan 2020.

[169] C. Waddington. *The strategy of the genes : a discussion of some aspects of theoretical biology / by C.H. Waddington. | Wellcome Collection.* Allen and Unwin, London, 1975.

[170] A. H. Lang, H. Li, J. J. Collins, and P. Mehta. Epigenetic Landscapes Explain Partially Reprogrammed Cells and Identify Key Reprogramming Genes. *PLOS Computational Biology*, 10(8):e1003734, aug 2014.

[171] P. Cahan et al. Computational Stem Cell Biology: Open Questions and Guiding Principles. *Cell Stem Cell*, 28(1):20–32, jan 2021.

2

Phiclust: A clusterability measure for single-cell transcriptomics reveals phenotypic subpopulations

The ability to discover new cell phenotypes by unsupervised clustering of single-cell transcriptomes has revolutionized biology. Currently, there is no principled way to decide whether a cluster of cells contains meaningful subpopulations that should be further resolved. Here we present phiclust (ϕ_{clust}), a clusterability measure derived from random matrix theory, that can be used to identify cell clusters with non-random substructure, testably leading to the discovery of previously overlooked phenotypes.

This chapter is based on **Mircea, M.**, Hochane, M., Fan, X., Chuva de Sousa Lopes, S.M., Garlaschelli, D., Semrau S., Phiclust: a clusterability measure for single-cell transcriptomics reveals phenotypic subpopulations. Genome Biology 23, 18 (2022). https://doi.org/10.1186/s13059-021-02590-x [1]

2.1 Background

Unsupervised clustering methods [2–5] are integral to most single-cell RNA-sequencing (scRNA-seq) analysis pipelines [6], as they can reveal distinct cell phenotypes. Importantly, all existing clustering algorithms have adjustable parameters that have to be chosen carefully to reveal the true biological structure of the data. If the data is over-clustered, many clusters are driven purely by technical noise and do not reflect distinct biological states. If the data is under-clustered, subtly distinct phenotypes might be grouped with others and will thus be overlooked. Furthermore, most analysis pipelines rely on qualitative assessment of clusters based on prior knowledge, which can hinder the discovery of new phenotypes. To assess the quality of a clustering quantitatively and help choose optimal parameters, some measures of clustering quality and clusterability have been proposed [7], most of which are not directly applicable to scRNA-seq data. For example, some existing methods rely on multimodality of the expression matrix, which is not always justified for scRNA-seq data, especially when considering highly dynamic systems. Other methods have input parameters, such as the optimal number of dimensions for dimensionality reduction, that cannot be easily determined. Also, general methods do not explicitly account for uninformative sources of variability, related to cell cycle progression or the stress response, for example, which can be important confounders. In the context of scRNA-seq, one of the most widely used measures is the silhouette coefficient [8]. This measure requires the choice of a distance metric to compute the similarity between cells. Notwithstanding its usefulness, it cannot be excluded that a partition of random noise obtains a high silhouette coefficient, indicating high clustering quality. Other measures based on distance metrics or the fit of probability densities suffer from similar issues and often only provide binary results instead of a quantitative score [9]. A different approach is pursued by ROGUE [10], a recently developed tool to assess clustering quality specifically in scRNA-seq data. ROGUE applies the concept of entropy on a per-gene basis to quantify the mixing of cell types. While a clear improvement over existing methods, ROGUE depends on a challenging step of selecting informative genes to explain the differences between cell types. It also assumes a particular noise distribution and requires the careful choice of an adjustable parameter. Here we present phiclust, a new clusterability measure for scRNA-seq data that addresses some of the shortcomings of existing methods. This measure is based on the angle ϕ between vectors of the noise-free signal and the measured, noisy signal. We consider clusterability to be the theoretically achievable agreement with the unknown ground truth clustering, for a given signal-to-noise ratio. (Below, we will describe in detail how we define "signal" and "noise" in this context.) Importantly, our measure can estimate the level of achievable agreement without knowledge of the ground truth. High clusterability (phiclust close to 1) means that multiple phenotypic subpopulations are present and that clustering algorithms should be able to distinguish them. Low clusterability (phiclust close to 0) means that the noise is too strong for even the best possible clustering algorithm to find any clusters accurately. If phiclust equals 0, the observed variability within a cluster is consistent with random noise. Any subclusters of such a cluster still have a phiclust of 0, which prevents over-clustering of random noise. Instead of assuming a certain noise distribution or relying on a selection of informative genes, our measure can be applied to arbitrary types of random noise and includes all genes in the analysis. This is made

possible by certain universal properties of random matrix theory (RMT) [11], which has been applied successfully in finance [12], physics [13] and recently also scRNA-seq data analysis [14].

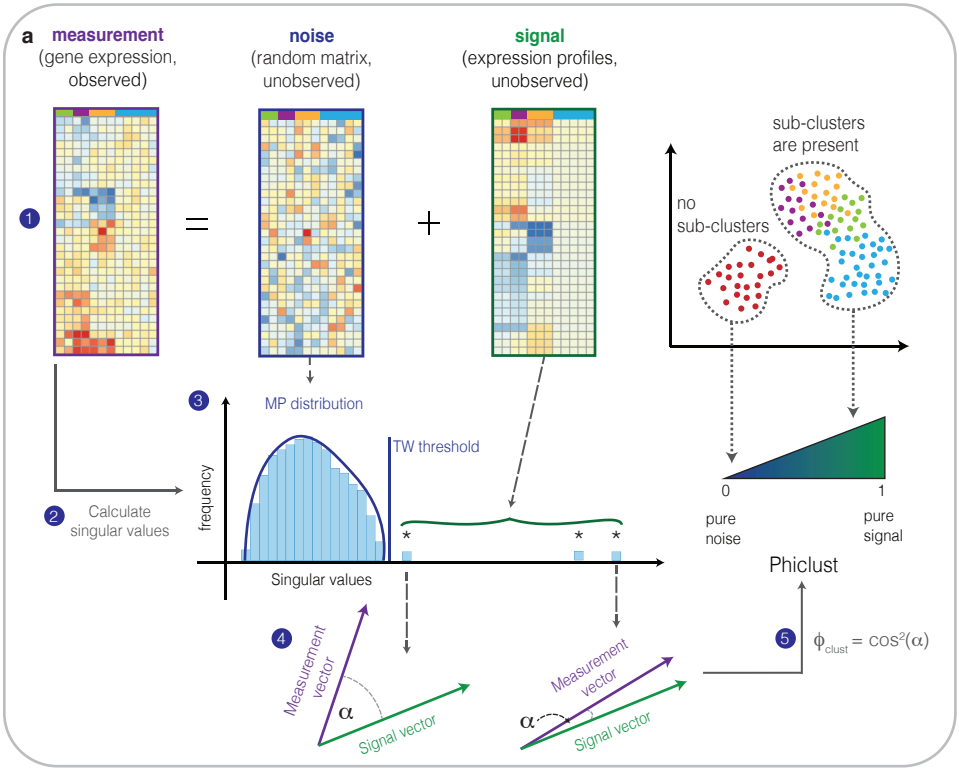


Figure 2.1: Concept of phiclust a Scheme illustrating the rationale behind phiclust.

Below we will use results of RMT on the singular value decomposition (SVD) of a single-cell gene expression matrix, where rows correspond to genes and columns correspond to cells. To get an intuitive understanding of RMT, it is useful to first consider the cell-cell correlation matrix, calculated from the gene expression profiles. We start from the null hypothesis that the data does not contain any structure and is produced by a random process. In the context of single-cell transcriptomics, "structure" means multiple, distinguishable clusters of cells, or phenotypes. RMT can predict, what the correlation matrix looks like, if the entries of the gene expression matrix are samples of random variables that are independent and identically distributed. Trivially, the diagonal elements of this correlation matrix are all equal to 1. The off-diagonal elements are not exactly 0, however, despite the absence of any meaningful structure in the data. Only in the limit of measuring an infinite number of (random) genes would the off-diagonal elements become identically 0 and the correlation matrix would become the identity matrix. In that case, the only eigenvalue of the correlation matrix is 1. RMT describes the properties of a correlation matrix for a finite ratio of cells

and genes. These correlation matrices are, in a sense, distributed "around" the identity matrix, which corresponds to an eigenvalue spectrum distributed around 1. Although the individual entries of the correlation matrix fluctuate from realization to realization, RMT shows that the eigenvalue spectrum is robust (a so-called "self-averaging" property) and an analytical expression for it can be obtained [15]. Likewise, RMT predicts that the singular value distribution of a purely random matrix is closely approximated by the Marchenko-Pastur (MP) distribution. This result holds true irrespective of the distribution of the random variable. This universal property of random matrices allows us to apply RMT to gene expression matrices obtained by scRNA-seq. Of course, any biologically interesting scRNA-seq measurement should contain structure, usually in the form of cell clusters. RMT allows us to regard singular values lying above the MP distribution as evidence for the rejection of the null hypothesis (i.e., the absence of structure in the data). The MP distribution is characterized by sharp upper and lower limits for the singular values of a random matrix, but is strictly valid only in the limit of infinite numbers of genes and cells (while keeping the cell-gene ratio fixed). For finite matrices, the largest and smallest singular values are distributed around those sharp limits, which is described by the Tracy-Widom distribution [16]. As explained above, the presence of structure manifests itself as singular values above the MP distribution (i.e., the prediction for a purely random matrix). Qualitatively, the magnitude of those outlying singular values corresponds to the magnitude of the differences between clusters. We can understand this relationship, if we assume that the measured gene expression matrix is the sum of a random matrix (the "noise") and a matrix of noise-free gene expression profiles (the "signal"), see Fig. 2.1. The bigger the difference in gene expression between phenotypes, the larger the magnitude of the non-zero singular values of the signal matrix. If the number of non-zero singular values (i.e., the rank of the signal matrix) is small compared to the dimensions of the matrix, low-rank perturbation theory [17] is applicable. This theory allows us to calculate the singular values of the measured gene expression matrix from the singular values of the signal matrix. Remarkably, knowledge of the complete signal matrix is not required for this calculation. phiclust is meant to help identify non-random (or deterministic) structure. At the level of a complete data set, for example of a complex tissue, clusters are typically easily discernible. However, if we zoom in on a single cluster, it is much more difficult to decide, whether the variability within that cluster corresponds to meaningful sub-structure (such as the presence of multiple phenotypes) or is consistent with random noise. Below, we will precisely define a notion of clusterability, based on the adjusted rand index, and show that it strongly correlates with phiclust. Furthermore, we will demonstrate that our measure compares favorably to the silhouette coefficient and ROGUE on simulated data and experimental data sets with known ground truth. (See Table S1 for a list of all used simulated and experimental data sets.) Finally, we will apply phiclust to scRNA-seq measurements of complex tissues and obtain new biological insights, which we validate with follow-up measurements.

2.2 Results

2.2.1 Phiclust is derived from first principles and does not have free parameters

To derive phiclust, we considered the measured gene expression matrix as a random matrix perturbed by the unobserved, noise-free gene expression profiles (Fig. 2.1). This is the exact opposite of the conventional approach, which considers random noise as a perturbation to a deterministic signal. Note that, in our approach, the random matrix contains both the biological variability within a phenotype as well as the technical variability (which is due to limited RNA capture and conversion efficiency, for example). Our point of view allows us to leverage well-established results from RMT [14, 18] and perturbation theory [17].

Figure 2.2 illustrates the basic principles that were applied to derive phiclust. RMT predicts that the SVD of a random noise matrix results in normal distributed singular vectors and a distribution of singular values that is closely approximated by the MP distribution, if the matrix is large enough (Fig. 2.2a, left column). Here, we consider the noise-free gene expression profiles of the cells in various phenotypes, as the "signal" that perturbs the random matrix and thus its singular value distribution. Since biological and technical variability are lumped into the random matrix, expression profiles are identical for cells that belong to the same phenotype. For example, in the case of two distinct phenotypes, the signal matrix has only one non-zero singular value (Fig. 2.2a, middle column). The observed (or measured) gene expression matrix is obtained as the sum of the random noise matrix and the noise-free gene expression profiles (2.2a, right column). The singular value distribution of the measured expression matrix has exactly one singular value above the upper limit that the theory predicts for a purely random matrix, the Tracy-Widom (TW) threshold. The outlying singular value and its associated singular vector correspond to the deterministic component of the measured expression matrix. The distribution of the remaining singular values (the "bulk") is still closely approximated by the MP distribution. Importantly, as the perturbation becomes larger, the value of the outlying singular value also increases (Fig. 2.2b). A larger perturbation means more distinct and therefore more easily clusterable phenotypes (compare the singular vectors in the middle row of Figs. 2.2a and b). The basic idea of phiclust is to use the magnitude of the outlying singular values to quantify clusterability.

Due to the universality of RMT, all described principles are independent of the particular noise distribution (see Fig. 2.2a-b for normal distributed noise and Fig. 2.2c-d for Poisson distributed noise). SVD of appropriately preprocessed real data sets therefore leads to singular value distributions with the same shape as obtained in simulations: a bulk closely approximated by the MP distribution and one or multiple outlying values. We found that data preprocessing has to comprise normalization and log-transformation, as well as gene-wise and cell-wise scaling (Fig. 2.3 a-d). SVD of raw data or log-transformed, normalized data typically results in a largest outlying singular value that is much larger than all others (Fig. 2.3 a,b). The corresponding singular vector reflects a global trend in the data and is called "market mode" in the context of stock market analysis [12, 19].

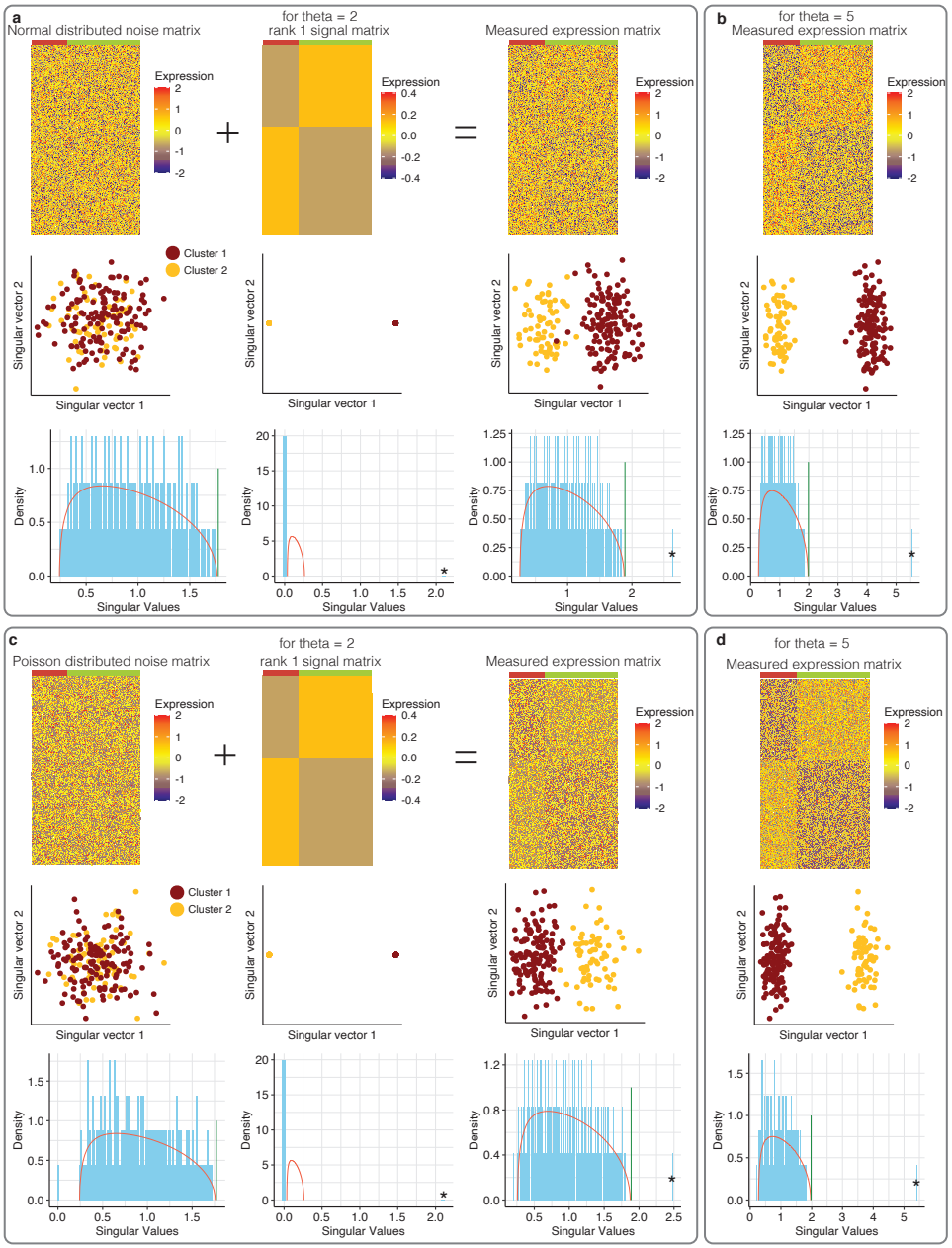


Figure 2.2: Basic principles of random matrix theory and perturbation theory a Top row: heatmaps of a random matrix drawn from a normal distribution, a rank 1 signal matrix with a singular value θ of 2, and the resulting expression matrix. Middle row: Singular vectors of the corresponding matrices. Bottom row: histograms of the corresponding singular values. Red line: MP distribution, green line: TW threshold. b Heatmap, singular vectors and singular values of an expression matrix constructed as in a, except the singular value of the signal matrix was 5. c,d Matrices, singular vectors and singular values obtained as in a and b, but the random matrix was drawn from a Poisson distribution.

Here, we call it "transcriptome mode", since it corresponds to an expression trend that is present across all cells, irrespective of cell type (such as, for example, high expression of particular cytoskeletal genes or essential enzymes and low expression of certain membrane receptors or transcription factors). The transcriptome mode is obviously not informative for clustering. Scaling shifts its singular value to 0, which effectively removes it from further analysis (Fig. 2.3c,d).

We tested for all data sets used in this study, whether the bulk of the singular value distribution of each cluster deviates significantly from the MP distribution after the described preprocessing (Kolmogorov-Smirnov test, Fig. 2.3e). For reasonably large clusters (> 50 cells), we only found one example of a (marginally significant) deviation from the MP distribution.

We next wanted to confirm, for real data, that the remaining outlying singular values reflect the strength of the signal, i.e., differences between the phenotypes. To that end, we extracted the gene expression profiles from two clusters in an experimental single-cell RNA-seq data set and added, as additional signal, a matrix with one non-zero singular value. As to be expected, SVD of the combined data results in one additional singular value, which increases with the strength of the perturbation (Fig. 2.3f-g). See Table S2 for a list of all outlying singular values of experimentally measured expression matrices as well as the corresponding signal matrices. All in all, these tests show that the basic principles of random matrix theory and perturbation theory are applicable to real single-cell RNA-seq data.

So far, we have shown that the values of the outlying singular values are, qualitatively, related to the differences between phenotypes. However, their magnitudes are difficult to interpret. Phiclust is derived from the outlying singular values and can be interpreted as a measure of clusterability, as we will show in the next section. More specifically, phiclust is defined as the squared cosine of the angle between the leading singular vector of the measured gene expression matrix and the corresponding singular vector of the unobserved, noise-free expression matrix. Low-rank perturbation theory is able to predict this angle using only the dimensions of the measured gene expression matrix and its singular values, but without knowledge of the noise-free expression profiles. See Additional File 2 for a detailed derivation. If the noise level is low compared to the signal, this angle will be small, since the measured gene expression matrix is then very similar to the noise-free signal. This would result in phiclust close to 1. As the level of noise increases, for a fixed signal, the singular vectors of the measured expression matrix and the noise-free signal become increasingly orthogonal and phiclust approaches 0. To illustrate the calculation of phiclust, we simulated data sets with realistic noise structure using the Splatter package [20] (Fig. 2.4a,b). As to be expected, increasing the number of genes that are differentially expressed between clusters makes the clusters more easily separable and leads to larger singular values outside of the MP distribution (Fig. 2.4a). By construction, this results in higher values of phiclust (Fig. 2.4b). Please refer to Table S2 for the numerical values of the outlying singular values in the simulated expression matrices as well as the corresponding signal matrix.

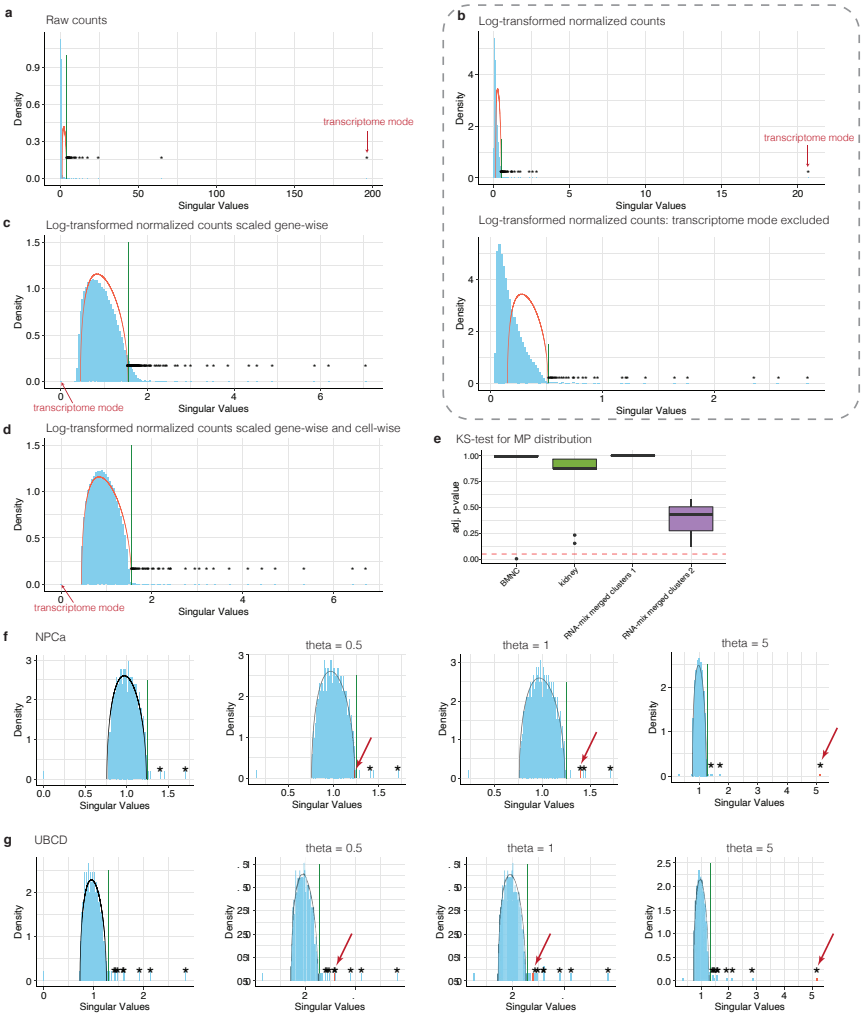


Figure 2.3: Importance of preprocessing for MP fit and effect of perturbation on singular value distribution Singular value (SV) distributions of the fetal kidney single-cell RNA-seq data set after different preprocessing steps. a Raw UMI counts. Arrow indicates transcriptome mode b Log-transformed, normalized UMI counts. Arrow indicates transcriptome mode. Right: Transcriptome mode was excluded. c Log-transformed, normalized data as in b, that was additionally centered gene-wise. The transcriptome mode, visible as the highest singular value in a and b appears close to 0 (indicated by the arrow). d Log-transformed, normalized, and gene-wise standardized data, as in c, that was additionally standardized cell-wise. The SV distribution coincides with the bulk of the MP distribution. This is not a fit: The MP distribution is completely determined by the dimensions of the matrix and has no free parameters. e Kolmogorov-Smirnov (KS) test of a significant difference between the bulk of the singular value distributions and the MP distribution. The boxplot shows the adjusted p-values of the KS test for each cluster per data set. Red dashed line indicates an adjusted p-value of 0.05. f Histogram of singular values for the NPCa cluster of the fetal kidney data set. Left: original values. Rest: Singular values of the NPCa expression matrix plus a rank 1 perturbation with increasing magnitudes of the perturbation (singular value θ of the perturbation = 0.5, 1 or 5). The red arrow indicates the singular value that stems from the additional perturbation. g Histogram of singular values for the UBCD cluster of the fetal kidney data set. Left: original values. Rest: Singular values of UBCD expression matrix plus a rank 1 perturbation with increasing magnitudes of perturbation (singular value θ of the pertrubation = 0.5, 1 or 5). The red arrow indicates the singular value that stems from the additional perturbation.

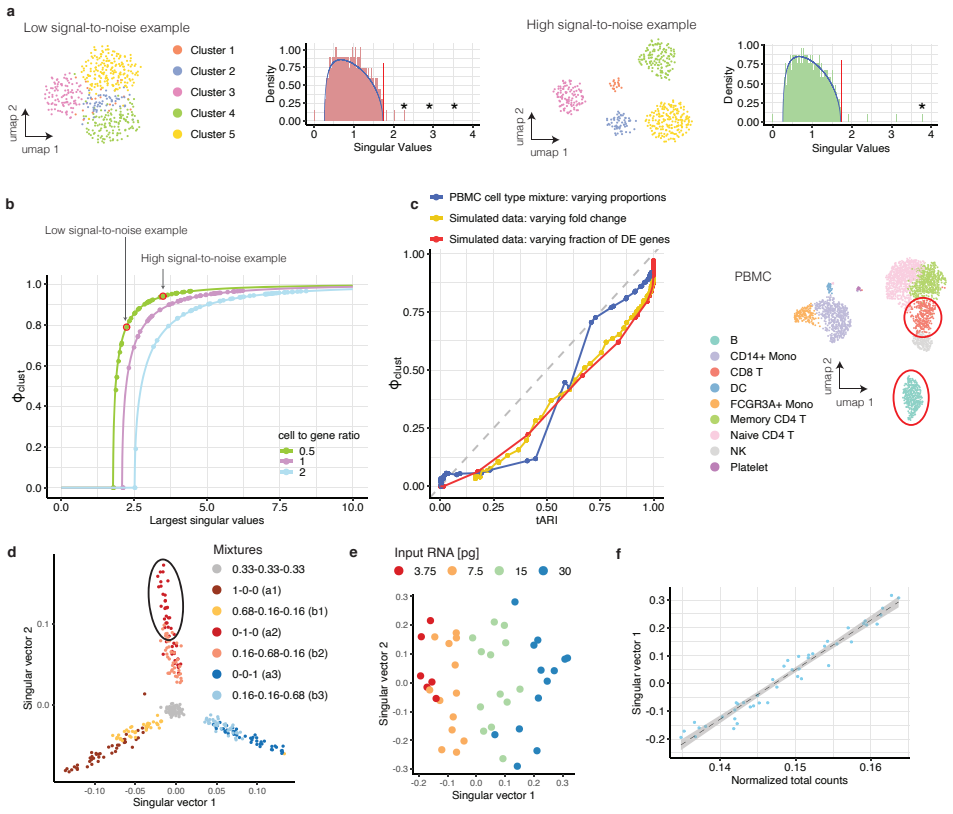


Figure 2.4: Phiclust is a proxy for the theoretically achievable adjusted rand index (tARI). a Singular value distributions of simulated data sets with 5 clusters and different levels of noise; Red: low signal-to-noise, Green: high signal-to-noise. The MP distribution is indicated by a solid blue line, the TW threshold is indicated by a red solid line and significant singular values are highlighted with asterisks. Inserts show UMAPs of the data. The data set with a higher signal-to-noise ratio has more significant singular values and those singular values are bigger. b Value of the largest singular value versus for simulated data. Arrows indicate where the examples from panel b are located. The relationship between the largest singular values and phiclust only depends on the dimensions of the expression matrix. Simulations with different cell-to-gene ratios are shown in different colors. c Phiclust versus theoretically achievable ARI (tARI). Red data points: Simulated data sets with two clusters. The number of differentially expressed (DE) genes was varied, the log fold change between clusters was fixed. Green data points: Simulated data sets with two clusters. The mean log fold change between clusters was varied, the number of differentially expressed genes was fixed. Blue data points: Two synthetic clusters were created by weighted averages of cells from two clusters in the PBMC data set. Cluster weights were varied. The grey dashed line indicates identity. Inset: UMAP of PBMC data set with the two clusters used indicated by red solid circles. d scRNA-seq of mixtures of RNA extracted from three different cell lines. Each data point is a mixture. For each mixture the entries of the first two singular vectors are plotted. Colors indicate different ratios of contributions from the three cell lines. e First two singular vectors of the cluster indicated by a black solid ellipse in f. The amount of mRNA per mixture [pg] is indicated in color. g Normalized total counts per mixture versus first singular vector of the cluster shown in g. Linear regression (dashed line) is used to regress out the correlation with the total counts. Grey area indicates standard deviation.

We would like to stress at this point that phiclust is derived from universal properties of perturbed random matrices, which can be considered first principles. By contrast, many other measures are developed based on empirical observations and justified post hoc by their usefulness. Phiclust is calculated using only the SVD and the dimensions of the expression matrix. Thus, it does not have any free, adjustable parameters, which would have to be chosen by the user or learned from the data.

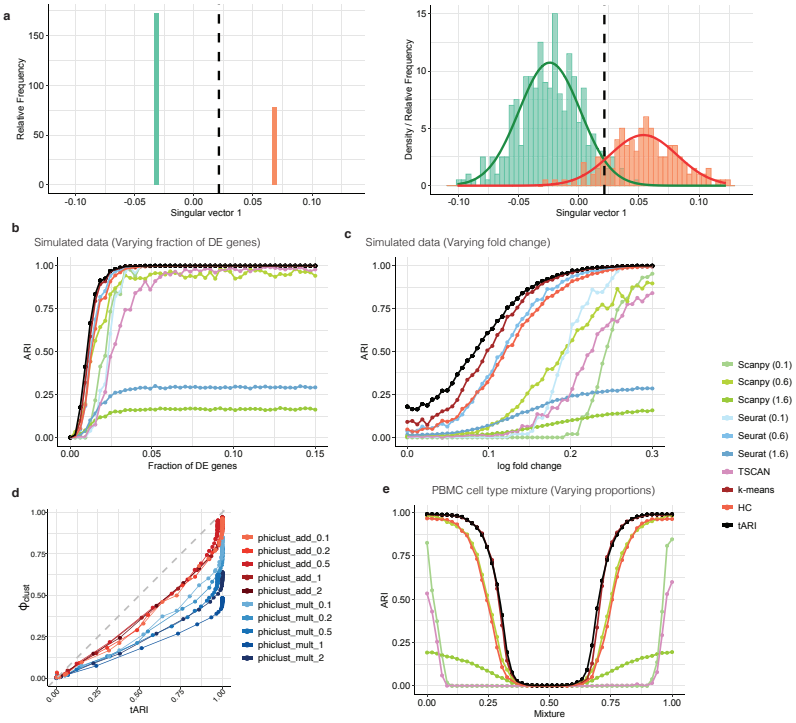


Figure 2.5: An upper limit to the achievable ARI can be estimated using a Bayes classifier. a Left: Histogram of the noise-free singular vector for a scenario with two clusters (or phenotypes) . Only the first singular vector is significant. The dashed line indicates a possible decision boundary. Right: Histogram of the first singular vector in the presence of noise. The color indicates to which simulated (ground truth) cluster the cells belong. Two normal distributions fitted separately to the singular vector entries belonging to the two clusters are shown as solid lines. The Bayesian error rate is estimated from the overlap of these two distributions and used to calculate the theoretical ARI (tARI). The dashed line indicates the optimal decision boundary. b ARI achieved by various clustering methods compared to the ground truth and tARI for simulated data with two clusters. The number of differentially expressed genes was varied. c ARI achieved by various clustering methods compared to the ground truth and tARI for simulated data with two clusters. The mean log fold change between clusters was varied. d tARI versus phiclust for simulated data sets with two clusters and different fractions of DE genes. Red curves: Values of phiclust for additive perturbation at different cell to gene ratios. Blue curves: Values of phiclust for multiplicative perturbation at different cell to gene ratios. Dashed grey line indicates diagonal. e ARI achieved by various clustering methods compared to the ground truth and tARI for PBMC cell type mixtures. Two synthetic clusters were created by weighted averages of cells from two clusters in the PBMC data set (see Fig. 1d). The mixture proportions were varied from 0 to 1. b,c,e The numbers in the legend indicate the resolution parameter used.

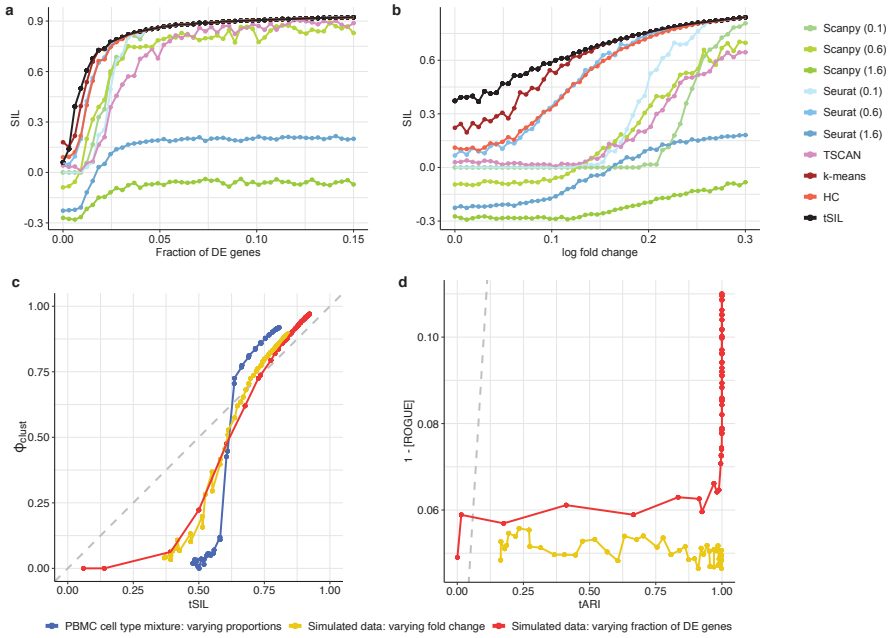


Figure 2.6: An approximate upper limit to the best possible silhouette coefficient and accordance of ROGUE with tARI. a Silhouette coefficient (SIL) achieved by various clustering methods and theoretical SIL (tSIL) for simulated data with two clusters. The number of differentially expressed (DE) genes was varied. b SIL achieved by various clustering methods and tSIL for simulated data with two clusters. The mean log fold change between clusters was varied. a,b The numbers in the legend indicate the resolution parameter used. c tSIL versus phiclust. Red data points: Simulated data sets with two clusters. The number of DE genes was varied, the log fold change between clusters was fixed. Green data points: Simulated data sets with two clusters. The log fold change between clusters was varied, the number of DE genes was fixed. Blue data points: Two synthetic clusters were created by weighted averages of cells from two clusters in the PBMC data set (see Fig. 2.4c). Cluster weights were varied. The Grey dashed line indicates identity. d tARI versus 1 - [ROGUE] score. Red data points: Simulated data sets with two clusters. The number of DE genes was varied, the log fold change between clusters was fixed. Green data points: Simulated data sets with two clusters. The log fold change between clusters was varied, the number of DE genes was fixed. Blue data points: Two synthetic clusters were created by weighted averages of cells from two clusters in a PBMC data set (see Fig. 2.4c). Cluster weights were varied. The Grey dashed line indicates identity.

2.2.2 Phiclust is a proxy for clusterability

To show that phiclust is a proxy for clusterability, we have to make the concept of clusterability more precise and quantifiable. We adopted the Adjusted Rand Index (ARI) [21] as a well-established measure for the agreement between an empirically obtained clustering and the ground truth. Next, we will argue that perfect agreement with the ground truth (ARI = 1) is not achievable in the presence of noise, even with the best conceivable clustering algorithm.

Take, for instance, the simplest possible case of two cell types, A and B. Without any noise (technical or biological), expression profiles within a cell type are identical and the data can be clustered perfectly. Correspondingly, the singular vector of the expression matrix has only two different entries Fig. 2.5a, left). Therefore, it is easy to find a threshold

that discriminates between the two cell types. In the presence of noise, however, there is a chance that the measured expression profile of a cell from cell type A looks more like cell type B and is therefore clustered with other cells from cell type B and vice versa. Correspondingly, the entries of the singular vector are now spread by the noise and can overlap (Fig. 2.5a, right). Even if we use the best possible threshold to discriminate between the two cell types, some cells will be necessarily misclassified, if the distributions overlap. This type of error is unavoidable (or irreducible) and known as Bayes error rate [22] in the context of statistical classification. From the overlap of the singular vector entries, we can calculate the Bayes error rate or, equivalently, a theoretically achievable ARI (tARI, see also Additional File 2). Of course, this is only possible for data with known ground truth. We first used simulated data to show empirically that commonly used clustering methods are not able to exceed the tARI (Fig. 2.5b,c). It therefore quantifies our notion of clusterability: With increased noise, tARI decreases and it is more challenging even for the best conceivable clustering algorithm to distinguish the difference between phenotypes. Importantly, phiclust is strongly correlated with the tARI (Fig. 2.4d) and thus allows us to estimate clusterability without knowing the ground truth.

So far, we have assumed additive noise (i.e., the measured gene expression is the sum of a random matrix and the noise-free expression matrix). Low-rank perturbation theory also makes a prediction for multiplicative noise (i.e., the measured gene expression is the product of a random matrix and the noise-free expression matrix). In that case, phiclust still scales approximately linearly with the tARI, but its dynamic range depends somewhat on the cell-to-gene ratio (Fig. 2.5d). To our knowledge, the noise generating mechanisms at work in scRNA-seq have not been pinpointed comprehensively. Therefore, we will continue to assume additive noise, noting that our approach can be easily adapted to multiplicative noise.

To test the relationship between phiclust and the tARI in experimentally measured data, we used an scRNA-seq data set of peripheral blood mononuclear cells (PBMCs) [23]. We chose two very distinct cell types and created new clusters as weighted, linear combinations of expression profiles from the two cell types. This approach allowed us to precisely control the difference between the newly created clusters, while maintaining the experimentally observed noise structure (Fig. 2.5e). Also for this data, phiclust strongly correlates with the tARI (Fig. 2.4d). As an alternative to the tARI, we also calculated the theoretically achievable silhouette coefficient [8] (tSIL), which considers the distances between the best possible clusters (Fig. 2.6 a-c). For a large range of simulation parameters, the tSIL has a smaller dynamic range than the tARI, which makes it less useful overall for assessing clusterability. In contrast to phiclust, ROGUE [10] does not show collinearity with the tARI (Fig. 2.6d). Therefore, ROGUE seems to implement a notion of clusterability that is distinct from our point of view.

2.2.3 Confounder regression removes unwanted variability

To further characterize the performance of phiclust on experimental data sets with known ground truth, we used a measurement of purified RNA from 3 cell types, mixed at different ratios [24] (Fig. 2.4e). We noticed a significant correlation between the amount of input RNA and the entries of the first singular vector of individual clusters (Fig. 2.4f). This might be explained by lowly expressed genes not being well-represented in the low-input libraries,

and the resulting differences in the expression profiles. In any case, the amount of input RNA seemed to be a confounding factor that could lead to high values of phiclust, even in the absence of meaningful subclusters. Correspondingly, we found a correlation between the singular vector entries and the number of total counts, despite normalization of the data (Fig. 2.4g). This is consistent with the finding that total counts are a confounding factor in scRNA-seq data that cannot be eliminated by normalization using one single scaling factor per cell [23, 25]. Different groups of genes scale differently with the total counts per cell. Therefore, a correlation with the total counts remains even after normalization.

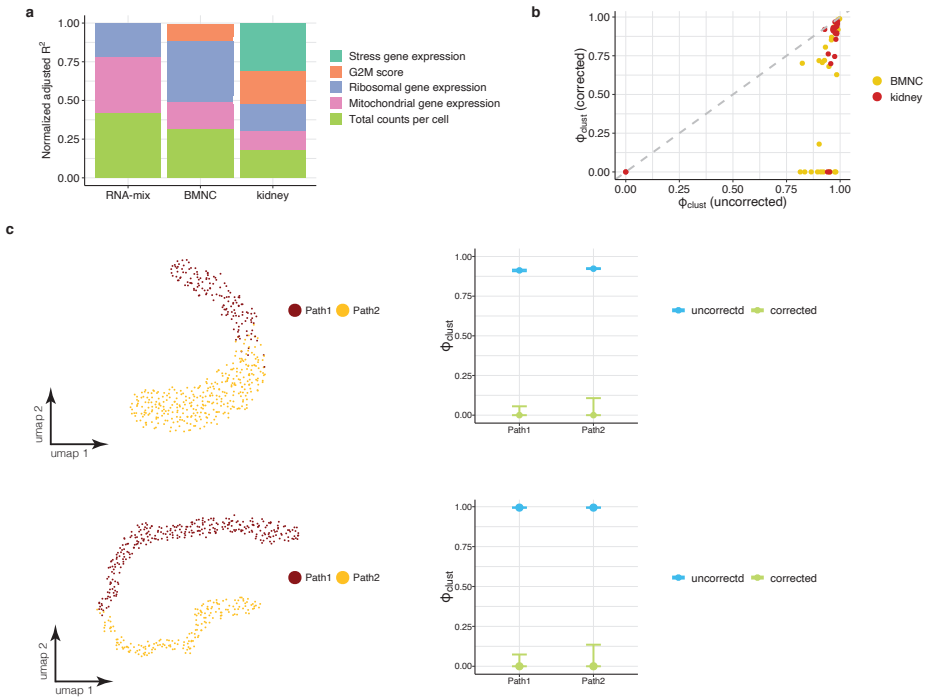


Figure 2.7: Correcting for nuisance parameters and unwanted variability. a Summary of adjusted R2 for several nuisance parameters in all experimental data sets. b Original (uncorrected) phiclust values vs phiclust corrected for the influence nuisance parameters in the BMNC and fetal kidney data sets. Dashed grey line indicates diagonal. c Two examples of differentiation paths with different numbers of differentially expressed genes. Left: UMAPs of simulated data sets with two differentiation paths. Right: Original (uncorrected) values of phiclust and phiclust values corrected by confounder regression using pseudotime as the only confounder.

More generally, there are various experimental and biological factors that drive artefactual or irrelevant variability in single-cell RNA-seq data [23, 26]. We therefore introduced a regression step that removes the influence of any nuisance variables, such as the number of total counts per cell, ribosomal gene expression, mitochondrial gene expression or cell cycle phase (see also Additional File 2). More specifically, we first regress the entries of a singular vector on one or multiple confounders. The fraction of variance explained by all confounder is then given by the adjusted R2 (coefficient of determination) of the linear

regression. Since the squared singular values can also be interpreted as the amount of variance explained, we correct them by multiplying with 1 − the adjusted R2 found in the confounder regression. (See Table S2 for a list of the uncorrected and corrected singular values for both simulated and experimental expression matrices.) The corrected singular values are then used to calculate phiclust.

Interestingly, the relative influence of the confounders considered in this study varied substantially between data sets (Fig. 2.7 a). For example, cell stress is a relevant confounder only in the kidney data set. This is likely related to the cell dissociation procedure, which is necessarily more aggressive for kidney tissue, compared to the other samples: bone marrow mononuclear cells (BMNCs) and purified RNA, extracted from cell cultures. Total counts and ribosomal gene expression explain most of the artefactual variance in BMNCs. This might be explained by high variability in the metabolic state of the cells. In Table S2 we list the R2 values of each considered confounder for each cluster. For real scRNA-seq data sets, confounder regression can lead to a significant reduction of phiclust (Fig. 2.7b, see Table S2 for the numerical values.) It is therefore an important part of the method.

Confounder regression can also help to analyze data sets that are not made up of regular clusters but contain irregularly shaped continua of gene expression. For example, in developmental and stem cell biology we commonly observe differentiation paths, which are large clusters with gradually changing expression profiles. Uncorrected phiclust values are high for such paths, which suggests meaningful subpopulations (2.7c,d). Depending on the biological question, it might in fact be desirable to cluster differentiation paths, for example, to separate a stem cell state from a differentiated cell type. For other applications, it could be preferable to treat a differentiation path as one cluster. In that case we can use pseudotime approaches [27] to infer a temporal order of the gene expression profiles and use the inferred pseudotime in the confounder regression. If all observed variability is explained by developmental dynamics, phiclust is reduced to 0 and thus no sub-clustering is suggested (2.7c,d).

2.2.4 Phiclust has high sensitivity for the detection of substructure

After correction for unwanted variability, we compared the performance of phiclust with other clusterability measures in the RNA mixture data set (Fig. 2.4e). Phiclust successfully indicated the presence or absence of subclusters for all tested combinations of the 7 original mixtures (Fig. 2.8). By contrast, ROGUE only indicated the presence of substructure when the merged clusters were very clearly distinguishable (Fig. 2.8 b,c). The silhouette coefficient was qualitatively similar to phiclust but its dynamic range was much smaller (Fig. 2.8, middle row). This might become critical in the case of highly similar phenotypes, which is precisely where phiclust might have an advantage. An example for this can be seen in Fig. 2.8b: the silhouette coefficients in the pure cluster are very similar to the merged clusters (which were composed of two original clusters). To compare phiclust with the silhouette coefficient in more detail, we carried out additional simulations (Fig. 2.9). First, we simulated 3 clusters and subsequently merged two of them. While phiclust clearly distinguished the merged cluster from the pure cluster, the silhouette coefficients were similar for both. Increasing the fraction of genes that are differentially expressed between the merged cluster increased the difference in silhouette coefficient, but only

gradually (Additional file 1: Fig. S7b). By contrast, phiclust jumped to values close to 1 for the merged cluster for very small fractions of differentially expressed genes (around 0.03). It is therefore the more sensitive measure. The silhouette coefficient strongly depends on the number of principal components used in dimensionality reduction (Fig. 2.9c), as well as the metric for distances between expression profiles (Fig. 2.9d). Phiclust does not depend on such user-defined parameters.

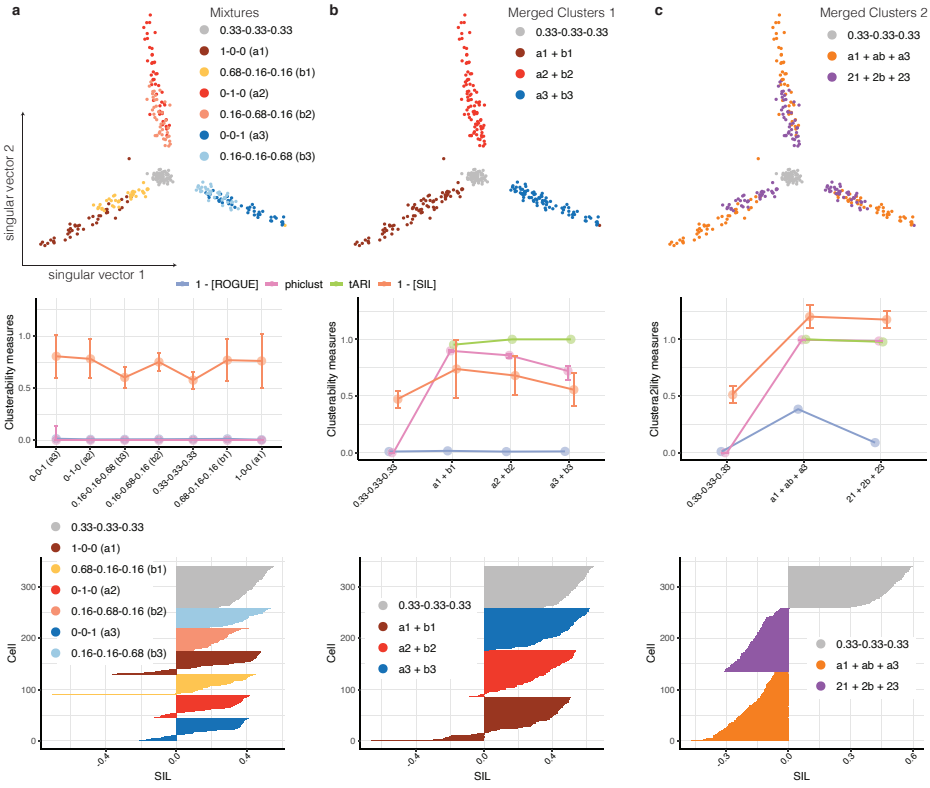


Figure 2.8: Phiclust outperforms other measures on experimental data. Clusters of mixtures of RNA extracted from three different cell lines were merged in different ways to vary the amount of variability in each merged cluster. Top: first two singular vectors of RNA mixture data. Colors indicate different ratios of contributions from the three cell lines. Middle: The values of phiclust (rose), 1 – silhouette coefficient [SIL] (orange), tARI (green) and 1 - [ROGUE] (blue) for each corresponding cluster. For the calculation of the error bars, see Methods. Bottom: Bar plot of silhouette coefficients for each cell, sorted by cluster. a Original RNA mixture. b Merged clusters. Red: 0-1-0 merged with 0.16-0.68-0.16. Blue: 0-0-1 merged with 0.16-0.16-0.68. Green: 1-0-0 merged with 0.68-0.16-0.16. c Violet: merged cluster contains mixtures 0.68-0.16-0.16, 0.16-0.68-0.16 and 0.16-0.16-0.68. Orange: merged cluster contains mixtures 1-0-0, 0-1-0, and 0-0-1.

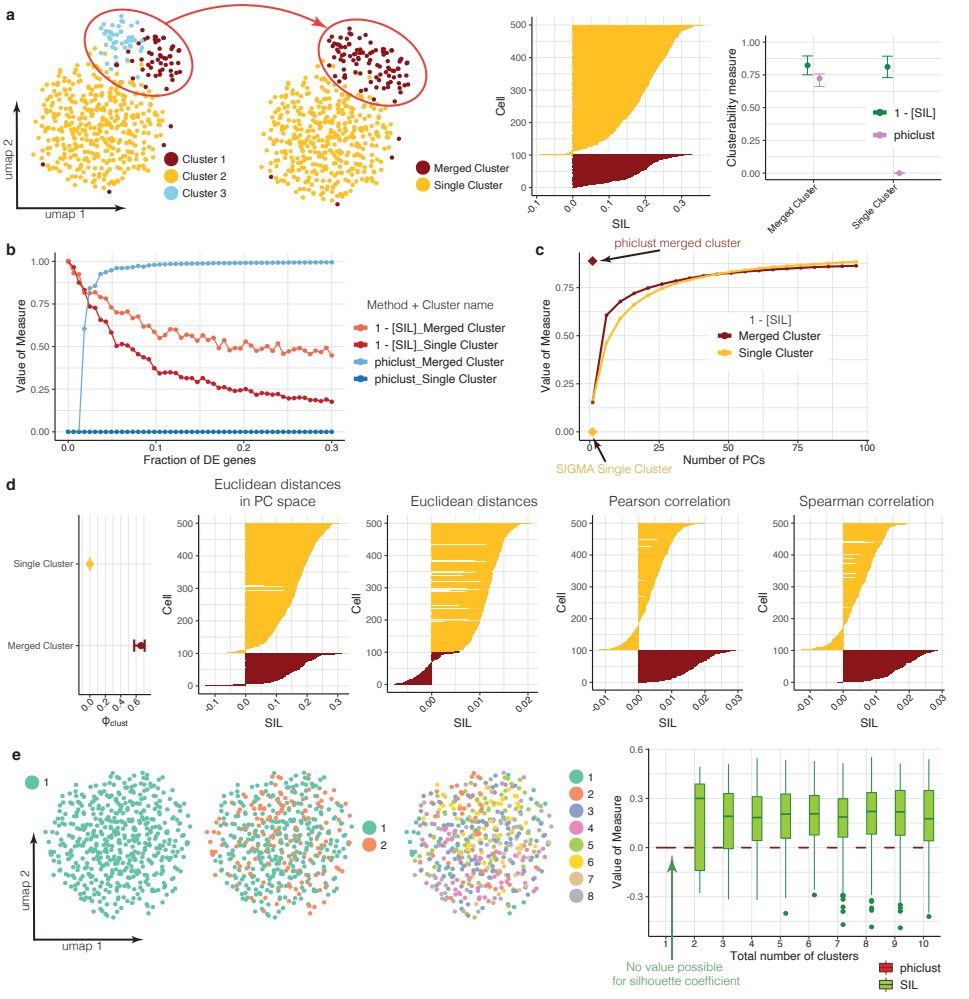


Figure 2.9: Phiclust outperforms the silhouette coefficient on simulated data. a Left: UMAP of 3 simulated clusters. Two clusters were merged to one, resulting in two clusters in total. Middle: Bar plot of silhouette coefficients for each cell, sorted by cluster. Right: phiclust value and average silhouette coefficient for each cluster. b Simulation of clusters as shown in a with different fractions of differentially expressed genes. Phiclust (blue) and average silhouette coefficient (red) for merged cluster and single cluster. c Simulation of clusters as shown in a. The average silhouette coefficient per cluster was calculated for Euclidean distances using different numbers of principal components. Value of phiclust for each cluster is indicated by diamond-shaped data points. d Simulation of clusters as shown in a. Leftmost graph: Value of phiclust for each cluster. Other graphs: Bar plot of silhouette coefficients for each cell, sorted by cluster, calculated with the following distance metrics: Euclidean distances in principal component space, Euclidean distances in the original space, Pearson correlation and Spearman correlation. e Simulation of a cluster consistent with random noise. Clustering was performed with k-means to obtain 2 to 10 clusters. Left: UMAP of simulated data with 0, 2 and 8 clusters. Right: Boxplot of the values of phiclust and the silhouette coefficient for each k-means clustering.

Most importantly, the silhouette coefficient cannot answer the question, whether an identified cluster contains meaningful substructure, as it requires partitioning into at least 2 sub-clusters. We simulated a cluster without any substructure and all variability was purely random (Fig.2.9e). The silhouette coefficient was maximal for a k-means clustering with k=2, which might prompt a user to conclude (wrongly) that there are 2 sub-clusters present. Phiclust, which does not require any further partitioning of the cluster, was 0, indicating correctly that the observed variability was consistent with random noise. All in all, these comparisons indicate that phiclust is a sensitive measure, which detects differences between highly similar phenotypes.

2.2.5 Genes responsible for the detected substructure can be identified

In full analogy to the reasoning outlined so far, our approach can also be used to characterize variability in gene space, for which we defined the conjugate measure g-phiclust (see Additional File 2 for the derivation). Above, we considered only the right singular vectors, where each entry corresponds to a cell in the data set. We therefore also call them "cell-singular vectors". In the simplest case of well separated clusters, entries in the cell singular vectors indicate the membership of a cell in a cluster or a group of clusters. For the left singular vectors, each entry corresponds to a gene. Therefore, we also call them "gene-singular vector". The squared cosine of the angle between the leading gene-singular vector in the measured gene expression matrix and the corresponding gene-singular vector of the noise-free signal matrix is g-phiclust . As for phiclust, data sets with higher signal-to-noise ratios are characterized by higher values of g-phiclust (Fig. 2.10a). "Signal" and "noise" are defined exactly as above: "noise" is a random matrix and the "signal" is a low-rank matrix consisting of noise-free expression profiles, where the strength of the signal (or difference between the clusters) corresponds to the magnitude of the non-zero singular values. A g-phiclust close to 0 would indicate that all observed differential gene expression can be explained by random noise. Larger values of g-phiclust indicate less overlap of the gene expression profiles between phenotypes. We therefore expect to find a bigger number of significantly differentially expressed (DE) genes and/or larger fold changes between phenotypes. We confirmed by simulations that genes with larger absolute entries in a gene-singular vector contribute more to the differences between the clusters separated along the corresponding cell-singular vector (Fig. 2.10b-d): For example, if two clusters (A and B) are separated along a cell-singular vector and cells in cluster A are characterized by positive entries, the genes with large positive entries in the corresponding gene-singular vector will be mostly expressed in cluster A. We call these "variance driving" genes. Our approach thus not only identifies relevant substructure in a cell cluster but can also reveal the genes responsible for it. In contrast to differential expression tests, the variance driving genes can be obtained before clustering and might help the user interpret the observed variability and make an informed decision on whether it is useful to sub-cluster the data. If the variance driving genes have enriched biological features (such as being involved in the same signaling pathway or cellular function), we can take that as additional evidence for biologically meaningful sub-population.

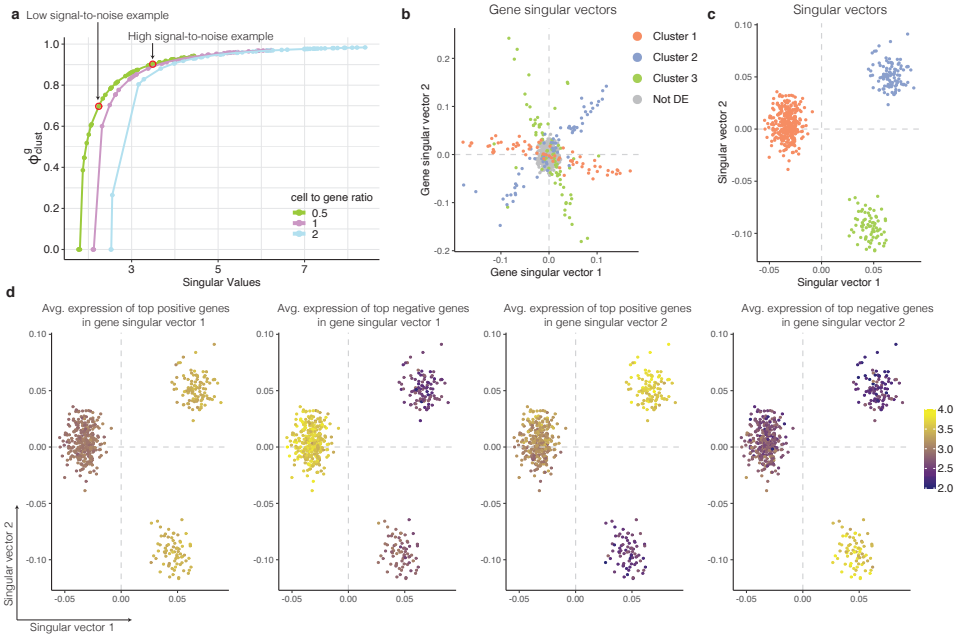


Figure 2.10: Variance-driving genes identified in gene singular vectors coincide with differentially expressed genes in a simulated data set. Genes with high absolute values in the gene singular vector contribute the most to the variability. a Value of the largest singular value versus the squared cosine of the angle between the gene singular vector of the signal matrix and the gene singular vector of the measured expression matrix (g-phiclust) in simulated data. Arrows indicate examples shown in Figure 1b. b First two gene-singular vectors. Differentially expressed genes of each cluster are indicated by color. c First two (cell-)singular vectors for the simulated data set shown in panel b. Dashed grey lines indicate the 0 value on each of the axes. Cell clusters are indicated by color. d First two singular vectors as in c. Dashed grey lines indicate the 0 value on each of the axes. The average log-transformed expression of the top 1% genes driving the variance is indicated by color. The 4 panels show, respectively, from left to right: genes corresponding to the highest values in gene singular vector 1, genes corresponding to the lowest values in gene singular vector 1, genes corresponding to the highest values in gene singular vector 2, and genes corresponding to the lowest values in gene singular vector 2.

2.2.6 Application of phiclust to a BMNC data set drives the discovery of biologically meaningful sub-clusters.

The most important application of phiclust, in our opinion, is to prioritize clusters for further sub-clustering and follow-up studies. For a complex tissue with dozens of clusters, it is not feasible to sub-cluster all of them and try to validate all resulting subpopulations. This is particularly inefficient, if many subclusters are in fact just driven by random noise. High values of phiclust nominate those clusters that likely have deterministic structure and are therefore worthwhile to be scrutinized experimentally in more detail. To demonstrate the application of phiclust and g-phiclust, we analyzed scRNA-seq measurements of complex tissues. In a data set of bone marrow mononuclear cells (BMNCs) [28] we calculated phiclust for the clusters reported by the authors (Fig. 2.11a,b).

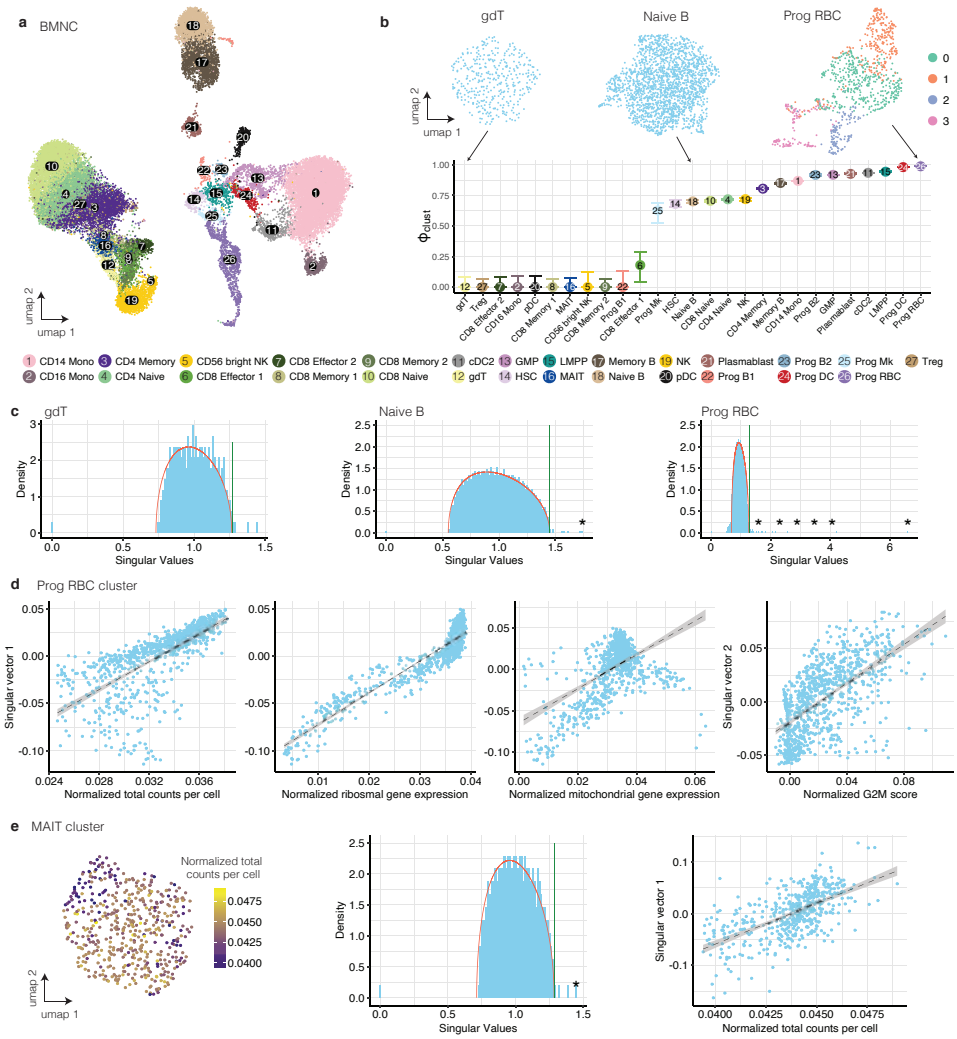
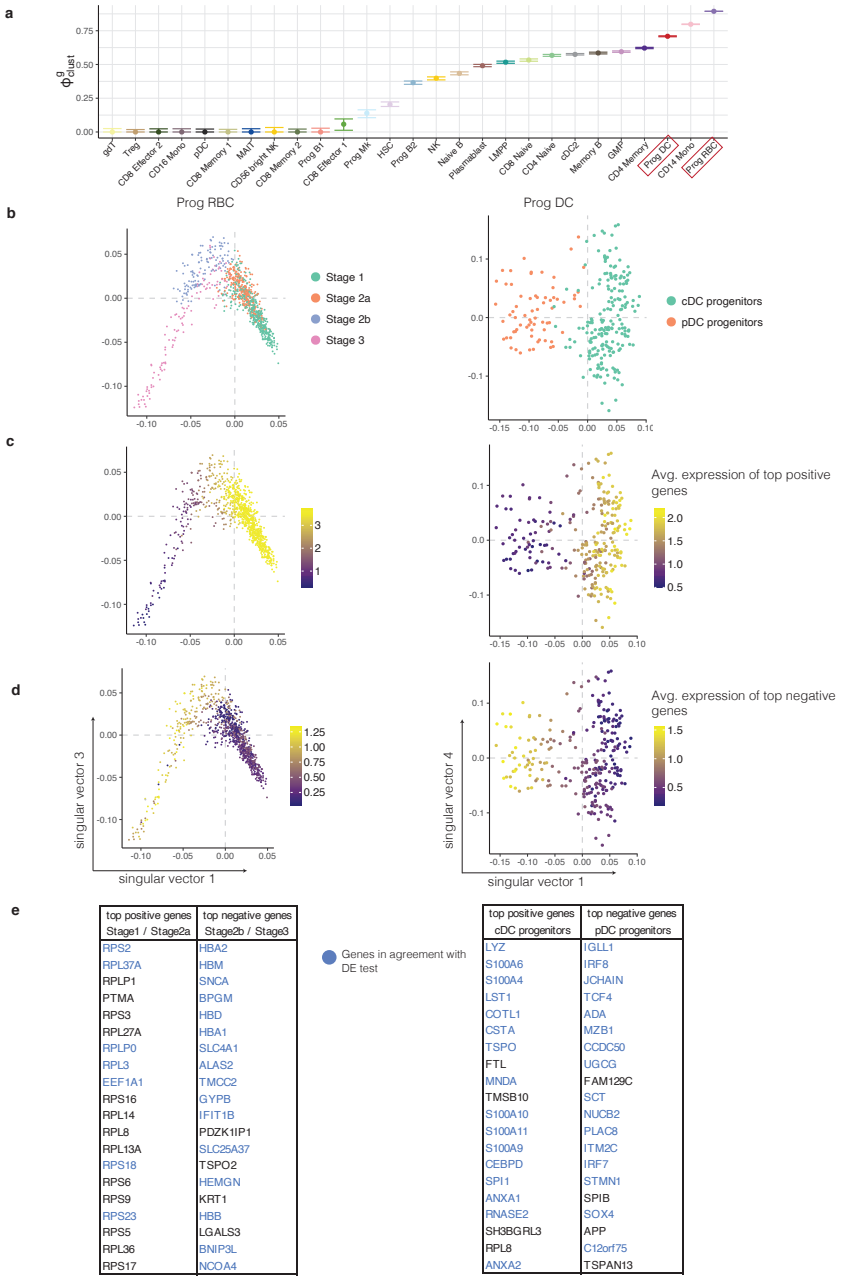


Figure 2.11: Application of phiclust to a BMNC data set drives the discovery of biologically meaningful sub-clusters. a UMAP of BMNC data set. b Phiclust for the BMNC data set. Error bars indicate the uncertainty obtained by resampling the noise. Inset: UMAP of clusters with low, intermediate, and high values of phiclust. c Singular value distribution, MP distribution (red line) and TW threshold (green line) of clusters with low, intermediate, and high values of phiclust. Significant singular values are highlighted with asterisks. In the gdT cluster, the singular vectors corresponding to the outlying singular values had normal distributed entries and were thus not significant. d First three graphs: First singular vector of the red blood cell progenitor cluster in the BMNC data set versus normalized total counts per cell, normalized expression of ribosomal genes, and normalized expression of mitochondrial genes. Rightmost graph: Second singular vector versus normalized G2M score. The dashed line indicates the linear regression and the grey area indicates the standard deviation. e Left: UMAP of the MAIT cell cluster in BMNC data set. The color indicates the normalized total counts per cell. Middle: singular value distribution, MP distribution (red line) and TW threshold (green line) for the MAIT cell cluster. The only significant singular value is indicated by an asterisk. Right: Normalized total counts per cell versus the singular vector associated with the significant singular value (here: first singular vector) in the MAIT cluster.



top positive genes Stage1 / Stage2a	top negative genes Stage2b / Stage3
RPS2	HBA2
RPL37A	HBM
RPLP1	SNCA
PTMA	BPGM
RPS3	HBD
RPL27A	HBA1
RPLP0	SLC4A1
RPL3	ALAS2
EEF1A1	TMCC2
RPS16	GYPB
RPL14	IFIT1B
RPL8	PDZK1IP1
RPL13A	SLC25A37
RPS18	TSPO2
RPS6	HEMGN
RPS9	KRT1
RPS23	HBB
RPS5	LGALS3
RPL36	BNIP3L
RPS17	NCOA4

top positive genes cDC progenitors	top negative genes pDC progenitors
LYZ	IGLL1
S100A6	IRF8
S100A4	JCHAIN
LST1	TCF4
COTL1	ADA
CSTA	MZB1
TSPO	CCDC50
FTL	UGCG
MNDA	FAM129C
TMSB10	SCT
S100A10	NUCB2
S100A11	PLAC8
S100A9	ITM2C
CEBPD	IRF7
SPI1	STMN1
ANXA1	SPIB
RNASE2	SOX4
SH3BGRL3	APP
RPL8	C12orf75
ANXA2	TSPAN13

Figure 2.12: Congruence between variance-driving genes and differentially expressed genes between sub-clusters in a BMNC data set. (Caption on the next page)

For all clusters, except the red blood cell (RBC) progenitor cluster, the bulk of the singular value distribution was well-described by the MP distribution. (In the RBC progenitors, we found several singular values below the lower limit of the MP distribution. These outliers did not influence the further analysis since we are only interested in singular vectors above the upper limit.) The first cell-singular vectors of all clusters were significantly correlated with several confounding factors (see Fig. 2.11d for RBC progenitors and Fig. 2.11e for MAIT cells). After correction for these confounding factors, phiclust corresponded well with a visual inspection of the cluster UMAPs (Fig. 2.11b): Where obvious clusters were present, phiclust was highest, while homogeneous, structure-less clusters resulted in a phiclust of 0. Reassuringly, many progenitor cell types received a high phiclust (indicating possible substructure) in agreement with the known higher variability in these cell types. Ranking existing clusters by g-phiclust resulted in a very similar order (Fig. 2.12a).
To confirm the presence of relevant substructure, we subclustered the two original clusters with the highest phiclust (Fig. 2.12 b-e). In the RBC progenitors, we identified 4 subsets that correspond to different stages of differentiation, ranging from erythroid precursors to highly differentiated RBCs, as identified previously [29]. In the dendritic cell (DC) progenitor cluster, two subclusters were identified, which correspond to precursors of classical or plasmacytoid DCs, respectively [30]. For both examples, the variance-driving genes found in the gene-singular vectors were localized to their corresponding clusters (Fig. 2.12 c,d) and overlapped strongly with differentially expressed genes found after subclustering (see Table S3).

Figure 2.12: Congruence between variance-driving genes and differentially expressed genes between sub-clusters in a BMNC data set. (Figure on the previous page) a g-phiclust for each cluster in the BMNC data set. b Singular vectors of the two clusters from the BMNC data set with the highest phiclust. The color indicates sub-clustering. Dashed grey lines indicate the 0 value on each of the axes. c Singular vectors of clusters shown in panel a with color indicating the average log-transformed gene expression of genes with the 1% highest values in the first gene singular vector. d Singular vectors of clusters shown in panel a with color indicating the average log-transformed gene expression of genes with the 1% lowest values in the first gene singular vector. e Genes driving the variance in the two clusters shown in b. These genes have the 20 highest/lowest values in the first gene singular vector respectively. In blue: top 20 upregulated genes based on differential expression (DE) test between the sub-clusters using findMarkers (from scran R package).

2.2.7 Phiclust reveals subpopulations in a fetal human kidney data set that can be confirmed experimentally

As a second example of our approach we analyzed a fetal human kidney data set we published previously [31]. In our original analysis, we were forced to merge several clusters, since we were unsure if we the observed variability was just noise. We hence wanted to use phiclust to find previously overlooked subpopulations. As for BMNCs, phiclust corresponded well with a qualitative assessment of cluster variability (Fig. 2.13 a): Clusters with visible sub-clusters had the highest values of phiclust. Ordering the clusters by g-phiclust resulted in a similar ranking as phiclust (Fig. 2.15a). Subclustering of the cluster with the highest phiclust, ureteric bud/collecting duct (UBCD), revealed a subset of cells with markers of urothelial cells (UPK1A, KRT7) (Fig. 2.13b, Fig. 2.15 b-e). Immunostaining of these two genes, together with a marker of the collecting system (CDH1), in week 15 fetal human kidney sections confirmed the presence of the urothelial subcluster (Fig. 2.14a, Fig. 2.16a).

Another cell type we did not find in our original analysis, are the parietal epithelial cells (PECs). They could now be identified within the SSBpr cluster (S-shaped body proximal precursor cells) (Fig. 2.13b, Fig. 2.15 b-e).

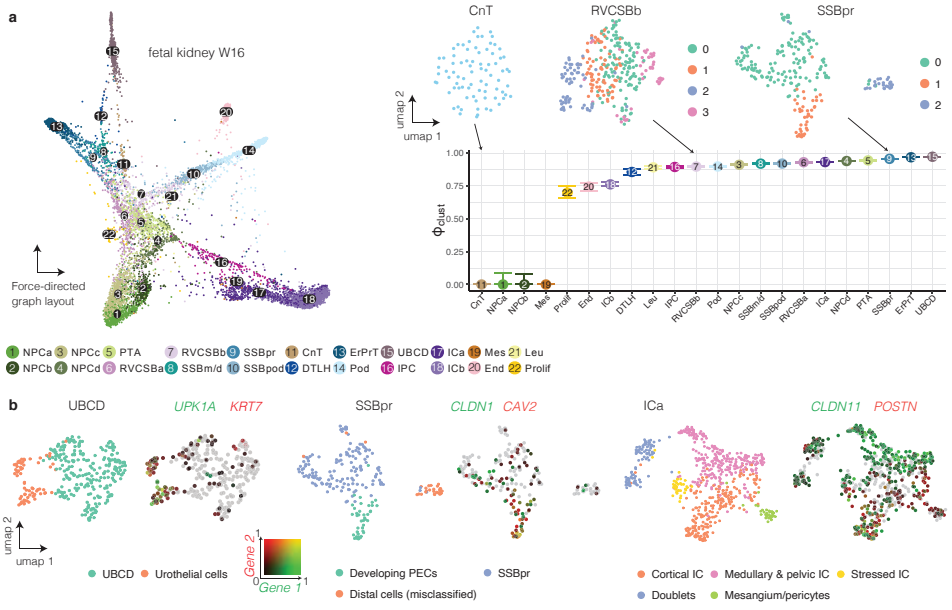


Figure 2.13: Phiclust reveals subpopulations in a human fetal kidney data a Force-directed graph layout and phiclust for the fetal kidney data set. Error bars indicate the uncertainty obtained by resampling the noise. Inset: UMAP of clusters with low, intermediate, and high values of phiclust. b UMAPs of the UBCD, SSBpr, and ICa clusters. Left: Colors indicate sub-clusters. Right: Colors indicate the log-normalized gene expression of the two indicated genes. One gene follows the red color spectrum, the other gene the green color spectrum. Absence of color indicates low expression in both genes, yellow indicates co-expression of both genes.

To reveal these cells in situ, we stained for AKAP12 and CAV2, which were among the top differentially expressed genes in this subcluster (Table S4), together with CLDN1, a known marker of PECs, and MAFB, a marker of the neighboring podocytes (Fig. 2.13d, Fig. 2.16b). Next to the PECs and proximal tubule precursor cells, SSBpr also contained a few cells that were misclassified in the original analysis, indicating the additional usefulness of phiclust as a means to identify clustering errors.

Further analysis of a cluster of interstitial cells (ICa) revealed multiple subpopulations (Fig. 2.13b, Fig. 2.15 b-e). Immunostaining showed that a POSTN-positive population is found mostly in the cortex, often surrounding blood vessels, whereas a SULT1E1-positive population is located in the inner medulla and papilla, often surrounding tubules (Fig. 2.14c, Fig. 2.16c). CLDN11, another gene identified by analysis of the gene-singular vectors (Fig. 2.15b-e), was found mostly in the medulla, but also in the outermost cortex. A more detailed, biological interpretation of the results can be found in Additional File 3.

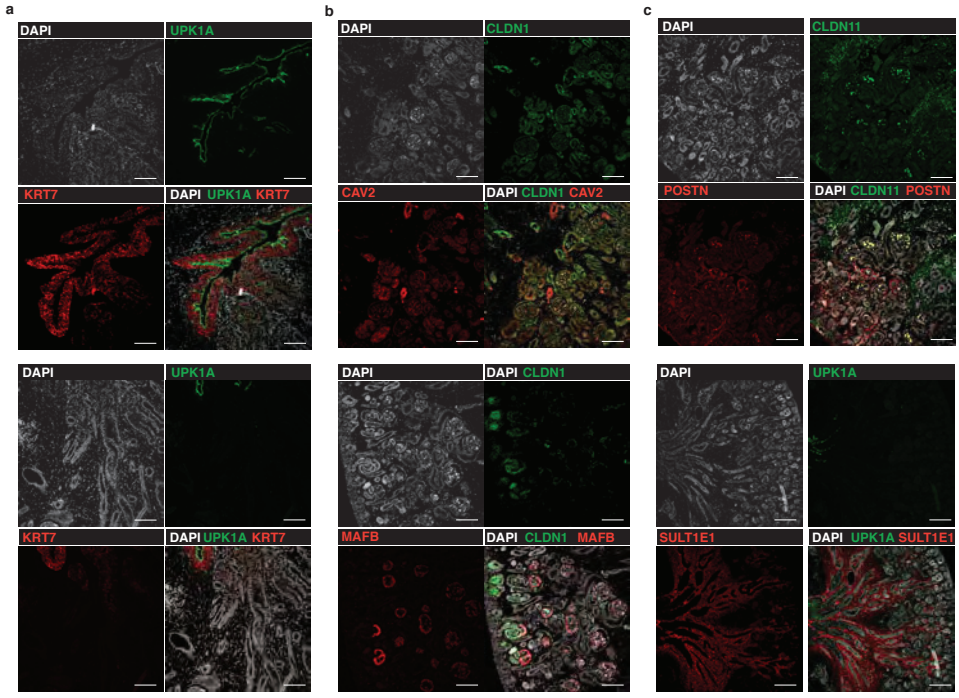
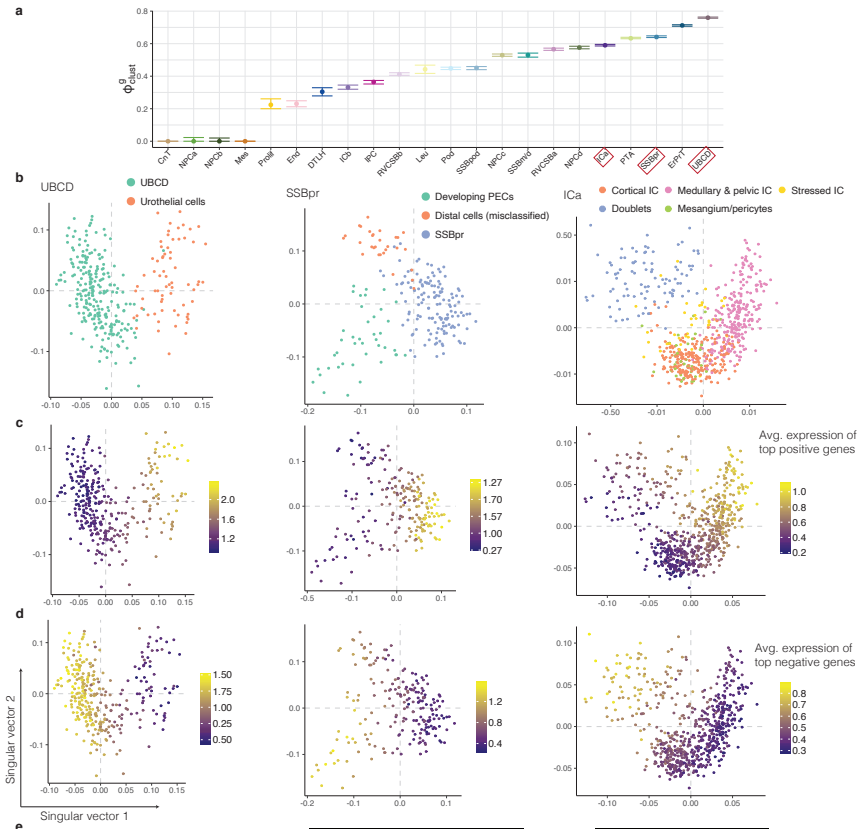


Figure 2.14: Subpopulations in a human fetal kidney data set revealed by phiclust can be confirmed experimentally.c-e Immunostainings of week 15 fetal kidney sections. c UPK1A and KRT7 are expressed in the urothelial cells of the developing ureter (upper panel) and absent from the tubules in the adjacent inner medulla (lower panel). d PECs express CLDN1 and CAV2 (upper panel), as well as CLDN1 at the capillary loop stage and later stages (lower panel). MAFB staining is found in podocytes and their precursors in the SSB (lower panel). e CLDN11 and POSTN are expressed in interstitial cells visualized by immunostaining (upper panel). SULT1E1 is expressed in the interstitial cells surrounding the ureter (marked by UPK1A) and the tubule in the inner medulla (lower panel). Scale bars: 100 µm.



top positive genes Ureteric epithelium	top negative genes UBCD
RPS6	TMSB4X
DHRS2	NGFRAP1
SPINK1	ACTB
S100A6	IGFBP7
HPGD	WFDC2
VSIG2	CLDN3
KRT7	NDUFA4
FXYD3	MEST
FBLN1	HINT1
S100P	STMN1
S100A11	PTMA
ADIRF	ACTG1
SNCG	BCAM
PVALB	VIM
UPK1A	STC1
UQCRQ	NREP
RPS18	CYR61
HMGCS2	PTGER1
FTH1	HNRNPA1
PSCA	HMGN1

top positive genes Proximal Progenitors	top negative genes PEC / Distal progenitors
LDHB	LMNA
MDK	ERP27
PTMA	HERPUD1
CADM1	FOSB
VCAN	MYL9
HNRNPA1	PDZK1IP1
GPC3	PPP1R15A
PSMA2	DNAJB1
H3F3A	KLF4
LHX1	EIF1
CA4	SELK
HMGB1	NR4A1
HSPD1	ZFP36
RPS2	AKAP12
ARGLU1	SERTAD1
HNRNPA2B1	CAPN2
MARCKSL1	MAFF
PODXL2	SDF2L1
EMX2	MLF1
GAPDH	CLDN3

top positive genes Medullary IC	top negative genes Mis-classified cells
TCF21	CXXC5
SULT1E1	CD24
TMSB4X	PAX2
PRSS23	LYPD1
LGALS1	FAM60A
COL3A1	EPCAM
CTSC	BST2
DCN	PTMA
MOXD1	BEX1
IGFBP7	CRABP2
PTN	CADM1
MFAP4	HMGB1
COL1A1	PAX8
TNC	DAPL1
MYLK	LHX1
PLAC9	HMGA1
SFRP1	H2AFZ
B2M	DEK
FBLN5	EZR
CLDN11	KRT18

Figure 2.15: Congruence between variance-driving genes and differentially expressed genes between sub-clusters in a fetal kidney data set. a g-phiclust for each cluster in the fetal kidney data set. b First two singular vectors of three clusters from the fetal kidney data set with high phiclust. The color indicates sub-clustering. Dashed grey lines indicate the 0 value on each of the axes. c First two singular vectors of clusters shown in panel a with color indicating the average log-transformed gene expression of genes with the 1% highest values in the first gene singular vector. d First two singular vectors of clusters shown in panel a with color indicating the average log-transformed gene expression of genes with the 1% lowest values in the first gene singular vector. e Genes driving the variance in the three clusters shown in b. These genes have the 20 highest/lowest values in the first gene singular vector respectively. In blue: top 20 upregulated genes based on differential expression (DE) test between the sub-clusters using findMarkers (from scran R package).

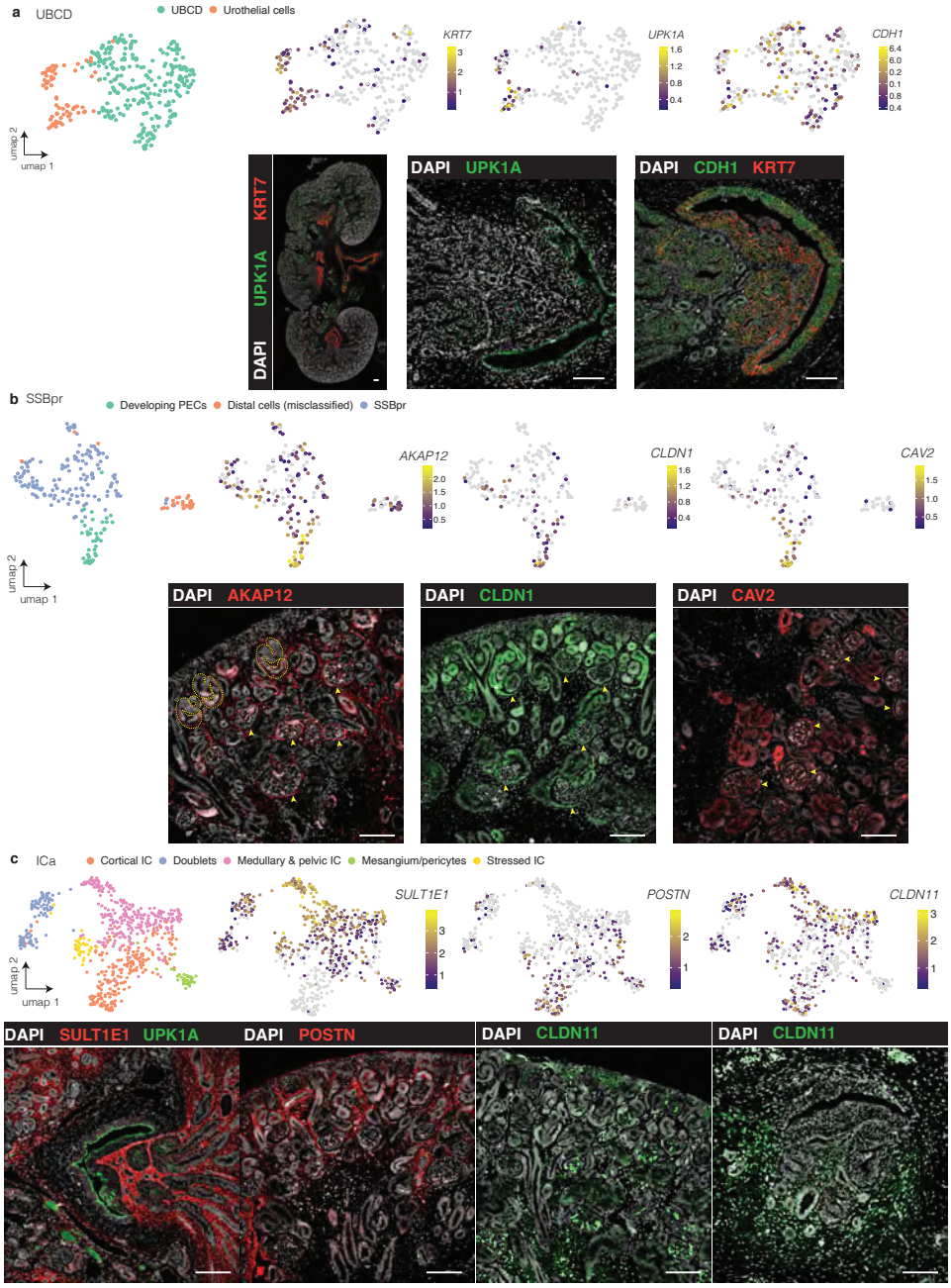


Figure 2.16: Immunostaining validates newly identified subclusters in fetal kidney data set. (Caption on the next page)

2.3 Discussion

Here, we presented phiclust, a clusterability measure that can help detect subtly different phenotypes in scRNA-seq data. Universal properties of the underlying theory make it possible to apply phiclust to arbitrary noise distributions, and the noise can be additive or multiplicative. Empirically, we find that the bulk of the singular value distribution of measured expression matrices is well-approximated by the MP distribution. This supports the assumption that the noise is generated by independent and identically distributed random processes.

While most of the technical and biological noise can likely be considered random, there are also known systematic errors and unwanted, confounding factors (such as the efficiency of RNA recovery, cell cycle phase etc.) Therefore, regressing out uninformative, deterministic factors, is an important part of the method.

The approach underlying phiclust also allows us to identify the genes that are most relevant for the biological interpretation of the observed variability. We found these genes to overlap strongly with differentially expressed genes identified after sub-clustering. The g-phiclust measure, a conjugate to phiclust, quantifies how distinguishable the expression profiles of different phenotypes are in the presence of noise.

The most important application of phiclust is the nomination of clusters for sub-clustering and subsequent experimental validation. All clusters that were nominated in the fetal kidney data set turned out to have subpopulations that could be validated by experiments: rare urothelial cells, which differ from nearby clusters in only a few genes; PECs and subtypes of interstitial cells, which had distinct spatial distributions.

There are several other methods that attempt to detect the presence of meaningful information in single-cell RNA-seq data. Below, we will compare phiclust to some of the most popular examples: the silhouette coefficient, ROGUE, robust PCA, the dip test and ZINB-WaVE.

The silhouette coefficient is a popular tool to assess clustering quality. In contrast to phiclust, this coefficient requires a (sub-)clustering and it cannot be used to decide, whether a cluster contains meaningful variability and should be sub-clustered further. As demonstrated, using the silhouette coefficient can lead to over-clustering of random noise as well as missing the presence of subtly different phenotypes. Likewise, phiclust appeared to be more sensitive than ROGUE, an entropy-based clusterability measure. Both ROGUE and the silhouette coefficient do not scale linearly with the tARI, which we introduced as an upper limit to the achievable agreement of an empirical clustering with the ground truth.

Figure 2.16: Immunostaining validates newly identified subclusters in fetal kidney data set. (Figure on the previous page) a-c Upper panels show UMAPs of the selected clusters in the fetal kidney data set. Log-normalized expression of selected genes is indicated by color. Lower panels show immunostainings of week 15 fetal kidney sections. a UBCD cluster. UPK1A, CDH1, and KRT7 expression is shown in a complete section (leftmost image) and in the urothelial epithelium. b SSBpr cluster. Expression of AKAP12, CLDN1 and CAV2 is shown. The dashed lines indicate S-shaped bodies, arrowheads indicate PECs in developing glomeruli c ICa cluster. Expression of SULT1E1 and UPK1A is visible around the ureter expression of POSTN is visible in cortical areas, CLDN11 is visible in the cortical area (CLDN11, left image) and around the ureter (CLDN11, right image). Scale bars: $100\mu m$.

Robust PCA [32, 33] decomposes a measured expression matrix into a sparse component and a low-rank component. Under the assumption that noise is sparse, the sparse component is identified with random noise. In our opinion, there is no reason to assume that the noise in scRNA-seq data is sparse, or sparser than the measured expression matrix itself. Likely, every non-zero gene expression measurement was corrupted by noise. Additionally, the remaining low-rank component cannot be identified as the noise-free signal. It is fundamentally impossible to reconstruct the noise-free signal from the measured expression because the noise is created by a random process. The low-rank component is therefore only a (noisy) approximation of the noise-free signal. Given the fundamental limit to signal reconstruction, the best thing we can do is quantify the closeness between signal and measured expression, as implemented by phiclust. In robust PCA, the low-rank matrix is often further subjected to dimensionality reduction, where it is difficult to determine the correct number of dimensions. phiclust does not require any dimensionality reduction and uses all available data.

The dip test [9], a method aimed at detecting the presence of clusters, tests whether there are multiple modes in the data. It can be applied directly to the distribution of distances between expression profiles or a low-dimensional representation of the data, such as principal component scores. The dip test will miss relevant variability, if it does not manifest itself as separate modes, which can easily occur, for example in the case of differentiation paths. It also just provides a binary result (modes present or not), whereas phiclust is a continuous measure and does not require the presence of modes.

ZINB-WaVE [25] performs dimensionality reduction based on a zero-inflated negative binomial distribution and is similar to principal component analysis, if no additional covariates are added to the model. ZINB-WaVE acknowledges the fact that principal components are prone to correlate with nuisance parameters, even after normalization. The problem is circumvented by adding such parameters as covariates to the model, which is similar to the confounder regression used for phiclust. However, the user has to decide the number of dimensions to use and currently there is no principled way to determine the optimal number. phiclust does not have any such adjustable parameters.

2.4 Conclusion

We hope that this manuscript will bring renewed awareness to random noise as a factor that imposes hard limits on clustering and identification of differentially expressed genes. We hope that quantitative measures of clusterability, such as phiclust, can play an important role in making single-cell RNA-seq analysis more reproducible and robust.

2.5 Methods

2.5.1 Preprocessing

Before applying the method to simulated or measured single-cell RNA-seq data sets, several preprocessing steps are necessary. The raw counts are first normalized and log-transformed. Next, the expression matrix is standardized, first gene-wise, then cell-wise. These steps assure the proper agreement of the bulk of the singular value distribution with the MP distribution (Additional file 1: Fig. S2). See also Supplementary Note, Section 3.1.

2.5.2 Phiclust

To derive phiclust, we assume that the expression matrix $\widetilde{X}$ measured by scRNA-seq, can be written as the sum of a random matrix X, which contains random biological variability and technical noise, and a signal matrix P, which contains the unobserved expression profiles of each cell:

$$\widetilde{X} = X + P$$

Note that in this decomposition, cells that belong to the same cell type (or phenotype) have identical expression profiles in the signal matrix P. Below we will show that multiplicative noise can be treated analogously.

We apply SVD to obtain the singular values, as well as the right and left singular vectors of $\widetilde{X}$. The left singular vectors span gene-space and the right singular vectors span cell-space. Hence, we call them gene-singular vectors and cell-singular vectors, respectively. If we use the term "singular vector" it is implied to mean cell-singular vector.

Considering the signal matrix P a perturbation to the random matrix X, we can apply results from both random matrix theory and low-rank perturbation theory. Random matrix theory [33, 34] predicts that the singular value distribution of X is a Marchenko-Pastur (MP) distribution [18, 19, 35], which coincides with the bulk of the singular value distribution [12–14] of $\widetilde{X}$. The singular values of $\widetilde{X}$ above the values predicted by the MP distribution characterize the signal matrix P. Since the agreement with the MP distribution holds strictly only for infinite matrices, we use two additional concepts to identify relevant singular values exceeding the range defined by the MP distribution. The Tracy-Widom [16, 36] (TW) distribution describes the probability of a singular value to exceed the MP distribution, if the matrix is finite. Additionally, since singular vectors of a random matrix are normally distributed, relevant singular vectors have to be significantly different from normal [14]. To test for normality we used the Shapiro-Wilk test.

We apply low-rank perturbation theory [17] to calculate the singular values (θ_i) of P from the outlying singular values (γ_i) of the measured expression matrix $\widetilde{X}$:

$$\theta_i(\gamma_i) = \sqrt{\frac{2c}{{\gamma_i}^2 - (c+1) - \sqrt{\left({\gamma_i}^2 - (c+1)\right)^2 - 4c}}}$$

where c is the cell-to-gene ratio, i.e. the total number of cells divided by the total number of genes.

The values of θ_i are then used to obtain the angles ϕ_i between the singular vectors of $\widetilde{X}$ and P. These angles are conveniently expressed in terms of their squared cosine as

$$\phi_i = cos(\alpha_i)^2 = 1 - \frac{c\left(1+\theta_i^2\right)}{\theta_i^2\left(\theta_i^2+c\right)}.$$

The leading singular vector of the measured expression matrix, which has the largest singular value, has the smallest angle to its unperturbed counterpart. The squared cosine of this smallest angle is then used as a measure of clusterability:

$$\phi_{clust} = \cos(min_i\phi_i)^2 = max_i \cos(\phi_i)^2, \quad phi_i \in [0, \frac{\pi}{2}]$$

For a detailed derivation of phiclust, see Additional File 2, Section 2.1-2.4.

Uncertainty of phiclust

The uncertainties for the values phiclust are estimated using a sampling approach. The basic idea is to approximate the signal matrix P and add new realizations of the noise matrix by sampling from a random distribution. The uncertainty is then obtained from the values phiclust calculated for this ensemble of sampled matrices. First, we decompose a simulated or measured expression matrix $\widetilde{X}$ into a noise matrix Xr and a matrix Xs that contains deterministic structure. Xs is constructed from the relevant singular vectors, which were identified as described in the previous section. Note that Xs contains noise and is thus different from the signal matrix P. To create an approximation Ps of the signal matrix P, we replace the singular values γ_i used in the construction of Xs with the singular values θ_i of P, calculated using low-rank perturbation theory as shown in the previous section. The entries of the noise matrix Xr have a mean of 0 and a standard deviation of 1, as a result of preprocessing. Since the results of RMT are valid irrespective of the particular noise distribution, we can create additional realizations of the noise matrix by sampling from a normal distribution with mean 0 and standard deviation 1. By adding sampled noise matrices to the approximated signal matrix Ps, we can create an ensemble of matrices with the same singular value spectrum as the original measured expression matrix but different realizations of the noise. The uncertainty for positive and negative deviations from the mean is then calculated as the standard deviation for at least 50 sampled matrices. See Supplementary Note, Section 2.4.3 for a detailed description.

Test for deviation from the MP distribution

To validate the use of the MP distribution, we test whether the bulk of the measured singular value distribution deviates significantly. Singular values are considered to be part of the bulk, if they are located below the MP upper bound and not associated with the transcriptome mode. We sample 1000 values from the MP distribution using the RMTstat R package (V 0.3) and subsequently test for similarity with the Kolmogorov-Smirnov test [35]. The resulting p-values are adjusted for multiple hypothesis testing with the Benjamini-Hochberg procedure [37].

Confounder regression

scRNA-seq data contains various confounding factors that drive uninformative variability. These either emerge from technical issues (such as the varying efficiency of transcript recovery, which cannot be fully eliminated by normalization) or biological factors (such as cell cycle phase, metabolic state, or stress). To account for these factors, a regression step, inspired by current gene expression normalization methods [23, 26], is included. We perform a linear regression by using each relevant singular vector as a response variable and the confounding factors as covariates. This is a valid approach because the singular vectors of the measured expression matrix contain normal distributed noise. The amount of variance explained by the nuisance parameters is then given by the value of the adjusted R2 (coefficient of determination) of this linear regression. To relate the regression result to the singular values, we consider the squared singular values (= eigenvalues) which correspond to the variance explained by the corresponding singular vectors / eigenvectors. Squared singular values are corrected by multiplication with (1 – adjusted R2) to retrieve the fraction of variance not explained by nuisance parameters. The square root of the result is the corrected singular vector. See also Supplementary Note, Section 3.2. For Additional file 1: Fig. S5a, each nuisance parameter was individually regressed on, to compare the influence of each factor.

Multiplicative noise

To model multiplicative noise, we use a rectangular random noise matrix X with the same dimensions as the measured expression matrix $\widetilde{X}$ and a square signal matrix P whose number of rows or columns is equal to the number of measured genes. The measured expression matrix $\widetilde{X}$ is then modeled as:

$$\widetilde{X} = (I + P)^{\frac{1}{2}} \; X,$$

Where I denotes the identity matrix. Importantly, the bulk of the singular vector distribution of the measured expression matrix $\widetilde{X}$ still follows the MP distribution in this model. The singular values of the signal matrix P are calculated from the outlying singular values of $\widetilde{X}$ by:

$$\theta_i = \frac{2c}{\lambda_i - c - 1 - \sqrt{(\lambda_i - a)(\lambda_i - b)}}$$

with $a, b = \left(1 \pm \sqrt{c}\right)^2$. The angles between the corresponding singular vectors of the measured expression matrix and the signal matrix are then calculated as: $\phi\sigma^i{}_{mult} = \frac{1}{\theta_i} \frac{{\theta_i}^2 - c}{\theta_{i(c+1)} + 2c}$
More information on multiplicative perturbation can be found in [38].

2.5.3 Clustering

Theoretically achievable clustering quality

To construct a Bayes classifier [22], which achieves the minimal error rate, we need to know the ground truth clustering. Hence, we used data simulated with Splatter [20], containing two clusters. For each ground truth cluster, we fit a multidimensional Gaussian to the corresponding entries of the singular vectors (see Additional file 1: Fig. S3a). We only

consider singular vectors with singular values larger than predicted by the MP distribution. For the fit, we use the mclust [39] R package (V 5.4.6). We then construct a classifier by assigning a cell to the cluster for which it has the highest value of the fitted Gaussian distribution. This classifier is thus approximately a Bayes classifier (for a true Bayes classifier, we would need to know the exact distributions of the singular vector entries). The ARI [21] calculated based on this classification is thus approximately the best theoretically achievable ARI (tARI). We also tested the silhouette coefficient [8] as a potential alternative to the ARI for quantifying our notion of clusterability. The silhouette coefficient was calculated on the first singular vector using Euclidean distances. In Additional file 1: Fig. S4 the silhouette coefficient averaged over all cells is reported. The theoretically achievable silhouette coefficient tSIL is defined as the silhouette coefficient of the Bayes classification described in the previous paragraph. The calculation of tARI and tSIL is described in more detail in Additional File 2, section 2.5.

Clustering methods

For the validation of the tARI and tSIL, several clustering methods were used on simulated data with two clusters. Seurat clustering [2] was performed with the Seurat R package with 10 principal components (PCs) and 20 nearest neighbors. Three different resolution parameters were used: 0.1, 0.6, and 1.6. Scanpy clustering [3] was performed with the scanpy python package with 10 PCs and 20 nearest neighbors. Three different resolution parameters were used: 0.1, 0.6, and 1.6. Hierarchical clustering [5] was performed on the first 10 PCs and Euclidean distances. The hierarchical tree was built with the Ward linkage and the tree was cut at a height where 2 clusters could be identified. K-means [4] was performed on the first 10 PCs using Euclidean distances and two centers. TSCAN [40] was calculated on the first 10 PCs. In Additional file 1: Fig. S7 k-means clustering was performed on the first 3 principal components and using Euclidean distances.

Clusterability measures

ROGUE [10] is an entropy-based clusterability measure. A null model is defined under the assumption of Gamma-Poisson distributed gene expression and its differential entropy is then compared to the actual differential entropy of the gene expression profile. For the RNA-mix data set ROGUE (V 1.0) was used with 1 sample (see Fig S6), "UMI" platform, and a span of 0.6. For the simulated data sets, ROGUE was used with k = 10 (Additional file 1: Fig. S4 d). The silhouette coefficient was calculated with the cluster R package (V 2.1.0) using euclidean distances in the space of the relevant singular vectors. The reported values for the silhouette coefficients are average values per cluster. The confidence intervals given in Additional file 1: Fig. S6 and S7 are standard deviations of its values per cluster.

2.5.4 Variance driving genes

Genes that drive the variance in the significant singular vectors can be used to explore the biological information in the sub-structures. Genes with large positive or negative entries in a gene-singular vector are localized in cells with high positive or negative entries in the corresponding cell-singular vector. It is also possible to assess the signal-to-noise ratio for the genes by calculating the angle between the gene singular vectors of the measured

expression matrix $\widetilde{X}$ and the gene singular vectors of the signal matrix P, given by15

$$\phi^g_{clust} = cos\left(\widetilde{\alpha\phi}\right)^2 = 1 - \frac{\left(c + \theta_i^2\right)}{\theta_i^2\left(\theta_i^2 + 1\right)},$$

2

where c is the cell-to-gene ratio. We call ϕ^g_{clust} the gene phiclust (g-phiclust). See Additional File 2, section 2.4 for a more detailed discussion.

2.5.5 Data sets

The simulated data sets in Additional file 1: Fig. S1 comprised 201 cells and 350 genes. The random noise matrix was sampled from a normal distribution with mean 0 and variance 1 in panels a and b, or from a Poisson distribution with parameter 1 in panels c and d. The rank 1 signal matrix was constructed from one cell-singular vector and one gene-singular vector. The cell-singular vector consisted of 67 entries equal to $1/\sqrt{N_{cell}}$ and all other entries equal to $-1/\sqrt{N_{cell}}$, where Ncell is the number of cells. The gene-singular vector consisted of 200 entries equal to $1/\sqrt{N_{gene}}$ and the rest equal to $-1/\sqrt{N_{gene}}$, where Ngene is the number of genes. The signal matrix was then created by matrix multiplication of the gene-singular vector and the transposed cell-singular vector times the singular value θ ($\theta = 2$ in a,c and $\theta = 5$ in b,d). In Additional file 1: Fig. S2f,g a rank 1 signal matrix was created similarly as described above. The cell-singular vector with a number of entries matching the number of cells in the cluster was constructed as before. The gene-singular vector was drawn from a normal distribution and subsequently normalized to unit length. The rank 1 signal matrix was then added to the preprocessed expression matrix of the indicated cluster. The remaining simulated data sets were produced with the splatter [20] R package (V 1.10.1). The parameters used for the simulation are shown in Table S1. For Fig. 1c,d, Additional file 1: Fig. S3b-e, Additional file 1: Fig. S4, and Additional file 1: Fig. S8a the simulations for each parameter were performed 50 times, each with a different seed. The results were averaged over the 50 runs. Confounder regression was performed for the total number of transcripts per cell. PBMC data [23] was downloaded from the 10x genomics website (). For the calculation of the tARI, clustering with Scanpy, TSCAN, k-means, and hierarchical clustering, preprocessing was performed with the scanpy python package (V 1.4.6) following the provided pipeline () for the filtering of cells and genes, normalization, and log-transformation as well as cluster annotation. For the clustering with Seurat, the provided Seurat pipeline was used () for preprocessing, such as cell and gene filtering, normalization, log-transformation and cluster annotation using the Seurat R package (V 3.1.5). CD8 T cells and B cells were extracted from the data and each cluster was standardized gene-wise and cell-wise before the calculation of the SVD. To remove any sub-structure in these clusters and before the reconstruction of the matrices from the SVD, singular values above the MP distribution were moved into the bulk, and the transcriptome mode (i.e. the singular vector that would have the largest singular value without normalization, see Supplementary Methods Note 1) was moved above the MP distribution. Then, two synthetic clusters containing 150 cells each were created from the cleaned-up original clusters. For cluster 1, a weighted average of a randomly picked B cell with expression profile c_B and a randomly picked CD8 T cell with expression profile $c_{CD8\ T}$ was calculated according to: $c_1 = \alpha \bullet c_B + (1 - \alpha) \bullet c_{CD8\ T}$. For cluster 2, the weights

were flipped: $c_2 = (1 - \alpha) \cdot c_B + \alpha \cdot c_{CD8\ T}$. α was chosen in a range from 0 to 1. α close to 0.5 produced highly similar clusters, while α close to 0 or 1 produced maximally different clusters (see Fig S3e). For each value of α, the procedure was repeated 50 times, each with a different seed for selecting 300 cells per cell type, and the results were averaged. RNA-mix data [24] was downloaded from the provided GitHub page. The data were normalized with the R scran package (V 1.14.6) and then log-transformed. Confounder regression was performed for the total number of transcripts, average mitochondrial gene expression, and average ribosomal gene expression. Two different merged clusters were created from the provided RNA mixtures as shown in Additional file 1: Fig. S6. The bone marrow mononuclear cell data set (BMNC) [28] was downloaded from the R package SeuratData (bmcite, V 0.2.1). Normalization and calculation of the G2M score [41] were performed with the Seurat R package (V 3.1.5). Confounder regression was performed for the log-transformed total number of transcripts, cell cycle score, and average expression of: mitochondrial genes and ribosomal genes (list obtained from the HGNC website). For the fetal kidney data set [31], the same preprocessing and normalization was used as reported previously (scran R package [42]). The data was then log-transformed and the G2M score was calculated with the Seurat R package. Confounder regression was performed for the log-transformed total number of transcripts, G2M scores, and the average expression of: mitochondrial genes, ribosomal genes, and stress-related genes [43].

2.5.6 Single cell data analysis

Embedding

Uniform Manifold Approximation and Projections [44] (UMAPs) for individual clusters were calculated with the R package umap (V 0.2.7.0) on the first 10 PCs, 20 nearest neighbors, min_ dist = 0.3, and Euclidean distances. The umap for BMNC data was calculated with the Seurat R package using 2000 highly variable genes (hvg), d = 50, k = 50, min.dist = 0.6 and metric = cosine. For the fetal kidney data set a force-directed graph layout was calculated using the scanpy python package. The graph was constructed using 100 nearest neighbors, 50 PCs, and the ForceAtlas2 layout for visualization.

Differential expression test

Differentially expressed genes within the sub-clusters found in Additional file 1: Fig. S9 and Additional file 1: Fig. S10 were calculated with the function findMarkers of the scran R package on log-transformed normalized counts. Genes with a false discovery rate below 0.05 were selected and then sorted by log2 fold change. In Figures S9e and S10e, genes with the top 20 highest/lowest values in the gene singular vectors are listed and colored blue if they correspond to the top 20 DE genes.

2.5.7 Staining

A human fetal kidney (female) at week 15 of gestation was used for immunofluorescence using the same procedure as reported previously [31]. The following primary antibodies were used: rabbit anti-UPK1A (1:35, HPA049879, Atlas Antibodies), mouse anti-KRT7 (1:200, # MA5-11986, Thermo Fisher Scientific), rabbit anti-CDH1 (1:50, SC-7870, Santa Cruz), rabbit anti-CLDN1 (1:100, # 717800, Thermo Fisher Scientific), goat anti-CAV2 (1:100, AF5788-SP, R& D Systems), mouse anti-AKAP12 (1:50, sc-376740, Santa Cruz), rabbit anti-

CLDN11 (1:50, HPA013166, SIGMA Aldrich), mouse anti-POSTN (1:100, sc-398631, Santa Cruz) and goat anti-SULT1E1 (1:50, AF5545-SP, R& D Systems). The secondary antibodies were all purchased from Invitrogen and diluted to 1:500: Alexa Fluor 594 donkey anti-mouse (A21203), Alexa Fluor 594 donkey anti-rabbit (A21207), Alexa Fluor 647 donkey anti-mouse (A31571), Alexa Fluor 647 donkey anti-rabbit (A31573), Alexa Fluor 647 donkey anti-goat (A21447). The sections were imaged on a Nikon Ti-Eclipse epifluorescence microscope equipped with an Andor iXON Ultra 888 EMCCD camera (Nikon, Tokyo, Japan).

2.5.8 Data availability

All sequencing data sets were obtained from publicly available resources. The BMNC data can be downloaded with the R package SeuratData, named "bmcite." The fetal kidney data is available in the GEO database under the accession number GSE114530. The PBMC data can be downloaded at https://cf.10xgenomics.com/samples/cell/pbmc3k/pbmc3k_filtered_gene_bc_matrices.tar.gz and the RNA-mix data is available at https://github.com/LuyiTian/sc_mixology, named "mRNAmix_qc". Supplementary tables are available in the online version at https://doi.org/10.1186/s13059-021-02590-x.

2.6 Supplementary Note 1

Introduction

Our aim is to develop a clusterability measure for scRNA-seq data. As we define more precisely in section (2.6), we consider clusterability to be the clustering quality that is optimally achievable, given a certain amount of noise in the data. Clustering quality can only be assessed quantitatively if the ground truth is known, which is strictly only the case for simulated data. A clusterability measure must thus be able to reflect clustering quality without knowledge of the ground truth. Such a measure would be highly useful, since it would allow us to detect the presence of meaningful (non-random) variability, and thus determine the necessity to sub-cluster measured data. For the development of this clusterability measure we will use concepts from random matrix theory and perturbation theory. In short, we decompose the single-cell gene expression matrix $\tilde{X}$ into a random matrix X, which contains technical and biological noise, and a signal matrix P, which contains the expression profiles of different cell types or states. Then, we apply perturbation theory, treating the signal matrix P as a low-rank perturbation of the noise matrix X. Perturbation theory then allows us to calculate the angle between the singular vectors of the measured single cell expression matrix $\tilde{X}$ and the corresponding singular vectors of the unobserved signal matrix P. The cosine of this angle constitutes a useful clusterability measure because a large value (small angle) indicates a high signal-to-noise ratio (and thus high clusterability) and a small value (large angle) indicates a low signal-to-noise-ratio (and thus low clusterability). We show empirically that this clusterability measure is a proxy for the theoretically achievable adjusted rand index [Fig. 1d].

In what follows, we first present our model of gene expression data (2.6) and introduce matrix decomposition (2.6). Subsequently, we introduce the Marchenko-Pastur (MP) distribution (2.6), which describes the eigenvalue spectrum of a random matrix and apply perturbation theory to link the (unobserved) signal matrix to the spectrum of the measured expression matrix (2.6). In section (2.6), we establish our notion of clusterability. In section (2.6), we describe the preprocessing steps necessary for the application of the theory to single-cell RNAseq data. Then, in section (2.6), we develop a method to remove the effect of nuisance variables (i.e. sources of systematic, non-random variability that should not drive clustering.) The complete algorithm can be found in section (2.6).

Phiclust

Model

Let $\tilde{X} \in \mathbb{R}^{M \times N}$ be the measured single-cell expression matrix with M the number of genes (rows) and N the number of cells (columns). We model the measurement $\tilde{X}$ as the sum of a random noise matrix $X \in \mathbb{R}^{M \times N}$ and a "signal" matrix $P \in \mathbb{R}^{M \times N}$.

$$\tilde{X} = X + P \tag{2.1}$$

In our model, X contains both technical and biological noise. For example, if there was only one cell type or cell state present in a data set, P would consist of identical columns. Note that we only observe the matrix $\tilde{X}$ experimentally. We will show below, that we can make a statement about the influence of the noise X on the signal P, without knowing X or P. To achieve that we invert the logic of conventional models: instead of modeling

the influence of random noise on the signal, we consider the influence of a deterministic perturbation on a random matrix. All results rely on matrix decomposition, which will be introduced next.

MATRIX DECOMPOSITION

1. Eigendecomposition

We first define the cell-cell correlation matrix. To that end, we assume that $\tilde{X}$ has been standardized cell-wise (i.e. column-wise) to mean 0 and standard deviation 1. The cell-cell correlation matrix $C \in [-1,1]^{N\times N}$ is then defined as:

$$C = \frac{1}{M-1}\tilde{X}^T\tilde{X} \tag{2.2}$$

The correlation matrix is a square and symmetric matrix which can hence, by the spectral theorem, undergo eigendecomposition into the form

$$C = V\Sigma V^T = \sum_{i=1}^{N} \lambda_i v_i v_i^T. \tag{2.3}$$

$V \in \mathbb{R}^{N\times N}$ contains the eigenvectors v_i of C in the columns and $\Sigma \in \mathbb{R}^{N\times N}$ is a diagonal matrix containing the eigenvalues λ_i of C. If $M < N$, then C is a singular matrix and will contain at least $N - M$ eigenvalues equal to 0, which is an important consideration for the definition of the Marchenko-Pastur distribution (see below).

In full analogy to the cell-cell correlation matrix we can define a gene-gene correlation matrix $\hat{C}$, now assuming that the expression matrix $\tilde{X}$ has been standardized gene-wise (row-wise) to mean 0 and standard deviation 1:

$$\hat{C} = \frac{1}{N-1}\tilde{X}\tilde{X}^T. \tag{2.4}$$

If $M > N$, then $\hat{C}$ is a singular matrix and will contain at least $M - N$ eigenvalues equal to 0. Therefore either C (if $M < N$) or $\hat{C}$ (if $M > N$) is a singular matrix (unless $M = N$) with at least $|N - M|$ eigenvalues equal to 0.

2. Singular value decomposition

To decompose the (rectangular) expression matrix $\tilde{X}$ into noise and signal, we use singular value decomposition:

$$\tilde{X} = \sum_{i=1}^{N} \gamma_i u_i v_i^T.$$

The v_i's are the right singular vectors of $\tilde{X}$ and correspond to the eigenvectors of the cell-cell correlation matrix. We will call them cell singular vectors or singular vectors in the following. The u_i's are the left singular vectors of $\tilde{X}$ and correspond to the eigenvectors of the gene-gene correlation matrix, which we will call gene singular vectors. The singular values are denoted by γ_i. The singular values of $\tilde{X}$ and the eigenvalues of the corresponding correlation matrix have a known connection given by:

$$\lambda_i = \gamma_i^2.$$

Random Matrix Theory

The Marchenko-Pastur (MP) distribution is widely used to reveal nonrandom properties of empirical correlation matrices in physics and finance [12, 13]. The MP distribution describes the distribution of eigenvalues of a random correlation matrix in the asymptotic limit [18, 19, 36] (for $N \longrightarrow \infty$ and $M \longrightarrow \infty$, $\frac{N}{M} < 1$). The entries of the random matrix are arbitrary as long as they are distributed identically and independently. scRNA-seq data are typically modeled by a Poisson, a negative binomial or a zero-inflated negative binomial distribution, which are in principle admissible in random matrix theory.

Theorem 1 (Marchenko-Pastur) *([18, 19, 36]) Let Y be a $M \times N$ matrix with entries that are independent identically distributed (i.i.d.), mean 0 and variance $v^2 < \infty$. The corresponding Wishart matrix is defined as $W = \frac{1}{M} Y^T Y$. For $N \longrightarrow \infty$, $M \longrightarrow \infty$ and $0 < c < 1$, where c is defined as $\frac{N}{M}$. The distribution of the eigenvalues λ of W is given by*

$$\mu(\lambda) = \frac{\sqrt{(b-\lambda)(\lambda-a)}}{2\pi c \lambda v^2} \, d\lambda \quad \text{if} \quad a \leq \lambda \leq b$$

For $c > 1$ the distribution has an additional number of 0 eigenvalues:

$$\mu(\lambda) = \frac{\sqrt{(b-\lambda)(\lambda-a)}}{2\pi c \lambda v^2} \mathbb{1}_{[a,b]} + (1 - \frac{1}{c}) \delta_0(\lambda)$$

with

$$a, b = v^2 \left[1 \pm \sqrt{c}\right]^2.$$

$\delta_0(\lambda)$ is the Dirac delta function, which is 1 if $\lambda = 0$ and 0 otherwise. For the correlation matrix we obtain $v = 1$ because the mean of all eigenvalues is 1.
This theorem places the eigenvalues of a random correlation matrix into a compact interval between $[a, b]$. All eigenvalues of an empirical correlation matrix that fall within this interval can be considered to be due to random noise. The presence of eigenvalues above this distribution indicates the existence of non-random structure in the data. An empirical (measured) correlation matrix can therefore be decomposed into a random part C^r and a signal part C^s [19]:

$$C = \sum_{\lambda \leq b} \lambda_i v_i v_i^T + \sum_{\lambda > b} \lambda_i v_i v_i^T = C^r + C^s$$

C^s contains the non-random and therefore biologically relevant correlations.

For the application of the MP distribution to an empirical correlation matrix we need to consider that the eigenvalues of a correlation matrix always sum up to 1. Thus, if there are eigenvalues above the MP distribution the bulk of the distribution (which is described by MP) will shift to the left. To approximately account for this shift, we introduce a modified MP-distribution as follows:

$$\mu^*(\lambda) = \frac{\mu(\lambda)}{\alpha},$$
$$a^* = \alpha a, \; b^* = \alpha b.$$

where $\alpha = 1 - \frac{\lambda_{max}}{N}$ and a^* and b^* replace a and b respectively.

We can formulate the MP distribution also for singular values, via a variable transform, and obtain the following density:

$$d\rho(\gamma) = \frac{\sqrt{(b-\gamma^2)(\gamma^2-a)}}{\pi\gamma c}\, d\gamma \quad \text{if} \quad \sqrt{a} \leq \gamma \leq \sqrt{b} \tag{2.5}$$

In this case, all singular values that lie within the compact interval of $[\sqrt{a}, \sqrt{b}]$ can be considered to arise from random noise and singular values above this threshold indicate deterministic biological relevant signal. Thus, we can decompose the matrix $\tilde{X}$ into two parts:

$$\tilde{X} = \sum_{\gamma \leq \sqrt{b}} \gamma_i u_i v_i^T + \sum_{\gamma > \sqrt{b}} \gamma_i u_i v_i^T = \tilde{X}^r + \tilde{X}^s \tag{2.6}$$

The first part $\tilde{X}^r$ is random noise, the second part $\tilde{X}^s$ contains relevant signal.

The MP theorem holds strictly only in the asymptotic limit, but provides a very good approximation for big enough N and M. For finite dimensions, there is however a non-zero probability that a random i.i.d matrix has eigenvalues above the MP distribution. That probability is described by the Tracy-Widom (TW) distribution.

Theorem 2 (Tracy-Widom) *([36]) For empirical correlation matrices of size $N \times N$ of i.i.d. random variables with a finite fourth moment, the distance between the upper edge of the spectrum of the MP distribution b and the largest eigenvalue λ_{max} converges towards the Tracy-Widom distribution*

$$\text{Prob}\left(\lambda_{\max} \leq b + \gamma N^{-2/3} u\right) = F_1(u),$$

where γ in this case is given by $\gamma = \sqrt{c}\, b^{2/3}$.

$F_1(u)$ is the TW distribution, the probability distribution of the re-scaled eigenvalues of a random Hermitian matrix. We are interested in the type-1 distribution which holds for Gaussian orthogonal ensembles [15]. The distribution function can not be explicitly stated but relies on numerical approximations.
The TW distribution can be formulated, as well, for the singular values via the variable transform:

$$\text{Prob}\left(\gamma_{max} \leq \sqrt{b + \gamma N^{-2/3} u}\right) = F_1(u), \tag{2.7}$$

Since we always work with finite matrices in practice, we use the TW distribution to discriminate between singular values that belong to noise and signal, respectively. Specifically, we use $u = 1$ as a cutoff, so that $F_1(1) \approx 0.95$. In other words, there is a probability of 0.05 that a singular value bigger than $\sqrt{b + \gamma N^{-2/3}}$ is observed, if the matrix is entirely random. If N is very low, the MP distribution is not a good approximation anymore. For $N < 50$, we create an empirical distribution of noise-related singular values, by permuting the entries

of the measured expression matrix $\tilde{X}$. For each permutation we calculate the singular values and note the largest singular value. The 95th quantile of the distribution of the largest singular values across permutations is then taken to be the cutoff between singular values stemming from noise and signal respectively.

To discriminate random from non-random matrix components we can also look at the singular vectors [14]. Singular vectors that correspond to random components are "de-localized" and their elements have the following distribution:

$$f(\psi) = (1 - \psi^2)^{\frac{N-3}{2}}$$

If N is large, this distribution can be estimated by a Gaussian distribution with mean zero and variance $\frac{1}{N}$.

$$f(\psi) \sim \frac{N}{\sqrt{2\pi}} e^{\frac{-N\psi^2}{2}} \tag{2.8}$$

In order to distinguish localized from de-localized singular vectors, we can therefore assess the normality of the singular vectors. In our implementation we use a Shapiro-Wilk test. We assign singular vectors that obtain a p-value < 0.01 or are associated to singular values far from the bulk (the highest 50% of signal singular values) to real variability above the MP distribution.

Perturbation theory

As explained above, we model the observed expression matrix $\tilde{X}$ as a random matrix X perturbed by a deterministic signal matrix P. There is an important difference between the perturbation matrix P in equation 2.1 and the matrix $\tilde{X}^s$ in equation 2.6. $\tilde{X}^s$ does contain biologically relevant information, but is still influenced by the effects of random noise, whereas the matrix P consists of the pure signal without any added noise. The only case where these two matrices are identical is when the singular vectors of the noise matrix X and the perturbation matrix P are linearly independent, which is rarely the case. It is thus not possible to recover the unobserved, noise-free signal matrix by using those singular vectors that are associated with the highest singular values.

While it is not possible to reconstruct the signal matrix from measured data, perturbation theory [17] establishes a simple relationship between the singular value of the observed expression matrix $\tilde{X}$ and those of the signal matrix P. P is assumed to have finite rank r. Its singular value decomposition is thus:

$$P = \sum_{i=1}^{r} \theta_i u_i v_i^T, \text{ where } r \ll N, M$$

For scRNA-seq data, we only have to consider singular values $\theta_i > 0$, which means that $\tilde{X}$ potentially has singular values above the MP distribution. Thus, we only need to consider the largest singular values of $\tilde{X}$.

Theorem 3 (Largest Singular Value for MP) *([17]) The r largest singular values $\gamma_i(\tilde{X})$ of the $M \times N$ perturbed matrix $\tilde{X}$ exhibit the following behaviour as $M, N \longrightarrow \infty$ and $\frac{N}{M} \longrightarrow c$: For each fixed $1 \leq i \leq r$,*

$$\gamma_i(\tilde{X}) \xrightarrow{a.s.} \begin{cases} \sqrt{\frac{(1+\theta_i^2)(c+\theta_i^2)}{\theta_i^2}} & \text{if } i \leq r \text{ and } \theta_i > c^{1/4}, \\ b & \text{otherwise} \end{cases} \tag{2.9}$$

Moreover, for each fixed $i > r$, we have that $\gamma_i\left(\tilde{X}_n\right) \xrightarrow{a.s.} b$.

This theorem establishes a functional relationship between the largest singular values γ_i of the measured expression matrix and the singular values θ_i of the signal matrix P. Note that if θ_i is smaller than or equal to $c^{\frac{1}{4}}$, the corresponding γ_i will be equal to b, which is the upper limit of the MP distribution. In other words, if the perturbation (signal) is too small, the singular value spectrum of the observed expression matrix $\tilde{X}$ will be just the MP distribution and hence, no meaningful signal can be extracted.

From the above formula we are able to calculate the singular values of the perturbation matrix P. These are the values that describe the actual variances of the signal matrix without any contribution of the noise. This is achieved by calculating the inverse function

$$\theta_i(\gamma_i) \xrightarrow{\text{a.s.}} \begin{cases} \sqrt{\frac{2c}{\gamma_i^2 - (c+1) - \sqrt{(\gamma_i^2-(c+1))^2 - 4c}}} & \text{if } \gamma_i > b, \\ c^{\frac{1}{4}} & \text{otherwise} \end{cases} \tag{2.10}$$

Phiclust

Next, we want to establish how the singular vectors of $\tilde{X}$ depend on the perturbation P. In section 2.6 it is described that the elements of the singular vectors will follow a Gaussian distribution for a random matrix and large N. The elements of the singular vectors of the perturbation P are deterministic and correspond to biological variance. The following theorem describes the scalar product between the singular vector of the perturbation P and the perturbed matrix $\tilde{X}$.

Theorem 4 (Norm of Projection of Largest Singular Vectors for MP) *([17]) Let $\tilde{v}$ the right unit singular vectors of $\tilde{X}$. Then, the norm of projection of the right singular vector is given by*

$$|\langle \tilde{v}_i, v_i \rangle|^2 \xrightarrow{a.s.} \begin{cases} 1 - \frac{c(1+\theta_i^2)}{\theta_i^2(\theta_i^2+c)} & \text{if } \theta_i \geq c^{1/4} \\ 0 & \text{otherwise} \end{cases} \tag{2.11}$$

This theorem shows the same qualitative behavior as equation 2.9. If the singular value θ_i of the perturbation matrix is below the threshold of $c^{\frac{1}{4}}$, the scalar product is zero, indicating that the perturbed matrix $\tilde{X}$ has no relationship to the perturbation P. In other words, no relevant signal can be extracted. In the other limit, when the scalar product goes to 1, the

singular vectors of the perturbation P are perfectly aligned with the singular vectors of the perturbed matrix $\tilde{X}$. Thus, random noise has a negligible influence on the signal.

The scalar product given by $|\langle \tilde{v}_i, v_i \rangle|^2$ is identical to the squared cosine of the angle between the vectors:

$$\phi_{\text{clust}} = cos(\alpha)^2 = \left(\frac{\tilde{v} \cdot v}{\|v\| \, \|\tilde{v}\|} \right)^2 = (\tilde{v} \cdot v)^2 = |\langle \tilde{v}_i, v_i \rangle|^2.$$

This holds because the singular vectors are assumed to have norm 1.

We propose ϕ_{clust} (phiclust) as a measure of clusterability in scRNA-seq data. If, for a given cluster, there are no values above the MP distribution the signal of the perturbation matrix P can not be recognized any more and phiclust will be zero. If there are singular values above the MP distribution, phiclust evaluates how closely related the singular vectors of the expression matrix $\tilde{X}$ are to those of the perturbation matrix P.
We obtain a value of phiclust for each singular value that can be found above the MP distribution. Each of them indicates the signal-to-noise ratio for the variance that the corresponding singular vector explains. Thus, the more singular values are above the MP distribution, the more variances can be found in the data and it can be interpreted as proportional to the number of clusters. In the definition of phiclust, we have decided to use the maximum of all angles, thus indicating the maximal clusterability that can be achieved from clustering.

G-phiclust
In accordance with the above definition of phiclust (2.6), we can also define the clusterability, or signal-to-noise ratio, for the gene space. The following theorem describes the equation.

Theorem 5 (Norm of Projection of Largest Singular Vectors for MP) *([17]) Let $\tilde{u}$ be the left unit singular vectors of $\tilde{X}$. Then, the norm of projection of the left singular vector by*

$$|\langle \tilde{u}_i, u_i \rangle|^2 \xrightarrow{a.s.} \begin{cases} 1 - \frac{\left(c+\theta_i^2\right)}{\theta_i^2(\theta_i^2+1)} & \text{if } \theta_i \geq c^{1/4} \\ 0 & \text{otherwise} \end{cases} \tag{2.12}$$

For the gene singular vector, ϕ^g_{clust} (g-phiclust) indicates how closely the variance among genes is related to the original variance in the perturbation matrix P. For each singular vector, the variance-driving genes correspond to those with the highest absolute loading in the corresponding gene singular vector. Cells with high positive or negative entries in the singular vector have high expression of genes with large positive or negative entries in the corresponding gene singular vector, respectively. This relationship is not a replacement for the calculation of differentially expressed genes, but merely indicates the genes that drive the variance across cells for each singular vector. Based on the value of g-phiclust, it is possible to evaluate how accurate the determination of differentially expressed genes will be. With a low signal-to-noise ratio, it is more likely to obtain genes differentially expressed that can be attributed to noise. As well as for phiclust, we obtain several angles,

one for each singular value above the MP distribution. Thus, genes driving the variances in gene singular vectors with a higher g-phiclust are more accurate. We decided, to be consistent, to define g-phiclust as the highest squared cosine of the angle.

Uncertainty of phiclust

The theory presented above holds as the expected value in the infinite limit, however we do not know about the variations within the finite limit. To address this, we constructed a confidence interval for the values of phiclust using the following sampling approach. The basic idea is to approximate the signal matrix P and add new realizations of the noise matrix by sampling from a random distribution. The standard deviation is then constructed from the values of phiclust calculated for this ensemble of sampled matrices.

First, the matrix $\tilde{X}$ is pre-processed as described in section 2.6. By applying the MP distribution, we then determine the singular values associated with signal and noise. We decompose the simulated or measured expression matrix $\tilde{X}$ into a noise matrix X^r and a matrix X^s that contains deterministic structure (see equation 2.6).

Then, we estimate the first two moments of X^r, which due to the pre-processing of the measured expression matrix are equal to a mean of 0 and a standard deviation of 1. It is thus possible, given the universality property of the MP distribution, to sample a new noise matrix X with the same two first moments (mean = 0 and variance = 1) from a normal distribution.

To approximate the perturbation matrix, we use the singular values λ_i of X^s to calculate the expected singular values θ_i of the perturbation matrix based on equation 2.10. We replace the singular values λ_i of the matrix X^s with those of the perturbation matrix θ_i and call it P^s. In this way we have created a perturbation matrix with the expected singular values θ_i and unit singular vectors. Note that P^s contains noise and is thus different from the signal matrix P. Luckily, low rank-perturbation theory is independent of the exact distribution of the signal singular vectors.

Together, we obtain a sample measurement matrix (Step 1):

$$\tilde{X}^* = X + P^s.$$

We next calculate the values phiclust of $\tilde{X}^*$ (Step 2). By sampling new values for the noise matrix X several times (~ 50), and repeating step 1 and 2, we are now able to estimate the influence of random variations, in finite limits, on the additive perturbation and thus on phiclust.

We can subsequently calculate the upper ϕ^{up}_{clust} and lower $\phi^{down}_{\text{clust}}$ standard deviation as follows. Let k be the number of values above the original value ϕ^*_{clust} and N the total number of sampled values then

$$\phi^{up}_{\text{clust}} = \left(\frac{1}{k-1} \sum_{\phi^*_{\text{clust}} \geq \phi_{\text{clust}}} (\phi^*_{\text{clust}} - \phi_{\text{clust}})^2 \right)^{1/2} \tag{2.13}$$

$$\phi^{down}_{\text{clust}} = \left(\frac{1}{N-k-1} \sum_{\phi^*_{\text{clust}} < \phi_{\text{clust}}} (\phi^*_{\text{clust}} - \phi_{\text{clust}})^2 \right)^{1/2} \tag{2.14}$$

are the upper and lower boundaries of the interval.

Clusterability

Assessing clustering quality

We use two different methods to assess clustering quality, the adjusted rand index (ARI) and the silhouette coefficient.

Assuming two partitions, A and B, of a set of N cells, the rand index is defined as[21]:

$$RI(A,B) = \frac{N_{11} + N_{00}}{\binom{N}{2}},$$

where N_{11} is the number of pairs of elements that are in the same cluster in A and in the same cluster in B. N_{00} is the number of pairs of elements that are in a different cluster in A and in a different cluster in B. The rand index takes values between 0 and 1, where 0 indicates the complete lack of agreement between the partitions and 1 would indicate identical partitions. Even a random clustering of elements produces a non-zero rand index. The ARI is defined in such a way, that its value is on average 0 for a pair of partitions with randomly permuted cluster labels. A positive ARI thus indicates that partitions agree more than expected to happen by random chance. Let partition A have K_A clusters of sizes a_i and partition B have K_B clusters of sizes b_j, then the adjusted rand index is defined as:

$$ARI(A,B) = \frac{RI(A,B) - \mathbb{E}[RI(A,B)]}{1.0 - \mathbb{E}[RI(A,B)]} = \frac{\binom{N}{2}\sum_{k,m=1}^{K_A K_B}\binom{n_{km}}{2} - \sum_{m=1}^{K_A}\binom{a_k}{2}\sum_{m=1}^{K_B}\binom{b_m}{2}}{\frac{1}{2}\binom{N}{2}\left[\sum_{k=1}^{K_A}\binom{a_k}{2}\sum_{m=1}^{K_B}\binom{b_m}{2}\right] - \sum_{k=1}^{K_A}\binom{a_k}{2}\sum_{m=1}^{K_B}\binom{b_m}{2}}$$

For synthetic data, we take a high ARI between a clustering and the ground truth partition to indicate a clustering of high quality.

Another useful measure for clustering quality is the silhouette coefficient. Let $a(i)$ be the mean distance from point i to all other data points in the same cluster and $b(i)$ be the mean distance from point i to all other points from different clusters, then the silhouette coefficient is defined as [8]:

$$s(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}}.$$

For the calculation of the distance, we consider the euclidean distance metric in the space spanned by the singular vectors that are associated with singular values above the MP distribution of the expression matrix $\tilde{X}$ (see 2.6). The final silhouette coefficient is taken as the mean value over all data points. For the calculation of the silhouette coefficient we use the *cluster R package (V 2.1.0).*

Theoretically achievable clustering quality

A perfect clustering would coincide with the ground truth and obtain an ARI of 1. Here we argue that such a perfect clustering is in general not achievable, if there is noise in the data. In other words there is always a finite Bayes error rate (also called irreducible error) for assigning cells to the appropriate cluster. To construct a Bayes classifier, which achieves the minimal error rate, we need to know the ground truth partition. Hence, we use simulated data. For each ground truth cluster, we fit a multidimensional Gaussian to

the elements of the singular vectors of the expression matrix $\tilde{X}$ that correspond to the cells in the respective cluster (see Additional file 1: Fig. S3a). We only consider singular vectors with singular values above the MP distribution. For the fit we use the *mclust R package (V 5.4.6)*. We then construct a classifier by assigning a cell to the cluster for which it has the highest value of the fitted Gaussian distribution. This corresponds to the best clustering one can achieve if the ground truth partition is known. We define the theoretically achievable adjusted rand index (tARI) as the ARI between this best achievable clustering and the ground truth partition. Similarly, we define the theoretically achievable silhouette coefficient (tSIL) as the silhouette coefficient of the best achievable clustering. Since we use the fitted Gaussian distributions instead of the actual (unknown) distribution of singular vector elements, the constructed classifier only approximates the Bayes classifier. However, we confirmed empirically, that the tARI defined above is an upper bound for all tested clustering methods, which comprises the currently most popular tools used for single-cell RNA-seq data [Additional file 1: Fig. S3 b, c].
The tARI embodies our notion of clusterability. We define high clusterability as a low Bayes error rate for cluster assignments, which corresponds to a high tARI. We show empirically that our clusterability measure is a proxy of the tARI and thus a way to assess clusterability without knowing the ground truth [Fig. 1d].

Application to single-cell RNA-seq data

Preprocessing of scRNA-seq data

In the following the necessary preprocessing steps for the application of the clusterability measure for scRNA-seq data are described.

Transcriptome Mode

The largest eigenvalue λ_1 of an expression matrix is typically much larger than all the other singular values and its corresponding singular vector has entries of equal sign, which often have similar magnitude (of order $\frac{1}{\sqrt{N}}$, which is the ideal value in the perfectly homogeneous case). This singular vector reflects a general, global trend in the data. This structure has been observed for many empirical data matrices. (In time series analysis of the stock market, this singular vector is called the "market mode" since it corresponds to a trend that is common across many stocks [19]). Here, we refer to this singular vector as "transcriptome mode" since it reflects a trend that is shared across the whole transcriptome (see Additional file 1: Fig. S2 a-d). In order to reduce the influence of this singular value on the calculation of the MP fit, we center the expression matrix $\tilde{X}$ gene-wise. As a result, the singular value of the transcriptome mode will be reduced to a value close to 0.

Normalization

The efficiency of the capture of transcripts and their conversion to cDNA is known to be highly variable between cells. Hence, single-cell gene expression data is usually normalized cell-wise. We have tested several normalization methods but none of them seemed sufficient to remove all technical variability in the data. Thus, in section 2.6 we describe a method to reduce these effects for our clusterability measure phiclust. Nevertheless, we normalize the expression to the total counts per cell and subsequently log-transform to stabilize the variance.

Gene distribution
Gene expression is typically modelled by a Poisson, negative binomial or zero inflated negative binomial distribution. However, the parameters of these distributions differ between genes, this violates the assumptions of the MP theorem, where all values are sampled from the same distribution. In practice, gene-wise standardization to a mean of 0 and standard deviation of 1 mostly circumvents this problem. Additionally, we have observed that there is a bias resulting from variations in cells. These biases are as well reduced by standardising the cells to a mean of 0 and standard deviation of 1 (see Additional file 1: Fig. S2 c,d). This is equivalent to calculating the eigenvalues and vectors of a correlation matrix instead of a covariance matrix.

Zero inflation
Another factor to be considered is the large amount of zero values in scRNA-seq data. These zeros might be on the one hand due to technical artefacts (low efficiency, dropout) or simply due to low, stochastic gene expression. After performing the above mentioned preprocessing steps we mostly do not observe deviations from the MP distribution. However, this is a known problem discussed within the framework of sparsity induced singular values. For single cell RNA-seq data an extensive analysis has been performed in [14], where the authors observe deviations from the MP distribution caused by sparsity. The authors suggest the exclusion of outlier genes that can be identified through the fit of the MP distribution. For phiclust we do not use this preprocessing step, however we do exclude genes that have a high expression in only a few number of cells.

Regressing out unwanted sources of variability (Confounder Regression)
scRNA-seq data suffers from several sources of technical variability that can obscure or even be mistaken for relevant biological signal. One of the most important of these is the variable efficiency of mRNA capture and cDNA conversion. The total number of detected transcripts per cell is typically taken as a proxy of this efficiency. There are also biological processes that can cause unwanted signal. Most cells are stressed due to the tissue dissociation necessary for single-cell library preparation. The percentage of expression coming from mitochondrial genes or the expression of marker genes for stress can be used to estimate the level of stress. Different metabolic states of cells might be reflected in the level of ribosomal gene expression and many genes fluctuate with the cell cycle. Here, we seek to establish a method to remove any effect of these nuisance variables on the clusterability measure.
We model the signal matrix P as a sum of relevant signal B and unwanted signal due to nuisance variables Y. Inspired by published approaches to expression data normalization [23, 26], we model the influence of Y by linear regression. This is a valid approach because the regression is performed on the singular vectors of $\tilde{X}$, which contain Gaussian distributed noise. Given the singular value decomposition of $\tilde{X}$ and singular vectors $\tilde{v}_i$,

$$\tilde{v}_i = \beta Z, \quad \text{with } \beta \in \mathbb{R}^k \tag{2.15}$$

where $Z \in \mathbb{R}^{N \times k}$ is a matrix of covariates, such as the total counts per cell, with k the number of covariates and N the number of cells. Each covariate is normalized to a length

of 1 such that the range agrees with the range of the singular vectors. The amount of variance explained by the nuisance parameters is then given by the value of the adjusted R squared (R^2_{adj}) of this linear regression. Since the eigenvalues of the cell-cell correlation matrix can be interpreted as the amount of variance explained, we reduce the eigenvalues λ_i by $\tilde{\lambda}_i = (1 - R^2_{adj})\lambda_i$. In the next step, we calculate adjusted singular values by $\tilde{\gamma}_i = \sqrt{\tilde{\lambda}_i}$ and use these adjusted singular values $\tilde{\gamma}_i$ for the consecutive steps in the calculation of the clusterability measure.

Algorithm

The procedure to obtain the clusterability measure involves the following steps:

1. Preprocess the single cell expression matrix as described in section 2.6:
   (a) Normalization
   (b) Log-transformation
   (c) Standardization gene-wise
   (d) Standardization cell-wise
2. Calculate the singular value decomposition of the gene expression matrix $\tilde{X}$.
3. Fit the MP distribution to the singular values (equation 2.5).
4. Determine singular values/vectors that correspond to non-random variability using the Tracy-Widom distribution (equation 2.7) or the Shapiro-Wilk test (equation 2.8), respectively.
5. Adjust the singular values for effects of nuisance variables by linear regression (equation 2.15).
6. Calculate the singular values θ_i of the signal matrix P using the inverse of equation 2.9, given by 2.10.
7. Calculate the projections of the singular vectors of the expression matrix $\tilde{X}$ on the corresponding singular vector of the signal matrix P with equations 2.11 for the singular vectors and 2.12 for the gene singular vectors.
8. The clusterability measure is the largest of the projections for the singular vectors obtained in the previous step.

2.7 Supplementary Note 2

Application of phiclust to our previously published single-cell RNA-sequencing study of the human fetal kidney [31] revealed two distinct groups of clusters (Fig. 3a). Connecting tubule (CnT), nephron progenitor cells-a (NPCa), nephron progenitor cells-b (NPCb), and mesangial cells (Mes) all obtained a phiclust of 0, which signified that these clusters consist of pure populations with homogeneous gene-expression profiles. The rest of the clusters obtained higher values of phiclust, indicating that they contained subpopulations that were previously overlooked. For further analysis, we explored all clusters with highest phiclust and chose to further investigate clusters in which new cell populations were identified: the ureteric bud/collecting duct (UBCD), the S-Shaped Body proximal precursor cells (SSBpr), and the Interstitial cells a (ICa), with phiclust of 0.97, 0.95, and 0.93, respectively.

UBCD The analysis of this cluster yielded two clearly separate subpopulations (Fig. 3b, Additional file 1: Fig. S10b). The bigger subpopulation contained developing collecting duct cells and their precursors (ureteric bud), indicated by the expression of genes such as WFDC2, AQP2, CLDN3, MMP7, and CALB1. In contrast, the smaller sub-cluster showed little or no expression of the aforementioned genes and was characterized by UPK1A and UPK1B, well-known markers of the urothelial epithelium, which constitutes the inner lining of the ureter. The presence of such cells in our data is plausible given that the whole fetal kidney was used in our sequencing experiment. Both DE analysis (Table S4) and inspection of the top variance-driving genes (Additional file 1: Fig. S10e) revealed SPINK1, UPK2, S100A6, KRT7, and KRT19 as additional markers. Staining of week 15 fetal kidney sections with UPK1A and KRT7 antibodies confirmed our interpretation (Fig. 3c, Additional file 1: Fig. S11a). UPK1A was restricted to the superficial urothelial cells in major and minor calyces as well as the developing ureter. KRT7 was expressed more broadly, across the superficial, intermediate, and basal urothelium. Both KRT7 and UPK1A were completely absent from the whole collecting system and the branching ureteric bud, marked by CDH1 (Additional file 1: Fig. S11a).

SSBpr Sub-clustering the SSBpr population showed the presence of 3 subpopulations (Fig. 3b, Additional file 1: Fig. S10b). One subpopulation contained markers of proximal cell precursors (GPC3, LHX1, CADM2) together with low expression of AMN and APOE (see Table S4), which is consistent with the original annotation of the cluster. A second subpopulation, contiguous to the previous one, showed the expression of CLDN1, which is expressed in the proximal epithelium, together with CITED2, expressed in developing podocytes. This suggested parietal epithelial cells (PECs) as the most likely cell type, as these cells were reported to share several markers with both proximal epithelium and podocytes [45]. To confirm this interpretation, we performed Immunostaining of CLDN1, as well as CAV2 and AKAP12 which were found by DE analysis (Fig. 3d, Additional file 1: Fig. S11b). Interestingly, CLDN1 was found in all segments of the S-shaped body except in the precursors of the PECs, which are the thin layer of cells at the lateral side of the proximal segment of the SSB. CLDN1 appeared in the parietal epithelium only at the capillary loop stage and continued to be expressed in all PECs in more mature glomeruli. CAV2 was present in the parietal epithelium in developing glomeruli, but also in the endothelial cells of both the glomerular capillaries and the surrounding vasculature. Intriguingly, CAV2

overlapped with CLDN1 only in a subpopulation of PECs in individual glomeruli, which might indicate previously unobserved heterogeneity within these cells in the developing kidney. Only AKAP12 marked the precursors of the PECs in S-shaped bodies and continued to be abundantly expressed. However, AKAP12, was not specific to PECs, as it was also expressed in interstitial cells in the cortex. Finally, a third, small and distinct subpopulation in the SSBpr cluster expressed distal tubule markers (SPP1, ODC1, IRX3, and S100A10), suggesting that these cells were misclassified during the original clustering. This shows that phiclust can pinpoint clustering errors, making it a useful tool for clustering quality control.

ICa This cluster consisted of 5 subpopulations (Fig. 3b, Additional file 1: Fig. S10b). All subpopulations expressed markers of the renal interstitium. One also expressed genes found uniquely expressed in other cell types (EPCAM, CD24, BST2, NNAT, DAPL1) and thus likely contains doublets. Another small subset was characterized by markers of mesangial cells (MGP, ACTA2, PDGFRB), suggesting that it contains mesangial cells erroneously grouped with the ICa or renal pericytes, which share a similar gene expression profile [46]. Another subpopulation showed high expression of non-specific genes related to components and regulators of microtubules together with metabolic, mitochondrial and stress-related genes (H2AFZ, TUBA1B, TYMS, STMN1, DUT, MT-CO3, MT-ND5). The two remaining subpopulations were clearly interstitial but their gene expression profiles could not be linked to known interstitial populations, likely due to the dearth of knowledge about the renal stroma. We hypothesized that these two subpopulations were localized in different regions of the kidney. To test this idea, we stained fetal kidney sections with POSTN, CLDN11, and SULT1E1 (Fig. 3e, Additional file 1: Fig. S11c), which were identified by DE analysis and inspection of the top variance-driving genes. SULT1E1 was highly expressed in the pelvic area in the immediate vicinity of the developing ureter, as well as the inner and outer medulla, preferentially surrounding tubules. This marker might thus indicate the medullary interstitium as well as pelvic smooth muscle cells. Staining with CLDN11 showed a higher signal in the medulla and papilla, similar to SULT1E1, but with a wider spatial distribution. In contrast to SULT1E1, CLDN11 was also expressed in groups of cortical interstitial cells, situated directly underneath the renal capsule, in the nephrogenic zone. CLDN11 might thus also be expressed by the interstitial progenitor cells or their immediate progeny. Lastly, POSTN was mainly found in the renal cortex surrounding tubules and glomerular microvasculature. POSTN was also expressed in cortical blood vessels with larger diameters together with their arborizations. POSTN is a secreted extracellular matrix protein known to be expressed in cardiac smooth muscle cells, as well as connective tissues. Here, POSTN might mark smooth muscle cells of the cortical vasculature.

In conclusion, a reanalysis of our previously published data showed the ability of phiclust to reveal overlooked subpopulations. Interestingly, phiclust identified sub-clusters with only a few cells (41 developing PECs, 68 urothelial cells, 29 distal cells), highlighting its sensitivity to relevant substructure hidden within a bigger cluster. Finally, phiclust was also useful to pinpoint clustering errors and the presence of doublets, which makes it useful for quality control prior to DE analysis.

References

[1] M. Mircea et al. Phiclust: a clusterability measure for single-cell transcriptomics reveals phenotypic subpopulations. *Genome Biology*, 23(1):1–24, dec 2022.

[2] R. Satija et al. Spatial reconstruction of single-cell gene expression data. *Nature Biotechnology*, 33(5):495–502, may 2015.

[3] F. A. Wolf, P. Angerer, and F. J. Theis. SCANPY: Large-scale single-cell gene expression data analysis. *Genome Biology*, 19(1):15, feb 2018.

[4] J. A. Hartigan and M. A. Wong. Algorithm AS 136: A K-Means Clustering Algorithm. *Applied Statistics*, 28(1):100, 1979.

[5] F. Murtagh and P. Legendre. Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion? *Journal of Classification*, 31(3):274–295, oct 2014.

[6] V. Y. Kiselev, T. S. Andrews, and M. Hemberg. Challenges in unsupervised clustering of single-cell RNA-seq data. *Nature Reviews Genetics*, 20(May):273–282, 2019.

[7] S. Ackerman, Margareta; Ben-David. Clusterability : A Theoretical Study. In M. van Dyk, David; Welling, editor, *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics*, volume 5, pages 1–8. PMLR, 2009.

[8] P. J. Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987.

[9] A. Adolfsson, M. Ackerman, and N. C. Brownstein. To cluster, or not to cluster: An analysis of clusterability methods. *Pattern Recognition*, 88:13–26, apr 2019.

[10] B. Liu et al. An entropy-based metric for assessing the purity of single cell populations. *Nature Communications*, 11(1):1–13, dec 2020.

[11] D. Paul and A. Aue. Random matrix theory in statistics: A review. *Journal of Statistical Planning and Inference*, 150:1–29, jul 2014.

[12] M. Potters, J.-P. Bouchaud, and L. Laloux. Financial Applications of Random Matrix Theory: Old Laces and New Pieces. *Acta Physica Polonica*, 35(9):2767–2784, 2005.

[13] O. Bohigas, M. J. Giannoni, and C. Schmit. Characterization of Chaotic Quantum Syectra and Universality of Level Fluctuation Laws. *Physical Review Letters*, 52(1):1–4, 1984.

[14] L. Aparicio, M. Bordyuh, A. J. Blumberg, and R. Rabadan. A Random Matrix Theory Approach to Denoise Single-Cell Data. *Patterns*, 1(3), 2020.

[15] G. Livan, M. Novaes, and P. Vivo. *Introduction to Random Matrices Theory and Practice*. Springer, Switzerland, 2018.

[16] C. A. Tracy and H. Widom. Level-spacing distributions and the Airy kernel. *Communications in Mathematical Physics*, 159(1):151–174, jan 1994.

[17] F. Benaych-Georges and R. R. Nadakuditi. The singular values and vectors of low rank perturbations of large rectangular random matrices. *Journal of Multivariate Analysis*, 111:120–135, 2012.

[18] E. P. Wigner. Characteristic Vectors of Bordered Matrices With Infinite Dimensions. Technical Report 3, 1955.

[19] M. Macmahon and D. Garlaschelli. Community Detection for Correlation Matrices. *Physical Review X*, 021006(5):1–34, 2015.

[20] L. Zappia, B. Phipson, and A. Oshlack. Splatter: Simulation of single-cell RNA sequencing data. *Genome Biology*, 18(1):174, sep 2017.

[21] A. J. Gates and Y.-Y. Ahn. The Impact of Random Models on Clustering Similarity. Technical report, 2017.

[22] K. Fukunaga. *Introduction to Statistical Pattern Recognition*. Elsevier, San Diego, 2 edition, 1990.

[23] C. Hafemeister and R. Satija. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biology*, 20(1):296, dec 2019.

[24] L. Tian et al. experiments. *Nature Methods*, 16(June), 2019.

[25] D. Risso et al. A general and flexible method for signal extraction from single-cell RNA-seq data. *Nature Communications 2018 9:1*, 9(1):1–17, jan 2018.

[26] D. Grün. Revealing dynamics of gene expression variability in cell state space. *Nature Methods*, 17(1):45–49, jan 2020.

[27] W. Saelens, R. Cannoodt, H. Todorov, and Y. Saeys. A comparison of single-cell trajectory inference methods. *Nature Biotechnology*, 37(5):547–554, may 2019.

[28] T. Stuart and R. Satija. Integrative single-cell analysis. *Nature Reviews Genetics*, 20(5):257–272, jan 2019.

[29] F. V. Mello et al. Maturation-associated gene expression profiles during normal human bone marrow erythropoiesis. *Cell Death Discovery*, 5(1):69, dec 2019.

[30] A. C. Villani et al. Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. *Science*, 356(6335), apr 2017.

[31] M. Hochane et al. Single-cell transcriptomics reveals gene expression dynamics of human fetal kidney development. *PLOS Biology*, 17(2):e3000152, feb 2019.

[32] B. Adamson et al. A Multiplexed Single-Cell CRISPR Screening Platform Enables Systematic Dissection of the Unfolded Protein Response. *Cell*, 167(7):1867–1882.e21, dec 2016.

[33] S. Bullett, T. Fearn, F. Smith, and I. E. Smolyarenko. An Introduction to Random Matrix Theory. *Advanced Techniques in Applied Mathematics*, pages 139–171, 2016.

[34] R. Mingo, James A; Speicher. *Free Probability and Random Matrices*. Springer New York LLC, 1 edition, 2017.

[35] K. Kendall and M. George. Kolmogorov–Smirnov Test. *The Concise Encyclopedia of Statistics*, pages 283–287, feb 2008.

[36] J. Bun, J.-p. Bouchaud, and M. Potters. Cleaning large correlation matrices : Tools from Random Matrix Theory. *Physics Reports*, 666:1–109, 2017.

[37] W. Haynes. Benjamini–Hochberg Method. *Encyclopedia of Systems Biology*, pages 78–78, 2013.

[38] F. Benaych-Georges and R. R. Nadakuditi. The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices. *Advances in Mathematics*, 227(1):494–521, 2011.

[39] L. Scrucca, M. Fop, T. B. Murphy, and A. E. Raftery. Mclust 5: Clustering, classification and density estimation using Gaussian finite mixture models. *R Journal*, 8(1):289–317, 2016.

[40] Z. Ji and H. Ji. TSCAN: Pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis. *Nucleic Acids Research*, 44(13):e117, jul 2016.

[41] I. Tirosh et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science*, 352(6282):189–196, apr 2016.

[42] A. T. Lun, K. Bach, and J. C. Marioni. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biology*, 17(1):1–14, apr 2016.

[43] S. C. van den Brink et al. Single-cell sequencing reveals dissociation-induced gene expression in tissue subpopulations. *Nature Methods*, 14(10):935–936, 2017.

[44] L. McInnes, J. Healy, and J. Melville. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *Journal of Open Source Software*, 3(29):861, feb 2018.

[45] S. S. Guhr et al. The expression of podocyte-specific proteins in parietal epithelial cells is regulated by protein degradation. *Kidney International*, 84(3):532–544, sep 2013.

[46] X. Wang et al. Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science*, 361(6400), jul 2018.

3

ETV2 upregulation marks the specification of early cardiomyocytes and endothelial cells during co-differentiation

The ability to differentiate human induced pluripotent stem cells (hiPSCs) efficiently into defined cardiac lineages, such as cardiomyocytes and cardiac endothelial cells, is crucial to study human heart development and model cardiovascular diseases in vitro. The mechanisms underlying the specification of these cell types during human development are not well-understood, which limits the fine-tuning and broader application of cardiac model systems. Here, we used the expression of ETV2, a master regulator of hematoendothelial specification in mice, to identify functionally distinct subpopulations during the co-differentiation of endothelial cells and cardiomyocytes from hiPSCs. Targeted analysis of single-cell RNA sequencing data revealed differential ETV2 dynamics in the two lineages. A newly created fluorescent reporter line allowed us to identify early lineage-restricted states and show that a transient ETV2-high state initiates the specification of endothelial cells. We further demonstrated that functional cardiomyocytes can originate, unexpectedly, from progenitors expressing ETV2 at a low level. Our study thus sheds light on the in vitro differentiation dynamics of two important cardiac lineages.

This chapter is based on Xu Cao, **Maria Mircea**, Gopala Krishna Yakala, Francijna E. van den Hil, Hailiang Mei, Christine L. Mummery, Stefan Semrau, and Valeria V. Orlova. ETV2 upregulation marks the specification of early cardiomyocytes and endothelial cells during co-differentiation. Manuscript submitted.

Significance statement

In vitro differentiation of cardiac cell types is of great importance for disease modeling and future regenerative medicine. Many of the relevant molecular mechanisms are currently not understood, which limits the efficiency and fine-tuning of existing differentiation protocols. Here, we focus on the master regulator ETV2 and show that its upregulation marks the specification of two cardiac cell types during co-differentiation. Using single-cell RNA-seq and a new fluorescent reporter line, we identify lineage-restricted subpopulations in the ETV2+ cells. Our study is the first to resolve ETV2 dynamics at the single-cell level in the context of in vitro human cardiac differentiation.

3.1 Introduction

In vivo, cardiomyocytes (CMs) and endothelial cells (ECs) originate from Mesp1+ progenitors specified during gastrulation. In mice, these cells appear in the primitive streak and subsequently migrate towards the lateral plate mesoderm around E6.5 [1–4]. It is still controversial when the segregation of CMs and ECs from their common progenitor occurs. Single-cell RNA-seq (scRNA-seq) of mouse Mesp1+ progenitors collected at E6.75 and E7.25 showed that these cells were already segregated into distinct cardiovascular lineages, including CMs and ECs [5]. However, other studies showed that multipotential progenitors were still present in Flk-1-expressing lateral plate mesoderm [6, 7]. These cells were the first to be recognized as multipotent cardiac progenitor cells (CPCs) [8]. Studies in mouse and chick showed that CPCs come from two different sources [9, 10]: the first and second heart field (FHF, SHF). The FHF in the cardiac crescent contributes to the primitive heart tube, which serves as a scaffold into which SHF cells can migrate before heart chamber morphogenesis. It has been shown that cells from the SHF are patterned before migration to give rise to different parts of the heart [2, 11]. CPCs from FHF and SHF can be distinguished by the expression of ISL1, which is specifically expressed in SHF [12]. NKX2-5 expressing CPCs in both FHF and SHF from E7.5 to E7.75 contribute to both CMs and ECs in the heart [13]. As a direct target of NKX2-5, ETV2 was found to be expressed in all ECs but not myocardium by E8.5 [14]. ETV2 was required for the development of endothelial and hematopoietic lineages and directly targets TAL1, GATA2, LMO2, TEK, NOTCH1, NOTCH4, and CDH5 [14–17]. In mouse embryonic stem cells (ESCs), VEGF-FLK1 signaling upregulates ETV2 expression to induce hemangiogenic specification via an ETV2 threshold-dependent mechanism [18]. ETV2 expression was also found to direct the segregation of hemangioblasts, and smooth muscle cells (SMCs) in mouse ESCs [19].
In human heart development, much less is known about the specification of endothelial and myocardial lineages from multipotent CPCs, both in terms of timing and gene regulatory mechanisms. More specifically, it is still unclear whether ETV2 also plays a role in the segregation of ECs and CMs from CPCs in humans. Overexpression of ETV2 converts human fibroblasts into endothelial-like cells [20] and ETV2 expression levels were manipulated in several studies to drive hiPSCs towards ECs in 2D, and 3D cultures [21–27]. Paik et al. performed scRNA-seq analysis of hiPSC-derived ECs (hiPSC-ECs), which made up less than 10% of the cells that expressed the cardiac maker TNNT2. The developmental dynamics of ECs and cardiac lineages as such were not further studied [28].

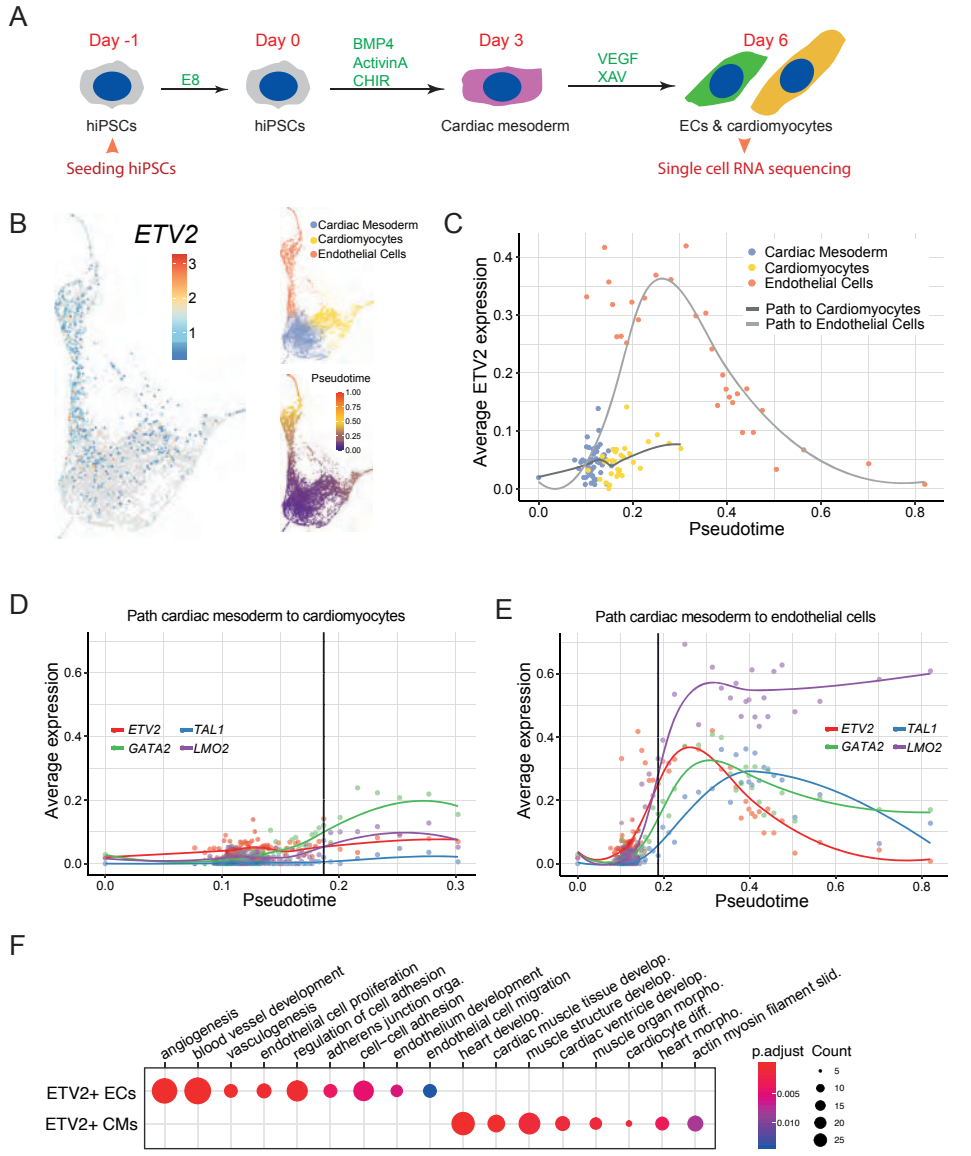


Figure 3.1: scRNA-seq analysis of EC and CM co-differentiation reveals transient ETV2 upregulation after bifurcation. (A) Schematic overview of the co-differentiation protocol from day -1 to day 6. Cells were collected for scRNA-seq on day 6. (B) Two-dimensional representation of the scRNA-seq data. Each data point is a single cell. Left: log2 transformed ETV2 expression is indicated by color. Top right: Cell identities are labeled with different colors. Bottom right: Pseudotime is indicated by color. (C) Average expression of ETV2 in bins of pseudotime for the developmental path of CMs or ECs . Cell identities are labeled with different colors. (D-E) Average expression of ETV2 and its direct target genes TAL1, GATA2, LMO across binned pseudotime along the developmental path of CMs (D) or ECs (E). Threshold (indicated in black) is set to the timepoint where the average ETV2 expression in EC reaches 0.25. (F) GO enrichment analysis of genes that were differentially expressed between ETV2+ ECs and ETV2+ CMs in the scRNA-seq dataset. 128 and 136 genes were upregulated in ETV2+ ECs and CMs respectively (Padjusted < 0.05). A complete list of GO terms can be found in Table S3. Color represents the Padjusted of the enrichment analysis and dot size represents the count of genes mapped to the GO term.

In a scRNA-seq study of hiPSC-ECs created with another differentiation protocol [29], ECs were collected at multiple time points. This study showed that endothelial and mesenchymal lineages had a common developmental origin in mesoderm cells, but the identity and differentiation potential of these cells was not characterized.
Previously, we found that MESP1+ progenitors derived from human ESCs could give rise to CMs, ECs and SMCs [30, 31]. We also developed a co-differentiation system for ECs and CMs from hiPSCs through a common cardiac mesoderm precursor [32]. Here we performed scRNA-seq analysis of this co-differentiation system to elucidate the relationship between ETV2 expression and specification of ECs and CMs from cardiac mesoderm. ETV2 expression was observed principally in the form of an initial pulse in the endothelial lineage but also in a subpopulation of the myocardial lineage. Using a newly generated $ETV2^{mCherry}$ hiPSC reporter line, we purified two subpopulations of ETV2+ cells and revealed their respective EC and CM expression characteristics by bulk RNA-seq. These sorted populations also showed distinct differentiation potentials towards CMs and ECs upon further differentiation with VEGF. In summary, this study depicted ETV2 dynamics during the segregation of human CMs and ECs differentiated from hiPSCs.

3.2 Results

3.2.1 ETV2 is upregulated after a bifurcation into CMs and ECs

To characterize the expression of ETV2 during co-differentiation of ECs and CMs [32] (Fig. 3.1A), we collected scRNA-seq data on day 6 of differentiation (Fig. 3.1B). We identified three distinct clusters: cardiac mesoderm, CMs, and ECs (Fig. 3.1B, top right and Table S3). Pseudotime analysis revealed cardiac mesoderm as the common developmental origin of CMs and ECs (Fig. 3.1B, bottom right). We found ETV2 to be highly expressed in the EC cluster, as well as a small fraction of cells in the cardiac mesoderm and CM clusters (Fig. 3.1B, left). We next focused on ETV2 expression dynamics along the developmental path from cardiac mesoderm to CMs and ECs. Notably, ECs extended to larger pseudotimes (0.15-0.8) compared to CMs (0.15-0.3), which might indicate faster differentiation dynamics in the EC lineage (Fig. 3.1C, 3.2A). After the bifurcation into ECs and CMs (around pseudotime 0.15), ETV2 increased only slightly in the CM lineage. In the EC lineage, however, it was initially strongly upregulated (until pseudotime 0.25), and subsequently declined to a similar level as in cardiac mesoderm (Fig. 3.1C). ETV2 downstream target genes, such as TAL1, GATA2, and LMO2 [17], were only slightly increased in the CM lineage (Fig. 3.1D), while in the EC lineage, they were highly induced and strongly correlated with ETV2 (Fig. 3.1E). Notably, TAL1, GATA2, and LMO2 only showed significant expression after ETV2 expression exceeded 0.25 in ECs, an expression level that was not reached in CMs (Fig. 3.2B-C). Endothelial specific genes KDR, CD34, SOX17, CDH5, and PECAM1 increased on the path from cardiac mesoderm to ECs (Fig. 3.2E). Most of these genes started to increase when ETV2 was already declining, as exemplified by CDH5 (Fig. 3.2D). Genes related to cardiac or muscle functionalities, like ACTC1, PDLIM5, HAND1, PKP2, and GATA4, most of which were already expressed in the cardiac mesoderm, were further increased in the CM lineage (Fig. 3.2F). Identification of genes that are differentially expressed between ETV2+ CMs and ECs showed enrichment in CM- and EC-specific genes, respectively (Fig. 3.1F, Table S4).

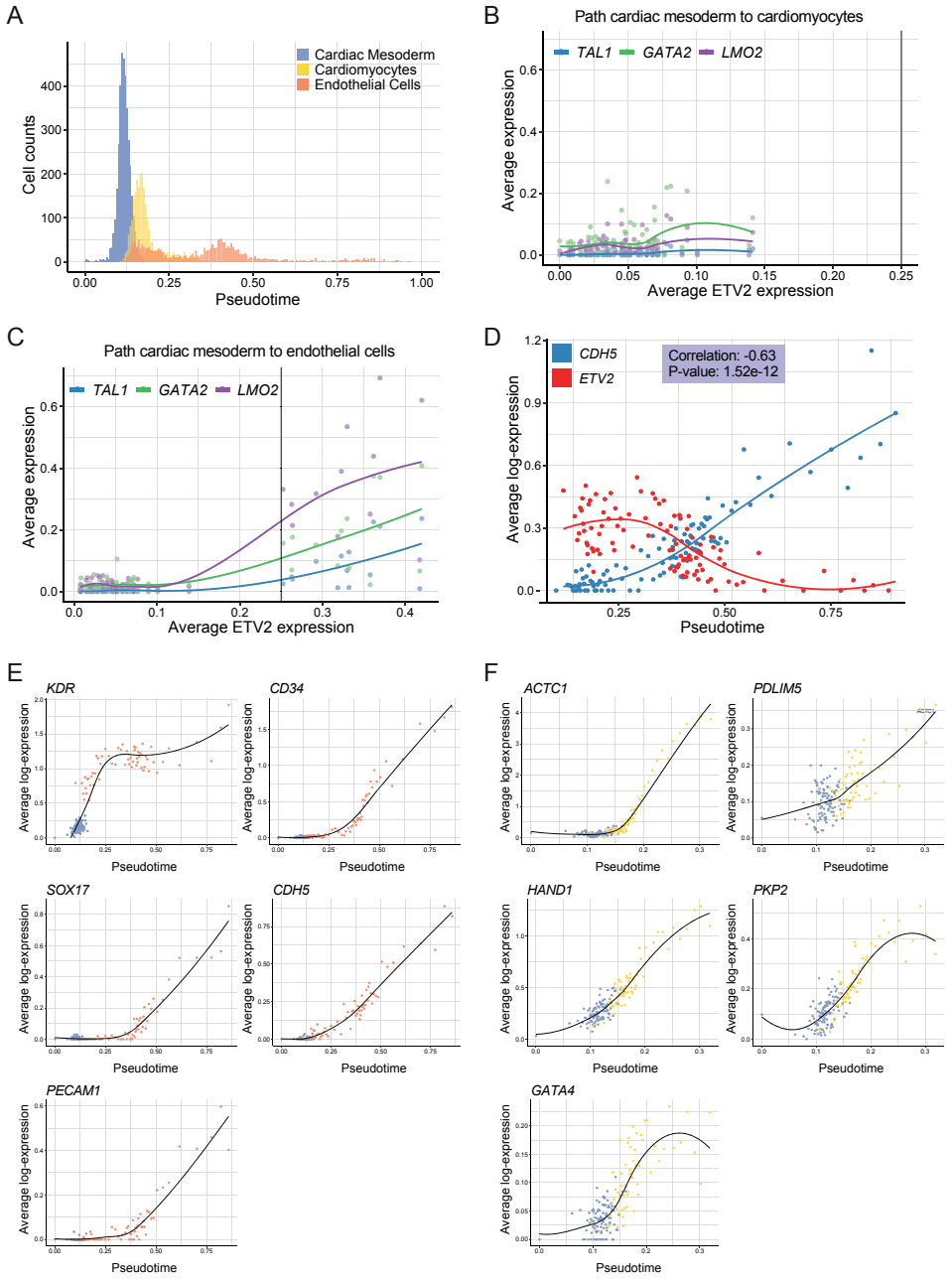


Figure 3.2: Pseudotime analysis of co-differentiation (A) Distributions of pseudotime for each cell cluster. (B-C) Average expression of ETV2 and ETV2 target genes along the developmental path of CMs (B) and ECs (C). Threshold (indicated in black) is set to the timepoint where the average ETV2 expression in the endothelial cell cluster reaches 0.25. (D) Average expression of CDH5 and ETV2 in the EC cluster across pseudotime. Binning and averaging was performed as for (B) and (C). The p-value for the correlation between CDH5 and ETV2 expression is based on the null hypothesis that the correlation is zero. (E) Expression of endothelial markers across pseudotime along the development path of ECs. (F) Expression of cardiac markers across pseudotime along the development path of CMs.

Taken together, these analyses confirmed the differentiation of cardiac mesoderm into CMs and ECs, which we had discovered previously. They also revealed the increase of ETV2 as a global indicator of early lineage separation and a transient pulse of high ETV2 at the beginning of EC specification.

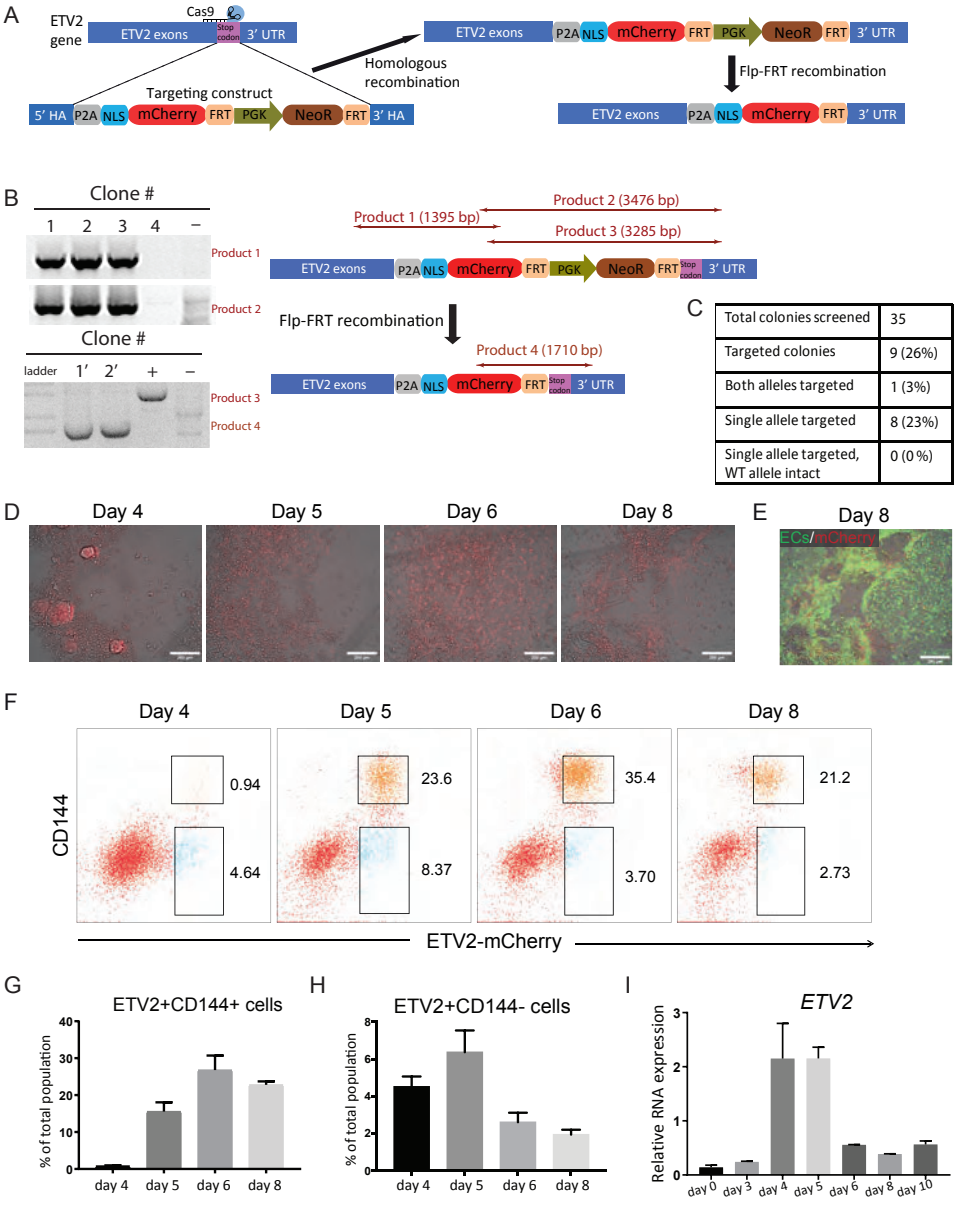


Figure 3.3: Generation and characterization of an ETV2mCherry hiPSC reporter line. (Caption on the next page.)

3.2.2 Generation and characterization of an ETV2^{mCherry} fluorescent hiPSC reporter line

In order to follow ETV2 expression in real-time, we introduced a fluorescent reporter for ETV2 in hiPSCs. P2A-mCherry with a nuclear localization signal (NLS) and a neomycin selection cassette was inserted into the endogenous ETV2 locus before the stop codon using CRISPR/Cas9-facilitated homologous recombination (Fig. 3.3A). After neomycin selection and excision of the selection cassette, targeted hiPSC clones were validated by PCR (Fig. 3.3B) and Sanger sequencing. Clones with ETV2^{mCherry} in both alleles were selected for downstream analysis (Fig. 3.3C). Time-lapse imaging of differentiating cultures showed the appearance of nuclear mCherry from day 4 of differentiation (Fig. 3.3D and supplemental online Video 1). Co-staining of live cells with EC-specific fluorescent agglutinin showed distinct EC clusters on day 8 of differentiation (Fig. 3.3E). Flow cytometry analysis at different stages of differentiation revealed upregulation of ETV2 (mCherry protein) as early as day 4 of differentiation, followed by the upregulation of the EC-specific marker CD144 (Fig. 3.3F). Quantification of percentages of single positive (SP; ETV2^{mCherry}+CD144-)(Fig. 3.3G) and double positive (DP; ETV2^{mCherry}+CD144+)(Fig. 3.3H) cells on day 4, 5, 6 and 8 of differentiation from at least three independent experiments showed a decrease and an increase of SP and DP cells respectively (Fig. 3.3G-H). mCherry protein remained present for a longer period than endogenous ETV2 mRNA (Fig. 3.3I), likely due to a longer half-life of the protein compared to the mRNA. This also explains the persistence of mCherry signal in both the DP and SP population.

Figure 3.3: Generation and characterization of an ETV2mCherry hiPSC reporter line (Figure on the previous page.) (A) Schematic of CRISPR/Cas9-Mediated Knock-in of mCherry into the ETV2 locus of hiPSCs. mCherry and Neomycin Resistance (NeoR) was inserted into the ETV2 locus through homologous recombination. Then NeoR cassette was removed by flpo recombinase. (B) PCR screening of targeted clones with correct insertion at the ETV2 locus. Two pairs of primers (for product 1 and 2) were used to confirm the integration of the construct. Clone 1, 2, 3 were correctly targeted and clone 4 was not targeted. Non-targeted hiPSCs (-) were included as negative control. Lower panel: The excision of the neomycin-resistance cassette was confirmed by PCR (product 3 and 4 for before and after excision). Clone 1' ad 2' were successfully excised. Genomic DNA before excision (+) and non-targeted hiPSCs (-) were included as positive and negative control, respectively. (C) Summary of CRISPR targeting efficiency. Out of 35 colonies screened, 1 colony was targeted in both alleles. 8 colonies were targeted in only one allele, but the other allele showed undesired mutations. (D) Overlay of bright-field and mCherry fluorescence images on differentiation day 4, 5, 6 and 8 of differentiation using the ETV2-mCherry reporter line. Scale bar 200 μm. (E) FITC-agglutinin staining of cells at the same location as shown in (D) on day 8. mCherry is in red and Agglutinin is in green. Scale bar 200 μm. (F) Fluorescence activated cell sorting (FACS) based on CD144 and ETV2mCherry expression on day 4, 5, 6 and 8 of differentiation. (G-H) Quantification of ETV2+CD144- (SP) ETV2+CD144+ (DP) cells by flow cytometry on differentiation day 4, 5, 6 and 8. (I) Quantification of ETV2 expression by qPCR on different days of differentiation. (F-G) Error bars are standard deviations calculated from three independent experiments.

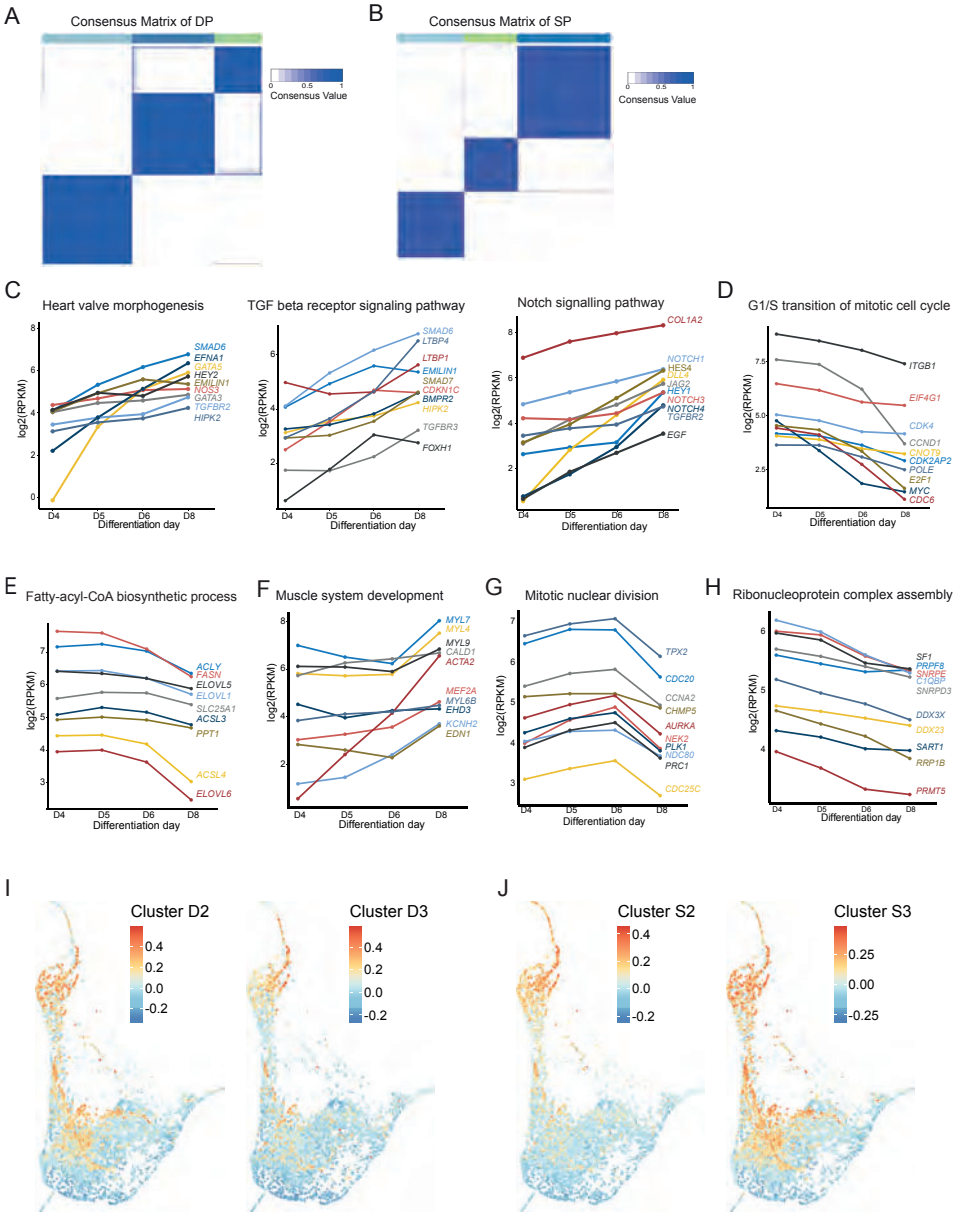


Figure 3.4: Bulk RNA-seq analysis of of EC and CM differentiation (A-B) Consensus clustering of the 3000 most variant genes across all DP samples (A) and all SP samples (B). All genes were divided into 3 clusters D1-D3 for DP (A) and S1-S3 for SP (B). Consensus value indicates similarity between two genes. (C-H) Representative GO terms enriched in cluster D1 (C), D2 (D), D3 (E), S1 (F), S2 (G) and S3 (H). Representative genes mapped to these GO terms and their expression levels from day 4 to day 8 are shown. (I-J) Low-dimensional representation of the scRNA-seq data. Each data point is a single cell. Mean expression of genes in cluster D1 (I) and S1 (J) in the scRNA-seq data is indicated by color. Gene expression was scaled gene-wise prior to averaging.

3.2.3 The ETV2^{mCherry} fluorescent reporter allows for the purification of differentiating cells with lineage-specific expression profiles

We next sorted DP and SP cells at different stages of differentiation (Fig. 3.3F) and performed bulk RNA-seq of at least three independent replicates. Principal component analysis (PCA) showed that DP and SP populations diverged progressively over time (Fig. 3.5A). Mapping of the bulk transcriptomes to the scRNA-seq data revealed that DP samples aligned on the EC branch and SP cells on the CM branch (Fig. 3.5B). Notably, SP and DP cells of later time points were further away from cardiac mesoderm, reflecting ongoing differentiation (Fig. 3.5B).

We next leveraged the higher sensitivity and accuracy of bulk RNA-seq compared to scRNA-seq, to get a more comprehensive and robust transcriptional characterization of the subpopulations. By consensus clustering of the most variable genes across DP or SP cells (3000 genes each) we found three gene clusters for each population, with distinct expression dynamics (Fig. 3.5C-D, Fig. 3.4A-B, Table S5). In DP cells, cluster D1 (1226 genes) expression increased over time. Gene Ontology (GO) terms enriched in cluster D1 included "angiogenesis", "Notch signaling pathway", "transforming growth factor beta receptor signaling pathway", "receptor-mediated endocytosis" and "developmental maturation" (Fig. 3.5E, Table S6). In accordance with this analysis, angiogenesis-related genes (CDH5, TIE1, TEK, EFNB2, SOX18, VEGFB, LEPR), Notch and transforming growth factor beta receptor signaling pathway related genes (COL1A2, NOTCH1, HES4, DLL4, JAG2, HEY1, NOTCH3, NOTCH4, TGFBR2, EGF) and heart valve morphogenesis related genes (SMAD6, EFNA1, GATA5, HEY2, EMILIN1, NOS3, GATA3) were upregulated over the course of differentiation in DP cells (Fig. 3.5G, Fig. 3.4C). In the scRNA-seq data, cluster D1 genes were specifically expressed in the EC cluster and showed increasing expression along pseudotime (Fig. 3.5I). Cluster D1 genes are thus likely involved in EC-specific functions. Cluster D2 (1127 genes), which was downregulated after day 4 (Fig. 3.5C), was enriched for cell cycle-related genes (ITGB1, CDK4, CCND1, CDK2AP2, MYC, CDC6)(3.4D). Cluster D3 (647 genes), which was downregulated after day 5-6 (Fig. 3.5C), contained cell proliferation- and fatty-acyl-CoA biosynthetic process-related genes (ACLY, FASN, ELOVL1, SLC25A1, ACSL3, ACSL4)(Fig. 3.4E). Genes in clusters D2 and D3 were more broadly expressed in the scRNA-seq data (Fig. 3.4I). Their dynamics likely reflect changes in proliferation and metabolism at the exit from the multipotent progenitor state.

In SP cells, cluster S1 (936 genes) increased over time and contained genes enriched for GO terms related to heart development and function (Fig. 3.5F, Table S6). In agreement with this observation, cardiac chamber and cardiac muscle development related genes (MYH6, HAND1, MYH10, TNNT2, NKX2-5, ISL1, TNNC1, MYOD, LMO4 and HEY1, MYL7, MYL4, ACTA2, KCNH2) were upregulated over the course of differentiation (Fig. 3.5H, Fig. 3.4F). Cluster S1 genes were highly expressed in the cardiac mesoderm and CMs clusters in the scRNA-seq data, which showed an increase over pseudotime in the CM lineage (Fig. 3.5J). These genes are thus likely involved in CM-specific functions. Cluster S2 (746 genes), which increased slightly until day 6 and was downregulated afterwards (Fig. 3.5D), contained mitotic nuclear division genes (TPX2, CDC20, NEK2, PLK1, PRC1 and CDC25C) (Fig. 3.4G).

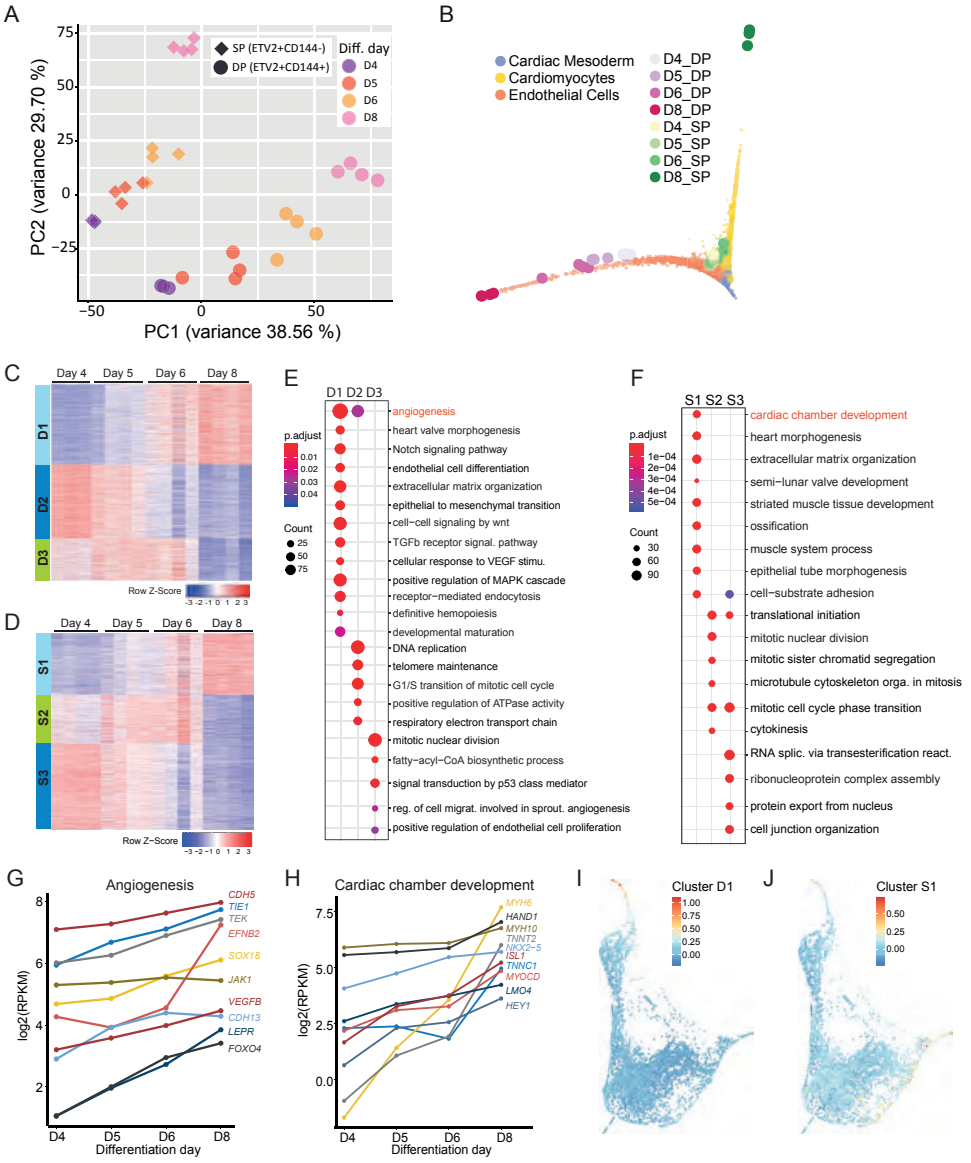


Figure 3.5: Bulk RNA-seq of the ETV2mCherry reporter line shows diverging transcriptional profiles. (A) PCA of all sorted DP and SP samples collected from three or four independent differentiations. (B) Low-dimensional representation (diffusion map) of scRNA-seq and bulk RNA-seq samples collected on day 4, 5, 6 and 8. The small data points correspond to individual cells, the large symbols correspond to bulk samples. Different clusters of cells or bulk samples are labelled with different colors. (C-D) Gene expression pattern in all DP (D) and SP (E) cells. The 3000 most variable genes across all DP or SP samples were identified and grouped into three clusters by consensus clustering. The genes in each cluster can be found in Table S5. The color scale represents relative expression (row-wise z-score). (E-F) GO enrichment analysis of each gene cluster of DP (E) and SP (F) samples. Representative GO terms are shown. The complete list of GO terms can be found in Table S6. Color represents the Padjusted of the enrichment analysis and dot size represents the count of genes mapped to the GO term. (G-H) Representative genes mapped to representative GO terms of clusters D1 (G) and S1 (H) and their expression levels from day 4 to day 8 are shown. (I-J) Low-dimensional representation of the scRNA-seq data. Each data point is a single cell. Mean expression of genes in cluster D1 (I) and S1 (J) in the scRNA-seq data is indicated by color. Gene expression was scaled gene-wise prior to averaging.

Cluster S3 (1318 genes), whose expression decreased continuously over time (Fig. 3.5D), contained transcription and translation process-related genes (SF1, SNRPE, DDX23, RRP1B and PRMT5) (3.4H). In the scRNA-seq data, genes from clusters S2 and S3 showed broader expression patterns compared to cluster S1 genes (Fig. 3.4J). The dynamics of clusters S2 and S3 likely reflect changes in proliferation and metabolism in the CM lineage, analogous to the role of clusters D2 and D3 in the EC lineage.
All in all, time-resolved bulk RNA-seq of sorted SP and DP populations confirmed that ETV2-positive cells contained transcriptionally distinct subpopulations. DP cells were part of the EC lineage, while SP cells corresponded to the CM lineage.

3.2.4 ETV2+ cells contain lineage-restricted subpopulations

Next, we wanted to find out how far the various subpopulations we identified differed in terms of their further differentiation potential. To that end, we sorted cells on ETV2 reporter levels shortly after the bifurcation (on day 5) and attempted to differentiate them further towards ECs by adding VEGF (Fig. 3.6A-B). After 5 days of additional differentiation, $ETV2^{mCherry}$+ cells produced more than 90% CD144+CD31+ ECs, while $ETV2^{mCherry}$- cells gave rise to only 10-15% ECs (3.6C-D). Only cells derived from $ETV2^{mCherry}$+ cells expressed endothelial-specific markers, as observed by qPCR and immunofluorescence (Fig. 3.6E-H). These cells also upregulated pro-inflammatory markers, such as ICAM-1 and E-Selectin upon TNF-α stimulation (Fig. 3.6I-L), as shown previously for hiPSC-derived ECs [33]. We could thus conclude that the majority of ETV2-positive cells on day 5 have a strong propensity to produce ECs.
Both the analysis of the scRNA-seq data and the time-resolved bulk RNA-seq of sorted cells identified a subpopulation of ETV2+ cells with CM characteristics. We strongly suspected that the differentiation potential of those cells would be restricted to the CM lineage. To test this hypothesis with our reporter line, we co-differentiated cells until day 7. We chose a later time point for this experiment because the majority of cells are past the bifurcation at this point, and it is, therefore, easier to identify the ETV2+ population that does not correspond to early ECs. We co-stained for CD144 and sorted the cells into DP, SP, and double negative (DN) populations. These subpopulations were then further cultured in the presence of VEGF until day 18 (Fig. 3.6A, M). The majority (>80%) of DP cells differentiated into CD144+CD31+ ECs, in agreement with the previous experiment (Fig. 3.6N-O). In contrast, more than 50% of SP and DN cells differentiated into cTnT+ CMs while very few ECs were detected (Fig. 3.6N-O). Interestingly, CMs derived from SP cells seemed to proliferate more and formed a monolayer composed of a contracting cell sheet, while CMs from DN cells proliferated to a lesser extent and produced only a few, separated clusters of contracting cells (supplemental online Video 2). Almost all DP cells on day 18 expressed the EC marker CD31, while only few cells derived from SP and DN cells were positive for CD31 (Fig. 3.6P-R). Most cells derived from SP and DN expressed CM-specific α-Actinin and cTnT and showed typical sarcomeric structures (Fig. 3.6Q-R). A small number of SP and DN-derived cells were also positive for the smooth muscle cell marker SM22, while negative for cardiac markers (data not shown). Furthermore, SP cell fraction-derived α-Actinin positive CMs were positive for SM22, possibly indicating their immaturity (Fig. 3.6R).

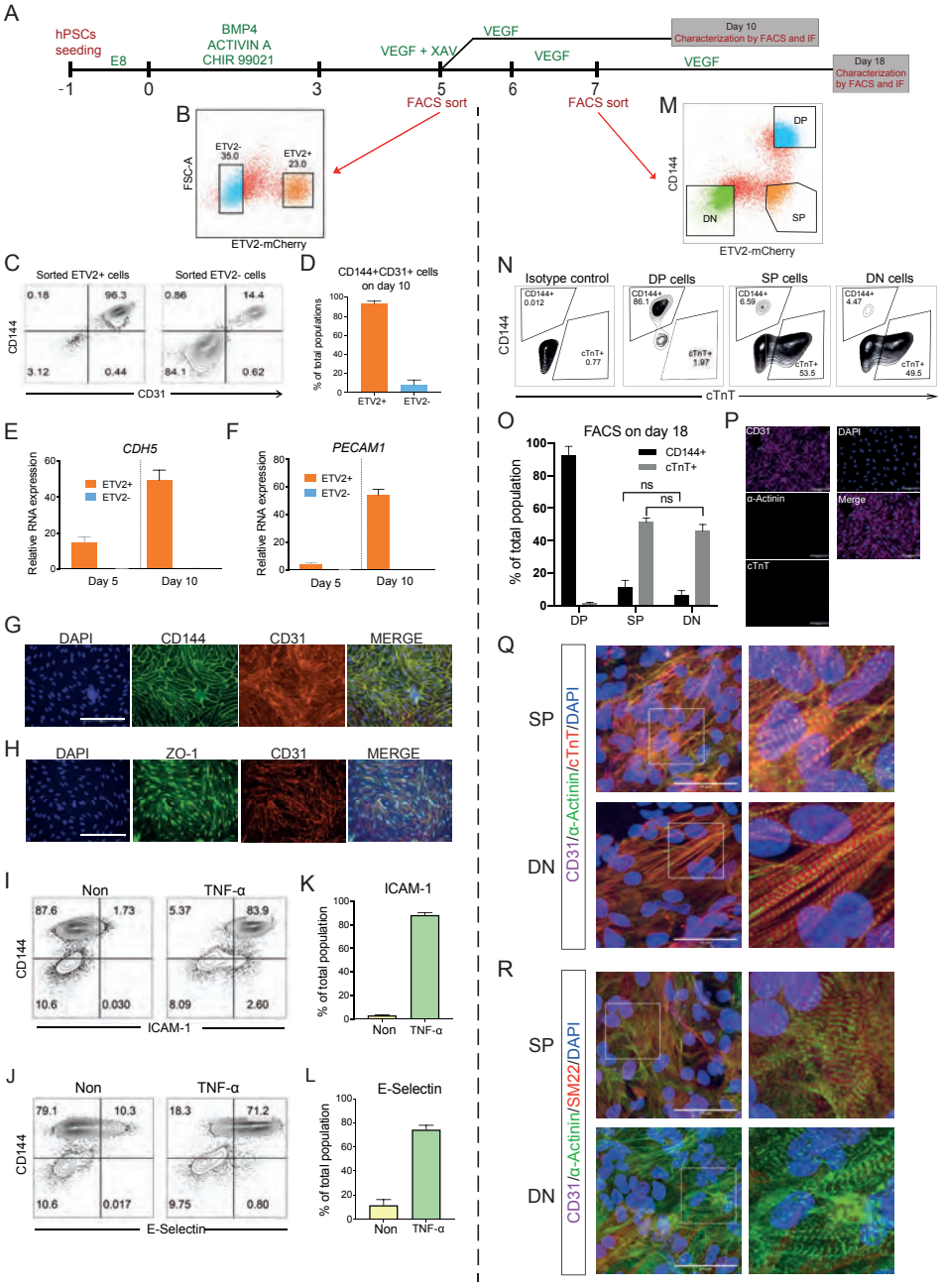


Figure 3.6: ETV2+ cells contain two lineage-restricted subpopulations (Caption on the next page.)

Taken together, the VEGF differentiation experiments showed that DP and SP cells are restricted to the EC and CM lineage respectively. DN cells were largely unable to give rise to EC, but produced CMs, albeit with lower efficiency than SP cells. Entering a transient state characterized by high ETV2 expression, thus seems necessary to initiate EC specification.

3.3 Discussion

In this study, we characterized the dynamics of EC and CM co-differentiation from hiPSCs [32]. ETV2 was identified as an early indicator of lineage separation and found to be strongly, but transiently, upregulated in ECs, in agreement with its fundamental role in hemangiogenic development [34]. Interestingly, ETV2 expression was also observed in a small population of cardiac mesoderm and CMs. This is reminiscent of a recent study where etv2 expression was observed in lateral plate mesoderm and the CM population in zebrafish [35]. In our experiments, expression of ETV2 target genes seemed to occur only beyond a certain level of ETV2 expression, although this observation could also be explained by a temporal delay between ETV2 upregulation and target gene expression. An ETV2 threshold in our system would be in line with previous reports of an ETV2 threshold in hemangiogenic specification [18, 19]. Our results thus support an ETV2 pulse- and threshold dependent specification of ECs.

With the ETV2^{mCherry} hiPSC reporter line generated to track, isolate and characterize ETV2+ cells, we showed that ETV2+ cells could give rise to both ECs and CMs. Over time, EC and CM precursors acquired more specific endothelial and myocardial identities, respectively, as well as downregulated cell cycle-related genes, which indicated exit from the progenitor state and further maturation.

In the DP subpopulation (EC precursors), several key angiogenesis and Notch signaling pathway genes, like LEPR, FOXO4, DLL4, NOTCH4 and EGF, strongly increased starting from day 4, indicating a specified EC fate but an immature state on day 4. These relatively late expressed genes could potentially be used as markers to distinguish early and late ECs during development in vitro or in vivo.

Figure 3.6: ETV2+ cells contain two lineage-restricted subpopulations (Figure on the previous page.) (A) Schematic of the differentiation protocol and cell sorting. ETV+ and ETV2- cells were sorted on day 5 and cultured in VEGF until day 10. DP, SP, DN were sorted on day 7 and cultured in VEGF until day 18. (B) Representative flow cytometry analysis ETV2-mCherry expression on day 5 and gates for cell sorting of ETV2+ and ETV2- cells are shown. (C) Flow cytometry analysis of endothelial markers CD144 and CD31 on day 10 of sorted ETV2+ and ETV2- cell differentiation. (D) Quantification of CD144+CD31+ cells in the total population on day 10 of sorted ETV2+ and ETV2- cell differentiation. (E-F) Quantification of CDH5 and PECAM1 expression in sorted ETV2+ and ETV2- cells on day 5 and day 10. (G-H) Immuno-staining of CD144, CD31 and cell-cell junctional marker ZO-1 for sorted ETV2+ cells on day 10. Scale bar 200 µm. (I-J) Flow cytometry analysis of ICAM1, E-Selectin and CD144 for sorted ETV2+ cells on day 10. Cells were stimulated with TNF-α for 24 h before analysis. (K-L) Quantification of CD144+ICAM-1+ (K) and CD144+E-Selectin+ (L) cells in the population on day 10. (M) Flow cytometry analysis of CD144 and ETV2-mCherry expression on day 7. DP, SP and DN cells were gated and sorted. (N) Flow cytometry analysis of CD144 and CM marker cTnT expression on day 18 of sorted DP, SP and DN cells. Isotype control antibodies were included as negative control. (O) Quantification of CD144+ ECs and cTnT+ CMs on day 18 of DP, SP and DN cell differentiation. (P) Immuno-staining of CD31, a-Actinin, cTnT and DAPI on day 18 of DP cell differentiation. Scale bar 50 µm. (Q-R) Immuno-staining of CD31, a-Actinin, cTnT, SM22 and DAPI on day 18 of SP and DN cells. Scale bar 50 µm. Error bars are ± SD of three independent experiments in (D-F, K-L, O).

Genes involved in heart development and definitive hematopoiesis were also upregulated during EC development, suggesting cardiac endothelial- and probably a mixture of hemogenic endothelial identity of these ECs. A better characterization of these cells' hematopoietic potential would be interesting, but is beyond the scope of this study. ECs that were further differentiated with VEGF showed a clear endothelial identity and were fully functional based on their inflammatory response upon TNF-α stimulation. Notably, they also expressed a number of cardiac markers like MEOX2, GATA4, GATA6 and ISL1, suggesting a cardiac specific EC identity [32].

The SP subpopulation (CM precursors) had already committed to a cardiac fate on day 4, as evidenced by the expression of cardiac genes HAND1, MYH10, NKX2-5, ISL1, TNNC1, MYOCD and LMO4. However, some crucial genes for CMs were still absent, including MYH6 and TNNT2. MYH6 encodes the major CM thick filament protein MHC-α and TNNT2 is routinely used as a CM marker. Both genes are essential for CM contractility and started to be expressed only after day 4. Their relatively late expression could allow us to identify early and late cardiac progenitors during cardiac development in future studies. CMs were still early progenitors on day 6 of the differentiation as no functionally contracting CMs were observed yet at this stage. Pseudotime analysis also suggested that ECs had differentiated further compared to CMs on day 6. After additional VEGF differentiation, SP cells gave rise to contracting CM, which provided direct evidence they were CM precursors. More importantly, it demonstrated that both ECs and CMs could be derived from ETV2+ progenitors, confirming the presence of a common precursor implied by our earlier studies [32].

Notably, ETV2- cells (DN population) also gave rise to contracting CMs after VEGF treatment, albeit less frequently than SP cells. This difference could be due to either the different cell growth rates or their different developmental origins (FHF vs. SHF). More work is needed to establish the identity of CMs from SP and DN populations in the future.

3.4 Conclusion

Transcriptomic analysis at bulk and single cell levels done in this study provide insight into the differentiation dynamics of two important human cardiac lineages and a rich data set for comparison with in vivo data. An ETV2 reporter system was generated in hiPSCs and utilized to identify a new subpopulation of early CM precursors that expressed ETV2.

3.5 Materials and Methods

3.5.1 Experimental Methods

hiPSC culture

The NCRM1 hiPSC line (NIH) was used in this study. This hiPSC control line was cultured in TeSR-E8 on Vitronectin XF and was routinely passaged once a week using Gentle Cell Dissociation Reagent (all from Stem Cell Technologies). Prior to targeting, NCRM1 hiPSCs were passaged as a bulk on feeders in hESC-food medium. RevitaCell (Life Technologies) was added to the medium (1:200) after every passage to enhance viability after single cell passaging with TrypLE (Life technologies).

Generation of hiPSC reporter line using CRISPR/Cas9

The p15a-cm-hETV2-P2A-NLS-mCherry-neo repair template plasmid was generated using overlap PCR and restriction-based cloning and ligation. The ETV2 homology arms were amplified from genomic DNA and the neomycin cassette flanked by two flippase recognition target (FRT) sites was amplified from the P15 backbone vector (kindly provided by Dr. Konstantinos Anastassiadis, Technical University Dresden). P2A-NLS-mCherry double-stranded DNA fragment was ordered from IDT. The sgRNA/Cas9 plasmid was modified from SpCas9-2A-Puro V2.0 plasmid (Addgene, Feng Zang).
NCRM1 hiPSCs were passaged with ratio 1:2 or 1:3 into 60 mm dishes to reach 60-70% confluency the next day for transfection. 20 µl lipofectamine (Invitrogen), 8 µg of repair template and 8 µg of sgRNA/Cas9 plasmid were diluted in 600 µl of Opti-MEM and added to each 60 mm dish. After 18 hours the medium was changed to hESC-food. After another 6 hours, G-418 (50 µg/ml) selection was started and was kept for 1 week. Survived cells were cultured in hESC-food and passage into 6-well plate for the transfection of Flp recombinase expression vector to remove the neomycin cassette [36]. 300 µl of Opti-MEM containing 10 µl lipofectamine and 4 µg CAGGs-Flpo-IRES-puro plasmid was added per well for 18 hours. Puromycin (0.5 µg/ml) selection was started 24 hours post transfection and lasted for 2 days. Once recovered, cells were passage into 96-well format for clonal expansion via limited dilution. Targeted clones were identified by PCR and sequencing. Primers outside the ETV2 homology arms and primers inside the targeting construct were used to confirm on-target integration. The absence of mutation within inserted sequence and untargeted allele were confirmed by Sanger sequencing (BaseClear).

Endothelial and myocardial lineages co-differentiation from hiPSCs

Endothelial and cardiac cells were induced from hiPSCs in a monolayer using CMEC protocol as described previously [32]. Briefly hiPSCs were split with a 1:12 ratio and seeded on 6-well plates coated with 75 µg/mL (growth factor reduced) Matrigel (Corning) on day -1. At day 0, cardiac mesoderm was induced by changing TeSR-E8 to BPEL medium [37], supplemented with 20 ng/mL BMP4 (R& D Systems), 20 ng/mL ACTIVIN A (Miltenyi Biotec) and 1.5 µM CHIR99021 (Axon Medchem). At day 3, cells were refreshed with BPEL supplemented 5 µM XAV939 (Tocris Bioscience) and 50 ng/ml VEGF (R& D Systems). From day 6 onwards, cells were refreshed every 3 days with BPEL medium supplemented with 50 ng/ml VEGF.

Fluorescence-activated cell sorting

For FACS sorting on day 5 of CMEC protocol, ETV2-mCherry positive and negative cells were sorted using FACSAria III (BD-Biosciences). Around 20k cells/cm2 were seeded on Fibronectin (from bovine plasma, 5µg/ml, Sigma Aldrich) coated plates. Cells were cultured in BPEL supplemented with VEGF (50 ng/ml) until day 10. The medium was refreshed every 3 days. For FACS sorting on day 7 of CMEC protocol, VEC+mCherry+ (DP), VEC-mCherry+ (SP) and VEC-mCherry- (DN) cells were sorted using FACSAria III. 1 million cells were seeded in each well of Matrigel-coated 12-well plate in BPEL supplemented with VEGF (50 ng/ml) and RevitaCell (1:200). Cells were refreshed 24 h after seeding and every three days afterwards with BPEL supplemented with VEGF (50 ng/ml).

Immunofluorescence staining and imaging

Cultured cells were fixed in 4% paraformaldehyde for 15 min, permeabilized for 10 min with PBS containing 0.1% Triton-X 100 (Sigma-Aldrich) and blocked for 1h with PBS containing 5% BSA (Sigma-Aldrich). Then cells were stained with primary antibody overnight at 4℃. The next day cells were washed three times (20 min each time) with PBS. After that cells were incubated with fluorochrome-conjugated secondary antibodies for 1h at room temperature and washed three times (20 min each time) with PBS. Then cells were stained with DAPI (Life Technologies) for 10 min at room temperature and washed once with PBS for 10min. Both primary and secondary antibodies were diluted in 5% BSA/PBS. Images were taken with EVOS FL AUTO2 imaging system (ThermoFischer Scientific) with 20x objective. For staining of ECs and CMs on day 18, images were taken with a Leica SP8WLL confocal laser-scanning microscope using a 63x magnification objective and Z-stack acquisition. Details of all antibodies that were used are provided in Table S1.

FACS analysis

Cells were washed once with FACS buffer (PBS containing 0.5% BSA and 2 mM EDTA) and stained with FACS antibodies for 30 min at 4oC. Samples were washed once with FACS buffer and analyzed on MACSQuant VYB (Miltenyi Biotech) equipped with a violet (405 nm), blue (488 nm) and yellow (561 nm) laser. The results were analyzed using Flowjo v10 (FlowJo, LLC). Details of all fluorochrome conjugated FACS antibodies are provided in Table S1.

Quantitative Real-Time Polymerase Chain Reaction (qPCR)

Total RNA was extracted using the NucleoSpin® RNA kit (Macherey-Nagel) according to the manufacturer's protocol. cDNA was synthesized using an iScript-cDNA Synthesis kit (Bio-Rad). iTaq Universal SYBR Green Supermixes (Bio-Rad) and Bio-Rad CFX384 real-time system were used for the PCR reaction and detection. Relative gene expression was determined according to the standard ΔCT calculation and normalized to housekeeping genes (mean of HARP and RPL37A). Details of all primers used are provided in Table S2.

3.5.2 Bulk RNA sequencing and analysis

Cells were sorted on differentiation day 4, 5, 6 and 8 for bulk RNA-Seq. Total RNA was extracted using the NucleoSpin® RNA kit (Macherey-Nagel). Whole transcriptome data were generated at BGI (Shenzhen, China) using the Illumina Hiseq4000 (100bp

paired end reads). Raw data was processed using the LUMC BIOPET Gentrap pipeline (https://github.com/biopet/biopet), which comprises FASTQ preprocessing, alignment and read quantification. Sickle (v1.2) was used to trim low-quality read ends (https://github.com/najoshi/sickle). Cutadapt (v1.1) was used for adapter clipping [38], reads were aligned to the human reference genome GRCh38 using GSNAP (gmap-2014-12-23) [39, 40] and gene read quantification with htseq-count (v0.6.1p1) against the Ensembl v87 annotation [41]. Gene length and GC content bias were normalized using the R package cqn (v1.28.1) [42]. Genes were excluded if the number of reads was below 5 reads in ≥ 90% of the samples. The final dataset comprised gene expression levels of 31 samples and 22,419 genes.

Differentially expressed genes were identified using generalized linear models as implemented in edgeR (3.24.3) [43]. P-values were adjusted using the Benjamini-Hochberg procedure and FDR ≤ 0.05 was considered significant. Analyses were performed using R (version 3.5.2). PCA plot was generated with the built-in R functions prcomp using transposed normalized RPKM matrix. Correlation among samples was calculated using cor function with spearman method and the correlation heatmap was generated with aheatmap function (NMF package).

Gene clusters were calculated with CancerSubtypes package [44]. The top 3000 most variable genes across all chosen samples were identified based on the most variant Median Absolute Deviation (MAD) using FSbyMAD function, then z_ score normalization was performed for each gene. K clusters were calculated using the K-means clustering of euclidean distance. Clustering was iterated 1000 times for K clusters in the range 2 to 10. Heatmap of genes in all clusters was generated using R basic heatmap function. Gene ontology enrichment for each cluster of genes was performed using compareCluster function of clusterProfiler package (v3.10.1) [45] and q ≤ 0.05 was considered significant.

3.5.3 Single-cell RNA sequencing and analysis

Sample preparation and sequencing

Cells were dissociated into single cells on day 6 of CMEC differentiation and loaded into the 10X Chromium Controller for library construction using the Single-Cell 3' Library Kit. Next, indexed cDNA libraries were sequenced on the HiSeq4000 platform. The mean reads per cell were reported as 28,499 in the first replicate and 29,388 in the second replicate.

Pre-processing

Both replicates of day 6 CMEC differentiation were merged into one data set. The average number of detected genes is 2643 and the average total expression per cell is 10382 (Fig. 3.2A-B). Then, undetected genes (> 1 UMI count detected in less than two cells) and cells with low number of transcripts were removed from further analysis (3.2A-B). This resulted in 5107 cells for the first replicate, and 3743 cells for the second replicate with 13243 genes each. Expression profiles were normalized with the scran package in R (V 1.10.2) using the method described in [46]. The 5% most highly variable genes (HVGs) for each replicate were calculated with scran after excluding ribosomal genes (obtained from the HGNC website without any filtering for minimum gene expression), stressed genes [47] and mitochondrial genes. For downstream analysis the top 5% HVGs were used after excluding proliferation [48] and cell cycle [49] related genes.

Cell cycle analysis

For each combined data set, cell cycle analysis was performed with the scran package using the cyclone function [50] on normalized counts. Cells with a G2/M score higher than 0.2 were considered to be in G2/M phase. Otherwise, they were classified as G1/S. Using this binary classifier as predictor, we regressed out cell cycle effects with the R package limma (V 3.42.2) [51] applied to log-transformed normalized counts. Then, the two replicates were batch corrected with the fast mutual nearest neighbors (MNN) correction method [52] on the cell cycle corrected counts, using the 30 first principal components and 20 nearest-neighbors.

Clustering

First, batch-corrected counts were standardized per gene and then used to create a shared nearest neighbour (SNN) graph with the scran R package (d = 30, k = 20). Louvain clustering was applied to the SNN graph using the igraph python package (V 0.7.1) with 0.4 for resolution parameter. This resulted in 5 clusters. Two of these 5 clusters were excluded from further analysis based on the expression of pluripotency markers [53].

Dimensionality reduction and pseudotime

Dimensionality reduction was performed using the python scanpy pipeline (V 1.4.6). First a 20 nearest-neighbors (knn, k=20) graph was created from diffusion components of the batch corrected data sets. Diffusion components are the eigenvectors of the diffusion operator which is calculated from Euclidean distances and a gaussian kernel. The aim is to find a lower dimensional embedding which considers the cellular progression. The graph was projected into two dimensions with the default force-directed graph layout and starting positions obtained from the partition-based graph abstraction (PAGA) output [54]. PAGA estimates connectivities between partitions and performs an improved version of diffusion pseudotime. Diffusion pseudotime [52, 54] was calculated on these graphs with root cells selected based on the graph layout from the “Cardiac Mesoderm” cluster.
Average gene expression trajectories were calculated by dividing the cells of each cluster into bins along pseudotime. 50 bins were created for cardiac mesoderm and 30 bins for each, endothelial cells and cardiomyocytes. Then, the average log-expression per bin was shown. The value of the threshold shown in Fig. 3.1D,E was determined by calculating the point in pseudotime where the average ETV2 expression was the lowest in the endothelial cell cluster before the peak expression, which corresponds to a value around 0.25 [48].

Differential expression analysis and identification of cluster maker genes

The R package edgeR (V 3.24.3, 31) [43] was used to perform differential expression analysis. We used raw counts and a negative binomial distribution to fit the generalized linear model. The covariates were comprised of 6 binary dummy variables that indicate the three remaining clusters per replicate and a variable that corresponds to the total number of counts per cell. Finally, p-values for each cluster considering both replicates were obtained and adjusted for multiple hypothesis testing with the Benjamini-Hochberg method.

Comparison to bulk RNA-sequencing data

Both replicates of normalized single cell counts were combined with bulk RNA-sequencing data. The MNN approach was used to correct between the two single-cell replicates using the 10% HVG per replicate and the bulk RNA-sequencing data, with d = 30 and k = 20. After batch correction a diffusion map was calculated on the MNN corrected values with default parameters.

Statistics

Statistical analysis was conducted with GraphPad Prism 7 software. Data are represented as mean± SD.

3.5.4 Data availability

The accession numbers for the bulk and single cell RNA sequencing datasets reported in this paper are https://www.ncbi.nlm.nih.gov/geo/ GEO: GSE157954 (bulk) and GEO: GSE202901 (single cell). Supplementary tables and videos are available at https://doi.org/10.5061/dryad.9p8cz8wkg.

Acknowledgements S.L. Kloet and E. de Meijer (Leiden Genome Technology Center) for help with 10X Genomics experiments (cell encapsulation, library preparation, single-cell sequencing, primary data mapping, and quality control). K. Anastassiadis (Technical University Dresden) for providing P15 backbone with a Neomycin resistance cassette surrounded by two FRT sequences and CAGGs-Flpo-IRES-puro vector.

Funding This project received funding from the European Union's Horizon 2020 Framework Programme (668724); European Research Council (ERCAdG 323182 STEMCARDIOVASC); Netherlands Organ-on-Chip Initiative, an NWO Gravitation project funded by the Ministry of Education, Culture and Science of the government of the Netherlands (024.003.001). M. M. and S.S. were supported by the Netherlands Organisation for Scientific Research (NWO/OCW,www.nwo.nl), as part of the Frontiers of Nanoscience (NanoFront) program. The computational work was carried out on the Dutch national e-infrastructure with the support of SURF Cooperative.

Disclosure of potential conflict of interest The authors indicated no potential conflicts of interest.

References

[1] W. P. Devine et al. Early patterning and specification of cardiac progenitors in gastrulating mesoderm. *eLife*, 3, oct 2014.

[2] F. Lescroart et al. Early lineage restriction in temporally distinct populations of Mesp1 progenitors during mammalian heart development. *Nature Cell Biology 2014 16:9*, 16(9):829–840, aug 2014.

[3] S. Zaffran et al. Right ventricular myocardium derives from the anterior heart field. *Circulation Research*, 95(3):261–268, aug 2004.

[4] Y. Saga et al. MesP1 is expressed in the heart precursor cells and required for the formation of a single heart tube. *Development*, 126(15):3437–3447, aug 1999.

[5] F. Lescroart et al. Defining the earliest step of cardiovascular lineage segregation by single-cell RNA-seq. *Science*, 359(6380):1177–1181, mar 2018.

[6] M. Ema, S. Takahashi, and J. Rossant. Deletion of the selection cassette, but not cis-acting elements, in targeted Flk1-lacZ allele reveals Flk1 expression in multipotent mesodermal progenitors. *Blood*, 107(1):111–117, jan 2006.

[7] D. J. Garry and E. N. Olson. A Common Progenitor at the Heart of Development. *Cell*, 127(6):1101–1104, dec 2006.

[8] P. P. Tam, M. Parameswaran, S. J. Kinder, and R. P. Weinberger. The allocation of epiblast cells to the embryonic heart and other mesodermal lineages: the role of ingression and tissue movement during gastrulation. *Development*, 124(9):1631–1642, may 1997.

[9] M. Buckingham, S. Meilhac, and S. Zaffran. Building the mammalian heart from two sources of myocardial cells. *Nature Reviews Genetics 2005 6:11*, 6(11):826–835, nov 2005.

[10] S. D. Vincent and M. E. Buckingham. How to Make a Heart: The Origin and Regulation of Cardiac Progenitor Cells. *Current Topics in Developmental Biology*, 90(C):1–41, jan 2010.

[11] D. Galli et al. Atrial myocardium derives from the posterior region of the second heart field, which acquires left-right identity as Pitx2c is expressed. *Development*, 135(6):1157–1167, mar 2008.

[12] C. L. Cai et al. Isl1 Identifies a Cardiac Progenitor Population that Proliferates Prior to Differentiation and Contributes a Majority of Cells to the Heart. *Developmental Cell*, 5(6):877–889, dec 2003.

[13] N. Paffett-Lugassy et al. Heart field origin of great vessel precursors relies on nkx2.5-mediated vasculogenesis. *Nature Cell Biology 2013 15:11*, 15(11):1362–1369, oct 2013.

[14] A. Ferdous et al. Nkx2-5 transactivates the Ets-related protein 71 gene and specifies an endothelial/endocardial fate in the developing embryo. *Proceedings of the National Academy of Sciences of the United States of America*, 106(3):814–819, jan 2009.

[15] S. De Val et al. Combinatorial regulation of endothelial gene expression by ets and forkhead transcription factors. *Cell*, 135(6):1053, dec 2008.

[16] D. Lee et al. ER71 Acts Downstream of BMP, Notch, and Wnt Signaling in Blood and Vessel Progenitor Specification. *Cell Stem Cell*, 2(5):497–507, may 2008.

[17] F. Liu et al. Induction of hematopoietic and endothelial cell program orchestrated by ETS transcription factor ER71/ETV2. *EMBO reports*, 16(5):654–669, may 2015.

[18] H. Zhao and K. Choi. A CRISPR screen identifies genes controlling Etv2 threshold expression in murine hemangiogenic fate commitment. *Nature Communications 2017 8:1*, 8(1):1–12, sep 2017.

[19] H. Zhao and K. Choi. Single cell transcriptome dynamics from pluripotency to FLK1+ mesoderm. *Development (Cambridge)*, 146(23), dec 2019.

[20] R. Morita et al. ETS transcription factor ETV2 directly converts human fibroblasts into functional endothelial cells. *Proceedings of the National Academy of Sciences of the United States of America*, 112(1):160–165, jan 2015.

[21] I. Elcheva et al. Direct induction of haematoendothelial programs in human pluripotent stem cells by transcriptional regulators. *Nature Communications 2014 5:1*, 5(1):1–11, jul 2014.

[22] A. G. Lindgren, M. B. Veldman, and S. Lin. ETV2 expression increases the efficiency of primitive endothelial cell derivation from human embryonic stem cells. *Cell Regeneration*, 4(1):4:1, jan 2015.

[23] V. S. Brok-Volchanskaya et al. Effective and Rapid Generation of Functional Neutrophils from Induced Pluripotent Stem Cells Using ETV2-Modified mRNA. *Stem Cell Reports*, 13(6):1099–1110, dec 2019.

[24] K. Suknuntha et al. Optimization of Synthetic mRNA for Highly Efficient Translation and its Application in the Generation of Endothelial and Hematopoietic Cells from Human and Primate Pluripotent Stem Cells. *Stem Cell Reviews and Reports*, 14(4):525–534, aug 2018.

[25] B. Cakir et al. Engineering of human brain organoids with a functional vascular-like system. *Nature Methods 2019 16:11*, 16(11):1169–1175, oct 2019.

[26] K. Wang et al. Robust differentiation of human pluripotent stem cells into endothelial cells via temporal modulation of ETV2 with modified mRNA. *Science Advances*, 6(30), jul 2020.

[27] B. Palikuqi et al. Adaptable haemodynamic endothelial cells for organogenesis and tumorigenesis. *Nature 2020 585:7825*, 585(7825):426–432, sep 2020.

[28] D. T. Paik et al. Large-scale single-cell RNA-seq reveals molecular signatures of heterogeneous populations of human induced pluripotent stem cell-derived endothelial cells. *Circulation Research*, 123(4):443–450, 2018.

[29] I. R. McCracken et al. Transcriptional dynamics of pluripotent stem cell-derived endothelial cell differentiation revealed by single-cell RNA sequencing. *European Heart Journal*, 41(9):1024–1036, mar 2020.

[30] S. C. Den Hartogh et al. Dual reporter MESP1 mCherry/w-NKX2-5 eGFP/w hESCs enable studying early human cardiac differentiation. *Stem Cells (Dayton, Ohio)*, 33(1):56–67, jan 2015.

[31] A. Moretti et al. Multipotent Embryonic Isl1+ Progenitor Cells Lead to Cardiac, Smooth Muscle, and Endothelial Cell Diversification. *Cell*, 127(6):1151–1165, dec 2006.

[32] E. Giacomelli et al. Three-dimensional cardiac microtissues composed of cardiomyocytes and endothelial cells co-differentiated from human pluripotent stem cells. *Development (Cambridge)*, 144(6):1008–1017, mar 2017.

[33] O. V. Halaidych et al. Inflammatory Responses and Barrier Function of Endothelial Cells Derived from Human Induced Pluripotent Stem Cells. *Stem Cell Reports*, 10(5):1642–1656, may 2018.

[34] N. Koyano-Nakagawa and D. J. Garry. Etv2 as an essential regulator of mesodermal lineage development. *Cardiovascular Research*, 113(11):1294–1306, sep 2017.

[35] B. Chestnut, S. Casie Chetty, A. L. Koenig, and S. Sumanas. Single-cell transcriptomic analysis identifies the conversion of zebrafish Etv2-deficient vascular progenitors into skeletal muscle. *Nature Communications 2020 11:1*, 11(1):1–16, jun 2020.

[36] A. Kranz et al. An improved Flp deleter mouse in C57Bl/6 based on Flpo recombinase. *genesis*, 48(8):512–520, aug 2010.

[37] E. S. Ng, R. Davis, E. G. Stanley, and A. G. Elefanty. A protocol describing the use of a recombinant protein-based, animal product-free medium (APEL) for human embryonic stem cell differentiation as spin embryoid bodies. *Nature Protocols 2008 3:5*, 3(5):768–776, apr 2008.

[38] M. Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, 17(1):10–12, may 2011.

[39] T. D. Wu and C. K. Watanabe. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*, 21(9):1859–1875, may 2005.

[40] T. D. Wu and S. Nacu. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics*, 26(7):873–881, apr 2010.

[41] A. Yates et al. Ensembl 2016. *Nucleic Acids Research*, 44(D1):D710–D716, jan 2016.

[42] K. D. Hansen, R. A. Irizarry, and Z. Wu. Removing technical variability in RNA-seq data using conditional quantile normalization. *Biostatistics*, 13(2):204–216, apr 2012.

[43] M. D. Robinson, D. McCarthy, Y. Chen, and G. K. Smyth. edgeR: differential expression analysis of digital gene expression data User's Guide. *Bioinformatics*, 26(October 2018):1–75, 2013.

[44] T. Xu et al. CancerSubtypes: an R/Bioconductor package for molecular cancer subtype identification, validation and visualization. *Bioinformatics (Oxford, England)*, 33(19):3131–3133, oct 2017.

[45] G. Yu, L. G. Wang, Y. Han, and Q. Y. He. clusterProfiler: an R package for comparing biological themes among gene clusters. *Omics : a journal of integrative biology*, 16(5):284–287, may 2012.

[46] A. T. Lun, K. Bach, and J. C. Marioni. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biology*, 17(1):1–14, apr 2016.

[47] S. C. van den Brink et al. Single-cell sequencing reveals dissociation-induced gene expression in tissue subpopulations. *Nature Methods*, 14(10):935–936, 2017.

[48] M. L. Whitfield, L. K. George, G. D. Grant, and C. M. Perou. Common markers of proliferation. *Nature Reviews Cancer 2006 6:2*, 6(2):99–106, feb 2006.

[49] B. Giotti, A. Joshi, and T. C. Freeman. Meta-analysis reveals conserved cell cycle transcriptional network across multiple human cell types. *BMC Genomics*, 18(1):1–12, jan 2017.

[50] A. Scialdone et al. Computational assignment of cell-cycle stage from single-cell transcriptome data. *Methods*, 85:54–61, sep 2015.

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

[52] L. Haghverdi et al. Diffusion pseudotime robustly reconstructs lineage branching. *Nature Methods 2016 13:10*, 13(10):845–848, aug 2016.

[53] L. Haghverdi, A. T. Lun, M. D. Morgan, and J. C. Marioni. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nature Biotechnology*, 36(5):421–427, 2018.

[54] F. A. Wolf et al. PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biology*, 20(1):1–9, mar 2019.

4

Tissue microenvironment partially removes signatures of developmental origin in a 3D in vitro model of cardiac endothelial cell differentiation

Endothelial cells (ECs) are ubiquitous across different organs of the body. They line the surface of different types of vessels, forming a tight barrier between a liquid and the surrounding tissue, and carry out various crucial functions. Dysfunction of ECs can therefore lead to a wide range of diseases. Importantly, the functions of ECs are highly organ-specific. For example, cardiac ECs line the blood vessels in the heart and have an essential role in nutrient transport. It is currently unknown, how ECs obtain their organ-specific characteristics. Developmental origin and tissue microenvironment are two possible determining factors, but their relative importance is unclear. This study uses human-induced pluripotent stem cells to derive ECs from two developmental origins: paraxial and cardiac mesoderm. We compare their characteristics and further integrate them into a cardiac microtissue, containing cardiomyocytes and fibroblasts, to investigate the influence of the surrounding cells on EC identity. Upon integration into the microtissue, the developmental origin is partially removed, and ECs from both developmental origins acquire an intramyocardial signature.

This chapter is based on Xu Cao*, **Maria Mircea***, Francijna E. van den Hil, Hailiang Mei, Konstantinos Anastassiadis, Christine L. Mummery, Stefan Semrau and Valeria V. Orlova. Tissue microenvironment partially removes signatures of developmental origin in a 3D in vitro model of cardiac endothelial cell differentiation. Manuscript to be submitted. (*contributed equally)

4.1 Introduction

All organs of the body rely on highly specialized cells to carry out their particular functions. Yet, there are cells with similar characteristics that appear across many different organs. A prime example are endothelial cells (ECs). ECs line the interior surfaces of vessels and form a tight barrier between a liquid, such as lymph or blood, and the surrounding tissue [1–6]. Due to their ubiquity, EC dysfunction is associated with a wide range of disease states, including atherosclerosis, diabetes, heart failure, hypertension and ischemia [7]. Despite the similarity in function, ECs from different vessels (artery, capillary, vein, lymphatic) show significant differences, reflecting the specific requirements of a particular environment. It is currently not well understood, how ECs (or other ubiquitous cell types) acquire their organ-specific characteristics. Developmental origin likely plays a role but cues from the microenvironment could be equally important. Using cardiac ECs as a model, this study explores the factors that lead to organ-specific molecular profiles.

In the adult heart, ECs constitute > 60% of nonmyocytes [8] and can be divided into two subtypes: endocardial ECs (eEC) and intramyocardial ECs (iECs). eECs form the innermost layer and serve as a barrier between blood and myocardium. iECs constitute the coronary vessels, which transport oxygen and nutrients to the heart. In the mouse heart, eECs and iECs are distinguished by specific expressions of the eEC markers Nfatc1, Npr3, Tmem100, Cdh11, Hapln1 [9–12], and the iEC markers Apln, Fabp4, Cd36 [9, 13–15]. Equivalent EC subtypes have also been identified in the human heart, by single-cell RNA sequencing (scRNA-seq)[16–18]. Some of the eEC and iEC markers identified in the mouse are conserved in humans, such as CDH11, NPR3 in eECs, and CD36, FABP4 in iECs [17].

Unlike eECs, which have a sole origin, iECs have been shown to have multiple origins. The first established origin is the proepicardium on the venous pole of the heart, which later migrates onto the heart and gives rise to the epicardium and a small number of iECs [13, 19, 20]. More recently, sinus venosus [13, 16] and endocardium [10, 15] were revealed as two additional, major sources of iECs, by lineage tracing in mice. Postnatally, the endocardium continues to give rise to iECs that generate the majority of coronary vessels in the myocardium closest to the endocardium [21].

In addition to distinct developmental origins, the cardiac microenvironment could also endow eECs and iECs with tissue specific signatures. eECs are formed initially in the heart tube and exposed to blood flow, while iECs are developed first as a vascular plexus in the myocardium and are surrounded by other cell types like cardiomyocytes and fibroblasts. Cardiomyocytes are known to produce a vast amount of vascular endothelial growth factor-A (VEGF-A) [22] and other factors like Angiopoetin-1 [23], nitric oxide (NO), endothelin-1 (ET-1), fibroblast growth factor (FGF)-2, urocortin, haemoxygenase and adenosine [24] , which are all key regulators of the EC phenotype. Other cardiac cell types, including fibroblast, vascular smooth muscle cells (VSMCs) and macrophages, also affect heart EC functions through either direct cell-cell contact or paracrine factors [25].

Most of what we know about EC development has been established in the mouse. Human *in vitro* systems provide a convenient, easily manipulated model to study human EC development. Human induced pluripotent stem cells (hiPSCs) provide an unlimited source of ECs [26, 27], which can be differentiated through different mesodermal origins including lateral plate mesoderm and paraxial mesoderm (PM) [28]. Several studies derived organ-

specific ECs from hiPSCs. Two groups claimed to have obtained brain microvascular ECs from hiPSCs through either co-differentiation or co-culture with neuronal cells [29, 30]. Our group previously developed a method to co-differentiate cardiomyocytes and ECs from hiPSCs through cardiac mesoderm (CM). These hiPSC-derived ECs expressed a number of cardiac specific genes like MEOX2, GATA4, GATA6 and ISL136, while tissue specific identities (of eEC or iEC) were still absent.

In vitro systems also allow probing the influence of the microenvironment, because different cell types can be easily combined. Recently our group established heart microtissues (MTs), a 3D cell culture model composed of ECs, cardiomyocytes and fibroblasts, all derived from hiPSCs [31, 32]. This model provides an ideal tool to investigate the influence of both the developmental origin and the cardiac specific microenvironment on the acquisition of a tissue specific EC identity. In this study, we derived ECs from hiPSCs through two mesodermal origins (CM and PM), using our established protocol [31] and an adapted protocol from the literature [28]. MTs were then generated using these two sources of ECs respectively. Interestingly, although newly differentiated ECs from two origins showed distinct identities, they became more similar after extended culture in MTs. Furthermore, based on eEC- and iEC-specific signatures extracted from a published single-cell RNA sequencing (scRNA-seq) dataset of human fetal heart [18], we observed an iEC rather than an eEC identity for both developmental origins after MT culture. In summary, this study shows that, although certain characteristics are inherited from progenitors, ECs efficiently adapt to the microenvironment and acquire new tissue-specific signatures. Our results provide new insights into the acquisition of organ/tissue-specific cell identities, which will inform the preparation of hiPSC-derived, organ specific ECs for disease modeling and drug development.

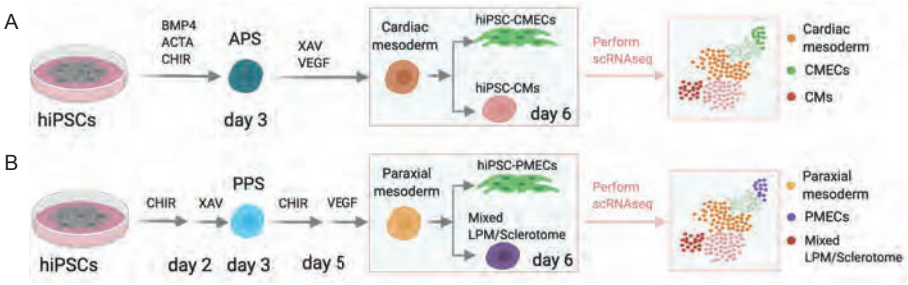


Figure 4.1: Single-cell RNA sequencing analysis of endothelial cells differentiated from cardiac and paraxial mesoderm(A-B) Schematic overview of CMEC (A) and PMEC (B) differentiation protocols until day 6. Cells were collected for scRNAseq on day 6. ACTA: activin-A. CHIR: CHIR99021. APS/PPS: anterior/posterior primitive streak. LPM: lateral plate mesoderm.

4

4.2 Results

4.2.1 Derivation of endothelial cells from different mesodermal origins

We set out to derive ECs from hiPSCs via two different mesodermal intermediates. To obtain cardiac mesoderm-derived ECs (CMECs), we used a protocol that was established previously in our group [31] (Figure 4.1A). Briefly, BMP4, Activin A (ACTA) and CHIR99021 (CHIR) were used to induce anterior primitive streak (APS) on day 3. Then, CMECs and early cardiomyocytes were induced on day 6 in the presence of XAV-939 (XAV) and VEGF. To obtain paraxial mesoderm-derived ECs (PMECs), we developed a new protocol (Figure 4.1B) based on a published approach [28]. Briefly, posterior primitive streak was derived through induction by CHIR for two days and subsequently XAV for one day. Next, PMECs and a mixed lateral plate mesoderm/sclerotome population were derived by exposure to CHIR for two days and VEGF for one day.

To facilitate the characterization of the newly developed PMEC protocol, we generated an hiPSC line harboring fluorescent reporters for PAX3 and MSGN1 ($PAX3^{Venus}$ $MSGN1^{mCherry}$) using CRISPR/Cas9 and the piggyBac transposon system. More than 70% of the cells expressed $MSGN1^{mCherry}$ on day 2 of PMEC differentiation (Figure 4.2 A, B). MSGN1 expression persisted in 50% of the cells until day 8. However, MSGN1 mRNA was only detectable on day 2 (Figure 4.2E). On day 5, around half of the cells started to express $PAX3^{Venus}$ (Figure 4.2A,C). Most of these cells also expressed $MSGN1^{mCherry}$. Both $PAX3^{Venus}$ and PAX3 mRNA were highly expressed from day 5 to day 8 (Figure 4.2A, C-E). On day 2, pan-mesoderm markers TBXT and MIXL1 were expressed in both the CMEC and the PMEC differentiation protocol. While cardiac genes (MESP1, GATA4 and NKX2-5) were exclusively expressed in CMEC differentiation, paraxial mesoderm related genes (MSGN1, TBX6, PAX3) were specifically expressed during PMEC differentiation (Figure 4.2E). Taken together, mRNA measurements and assessment of reporter fluorescence suggested that our new protocol produces ECs with paraxial mesoderm characteristics.

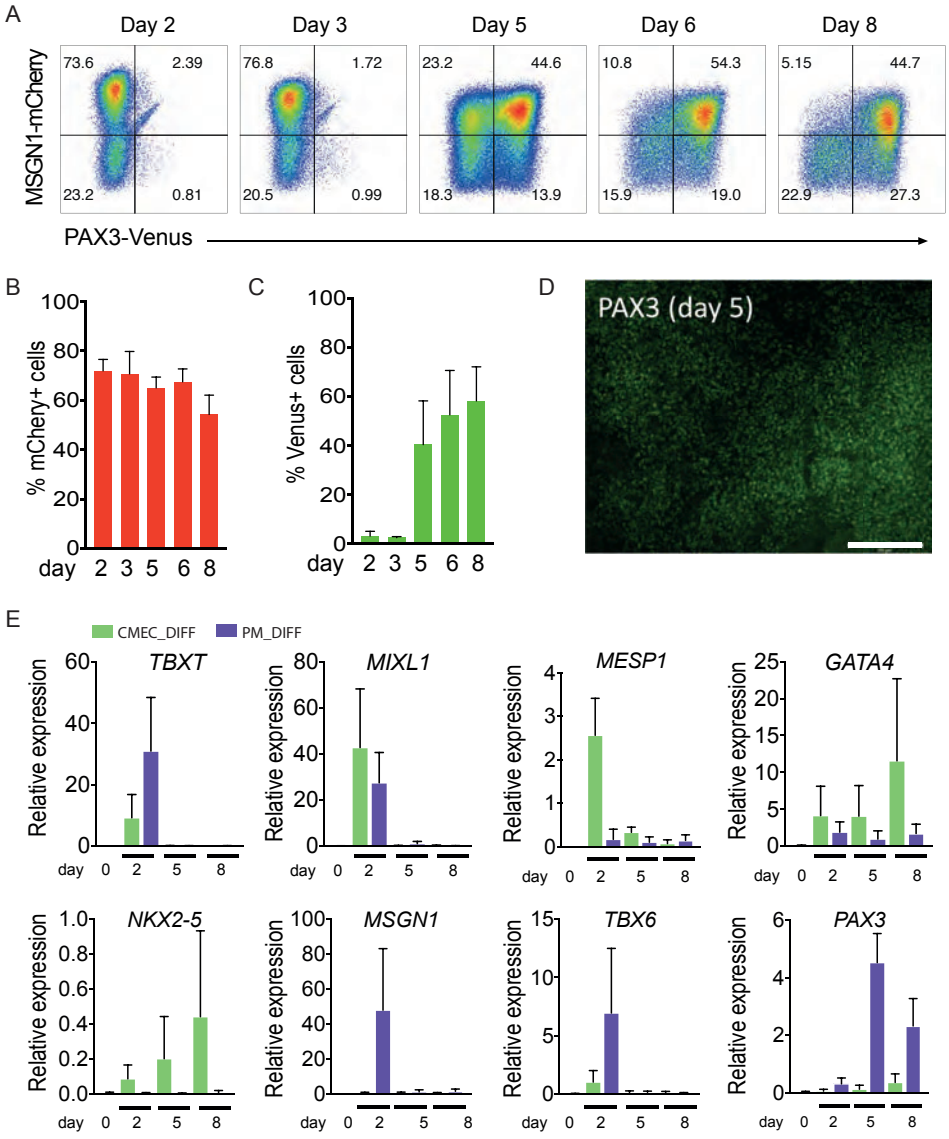


Figure 4.2: Characterization of PMEC differentiation using $MSGN1^{mCherry}$ $PAX3^{Venus}$ dual reporter line(A) FACS analysis of $PAX3^{Venus}$ and $MSGN1^{mCherry}$ expression at day 2, 3, 5, 6 and 8 of PMEC differentiation. (B-C) Quantification of percentages of mCherry+ (B) and Venus+ (C) cells in the total population by flow cytometry on day 2, 3, 5, 6 and 8. (D) Representative fluorescence image of $PAX3^{Venus}$ expression on day 5 of PMEC differentiation. Scale bar represents 200 μm. (E) Quantification of TBXT, MIXL1, MESP1, GATA4, NKX2-5, MSGN1, TBX6 and PAX3 expression by qPCR on day 0, 2, 5 and 8 of CMEC (green) and PMEC (purple) differentiation.

4.2.2 Transcriptomic profiling of CMECs and PMECs

To compare CMECs and PMECs more broadly, cells expressing the EC marker VEC were sorted on day 6 and day 8 of both protocols and characterized by bulk RNA sequencing (RNA-seq) (Figure 4.3A). Principle component analysis (PCA) showed that CMECs and PMECs clustered separately along PC1, and day 6 and day 8 were separated along PC2 (Figure 4.2B). On day 6, 3307 and 2592 genes were significantly differentially upregulated ($p_{adjusted} \leq 0.05$, fold-change ≥ 2) in CMECs and PMECs respectively (Table S1). Gene ontology (GO) analysis showed that cardiac related genes were specifically upregulated in day 6 CMECs (CMEC_D6), while genes related to skeletal system development and function were specifically upregulated in day 6 PMECs (PMEC_D6) (Figure 4.2C, Table S2). Genes involved in heart development, like GATA4, GATA5, TBX3, ISL1 and MYH6, were highly expressed in day 6 and day 8 CMECs. TBX3, ISL1 and MYH6 were upregulated from day 6 to day 8 in CMECs. Essential genes for skeletal muscle development like PAX3, TBX1, FOXC1, EYA1 and MEOX1 were majorly expressed in day 6 and day 8 PMECs. FOXC1 and EYA1 were upregulated from day 6 to day 8, while TBX1 and MEOX1 were downregulated (Figure 4.2D). In summary, unbiased expression analysis by bulk RNA-seq confirmed the respective mesodermal origins of the two derived EC cultures.

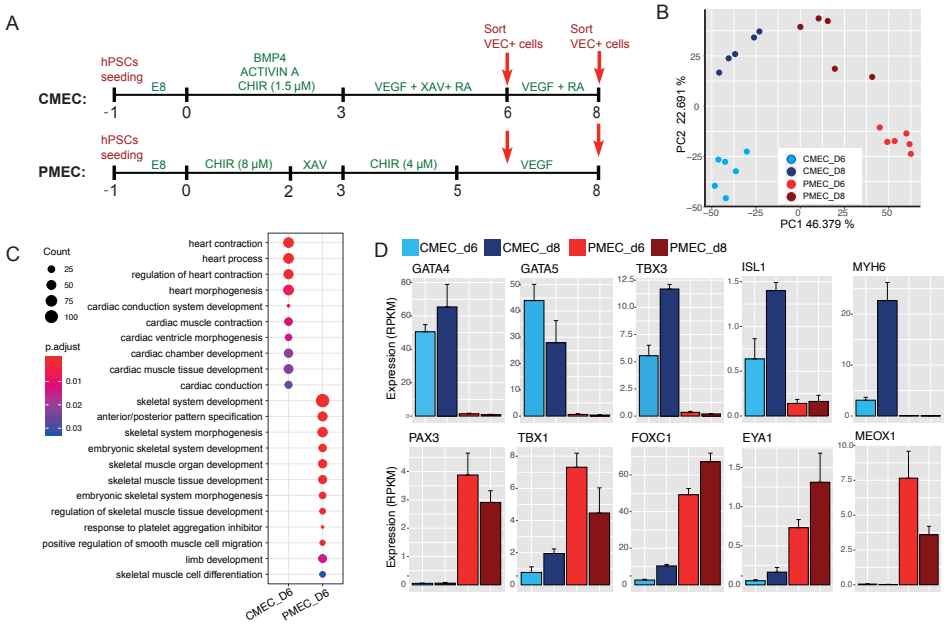


Figure 4.3: Characterization of hiPSC-ECs differentiated using CMEC and PMEC protocols(A) Schematic overview of CMEC and PMEC differentiation protocols from day -1 to day 8. VEC+ cells were sorted on day 6 and 8 from both protocols for bulk RNAseq. (B) PCA analysis of hiPSC-ECs sorted on day 6 and 8 of CMEC and PMEC protocol. (C) GO enrichment analysis for differentially expressed genes between CMECs and PMECs on day 6 of differentiation. The Complete list of GO terms can be found in Table S2. Color represents the enrichment p-value adjusted for multiple hypothesis testing and dot size represents the number of genes mapped to the GO term. (D) Normalized gene expression levels (RPKM) of cardiac and skeletal-related genes in CMECs and PMECs on day 6 and 8. Error bars indicate standard deviation.

4.2.3 Characterization of CMEC and PMEC differentiation by single-cell RNA-seq

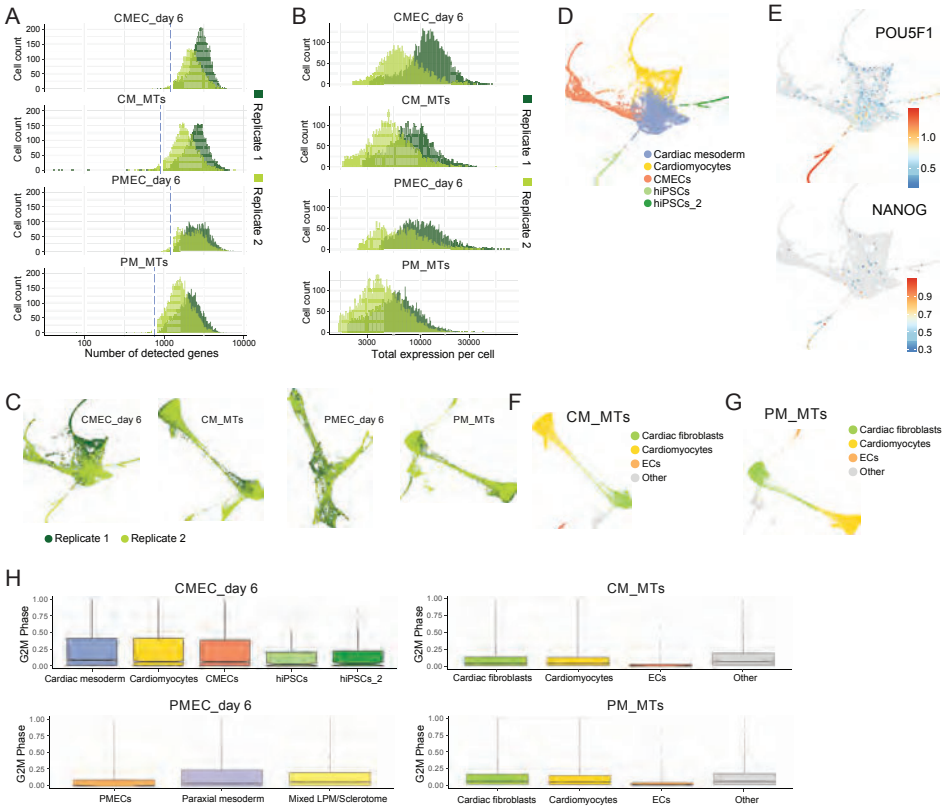


Figure 4.4: Quality control of scRNAseq datasets (A-B) Distribution of number of detected genes (A) and total expression (B) in each cell of the scRNAseq datasets. Dotted blue line indicates quality control threshold for datasets. Two different batches are labelled with different colors. (C) Two different batches of cells collected for each scRNAseq dataset are visualised using force-directed graph layout. (D) scRNAseq data of CMECs on day 6 is visualized using PAGA. Five cell clusters were identified and labelled with different colors. (E) Expression of pluripotency genes POU5F1 and NANOG in theCMEC dataset on day 6 is shown in PAGA plot. Color represents log transformed expression value. (F-G) scRNAseq data of CM_MTs (F) and PM_MTs (G) are visualized using PAGA. Four cell clusters were identified. Cluster labelled with "Other" was excluded in the downstream analysis in both datasets. (H) Boxplot of G2M phase-score in individual clusters of each dataset.

To reconstruct the differentiation trajectories of ECs, single-cell RNAseq (scRNA-seq) was performed on day 6 of CMEC and PMEC differentiation for two independent biological replicates (Figure 4.1, A-B, Figure 4.4). The replicates appeared highly similar in a low-dimensional representation (Figure 4.4C) and were therefore combined for further analysis. Undifferentiated hiPSCs remaining in the culture were excluded from further analysis (Figure 4.4D-E). In the CMEC differentiation data set, cells were grouped in 3 clusters, which we identified as cardiac mesoderm, cardiomyocytes and CMECs by marker gene analysis (Figure 4.5A, Table S3). The cardiac mesoderm cluster was characterized by mesoderm and early cardiac genes, such as MESP1, SMARCD3, ABLIM1, TMEM88, ISL1, MYL5, as

well as the cell cycle-related genes CDK6 and NEK2. The CMEC cluster was characterized by EC markers (CDH5, CD34, KDR, HEY2, TEK, TIE1, ACVRL1, SOX17, ENG, ICAM2, PECAM1). Cardiomyocytes were identified by expression of cardiomyocyte-specific genes, including MYL4, TNNI1, MYL7, ACTA2, TNNT2, HAND2 and NKX2-5 (Figure 4.5B, 4.6A-B). Pseudo-time analysis showed that both CMECs and cardiomyocytes differentiated from cardiac mesoderm, and CMECs progressed further compared to cardiomyocytes (Figure 4.5C).

In the PMEC differentiation data set, all cells were divided into 3 clusters, which were interpreted as paraxial mesoderm, PMECs and mixed lateral plate mesoderm (LPM)/sclerotome (Figure 4.5D, Table S2) using marker gene analysis (Table S3). The paraxial mesoderm cluster was characterized by posterior primitive streak and dermomyotome genes, such as MEOX1, PDGFRB, SIX1, CRABP2, NR2F1, EYA1, FOXC1 and PAX3. PMECs were characterized by EC markers, like ETV2, CDH5, CD34, KDR, ENG, SOX17, PLVAP, APLN, NRP1. The mixed LPM/sclerotome cluster was characterized by LPM and sclerotome specific genes, such as TMEM88, HAND1, TNNI1, PRRX1, ACTA2, DES, FOXH1, LEF1 and JAG1 (Figure 4.5E, 4.6C-D). Pseudo-time analysis showed that both PMECs and LPM/Sclerotome developed from paraxial mesoderm (Figure 4.5F). All in all, clustering and marker gene analysis of the scRNA-seq data confirmed the bulk RNA-seq results and revealed high similarity in developmental dynamics between the two differentiation protocols.

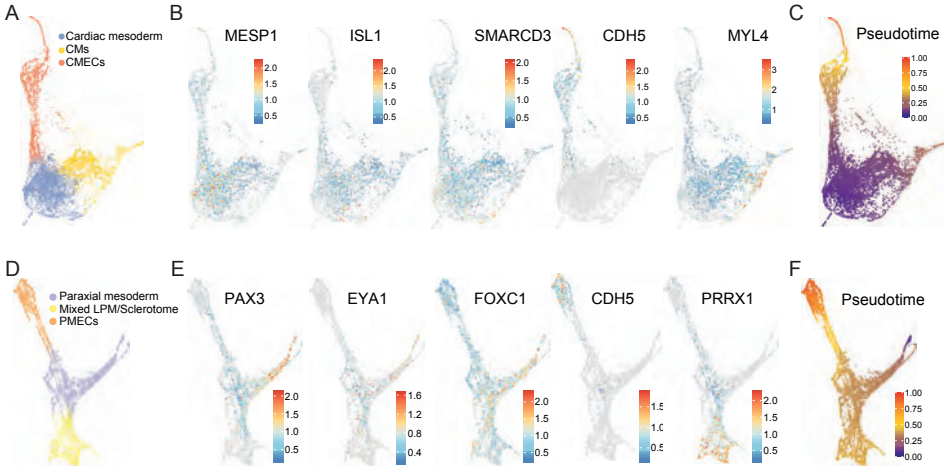


Figure 4.5: Single-cell RNA sequencing analysis of endothelial cells differentiated from cardiac and paraxial mesoderm (A) scRNAseq data of CMECs on day 6 is visualized using a force-directed graph layout (FGL). Three clusters of cells were identified. (B) FGL plots show expression (log transformed) patterns of MESP1, ISL1, SMARCD3, CDH5, MYL4 in CMEC population. Color represents log transformed expression value. (C) Pseudotime analysis of CMEC dataset on day 6 of differentiation. (D) scRNAseq data of PMECs on day 6 is visualized using FGL. Three clusters of cells were identified. (E) FGL plots show expression (log transformed) patterns of PAX3, EYA1, FOXC1, CDH5, PRRX1 in PMEC population. Color represents log transformed expression value. (F) Pseudotime analysis of PMEC dataset on day 6 of differentiation.

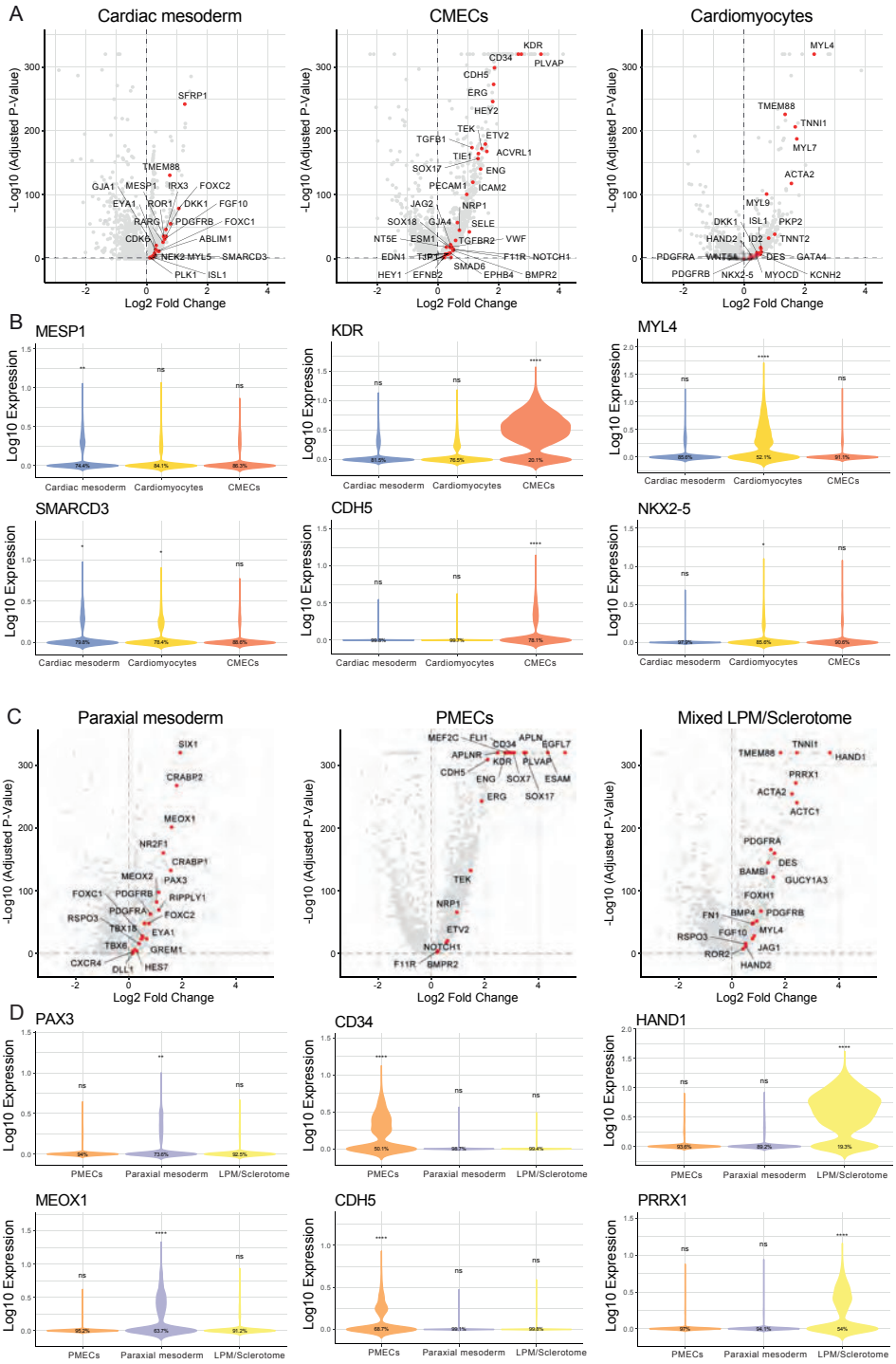


Figure 4.6: Single-cell RNA-sequencing analysis of CMEC and PMEC datasets on day 6 (Caption on the next page.)

4.2.4 hiPSC-ECs acquired organ-specific signatures in a cardiac microenvironment

Being able to produce ECs with properties corresponding to their mesodermal origins finally enabled us to test in how far the cellular microenvironment can either reinforce or reverse this specification. Specifically, we set out to mimic the cardiac microenvironment in vitro using a protocol for creating cardiac microtissues (MTs), which was published previously by our group [32]. Briefly, CD34+ CMECs or PMECs were sorted on day 6 and combined with hiPSC-derived cardiomyocytes (hiPSC-CMs) and hiPSC-derived fibroblasts (hiPSC-CFs) at a ratio of 3:14:3 to form MTs. MTs made from CMECs (CM_MTs) and PMECs (PM_MTs) were collected after 21 days for scRNA-seq (Figure 4.7A). Two independent biological replicates were deemed highly similar (Figure 4.4C) and were therefore combined for further analysis. Both CM_MTs and PM_MTs datasets were divided into three clusters that correspond to hiPSC-CFs, hiPSC-CMs and hiPSC-ECs (Figure 4.7B). Marker genes identified for each cluster, confirmed the cluster identities (Table S6).

In both cases, a fourth cluster of cells with an uninterpretable signature was ignored (Figure 4.4F-G). Next, we compared the CMECs from monoculture differentiation on day 6 (CMEC_day 6) with the CMECs in MTs (CMECs_MT). Most intramyocardial makers including CLDN5, GMFG, APLNR, CD36, NOTCH4, OIT3, IGFBP3, ARHGAP18, A2M and BCAM and several endocardial markers (TFPI2, EDN1, ECE1, FOXP1, FOXC1) were upregulated in CMEC_MT (Figure 4.7C-D, 4.8A, Table S7). However, the differences in endocardial marker expression were smaller compared to intramyocardial markers (Figure 4.7C-D, 4.8A, Table S7). Then, we compared PMECs from monoculture differentiation on day 6 (PMECs_day 6) to PMECs in MTs (PMECs_MT). Most intra-myocardial markers, including CLDN5, GMFG, NOTCH4, IGFBP3, ARHGAP18, A2M and BCAM, and some endocardium markers (TFPI2, EDN1, TEK, ECE1, FOXP1, ALDH2, FZD6) were upregulated in PMEC_MT (Figure 4.7C, 4.8B, Table S7).

Compared to PMECs_MT, CMECs_MT expressed higher levels of intra-myocardial markers, especially APLNR, CD36, OIT3, ARHGAP18, A2M, BCAM which were barely expressed in the majority of PMECs_MT cells (Figure 4.7F, Table S8). Although most endocardium markers were also higher in CMECs_MT than in PMECs_MT, their average expression levels were lower in general compared to intra-myocardial markers (Figure 4.8C, Table S8). Notably, most endocardial makers (including CDH11, FOXC1, FZD6, TMEM100 and NPR3) were barely expressed in both CMECs_MT and PMECs_MT, (Figure 4.8C). Overall, heart tissue-specific genes, especially intramyocardial markers, were upregulated in hiPSC-ECs upon extended culture in the cardiac microenvironment of the MT.

Figure 4.6: scRNAseq analysis of CMEC and PMEC datasets on day 6 (Figure on previous page.) (A) Volcano plots showing fold changes and adjusted p-values for differential gene expression between a specific cluster in the CMEC dataset and all other cells in that dataset. Representative, significantly up-regulated genes ($P_{adjusted} \leq 0.05$ & fold change ≥ 1.2) are labelled in red. (B) MESP1, SMARCD3, KDR, CDH5, MYL4 and NKX2-5 expression (log transformed) in three clusters of CMEC dataset on day 6. (C) Volcano plots showing fold changes and p-values for differential gene expression between a specific cluster in the CMEC dataset and all other cells in that dataset. Representative significantly up-regulated genes ($P_{adjusted} \leq 0.05$ & fold change ≥ 1.2) are labelled in red. (D) PAX3, MEOX1, CD34, CDH5, LEF1 and PRRX1 expression (log transformed) in three clusters of the PMEC dataset on day 6.

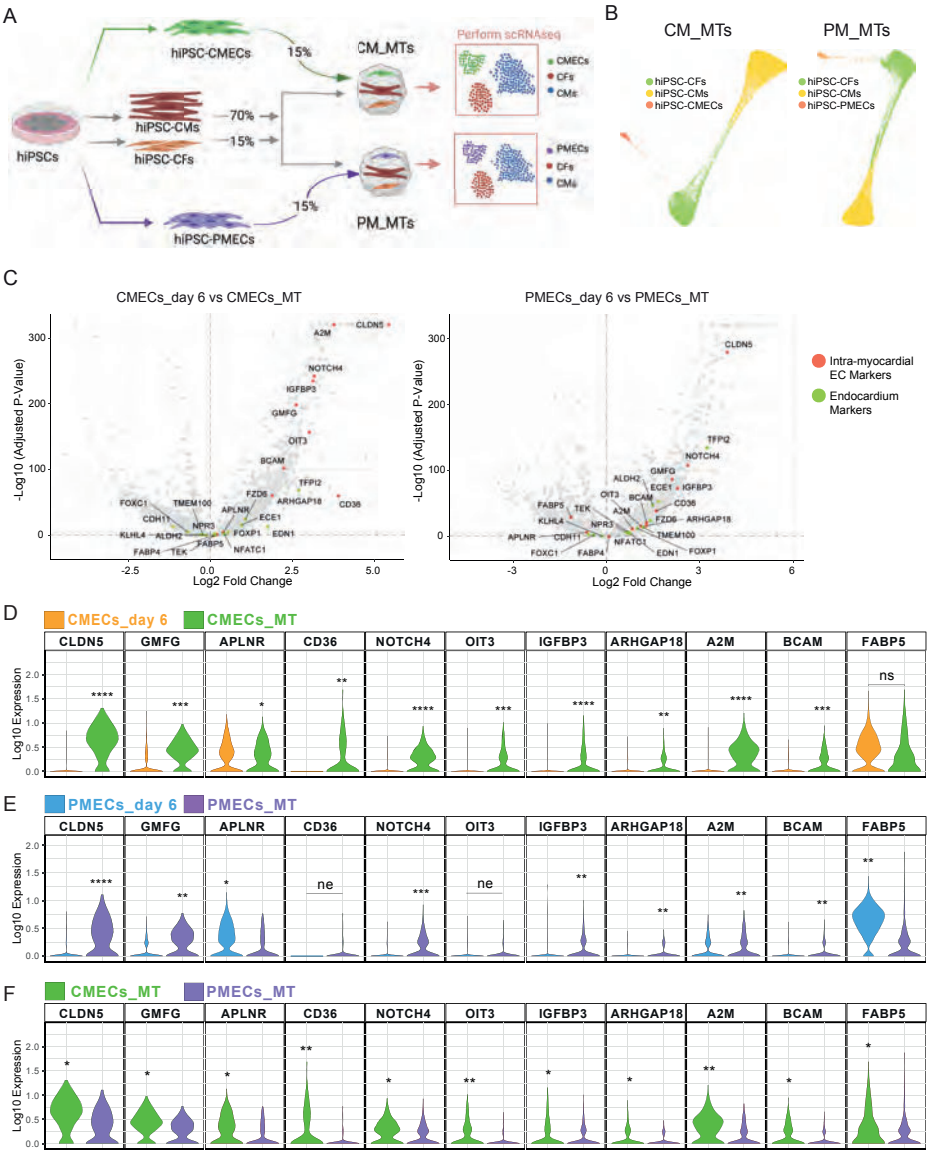


Figure 4.7: hiPSC-ECs acquired organ-specific signatures in a cardiac microenvironment (A) Schematic overview; generation of cardiac microtissues (MTs) from hiPSC-CMs, hiPSC-CFs and hiPSC-ECs. CMECs and PMECs were used for CM_MTs and PM_MTs, respectively. MTs were collected after 21 days for scRNAseq. (B) scRNAseq data of CM_MTs (left) and PM_MTs (right) were visualized using force-directed graph layout. Three clusters of cells were identified in both datasets. (C) Volcano plot shows fold changes and p values of all genes tested between two selected clusters: CMECs_day 6, CMECs_MTs (left), and PMECs_day 6, PMECs_MTs (right). Representative intra-myocardial and endocardial markers that are differentially expressed ($P_{adjusted} \leq 0.05$) are labelled in red and green respectively. (D-F) Differential expression tests between CMECs_day 6 and CMECs_MT (D), PMECs_day 6 and PMECs_MT (E), CMECs_MT and PMECs_MT (F) for representative intra-myocardial EC markers. ns: $p \geq 0.05$; * p ≤ 0.05; ** p ≤ 1e-10; *** p ≤ 1e-100; **** p ≤ 1e-200. Clusters with higher expression value were indicated with stars. ne: not expressed (0 counts) in ≥ 85% of cells in both groups.

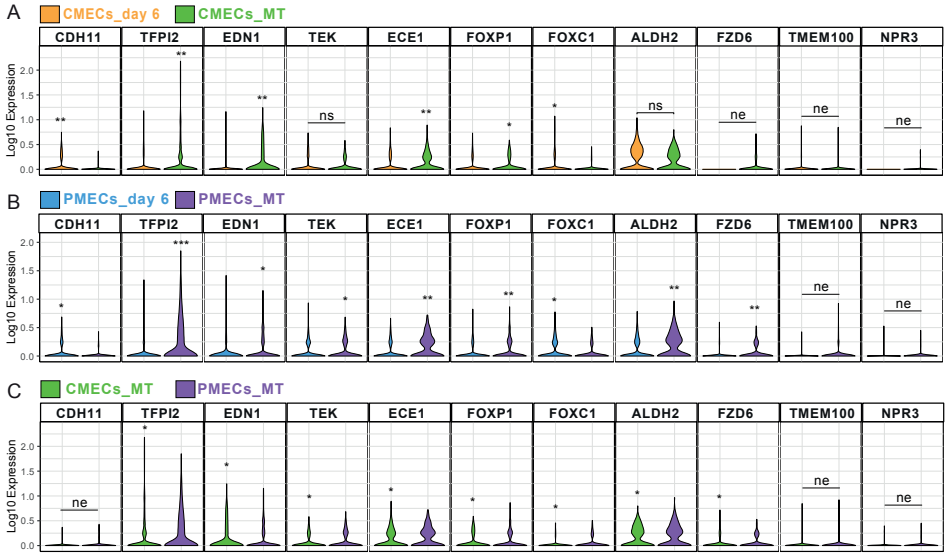


Figure 4.8: Characterization of endocardial signatures of hiPSC-ECs on day 6 and in MTs (A-C) Differential expression tests between cluster CMECs_day 6 and CMECs_MT (C), PMECs_ day 6 and PMECs_MT (D), CMECs_MT and PMECs_MT (E) for representative endocardial EC markers. ns: $p \geq 0.05$; * p ≤ 0.05; ** p ≤ 1e-10; *** p ≤ 1e-100; **** p ≤ 1e-200. Clusters with higher expression value were indicated with stars. ne: not expressed (0 counts) in ≥ 85% of cells in both groups.

4.2.5 Extraction of organ-specific signatures of human fetal heart ECs from a published scRNA-seq dataset

To assess how closely ECs in MTs assume an organ-specific identity, we sought to compare them to primary heart ECs. To that end, we re-analyzed a published scRNA-seq dataset of the fetal human heart (EGAS0000100399) [18] and identified the expression signature of heart ECs. The whole cell population was divided into 14 clusters. Our interpretation of these cluster deviated from the published annotation in two cases (Figure 4.9A): 1. The original endothelium/pericytes/adventia cluster (cluster 10) was reannotated as intramyocardial ECs, based on differentially expressed markers such as A2M, CD36, APLNR, ARHGAP18, IGFBP3, CLDN5, FABP4 and FABP5. 2. The cluster annotated as capillary endothelium in the original publication (cluster 0) was reannotated as endocardium, due to the presence of differentially expressed markers like NPR3, ALDH2, CDH11, ECE1, TMEM100, FOXC1 and EDN1 (Figure 4.9B, Table S5). Supporting the differential expression test, UMAP visualization of representative intra-myocardial and endocardial markers showed specific expression in the respective clusters (Figure 4.9C-D). In conclusion, our clustering and differential expression analysis revealed distinct endothelial tissues in the fetal human heart.

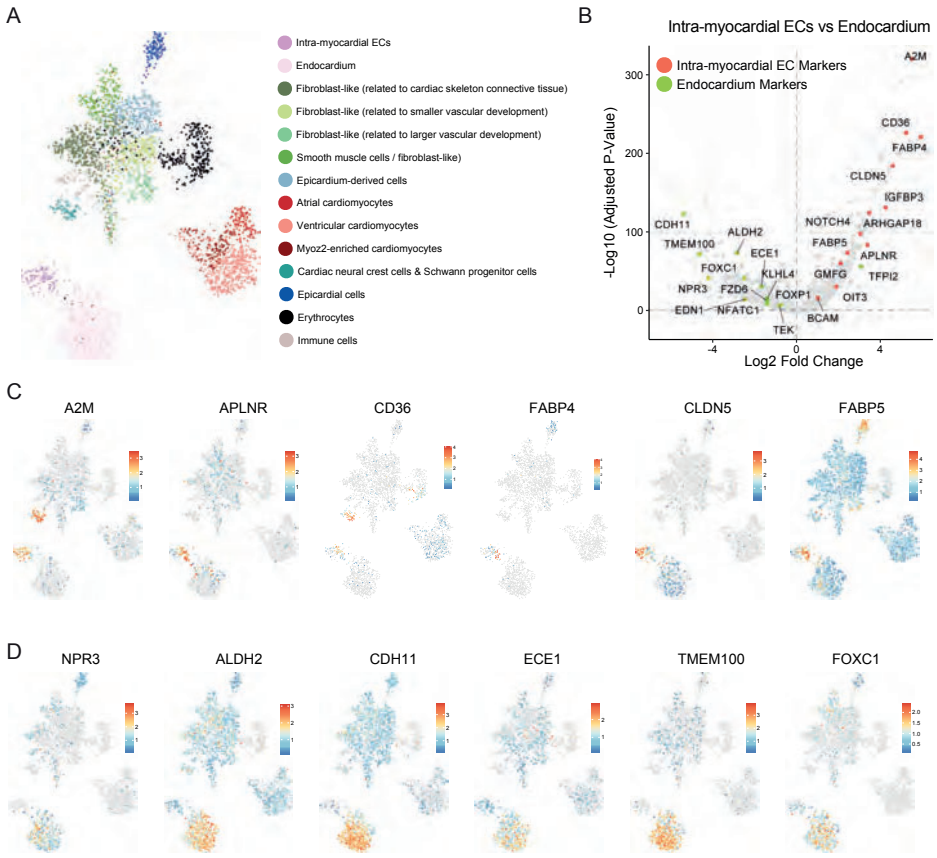


Figure 4.9: Re-analysis of a published scRNAseq dataset to identify organ specific signatures of human fetal heart ECs (A) Dimensionality reduction of scRNAseq data of human fetal heart using Uniform MAnifold Approximation and Projection (UMAP). 14 cell clusters were identified and named based on their identities according to the original publication except for two EC clusters: intra-myocardial ECs and endocardium. (B) Volcano plot showing fold changes and adjusted p-values for differential gene expression between intra-myocardial ECs and endocardium. Representative differentially expressed genes ($P_{adjusted} \leq 0.05$) that are known to be intra-myocardial or endocardial markers are labelled in red and green, respectively. (C-D) Expression (log transformed) of representative intra-myocardial EC markers (C) and endocardium markers (D) in individual cells (UMAPs).

4.2.6 Both CMECs and PMECs acquired intra-myocardial identity in MT culture

To obtain a clear view of the similarities between ECs in MTs and primary fetal heart ECs, we combined the CM_MT or PM_MT dataset with the published *in vivo* data (Figure 4.10A-B). In both cases, hiPSC-CFs in MTs (CF_MT) cluster together with fetal heart fibroblast-like cells; hiPSC-CMs in MTs (CM_MT) cluster together with fetal heart ventricular cardiomyocytes. Notably, both CMECs_MT and PMECs_MT cluster together with fetal heart intra-myocardial ECs rather than endocardium (Figure 4.10A-D).

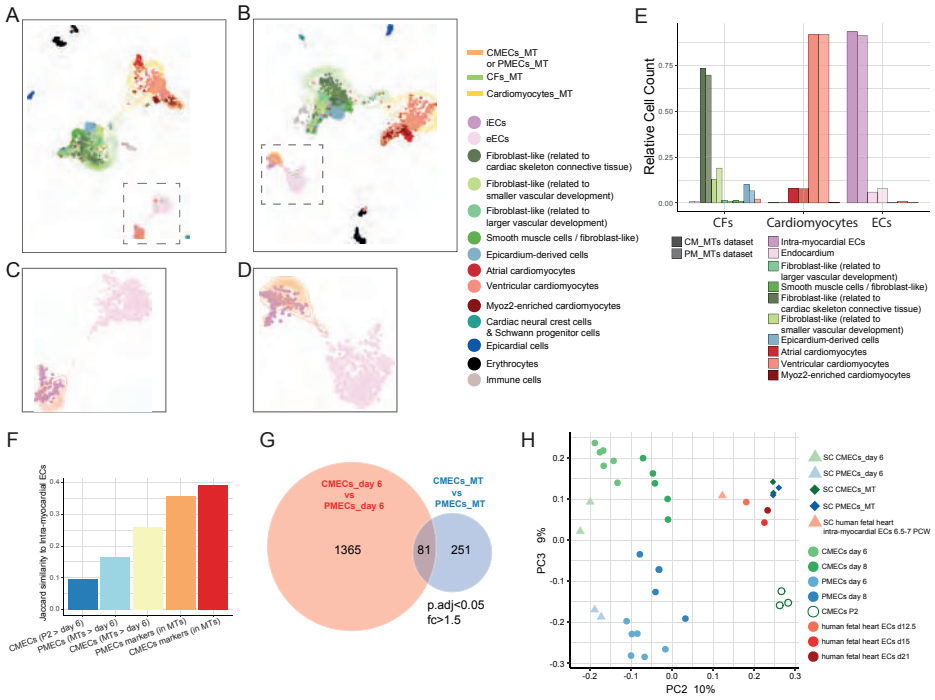


Figure 4.10: Both CMECs and PMECs acquired intra-myocardial identity in a cardiac microenvironment (A-B) The published human fetal heart scRNAseq dataset (EGAS0000100399) was combined with the CM_MTs (A) or PM_MTs (B) dataset and visualized using Uniform Manifold Approximation and Projection (UMAP). EC clusters are marked with dashed boxes. (C-D) Zoom-ins of EC clusters marked in (A-B), showing overlap of CMECs_MT (C) and PMECs_MT (D) with intramyocardial ECs of the fetal heart dataset. Cell clusters are labeled with different colors. Cells in the fetal heart dataset are represented with dots and cells in MT datasets are represented with contour lines. (E) The nearest clusters to the published dataset were identified for each cell of CM_MTs and PM_MTs. The result of knn-assignment is visualized in the bar plot. Cells from CM_MTs and PM_MTs are labelled in dark and light colors respectively. Identities of each cell clusters of the published dataset are shown with different colors. (F) Jaccard similarity to the amrker genes of intra-myocardial ECs from the published fetal heart dataset was calculated for each group of genes. CMEC markers (in MTs): specific markers of CMECs within CM_MT dataset; PMEC markers (in MTs): specific markers of PMECs within PM_MT dataset; CMECs (MTs ≥ day 6): differentially expressed genes (DEGs) that are higher in CMECs_MT compared to CMECs_day 6; PMECs (MTs ≥ day 6): DEGs that are higher in PMECs_MT compared to PMECs_day 6; CMECs (P2 ≥ day 6): DEGs that are higher in passage two CMECs compared to CMECs_day 6. (G) Venn diagram showing numbers and overlap of DEGs ($P_{adjusted} \leq 0.05$ and foldchange ≥ 1.5) between CMECs and PMECs from day 6 (in red) and MTs (in blue). (H) PCA of different EC populations in scRNAseq (triangle and diamond) and bulk RNAseq (circle) datasets using the marker genes of the intra-myocardial ECs from the in vivo data set. Average expression values of all cells in the cluster were used for the scRNAseq data.

To quantify our observation, we calculated the distances (in expression space) between each cell in MTs and the fetal heart dataset. This calculation showed that CF_MT cells are closest to fibroblast-like cells in vivo (related to cardiac skeleton connective tissue); CM_MT cells are closest to ventricular cardiomyocytes; and CMECs_MT as well as PMECs_MT are closest to intra-myocardial ECs in vivo. Annotating the in vitro cells based on the closest in vivo neighbors revealed that cell type identities were very similar in CM_MTs and PM_MTs (Figure 4.10E). Correspondingly, the set of markers of the CMECs_MT and

the PMECs_MT cluster showed a high overlap (Jaccard similarity) with the markers of intra-myocardial ECs we extracted from the in vivo data set. The gene set upregulated in CMECs_MT (compared to CMECs_day 6) had a higher overlap with intra-myocardial EC markers than the set of genes upregulated in PMECs_MTs (compared to PMECs_day 6). Genes that were upregulated in passage two CMECs compared to CMECs_day 6 overlapped the least with intra-myocardial EC markers (Figure 4.10F). CMECs_MT and PMECs_MT thus both resembled intra-myocardial ECs but a difference between the two differentiation systems remained. To quantify this difference directly, we used differential gene expression analysis. On day 6 of differentiation, 1446 genes were differentially expressed between CMECs and PMECs (Table S9), while only 332 genes were differentially expressed between CMECs_MT and PMECs_MT (Table S8). 81 genes were shared between the two sets (Figure 4.10G).
Next, all EC clusters from bulk and single cell RNA-seq datasets were combined and visualized using principal component analysis (PCA) (Figure 4.10H). CMECs_day 6 and PMECs_day 6 cluster far part, while CMECs_MT and PMECs_MT clustered closely together. Bulk and single cell RNA-seq samples clustered together for both CMECs and PMECs. CMECs_MT and PMECs_MT were found close to fetal heart intra-myocardial ECs and fetal heart ECs sequenced in our previous study ([32](Figure 4.10H). All in all, the integrated analysis of the in vitro and in vivo data sets revealed that the cardiac microenvironment, mimicked by cardiac MT, partially removed expression differences due to distinct mesodermal origins.

4.3 Discussion

In this study we set out to delineate the possible factors that can confer organ-specific characteristics to cardiac ECs. In principle, those characteristics could be inherited from developmental precursors or induced by the tissue microenvironment. We investigated the contributions from both factors using hiPSC-derived ECs and cardiac MTs as tools.
To model different developmental origins, CMECs and PMECs were derived through CM and PM respectively. CMECs obtained a clear cardiac phenotype but no tissue-specific (eEC or iEC) identity. PMECs expressed a number of limb/skeletal muscle specific genes. Both protocols thus resulted in early or immature organ-specific identities. However, we cannot conclude that these identities were conferred by the respective mesoderm precursors in an entirely cell-autonomous manner, since other cell types were co-differentiated in both protocols and could have influenced the ECs. To exclude the contribution from other cells in the culture, early EC progenitors would have to be purified and further differentiated to establish whether organ-specific identities are present.
To model the influence of cell-extrinsic factors, we took advantage of our cardiac MT model which mimics the heart-specific microenvironment better than other in vitro models [3]. Both CMECs and PMECs acquired an iEC identity after incorporation into MTs and continued culture. This result supports that high plasticity [5, 6] allows ECs to adapt efficiently to signals from the microenvironment.
Notably, the iEC identity was more pronounced in CMECs_MT than in PMECs_MT, as it might take extra steps and therefore more time for PMECs to adopt a cardiac fate, compared to CMECs. Importantly, extended monoculture of CMECs did not result in an iEC identity,

suggesting that the heart-specific microenvironment, modeled by cardiac MTs, caused the specification. Some of the genes upregulated during MT culture are essential for EC function. For example, CLDN5 and NOTCH4 were among the most highly upregulated genes in both protocols. CLDN5 is critical for the tight junction and barrier functions of endothelium [8], while NOTCH signaling plays key roles in both vascular morphogenesis and adult endothelium homeostasis [9]. These results might indicate that ECs also mature functionally in MTs.

ECs have been incorporated into hPSC-derived organoids of liver [10, 11], intestine [12], brain [13–15], pancreas [16] and kidney [17, 18] by either codifferentiation or aggregation. In most studies, organ/tissue-specific identities were barely investigated. Camp et al. [33] used an approach similar to our MTs, combining hepatic, stromal, and endothelial cells to make liver organoids [11]. Interestingly, that study reported transcriptomic changes in ECs upon coculture in liver organoids, suggesting further maturation. Similarly, coculture with neuronal cells was found to endow hPSC-derived ECs with a brain microvascular identity [34]. Together with these previous studies, our results support the notion that the microenvironment plays an essential role in developing an organ/tissue-specific identity of ECs. More work is needed to investigate the molecular mechanisms underlying the influence of the microenvironment, which likely include direct cell-cell contact, secretion of paracrine factors and modulation of the extracellular matrix.

We hope that our findings will guide the derivation of organ-specific ECs in the future and lay the foundation for various biomedical applications.

4.4 Materials and Methods

4.4.1 Experimental Methods

hiPSC culture

The NCRM1 hiPSC line (NIH) was used for all experiments in this study. The cells were cultured in TeSR-E8 on Vitronectin XF and was routinely passaged once a week using Gentle Cell Dissociation Reagent (all from Stem Cell Technologies). Prior to targeting, were grown on feeders in maintenance medium. RevitaCell (Life Technologies) was added to the medium (1:200) after every passage to enhance viability after single cell passaging with TrypLE (Life technologies).

Generation of $PAX3^{Venus}MSGN1^{mCherry}$ hiPSC dual reporter line

The $PAX3^{Venus}$ reporter was introduced first by CRISPR/Cas9 as follows: NCRM1 hiPSCs were passaged using split ratios of 1:2 or 1:3. The cells were transfected in 60 mm dishes after reaching 60-70% confluency. For transfection, 20 μl lipofectamine (Invitrogen), 8 μg repair template and 8 μg sgRNA/Cas9 plasmid were diluted in 600 μl Opti-MEM and added to each 60 mm dish. After 18 ho the medium was changed to maintenace medium. After another 6 h G-418 (50 μg/ml) selection was started and kept for 1 week. Surviving cells were cultured in maintenance medium and grown in 6-well plates for the transfection with a Flp recombinase expression vector to remove the neomycin cassette. 300 μl Opti-MEM containing 10 μl lipofectamine and 4 μg CAGGs-Flpo-IRES-puro plasmid was added per well for 18 h. Puromycin (0,5 μg/ml) selection was started 24 h post transfection and lasted for 2 days. Once recovered, individual clones were expanded by limiting dilution in 96-well plates. Targeted clones were identified by PCR and Sanger sequencing (BaseClear).
Next, the $MSGN1^{mCherry}$ reporter was integrated into the genome of the $PAX3^{Venus}$ reporter line using a piggyBac transposon system created by Katrin Neumann, Konstantinos Anastassiadis (Biotechnology Center TU Dresden).

Endothelial cell differentiation from hiPSCs

Endothelial cells with a cardiac mesoderm origin were induced from hiPSCs in a monolayer using the CMEC protocol as described previously [31]. Briefly hiPSCs were split at a 1:12 ratio and seeded in 6-well plates coated with 75 μg/mL growth factor reduced Matrigel (Corning) on day -1. On day 0, cardiac mesoderm was induced by changing from TeSR-E8 medium to BPEL medium [35], supplemented with 20 ng/mL BMP4 (R&D Systems), 20 ng/mL ACTIVIN A (Miltenyi Biotec) and 1.5 μM CHIR99021 (Axon Medchem). From day 3, the cells were grown in BPEL medium supplemented 5 μM XAV939 (Tocris Bioscience) and 50 ng/ml VEGF (R&D Systems), with or without 100 μM retinoic acid (MERCK) and the medium was refreshed every 3 days.
To derive endothelial cells with paraxial mesoderm origin, hiPSCs were split at a 1:12 ratio and seeded in 6-well plates coated with 75 μg/mL growth factor reduced Matrigel on day -1. On day 0, paraxial mesoderm was induced by changing from TeSR-E8 to BPEL medium, supplemented with 8 μM CHIR99021. On day 2, the medium was exchanged with BPEL medium supplemented with 5 μM XAV939. On day 3, the medium was exchanged with BPEL medium supplemented with 4 μM CHIR99021. From day 5 onwards, cells were grown in BPEL medium supplemented with 50 ng/ml VEGF and the medium was refreshed every 3 days.

Fluorescence-activated cell sorting

Cells were dissociated with TrypLE on day 6 and 8 of the CMEC or PMEC protocol and stained with a VE-Cadherin (VEC) antibody conjugated with PE (R&D Systems). Then VEC-positive cells were purified using a FACSAria III cell sorter. Total RNA was extracted right after sorting using the NucleoSpin®RNA kit (Macherey-Nagel).

Generation of 3D Cardiac Microtissues (MTs)

Cardiac MTs were generated from hiPSC-derived ECs (hiPSC-ECs), hiPSC-derived cardiac fibroblasts (hiPSC-CFs) and hiPSC-derived cardiomyocytes (hiPSC-CMs) as previously described [32]. Briefly, on day 6 of CMEC or PMEC differentiation, CD34-positive hiPSC-ECs were isolated using a Human cord blood CD34 Positive selection kit II (StemCell Technologies) following the manufacturer's instructions. On the day of MT formation, freshly isolated hiPSC-ECs and cultured hiPSCs-CFs and hiPSC-CMs were combined together (70% hiPSC-CMs, 15% hiPSC-ECs and 15% hiPSCs-CFs) at a concentration of 5000 cells per 50μl in BPEL medium supplemented with VEGF (50 ngml) and FGF2 (5 ngml). Cell suspensions were seeded in V-bottom 96-well microplates (Greiner bio-one) and centrifuged for 10 min at 1100 rpm. MTs were incubated at 37 $^{\circ}$C, 5% CO_2 for 21 days. The medium was refreshed every 3 to 4 days. single-cell RNA-sequencing analysis of MTs was performed after 21 days.

Flow cytometry analysis

Cells were dissociated with TrypLE, washed once with flow cytometry buffer (PBS containing 0.5% BSA and 2 mM EDTA) and analyzed on a MACSQuant VYB (Miltenyi Biotech) equipped with a violet (405 nm), blue (488 nm) and yellow (561 nm) laser. The results were analyzed using Flowjo v10 (FlowJo, LLC).

Quantitative Real-Time Polymerase Chain Reaction (qPCR)

Total RNA was extracted using the NucleoSpin®RNA kit according to the manufacturer's protocol. cDNA was synthesized using an iScript-cDNA Synthesis kit (Bio-Rad). iTaq Universal SYBR Green Supermixes (Bio-Rad) and Bio-Rad CFX384 real-time system were used for the PCR reaction and detection. Relative gene expression was determined according to the standard ΔCT calculation and normalized to housekeeping gene RPL37A.

4.4.2 Bulk RNA sequencing (RNA-seq) and analysis

Bulk RNAseq of passage two CMECs (CMECs P2) and human fetal heart ECs from week 12.5 (W12.5), W15 and W21 of gestation were performed in our previous study [32] and obtained from GEO (accession number GSE116464).

Bulk RNAseq of day 6 and 8 of CMEC and PMEC differentiation were performed by BGI (Shenzhen, China) using the Illumina Hiseq4000 sequencer (100bp paired end reads). Raw data was processed using the LUMC BIOPET Gentrap pipeline (https://github.com/biopet/biopet), which comprises FASTQ preprocessing, alignment and read quantification. Sickle (v1.2) was used to trim low-quality read ends (https://github.com/najoshi/sickle). Cutadapt (v1.1) was used for adapter clipping [36]. Reads were aligned to the human reference genome GRCh38 using GSNAP (gmap-2014-12-23) [37, 38] and htseq-count (v0.6.1p1) was for quantification using the Ensembl v87 annotation [39]. Biases related to gene length and GC content were corrected by conditional quantile normalization using the R package

cqn (v1.28.1) [40]. Genes were excluded if read count was below 5 in ≥90% of the samples. Differentially expressed genes were identified using a generalized linear model as implemented in edgeR (3.24.3) [41]. P-values were adjusted for multiple hypothesis testing using the Benjamini-Hochberg procedure and q-values of ≤ 0.05 were considered significant. Analyses were performed using R (version 3.5.2). The Principal Component Analysis (PCA) plot was generated with the built-in R functions 'prcomp'. Spearman correlation between samples was calculated using the 'cor' function and the correlation heatmap was generated with a heatmap function from the NMF package. Gene ontology term enrichment was performed using the compareCluster function of the clusterProfiler package (v3.10.1) [42] and q-values ≤ 0.05 were considered significant.

4.4.3 Single-cell RNA sequencing and analysis

Library preparation and sequencing

Cells were dissociated into single cells differentiation and loaded into the 10X Chromium Controller for library construction using the Single-Cell 3' Library Kit, Version 2 Chemistry (10x Genomics) according to the manufacturer's protocol. Next, indexed cDNA libraries were sequenced on the HiSeq4000 platform. Single-cell expression was quantified using unique molecular identifiers (UMIs) by 10x Genomics' "Cell Ranger" software.
Mean reads per cell for all eight data sets:
CMEC (R1): 28,499; CMEC (R2): 29,388; PMEC (R1): 31,860; PMEC (R2): 38,415; CM_MT (R1): 39,319; CM_MT (R2): 29,741; PM_MT (R1): 36,726; PM_MT (R2): 26,421.

Single-cell RNA-sequencing data pruning and normalization

For data pruning and normalization, the two replicates of each of the 4 conditions (CMEC, PMEC, CM_MT and PM_MT) were combined without batch correction. Cells with a number of genes per cell below a certain threshold (1200 (CMEC), 1200 (PMEC), 900 (CM_MT), 750 (PM_MT), see Figure 4.4) were removed. Genes expressed in less than 2 of the remaining cells with a count of at most 1 were excluded from further analysis. Each combined data set was normalized using the R package scran (V 1.14.6) [43]. Highly variable genes (HVGs) were calculated (using 'improvedCV2' from the scran package) for each replicate of the combined data sets after excluding ribosomal genes, stress markers [44] and mitochondrial genes. For downstream analysis the top 5% HVGs were used after excluding proliferation [45] and cell cycle [46] related genes.

Cell cycle analysis and batch correction

For each combined data set, cell cycle analysis was performed with the scran package using the 'cyclone' function [47] on normalized counts (Figure 4.4H). Cells with a G2/M score higher than 0.2 were considered to be in G2/M phase. Otherwise, they were classified as G1/S. Using this binary classifier as predictor, we regressed out cell cycle effects with the R package limma (V 3.42.2) [48] applied to log-transformed normalized counts. Then, for each combined data set, the two replicates were batch corrected with the fast mutual nearest neighbors correction method (MNN) [49] on the cell cycle corrected counts, using the 30 first principal components and 20 nearest-neighbors (Figure 4.4C).

Clustering

For each combined data set, batch-corrected counts were standardized per gene and then used to create a shared nearest neighbour (SNN) graph with the scran R package (d = 30, k = 2). Louvain clustering was applied to the SNN graph using the igraph python package (V 0.7.1) with these resolution parameters: 0.4 (CMEC), 0.4 (CM_MT), 0.3 (PMEC), 0.1 (PM_MT). For the CMEC data set, this resulted in 5 clusters (Figure 4.4D). Two of these 5 clusters were excluded from further analysis based on the expression of pluripotency markers (Figure 4.4E). For CM_MT and PM_MT, clustering resulted in 4 clusters (Figure 4.4F and 4.4G), where one cluster was excluded from further analysis, because it was mainly present in one of the two replicates. Additionally, the attempt to map this cluster to in vivo data resulted in mostly unassigned cell types (plot not shown). For PMEC, clustering resulted in 3 clusters.

Dimensionality reduction and pseudotime

Dimensionality reduction was performed using the python scanpy pipeline (V 1.4.6). For both data sets (CMEC and PMEC) a 20 nearest-neighbors (knn, k=20) graph was created from diffusion components of the batch corrected data sets. Diffusion components are the eigenvectors of the diffusion operator which is calculated from Euclidean distances and a Gaussian kernel. The aim is to find a lower dimensional embedding which considers the dynamics of differentiation. Both graphs were projected into two dimensions using a force-directed graph layout and starting positions obtained from the partition-based graph abstraction (PAGA) output [50]. PAGA estimates connectivities between partitions and performs an improved version of diffusion pseudotime. Diffusion pseudotime [50, 51] was calculated on these graphs with root cells selected from the "Cardiac Mesoderm" cluster in CMEC, and the "Paraxial Mesoderm" cluster in PMEC.

For CM_MT and PM_MT, the knn graphs (k=50 for PM_MT, k=100 for CM_MT) were created from the first 30 principal components of the batch-corrected data sets. These graphs were projected into two dimensions with a force-directed graph layout and starting positions from the PAGA output.

In vivo data analysis and mapping

The in vivo data set, downloaded from https://www.spatialresearch.org/resources-published-datasets/doi-10-1016-j-cell-2019-11-025/, contains a 6.5 PCW (postcoitum weeks) human fetal cardiac tissue sample. The clusters and cluster annotations were obtained from the original publication [18]. The data set was normalized with the scran R package and HVGs were calculated as described above. Dimensionality reduction was performed with the R package umap (V 0.2.5.0) using 20 nearest-neighbors, min_dist = 0.7 and Euclidean distances.

Differential expression analysis

All differential expression tests were performed with edgeR (V 3.28.1) [41] using a negative binomial regression and raw counts. The predictors in the regression were: cluster and replicate (both discrete variables), as well as the total number of counts per cell.

For marker gene analysis (Figures 4.6A and 4.6C), p-values were obtained for a contrast between the cluster of interest and all other clusters using regression coefficients averaged over the replicates. For tests between different data sets (Figure 4.7C), the corresponding

endothelial cell cluster was extracted from each data set. Then, a contrast between MT and day 6 was calculated by averaging over the predictors of both replicates. For the in vivo test (Figure 4.9B), intra-myocardial EC and endocardium clusters were extracted from the data set to calculate the contrast between them. P-values were adjusted for multiple hypothesis testing with the Benjamini-Hochberg procedure.

Comparison to the in vivo data set

CM_MT and PM_MT data sets were mapped on the in vivo data set using the MNN method (d = 30 principal components, k = 100 nearest neighbors). First, in vitro replicates were mapped to each other, then the in vivo data was mapped on the combined in vitro data, using normalized, log- transformed counts and the 10% top HVGs of the in vivo data set. Dimensionality reduction was performed with the R package umap using 100 nearest-neighbors, min_dist = 0.3 and Euclidean distance.

K-nearest-neighbour (KNN) assignment was performed in the batch corrected, principal component space (30 PCs). The 100 nearest-neighbors in the in vivo data set based on Euclidean distances were calculated for each in vitro cell. The in vitro cell was ascribed the cell type most abundant among the 100 in vivo neighbors. Each such assignment received a confidence score, which is the number of in vivo neighbors with that cell type divided by the number of all nearest neighbors (=100). A cell was not ascribed a cell type if either the average distance to its nearest neighbour exceeded a certain threshold (determined by the long tail of the histogram of average distances: 0.35), or the assignment had a confidence score smaller than 0.5. In addition, clusters containing less than 10 cells were not ascribed a cell type.

For the Jaccard similarity measure, marker genes of each differential expression test were selected with adjusted p-value ≤ 0.05. The remaining genes were ranked by log2 fold-change and the first 478 genes were selected for analysis. Then, the Jaccard distances were calculated between the marker genes of intra-myocardial endothelial cells and each of the other gene sets. For principal component analysis (Figure 4.10H), human fetal bulk samples [32] and in vitro bulk samples were combined with the single cell data sets. For each single-cell data set, the endothelial cells were extracted and the sum per gene over all cells was calculated. Then, bulk and single cell samples were log-transformed and combined into one data set. Principal component analysis was applied on the gene-wise standardized data set, using marker genes of the intra-myocardial endothelial cells from the in vivo data set.

4.4.4 Data availability

The accession numbers for the bulk and single-cell RNA-sequencing datasets reported in this paper are https://www.ncbi.nlm.nih.gov/geo/. Supplementary tables are available at https://doi.org/10.5061/dryad.9p8cz8wkg.

Funding This project received funding from the European Union's Horizon 2020 Framework Programme (668724); European Research Council (ERCAdG 323182 STEMCARDIOVASC); Netherlands Organ-on-Chip Initiative, an NWO Gravitation project funded by the Ministry of Education, Culture and Science of the government of the Netherlands (024.003.001). M. M. and S.S. were supported by the Netherlands Organisation for Scientific

Research (NWO/OCW,www.nwo.nl), as part of the Frontiers of Nanoscience (NanoFront) program. The computational work was carried out on the Dutch national e-infrastructure with the support of SURF Cooperative.

Disclosure of potential conflict of interest The authors indicated no potential conflicts of interest.

References

[1] J. Kalucka et al. Single-Cell Transcriptome Atlas of Murine Endothelial Cells. *Cell*, 180(4):764–779.e20, feb 2020.

[2] A. Jambusaria et al. Endothelial heterogeneity across distinct vascular beds during homeostasis and inflammation. *eLife*, 9, jan 2020.

[3] R. Marcu et al. Human Organ-Specific Endothelial Cell Heterogeneity. *iScience*, 4:20, jun 2018.

[4] D. J. Nolan et al. Molecular Signatures of Tissue-Specific Microvascular Endothelial Cell Heterogeneity in Organ Maintenance and Regeneration. *Developmental Cell*, 26(2):204–219, jul 2013.

[5] M. Potente and T. Mäkinen. Vascular heterogeneity and specialization in development and disease. *Nature Reviews Molecular Cell Biology 2017 18:8*, 18(8):477–494, may 2017.

[6] D. T. Paik et al. Single-Cell RNA Sequencing Unveils Unique Transcriptomic Signatures of Organ-Specific Endothelial Cells. *Circulation*, 142(19):1848–1862, nov 2020.

[7] R. J. Esper et al. Endothelial dysfunction: a comprehensive appraisal. *Cardiovascular Diabetology*, 5:4, feb 2006.

[8] A. R. Pinto et al. Revisiting cardiac cellular composition. *Circulation Research*, 118(3):400–409, 2016.

[9] H. Zhang et al. Endocardium Minimally Contributes to Coronary Endothelium in the Embryonic Ventricular Free Walls. *Circulation Research*, 118(12):1880–1893, jun 2016.

[10] B. Wu et al. Endocardial Cells Form the Coronary Arteries by Angiogenesis through Myocardial-Endocardial VEGF Signaling. *Cell*, 151(5):1083–1096, nov 2012.

[11] S. Somekawa et al. Tmem100, an ALK1 receptor signaling-dependent gene essential for arterial endothelium differentiation and vascular morphogenesis. *Proceedings of the National Academy of Sciences of the United States of America*, 109(30):12064–12069, jul 2012.

[12] M. C. Puri, J. Partanen, J. Rossant, and A. Bernstein. Interaction of the TEK and TIE receptor tyrosine kinases during cardiovascular development. *Development*, 126(20):4569–4580, oct 1999.

[13] H. I. Chen et al. The sinus venosus contributes to coronary vasculature through VEGFC-stimulated angiogenesis. *Development (Cambridge)*, 141(23):4500–4512, dec 2014.

[14] X. Tian et al. Subepicardial endothelial cells invade the embryonic ventricle wall to form coronary arteries. *Cell Research 2013 23:9*, 23(9):1075–1090, jun 2013.

[15] K. Red-Horse, H. Ueno, I. L. Weissman, and M. A. Krasnow. Coronary arteries form by developmental reprogramming of venous cells. *Nature 2010 464:7288*, 464(7288):549–553, mar 2010.

[16] H. Suryawanshi et al. Cell atlas of the foetal human heart and implications for autoimmune-mediated congenital heart block. *Cardiovascular Research*, 116(8):1446–1457, jul 2020.

[17] Y. Cui et al. Single-Cell Transcriptome Analysis Maps the Developmental Track of the Human Heart. *Cell Reports*, 26(7):1934–1950.e5, feb 2019.

[18] M. Asp et al. A Spatiotemporal Organ-Wide Gene Expression and Cell Atlas of the Developing Human Heart. *Cell*, 179(7):1647–1660.e19, dec 2019.

[19] R. E. Poelmann et al. Development of the cardiac coronary vascular endothelium, studied with antiendothelial antibodies, in chicken-quail chimeras. *Circulation Research*, 73(3):559–568, 1993.

[20] T. C. Katz et al. Distinct Compartments of the Proepicardial Organ Give Rise to Coronary Vascular Endothelial Cells. *Developmental Cell*, 22(3):639–650, mar 2012.

[21] X. Tian et al. Vessel formation. De novo formation of a distinct coronary vascular population in neonatal heart. *Science (New York, N.Y.)*, 345(6192):90–94, jul 2014.

[22] F. J. Giordano et al. A cardiac myocyte vascular endothelial growth factor paracrine pathway is required to maintain cardiac function. *Proceedings of the National Academy of Sciences of the United States of America*, 98(10):5780–5785, may 2001.

[23] N. L. Ward et al. Angiopoietin 1 expression levels in the myocardium direct coronary vessel development. *Developmental Dynamics*, 229(3):500–509, mar 2004.

[24] D. Tirziu, F. J. Giordano, and M. Simons. Cell Communications in the Heart. *Circulation*, 122(9):928–937, aug 2010.

[25] F. Perbellini, S. A. Watson, I. Bardi, and C. M. Terracciano. Heterocellularity and Cellular Cross-Talk in the Cardiovascular System. *Frontiers in Cardiovascular Medicine*, 5:143, nov 2018.

[26] V. V. Orlova et al. Functionality of endothelial cells and pericytes from human pluripotent stem cells demonstrated in cultured vascular plexus and zebrafish xenografts. *Arteriosclerosis, Thrombosis, and Vascular Biology*, 34(1):177–186, jan 2014.

[27] V. V. Orlova et al. Generation, expansion and functional analysis of endothelial cells and pericytes derived from human pluripotent stem cells. *Nature Protocols 2014 9:6*, 9(6):1514–1531, may 2014.

[28] K. M. Loh et al. Mapping the pairwise choices leading from pluripotency to human bone, heart and other mesoderm cell-types. *Cell*, 166(2):451, jul 2016.

[29] H. Minami et al. Generation of Brain Microvascular Endothelial-Like Cells from Human Induced Pluripotent Stem Cells by Co-Culture with C6 Glioma Cells. *PLOS ONE*, 10(6):e0128890, jun 2015.

[30] E. S. Lippmann et al. Derivation of blood-brain barrier endothelial cells from human pluripotent stem cells. *Nature Biotechnology 2012 30:8*, 30(8):783–791, jun 2012.

[31] E. Giacomelli et al. Three-dimensional cardiac microtissues composed of cardiomyocytes and endothelial cells co-differentiated from human pluripotent stem cells. *Development (Cambridge)*, 144(6):1008–1017, mar 2017.

[32] E. Giacomelli et al. Human-iPSC-Derived Cardiac Stromal Cells Enhance Maturation in 3D Cardiac Microtissues and Reveal Non-cardiomyocyte Contributions to Heart Disease. *Cell Stem Cell*, 26(6), 2020.

[33] J. G. Camp et al. Multilineage communication regulates human liver bud development from pluripotency. *Nature 2017 546:7659*, 546(7659):533–538, jun 2017.

[34] M. Buckingham, S. Meilhac, and S. Zaffran. Building the mammalian heart from two sources of myocardial cells. *Nature Reviews Genetics 2005 6:11*, 6(11):826–835, nov 2005.

[35] E. S. Ng, R. Davis, E. G. Stanley, and A. G. Elefanty. A protocol describing the use of a recombinant protein-based, animal product-free medium (APEL) for human embryonic stem cell differentiation as spin embryoid bodies. *Nature Protocols 2008 3:5*, 3(5):768–776, apr 2008.

[36] M. Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, 17(1):10–12, may 2011.

[37] T. D. Wu and C. K. Watanabe. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*, 21(9):1859–1875, may 2005.

[38] T. D. Wu and S. Nacu. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics*, 26(7):873–881, apr 2010.

[39] A. Yates et al. Ensembl 2016. *Nucleic Acids Research*, 44(D1):D710–D716, jan 2016.

[40] K. D. Hansen, R. A. Irizarry, and Z. Wu. Removing technical variability in RNA-seq data using conditional quantile normalization. *Biostatistics*, 13(2):204–216, apr 2012.

[41] M. D. Robinson, D. McCarthy, Y. Chen, and G. K. Smyth. edgeR: differential expression analysis of digital gene expression data User's Guide. *Bioinformatics*, 26(October 2018):1–75, 2013.

[42] G. Yu, L. G. Wang, Y. Han, and Q. Y. He. clusterProfiler: an R package for comparing biological themes among gene clusters. *Omics : a journal of integrative biology*, 16(5):284–287, may 2012.

[43] A. T. Lun, K. Bach, and J. C. Marioni. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biology*, 17(1):1–14, apr 2016.

[44] S. C. van den Brink et al. Single-cell sequencing reveals dissociation-induced gene expression in tissue subpopulations. *Nature Methods*, 14(10):935–936, 2017.

[45] M. L. Whitfield, L. K. George, G. D. Grant, and C. M. Perou. Common markers of proliferation. *Nature Reviews Cancer 2006 6:2*, 6(2):99–106, feb 2006.

[46] B. Giotti, A. Joshi, and T. C. Freeman. Meta-analysis reveals conserved cell cycle transcriptional network across multiple human cell types. *BMC Genomics*, 18(1):1–12, jan 2017.

[47] A. Scialdone et al. Computational assignment of cell-cycle stage from single-cell transcriptome data. *Methods*, 85:54–61, sep 2015.

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

[49] L. Haghverdi, A. T. Lun, M. D. Morgan, and J. C. Marioni. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nature Biotechnology*, 36(5):421–427, 2018.

[50] F. A. Wolf et al. PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biology*, 20(1):1–9, mar 2019.

[51] L. Haghverdi et al. Diffusion pseudotime robustly reconstructs lineage branching. *Nature Methods 2016 13:10*, 13(10):845–848, aug 2016.

5

A gastruloid model of the interaction between embryonic and extra-embryonic cell types

Stem-cell derived in vitro systems, such as organoids or embryoids, hold great potential for modeling in vivo development. Full control over their initial composition, scalability and easily measurable dynamics make those systems useful for studying specific developmental processes in isolation. Here we report the formation of gastruloids consisting of mouse embryonic stem cells (mESCs) and extraembryonic endoderm (XEN) cells. These XEN-enhanced gastruloids (XEGs) exhibit the formation of neural epithelia, which are absent in gastruloids derived from mESCs only. By single-cell RNA-seq, imaging and differentiation experiments, we demonstrate the neural characteristics of the epithelial tissue. We further show that the mESCs induce the differentiation of the XEN cells to a visceral endoderm-like state. Finally, we demonstrate that local inhibition of WNT signaling and production of a basement membrane by the XEN cells underlie the formation of the neuroepithelial tissue. In summary, we establish XEGs to explore heterotypic cellular interactions and their developmental consequences in vitro.

This chapter is based on Bérenger-Currias NM, **Mircea M**, Adegeest E, et al. A gastruloid model of the interaction between embryonic and extra-embryonic cell types. Journal of Tissue Engineering. (2022). doi:10.1177/20417314221103042. [1]

5.1 Introduction

Multicellular in vitro systems have become a major focus of biology and bioengineering over the last few years. Stem cell-derived systems, such as embryoids and organoids show complex organization and have the potential to serve as models for in vivo development [2–5]. Among the most prominent examples of such model systems are gastruloids. These aggregates of mouse or human embryonic stem cells (ESCs) recapitulate elements of embryonic development, such as body axis formation and extension [6–10]. Notably, gastruloids do not contain extraembryonic cells, which provide numerous signaling inputs during gastrulation in vivo [11]. The remarkable self-organizing capabilities of ESCs thus raise questions about the precise role of extraembryonic tissues in gastrulation. Here, we will focus on the extraembryonic endoderm, which derives from the primitive endoderm (PrE) in vivo. At the blastocyst stage, prior to implantation of the embryo in the uterine wall, the PrE overlays the developing epiblast, which gives rises to all embryonic tissues (see Fig. 5.1a for a schematic of early mouse development.) Subsequently, the PrE differentiates into the Parietal Endoderm (PE), which covers the inside of the blastocoel cavity [12] and the Visceral Endoderm (VE), which surrounds the embryo until the formation of the visceral yolk sac and integration of some VE cells in the embryonic gut [13, 14]. Another subpopulation of the VE, the Anterior Visceral Endoderm (AVE) is involved in the establishment of the embryo's body axes [15, 16]. In this study, we set out to develop an in vitro model system for the interaction between the extraembryonic endoderm and the gastrulating embryo. As a proxy for the extraembryonic endoderm in vivo, we used XEN cells, which can be derived from the PrE cells in blastocysts. XEN cells have been previously incorporated in embryoid systems [17–21] that model the earliest stages of development. Here, we wanted to explore, whether the role of the extraembryonic endoderm in gastrulation can be modeled by adding XEN cells to the gastruloid model system. Below, we report that aggregates of mESCs and XEN cells can produce columnar neural epithelia. Using multiple markers, perturbation of the signaling pathways that play a role in neural development in vivo, and further differentiation to neural organoids, we confirmed that the epithelial structures indeed have neural characteristics. By single-cell RNA-seq, we identified differences in composition and molecular profiles between our new model system and regular gastruloids. We then established that a majority of XEN-derived cells become visceral endoderm-like due to co-differentiation with the mESCs. Finally, we showed that XEN cells promote epithelia formation by local, DKK1 mediated, WNT inhibition, as well as through production of a basement membrane. Our study thus highlights the complex interplay between embryonic and extraembryonic cells and explores possible mechanisms underlying their interaction.

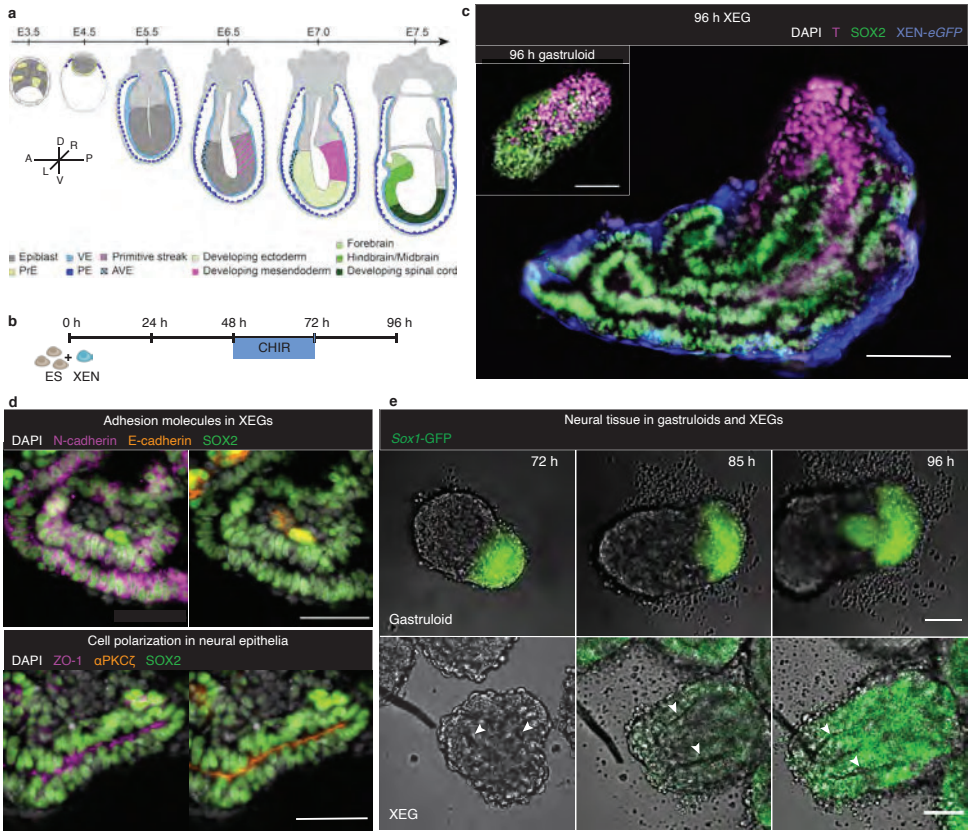


Figure 5.1: XEN cells induce neuroepithelial structures in XEN enhanced gastruloids. a, Schematic of early mouse embryonic development. Tissues discussed in this manuscript are indicated with color. A: anterior, P: posterior, D: dorsal, V: ventral, L: left, R: right. b, Schematic of the culture protocol: at 0 h, 200 cells (150 ESCs and 50 XEN cells) were aggregated; CHIR99021 was added between 48 h and 72 h after cell seeding to activate the WNT pathway; cell aggregates were cultured until 96 h. c, T and SOX2 expression at 96 h in XEGs Inset: Aggregate resulting from the standard gastruloid protocol (without XEN cells) at 96 h. Z-projections of wholemount immunostaining. Scale bars: 100 µm. d, Expression of SOX2, E-cadherin, N-cadherin, ZO-1 and $\alpha PKC\zeta$ in XEGs at 96 h (immunostaining of cryosections). Scale bar: 50 μm. e, Live-cell imaging of SOX1 expression in a gastruloid (top panels, scale bar: 20 μm) and a XEG (bottom panels, scale bar: 50 μm), grown with Sox1-GFP mESCs (see Supplementary Videos 3-8). In all images, a single z-plane is shown. The arrows indicate epithelial structures. c-d, Cell nuclei were stained with DAPI.

5.2 Results

5.2.1 XEN cells induce neuroepithelial structures in XEN enhanced gastruloids

We first implemented the original mouse gastruloid protocol4, in which mESCs are aggregated in N2B27 media and exposed to a 24 h pulse of CHIR99021 (CHIR), which activates the WNT pathway. After 96 h, this protocol results in elongated gastruloids. As reported before [6–8], 96 h gastruloids contained localized compartments, marked by Brachyury

(T) or SOX2, respectively (Fig. 5.1c, inset). These compartments are believed to resemble early in vivo mesendodermal (T) or neural progenitor (SOX2) cell types. Starting from the gastruloid protocol, we developed a new system by aggregating mESCs and XEN cells, keeping all other experimental conditions the same (Fig. 5.1b). We call our mixed aggregates "XEN Enhanced Gastruloids" (XEGs). Like gastruloids, 96 h XEGs showed an elongated morphology and localized T-positive and SOX2-positive compartments. However, unlike in gastruloids, SOX2-positive cells in XEGs were organized in columnar epithelia surrounding one or several lumina (Fig 5.1c).

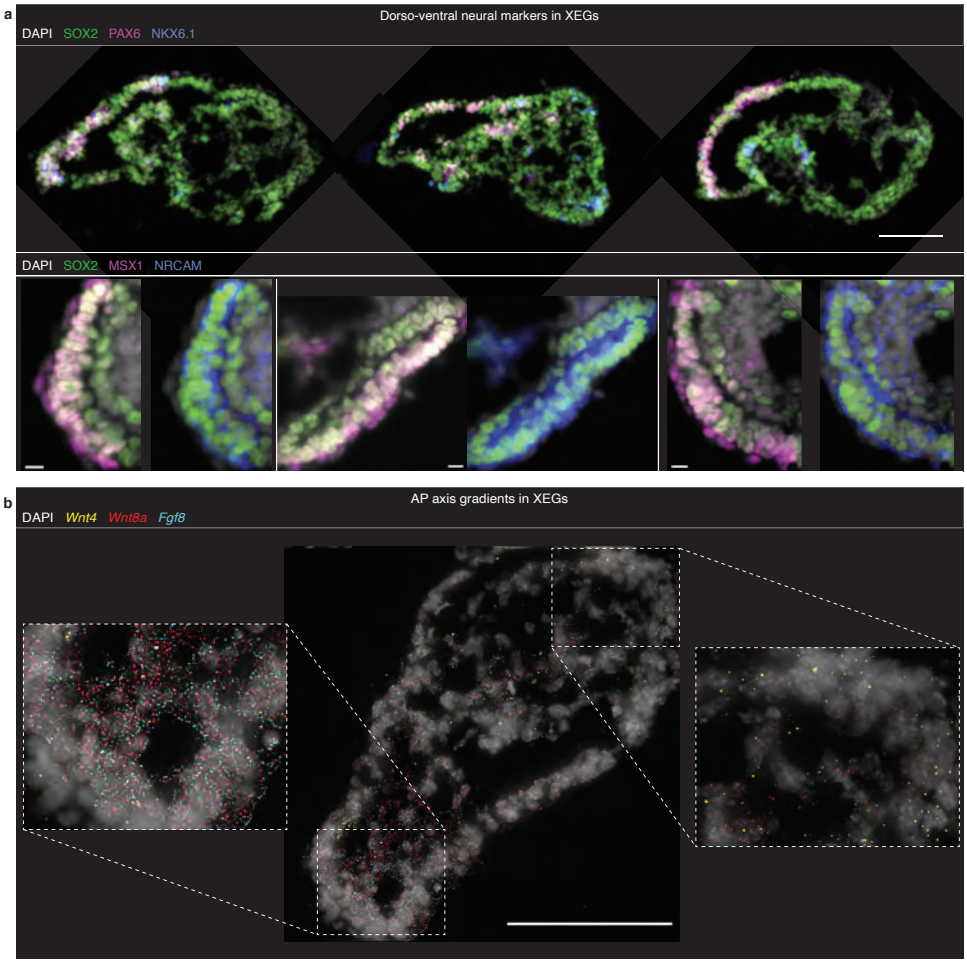


Figure 5.2: Neural epithelia in XEGs are heterogeneous and contain subpopulations with dorsal or ventral characteristics. a, Expression of dorsal (PAX6, MSX1) and ventral (NKX6.1) neural tube markers and a neural cell adhesion molecule (NRCAM) in 96 h XEGs (immunostaining of sections). Top, expression of PAX6 and NKX6.1. Scale bar: 100 μm. Bottom, zoomed pictures of neural epithelia showing the expression of MSX1 and NRCAM. Scale bars: 20 μm. b, Wnt4, Wnt8a and Fgf8 expression in XEGs at 96 h, visualized by smFISH on sections. Each diffraction limited dot is a single mRNA molecule. Scale bar: 100 μm. a-b, Cell nuclei were stained with DAPI.

Expression of the broadly expressed neural marker SOX2 and the striking morphology suggested that the observed structures resemble neural epithelia. The lack of pluripotency marker expression (Supplementary Fig. 5.1a) excluded that the structures were formed by remaining undifferentiated cells. The presence of N-cadherin and absence of E-cadherin in the epithelia (Fig. 5.1d, top) is consistent with the known switch from E- to N-cadherin during neural differentiation in vivo [22] and in vitro [23]. We could also observe that the epithelial cells were polarized and expressed apical markers ZO-1 and aPKC (Fig. 5.1d, bottom), consistent with neural epithelia in vivo [24]. Finally, we detected the neural progenitor markers PAX6 and NKX6.1 [25] in a subpopulation of epithelial cells (Supplementary Fig. 5.1b). Combined, these results suggest that the observed structures in XEGs have the characteristics of neural epithelia.

To understand how these structures formed, we used time-lapse microscopy of developing XEGs. Around 48 h after seeding, cells formed rosette-like shapes (Fig. 5.3c, Supplementary Video 1), which resembled structures found in Matrigel-embedded mESCs [26, 27] and indicated a mesenchymal-epithelial transition. Subsequently, a columnar epithelium was formed. Then, lumina opened at different places and merged between 48 h and 72 h (Fig. 5.3d, Supplementary Video 2). During the final 24 h, the epithelium kept extending and differentiated further, as revealed, by the expression of the neural progenitor marker SOX1 [28] (Fig. 5.1e, bottom; Supplementary Videos 3-5). A SOX1 positive cell population also appeared in gastruloids within the same time frame, but, importantly, remained unorganized (Fig. 5.1e, top; Supplementary Videos 6-8).

To explore the robustness of the protocol and identify optimal conditions for the formation of epithelial structures, we tested different ratios of mESCs and XEN cells (Fig. 5.3e, f). Interestingly, even the smallest proportion of XEN cells tested (1:5), was able to induce some epithelia formation. On the other hand, elongation and symmetry breaking were inhibited when the proportion of XEN cells exceeded 1:2. A ratio of 1:3 gave optimal results, with the concurrence of SOX2-positive epithelia and T-positive cells in nearly all aggregates.

5.2.2 Neuroepithelial cells in XEGs are heterogeneous and show further specification

To establish whether the neuroepithelial cells are a homogeneous population of progenitors or have undergone further specification, we carried out additional immunostaining. In subpopulations of cells, we observed the expression of PAX6, MSX1 and ASCL1 (Fig. 5.2a, Fig. 5.3b), which can be found in dorsal progenitors in the developing neural tube [29]. Notably, these markers were localized close to the XEN-derived cells at the outside of the XEGs. By contrast, the ventral marker NKX6.1 was found only sporadically and did not show any preferential spatial localization (Fig. 5.2a). Finally, the neural adhesion molecule NRCAM was ubiquitously expressed in epithelial cells, further supporting their neural character (Fig. 5.2a).

We also attempted to establish a possible specification related to the anteroposterior axis in vivo. Using single-molecule FISH, we observed the expression of Wnt4, Wnt8a and Fgf8 as gradients along the long axis of XEGs (Fig. 5.2b), which resembled similar anteroposterior gradients found in vivo [30]. However, important canonical markers of the most anterior part of the embryo (OTX2, LEFTY1, EN1, ZIC1) could not be detected in XEGs (data not

shown). Taken together, our measurements indicated that neuroepithelial cells in XEGs are heterogeneous and contain subpopulations that might correspond to either dorsal or ventral neural progenitors.

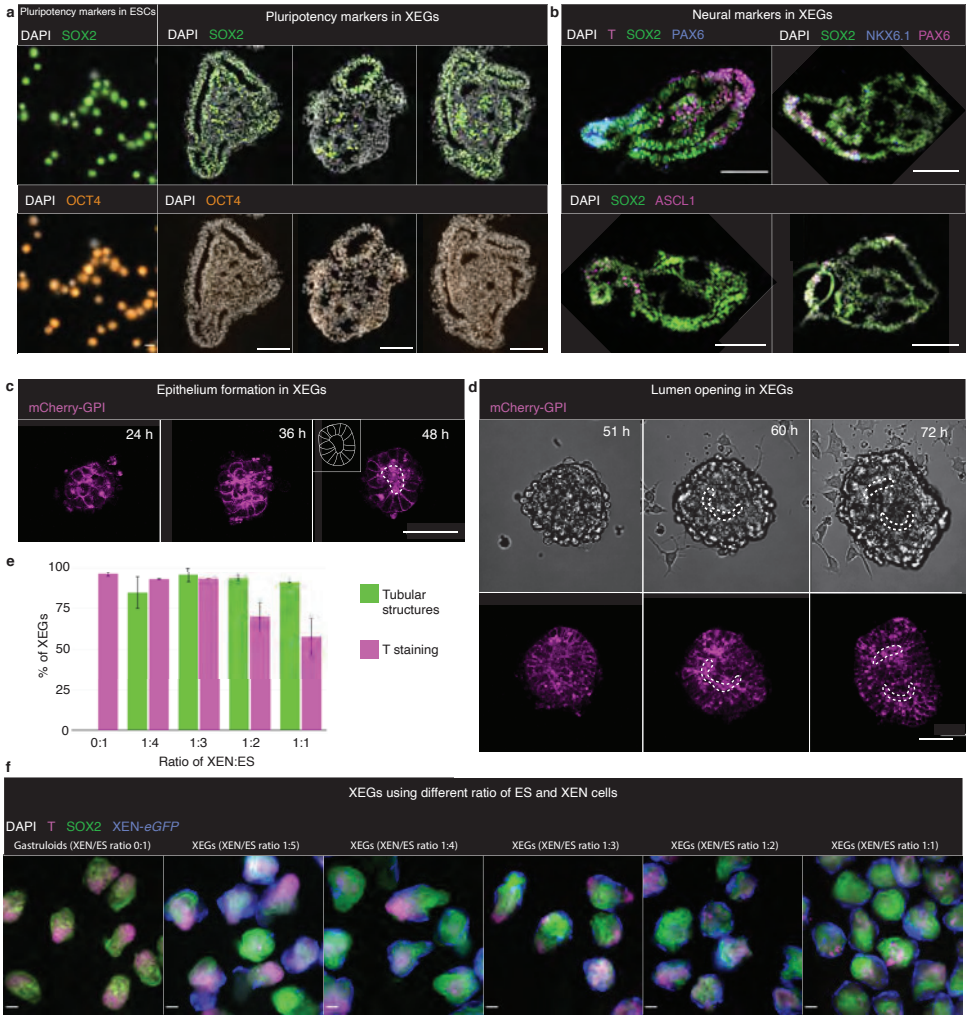


Figure 5.3: Optimization and characterization of XEGs. a, Expression of SOX2 and OCT4 (immunostaining) in cultured ESCs (left, scale bar: 10 μm) and sections of XEGs at 96 h (right, scale bars: 100 μm). b, Expression of T, SOX2, PAX6, ASCL1 and NKX6.1 (immunostaining) in sections 96 h XEGs. Scale bars: 100 μm. c-d, Live-cell imaging of morphological changes in XEGs grown from mCherry-GPI expressing mESCs. mCherry-GPI is localized to the cell membrane. In all images, a single z-plane is shown. Scale bars: 50 μm. c, Rosette formation. The center of the rosette is indicated by a dashed line. Inset: tracing of cell outlines. See also Supplementary Video 1. d, Cavitation of rosettes. The top row shows the brightfield channel, the bottom row shows the mCherry channel. Dashed lines indicate the opening lumen. See also Supplementary Video 2. e, Average fraction of aggregates showing epithelial structures and T staining at 96 h for different starting ratios of ESCs and XEN cells (n = 2 experiments, error bars show standard deviation). f, SOX2 and T expression in gastruloids and XEGs with different starting ratios of ESCs and XEN cells (z-projection of whole mount immunostaining). Scale bars: 100 μm. a, b, f, Cell nuclei were stained with DAPI.

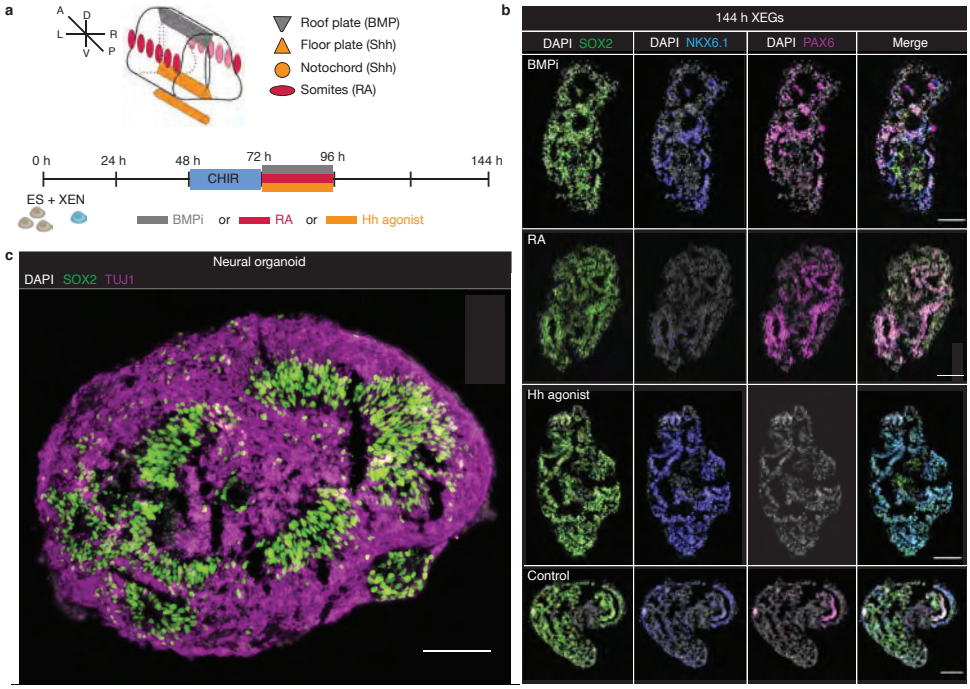


Figure 5.4: Signaling perturbation experiments and continued differentiation confirm neural character. a, Top: schematic of signaling sources patterning the developing neural tube in vivo. A: anterior, P: posterior, D: dorsal, V: ventral, L: left, R: right. Bottom: time line of the signaling experiments. XEGs were treated from 72 h to 96 h, with either BMP pathway inhibitor (BMPi), retinoic acid (RA) or hedgehog pathway agonist (Hh agonist). The XEGs were then allowed to grow for an additional 48 h before staining. b, Expression of SOX2, NKX6.1 and PAX6 in XEGs at 144 h, treated with the indicated factors (immunostaining of sections). n = 3 experiments. Scale bars: 100 μm. c, Expression of SOX2 and TUJ1 in XEGs, 8 days after cell seeding, differentiated according to a cerebral organoid protocol for 4 days (immunostaining of sections). Scale bar: 100 μm. b-c, Cell nuclei are stained with DAPI.

5.2.3 Signaling perturbation experiments and further differentiation support the neuroepithelial character

To further characterize the neuroepithelial structures, we tested how they respond to signaling inputs found in vivo. Specifically, we explored the response to BMP pathway inhibition, as well as Sonic Hedgehog (Shh) and retinoic acid (RA) pathway activation (Fig. 5.4a, b). BMP signaling is known to prevent premature neural specification [31] and to be involved in dorsal patterning of the neural tube [32]. In XEGs, BMP inhibition resulted in an increased number of cells expressing the neural progenitor markers SOX2 and PAX6, as well as NKX6.1, which is expressed in ventral progenitors in the developing neural tube. Sonic hedgehog, produced in vivo by the notochord and the floor plate (see schematic in Fig. 5.4a), is known to be necessary for the patterning of the ventral part of the neural tube [33]. The activation of the Hedgehog signaling pathway led to a higher frequency of cells expressing a ventral marker (NKX6.1) in XEGs. RA, involved in anteroposterior and dorsoventral patterning [34], strongly increased the number of cells expressing PAX6,

which is expressed in dorsal progenitors in the neural tube in vivo [35]. The neuroepithelial structures in XEGs thus responded to signaling inputs as expected from in vivo development.

To test the developmental potential of the neural progenitors further, we sought to differentiate them to more advanced states. Within four days of additional culture in cerebral organoid differentiation media [36], XEGs developed a layered organization of neural progenitors (SOX2+/PAX6+) and neurons (TUJ1+/CTIP2+/PAX2+), surrounding cavities, reminiscent of the organization of the developing dorsal spinal cord [29, 37] (Fig. 5.4c, Fig. 5.5a-c). Interestingly, we also observed a population of cells expressing the endothelial marker CD31 (Fig. 5.5d). This might indicate that non-neural cells remained and might have differentiated further. Those CD31+ cells could specifically represent an early stage of vasculature. Taken together, the signaling perturbation and differentiation experiments confirmed the neural potential of the epithelia.

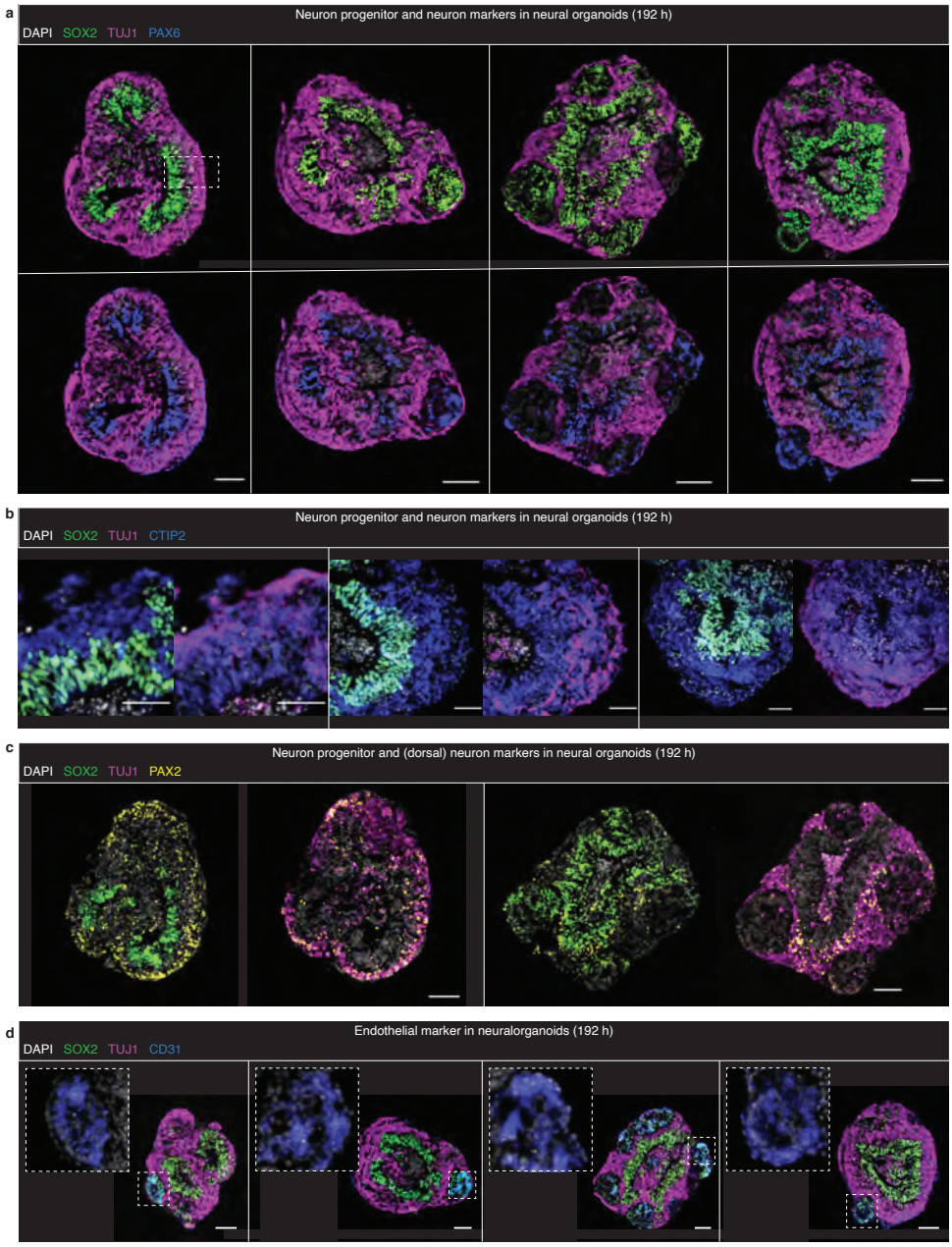


Figure 5.5: Developing spinal cord markers are expressed in XEGs further differentiated with a cerebral organoid protocol. a-d, Immunostaining of sections of XEGs on day 8 after cell seeding. a, TUJ1, a neuron marker, and SOX2 (top) and PAX6 (bottom), neural progenitor markers. The dashed box highlights an example of layered organization reminiscent of the ventricular and mantle zones in the developing spinal cord. b, TUJ1, SOX2 and CTIP2, a neuron marker. c, TUJ1, SOX2 and PAX2, a marker of dorsal neurons in the developing spinal cord. d, TUJ1, SOX2 and CD31, an endothelial marker. Insets show clusters of cells positive for CD31. a-d, Cell nuclei were stained with DAPI. Scale bars: 100 μm.

5.2.4 Single-cell RNA-seq reveals the transcriptional profiles of XEG cells

Having focused on the most striking, morphological difference between gastruloids and XEGs, we wanted to take a more comprehensive approach to reveal additional differences between the two model systems. To that end, we used single-cell RNA-sequencing (scRNA-seq) (Fig. 5.6a-e). By mapping the data to single-cell transcriptomes of mouse embryos from E6.5 to E8.5 [38] (Fig. 5.7a,b) we classified the transcriptional identity of the cells (Fig. 5.8a,b). Except for the least abundant cell types, the distribution of cell types was consistent across two biological replicates (Fig. 5.8c). Expression of known markers confirmed the classification by mapping to in vivo data (Fig. 5.7c, Supplementary Table 1). Most cell types belonged to the E8.0 or E8.5 embryo (Fig. 5.7d), which might indicate that in vitro differentiation proceeded roughly with the same speed as in vivo development.
Neuromesodermal progenitors (NMPs) and spinal cord-like cells were the most abundant in both model systems (Fig. 5.8c). Gastruloids thus already contain cells of the neural lineage, which, however, seem to lack organization (Fig. 5.1c, inset). To identify the cells forming epithelial structures in XEGs, we used the neural markers Sox2, Pax6 and Nkx6.1 [25], which we had detected by immunostaining (Fig. 5.3b). We found these markers to be co-expressed in cells classified as "spinal cord" and "brain" in the scRNA-seq data (Fig. 5.8d, Fig. 5.7e), confirming their neural ectoderm identity. While NMPs also expressed Sox2 and Nkx6.1, neuroepithelial structures were clearly distinguishable by the presence of Pax6 and the absence of T.
We next asked, whether there are any subpopulations in the neural ectoderm-like cells, as hinted at by our immunostaining results (Fig. 5.2). Differential gene expression analysis between spinal cord-like cells and other cells in XEGs identified markers of both the dorsal and ventral neural tube (Fig. 5.9a, Supplementary Table 2). A comparison between XEGs and gastruloids revealed that neural ectoderm-like cells (including spinal cord and clusters identified as neuroectoderm or brain) expressed more dorsal markers in XEGs (Fig. 5.9b, Supplementary Table 3). This dorsal identity was confirmed by mapping the neural ectoderm-like cells to single-cell expression profiles of in vivo neural tube [37] (Fig. 5.9c). The majority of neural ectoderm-like cells from XEGs turned out to be more similar to dorsal progenitors in vivo. To reveal subpopulations, we clustered the neural ectoderm-like cell using a curated list of genes that are dorsoventral axis markers in the developing neural tube [29, 37, 39] (Fig. 5.8e). This analysis resulted in 5 clusters with distinct dorsoventral characteristics (Fig. 5.8f). Two clusters (1 and 2) had a more ventral identity, two clusters (4 and 5) had dorsal transcriptional characteristics and cluster 3 expressed both dorsal and ventral markers at a low level. Roughly one third of the cells had a dorsal identity (Fig. 5.8g), which is qualitatively consistent with our immunostaining results (Fig. 5.2). Overall, both model systems showed a diverse cell type distribution, also comprising a variety of mesodermal cell types. Thus, XEG cells are not globally biased towards the neural fate, as occurring in other protocols for induction of neural epithelia [27, 40, 41]. On the contrary, XEGs even contained a bigger proportion of mesoderm-like cells, compared to gastruloids (Fig. 5.8c, Fig. 5.10a). While paraxial, intermediate and somitic mesoderm-like clusters were present in both model systems, only XEGs contained cells transcriptionally resembling primitive streak, nascent mesoderm, pharyngeal mesoderm and hematoendothelial progenitors. To confirm the presence of mesodermal cell types

in XEGs, we focused on two genes, Tbx6 and Pax2, which are markers of nascent and intermediate mesoderm, respectively. Our single-cell RNA-seq data showed expression of both genes in subpopulations of XEG cells (Fig. 5.10b) and immunostaining confirmed their presence (Fig. 5.10c). However, we did not observe any tissue-level organization of those cells in XEGs. Taken together, these results suggest that the XEN-derived cells in XEGs also have an effect on the mesodermal cell population.

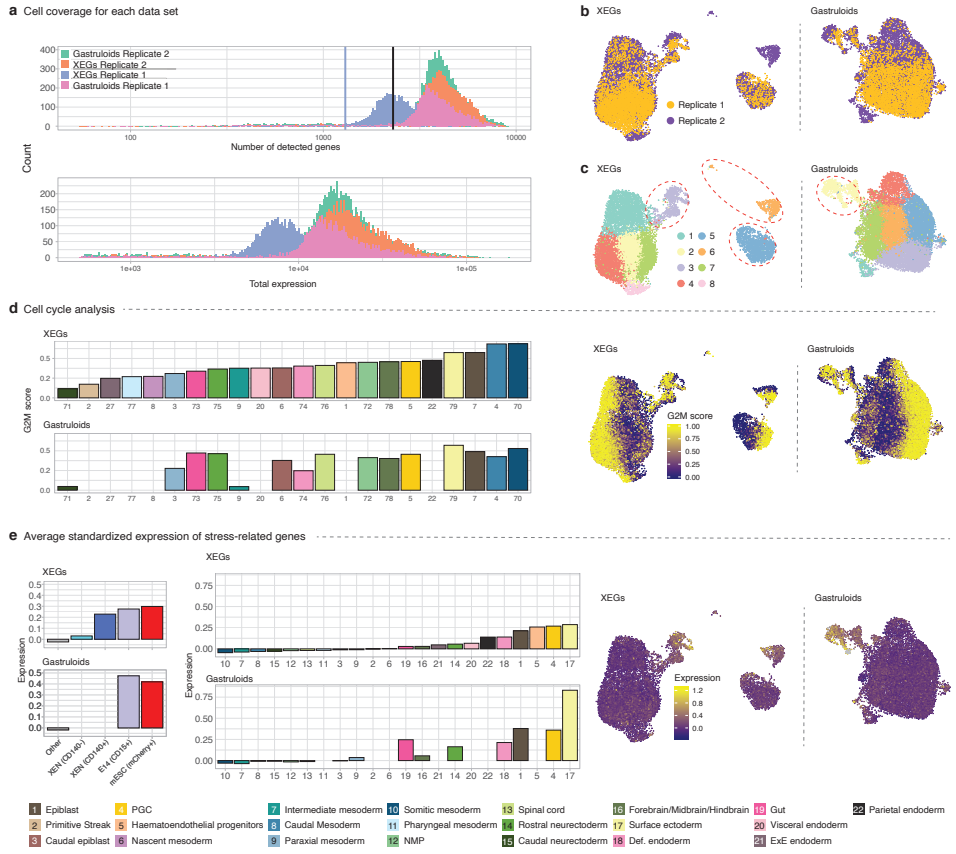


Figure 5.6: Quality control of single-cell RNA-seq data. a, Top, number of detected genes per cell in each replicate; the blue line indicates a quality control threshold for XEGs from replicate 1 and the black line for the remaining datasets. Bottom, total expression per cell for each dataset. b, UMAP of cells in XEGs and gastruloids, colored by replicate. c, UMAP of cells in XEGs and gastruloids, colored by Louvain clustering. The encircled clusters contain the spiked-in cells. d, Left, average G2M scores for each cell type. Right, UMAPs of cells in XEGs and gastruloids colored by G2M score. e, Left, average standardized expression of stress-related genes in spike-in cells. Middle, expression of stress-related genes by cell type. Right, UMAPs of cells in XEGs and gastruloids with expression of stress-related genes indicated by color. b-e, UMAPs contain both replicates, batch corrected.

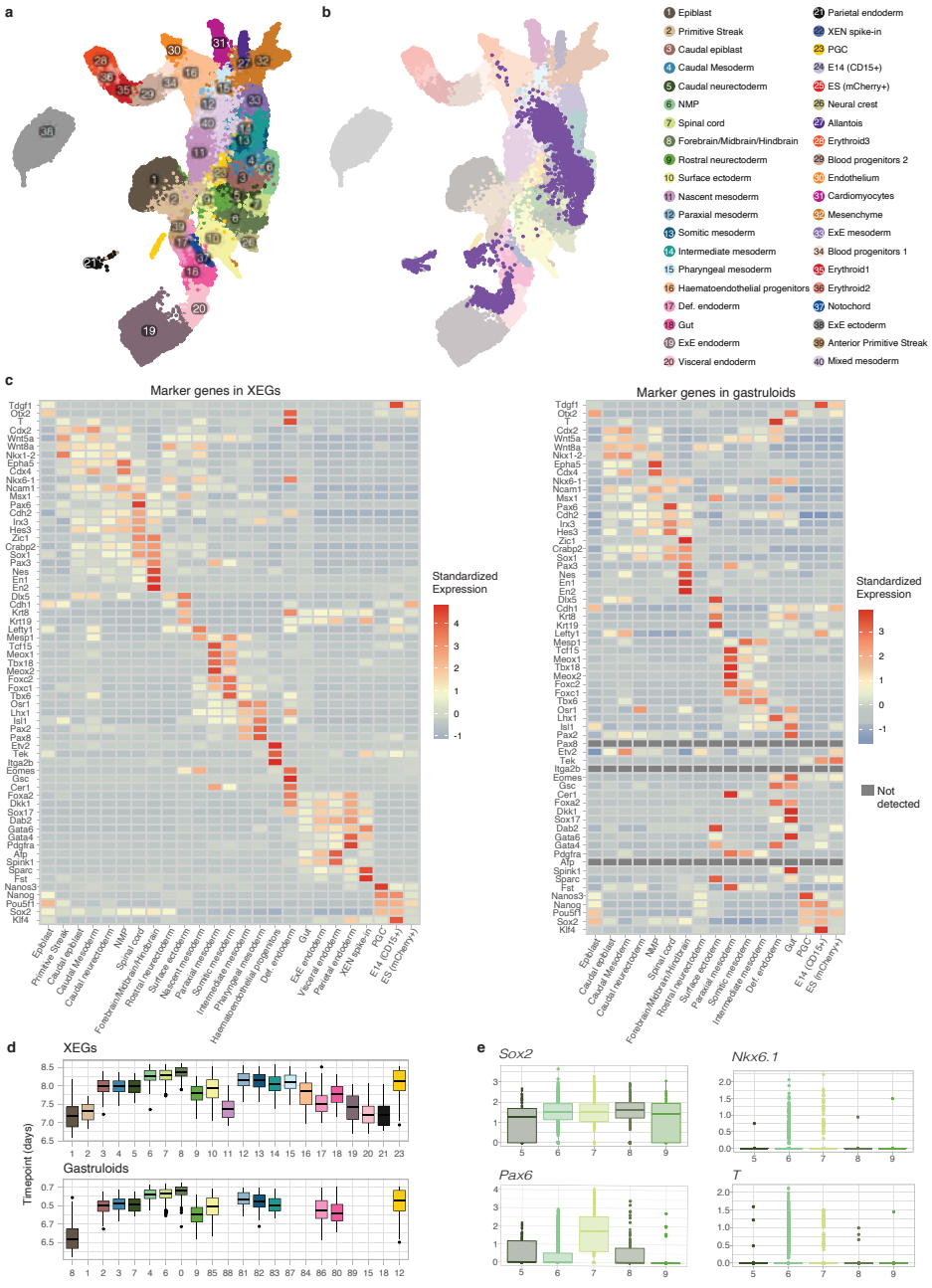


Figure 5.7: Single-cell RNA-seq resolves cell type diversity in XEGs and gastruloids. a, UMAP of the Pijuan-Sala et al. [38] dataset with cell types indicated by color. b, MNN mapping of XEG cells from replicate 2 (opaque color) to the Pijuan-Sala et al. dataset (pale colors), as an example for the mapping procedure. c, Heat map of standardized expression of genes associated with mouse embryonic development in XEGs and gastruloids. References describing the in vivo expression of the genes are given in Supplementary Table 1. d, Developmental age of cell types based on mapping to in vivo data. e, Expression of Sox2, Pax6 and Nkx6.1 and T in XEGs, as measured by single-cell RNA-seq.

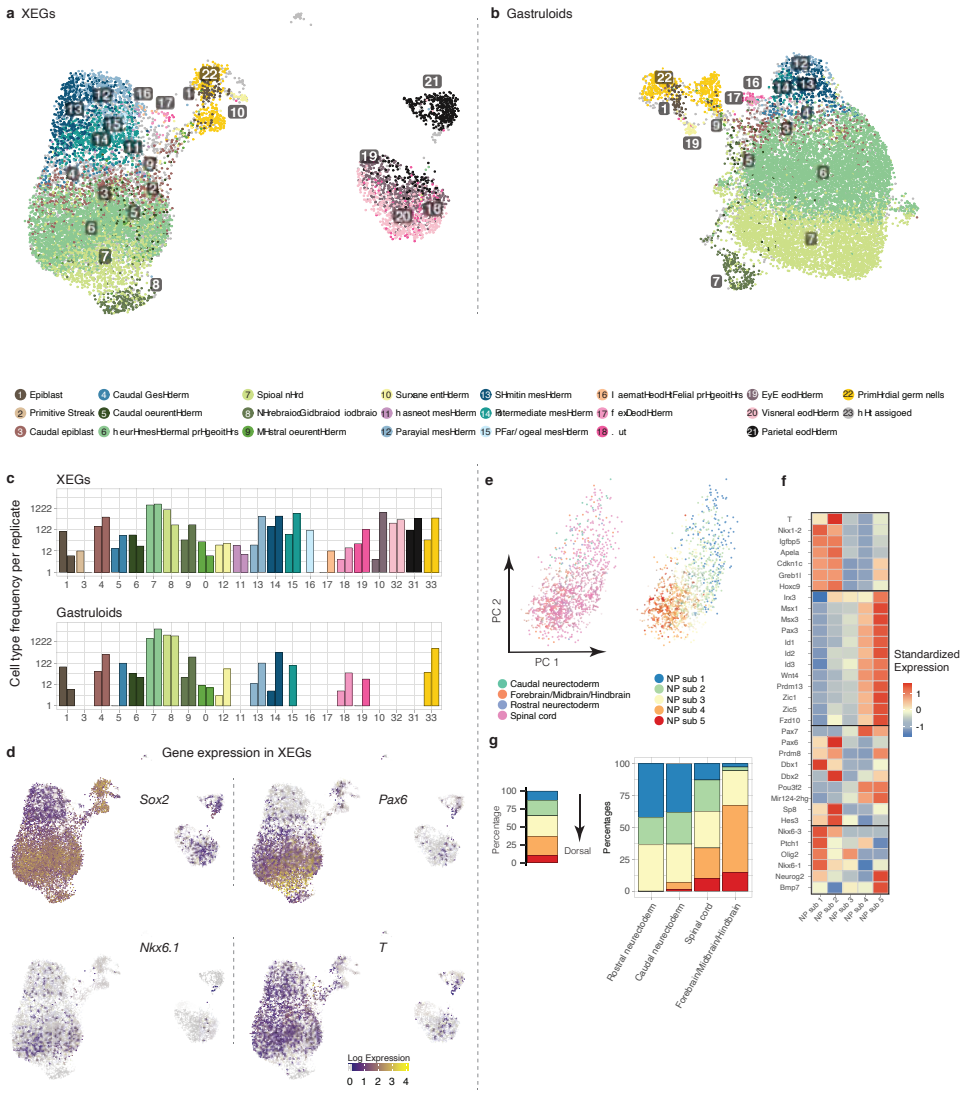


Figure 5.8: Single-cell RNA-seq reveals the transcriptional profiles of XEG cells. a,b, UMAP of cells in XEGs and gastruloids colored by cell type based on the mapping to in vivo data shown in Supplementary Fig. 4a,b. c, Cell type frequencies for both replicates in XEGs and gastruloids. d, Sox2, Pax6, Nkx6.1 and T log-expression levels indicated by color in UMAPs of XEGs. e, Principal Component Analysis of neural ectoderm-like cells in XEGs. Left: Colored by cell type based on the mapping to in vivo data. Right: Colored by neural progenitor subtypes (NP sub) found by sub-clustering. f, Heatmap of standardized expression of dorsoventral markers in the sub-clusters of neural ectoderm-like cells shown in e. g, Relative frequency (percentage) of sub-clusters. Left: For all neural ectoderm-like cells. Right: Per cell type (based on mapping to in vivo data).

We repeated this analysis on integrated neural ectoderm-like cells from gastruloids and XEGs. Strikingly, the clusters with clear dorsal characteristics were almost exclusively comprised of XEG cells (Fig. 5.9d). XEN-derived cells thus seem to promote dorsal spec-

ification in a subpopulation of neural ectoderm-like cells. Judged by the expression of Hox genes (Fig. 5.9e), there was no overt difference in anterior-posterior characteristics between the neural ectoderm-like clusters in XEGs. However, scRNA-seq did confirm the heterogeneous expression of Fgf8, Wnt4 and Wnt8a (Fig. 5.9f) we had observed by smFISH (Fig. 5.2b), hinting at additional subpopulations related to the anteroposterior axis in vivo.

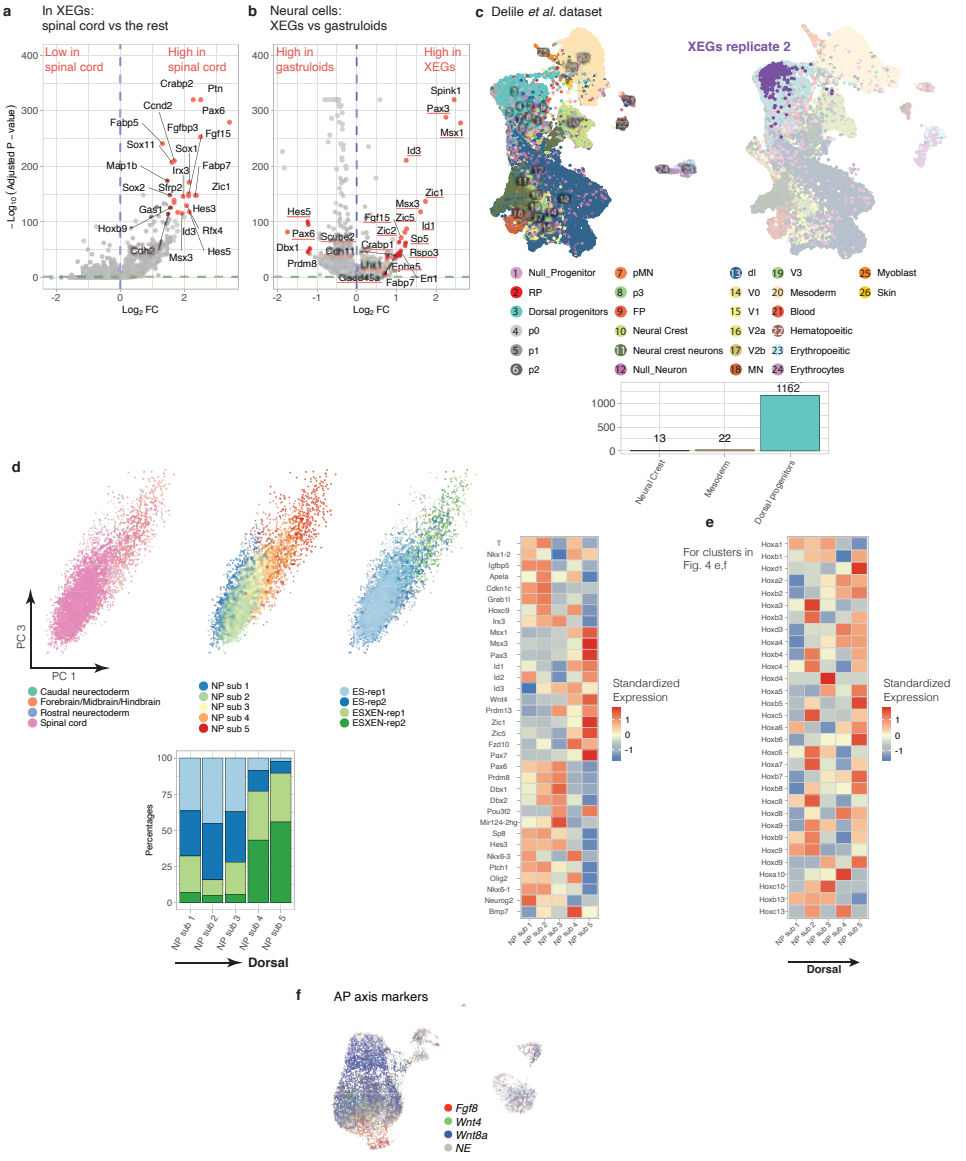


Figure 5.9: Neuroepithelial cells in XEGs are biased towards a dorsal expression profile. (Caption on the next page)

5.2.5 Most XEN cells become VE-like in XEGs

Compared to gastruloids, XEGs additionally contained extraembryonic endoderm cell types (Fig. 5.8c). By using GFP-expressing XEN cells in XEGs (Fig. 5.11a), we established that those cell types were exclusively differentiated from XEN cells. By comparison to undifferentiated XEN cells, which were spiked into the scRNA-seq samples, we studied the transcriptional changes in XEN-derived cells. Undifferentiated XEN cells mostly mapped to PE [38] (Fig. 5.12a, Fig. 5.11b), consistent with a previous study [42, 43]. Their derivatives in XEGs mapped to multiple kinds of extraembryonic endoderm: PE, embryonic VE and extraembryonic VE. Interestingly, some also mapped to gut, reminiscent of the contribution of VE to the gut in vivo [13, 14]. The identification of those cell types was confirmed by mapping our scRNA-seq data to an endoderm-focused scRNA-seq dataset [14] (Fig. 5.11c,d). Quantification revealed that, on average, 8% of the initially PE-like XEN cells acquired a gut-like and 66% a VE-like transcriptomic profile (29% embryonic VE, 37% extraembryonic VE) when co-cultured in XEGs. However, 25% retained a PE-like transcriptome (Fig. 5.12a). Differential gene expression analysis between undifferentiated XEN cells and XEN-derivatives revealed several differences (Fig. 5.11e, Supplementary Table 4). PE markers were less expressed in XEN-derivatives, while VE markers were highly expressed, suggesting that most XEN cells differentiate from a PE to a VE-like state in XEGs. To validate this finding experimentally, we performed single-molecule FISH of Dab2, Fst and Spink1 (Fig. 5.12b). Dab2 is a pan-extraembryonic endoderm marker [44], which is exclusively expressed in undifferentiated XEN cells and XEN-derived cell types in our scRNA-seq data set (Fig. 5.7c). Within the extraembryonic endoderm, Fst is expressed in the PE [45], whereas Spink1 is found in the VE [46]. The smFISH measurement showed that XEN-derived cells in XEGs only expressed Dab2 and Spink1, while undifferentiated XEN cells broadly co-expressed all markers. Some XEN cells in XEGs were also highly expressing E-cadherin, known to be expressed in VE [47] (Fig. 5.12c).

Figure 5.9: Neuroepithelial cells in XEGs are biased towards a dorsal expression profile. (Figure on the previous page) a, Gene expression differences between cells classified as "spinal cord" and all other cells in XEGs (fold-change vs p-value). Named genes are expressed in the neural tube according to previous studies (Supplementary Table 2). b, Gene expression differences between neural ectoderm-like cells in gastruloids and XEGs (fold-change vs p-value). Underlined genes are expressed in the dorsal part of the neural tube according to previous studies (Supplementary Table 3). c, Left, UMAP of the cells in the Delile et al. dataset [37], colored by cell type. Right, MNN mapping of cells classified as "spinal cord" in replicate 2 XEGs (opaque color) to the Delile et al. dataset (pale colors), as an example of the mapping procedure. Bottom: Absolute cell type frequency as a result of MNN mapping. d, Top row: Principal component analysis of neural ectoderm-like cells from all 4 data sets (XEGs and gastruloids) after integration. Colors from left to right: Cell types based on mapping to in vivo data, sub-clusters and data sets. Bottom: Relative frequency (percentage) of data sets in each sub-cluster. Right: Heatmap of standardized expression of dorsoventral markers in the sub-clusters shown in the PCA plot. e, Heatmap of standardized expression of Hox genes in neural ectoderm-like sub-clusters shown in Fig. 4e. f, UMAP of cells in XEGs with log-expression of the genes Wnt4, Wnt8a and Fgf8 indicated by colors. The UMAP shows both replicates, batch corrected.

However, the more anterior VE marker Hhex [48] was not detected by single-molecule FISH (Fig. 5.12d). Exposing undifferentiated XEN cells to CHIR in the same way as XEGs did not cause differentiation (Fig. 5.12d), which suggests that the interaction with mESCs plays a role. Cell-cell communication analysis of our scRNA-seq data with CellPhoneDB [49] suggested that mESCs and some mesodermal cell types in XEGs signal to the XEN-derived cells via the BMP4 pathway (Fig. 5.11f). This result is consistent with a previous study showing the differentiation of XEN cells in monolayer culture to a VE-like state with BMP [50].
Taken together, these results suggest that the mESCs or their derivatives induce the differentiation of the XEN-derived cells in XEGs. While undifferentiated XEN cells have both PE and VE characteristics, the majority (66%) of these cells becomes more VE-like. This effect is possibly mediated by BMP signaling originating in the mESCs.

5.2.6 XEN CELLS GUIDE SYMMETRY BREAKING BY LOCAL INHIBITION OF WNT SIGNALING

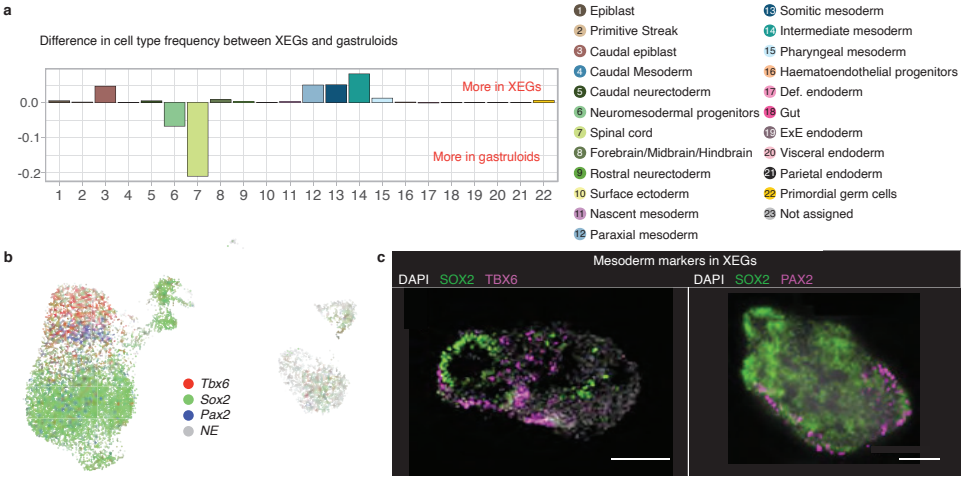


Figure 5.10: XEGs and gastruloids differ in cell type distribution. a, Differences between relative frequencies of cell types in XEGs and gastruloids. b, UMAP of cells in XEGs (both replicates, batch corrected) with log expression of Tbx6, Sox2 and Pax2 indicated by color. c, Expression of mesoderm markers. Left, TBX6 expression in a 96 h XEG (immunostaining of a section). Right, PAX2 expression in a 96 h XEG (wholemount immunostaining). Scale bars: 100 µm. Nuclei were stained with DAPI.

Although XEN-derived cells in XEGs did not express canonical AVE markers (Fig. 5.7c), we were wondering if they might effectively carry out an AVE-like function. XEN cells always formed the outermost layer (Fig. 5.1c, Fig. 5.3), resembling in vivo organization. Focusing on XEGs partially covered with XEN cells, we observed that epithelial structures were always adjacent to the XEN cells, while the T-positive population was on the opposite side (Fig. 5.3f). Notably, this organization was already established at 72 h, when aggregates are still spherical (Fig. 5.13a). This observation suggested that XEN cells guide symmetry breaking by a local effect on the adjacent mESCs.
We speculated that this effect might be mediated by a basement membrane produced by the XEN cells. As established above, cells were polarized early during XEG development,

prior to forming a columnar epithelium (Fig. 5.3c, d). This epithelium was supported by a basement membrane containing laminin and collagen (Fig. 5.13b, c), which were mostly produced by the XEN cells (Fig. 5.14a). CellPhoneDB analysis of the scRNA-seq data supported the existence of laminin signaling between XEN-derived cells and multiple mESC-derived cell types (Fig. 5.14b). It has been demonstrated previously, for small aggregates of mESCs, that the presence of an extracellular matrix can be sufficient for polarization and lumen formation [26, 27, 41]. Exposing gastruloids to a soluble basement membrane extract (Geltrex) did result in some, fragmented epithelia, if exposure was started after WNT activation (Fig. 5.14c). If exposure was started earlier, localized T-positive or SOX2-positive compartments were not formed, possibly due to unknown factors in the basement membrane extract that interfered with WNT activation. The results of the Geltrex experiments are consistent with the notion that the basement membrane provided by the XEN cells plays a role in epithelium formation.

To test whether XEN cells produced other, diffusible factors that were also involved, we grew gastruloids in medium conditioned by XEN cells (Fig. 5.13d). We observed that the gastruloids did not elongate and had a T-positive cell population that was restricted to the center of the aggregate. We hypothesized that the WNT inhibitor DKK1, which is expressed in XEN cells (Fig. 5.13e; Fig. 5.15a), might be one of those factors. In vivo, DKK1 is expressed by the AVE and limits the growth of the primitive streak [51], together with the NODAL antagonists CER1 and LEFTY1, which are not expressed in XEGs (Fig. 5.13e, Fig. 5.7c). Growing gastruloids in the presence of DKK1 resulted in a round morphology, with the T-positive cells confined to the center, as observed for XEN-conditioned medium (Fig. 5.13f, Fig. 5.15b). Factors limiting the primitive streak expansion in vivo are also known to preserve the anterior part of the epiblast and are thereby necessary for proper ectoderm domain differentiation [52]. Thus, we wanted to explore, if DKK1 could have a similar role in XEGs and bias differentiation towards the ectodermal lineage. Growing XEGs with the DKK1 inhibitor WAY-262611 [53] led to XEGs with elongated shapes but without epithelial structures (Fig. 5.13g, Fig. 5.15c). Since growing XEGs without CHIR resulted in similar epithelial structures as in regular XEGs (Fig. 5.15d), XEN cells likely suppressed pre-existing, low-level endogenous WNT activity [7, 54].

Finally, we noticed that XEN-derived cells highly expressed BMP2 (Fig. 5.15a) and that several of the dorsal markers expressed in XEGs are induced by BMP signaling (Supplementary Table 3). Cell-cell communication analysis of our scRNA-seq data supported the presence of BMP2 signaling between XEN-derived cells and multiple mESC-derived cell types, including neural ectoderm-like cells (Fig. 5.15e). Thus, XEN-derived cells might also contribute to the dorsal characteristics of the neural progenitor cells in XEGs.

All combined, our experiments suggest that XEN cells guide symmetry-breaking by local inhibition of cell differentiation into a T-positive population. Diffusible factors, including DKK1, and the presence of a basement membrane both seem to contribute to the formation of the neuroepithelial structures.

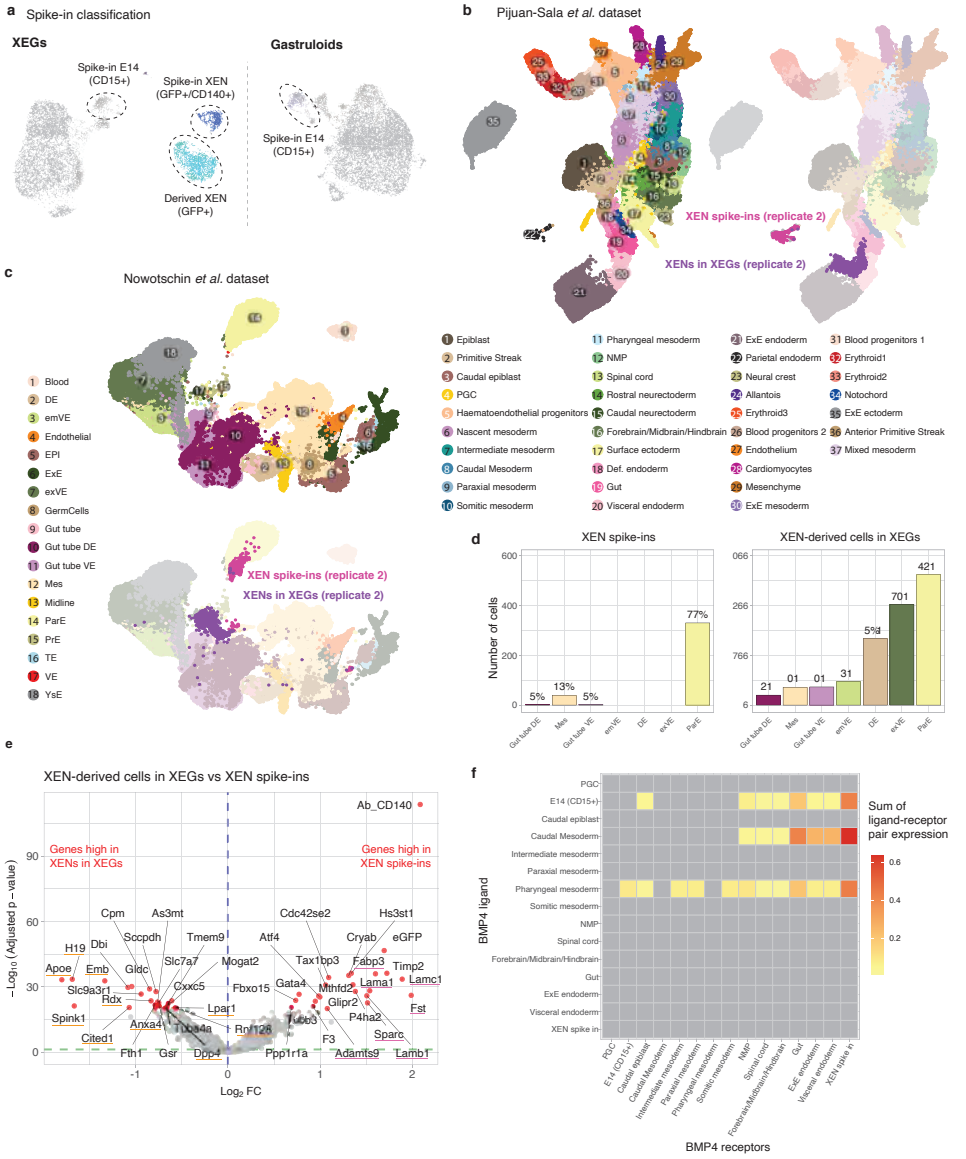


Figure 5.11: Most XEN-derived cells become visceral endoderm-like in XEGs. a, UMAP of cells in XEGs and gastruloids with spiked-in cells and XEN derived cells highlighted by color and circled (replicate 2). b, c, UMAPs of the Pijuan-Sala [38] or Nowotschin [14] dataset, respectively. XEN spike-ins and XEN-derived cells from XEG replicate 2 (opaque colors) are mapped to the in vivo datasets (pale colors). d, Cell type frequencies of XEN spike-ins and XEN derived cells in XEGs, resulting from knn assignments based on the mapping in (c). e, Gene expression differences between XEN spike-ins and XEN-derived cells in XEGs (fold-change vs p-value). Orange and pink lines indicate genes with PE-like and VE-like identity, respectively (see Supplementary Table 4). f, Sum of expression of BMP4 ligand-receptor pairs for cell types with significant communication identified by CellPhoneDB analysis. DE: definitive endoderm, emVE: embryonic visceral endoderm, EPI: epiblast, ExE: extraembryonic ectoderm, exVE: extraembryonic visceral endoderm, Mes: mesoderm, ParE: parietal endoderm, PrE: primitive endoderm, TE: trophectoderm, VE: visceral endoderm, YsE: yolk sac endoderm.

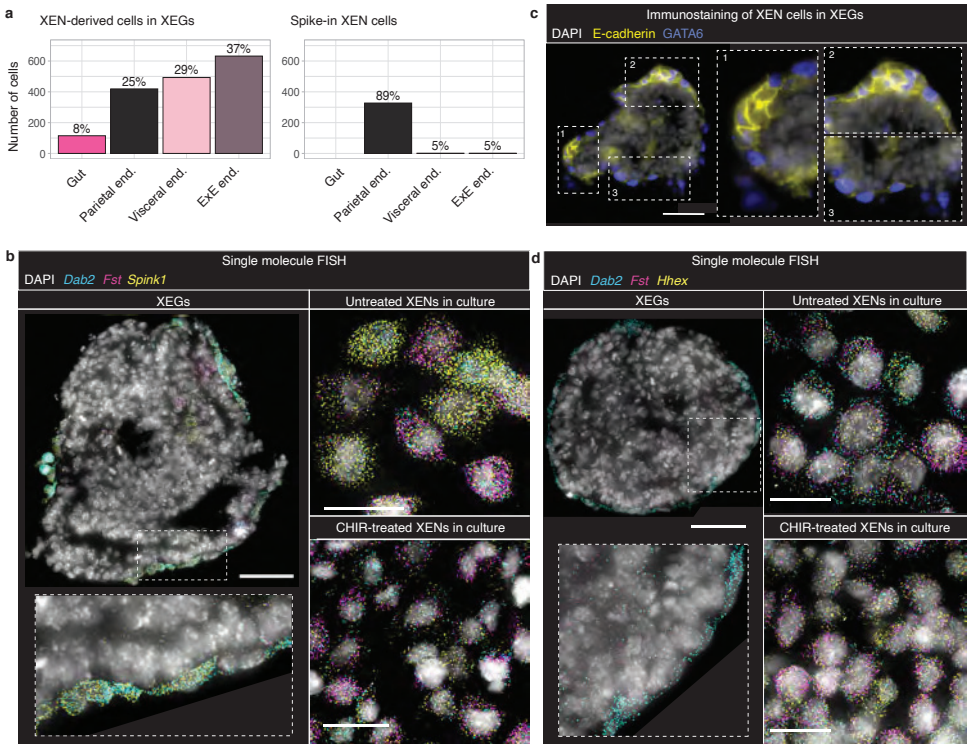


Figure 5.12: Most XEN cells become VE-like in XEGs. a, Left, cell types of XEN-derived cells in XEGs. Cells were classified as gut, PE ("parietal end."), embryonic VE ("visceral end.") or extraembryonic VE ("ExE end.") by mapping to the data set from Pijuan-Sala et al. [38]. Right, cell types of spiked-in XEN cells. b, Dab2, Spink1 and Fst expression in a section of an XEG at 96 h (left, scale bar: 50 μm), in XEN cells cultured under standard maintenance conditions (top right, scale bar: 20 μm) and in XEN cells treated with CHIR according to the XEG protocol (bottom right, scale bar: 20 μm). Expression was visualized by smFISH. Each diffraction limited dot is a single mRNA molecule. c, Expression of E-cadherin in XEGs at 96 h (immunostaining of sections). XEN cells were localized by expression of GATA6. Scale bars: 50 μm. d, Dab2, Fst and Hhex expression in a section of an XEG at 96 h (left, scale bar: 50 μm), in XEN cells cultured under standard maintenance conditions (top right, scale bar: 20 μm) and in XEN cells treated with CHIR according to the XEG protocol (bottom right, scale bar: 20 μm). Expression was visualized by smFISH. Each diffraction limited dot is a single mRNA molecule. b-d, Cell nuclei were stained with DAPI. The dashed boxes are shown at a higher magnification in the insets.

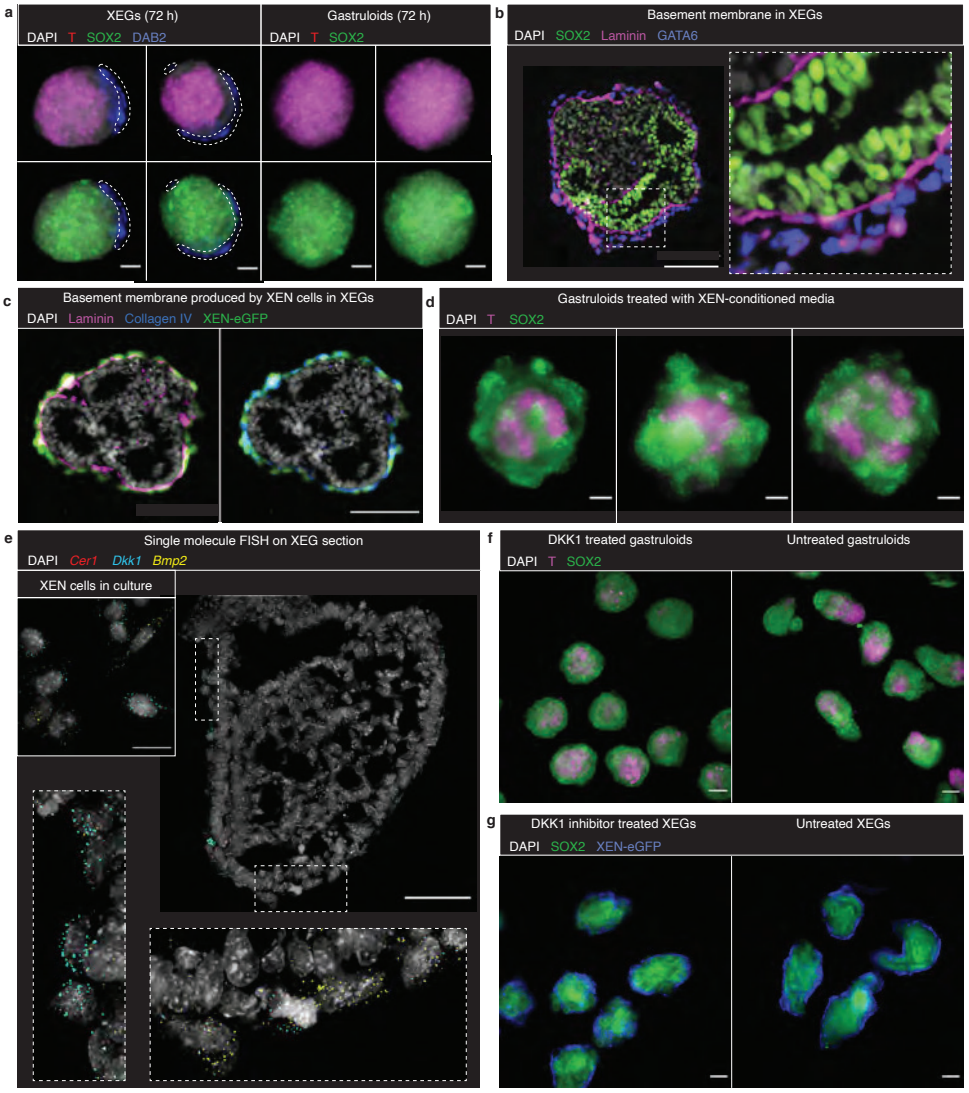


Figure 5.13: XEN cells guide symmetry breaking by local WNT inhibition and contribution to a basement membrane. a, T and SOX2 expression in XEGs (left) and gastruloids (right) at 72 h (z-projection of whole mount immunostaining). XEN cells were localized by expression of DAB2 and are indicated by a dashed outline. b, Expression of SOX2 and laminin in XEGs at 96 h (immunostaining of sections). XEN cells were localized by expression of GATA6. The dashed box is shown at a higher magnification in the inset. Scale bar: 50 μm. c, Expression of collagen IV and laminin in XEGs at 96 h (immunostaining of sections). XEN cells were localized by expression of H2B-GFP. Scale bar: 100 μm. d, T and SOX2 expression in gastruloids grown in XEN-conditioned media at 96 h (z-projection of whole mount immunostaining). e, Cer1, Dkk1 and Bmp2 expression in a section of a XEG at 96 h (scale bar: 50 μm) or XEN cells cultured under standard maintenance conditions (inset, scale bar: 20 μm). Expression was visualized by smFISH. Each diffraction limited dot is a single mRNA molecule. The dashed boxes are shown at a higher magnification in the insets. f, T and SOX2 expression in 96 h gastruloids treated with 200 ng/mL DKK1 and untreated (z-projection of whole mount immunostaining). Scale bars: 100 μm. g, T and SOX2 expression in 96 h XEGs treated with 0.25 μM DKK1 inhibitor WAY-262611 and untreated (z-projection of whole mount immunostaining). Scale bars: 100 μm. a-g, Cell nuclei were stained with DAPI.

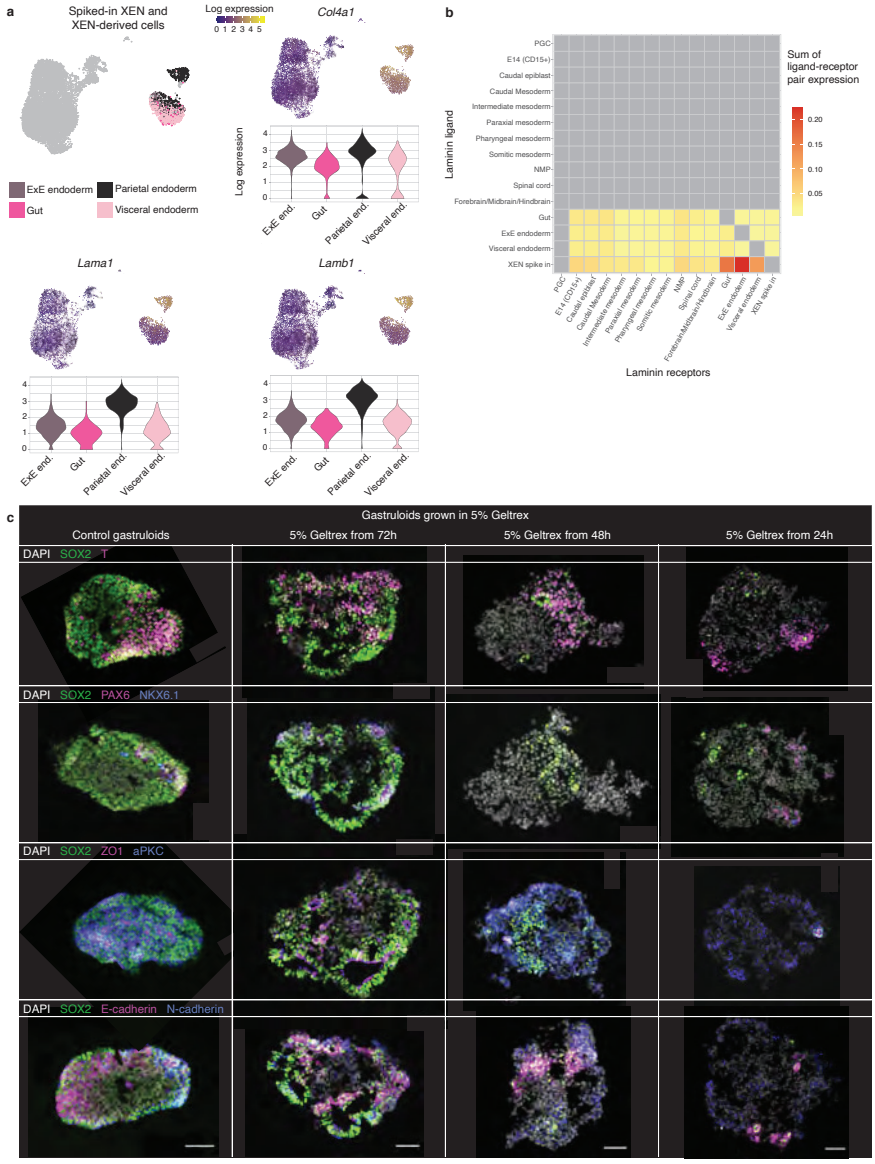


Figure 5.14: Basement membrane components produced by XEN cells, play a role in epithelia formation. a, Expression of genes in spiked-in XEN cells and XEN-derived cells in XEGs. Top left, UMAP with spiked-in XEN cells and XEN-derived cells colored by cell type (gut, parietal endoderm (parietal end.), embryonic VE (visceral end.) or extraembryonic VE (ExE end.)). Expression of basement membrane components collagen IV (Col4a1, top right), laminin alpha 1 (Lama1, bottom left) and laminin beta 1 (Lamb1, bottom right). UMAPs indicate log expression by color and contain both replicates, batch corrected. A violin plot of log expression in XEN-derived cell types is shown below the UMAP for each gene. b, Sum of expression of Laminin ligand-receptor pairs for cell types with significant communication identified by CellPhoneDB analysis. c, Expression of SOX2, T, PAX6, NKX6.1, ZO-1, $\alpha PKC\eta$, E-cadherin and N-cadherin in 96 h gastruloids grown in 5% Geltrex from 24 h, 48 h or 72 h onwards (immunostaining of cryosections). Scale bars: 50 µm. Cell nuclei were stained with DAPI.

5.3 Discussion

In this study we explored how the interaction between embryonic and extraembryonic cells in a multicellular in vitro system can lead to the formation of neuroepithelial tissue. While the neuroepithelial cells resembled in vivo neural progenitors transcriptionally, their organization was lacking, compared to embryonic neural tissue. In vivo, the neural tube forms via two distinct mechanisms [55]. During primary neurulation, the main part of the neural tube is formed by the folding of the neural plate, an epithelial sheet of neural ectoderm cells. Secondary neurulation, which gives rise to the most posterior part of the neural tube, works differently: mesenchymal cells condense to an epithelial rod which cavitates to form a tube [56, 57]. The two parts of the tube are then connected during junctional neurulation [58]. While we did not observe cell rearrangements characteristic of primary neurulation, the rosette formation seen in XEGs was reminiscent of secondary neurulation [55], which gives rise to the posterior neural tube. We could successfully differentiate XEGs further towards neural organoids that showed layered organization reminiscent of the developing spinal cord, which derives from the posterior neural tube.

The recently reported Trunk-Like Structures (TLS) [39], another gastruloid-derived model system, produce neural tube-like tissues, together with mesodermal tissue resembling somites. Notably, TLS are formed exclusively from mESCs and are grown in 5% Matrigel from 96 h onwards. Interestingly, the majority of neural tube cells in TLS had dorsal characteristics, as we also observed in XEGs. It will be interesting to explore, if the same mechanisms cause this phenomenon in both model systems. Another recent approach uses BMP4-treated ESCs as signaling centers to elicit neural tube-like tissues and other embryonic structures in untreated ESCs [14].

The fact that the majority of XEN cells becomes VE-like in XEGs clearly shows that there are reciprocal interactions between the co-differentiating, embryonic and extraembryonic cells. This observation supports the notion that such interactions are necessary for proper development, as previously observed in vivo [13, 59]. Recently, tissue-level organization has been achieved in vitro by exogenous induction of relevant signaling pathways [17, 40, 60]. XEN cells represent a potential alternative way to augment existing developmental in vitro systems, by providing a basement membrane and extraembryonic signaling inputs. Finally, with their large diversity of cell types, XEGs could be a starting point for developing more complex models containing all three germ layers as well as extraembryonic cells. Specifically, the CD31 positive endothelial cells observed in the neural organoids obtained from XEGs might be able to form a vascular network, if additional signaling cues are given [61]. In conclusion, in this study we showed how the gastruloid system can be used to explore complex heterotypic cell-cell interactions.

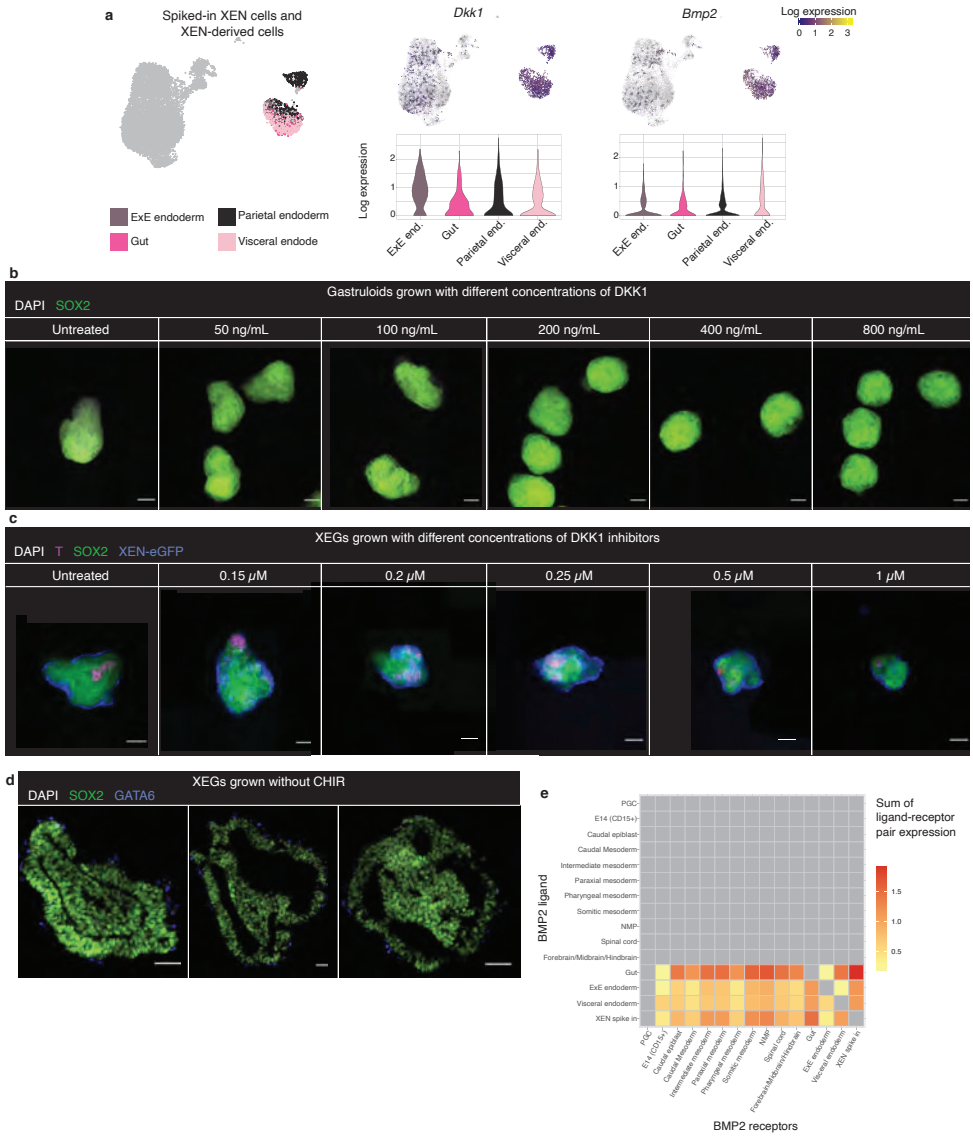


Figure 5.15: The WNT inhibitor DKK1, expressed exclusively by XEN cells and XEN-derived cells, plays a role in epithelia formation. a, Expression of genes in spiked-in XEN cells and XEN-derived cells. Left, UMAP with spiked-in XEN cells and XEN-derived cells colored by cell type (gut, parietal endoderm (parietal end.), embryonic VE (visceral end.) or extraembryonic VE (ExE end.)). Right, expression of signaling factors Dkk1 and Bmp2. UMAPs indicate log-expression by color and contain both replicates, batch corrected. A violin plot of log expression in XEN-derived cell types is shown below the UMAP for each gene. b, Expression of SOX2 in 96 h gastruloids treated with various concentrations of DKK1 between 24 h and 96 h (wholemount immunostaining). Scale bars: 100 µm. c, Expression of SOX2 and T in 96 h XEGs treated with various concentrations of DKK1 inhibitor WAY-262611 between 24 h and 96 h (wholemount immunostaining). d, Expression of SOX2 in XEGs grown without CHIR (immunostaining of cryosections). No specific T staining could be detected (data not shown). XEN cells were localized by expression of GATA6. Scale bars: 50 µm. e, Sum of expression of BMP2 ligand-receptor pairs for cell types with significant communication identified by CellPhoneDB analysis.

5.4 Methods

5.4.1 Experimental Methods

Cell culture

All cell lines were routinely cultured in KO DMEM medium (Gibco) supplemented with 10% ES certified FBS (Gibco),0.1 mM 2-Mercaptoethanol (Sigma-Aldrich), 1 × 100 U/mL penicillin/streptomycin, 1× MEM Non-Essential Amino Acids (Gibco), 2 mM L-glutamine (Gibco), 1000 U/mL mouse LIF (ESGRO). Cells were passaged every other day and replated in tissue-culture treated dishes coated with 0.2% gelatin. E14 mouse ES cells were provided by Alexander van Oudenaarden. The Sox1GFPiresPac mouse ES cell line was created by Mario Stavridis and Meng Li in the group of Austin Smith [62] and provided by Sally Lowell. XEN and XEN-eGFP were provided by Christian Schröter [43]. All cell lines were regularly tested for mycoplasma infection. The ES-mCherry-GPI cell line was obtained by introducing a mCherry-GPI transgene in the PdgfraH2B-GFP cell line, provided by the group of Anna-Katerina Hadjantonakis [63].

Differentiation

Gastruloids The gastruloid differentiation protocol was adapted from van den Brink et al. [6]. ES cells were collected from tissue-culture treated dishes by trypsinization, gentle trituration with a pipet and centrifugation (1200 r.p.m., 3 min). After collection, cells were resuspended in 2 mL of freshly prepared, prewarmed N2B27 medium: DMEM/F12 (Life technologies) supplemented with 0.5 × N2 supplement (Gibco), 0.5 × B27 supplement (Gibco), 0.5 mM L-glutamine (Gibco), 1 × 100 U/mL penicillin/streptomycin (Gibco), 0.5 × MEM Non-Essential Amino Acids (Gibco), 0.1 mM 2-Mercaptoethanol (Sigma-Aldrich). Cells were counted to determine the cell concentration. For gastruloids, 200 ES cells were seeded in 40 μL of N2B27 in each well of a round-bottom low-adherence 96-well plate. 48 h after seeding, 150 μL of prewarmed N2B27 supplemented with 3 μM of GSK3 inhibitor (CHIR99021, Axon Medchem) was added to each well. 72 h after seeding, 150 µL of medium was removed from each well and replaced by 150 μL of preheated N2B27. Gastruloids were collected at 96 h after seeding and fixed with 4% paraformaldehyde (PFA, Alfa Aesar) overnight at 4 °C.

For the experiments with gastruloids grown in Geltrex, cell aggregates were cultured with medium supplemented with 5% LDEV-Free, hESC-Qualified, reduced growth factor Geltrex (Gibco) from 24 h, 48 h or 72 h until the end of the protocol. At 96 h, gastruloids were washed twice with PBS supplemented with 1% BSA, then fixed with 4% PFA overnight at 4 °C.

XEN Enhanced Gastruloids (XEGs) ES and XEN cells were collected from tissue-culture treated dishes by trypsinization, gentle trituration with a pipet and centrifugation (1200 r.p.m., 3 min). After collection, cells were resuspended in 2 mL of fresh and prewarmed N2B27 medium. Cells were counted to determine cell concentration. For XEGs, several ratios of XEN and ES cells were tested (1:1, 1:2, 1:3, 1:4, 1:5) and compared with the regular gastruloid condition (0:1). The total number of cells was fixed at 200. Over two separate experiments, the proportion of organoids showing T staining and epithelial structures was quantified (total number of embryonic organoids 1:1=179, 1:2=143, 1:3=143, 1:4=140,

0:1=130) and the optimal ratio was determined to be 1:3 (Fig. 5.3e,f). A total of 200 cells (150 ES cells and 50 XEN cells) was seeded in 40 µL of N2B27 in each well of a round-bottom low-adherence 96-well plate. 48 h after seeding, 150 µL of prewarmed N2B27 supplemented with 3 µM of GSK3 inhibitor (CHIR99021, Axon Medchem) was added to each well. 72 h after seeding, 150 µL of medium was removed from each well and replaced by 150 µL of prewarmed N2B27. XEGs were collected at 96 h after seeding and fixed with 4% PFA overnight at 4 °C.
For the experiment of XEGs grown without GSK3 inhibitor, cells were seeded as usual. At 48 h, 150 µL of preheated N2B27 was added to each well. At 72 h, 150 µL of medium was removed from each well and replaced by 150 µL of prewarmed N2B27. XEGs were collected at 96 h after seeding. For the smFISH control experiments, XEN cells were seeded at low density in N2B27 medium. At 48 h the medium was replaced by prewarmed N2B27 supplemented with 3 µM of GSK3 inhibitor. 72 h after seeding, the medium was replaced with prewarmed N2B27. Cells were fixed at 96 h with 4% PFA for 1 h at 4 °C.

Neural differentiation For neural differentiation, a protocol for creating cerebral organoids was adapted from Lancaster et al. [36]. Instead of collecting XEGs at 96 h, the medium was replaced by cerebral organoid differentiation medium: DMEM-F12 (Life technologies), Neurobasal (Gibco), 0.5 × B27 supplement containing vitamin A (Gibco), 0.5 × N2 supplement (Gibco), 2.5 μM/mL Insulin, 2mM L-glutamine (Gibco), 0.5 × cMEM-Non-EssentialcAmino Acids (Gibco), 1 × 100 U/mL penicillin-streptomycin and 0.05 mM 2-Mercaptoethanol (Sigma-Aldrich). At 168 h, aggregates were collected and transferred, with fresh medium, into 10 cm dishes on an orbital shaker installed in the incubator (85 r.p.m.). Aggregates were grown until 192 h (8 days) during which medium was refreshed every other day until collection. Collected aggregates were fixed with 4% PFA for 48 h at 4 °C.

Signaling experiments

In the signaling experiments with XEGs, aggregates were treated between 72 h and 96 h with either LDN193189 (BMPi, 100 nM, Reagents Direct), a potent BMP pathway inhibitor, Purmorphamine (1 µM, STEMCELL Technologies), a small molecule agonist of the hedgehog pathway, Retinoic acid (RA, 100 nM, Sigma-Aldrich) or DMSO (0.1% final concentration, Sigma Aldrich) as a vehicle control. For this experiment, the XEGs were allowed to grow for an additional 48 h before fixation (144 h total growth) and preparation for staining (see Immunostaining).
DKK1 signaling pathways perturbation was performed in two ways, using DKK1 (Sigma-Aldrich) for activation in gastruloids and Way262611 (Sigma-Aldrich) for inhibition in XEGs. Gastruloids and XEGs were seeded according to the usual protocols. At 24 h, 40 µL of N2B27 supplemented with various concentration of DKK1 or Way262611 respectively, were added to each well. Next steps of the protocol were performed using N2B27 supplemented with DKK1 or Way262611. Aggregates were fixed at 96 h with 4% PFA overnight at 4 °C.

Immunostaining

Fixation and blocking After collection, gastruloids and XEGs were fixed in 4% PFA at 4 °C overnight. Tissue resulting from the cerebral organoid protocol was fixed under the same conditions, but for 48 h. After fixation, samples were washed three times in

washing solution (PBS, 1% bovine serum albumin (BSA)) and incubated at 4 °C in blocking buffer (PBS, 1% BSA, 0.3% Triton-X-100) for a minimum of 16 h. Samples for smFISH were washed 3 times in PBS after fixation and stored in 70% ethanol at 4 °C. To stain E14 cells for pluripotency markers, cells in suspension were fixed for 30 min in 4% PFA at 4 °C, washed three times in washing solution at RT and incubated in blocking buffer for 1 h at 4 °C.

Whole-mount immunolabeling and clearing Immunolabeling and clearing of gastruloids and XEGs were based on the protocol described by Dekkers et al. [64]. Briefly, after fixation and blocking, samples were incubated with primary antibodies at 4 °C overnight on a rolling mixer (30 r.p.m.) in organoid washing buffer (OWB) (PBS, 2% BSA, 0.1% Triton-X-100) supplemented with 0.02% sodium dodecyl sulfate (SDS), referred to as OWB-SDS. The following primary antibodies were used: rat anti-SOX2 (1:200, 14-9811-82, Thermo Fisher Scientific), goat anti-T (1:200, sc-17745, Santa Cruz Biotechnology), goat anti-T (1:100, AF2085, R&D systems), mouse anti-DAB2 (1:100, 610464, BD Biosciences). The next day, samples were washed three times for 2 h in OWB-SDS at RT, followed by incubation with secondary antibodies (donkey anti-goat Alexa Fluor 488 (1:200, A-11055, Thermo Fisher Scientific), donkey anti-rat Alexa Fluor 488 (1:200, A-21208, Thermo Fisher Scientific), donkey anti-goat Alexa Fluor 555 (1:200, A-21432, Thermo Fisher), donkey anti-mouse Alexa Fluor 555 (1:200, A-31570, Thermo Fisher Scientific), chicken anti-rat Alexa Fluor 647 (1:200, A-21472, Thermo Fisher Scientific)) and 4',6-diamidino-2-phenylindole (DAPI, 1 µg/mL, Merck) in OWB-SDS at 4 °C overnight on a rolling mixer (30 r.p.m.), protected from light. Finally, samples were washed three times for 2 h in OWB-SDS at RT. Clearing was performed by incubation in fructose-glycerol clearing solution (60% vol/vol glycerol, 2.5 M fructose) for 20 min at RT. Samples were imaged directly after clearing or stored at 4 °C in the dark.

Cryo-sectioning and immunolabeling of sections Prior to cryosectioning, fixed and blocked samples were incubated sequentially in sucrose solutions (10, 20 and 30%) for 30 min (gastruloids and XEGs) or 2 h (neural organoids) at 27 °C, and embedded in optimal cutting temperature (OCT) compound. Samples in OCT were placed on dry ice for rapid freezing, and stored at -80 °C prior to cryosectioning. Samples were cut to cryosections (10 µm thickness) using a cryostat (Thermo Fisher Scientific, USA) and cryosections were placed on poly-L-lysine coated glass slides (Merck). The slides were stored directly at -80 °C. For immunofluorescence staining, slides were thawed and rinsed with PBS for 10 min at RT to dissolve the OCT. Subsequently, slides were incubated overnight at 4 °C with the following primary antibodies diluted in blocking buffer: rat anti-SOX2 (1:200, 14-9811-82, Thermo Fisher Scientific), goat anti-T (1:200, sc-17745, Santa Cruz Biotechnology), mouse anti-N-cadherin (1:200, 33-3900, Thermo Fisher Scientific), rabbit anti-E-cadherin (1:200, 3195, Cell Signaling Technology), rabbit anti-PAX6 (1:100 (cerebral organoids) or 1:200 (gastruloids, XEGs), 42-6600, Thermo Fisher Scientific), mouse anti-NKX6.1 (1:200, F55A12, Developmental Studies Hybridoma Bank), rabbit anti-NKX6.1 (1:200, HPA036774, Merck), mouse anti-TUJ1 (1:200, 801202, BioLegend), goat anti-PAX2 (1:200, AF3364, R&D Systems), goat anti-TBX6 (1:200, AF4744, R&D Systems), mouse anti-ASCL1 (1:200, 14-5794-80, Thermo Fisher Scientific), rat anti-CTIP2 (1:200, ab18465, abcam), rabbit anti-CD31 (1:50, ab28364, Abcam), rabbit anti-GATA6 (1:200, PA1-104, Thermo Fisher Scientific), goat

anti-GATA6 (1:200, AF1700, R&D Systems), rabbit anti-Laminin (1:200, PA1-16730, Thermo Fisher Scientific), mouse anti-OCT4 (1:200, MA1-104, Thermo Fisher Scientific). The next day, the slides were washed twice for 10 min in PBS at RT. Subsequently, the slides were incubated with secondary antibodies (donkey anti-goat Alexa Fluor 488 (1:200, A-11055, Thermo Fisher Scientific), donkey anti-rat Alexa Fluor 488 (1:200, A-21208, Thermo Fisher Scientific), donkey anti-goat Alexa Fluor 555 (1:200, A-21432, Thermo Fisher), donkey anti-mouse Alexa Fluor 555 (1:200, A-31570, Thermo Fisher Scientific), chicken anti-rat Alexa Fluor 647 (1:200, A-21472, Thermo Fisher Scientific), donkey anti-rabbit Alexa Fluor 647 (1:200, A-31573, Thermo Fisher Scientific)) and DAPI (1 µg/mL, Merck) in blocking buffer for 4 h at 4 °C, and washed three times for 10 min at RT. Slides were mounted in ProLong™ Gold Antifade Mountant (Thermo Fisher Scientific) and imaged after 24-48 h.

Immunolabeling of E14 cells After fixation and blocking, E14 cells were incubated with the following primary antibodies in blocking buffer overnight at 4 °C: rat anti-SOX2 (1:200, 14-9811-82, Thermo Fisher Scientific) and mouse anti-OCT4 (1:200, MA1-104, Thermo Fisher Scientific). The next day, cells were washed three times in washing solution for 5 min at RT and incubated with secondary antibodies (donkey anti-rat Alexa Fluor 488 (1:200, A-21208, Thermo Fisher Scientific) and donkey anti-mouse Alexa Fluor 555 (1:200, A-31570, Thermo Fisher Scientific)) and DAPI (1 µg/mL, Merck) in blocking buffer for 3 h at 4 °C. Finally, the cells were washed three times in washing solution for 5 min at RT and imaged directly.

Single-molecule fluorescence in-situ hybridization (smFISH)

smFISH was performed as described previously [65]. Briefly, samples were fixed with PFA and stored in 70% ethanol, as described above. Custom designed smFISH probes for Dab2, Fst, Hhex and Spink1 (BioCat, Supplementary Table 5), labeled with Quasar 570, CAL Fluor Red 610, or Quasar 670, were incubated with the samples overnight 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; Thermo Fisher Scientific). Samples were washed twice for 30 min at 30 °C with wash buffer (25% formamide, 2X SSC). The wash buffer was supplemented with DAPI (1 μg/mL) in the second wash step. All solutions were prepared with RNAse-free water. Finally, the samples were mounted in ProlongGold (Life Technologies) and imaged when hardened (sections) or immediately (ibidi dishes). All components are from Sigma-Aldrich unless indicated.

Imaging

Fixed and stained samples were imaged on a Nikon Ti-Eclipse epifluorescence microscope equipped with an Andor iXON Ultra 888 EMCCD camera and dedicated, custom-made fluorescence filter sets (Nikon). Primarily, a 10× / 0.3 Plan Fluor DLL objective, a 20× / 0.5 Plan Fluor DLL objective, or a 40× / 1.3 Super Fluor oil-immersion objective (Nikon) were used. To image complete sections of neural organoids, multiple adjacent fields of view were acquired and combined using the tiling feature of the NIS Elements software (Nikon). Z-stacks were collected of whole-mount gastruloids and XEGs with distances of 10 μm between planes. For smFISH measurements, z-stacks were collected with a distance of 0.2 μm between planes in four fluorescence channels (DAPI, Quasar 570, CAL Fluor Red

610, Quasar 670) using a 100× /1.45 Plan Apo Lambda oil (Nikon) objective. Time lapses to observe the formation of epithelial structures were performed 24 h and 48 h after cell seeding, on XEGs grown from the mCherry-GPI ES cell line. XEGs were transferred to a glass-bottom μ-Slide imaging chamber (ibidi) and imaged every 30 min for 24 h with a Nikon Eclipse Ti C2+ confocal laser microscope (Nikon, Amsterdam, The Netherlands), equipped with lasers at wavelengths 408, 488 and 561, an automated stage and perfect focus system at 37°C and 5% CO2. Images were acquired with a Nikon 20x Dry Plan Apo VC NA 0.75 objective. To track SOX1 expression in gastruloids and XEGs during the 24 h growth after the GSK3 inhibitor pulse, 72 h gastruloids and XEGs grown from the Sox1GFPiresPac ES cell line were transferred to a glass-bottom μ-Slide imaging chamber (ibidi) and imaged every 40 min for 24 h, while temperature and CO2 levels were maintained at 37 °C and 5%, respectively, by a stage top incubator (INUG2-TIZW-SET, Tokai Hit) mounted on the Nikon Ti-Eclipse epifluorescence microscope.

Single-cell RNA-seq library preparation and sequencing

For each replicate, 96 pooled gastruloids and 96 pooled XEGs were collected from a round-bottomed low-adherence 96-well plate in 15 mL Falcon tubes and pelleted by gentle centrifugation (500 r.p.m. for 2 min). No final aggregate was excluded from the collection. After washing with cold PBS, samples were resuspended in N2B27. Cells were then dissociated by 5 min incubation in TrypLE (Gibco) and gentle trituration with a pipet, centrifuged and resuspended in 1 mL of cold N2B27. Cells were counted to determine cell number and viability. For the first replicate, ES-mCherry-GPI were spiked in at a frequency of 5%. For the second replicate, E14 cells were collected from culture dishes and incubated for 30 min at 4 °C with CITE-seq cell hashing [66] antibody Ab_CD15 (1:200) (Biolegend). XEN-eGFP were collected from culture plates and incubated for 30 min at 4 °C with CITE-seq cell hashing antibody Ab_CD140 (1:200) (Biolegend). In the gastruloid sample, labeled E14 cells were spiked in at a frequency of 5%, whereas in the XEG sample labeled E14 and XEN-eGFP were spiked in, both at a frequency of 5%. High viability of the cells in all samples was confirmed before 10X library preparation. Single-cell RNA-seq libraries were prepared using the Chromium Single Cell 3' Reagent Kit, Version 3 Chemistry (10x Genomics) according to the manufacturer's protocol. CITE-seq libraries were prepared according to the CITE-seq protocol from New York Genome Center version 2019-02-13. Libraries were sequenced paired end on an Illumina Novaseq6000 at 150 base pairs.

5.4.2 Computational methods

Analysis of single-cell RNA-sequencing data

Single-cell RNA-seq data pruning and normalization Cells with a low number of transcripts were excluded from further analysis based on the histograms in Fig. 5.6a (count < 1300 for replicate 1 of the XEG experiment and count < 2300 for the other datasets). Genes expressed in less than 2 cells (across merged replicates) were excluded from further analysis. The final XEG dataset contains 14286 genes and 4591 or 6857 cells for replicate 1 or 2, respectively. The gastruloid dataset contains 14384 genes and 4233 or 8363 cells per replicate. The two datasets were normalized using the scran R-package (V 1.10.2 [67]). Gene variabilities were calculated (improvedCV2, scran) for each replicate separately, after excluding ribosomal genes [Ribosomal Protein Gene Database,

http://ribosome.med.miyazaki-u.ac.jp/], exogenously expressed genes and the cell hashing antibodies. The 10% most highly variable genes (HVG) were selected based on variability p-values.

Dimensionality reduction For each of the two datasets, the two replicates were batch corrected with the fast mutual nearest neighbors (MNN) method implemented in the scran R-package [68], using the union of the 10% HVG of the two replicates and log-transformed normalized counts with d = 120 (number of principal components) and k = 50 (number of nearest neighbours). For dimensionality reduction, a uniform manifold approximation and projection (UMAP) was calculated on the batch corrected data using the R-package UMAP (V 0.2.3.1) with n = 50, min_dist = 0.7 and using the cosine distance measure.

Identification of spike-in cells Cells with any expression of mCherry were annotated as ES (mCherry+). The remaining spike-in cells, E14 (CD15+) and XEN spike-in (CD140+) (see Single-cell RNA-seq library preparation and sequencing), could not be determined by the expression level of the antibody alone. We therefore chose to assign spike-ins based on clusters. For each of the two datasets, a shared nearest neighbor graph was constructed from the batch corrected data (see Dimensionality reduction) with scran using k = 20 and d = 30. Louvain clustering was performed on the constructed graphs with the R-package igraph (V1.2.4.1), which resulted in 8 clusters for XEGs and 7 clusters for gastruloids (see Fig. 5.6c). We identified 3 out of the 8 clusters in XEGs based on literature markers and spike-in gene expression. One cluster out of these three was mainly comprised of mESCs, due to high Ab_CD15 expression and mCherry positive cells. Cells that had an expression of Ab_CD15 > 50 and were part of this cluster were considered spiked-in E14 and annotated as E14 (CD15+). The other two clusters were both eGFP positive, where one of them had a higher Ab_CD140 expression and was thus annotated as XEN spike-in (Ab_CD140+). The second cluster was annotated as XEN derived (Ab_CD140-). Similarly, for gastruloids, one of the 7 clusters was comprised of mainly mESCs based on literature markers and spike-in gene expression. Cells that had an expression of Ab_CD15 > 100 and were part of this cluster were considered spiked-in E14 and annotated as E14 (CD15+).

Analysis of cell cycle and stress-related genes For each of the two datasets, cell cycle analysis was performed with the scran package using the cyclone function [69] on the normalized counts. Cells in G2M phase were distributed evenly across all clusters and thus the clustering was not biased by cell cycle. No other separate cluster that consisted entirely of cell cycle related cells appeared. For the analysis of stress-related genes, a list of known stress genes [70] was used to calculate the average standardized expression per cell based on normalized counts. Stress-related genes were mainly found within the spike-in cells and there was no other separate cluster that consisted entirely of highly stressed cells.

Mapping to in vivo datasets Our datasets were mapped to three different in vivo datasets.

Pijuan-Sala et al. dataset
The Pijuan-Sala et al. dataset [38], which was downloaded from https://content.cruk.cam.

ac.uk/jmlab/atlas_data.tar.gz, consists of 9 timepoints from E6.5 to E8.5. The data was normalized by size factors provided by the authors. Cells with no cell type assignment were excluded from further analysis. The 10% HVG were calculated (improvedCV2, scran package) on the remaining cells excluding sex genes, similar to Pijuan-Sala et al.'s method. Cells in the "mixed_gastrulation" cluster were also excluded. MNN mapping was applied to log-transformed normalized counts of the 10% HVG. First, in vivo timepoints were mapped to each other in decreasing order. Then, each of our four datasets was mapped separately to the combined Pijuan-Sala et al. dataset (MNN method with d = 120, k = 50). K-nearest-neighbor (knn) assignment was performed in the batch corrected principal component space. For each cell in our datasets, the 50 nearest neighbors in the in vivo dataset, based on Euclidean distances, were calculated. Each cell was assigned the most abundant cell type within the knn, if certain distance and confidence score conditions were met. This confidence score was calculated for each cell as the number of the most abundant cell type divided by the total number of neighbors (k=50). A cell was annotated as "Not assigned" if either, the average distance to its nearest neighbor exceeded a certain threshold (determined by the long tail of the histogram of average distances for each of our datasets separately) or the assignment had a confidence score less than 0.5. Additionally, we placed cells in "Not assigned" if they were assigned to clusters with less than 10 cells, or to the cluster "Blood progenitors 2" (because this cluster did not show distinct expression of known literature markers). This resulted in 22 assigned clusters for XEGs and 15 assigned clusters for gastruloids. For each cell in our dataset we calculated the average and the standard deviation of the developmental age of the knn.

Nowotschin et al. dataset

The Nowotschin et al. dataset [14], which was downloaded from https://endoderm-explorer.com/, consists of 6 timepoints from E3.5 to E8.75. The data was normalized (scran) and the 10% HVG were calculated (improvedCV2, scran package). First, MNN was applied to the Nowotschin et al. dataset in increasing order of the timepoints (using log-transformed normalized counts of the 10% HVG, d = 150, k =50). Then, XEN cells from our XEG dataset (XEN spike-ins (CD140+) and XEN derived (CD140-)) were mapped to the MNN-corrected Nowotschin et al. dataset. Knn assignment was performed as described above and resulted in 7 assigned clusters.

Delile et al. dataset

The Delile et al. dataset [37], which was downloaded from https://github.com/juliendelile/MouseSpinalCordAtlas, consists of 5 timepoints from E9.5 to E13.5. Cells that had a cell type assignment of "Null" or "Outlier" were excluded from further analysis. The data was normalized (scran) and the 10% HVG were calculated. First, MNN was applied to the Delile et al. dataset in order of increasing timepoints (log-transformed normalized counts of the 10% HVG, d = 120, k =50). Then, we mapped neural ectoderm-like clusters, identified through the mapping to the Pijuan-Sala et al. dataset ("Rostral neurectoderm", "Caudal neurectoderm", "Spinal cord" and "Forebrain/Midbrain/Hindbrain") to the MNN corrected Delile et al. dataset separately for each of our replicates. Knn assignment was performed as described above and resulted in 3 clusters for XEGs and 3 clusters for gastruloids.

Differential expression analysis For the differential expression test between "spike-in XENs" and "XENs in XEGs" a Welch t-test (implemented in findMarkers, scran R package) was conducted on the normalized log-transformed counts. The test was performed on XEGs from replicate 2. "spike-in XENs" were chosen as the 100 cells with highest Ab_CD140 expression and "XENs in XEGs" were the 100 cells with lowest Ab_CD140 expression within the XEN identified cells.
For the differential expression test between XEGs and gastruloids, a negative binomial regression was performed (R package edgeR V 3.24.3 [71]). Based on the knn assignment to the Pijuan-Sala et al. dataset, all "neural ectoderm-like" clusters ("Rostral neurectoderm", "Caudal neurectoderm", "Spinal cord" and "Forebrain/Midbrain/Hindbrain") were extracted from our four datasets (XEGs: 975 cells in replicate 1 and 357 cells in replicate 2; gastruloids: 2134 cells in replicate 1 and 2106 cells in replicate 2). Raw counts were used for the regression with these four subsets as dummy variables and a variable corresponding to the total number of counts per cell. P-values were obtained for the contrast between XEGs and gastruloids using the average regression coefficients among variables of both replicates.
Similarly, for the differential expression test of the "Spinal cord" in XEGs, a negative binomial regression was used. Cells were excluded from the test if either their cell type occurred in less than 10 cells per replicate, or if the cells were annotated as "Not assigned", leaving a total of 13 cell types (7742 cells) to be considered. For each cell type and each replicate a dummy variable was created and a variable corresponding to the total number of counts per cell. Then, p-values were obtained for the contrast between the average regression coefficients of the two replicates of the "Spinal cord" cluster and the average regression coefficients of all other variables considered in the test.
For all differential expression tests p-values were adjusted for multiple hypothesis testing with the Benjamini-Hochberg method.

Sub-clustering of neural ectoderm-like cells Neural ectoderm-like cells ("Rostral neurectoderm", "Caudal neurectoderm", "Spinal cord" and "Forebrain/Midbrain/Hindbrain") were extracted from the XEG data sets for Fig. 4e-g and Fig. S5d. A curated list of genes that are dorsoventral axis markers in the developing neural tube [29, 37, 39] was used for all analysis steps (see Fig. 5.8f for the complete list). First, replicates were integrated with MNN using d=5 and k=20. The UMAP was created from the MNN corrected subspace with 20 nearest neighbors, min_dist = 0.3 and cosine metric. K-means clustering was performed on the MNN corrected subspace using Euclidean distances and 5 centers.
For Fig. 5.9d, neural ectoderm-like cells were extracted from XEG and gastruloid data sets. As before, only genes listed in the heatmap in Fig. 5.8f were used for all analysis steps. MNN mapping was performed using d=15 and k=20 in the following sequence: XEG replicate 2, XEG replicate 1, gastruloid replicate 2 and gastruloid replicate 1. UMAP and clustering was performed as described before. To correct for the difference in the number of cells coming from the 4 samples, first, relative frequencies for the 5 sub-clusters were calculated per sample. These frequencies were then normalized by dividing by the sum of relative frequencies for a specific sub-cluster.

Cell-cell interaction analysis with CellPhoneDB CellPhoneDB [49] was applied to the raw counts of replicate 2 of the XEG data set. All mouse gene names were converted

to human gene names with the biomaRt R package. All clusters, assigned through the mapping to the Pijuan-Sala et al. dataset, were used. Finally, results containing the ligands of interest (BMP2, BMP4 and LAM) were extracted. For each pair of cell types with significant communication (p-value < 0.05), the expression of all significant ligand-receptor pairs was summed. The expression of a ligand-receptor pair was taken to be the average of ligand and receptor expression.

Image analysis

Image stacks of whole-mount immunostained gastruloids and XEGs, and images of immunostained sections were pre-processed by background subtraction (rolling ball, radius: 50 pixels = 65 μm (10× objective), 32 μm (20× objective) or 16 μm (40× objective)) in the channels that showed autofluorescent background using ImageJ [72]. When background subtraction in images of sections did not result in proper removal of autofluorescent background signal, the Enhance Local Contrast (CLAHE) tool was used in ImageJ [72]. smFISH image stacks were pre-processed by applying a Laplacian of Gaussian filter (σ = 1) to the smFISH channels using scikit-image (v0.16.1) [73]. For all image stacks, a maximum projection was used to obtain a 2D representation. To show a single object per image, images were cropped around the object of interest.

5.4.3 Data availability

The single-cell RNA sequencing datasets generated in this study are available in the Gene Expression Omnibus repository, GSE141530. Supplementary tables and videos are available in the online version of the manuscript at https://doi.org/10.1177/20417314221103042.

Acknowledgements We are thankful to Alfonso Martinez Arias for insightful discussions and feedback on the manuscript. We acknowledge Anna-Katerina Hadjantonakis for helpful input at various stages of the project. We also thank Dr. Sylvia Le Dévédec and Hans de Bont of the Leiden University Cell Observatory for their support in this work. N. B.-C., M. M., P. v.d. B., M. F. and S.S. were supported by the Netherlands Organisation for Scientific Research (NWO/OCW, www.nwo.nl), as part of the Frontiers of Nanoscience (NanoFront) program. E.A. acknowledges support by a Stichting voor Fundamenteel Onderzoek der Materie (FOM, www.nwo.nl) projectruimte grant (16PR1040). M.H. acknowledges support by a Netherlands Organisation for Scientific Research (NWO/OCW, www.nwo.nl) VIDI grant (016.Vidi.189.007). This work was carried out on the Dutch national e-infrastructure with the support of SURF Cooperative. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author contributions N.B.-C., E.A. and M.H. cultured gastruloids and XEGs. N.B-C., E.A. and M.H. performed signaling experiments and immunostaining and analyzed the resulting images, N.B-C. prepared samples for single-cell RNA sequencing and interpreted the sequencing data, M.M. performed the computational analysis of the single-cell RNA sequencing data, P.v.d.B. contributed to the computational analysis of single-cell RNA sequencing data and carried out the smFISH measurements, M.F. supported all experiments and performed all cryosectioning, N.B.-C., M. M., E.A., P.v.d.B. and M.H. produced figures, N.B.-C., M. M., E.A., P.v.d.B., M.H. and M.F. contributed to the manuscript, T.I., S.T. and

S.S. conceived the study and acquired funding. S.S. interpreted the data and wrote the manuscript. All authors discussed the results and commented on the manuscript at all stages.

Competing interest The authors declare no competing interests.

Code availability Custom R and python code used to analyze the data is available from the authors upon request.

Correspondence and requests for materials Should be addressed to S.S., S.T., T.I., or M.H.

References

[1] N. M. Bérenger-Currias et al. A gastruloid model of the interaction between embryonic and extra-embryonic cell types:. *Journal of Tissue Engineering*, 13, jun 2022.

[2] M. Simunovic and A. H. Brivanlou. Embryoids, organoids and gastruloids: new approaches to understanding embryogenesis. *Development*, 144(6):976–985, mar 2017.

[3] S. Vianello and M. P. Lutolf. Understanding the Mechanobiology of Early Mammalian Development through Bioengineered Models. *Developmental Cell*, 48(6):751–763, mar 2019.

[4] M. N. Shahbazi, E. D. Siggia, and M. Zernicka-Goetz. Self-organization of stem cells into embryos: A window on early mammalian development. *Science*, 364(6444):948–951, jun 2019.

[5] N. Moris et al. An in vitro model of early anteroposterior organization during human development. *Nature 2020 582:7812*, 582(7812):410–415, jun 2020.

[6] S. C. Van Den Brink et al. Symmetry breaking, germ layer specification and axial organisation in aggregates of mouse embryonic stem cells. *Development (Cambridge)*, 141(22):4231–4242, nov 2014.

[7] D. A. Turner et al. Anteroposterior polarity and elongation in the absence of extraembryonic tissues and of spatially localised signalling in gastruloids: Mammalian embryonic organoids. *Development (Cambridge)*, 144(21):3894–3906, nov 2017.

[8] L. Beccari et al. Multi-axial self-organization properties of mouse embryonic stem cells into gastruloids. *Nature 2018 562:7726*, 562(7726):272–276, oct 2018.

[9] S. C. van den Brink et al. Single-cell and spatial transcriptomics reveal somitogenesis in gastruloids. *Nature*, 582(7812):405–409, feb 2020.

[10] A. Warmflash et al. A method to recapitulate early embryonic spatial patterning in human embryonic stem cells. *Nature Methods 2014 11:8*, 11(8):847–854, jun 2014.

[11] S. Pfister, K. A. Steiner, and P. P. Tam. Gene expression pattern and progression of embryogenesis in the immediate post-implantation period of mouse development. *Gene Expression Patterns*, 7(5):558–573, apr 2007.

[12] B. L. Hogan, A. R. Cooper, and M. Kurkinen. Incorporation into Reichert's membrane of laminin-like extracellular proteins synthesized by parietal endoderm cells of the mouse embryo. *Developmental Biology*, 80(2):289–300, dec 1980.

[13] G. S. Kwon, M. Viotti, and A. K. Hadjantonakis. The Endoderm of the Mouse Embryo Arises by Dynamic Widespread Intercalation of Embryonic and Extraembryonic Lineages. *Developmental Cell*, 15(4):509–520, oct 2008.

[14] S. Nowotschin et al. The emergent landscape of the mouse gut endoderm at single-cell resolution. *Nature*, 569(7756):361–367, apr 2019.

[15] M. Madabhushi and E. Lacy. Anterior visceral endoderm directs ventral morphogenesis and placement of head and heart via BMP2 expression. *Developmental Cell*, 21(5):907–919, nov 2011.

[16] I. Bedzhov et al. Development of the anterior-posterior axis is a self-organizing process in the absence of maternal cues in the mouse embryo. *Cell Research 2015 25:12*, 25(12):1368–1371, sep 2015.

[17] B. Sozen et al. Self-assembly of embryonic and two extra-embryonic stem cell types into gastrulating embryo-like structures. *Nature Cell Biology 2018 20:8*, 20(8):979–989, jul 2018.

[18] S. E. Harrison et al. Assembly of embryonic and extraembryonic stem cells to mimic embryogenesis in vitro. *Science*, 356(6334), apr 2017.

[19] N. C. Rivron et al. Blastocyst-like structures generated solely from stem cells. *Nature 2018 557:7703*, 557(7703):106–111, may 2018.

[20] B. Mathew et al. Mouse ICM Organoids Reveal Three-Dimensional Cell Fate Clustering. *Biophysical Journal*, 116(1):127–141, jan 2019.

[21] G. Amadei et al. Inducible Stem-Cell-Derived Embryos Capture Mouse Morphogenetic Events In Vitro. *Developmental Cell*, 56(3):366–382.e9, feb 2021.

[22] K. Hatta and M. Takeichi. Expression of N-cadherin adhesion molecules associated with early morphogenetic events in chick development. *Nature 1986 320:6061*, 320(6061):447–449, 1986.

[23] K. Punovuori et al. N-cadherin stabilises neural identity by dampening anti-neural signals. *Development (Cambridge)*, 146(21), nov 2019.

[24] S. Tsuda et al. FAK-mediated extracellular signals are essential for interkinetic nuclear migration and planar divisions in the neuroepithelium. *Journal of Cell Science*, 123(3):484–496, feb 2010.

[25] T. M. Jessell. Neuronal specification in the spinal cord: inductive signals and transcriptional codes. *Nature Reviews Genetics 2000 1:1*, 1(1):20–29, 2000.

[26] I. Bedzhov and M. Zernicka-Goetz. Self-organizing properties of mouse pluripotent cells initiate morphogenesis upon implantation. *Cell*, 156(5):1032–1044, feb 2014.

[27] A. Meinhardt et al. 3D reconstitution of the patterned neural tube from embryonic stem cells. *Stem Cell Reports*, 3(6):987–999, dec 2014.

[28] L. H. Pevny, S. Sockanathan, M. Placzek, and R. Lovell-Badge. A role for SOX1 in neural determination. *Development*, 125(10):1967–1978, may 1998.

[29] A. Sagner and J. Briscoe. Establishing neuronal diversity in the spinal cord: A time and a place. *Development (Cambridge)*, 146(22), nov 2019.

[30] I. Olivera-Martinez et al. Major transcriptome re-organisation and abrupt changes in signalling, cell cycle and chromatin regulation at neural differentiation in vivo. *Development*, 141(16):3266–3276, aug 2014.

[31] A. Di-Gregorio et al. BMP signalling inhibits premature neural differentiation in the mouse embryo. *Development*, 134(18):3359–3369, sep 2007.

[32] K. J. Lee, M. Mendelsohn, and T. M. Jessell. Neuronal patterning by BMPs: a requirement for GDF7 in the generation of a discrete class of commissural interneurons in the mouse spinal cord. *Genes and Development*, 12(21):3394, nov 1998.

[33] E. Dessaud, A. P. McMahon, and J. Briscoe. Pattern formation in the vertebrate neural tube: a sonic hedgehog morphogen-regulated transcriptional network. *Development*, 135(15):2489–2503, aug 2008.

[34] M. Maden. Retinoid signalling in the development of the central nervous system. *Nature Reviews Neuroscience 2002 3:11*, 3(11):843–853, 2002.

[35] J. Ericson et al. Pax6 Controls Progenitor Cell Identity and Neuronal Fate in Response to Graded Shh Signaling. *Cell*, 90(1):169–180, jul 1997.

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

[37] J. Delile et al. Single cell transcriptomics reveals spatial and temporal dynamics of gene expression in the developing mouse spinal cord. *Development*, 146(12):dev173807, jun 2019.

[38] B. Pijuan-Sala et al. A single-cell molecular map of mouse gastrulation and early organogenesis. *Nature*, 566(7745):490–495, feb 2019.

[39] J. V. Veenvliet et al. Mouse embryonic stem cells self-organize into trunk-like structures with neural tube and somites. *Science*, 370(6522), dec 2020.

[40] T. Haremaki et al. Self-organizing neuruloids model developmental aspects of Huntington's disease in the ectodermal compartment. *Nature Biotechnology 2019 37:10*, 37(10):1198–1208, sep 2019.

[41] A. Ranga et al. Neural tube morphogenesis in synthetic 3D microenvironments. *Proceedings of the National Academy of Sciences of the United States of America*, 113(44):E6831–E6839, nov 2016.

[42] K. Brown et al. A Comparative Analysis of Extra-Embryonic Endoderm Cell Lines. *PLOS ONE*, 5(8):e12016, 2010.

[43] T. Kunath et al. Imprinted X-inactivation in extra-embryonic endoderm cell lines from mouse blastocysts. *Development*, 132(7):1649–1661, apr 2005.

[44] D. H. Yang et al. Disabled-2 Is Essential for Endodermal Cell Positioning and Structure Formation during Mouse Embryogenesis. *Developmental Biology*, 251(1):27–44, nov 2002.

[45] A. Feijen, M. J. Goumans, and A. J. Van Den Eijnden-van Raaij. Expression of activin subunits, activin receptors and follistatin in postimplantation mouse embryos suggests specific developmental functions for different activins. *Development*, 120(12):3621–3637, dec 1994.

[46] J. Hou et al. A systematic screen for genes expressed in definitive endoderm by Serial Analysis of Gene Expression (SAGE). *BMC Developmental Biology*, 7(1):1–13, aug 2007.

[47] A. Wang et al. Nonmuscle myosin II isoform and domain specificity during early mouse development. *Proceedings of the National Academy of Sciences of the United States of America*, 107(33):14645–14650, aug 2010.

[48] P. Q. Thomas, A. Brown, and R. S. Beddington. Hex: a homeobox gene revealing peri-implantation asymmetry in the mouse embryo and an early transient marker of endothelial cell precursors. *Development*, 125(1):85–94, jan 1998.

[49] M. Efremova and S. A. Teichmann. Computational methods for single-cell omics across modalities. *Nature Methods*, 17(1):14–17, jan 2020.

[50] A. Paca et al. BMP signaling induces visceral endoderm differentiation of XEN cells and parietal endoderm. *Developmental Biology*, 361(1):90–102, jan 2012.

[51] C. Kimura-Yoshida et al. Canonical Wnt signaling and its antagonist regulate anterior-posterior axis polarization by guiding cell migration in mouse visceral endoderm. *Developmental Cell*, 9(5):639–650, nov 2005.

[52] R. M. Arkell and P. P. Tam. Initiating head development in mouse embryos: integrating signalling and transcriptional activity. *Open Biology*, 2(MARCH), 2012.

[53] J. C. Pelletier et al. (1-(4-(Naphthalen-2-yl)pyrimidin-2-yl)piperidin-4-yl)methanamine: A wingless β-catenin agonist that increases bone formation rate. *Journal of Medicinal Chemistry*, 52(22):6962–6965, nov 2009.

[54] D. ten Berge et al. Wnt Signaling Mediates Self-Organization and Axis Formation in Embryoid Bodies. *Cell Stem Cell*, 3(5):508–518, nov 2008.

[55] L. A. Lowery and H. Sive. Strategies of vertebrate neurulation and a re-evaluation of teleost neural tube formation. *Mechanisms of Development*, 121(10):1189–1197, oct 2004.

[56] G. C. Schoenwolf. Histological and ultrastructural studies of secondary neurulation in mouse embryos. *American Journal of Anatomy*, 169(4):361–376, apr 1984.

[57] J. F. Colas and G. C. Schoenwolf. Towards a cellular and molecular understanding of neurulation. *Developmental Dynamics*, 221(2):117–145, jun 2001.

[58] A. Dady et al. Junctional Neurulation: A Unique Developmental Program Shaping a Discrete Region of the Spinal Cord Highly Susceptible to Neural Tube Defects. *Journal of Neuroscience*, 34(39):13208–13221, sep 2014.

[59] P. Thomas and R. Beddington. Anterior primitive endoderm may be responsible for patterning the anterior neural plate in the mouse embryo. *Current Biology*, 6(11):1487–1496, nov 1996.

[60] P. F. Xu et al. Construction of a mammalian embryo model from stem cells organized by a morphogen signalling centre. *Nature Communications 2021 12:1*, 12(1):1–22, jun 2021.

[61] G. Rossi et al. Capturing Cardiogenesis in Gastruloids. *Cell Stem Cell*, 28(2):230–240.e6, feb 2021.

[62] Q. L. Ying et al. Conversion of embryonic stem cells into neuroectodermal precursors in adherent monoculture. *Nature Biotechnology 2003 21:2*, 21(2):183–186, jan 2003.

[63] J. Artus, J. J. Panthier, and A. K. Hadjantonakis. A role for PDGF signaling in expansion of the extra-embryonic endoderm lineage of the mouse blastocyst. *Development*, 137(20):3361–3372, oct 2010.

[64] J. F. Dekkers et al. High-resolution 3D imaging of fixed and cleared organoids. *Nature Protocols 2019 14:6*, 14(6):1756–1771, may 2019.

[65] S. Semrau et al. FuseFISH: Robust detection of transcribed gene fusions in single cells. *Cell Reports*, 6(1):18–23, jan 2014.

[66] M. Stoeckius et al. Simultaneous epitope and transcriptome measurement in single cells. *Nature Methods*, 14(9):865–868, aug 2017.

[67] A. T. Lun, K. Bach, and J. C. Marioni. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biology*, 17(1):1–14, apr 2016.

[68] L. Haghverdi, A. T. Lun, M. D. Morgan, and J. C. Marioni. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nature Biotechnology*, 36(5):421–427, 2018.

[69] A. Scialdone et al. Computational assignment of cell-cycle stage from single-cell transcriptome data. *Methods*, 85:54–61, sep 2015.

[70] S. C. van den Brink et al. Single-cell sequencing reveals dissociation-induced gene expression in tissue subpopulations. *Nature Methods*, 14(10):935–936, 2017.

[71] M. D. Robinson, D. McCarthy, Y. Chen, and G. K. Smyth. edgeR: differential expression analysis of digital gene expression data User's Guide. *Bioinformatics*, 26(October 2018):1–75, 2013.

[72] C. A. Schneider, W. S. Rasband, and K. W. Eliceiri. NIH Image to ImageJ: 25 years of image analysis. *Nature Methods 2012 9:7*, 9(7):671–675, jun 2012.

[73] S. Van Der Walt et al. Scikit-image: Image processing in python. *PeerJ*, 2014(1):e453, jun 2014.

6

Gene regulatory network inference with physics informed neural networks

One of the main goals of developmental biology is to reveal the gene regulatory networks (GRNs) underlying the robust differentiation of multipotent progenitors into precisely specified cell types. Most existing methods to infer GRNs from experimental data have limited predictive power as the inferred GRNs merely reflect gene expression similarity or correlation. Here, we demonstrate, how physics-informed neural networks (PINNs) can be used to infer the parameters of GRNs that provide mechanistic understanding of biological processes. Specifically we study GRNs that exhibit bifurcation behavior and can therefore model cell differentiation. We show that PINNs outperform regular feed-forward neural networks on the parameter inference task and analyze two relevant experimental scenarios: 1. a system with cell communication for which gene expression trajectories are available and 2. snapshot measurements of a cell population in which cell communication is absent. Our analysis will inform the design of future experiments to be analyzed with PINNs and provides a starting point to explore this powerful class of neural network models further.

Maria Mircea, Diego Garlaschelli and Stefan Semrau. Gene regulatory network inference with physics informed neural networks. Manuscript in preparation.

6.1 Introduction

Since the advent of single-cell molecular profiling, developmental biology has been inundated with high-dimensional data we are still learning to make sense of. Various machine learning methods have been used to find patterns in single-cell data, such as cell types or differentiation paths [1, 2]. Notwithstanding the great success of these methods, it remains difficult to infer mechanistic insights or quantitative, predictive models from single-cell data. Yet, one of the main goals of developmental biology is to understand the gene regulatory mechanisms underlying the robust differentiation of precisely defined cell types from multipotent progenitors [3] (see Chapter 1).

A common approach to the predictive mathematical modeling of differentiation uses the framework of dynamical systems theory [4–7]. In the context of differentiation, the dynamical system governs the abundance of gene products in the cell and stable attractor states are interpreted as cell types. Under certain conditions, the system can be represented by a quasi-potential [8]. This potential is the mathematical equivalent of Waddington's landscape [9], a seminal, qualitative model of differentiation in which the valleys in the landscape correspond to different cell types. In most models, the dynamical system has the structure of a network, in which the nodes are gene products, typically transcription factors, and the edges indicate interactions between them. Ideally, such a gene regulatory network (GRN) model should be able to predict the outcome of a differentiation process, given the initial cell state and external cues. Simulations using small GRNs with 2 – 5 nodes indeed exhibit bifurcations resembling actual differentiation processes [10–14].

As the parameter space grows quickly with the the size of a GRN, it can be tedious to find regimes with relevant behavior. A large body of work has therefore been devoted to inferring GRNs from measurements, typically transcriptomics or proteomics data sets. Most recently, single-cell data has been leveraged to that end [15]. Many inference methods use measures of similarity or correlation between genes and prior biological knowledge, most often about protein-protein binding affinities or the targets of transcription factors [16–19]. These methods can infer the existence of correlative or even causal relationships, especially if chromatin accessibility is taken into consideration [20]. However, they are typically unable to infer interaction strengths and are thereby lacking in predictive power. In fact, if only single-cell snapshot data is used and there is no prior biological knowledge, there are fundamental limits to GRN inference [21]. One should therefore 1. use time-resolved data that ideally contains information about the trajectories of individual cells and 2. constrain the inference problem with assumptions about the GRN. Seminal work using a Boolean network approach [22] or, more recently, catastrophe theory and approximate Bayesian computation [23], have successfully inferred predictive GRN models from time resolved data.

Another class of machine learning tools that have become extremely important in many fields are neural networks (NNs). These have been highly successful in pattern recognition and classification tasks [24] and are used extensively to interpret single-cell omics data [25, 26]. Naturally, NNs have also been used to infer GRNs from measurements [27, 28]. However, existing NN methods require GRNs obtained by other means as training data, which might limit the fidelity of the inferred GRNs. The optimal NN method would only use gene expression as training data while allowing us to implement prior knowledge or assumptions about the GRN to make the inference problem feasible. The recently developed

physics-informed neural networks (PINNs) [29–31] enable us to do just that. PINNs can solve a broad range of differential equations and also infer undetermined paramters. They have been applied successfully to various systems biology tasks [32–34].
In this chapter, we explore in how far PINNs can be used to infer GRNs. In our case, the differential equations to be solved by the PINN are defined by GRN topology and the mathematical expressions describing the interactions between the genes. Given gene expression measurements as training data, PINNs should be able to infer gene interactions. Fig. 6.1 shows a schematic of the inference procedure. Once the parameters have been learned, the dynamical system can then be used to make predictions. PINNs should thus allow us to gain mechanistic insights from measured expression data. As a proof of concept, we simulate data based on a GRN model recently introduced by Stanoev et al. [35]. At its core this GRN has two mutually inhibiting genes, u and v, which can be seen as the master transcription factors governing two alternative cell fates. This network motif has been studied intensively since it is one of the simplest motifs that exhibit bifurcations [7, 10, 14, 36]. Historically, this network motif was proposed decades ago to describe the mutual inhibition of gap genes [37]. The genetic toggle switch has been studied extensively in theory and demonstrated experimentally in organisms as simple as bacteria [38]. Kaneko et al. linked cell autonomous regulation with cell-cell interactions to show that cell division can drive differentiation and the formation of spatial patterns [39, 40]. The basic idea is the following: The idea is the following: Depending on the network parameters, a single stable attractor, interpreted as a multilineage primed (mlp) state can split into two stable attractors, which correspond to two different lineages. Such bifurcations successively create the large diversity of cell types in an adult organism, starting from the one-cell embryo [41]. In addition to the cell intrinsic mutual inhibition of the master transcription factors u and v, the model by Stanoev et al. also implements cell communication, which plays an essential role in development [42, 43] (see Chapter 3-5). In the Stanoev model, cells communicate via a diffusible signaling molecule s. This molecule is assumed to be activated by u and in turn inhibits u in both an autocrine and paracrine manner. The differential equations defining this GRN are shown in Fig. 6.2a. With this system, Stanoev et al. showed that population size can effectively serve as a control parameter that can bring the system from a homogeneous, progenitor state to a heterogeneous, differentiated state. This mechanism is an interesting alternative to the previously suggested noise-driven fate decisions [44, 45]. In certain parameter regimes, the GRN also creates regular, Turing-like spatial patterns.
In this chapter, we first explore the qualitative behavior of the GRN introduced by Stanoev et al. Subsequently we demonstrate that a naive feed-forward NN architecture and training based on simulations in a limited parameter regime are insufficient for robust GRN inference. We then explore how accurately PINNs can infer GRN parameters. Surprisingly, we find that it is not necessary to use all variables of the GRN for training. In other words, the measurement of a subset of genes across time can be sufficient for GRN inference. Lastly, we investigate a simpler system without cell communication. We study in how far GRN inference is still possible, if only snapshot data at discrete time points is available. This scenario is highly relevant as it describes the typical single-cell profiling experiment. This chapter thus provides a thorough assessment of PINNs for the purpose of GRN inference.

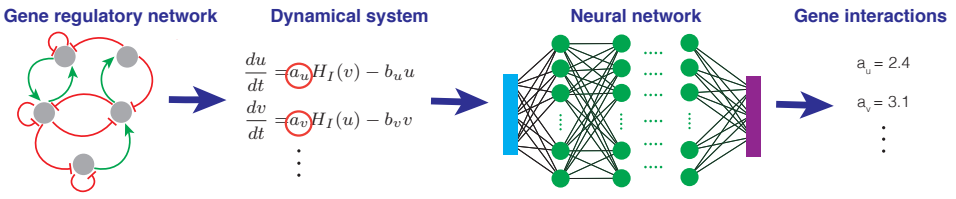


Figure 6.1: Gene regulatory network (GRN) inference with neural networks (NNs) First, a particular topology of the GRN is assumed. Together with the functional form of the interactions, the GRN topology defines a set of differential equations with undetermined parameter values. Next, a NN is trained on experimental or simulated time-series data. The parameters learned during training set the strength of interactions between genes. The fully determined dynamical system can then be used for predictions.

6.2 Results

6.2.1 Cell communication drives bifurcations in a GRN model of differentiation

We first set out to recapitulate the qualitative behavior of the GRN reported by Stanoev et al. to find interesting parameter regimes and establish a ground truth for GRN inference. We placed between 1 and 15 cells on a regular square grid and allowed cell communication between nearest neighbors, unless otherwise indicated by edges in 6.2b). The corresponding system of differential equations (shown in Fig. 6.2a) was solved by numerical integration. In the two-cell configuration, the parameter a_u, which sets the inhibition of u by v is a control parameter that can elicit a bifurcation (Fig. 6.2c). For low values of a_u there is one stable state. In this state, both cells have identical concentrations of u and v. Following Stanoev et al., we interpret a homogeneous state, where cells have identical, intermediate expression of both u and v as the mlp state. At a particular, critical value of a_u, two additional stable states appear through saddle node bifurcations. In one of these states, cell 1 has a high level of u but a low level of v, while cell 2 has a low level of u but a high level of v. In the other stable state, the expression patterns of cell 1 and 2 are reversed. These two states are thus considered differentiated states. Due to the mutual inhibition between u and v we will always find anti-correlation between the two genes, outside of the mlp state. Increasing a_u further, at a second critical point, the mlp state becomes unstable through a subcritical pitchfork bifurcation. Between a_u values of 2.33 and 2.69 the two differentiated states are the only stable states of the system. In summary, when controlled by a_u the system goes from the mlp state for low values of a_u via a small interval with three stable states (mlp, two differentiated states) to a wider interval with the two differentiated states as the only stable states. In other words, for differentiation to occur, a certain level of mutual inhibition between u and v is necessary. At even higher values of a_u additional bifurcations occur and the system returns to a single stable state. We will not explore this behavior at high a_u any further here, since it is unphysiological: In the real biological system, differentiated states are likely stabilized by other means (such as epigenetic marks), which would prevent the reversal to a single stable state.

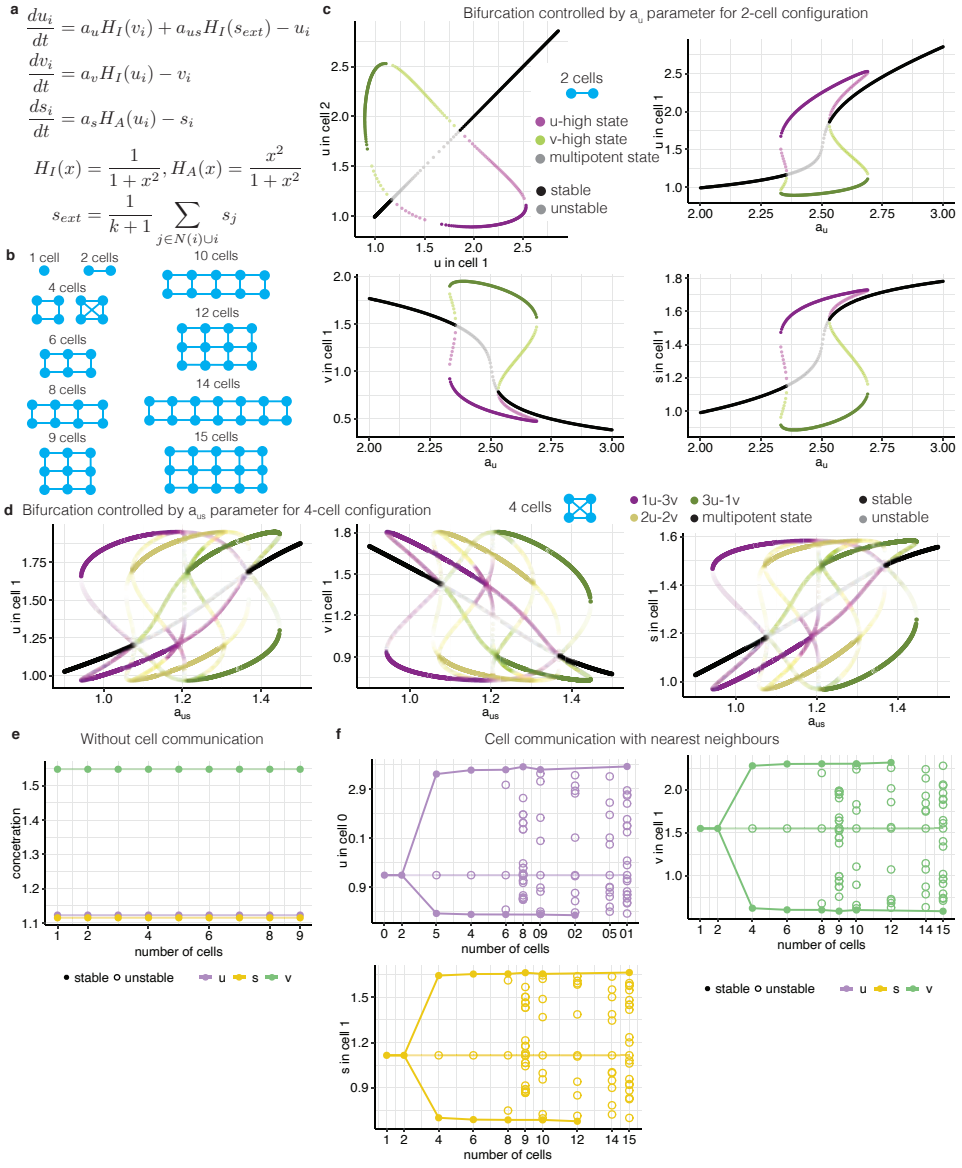


Figure 6.2: Cell communication drives bifurcations in a GRN model of differentiation. (Caption on the next page.)

For larger number of cells, the behavior of the system becomes more complex. We studied in detail a population of 4 cells where each cell was allowed to communicate with all other cells (Fig. 6.2d). We varied the strength of the intercellular communication, which is parameterized by a_{us}, and found that it can serve as a control parameter, similar to a_u. Starting from low values of a_{us} for which the mlp state is the only stable state, a sequence of bifurcations leads to the appearance and disappearance of several additional stable states. The first stable states to appear have one cell with high levels of u (low levels of v) and 3 cells with high levels of v (low levels of u). With increasing values of a_{us} the balance shifts to more cells with high u. At the extreme there is an interval of a_{us} with only two stable states in which 3 cells have a high level of u and one cell has a high level of v. Interestingly, the differentiated states are never homogeneous for the set of parameters used here.
Instead of modulating the strength of cell communication by tuning a_{us}, the system can also be driven out of the mlp state by increasing cell number, which would happen naturally through cell division. Without cell communication the system remains in the mlp state, irrespective of cell number (Fig. 6.2). With cell communication, a symmetry-breaking event occurs and two new stable states appear starting from 4 communicating cells, for the particular parameter set used (Fig. 6.2f). For 1 or 2 cells, only the mlp state is stable, and there are no other steady states. From 4 cells on, this state becomes unstable. As demonstrated previously by Stanoev et al. differentiation can occur simply after a certain number of cell divisions without changing the topology of the GRN or imposing a change of its parameters by external cues. Interestingly, more steady states appear in the system with increasing cell numbers. In the following, we will use simulations based on the 4-cell configuration with communication between all cells as ground truth training data for GRN inference with NNs.

Figure 6.2: Cell communication drives bifurcations in a GRN model of differentiation. (Figure on the previous page.) **a** System of differential equations corresponding to the GRN model by Stanoev et al. The mutual inhibition of the two master transcription factors u and v as well as the inhibition of u by the signaling molecule s are modeled with repressive Hill functions H_I. The cell autonomous activation of signalling molecule s by u is modeled with an activating Hill function H_A. i is the cell index. s_{ext} is the level of s averaged over cell i and its neighbors (typically nearest neighbors, unless otherwise indicated by the edges in panel b). The degradation rate for u, v and s is assumed to be identical, and time was rescaled with the inverse degradation rate, so that the rate does not appear explicitly in the equations. **b** Studied configurations of cells. Edges indicate cell communication. **c** Results for the 2-cell configuration. Several bifurcations are driven by the parameter a_u, which sets the strength of the inhibition of u by v. **d** Results for the 4-cell configuration with communication between all cells. Bifurcations are controlled by the parameter a_{us} which determines the strength of inter-cellular communication. Colors distinguish stable states with different ratios of u- and v-high cells. **e,f** Steady states (both stable and unstable) for the cell configurations shown in panel b without cell communication (panel **e**) or with cell communication (panel **f**). The following parameters were used: $a_u = 2.4,\ a_v = 3.5,\ a_s = 2, a_{us} = 1$.

6.2.2 Feedforward NN regression is unsuitable for GRN parameter inference

NNs have showed impressive performance in a large variety of supervised learning tasks [46]. The power of NNs usually relies on the existence of a large amount of high quality training data. Our first, naive idea was therefore to simulate expression trajectories, based on the dynamical system discussed above (see Fig. 6.2a), with randomly sampled parameters and use these trajectories to train a feedforward NN regression model (Fig. 6.3a). The input layer of this NN consists of the trajectories of u,v and s for n cells and k time points. Training samples are therefore vectors of length $3 \cdot n \cdot k$. The output nodes correspond to the 4 parameters of the GRN, a_u, a_v, a_s and a_{us}. Input and output layer were connected by several, fully-connected hidden layers.

To test this approach we used a configuration of 4 cells, with communication between all cells (as in Fig. 6.2d), and simulated 1000 trajectories with 25 time points for all variables in all cells. Parameters were sampled uniformly from intervals chosen such that trajectories from both the mlp as well as the differentiated regime were created (Fig. 6.3b). Initial states were also chosen randomly within reasonable intervals (see Methods). With this setup, the NN seemed to converge quickly and training was stopped after 1000 epochs (Fig. 6.3c). To create the test data we simulated 50 sets of 20 trajectories where the parameters were identical for each trajectory in a set, but the initial states were chosen randomly. Comparison of the parameter values used to simulate the trajectories (ground truth values) with the parameter values inferred by the NN model (Fig. 6.3d) revealed good accuracy of the model. Large systematic biases were absent for most parameter values. The random initial conditions contributed to the observed spread around the true values, which might limit the precision of the model.

At first glance, the simple feedforward architecture seemed to perform well. We next wanted to test, how important it is that the training data covers the different regimes of the dynamical system. When we trained the model with trajectories from the bistable, differentiated regime we observed that the model performed poorly for test trajectories outside of that regime (Fig. 6.3e). As the model is agnostic to the differential equations governing the dynamical system, it was unable to extrapolate beyond the parameter ranges it was trained on. In other words, if trained on a particular regime of the dynamical system, the NN model learns the behavior of that regime and does not generalize well. It would therefore be crucial to cover a large enough area of parameter space with the training data. Importantly, we were only able to identify the correct parameter ranges, because the system is relatively simple, allowing us to obtain a detailed understanding of its qualitative behavior (see Fig. 6.2). In an experimental setup, the relevant parameter ranges are usually unknown and it is typically hard to tune individual parameters. The naive feedforward NN regression model is therefore unsuitable for inferring GRN parameters from experimental data.

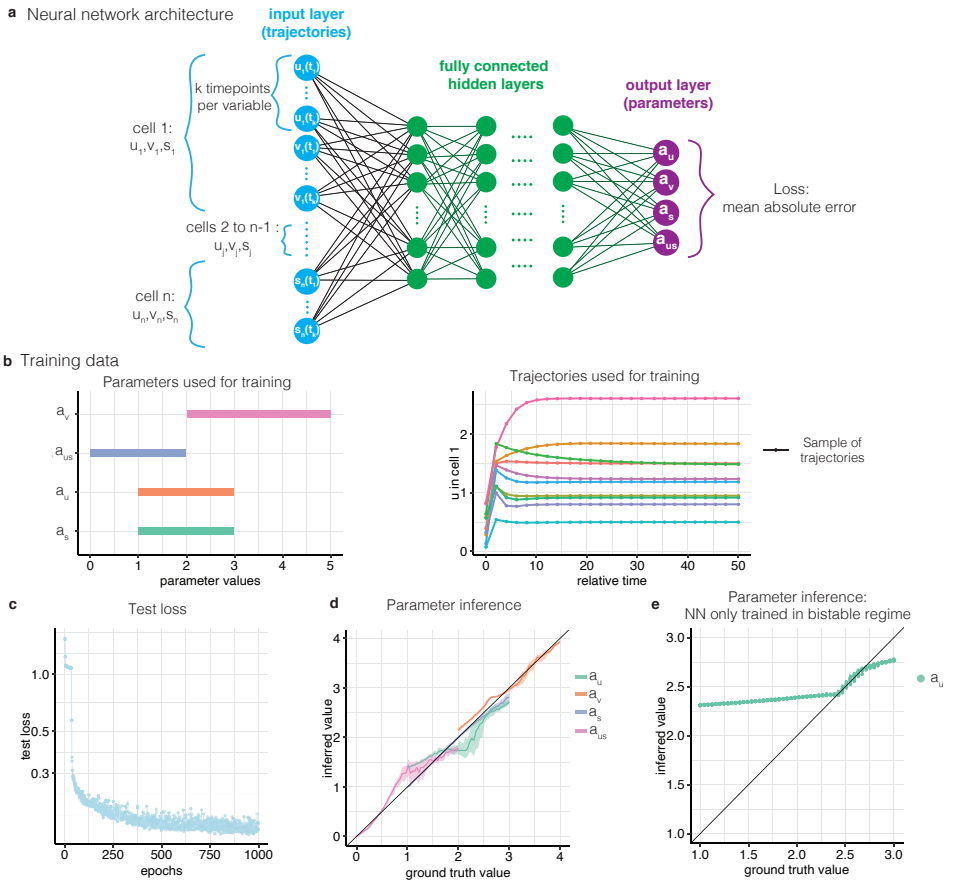


Figure 6.3: Feedforward NN regression is unsuitable for GRN parameter inference a Architecture of the feedforward NN. The mean absolute error was used for optimization. **b** Training data. Left: Parameter ranges used for creating simulated trajectories. Right: 10 example trajectories. **c** Test loss during training of the NN. **d,e** Ground truth parameter values (used for simulating the trajectories) versus parameter values inferred by the NN. In **d**, training trajectories covered the mlp as well as the bistable, differentiated regime. In **e** the training trajectories came exclusively from the bistable regime.

6.2.3 PHYSICS INFORMED NEURAL NETWORKS CAN INFER GRN PARAMETERS FROM PARTIAL AND NOISY DATA.

To ameliorate the reliance of NNs on large mounts of training data that represent all regimes of the dynamical system, we have to constrain the inference problem in a meaningful way. Ideally, the NN model should be aware of and respect the underlying differential equations. Physics-informed NNs (PINNs) leverage automated differentiation to solve a broad class of differential equations [29, 31]. The input layer of a PINN is composed of the independent variables of the differential equations (such as, for example, space and time for many applications in physics).

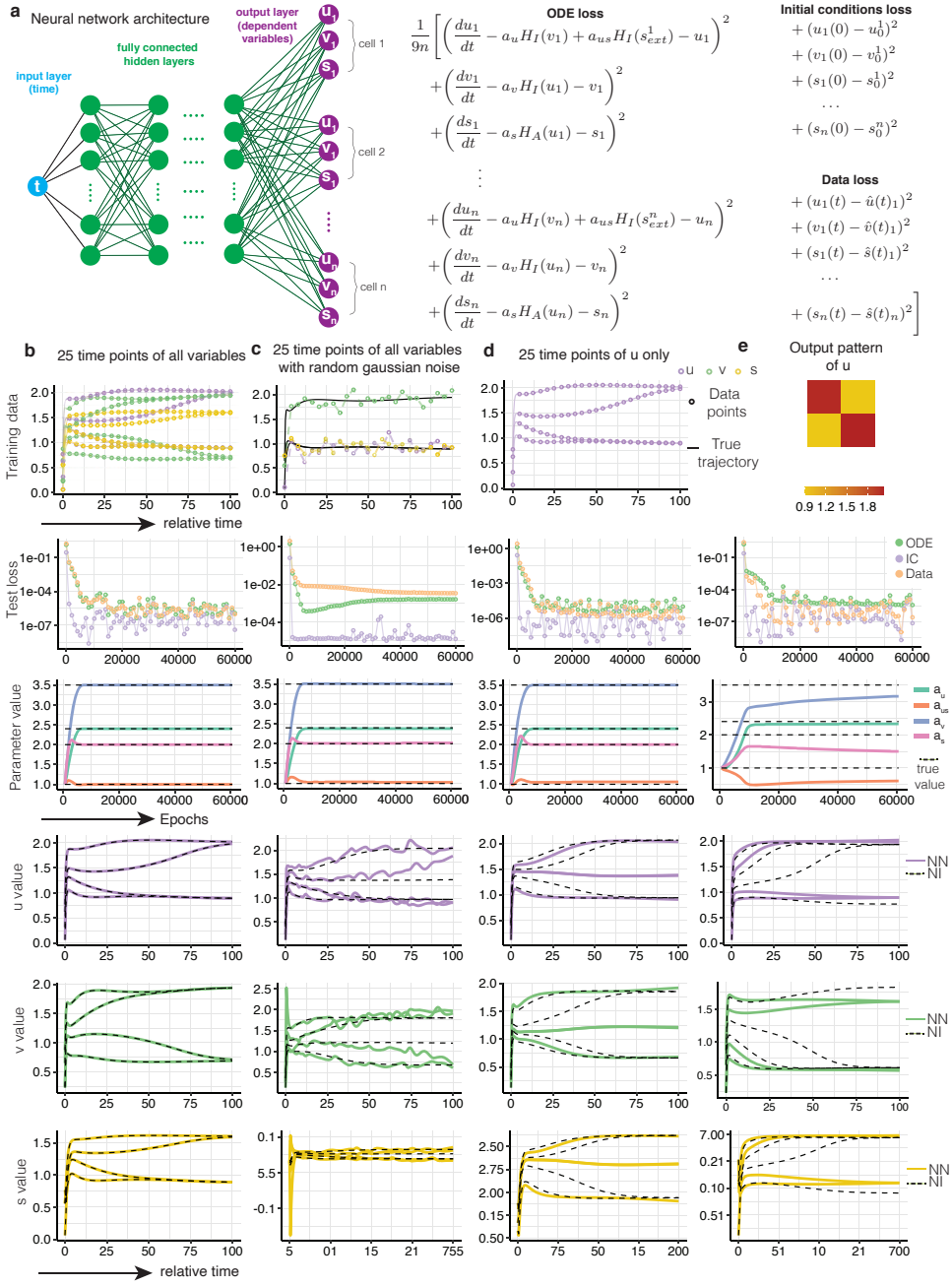


Figure 6.4: Physics informed neural network to infer gene regulatory networks. (Caption on the next page.)

The PINN is then trained such that the output layer approximates a solution to the differential equations for arbitrary points (in space and time) given as input. Fulfillment of the (ordinary) differential equations as well as initial and boundary conditions is ensured by appropriate loss terms, called ODE, IC and BC loss, respectively. During training, residual points on which the loss terms are evaluated are chosen randomly or in a way that adapts to the particular differential equations [30]. For this 'forward problem' of finding a solution of fully determined differential equations, no training data is necessary. PINNs can also infer undetermined parameters of the differential equations ('inverse problem'), which does require measured or simulated training data and a corresponding loss term ('data loss') that penalizes deviation of the solution from that data. The loss terms that ensure fulfillment of the differential equations strongly constrain the output space of the NN and thereby reduce the variance of the parameter inference. The issue of poor generalizability we observed with feedforward NN regression (Fig. 6.3e) should therefore be absent in PINNs.

To explore whether PINNs can successfully infer GRN parameters ('inverse problem'), we implemented the architecture shown in Fig. 6.4a with the DeepXDE package [30]. The input layer of the NN consists of only one node, which corresponds to time, and the output layer contains all dependent variables (u,v and s in all cells). As above, we used the 4-cell configuration with communication between all cells as a proof-of-concept. To generate training data we simulated trajectories with identical parameters but randomly drawn initial states. To explore the limitations of the PINN, we added noise and/or subset the data (Fig. 6.4b-e). Starting from noise-free trajectories with 25 time points per variable (Fig. 6.4b), we added Gaussian noise, since measurements are likely noisy due to biological and technical variability (Fig. 6.4c). We also explored training the PINN with a subset of variables as it is typically difficult to obtain measurements of all relevant dependent variables in experiments (Fig. 6.4d). Lastly, we studied training the model on the first and last time points only. For the set of parameters used here, the system has closely approached a stable steady state with two u-high and two v-high cells (Fig. 6.4e, top) by the last time point. This scenario is relevant for measurements with only one or a few time points or if the system is practically always in a stable steady state. In Fig. 6.4b-e we give examples of model behavior for different training scenarios. A systematic exploration and quantification of model performance is presented in Fig. 6.5.

Figure 6.4: Physics informed neural network to infer gene regulatory networks. (Figure on the previous page.) a Architecture of the PINN. The input to the network is time and the output consists of all dependent variables of the dynamical system. The PINN is optimized via a loss function that considers the differential equations (ODE loss), the initial conditions (IC loss) and training data (data loss). **b, c, d, e** The *first row* shows examples of training scenarios. A GRN with 4 cells that all communicate with each other was used. **b** Training on noise-free trajectories of all dependent variables with 25 fixed time points. **c** Training on trajectories shown on the left with added Gaussian noise. Only the trajectories in one cell are shown. **d** Training on noise-free trajectories of u only. **e** Only the first and last time point of the u trajectories in all 4 cells were used for training. The *second row* shows the resulting test losses. Colours indicate the different loss terms. *Row 4* shows the inferred parameters and *rows 3 - 6* show the approximated trajectories for the four scenarios. In the trajectory plots solid lines are trajectories approximated by the PINN and dashed lines are trajectories calculated by numerical integration using the inferred parameters.

When using complete trajectories for training, the PINN converges robustly after a few epochs and all three loss terms have similar convergence rates (Fig. 6.4b, second row). The inferred parameters are close to the ground truth parameters (Fig. 6.4b, third row) and the trajectories approximated by the PINN coincide with the trajectories calculated by numerical integration using the inferred parameters (Fig. 6.4b, rows 4-6). In contrast to feedforward NN regression (Fig. 6.3), which required many training samples, the PINN needs only one set of trajectories for accurate GRN inference. As to be expected, noise reduced the performance of the model, likely due to over-fitting, which can be seen for the inferred trajectories in Fig. 6.4c. Model performance was also compromised when only one dependent variable was used for training (Fig. 6.4d). Providing only the initial and final time point presented the biggest challenge for the PINN (Fig. 6.4e): The trajectories approximated by the PINN show large discrepancies with the trajectories calculated by numerical integration using the same (inferred) parameters. Hence, the trajectories approximated by the PINN are not a proper solution of the differential equations. Surprisingly, the inferred parameters were still roughly correct.

For a more systematic and quantitative assessment of model performance we tested 84 different conditions and considered: 1. the mean squared error between trajectories approximated by the PINN and trajectories found through the numerical integration using the inferred parameters, 2. the test loss and 3. the relative error of the inferred parameters (Fig. 6.5). For each condition we averaged over 10 runs with identical GRN parameters but randomly drawn initial states. First, we focused on the training scenarios that utilized all time points (Fig. 6.5b-d). As to be expected, increasing levels of noise reduced model performance (Fig. 6.5b).

In an attempt to mitigate over-fitting to the noisy training data, we introduced weights for the three loss terms and gave the ODE loss a 1000-times higher weight. Weighting improved trajectory approximation, but did not have a strong influence on parameter inference. Removing dependent variables from the training set had a strong and systematic effect on parameter inference and trajectory approximation was similarly affected when no weights were used (Fig. 6.5c). Weighting strongly improved trajectory approximation when only one dependent variable was used for training. Importantly, the relative errors of the parameter values depended on the set of variables used for training (Fig. 6.5d). For example, when only u and v were used, the parameters a_u and a_v were inferred more accurately than the parameters a_{us} and a_s. Conversely, when only u and s were used, a_{us} and a_s had a smaller error than the other parameters. Learning from only the first and last time point of the training data was overall a harder task for the PINN (Fig. 6.5e-g), but we observed similar trends for the dependence of model performance on noise (Fig. 6.5f) or the number of dependent variables used for training (Fig. 6.5g). Surprisingly, parameter inference from two time points was almost as accurate as when the whole trajectories were used for training, while the PINN's approximation of the trajectories was compromised. In summary, the PINN was able to infer GRN parameters even when only partial data was supplied for training.

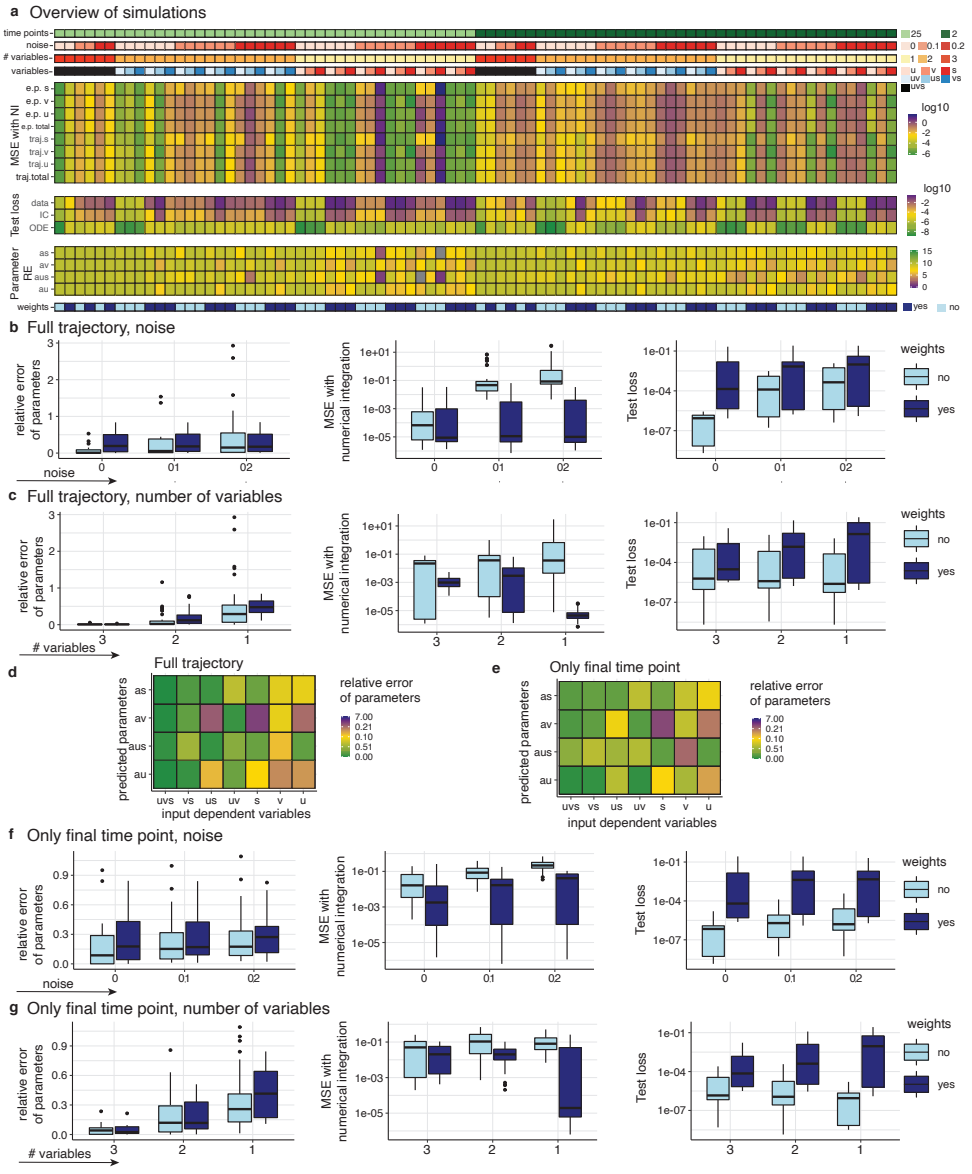


Figure 6.5: PINNs can infer GRN parameters from partial and noisy data. a Overview of all simulation results. The following parameters of the training data set were varied: the number of time points (either 25 per variable or only the first and last time point), the amount of Gaussian noise (standard deviation of 0, 0.1 or 0.2 with a mean of 0) and the dependent variables ($[u,v,s],[u,v],[u,s],[v,s],[u],[v],[s]$). When weights were given to the loss terms the ODE loss was weighted with a factor 1000. e.p.: end point, traj: full trajectory. **b-d** Dependence of PINN performance on noise level **b**, number **c** and identity **d** of dependent variables used for training when the complete trajectories were used. **e-g** Same performance comparisons as in **b-d** but the PINN was trained only on the initial and final time point of the trajectories.

6.2.4 PINNs can infer GRN parameters from snapshot data in the absence of cell communication.

Most high-throughput single-cell profiling assays are destructive, which prevents the measurement of single-cell trajectories. These assays therefore only provide "snapshots" of the system dynamics. Additionally, in conventional single-cell omics experiments, any information about the spatial arrangement of the cells is lost. Therefore we wanted to explore, how a PINN would perform when trained with snapshot data that lacks spatial resolution. As any parameter related to cell communication is unlikely to be estimated well in such a scenario, we considered a simpler dynamical system of two mutually inhibiting genes, u and v, without cell communication [10] (Fig. 6.6a). The parameters I_u and I_v modulate the inhibition of u or v, respectively, by the other gene. For a particular set of parameters, the dynamical system is bistable and leads to the cell autonomous differentiation into either a u-high or a v-high state. Cells will be attracted to one of these stable steady states depending on their initial state. As trajectories cannot be obtained from destructive snapshot measurements it is more natural to model the system at the level of a population of cells and consider a bivariate probability density of u and v. In the absence of noise, the dynamics of the probability density is completely determined by the conservation of probability (Fig. 6.6a), which is therefore the differential equation that must be fulfilled by the PINN.

Fig. 6.6b shows the architecture of the PINN together with the loss terms. The input layer is now composed of three nodes, corresponding to time, u and v. The only output node is the probability density at the point $[u,v,t]$ given as input. As before, the loss considers the governing differential equations, initial conditions and the training data. For training we simulated 1000 trajectories with initial values of u and v randomly drawn from a bivariate normal distribution centered around 1.5. As mentioned above, the parameters of the GRN and the distribution of the initial states are chosen such that trajectories tend to one of two stable steady states (Fig. 6.6c). The simulated trajectory positions were binned for each time point and the probability densities were approximated by the relative frequencies (Fig. 6.6d). As intended, the initial probability density, a normal distribution, developed into a bimodal distribution, reflecting the existence of two stable steady states with anti-correlated expression of u and v. Using these simulations we trained the PINN, leaving the parameters I_u and I_v undetermined. The PINN converged quickly (Fig. 6.6e) and the parameters were inferred with reasonable precision (Fig. 6.6f). The probability densities approximated by the PINN show qualitatively the same dynamics as the densities used for training (Fig. 6.6g). However, the PINN approximation of the probability density had a few negative values and was not always properly normalized, which could likely be improved by additional constraints. All in all, the PINN was able to infer GRN parameters from snapshot data for a model without cell communication.

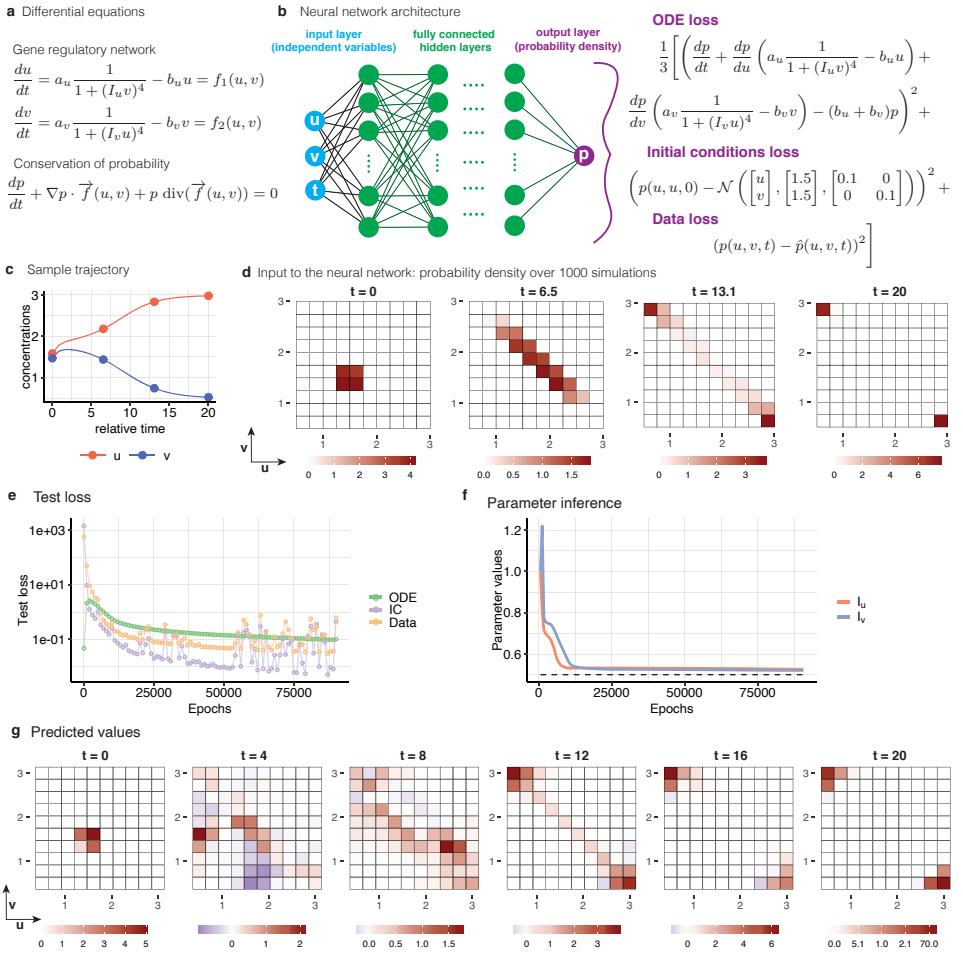


Figure 6.6: PINNs can infer GRN parameters from snapshot data in the absence of cell communication. **a** Top: Differential equations describing two, mutually inhibiting genes u and v. Bottom: Differential equation describing the conservation of probability. **b** Architecture of the PINN. The input nodes correspond to u , v and time t. The output is the probability density. The PINN is optimized via a loss function that considers the differential equations, the initial conditions and the training data. **c** Example trajectory for the differential equations shown in panel a. **d** Probability density of the values u and v at four different time points, generated by repeated simulation of trajectories as in c. This simulation was used as training data. **e** Test loss of PINN training using the data shown in d. **f** Convergence of the parameters I_u and I_v during PINN training. **g** Probability densities approximated by the PINN at 6 different time points.

6.3 Discussion

The inference of GRNs from noisy and usually incomplete measurements is a long-standing challenge which inspired the development of many different approaches. In this chapter, we studied the performance of PINNs in this context.

PINNs are general tools that approximate the solutions to a broad class of differential

equations. To apply them to GRN inference requires expressing the dynamical system defined by the GRN as a set of differential equations. To that end, specific expressions that model the gene interactions have to be assumed. Here, we used Hill functions with fixed Hill coefficients for both activation and inhibition. We selected a subset of relevant parameters to be learned by the PINN, but it might be interesting to leave more parameters undetermined, especially the Hill coefficients. Most importantly, we used the same network topology for simulation and training the PINN: The same genes were connected with the same type of interaction (either activating or inhibiting). In principle, one could base the training on a fully connected network and model each interaction as the sum of activating and inhibiting expressions. Such a setup would leave network topology unconstrained and it would be interesting to explore, if a PINN could infer it from the data. In this context, it might be useful to add a regularization term to the loss function such that only the strongest interactions are selected and the inferred GRN is sparse.

As a proof of concept we studied a minimal GRN with two mutually inhibiting genes. Such a GRN exhibits a bifurcation that models the differentiation of a multipotent progenitor into one of several differentiated cell types. While several pairs of master transcription factors that govern such bifurcations have been identified in experiments, real GRNs contain other relevant genes. It would therefore be useful to determine, how much and what kind of experimental data would be necessary to infer a much larger GRN with a PINN.

In this chapter we first studied an experimental scenario in which the trajectories of individual cells are measured and demonstrated that a PINN outperforms a simple feedforward NN regression model. While the feedforward NN requires many training samples that must cover all dynamical regimes of the GRN, the PINN efficiently infers GRN parameters from a single sample, at the cost of making assumptions or using prior knowledge about the GRN. The PINN was able to infer parameters even if only a subset of dependent variables was used for training. The relative errors of the inferred parameter values depended on the identity of the used variables, which is important to keep in mind for optimal experiment design. As the training of the PINN becomes computationally more costly with increasing number of cells, it would be interesting to explore, whether using measurements of a subset of cells for PINN training is sufficient for GRN inference. Possibly, that would require a kind of mean field approximation of the cells that are not used for training. Surprisingly, parameter inference was still possible when we only used the initial and final time point of the trajectories for training. However, in this case, the approximate trajectories provided by the PINN did not fulfill the differential equations as they deviated from trajectories calculated by numerical integration using the inferred parameters. It seems that the loss terms related to additional time points support the ODE loss in ensuring fulfillment of the differential equation. Inferring the GRN from the final state, which is in this case essentially a spatial pattern of differentiated cell types, would be very useful not only to study morphogenesis but also to inform synthetic biology applications. Recently, NNs were used to implement a cellular automaton that models morphogenesis [47]. Impressively, it was shown that providing a desired spatial pattern as training data is sufficient to train the NN such that the automaton robustly develops into that spatial pattern. While this is certainly a conceptually important feat, cellular automata are only rough approximations of real biological dynamics and it would be preferable to achieve a similar performance with GRNs. We speculate that constraining the final state to be a globally stable steady

state, potentially by using a Lyapunov function [48], might help in that respect.
In the final section of this chapter we studied a scenario in which only snapshot data of cell populations are available, which is the case for single-cell RNA-sequencing experiments [49]. As spatial context is not available in this scenario, we described the system at the population level, with probability densities of gene abundances. We showed that a PINN was able to infer a simple GRN without cell communication. While parameter inference was successful, the probability density approximated by the PINN sometimes had negative values and was not always properly normalized. Positive values could be enforced by adding an additional term to the loss function or using a rectified linear unit (ReLU) as activation function in the output layer. To ensure proper normalization one could reformulate the differential equations such that the dependent variable is a normalized function [50]. Another option is to discretize the domains of u and v and add a constraint on the sum of the resulting discrete densities via an additional loss term [51]. As there was no noise in the dynamical system used here, the relevant differential equation was simply given by the conservation of probability. In the presence of noise, one would have to use the Fokker-Planck equation which the PINN should be able to fulfill without problems.
Next to the two experimental scenarios studied in this chapter, there are a few others, which are currently very popular and would therefore be worthwhile to explore in future work. Many snapshot measurements of highly dynamical systems, such as developing tissues, in fact contain dynamical information: Pseudotime methods have been used to establish developmental progression from snapshot data [52, 53]. It would be interesting to work out, how pseudotime information could be leveraged for GRN inference with PINNs. Spatially resolved omics modalities are also being used extensively at the moment. To infer GRNs with cell communication from such data will be an interesting challenge.
In summary, we have established that PINNs can be used for the accurate inference of GRNs. PINNs thus present an exciting, new way to obtain mechanistic insights from single-cell data. We hope that our work will stimulate colleagues from mathematics, physics and biology to collaborate on the many fascinating problems presented by single-cell developmental biology.

6.4 Materials and Methods

Python (V 3.9) was used for all computations. NNs were implemented with *tensorflow* (V.2.6). All figures were generated with R (V 4.1).

6.4.1 Inference of a GRN with cell communication from trajectories

Differential equations

The set of coupled differential equations in Fig. 6.2a were adopted from [35]. For each cell, there are two mutually inhibiting genes, u and v. Additionally, a signalling molecule s that is stimulated by u, inhibits u in an autocrine and paracrine way. In the model considered here, all signalling molecules of neighboring cells contribute equally. The dynamics of each cell $i \in \{1,...,N\}$ is governed by this system of differential equations

$$\begin{aligned} \frac{du_i}{dt} &= a_u H_I(v_i) + a_{us} H_I(s_{ext}) - u_i \\ \frac{dv_i}{dt} &= a_v H_I(u_i) - v_i \\ \frac{ds_i}{dt} &= a_s H_A(u_i) - s_i, \end{aligned} \tag{6.1}$$

where s_{ext} denotes signalling molecule abundance averaged over the neighbors: $s_{ext} = \frac{1}{k+1}\sum_{j\in N(i)\cup i} s_j$ with k the number of neighbours. H_I and H_A are the inhibiting and activating Hill functions, respectively, defined here with a fixed Hill coefficient of 2:

$$H_I(x) = \frac{1}{1+x^2}, \quad H_A(x) = \frac{x^2}{1+x^2} \tag{6.2}$$

This results in a set of coupled differential equations with $3N$ equations and $3N$ variables. The parameters a_u, a_v, a_s and a_{us} are the same for each cell. Time was rescaled by an inverse degradation rate which was assumed to be identical for all genes. An additional parameter λ, used by Stanoev et al. to control the speed of the temporal evolution, was set to 1.

We distributed the cells on a regular grid, generated with *python-igraph* (V 0.9), and cell communication was typically restricted to nearest neighbors, unless otherwise indicated by edges in the graph.

Steady states

Steady states were found with a multi-start optimization algorithm using *scipy*. The stability was calculated based on the Jacobian matrix evaluated around the steady state. Bifurcation analysis was performed by repeating the optimization algorithm for each value of the control parameter (either a_u or a_{us}). For the 2-cell and 4-cell configurations, 500 initial values were chosen uniformly in the interval $[0,3]$ for each of the $3N$ variables. In the study of the effect of cell number (Fig. 6.2e,f) 100 initial values were used.

Steady states were either mlp, where each cell was an mlp, or differentiated, consisting of u-high and v-high cells. The mlp steady state was identified based on comparison with gene expression in the 1-cell configuration with the same parameters. If the relative error

(comparing to the 1-cell mlp state) for all dependent variables was below 0.1, a steady state was annotated as mlp. To identify u-high and v-high cells in a differentiated steady state, the ratio of u and v in individual cells was compared to the ratio of u and v in the mlp steady state for the same parameters. If the ratio of u and v in a cell was larger than in the mlp steady state, the cell was consider to be a 'u-high' cell, if the ratio was smaller than in the mlp steady state, the cell was considered to be 'v-high'.

Data simulation

Data points were generated by numerical integration (NI) of the differential equations for a given set of parameters. We used scipy (V 1.7) with the explicit Runge-Kutta method for integration. The initial conditions were chosen randomly from a uniform distribution in the interval $[0,1]$. Numerical integration was performed on 100 equidistant time points in the interval $[0,T]$. When the entire trajectory was used for training, a subset of 25 equidistant time points was used. When only the initial and final state were used as input, we considered the values at $t = 0$ and $t = T$.

6.4.2 Feed-forward neural network

The feed-forward NN architecture is depicted in Fig. 6.3a. The NN input takes 25 time points for each variable. The GRN consisted of 4 cells, which all communicate with each other, resulting in 12 independent variables in total. Thus, the input layer of the NN consists of 300 nodes. We found that a 20% dropout rate in the input layer prevented over-fitting during training. The NN has 4 fully-connected hidden layers with 32 nodes each and Rectified Linear Unit (ReLu) activation functions. The output layer has 4 nodes for the results shown in Fig. 6.3d, corresponding to the parameters a_u,a_v,a_s and a_{us}, and one node for the result shown in Fig. 6.3e, corresponding to the parameter a_u. The loss function was defined using the mean absolute error, which slightly outperformed the mean squared error. For optimization the Adam optimizer was used.
For training and testing, trajectories were generated by numerical integration using randomly drawn initial states from the interval $[0,1]$. For Fig. 6.3d, 1000 training trajectories were generated with parameters chosen randomly from uniform distributions over the following intervals: $a_u \in [1,3]$, $a_v \in [2,5]$, $a_s \in [1,3]$ and $a_{us} \in [0,2]$. To generate trajectories for testing in Fig. 6.3d, parameters were chosen from the same parameter ranges. For each run, 3 out of the 4 parameters were kept fixed ($a_u = 2.4$, $a_v = 3.5, a_u = 2, a_{us} = 1$), while the remaining parameter was varied. For each parameter, 50 values were taken from the above intervals equidistantly and each set of parameters was used 20 times with different initial values for the numerical integration.
For Fig. 6.3e, 1000 training trajectories were generated with three parameters fixed ($a_v = 3.5, a_u = 2, a_{us} = 1$) and a_u sampled from the interval $[2.4, 2.7]$, where the dynamical system has multiple steady states in the 4-cell configuration. The trajectories used for testing were generated with the same fixed parameters, but 50 values of a_u were equidistantly drawn from the interval $[1,3]$. Each parameter value was used 20 times with different initial conditions for the numerical integration.

6.4.3 Physics Informed Neural Network

All PINNs were implemented using *DeepXDE* (V 0.14) [30]. The network architecture is shown in Fig. 6.4a. The input layer consists of only one node, which corresponds to the time t. The hidden layers are 4 fully connected layers with 40 nodes each. The output layer has $3N$ nodes, where N is the number of cells, corresponding to the $3N$ dependent variables of the dynamical system. In Fig. 6.5, a configuration of 4 cells that all communicate with each other was implemented, which results in an output layer of size 12. *tanh* was used as the activation function.

The loss function consists of three terms. The first penalizes deviation from the differential equations:

$$\left(\frac{du_1}{dt} - a_u H_I(v_1) + a_{us} H_I(s^1_{ext}) - u_1\right)^2 + \left(\frac{dv_1}{dt} - a_v H_I(u_1) - v_1\right)^2 + \left(\frac{ds_1}{dt} - a_s H_A(u_1) - s_1\right)^2 +$$

$$\ldots$$

$$\left(\frac{du_n}{dt} - a_u H_I(v_n) + a_{us} H_I(s^n_{ext}) - u_n\right)^2 + \left(\frac{dv_n}{dt} - a_v H_I(u_n) - v_n\right)^2 + \left(\frac{ds_n}{dt} - a_s H_A(u_n) - s_n\right)^2 \tag{6.3}$$

The second loss term considers the initial conditions, using the mean squared error. The last term in the loss function includes the training data, also using the mean squared error. For time points within the trajectory we defined boundary conditions using the *PointSet* object from the *DeepXDE* package. This object allows the user to supply measured data at any point in the input domain and the coresponding loss term will be added to the loss function. In order to specify the final time point in the *DeepXDE* framework, we used the Dirichlet boundary condition object. In this way, we could set values for the time domain at the initial point $t = 0$ and the final point $t = T$.

To create the results shown in Fig. 6.5, PINNs were trained using 84 different scenarios, 10 times each, with randomly selected initial conditions for the differential equations. Results were averaged over the 10 simulations. The following properties of the training data and PINN were varied:

(a) number of time points used for training: 25 (full trajectory) or 2 (initial and final state)
(b) noise level: no noise; Gaussian with mean 0, standard deviation 0.1; Gaussian with mean 0, standard deviation 0.2. Negative values resulting from addition of noise addition were set to 0.
(c) number of dependent variables used for training: 1,2 or 3
(d) identity of dependent variables used for training: $[u, v, s], [u, v], [u, s], [v, s], u, v, s$
(d) weights: no weights or ODE loss weighted with factor 1000

Neural network validation

We used three measures to quantify the performance of the PINN. First, we computed the relative error between the inferred parameters and the true parameters. Second, we calculated the mean squared error between the PINN approximation of the trajectories and trajectories obtained by numerical integration using parameters and initial conditions inferred by the PINN. Lastly, we considered the test loss.

6.4.4 Inference of a GRN without cell communication from snapshot data

Differential equations

Two mutually inhibiting genes, u and v, were modeled with expressions used to describe lateral inhibition [10]. The differential equations are given by:

$$\begin{aligned}\frac{du}{dt} &= a_u \frac{1}{1+(I_u v)^4} - b_u u = f_1(u,v)\\ \frac{dv}{dt} &= a_v \frac{1}{1+(I_v u)^4} - b_v v = f_2(u,v)\end{aligned} \tag{6.4}$$

We used a set of parameters for which this dynamical system is bistable: $a_u = a_v = 1.5$, $I_u = I_v = 0.5$ and $b_u = b_v = 0.5$. To model inhibition we used an inhibiting Hill function with Hill coefficient 4.

We describe the system at the population level with the joint probability density for the abundance of u and v. The time evolution of this probability density is governed by the conservation of probability:

$$\frac{\partial p}{\partial t} + \nabla p \cdot f(u,v) + p \,\mathrm{div}(f(u,v)) = 0, \tag{6.5}$$

where $f(u,v)$ is defined as

$$f(u,v) = \begin{bmatrix} f_1(u,v) \\ f_2(u,v) \end{bmatrix} \tag{6.6}$$

Plugging in $f(u,v)$ gives the following differential equations:

$$\frac{dp}{dt} + \frac{dp}{du}\left(a_u \frac{1}{1+(I_u v)^4} - b_u u\right) + \frac{dp}{dv}\left(a_v \frac{1}{1+(I_v u)^4} - b_v v\right) - (b_u + b_v)p = 0 \tag{6.7}$$

Data simulation

Training data was created by numerical integration as described in section 6.4.1. The initial conditions were sampled from a normal distribution with mean 1.5 and standard deviation 0.1. Trajectories with 50 time points in the interval $[0,20]$ were obtained. 4 equidistant time points were used for further computations. In order to create a probability density function, 1000 trajectories were generated. The probability density was then estimated by the relative frequencies calculated for 100 bins with u and v in the interval $[0.5,3]$, for each time point. The values for u and v in each bin were taken as the bin's midpoint.

Physics Informed Neural Network

A PINN with the architecture shown in Fig. 6.6a was implemented with *DeepXDE*. The PINN takes three independent variables as input (u, v and t). The hidden layers are 4 fully connected layers with 40 nodes each and *tanh* activation functions. The output layer has only one node, which corresponds to the probability density $p(u,v,t)$. Initial conditions were defined at $t = 0$ as a bivariate normal distribution with mean $\begin{bmatrix}1.5\\1.5\end{bmatrix}$ and

variance $\begin{bmatrix} 0.1 & 0 \\ 0 & 0.1 \end{bmatrix}$. Simulated data was added, as described previously, with the *PointSet* object in *DeepXDE*. As before, the loss function was composed of 3 terms that consider the differential equations, the initial conditions and the training data, using the mean squared error. For training, 500 points were chosen randomly from the joint domain of u and v to define the initial condition loss, 100 points from the simulated data were chosen for the data loss and 1000 points from the joint domain of u, v and t were chosen randomly to define the differential equation loss. We weighted the data loss with a factor of 1000. The parameters a_u, a_v, b_u and b_v were fixed and the parameters I_u and I_v were inferred by the PINN.
To visualize the PINN approximation of the probability density, 6 time points $(0, 4, 8, 12, 16, 20)$ were selected and the approximated density was plotted on the u, v-grid that was used during training.

Funding M. M. and S.S. were supported by the Netherlands Organisation for Scientific Research (NWO/OCW,www.nwo.nl), as part of the Frontiers of Nanoscience (NanoFront) program. The computational work was carried out on the Dutch national e-infrastructure with the support of SURF Cooperative.

Disclosure of potential conflict of interest The authors indicated no potential conflicts of interest.

References

[1] L. Yu, Y. Cao, J. Y. Yang, and P. Yang. Benchmarking clustering algorithms on estimating the number of cell types from single-cell RNA-sequencing data. *Genome Biology*, 23(1):1–21, dec 2022.

[2] W. Saelens, R. Cannoodt, H. Todorov, and Y. Saeys. A comparison of single-cell trajectory inference methods. *Nature Biotechnology*, 37(5):547–554, may 2019.

[3] M. Mircea and S. Semrau. How a cell decides its own fate: a single-cell view of molecular mechanisms and dynamics of cell-type specification. *Biochemical Society Transactions*, dec 2021.

[4] A. Guillemin and M. P. H. Stumpf. Non-equilibrium statistical physics, transitory epigenetic landscapes, and cell fate decision dynamics. *Mathematical Biosciences and Engineering*, 17(6):7916–7930, nov 2020.

[5] P. Greulich, R. Smith, and B. D. MacArthur. The physics of cell fate. In *Phenotypic Switching*, pages 189–206. Elsevier, jan 2020.

[6] L. Xu and J. Wang. Quantifying Waddington landscapes, paths, and kinetics of cell fate decision making of differentiation/development. In *Phenotypic Switching*, pages 157–187. Elsevier, jan 2020.

[7] S. Huang. The molecular and mathematical basis of Waddington's epigenetic landscape: A framework for post-Darwinian biology? *BioEssays*, 34(2):149–157, feb 2012.

[8] J. X. Zhou, D. S. Aliyu, E. Aurell, and S. Huang. Quasi-potential landscape in complex multi-stable systems. *Journal of The Royal Society Interface*, 9(77):3539–3553, dec 2012.

[9] C. Waddington. *The strategy of the genes : a discussion of some aspects of theoretical biology / by C.H. Waddington. | Wellcome Collection.* Allen and Unwin, London, 1975.

[10] J. E. Ferrell. Bistability, Bifurcations, and Waddington's Epigenetic Landscape. *Current Biology*, 22(11):R458–R466, jun 2012.

[11] N. Saiz et al. Growth factor-mediated coupling between lineage size and cell fate choice underlies robustness of mammalian development. *eLife*, 9:1–38, jul 2020.

[12] D. Raina et al. Cell-cell communication through FGF4 generates and maintains robust proportions of differentiated cell fates in embryonic stem cells, feb 2020.

[13] M. K. Franke and A. L. Maclean. A single-cell resolved cell-cell communication model explains lineage commitment in hematopoiesis. *bioRxiv*, page 2021.03.31.437948, apr 2021.

[14] S. Huang, Y. P. Guo, G. May, and T. Enver. Bifurcation dynamics in lineage-commitment in bipotent progenitor cells. *Developmental Biology*, 305(2):695–713, may 2007.

[15] A. Pratapa et al. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nature Methods*, 17(2):147–154, feb 2020.

[16] V. A. Huynh-Thu, A. Irrthum, L. Wehenkel, and P. Geurts. Inferring regulatory networks from expression data using tree-based methods. *PLoS ONE*, 5(9):12776, 2010.

[17] S. Aibar et al. SCENIC: Single-cell regulatory network inference and clustering. *Nature Methods*, 14(11):1083–1086, oct 2017.

[18] N. Papili Gao, S. M. Ud-Dean, O. Gandrillon, and R. Gunawan. SINCERITIES: Inferring gene regulatory networks from time-stamped single cell transcriptional expression profiles. *Bioinformatics*, 34(2):258–266, jan 2018.

[19] X. Qiu et al. Inferring Causal Gene Regulatory Networks from Coupled Single-Cell Expression Dynamics Using Scribe. *Cell Systems*, 10(3):265–274.e11, mar 2020.

[20] K. Kamimoto, C. M. Hoffmann, and S. A. Morris. CellOracle: Dissecting cell identity via network inference and in silico gene perturbation, feb 2020.

[21] C. Weinreb et al. Fundamental limits on dynamic inference from single-cell snapshots. *Proceedings of the National Academy of Sciences of the United States of America*, 115(10):E2467–E2476, mar 2018.

[22] S. J. Dunn et al. Defining an essential transcription factor program for naïve pluripotency. *Science*, 344(6188):1156–1160, jun 2014.

[23] M. Sáez et al. Statistically derived geometrical landscapes capture principles of decision-making dynamics during cell fate transitions. *Cell Systems*, sep 2021.

[24] O. I. Abiodun et al. Comprehensive Review of Artificial Neural Network Applications to Pattern Recognition. *IEEE Access*, 7:158820–158846, 2019.

[25] C. Angermueller, T. Pärnamaa, L. Parts, and O. Stegle. Deep learning for computational biology. *Molecular Systems Biology*, 12(7):878, jul 2016.

[26] G. Eraslan et al. Single-cell RNA-seq denoising using a deep count autoencoder. *Nature Communications*, 10(1), 2019.

[27] Y. Yang, Q. Fang, and H. B. Shen. Predicting gene regulatory interactions based on spatial gene expression data and deep learning. *PLOS Computational Biology*, 15(9):e1007324, 2019.

[28] J. Wang et al. Inductive inference of gene regulatory network using supervised and semi-supervised graph neural networks. *Computational and Structural Biotechnology Journal*, 18:3335–3343, jan 2020.

[29] M. Raissi, P. Perdikaris, and G. Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, feb 2019.

[30] L. Lu, X. Meng, Z. Mao, and G. E. Karniadakis. DeepXDE: A deep learning library for solving differential equations. *SIAM Review*, 63(1):208–228, jul 2019.

[31] G. E. Karniadakis et al. Physics-informed machine learning. *Nature Reviews Physics 2021 3:6*, 3(6):422–440, may 2021.

[32] J. H. Lagergren et al. Biologically-informed neural networks guide mechanistic modeling from sparse experimental data. *PLOS Computational Biology*, 16(12):e1008462, dec 2020.

[33] A. Yazdani, L. Lu, M. Raissi, and G. E. Karniadakis. Systems biology informed deep learning for inferring parameters and hidden dynamics. *PLOS Computational Biology*, 16(11):e1007575, nov 2020.

[34] M. AlQuraishi and P. K. Sorger. Differentiable biology: using deep learning for biophysics-based and data-driven modeling of molecular mechanisms. *Nature Methods 2021 18:10*, 18(10):1169–1180, oct 2021.

[35] A. Stanoev, C. Schröter, and A. Koseska. Robustness and timing of cellular differentiation through population-based symmetry breaking. *Development*, 148(3):dev197608, feb 2021.

[36] S. Bessonnard et al. Gata6, Nanog and Erk signaling control cell fate in the inner cell mass through a tristable regulatory network. *Development (Cambridge)*, 141(19):3637–3648, oct 2014.

[37] J. Reinitz and D. H. Sharp. Mechanism of eve stripe formation. *Mechanisms of Development*, 49(1-2):133–158, jan 1995.

[38] T. S. Gardner, C. R. Cantor, and J. J. Collins. Construction of a genetic toggle switch in Escherichia coli. *Nature 2000 403:6767*, 403(6767):339–342, jan 2000.

[39] C. Furusawa and K. Kaneko. Emergence of rules in cell society: differentiation, hierarchy, and stability. *Bulletin of mathematical biology*, 60(4):659–687, 1998.

[40] Y. Goto and K. Kaneko. Minimal model for stem-cell differentiation. *Physical Review E*, 88(3):032718, sep 2013.

[41] B. Xia and I. Yanai. A periodic table of cell types. *Development (Cambridge)*, 146(12), jun 2019.

[42] E. Giacomelli et al. Human-iPSC-Derived Cardiac Stromal Cells Enhance Maturation in 3D Cardiac Microtissues and Reveal Non-cardiomyocyte Contributions to Heart Disease. *Cell Stem Cell*, 26(6):862–879.e11, jun 2020.

[43] N. M. Bérenger-Currias et al. A gastruloid model of the interaction between embryonic and extra-embryonic cell types:. *Journal of Tissue Engineering*, 13, jun 2022.

[44] M. B. Elowitz, A. J. Levine, E. D. Siggia, and P. S. Swain. Stochastic gene expression in a single cell. *Science*, 297(5584):1183–1186, aug 2002.

[45] H. Safdari et al. Noise-driven cell differentiation and the emergence of spatiotemporal patterns. *PLOS ONE*, 15(4):e0232060, apr 2020.

[46] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. *Proceedings of Machine Learning Research*, 2010.

[47] A. Mordvintsev, E. Randazzo, E. Niklasson, and M. Levin. Growing Neural Cellular Automata. *Distill*, 5(2):e23, feb 2020.

[48] L. Haustenne et al. On stability analysis of genetic regulatory networks represented by delay-differential equations. *IFAC-PapersOnLine*, 48(1):453–457, jan 2015.

[49] A. A. Kolodziejczyk et al. The Technology and Biology of Single-Cell RNA Sequencing, may 2015.

[50] W. I. T. Uy and M. D. Grigoriu. Neural network representation of the probability density function of diffusion processes. *Chaos*, 30(9), jan 2020.

[51] Y. Xu et al. Solving Fokker-Planck equation using deep learning. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 30(1):013133, jan 2020.

[52] L. Haghverdi et al. Diffusion pseudotime robustly reconstructs lineage branching. *Nature Methods 2016 13:10*, 13(10):845–848, aug 2016.

[53] Z. Ji and H. Ji. TSCAN: Pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis. *Nucleic Acids Research*, 44(13):e117, jul 2016.

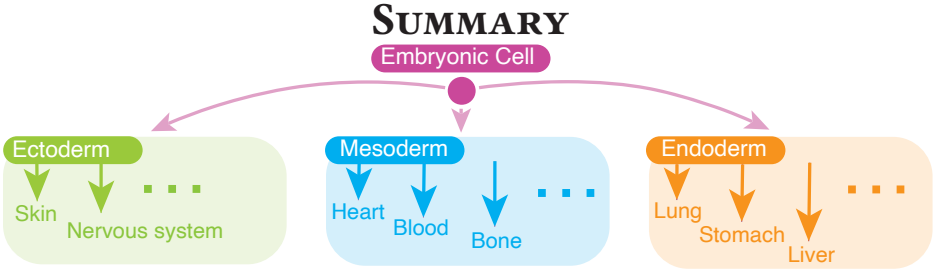


Figure 6.7: The three cell types that occur during gastrulation.

embryonic cells:
Cells of the embryo that have the potential to become any cell of the body.
gastrulation:
The point where three important cell types arise: endoderm, mesoderm and ectoderm.

Human life starts from a single fertilized egg which develops into around 30 trillion cells that make up our body. The egg in the beginning is only a single cell with the potential to divide and change in appearance, function, and composition. These changes are necessary to develop into a complete human, where cells of the heart are very different from cells in the lung or brain. It has been estimated that we have around 200 different cell types in our bodies. Cells at an early stage after fertilization still have the potential to become any cell of the body, and we call them **embryonic cells**. The path from the embryonic cell to a fully specialized cell contains many intermediate stages. We can imagine this path as our way to find a career. As a child, we can still go on to become anything we like, but throughout school and university, we make decisions that further narrow down our field of specialization. With every decision we make, we come one step closer to a specialized field and at the same time it becomes less likely to change career path. For example, one of the first and most crucial decision events in development is **gastrulation** (see Fig. 6.7). At this point, three different cell types emerge from the embryonic cells. This would be similar to making a choice of university studies. This choice will, in most cases, influence the rest of our lives. At this point, the cell can choose between either of the three cell types: endoderm, mesoderm and ectoderm. This choice will narrow down the future career choices: For example, mesoderm will become the heart, blood cells, or bone tissue; Ectoderm will become the nervous system or skin; Endoderm will become the liver, stomach or lungs. In this summary, I want to guide you further through the journey of the embryonic cell.

Information flow in a cell:
DNA → RNA → protein

To really understand how a single cell makes decisions to become a specialized cell type, we have to look into the inside of the cell. We have to figure out where this information is stored and how it is used. Every cell in our body contains the same DNA, which is conserved throughout the cell's life time. It carries information about the building blocks of our bodies. More precisely, it consists of many genes encoded into the DNA. Every gene contains different information, for example about the eye color, hair color or height. It also contains genes that specify cell types. Thus, for each cell type different genes have to become active. In order to do that, the information that the cell currently needs, is passed down from the DNA to the RNA. Every cell in our body has different RNA

molecules. The last information transfer is then, from RNA to protein. It is assumed that the DNA contains around 20.000 genes that code for a protein, and our cells can have as many as 500.000 different proteins. Proteins are the most important components in the cell because they have a wide range of functions. Proteins are like the employees in a company, each one of them has a different job to do and for the company to work everyone must do their job. We can see the inside of an embryonic cell in development like a start-up company. It has to make a lot of changes to its structure and employees until it becomes a well-established firm. With the help of all the employees in the company, the company will in the end make a decision for the next stage. One important kind of protein, in this regard, is called **transcription factor** (TFs). TFs decide which genes will be active and which will remain unused in a cell. They control the protein composition in the cell and can be viewed as the managers in the company. They can hire and fire the people in the company. TFs which can control a high number of genes are called **master TFs** and would therefore be directors of the company. Master TFs in particular often give the cell its identity, meaning that different sets of master TFs are at play in distinct cell types. So, each company has their own group of directors and together they decide on the company's brand. TFs can also control each other and work in parallel. In this way, they form a network called a **gene regulatory network**. In this scenario, you can imagine the management department having to hire or restructure themselves. If several directors are arguing for a position, then the winning group of directors can decide the new company brand.

In order to know which proteins are active inside the cell, they have to be measured. Preferably, we would like to measure all the proteins in a cell to understand exactly what is going on, but that is not yet possible. What is possible, is to measure the RNA molecules in a single cell with a method called **single-cell RNA sequencing** (scRNA-seq). The largest data set produced so far with this method contained 1.3 million brain cells. As proteins are produced from RNA, we can get an estimate of the number of proteins in a cell by measuring the number of RNAs. scRNA-seq has given unprecedented insights into the definition of a cell type and the development of an embryonic cell. Cell types are now defined based on the number of RNAs and many new cell types were discovered in the past decades in this way.

transcription factors:
A protein that can control the activity of some genes.
master TFs:
TFs that can control the activity of many genes.
gene regulatory network:
A network of proteins that activate and suppress each other.
single-cell RNA sequencing:
A method to measure the entire RNA in every single cell.

Chapter 1 of this thesis reviews cell identity in more detail and explains the changes that occur during the development of a cell. It outlines many single-cell techniques currently used to measure DNA, RNA, proteins, and chemical modifications in every single cell. Lastly, it introduces mathematical models for cell development to better understand the mechanisms and causal relationships between genes.

Chapter 2 derives a method named phiclust to evaluate the purity of a cell type in scRNA-seq data. As we can measure all RNA molecules in a cell, it is important to evaluate when two cells are of the same type. Every cell has different RNA molecules due to random fluctuations and processes unrelated to cell identity. Our new method can decide if the differences in RNA molecules between cells are different from random fluctuations. If

the differences are larger, the cells are of different types. We used phiclust on scRNA-seq data of a developing kidney and discovered previously overlooked cell types. In this way, phiclust can help classify and identify the many cell types in our body.

During development, cells are influenced by many factors that can change the cells' decision. Just like for us, when we decide on a career path, it matters what environmental influences we were exposed to. The same happens for cells: Their decision depends on the decisions of other cells in their environment. Naturally, an embryonic cell is inside the womb of the mother. Studying cells in the natural context is called **in vivo**. For ethical reasons, it is impossible to study human development in vivo. Additionally, many processes happen simultaneously in vivo that we can not control and we also do not know their influence. Thus, we decided to look at **in vitro** systems. In vitro refers to studies outside the normal biological context in a controlled lab environment. To study embryonic cells outside the body, they must be extracted and cultured in a dish. Once they are stable in vitro, we call them embryonic stem cells. Embryonic stem cells still maintain their potential to become any cell type of the body but have been extracted from their original environment. Also for ethical reasons, biologists often use **induced pluripotent stem cells** (iPSCs). These cells are taken from mature cells of the body, for example, the skin. Then, they are reverse-engineered to the characteristics of a stem cell and regain the potential to become any cell type of the body. In summary: *embryonic cells in vivo are embryonic stem cell in vitro.*

Chapter 3 studies the influence of the master TF ETV2 in blood vessel cells in vitro. Cells that form the vessels in our body are called **endothelial cells**. In general, their function is to create a barrier between a liquid, for example, blood, and the surrounding tissue. In particular, blood vessel-forming endothelial cells are important for the proper functioning of the heart. To understand the decision process of an endothelial cell, iPSCs were put into a dish and stimulated by specific chemical cues to become endothelial cells. But as described above, there are intermediate stages in the journey of the embryonic cell, which have to be respected in vitro. Using the career path metaphor, this means that first, it has to choose one of three basic career paths that cells have. For an endothelial cell this is mesoderm. In order to become a blood vessel, it has to specialize, like during master studies, on **cardiac** mesoderm (related to the heart). After its master studies, it can now choose the job of an endothelial cell. Surprisingly, we found that in the experiments, two cell types arise: endothelial cells and heart muscle cells which are responsible for the contractions in the heart. This means that after the master studies the cells could still choose between two different jobs. Using scRNA-seq, we showed that the decision depends on the master TF ETV2. If there is less ETV2 the iPSCs will become heart muscle cells and if there is more ETV2 they will become endothelial cells. Here, we can see how important master TFs are in the decision of the cell. As a master TF, ETV2 is able to manage the activity of other genes. So, if enough ETV2 is in the cell, then it will activate genes of the endothelial cell identity and in this way re-brand the cell. After re-branding, ETV2 is no longer needed and will be fired from

in vivo:
A process where cells are in their natural biological context.
in vitro:
A process where cells are outside their biological context in a dish.
induced pluripotent stem cells: Mature cells reverse-engineered to stem cells.
endothelial cells:
Cells that form vessels in the body.
cardiac:
Related to the heart.

the director position. This is an example of a gene regulatory network and how it changes as the cell makes choices.

As mentioned above, environmental factors can influence a cell's decision. One way in which this happens is by communication with each other. Cells communicate for example via specific proteins named **ligands**. This type of protein has the ability to leave the cell and wander off to the next cell. Once it arrives at another cell it changes the gene regulatory network of the receiving cell. In other words, this messaging protein, can influence the decisions of the managers in another company. In the first half of this thesis, the focus was on changes in the internal compositions of the cells. In the second half, we will extend this by investigating the influence of cellular communication on the cell's composition. We will study cell types combined in vitro that naturally would occur together in vivo to understand how they influence each other.

ligands:
Proteins that are used by cells to communicate.
paraxial:
Cells situated alongside an axis.
extraembryonic endoderm cells:
Cells surrounding the outside of the embryo.

Chapter 4 compares the influence of developmental origin to cellular communication in endothelial cells. This chapter is a continuation of *Chapter 3*, where heart muscle and endothelial cells were derived. In vivo, the human heart comprises multiple cell types, including heart muscle cells, blood vessel cells (endothelial cells), and connective tissue. So, these three cell types were put together in a dish in vitro to create a heart tissue environment. But endothelial cells not only form blood vessels but can for example also form lymph vessels that transport lymph instead of blood in the body. We wanted to understand how endothelial cells decide which type of vessel to become: Are they either influenced by the communication with neighboring cells or influenced by the cell's developmental origin (previous career choices)? For this reason, a second experiment was added where endothelial cells specialized in their master studies on **paraxial** mesoderm instead of cardiac mesoderm. Endothelial cells stemming from paraxial mesoderm usually form lymphatic vessels rather than blood vessels. Then, both endothelial cells, from either background, were put into the same heart tissue environment. With scRNA-seq, we saw that the RNA profiles of both endothelial cells were very similar after integration into the heart tissue environment. This result shows that cellular communication is more important than the developmental background for endothelial cells. It also shows in general that environmental influences have a high impact on the decisions of cells.

In a second example in *Chapter 5*, we show how cellular communication can cause the formation of a neural tube in vitro. We investigated cellular communication in a well-known in vitro model system, namely gastruloids. This system is a model for gastrulation which is the moment of choice between the three university studies of the cell (see Fig. 6.7): endoderm, ectoderm and mesoderm. We used mouse embryonic stem cells to reproduce this important decision event. The embryonic stem cells are able to form all three cell types by themselves in a dish. But we know from in vivo studies that the embryo is surrounded by a layer of cells called **extraembryonic endoderm cells** (XEN). When embryonic stem cells were combined with XEN cells in vitro, the XEN cells formed an outer layer around the stem cells, just like in vivo. Even more striking was that the stem cells started forming structures that looked like a tube. These structures have not been observed without the

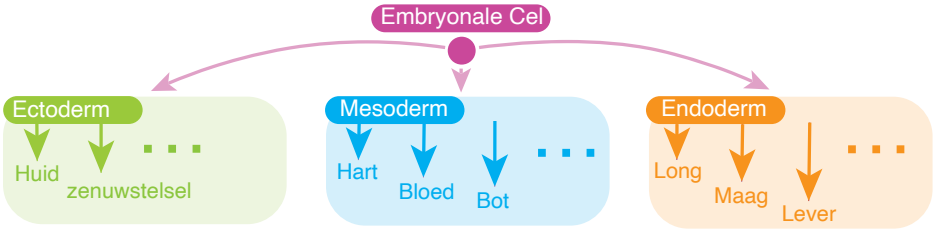
addition of XEN cells. Thus, they must be a direct result of the communication between the two types of cells. With scRNA-seq and imaging, we characterized these structures. We found that they evolve from ectoderm, which is responsible for creating the nervous system. These findings indicate the involvement of XEN cells in the formation of neural tube-like structures. Additionally, we observed that also the mouse embryonic stem cells cause changes in the RNA of XEN cells. This experiment shows that communication between cells is already very important in the early stages of a cell's path.

dynamical systems:
Equations that evolve through time.

In the previous chapters, we have characterized elements involved in a cell's decisions with scRNA-seq and imaging. However, to understand the causal relationships in a gene regulatory network we want to formulate mathematical models. Therefore, we set out to use this data to inform mathematical equations about gene interactions during development in *Chapter 6*. These equations describe the time evolution of each gene in the gene regulatory network and are called **dynamical systems**. In a dynamical system, parameters determine the strength of interactions between genes or molecules. We employed physics-informed deep neural networks (PINNs) to determine these parameters with measured data. A neural network is a machine learning technique usually used successfully for pattern recognition. PINNs combine machine learning with physical laws. In this way, we can learn patterns from the data while conforming to predefined mathematical equations. We covered two relevant experimental scenarios: One where cells communicate and measurements of single cells are available. This data can, for example, be generated by imaging proteins through time. In the other scenario, cells do not communicate, and only snapshot data is obtainable. This model could then incorporate, for instance, scRNA-seq measurements. This analysis provides a starting point for acquiring mechanistic insights into the development of an embryonic cell by utilizing different data sets.

I hope this summary has allowed you to get a glimpse of the very complex journey of an embryonic cell. We showed how to define cell identity based on scRNA-seq data, and took a closer look at the influence of master TFs. Then, we added cellular communication to the experiments and found that it crucially influences a cell's developmental path. Lastly, we wanted to combine all this information into a mathematical model that allows us to understand the mechanisms involved in the cell's decisions. There is still much left to study before we can fully understand how a cell makes its decisions.
May the embryonic cell always make the right choice!

Samenvatting



Figuur 6.8: De drie celtypen die tijdens de gastrulatie voorkomen.

embryonale cellen:
Cellen van het embryo die de potentie hebben om elke cel van het lichaam te worden.
gastrulatie:
Het punt waar drie belangrijke celtypes ontstaan: endoderm, mesoderm en ectoderm.

Het menselijk leven begint met één bevruchte eicel die zich ontwikkelt tot de ongeveer 30 biljoen cellen waaruit ons lichaam is opgebouwd. In het begin is het ei slechts één enkele cel met het potentieel om zich te delen en te veranderen in uiterlijk, functie en samenstelling. Deze veranderingen zijn nodig om zich tot een volledig mens te ontwikkelen, waarbij cellen van het hart heel anders zijn dan cellen in de longen of de hersenen. Er wordt geschat dat wij ongeveer 200 verschillende celtypes in ons lichaam hebben. Cellen in een vroeg stadium na de bevruchting hebben nog het potentieel om elke cel van het lichaam te worden. Deze noemen we **embryonale cellen**. De weg van de embryonale cel naar een volledig gespecialiseerde cel bevat vele tussenstadia. We kunnen ons dit pad voorstellen als onze weg naar een carrière. Als kind kunnen we nog alles worden wat we willen, maar op school en aan de universiteit nemen we beslissingen die ons specialisatiegebied verder inperken. Met elke beslissing die we nemen komen we een stap dichter bij een gespecialiseerd vakgebied en tegelijkertijd wordt het minder waarschijnlijk om van loopbaan te veranderen. Een van de eerste en meest cruciale beslissingsmomenten in de ontwikkeling is bijvoorbeeld de **gastrulatie** (zie Fig. 6.8). Op dat moment ontstaan drie verschillende celtypes uit de embryonale cellen. Dit zou vergelijkbaar zijn met het maken van een keuze voor een universitaire studie. Deze keuze zal, in de meeste gevallen, de rest van ons leven beïnvloeden. Op dit punt kan de cel kiezen tussen één van de drie celtypen: endoderm, mesoderm en ectoderm. Deze keuze zal de toekomstige loopbaankeuzes beperken: Mesoderm wordt bijvoorbeeld het hart, de bloedcellen of het botweefsel; Ectoderm wordt het zenuwstelsel of de huid; Endoderm wordt de lever, de maag of de longen. In deze samenvatting wil ik u verder begeleiden op de reis van de embryonale cel.

Informatiestroom in een cel:
DNA → RNA → eiwitten

Om werkelijk te begrijpen hoe een enkele cel beslissingen neemt om een gespecialiseerd celtype te worden, moeten we het binnenste van de cel bekijken. We moeten uitzoeken waar deze informatie is opgeslagen en hoe deze wordt gebruikt. Elke cel in ons lichaam bevat hetzelfde DNA, dat gedurende het hele leven van de cel bewaard blijft. Het draagt

informatie over de bouwstenen van ons lichaam. Om preciezer te zijn bestaat het uit vele genen die in het DNA zijn gecodeerd. Elk gen bevat andere informatie, bijvoorbeeld over de kleur van de ogen, de haarkleur of de lengte. Het bevat ook genen die celtypes specificeren. Voor elk celtype moeten dus andere genen actief worden. Om dat te doen, wordt de informatie die de cel op dat moment nodig heeft doorgegeven van het DNA naar het RNA. Elke cel in ons lichaam heeft andere RNA-moleculen. De laatste informatieoverdracht vindt daarna plaats van RNA naar eiwit. Er wordt aangenomen dat het DNA ongeveer 20.000 genen bevat die coderen voor een eiwit, en onze cellen kunnen wel 500.000 verschillende eiwitten bevatten. Eiwitten zijn de belangrijkste bestanddelen van de cel omdat zij een groot aantal functies hebben. Eiwitten zijn als de werknemers in een bedrijf, elk van hen heeft een andere taak te vervullen en om het bedrijf te laten werken moet iedereen zijn werk doen. We kunnen de binnenkant van een embryonale cel in ontwikkeling zien als een startend bedrijf. Het moet veel veranderingen aanbrengen in zijn structuur en werknemers tot het een gevestigde onderneming wordt. Met de hulp van alle werknemers in het bedrijf zal het bedrijf uiteindelijk een beslissing nemen voor de volgende fase. Een soort eiwit dat in dit verband belangrijk is, wordt **transcriptiefactor** (TF) genoemd. TF's beslissen welke genen actief zullen zijn en welke ongebruikt zullen blijven in een cel. Zij controleren de eiwitsamenstelling in de cel en kunnen worden gezien als de managers in het bedrijf. Zij kunnen de werknemers in het bedrijf aannemen en ontslaan. TF's die een groot aantal genen kunnen controleren, worden **master TF's** genoemd en zouden dus directeuren van het bedrijf zijn. Vooral master TF's geven de cel vaak zijn identiteit, wat betekent dat in verschillende celtypes verschillende sets van master TF's een rol spelen. Zo heeft elk bedrijf zijn eigen groep van directeuren en samen beslissen zij over het merk van het bedrijf. TF's kunnen elkaar ook controleren en parallel werken. Op die manier vormen ze een netwerk dat een **genregulerend netwerk** wordt genoemd. In dit scenario kunt u zich voorstellen dat de directie meer moet inhuren of zich moet herstructureren. Als verschillende directeuren om een positie twisten, dan kan de winnende groep directeuren beslissen over het nieuwe bedrijfsmerk.

Om te weten welke eiwitten in de cel actief zijn, moeten ze worden gemeten. Het liefst zouden we alle eiwitten in een cel willen meten om te begrijpen wat er precies aan de hand is, maar dat is nog niet mogelijk. Wat wel mogelijk is, is het meten van de RNA-moleculen in een enkele cel met een methode die **scRNA-seq** (Single Cell RNA Sequencing) wordt genoemd. De grootste verzameling van gegevens die tot dusver met deze methode is geproduceerd bevatte 1,3 miljoen hersencellen. Aangezien eiwitten worden gemaakt uit RNA, kunnen we een schatting krijgen van het aantal eiwitten in een cel door het aantal RNA's te meten. scRNA-seq heeft ongekende inzichten opgeleverd in de definitie van een celtype en de ontwikkeling van een embryonale cel. Celtypen worden nu gedefinieerd op basis van het aantal RNA's en vele nieuwe celtypen zijn in de afgelopen decennia op deze manier ontdekt.

transcriptiefactoren:
Een eiwit dat de activiteit van sommige genen kan controleren.
master TFs:
TF's die de activiteit van veel genen kunnen regelen.
genregulerend netwerk:
Een netwerk van eiwitten die elkaar activeren en onderdrukken.
single-cell RNA sequencing:
Een methode om het volledige RNA in elke cel te meten.

In *hoofdstuk 1* van dit proefschrift wordt de celidentiteit in meer detail besproken en wordt uitgelegd welke veranderingen zich voordoen tijdens de ontwikkeling van een cel.

Het schetst vele single-cell technieken die momenteel worden gebruikt om DNA, RNA, eiwitten en chemische modificaties in elke enkele cel te meten. Tenslotte introduceert het wiskundige modellen voor celontwikkeling om de mechanismen en oorzakelijke verbanden tussen genen beter te begrijpen.

Hoofdstuk 2 leidt een methode af, genaamd phiclust, om de zuiverheid van een celtype in scRNA-seq gegevens te evalueren. Aangezien we alle RNA moleculen in een cel kunnen meten, is het belangrijk te evalueren wanneer twee cellen van hetzelfde type zijn. Elke cel heeft verschillende RNA-moleculen als gevolg van willekeurige fluctuaties en processen die geen verband houden met celidentiteit. Onze nieuwe methode kan bepalen of de verschillen in RNA-moleculen tussen cellen verschillend zijn van willekeurige fluctuaties. Als de verschillen groter zijn, dan zijn de cellen van verschillende types. We hebben phiclust gebruikt op scRNA-seq data van een ontwikkelende nier en ontdekten celtypes die voorheen over het hoofd zijn gezien. Op deze manier kan phiclust helpen bij het classificeren en identificeren van de vele celtypen in ons lichaam.

Tijdens de ontwikkeling worden cellen beïnvloed door vele factoren die de beslissing van de cellen kunnen veranderen. Net als voor ons, wanneer we beslissen over een carrièrepad, maakt het uit aan welke omgevingsinvloeden we zijn blootgesteld. Hetzelfde gebeurt voor cellen: Hun beslissing hangt af van de beslissingen van andere cellen in hun omgeving. Een embryonale cel bevindt zich natuurlijk in de baarmoeder van de moeder. Het bestuderen van cellen in de natuurlijke context wordt **in vivo** genoemd. Om ethische redenen is het onmogelijk de menselijke ontwikkeling in vivo nader te bestuderen. Bovendien vinden in vivo veel processen gelijktijdig plaats die we niet kunnen controleren en waarvan we ook niet weten wat hun invloed is. Daarom hebben we besloten te kijken naar systemen **in vitro**. In vitro verwijst naar studies buiten de normale biologische omgeving in een gecontroleerde laboratoriumomgeving. Om embryonale cellen buiten het lichaam te bestuderen, moeten ze worden geëxtraheerd en gekweekt in een schaal. Zodra ze in vitro stabiel zijn, noemen we ze embryonale stamcellen. Embryonale stamcellen behouden nog steeds hun potentieel om elk celtype van het lichaam te worden, maar ze zijn wel uit hun oorspronkelijke omgeving gehaald. Ook om ethische redenen maken biologen vaak gebruik van **geïnduceerde pluripotente stamcellen** (iPSC's). Deze cellen worden uit rijpe cellen van het lichaam gehaald, bijvoorbeeld uit de huid. Vervolgens worden zij omgevormd tot de kenmerken van een stamcel en krijgen zij het potentieel om elk celtype van het lichaam te worden. Samengevat: *embryonale cellen in vivo zijn embryonale stamcellen in vitro.*

in vivo:
Een proces waar cellen in hun natuurlijke biologische omgeving zijn.
in vitro:
Een proces waarbij cellen buiten hun biologische omgeving in een schaaltje liggen.
geïnduceerde pluripotente stamcellen: Volgroeide cellen die omgevormd zijn tot stamcellen.
endotheliale cellen:
Cellen die vaten vormen in het lichaam.

In *hoofdstuk 3* wordt de invloed van master TF ETV2 op bloedvatcellen in vitro bestudeerd. Cellen die de bloedvaten in ons lichaam vormen worden **endotheelcellen** genoemd. In het algemeen is hun functie het vormen van een barrière tussen een vloeistof, bijvoorbeeld bloed, en het omringende weefsel. In het bijzonder zijn bloedvatvormende endotheelcellen belangrijk voor de goede werking van het hart. Om het besluitvormingsproces van een endotheelcel te begrijpen, werden iPSC's in een schaaltje gelegd en door specifieke

chemische signalen gestimuleerd om endotheelcellen te worden. Maar zoals hierboven beschreven, zijn er tussenstadia in de reis van de embryonale cel, die in vitro moeten worden gerespecteerd. Om de metafoor van het carrièrepad te gebruiken, betekent dit dat de cel eerst één van de drie basis carrièrepaden moet kiezen die voor cellen mogelijk zijn. Voor een endotheelcel is dit het mesoderm. Om een bloedvat te worden moet hij zich, bijvoorbeeld tijdens zijn masterstudie, specialiseren in het **cardiaal** mesoderm (dat verband houdt met het hart). Na zijn masterstudie kan hij nu kiezen voor de taak van een endotheelcel. Verrassend is dat bij de experimenten twee celtypes ontstaan: endotheelcellen en hartspiercellen die verantwoordelijk zijn voor de samentrekkingen van het hart. Dit betekent dat de cellen na de masterstudies nog steeds konden kiezen tussen twee verschillende taken. Met behulp van scRNA-seq hebben we aangetoond dat de beslissing afhangt van de master TF ETV2. Als er minder ETV2 is zullen de iPSCs hartspiercellen worden en als er meer ETV2 is zullen ze endotheelcellen worden. Hier kunnen we zien hoe belangrijk master TFs zijn in de beslissing van de cel. Als een master TF is ETV2 in staat de activiteit van andere genen te beheren. Als er voldoende ETV2 in de cel is zal het dus genen van de endotheliale celidentiteit activeren en op deze manier het merk van de cel te veranderen. Na deze verandering is ETV2 niet langer nodig en wordt het uit de regiefunctie ontslagen. Dit is een voorbeeld van een gen-regulerend netwerk en hoe het verandert als de cel keuzes maakt.

cardiaal:
Gerelateerd aan het hart.
liganden:
Eiwitten die door cellen worden gebruikt om te communiceren.
paraxiaal:
Cellen die langs een as liggen.

Zoals hierboven vermeld, kunnen omgevingsfactoren de beslissing van een cel beïnvloeden. Een manier waarop dit gebeurt is door communicatie met elkaar. Cellen communiceren bijvoorbeeld via specifieke eiwitten, genaamd **liganden**. Dit type eiwit heeft de eigenschap de cel te verlaten en naar de volgende cel te gaan. Als het bij een andere cel aankomt, verandert het het gen-regulerende netwerk van de ontvangende cel. Met andere woorden, dit bericht-eiwit kan de beslissingen van de managers in een ander bedrijf beïnvloeden. In de eerste helft van dit proefschrift lag de nadruk op veranderingen in de interne samenstelling van de cellen. In de tweede helft zullen we dit uitbreiden door de invloed van cellulaire communicatie op de celsamenstelling te onderzoeken. We zullen in vitro celtypes bestuderen die van nature in vivo samen zouden voorkomen om te begrijpen hoe ze elkaar beïnvloeden.

In *hoofdstuk 4* wordt de invloed van de ontwikkelingsoorsprong op de cellulaire communicatie in endotheelcellen vergeleken. Dit hoofdstuk is een voortzetting van *hoofdstuk 3*, waarin hartspiercellen en endotheelcellen werden afgeleid. In vivo bestaat het menselijk hart uit meerdere celtypes, waaronder hartspiercellen, bloedvatcellen (endotheelcellen), en bindweefsel. Daarom werden deze drie celtypes samengebracht in een schaaltje in vitro om een hartweefselomgeving te creëren. Endotheelcellen vormen echter niet alleen bloedvaten, maar kunnen bijvoorbeeld ook lymfevaten vormen die lymfe in plaats van bloed vervoeren in het lichaam. Wij wilden begrijpen hoe endotheelcellen beslissen welk type bloedvat zij worden: Worden ze ofwel beïnvloed door de communicatie met naburige cellen of beïnvloed door de ontwikkelingsoorsprong van de cel (vorige carrièrekeuzes)? Om deze reden werd een tweede experiment toegevoegd waarbij endotheelcellen zich specialiseerden in het **paraxiale** mesoderm in plaats van het cardiale mesoderm. Endo-

theelcellen afkomstig van het paraxiale mesoderm vormen gewoonlijk lymfevaten in plaats van bloedvaten. Vervolgens werden beide endotheelcellen, van beide achtergronden, in dezelfde hartweefselomgeving gebracht. Met scRNA-seq zagen we dat de RNA-profielen van beide endotheelcellen zeer vergelijkbaar waren na integratie in het hartweefselmilieu. Dit resultaat toont aan dat cellulaire communicatie voor endotheelcellen belangrijker is dan de ontwikkelingsachtergrond. Het laat ook in het algemeen zien dat omgevingsinvloeden een groot effect hebben op de beslissingen van cellen.

In een tweede voorbeeld in *hoofdstuk 5* laten we zien hoe cellulaire communicatie kan leiden tot de vorming van een neurale buis in vitro. We onderzochten cellulaire communicatie in een bekend in vitro modelsysteem, namelijk gastruloïden. Dit systeem is een model voor gastrulatie dat het moment bij uitstek is tussen de drie universitaire studies van de cel (zie Fig. 6.8): endoderm, ectoderm en mesoderm. Wij hebben gebruik gemaakt van embryonale stamcellen van muizen om dit belangrijke beslissingsmoment te reproduceren. De embryonale stamcellen zijn in staat om alle drie celtypes zelf te vormen in een schaaltje. Maar we weten uit in vivo studies dat het embryo omgeven is door een laag cellen die we **extraembryonale endodermcellen** (XEN) noemen. Wanneer embryonale stamcellen in vitro werden gecombineerd met XEN-cellen, vormden de XEN-cellen een buitenste laag rond de stamcellen, net als in vivo. Nog opvallender was dat de stamcellen structuren begonnen te vormen die leken op een buis. Deze structuren zijn niet waargenomen zonder toevoeging van XEN-cellen. Ze moeten dus een direct gevolg zijn van de communicatie tussen de twee celtypen. Met scRNA-seq en microscopie hebben wij deze structuren gekarakteriseerd. We ontdekten dat ze evolueren uit het ectoderm, dat verantwoordelijk is voor het ontstaan van het zenuwstelsel. Deze bevindingen wijzen op de betrokkenheid van XEN-cellen bij de vorming van neurale buis-achtige structuren. Bovendien stelden wij vast dat ook de muizenembryonale stamcellen veranderingen veroorzaken in het RNA van XEN-cellen. Dit experiment toont aan dat communicatie tussen cellen al in de vroege stadia van de ontwikkeling van een cel zeer belangrijk is.

In de vorige hoofdstukken hebben we met scRNA-seq en microscopie elementen gekarakteriseerd die betrokken zijn bij de beslissingen van een cel. Echter willen we ook wiskundige modellen formuleren om de causale relaties in een gen-regulerend netwerk te begrijpen. Daarom zijn we begonnen deze gegevens te gebruiken om wiskundige vergelijkingen op te stellen over geninteracties tijdens de ontwikkeling in *Hoofdstuk 6*. Deze vergelijkingen, die **dynamische systemen** worden genoemd, beschrijven de tijdsevolutie van elk gen in het genregulerend netwerk. In een dynamisch systeem bepalen parameters de sterkte van interacties tussen genen of moleculen. Wij gebruikten fysisch geïnformeerde diepe neurale netwerken (PINNs) om deze parameters met gemeten gegevens te bepalen. Een neuraal netwerk is een techniek voor machinaal leren die gewoonlijk met succes wordt gebruikt voor patroonherkenning. PINNs combineren machinaal leren met natuurkundige wetten. Op die manier kunnen we patronen in de gegevens vinden terwijl we ons houden aan vooraf bepaalde wiskundige vergelijkingen. Wij hebben twee relevante experimentele scenario's behandeld: Eén waarbij cellen communiceren en metingen van enkele cellen beschikbaar zijn. Deze gegevens kunnen bijvoorbeeld worden gegenereerd door eiwitten in

extraembryonale cellen van het endoderm:
Cellen die de buitenkant van het embryo vormen.
dynamische systemen:
Vergelijkingen die evolueren in de tijd.

de tijd in beeld te brengen. In het andere scenario communiceren cellen niet, en zijn alleen momentopnamen beschikbaar. In dit model kunnen dan bijvoorbeeld scRNA-seq-metingen worden gedaan. Deze analyse biedt een uitgangspunt voor het verwerven van mechanistische inzichten in de ontwikkeling van een embryonale cel door gebruik te maken van verschillende datasets.

Ik hoop dat deze samenvatting u een glimp heeft laten zien van de zeer complexe reis van een embryonale cel. We hebben laten zien hoe we de celidentiteit kunnen bepalen op basis van scRNA-seq gegevens, en we hebben de invloed van master TF's nader bekeken. Vervolgens hebben we cellulaire communicatie aan de experimenten toegevoegd en ontdekt dat deze een cruciale invloed heeft op het ontwikkelingstraject van een cel. Ten slotte hebben we al deze informatie gecombineerd in een wiskundig model dat ons in staat stelt de mechanismen te begrijpen die betrokken zijn bij de beslissingen van de cel. Er valt echter nog veel te onderzoeken voordat we volledig begrijpen hoe een cel zijn beslissingen neemt.
Moge de embryonale cel altijd de juiste keuze maken!

Zusammenfassung

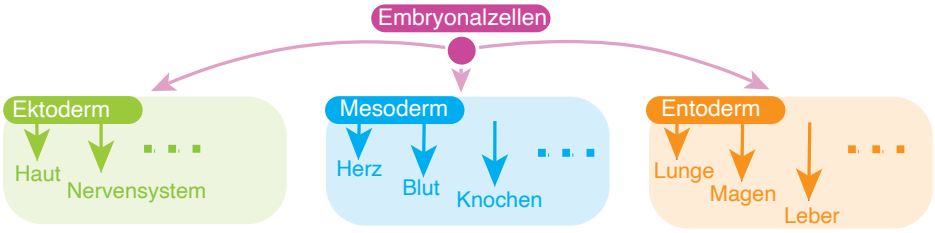


Abbildung 6.9: Die drei Zelltypen, die während der Gastrulation entstehen.

Embryonalzellen:
Zellen des Embryos, die das Potenzial haben, sich zu einer beliebigen Zelle des Körpers zu entwickeln.
Gastrulation:
Der Punkt, an dem drei wichtige Zelltypen entstehen: Entoderm, Mesoderm und Ektoderm.

Das menschliche Leben beginnt mit einer einzigen befruchteten Eizelle, die sich zu etwa 30 Billionen Zellen entwickelt, aus denen unser Körper besteht. Die Eizelle ist zu Beginn nur eine einzige Zelle mit dem Potenzial, sich zu teilen und ihr Aussehen, ihre Funktion und ihre Zusammensetzung zu verändern. Diese Veränderungen sind notwendig, um sich zu einem vollständigen Menschen zu entwickeln, wobei sich die Zellen des Herzens stark von den Zellen der Lunge oder des Gehirns unterscheiden. Man schätzt, dass es in unserem Körper etwa 200 verschiedene Zelltypen gibt. Zellen in einem frühen Stadium nach der Befruchtung haben noch das Potenzial, sich zu jeder Zelle des Körpers zu entwickeln, und wir bezeichnen sie als **embryonale Zellen**. Der Weg von der embryonalen Zelle zu einer voll spezialisierten Zelle enthält viele Zwischenstufen. Wir können uns diesen Weg als unseren Weg zur Berufsfindung vorstellen. Als Kind können wir noch alles werden, was wir wollen, aber während der Schul- und Universitätszeit treffen wir Entscheidungen, die unser Spezialisierungsfeld weiter eingrenzen. Mit jeder Entscheidung, die wir treffen, kommen wir einem Fachgebiet einen Schritt näher, und gleichzeitig wird es unwahrscheinlicher, den Berufsweg zu wechseln. Eines der ersten und wichtigsten Entscheidungsereignisse in der Entwicklung ist zum Beispiel die **Gastrulation** (siehe Abb. 6.9). Zu diesem Zeitpunkt entstehen drei verschiedene Zelltypen aus den embryonalen Zellen. Dies ist vergleichbar mit der Wahl eines Universitätsstudiums. Diese Wahl wird in den meisten Fällen den Rest unseres Lebens beeinflussen. Zu diesem Zeitpunkt kann die Zelle zwischen einem der drei Zelltypen wählen: Entoderm, Mesoderm und Ektoderm. Diese Wahl schränkt die künftige Berufswahl ein: Mesoderm wird zum Beispiel das Herz, die Blutzellen oder das Knochengewebe; Ektoderm wird das Nervensystem oder die Haut; Entoderm wird die Leber, der Magen oder die Lunge. In dieser Zusammenfassung möchte ich Sie weiter durch die Reise der embryonalen Zelle begleiten.

Informationsfluss in einer Zelle:
DNS → RNS → Protein

Um wirklich zu verstehen, wie eine einzelne Zelle Entscheidungen trifft, um zu einem spezialisierten Zelltyp zu werden, müssen wir in das Innere der Zelle schauen. Wir

müssen herausfinden, wo diese Information gespeichert ist und wie sie verwendet wird. Jede Zelle in unserem Körper enthält dieselbe DNS, die während der gesamten Lebenszeit der Zelle erhalten bleibt. Sie enthält Informationen über die Bausteine unseres Körpers. Genauer gesagt besteht sie aus vielen Genen, die in der DNS kodiert sind. Jedes Gen enthält unterschiedliche Informationen, zum Beispiel über die Augenfarbe, Haarfarbe oder Körpergröße. Sie enthält auch Gene, die Zelltypen bestimmen. Für jeden Zelltyp müssen also unterschiedliche Gene aktiv werden. Dazu werden die Informationen, die die Zelle gerade braucht, von der DNS an die RNS weitergegeben. Jede Zelle in unserem Körper hat unterschiedliche RNS-Moleküle. Die letzte Informationsübertragung ist dann die von der RNS zum Protein. Man geht davon aus, dass die DNS etwa 20.000 Gene enthält, die für ein Protein kodieren, und unsere Zellen können bis zu 500.000 verschiedene Proteine haben. Proteine sind die wichtigsten Bestandteile der Zelle, weil sie eine Vielzahl von Funktionen haben. Proteine sind wie die Angestellten in einem Unternehmen, jeder von ihnen hat eine andere Aufgabe, und damit das Unternehmen funktioniert, muss jeder seine Aufgabe erfüllen. Wir können uns das Innere einer embryonalen Zelle in der Entwicklung wie ein Start-up-Unternehmen vorstellen. Es muss viele Änderungen an seiner Struktur und seinen Mitarbeitern vornehmen, bis es zu einem etablierten Unternehmen wird. Mit Hilfe aller Mitarbeiter des Unternehmens wird das Unternehmen am Ende eine Entscheidung für die nächste Phase treffen. Eine Art von Protein, das in diesem Zusammenhang wichtig ist, heißt **Transkriptionsfaktor** (TF). Die TF entscheiden, welche Gene in einer Zelle aktiv werden und welche ungenutzt bleiben. Sie kontrollieren die Proteinzusammensetzung in der Zelle und können als die Manager in einem Unternehmen angesehen werden. Sie können die Mitarbeiter des Unternehmens einstellen und entlassen. TF, die eine große Anzahl von Genen kontrollieren können, werden als "Master-TF" bezeichnet und wären somit die Direktoren des Unternehmens. Insbesondere die Master-TF verleihen der Zelle oft ihre Identität, was bedeutet, dass in verschiedenen Zelltypen unterschiedliche Gruppen von Master-TF im Spiel sind. So hat jedes Unternehmen seine eigene Gruppe von Direktoren, die gemeinsam über die Marke des Unternehmens entscheiden. Die TF können sich auch gegenseitig kontrollieren und parallel arbeiten. Auf diese Weise bilden sie ein Netzwerk, das als **Genregulationsnetzwerk** bezeichnet wird. In diesem Szenario können Sie sich vorstellen, dass die Geschäftsleitung sich selbst einstellen oder umstrukturieren muss. Wenn sich mehrere Direktoren um eine Position streiten, kann die Direktorengruppe, die den Zuschlag erhält, die neue Unternehmensmarke bestimmen.

Um zu wissen, welche Proteine in der Zelle aktiv sind, müssen sie gemessen werden. Am liebsten würden wir alle Proteine in einer Zelle messen, um genau zu verstehen, was vor sich geht, aber das ist noch nicht möglich. Was jedoch möglich ist, ist die Messung der RNS-Moleküle in einer einzelnen Zelle mit einer Methode, die als **single-cell RNA sequencing** (scRNA-seq) bezeichnet wird. Der größte Datensatz, der bisher mit dieser Methode erzeugt wurde, umfasste 1,3 Millionen Gehirnzellen. Da Proteine aus RNS hergestellt werden, können wir die Anzahl der Proteine in einer Zelle schätzen, indem wir die Anzahl der RNS messen. scRNA-seq hat beispiellose Einblicke in die Definition eines Zelltyps und die

Transkriptionsfaktoren:
Ein Protein, das die Aktivität einiger Gene steuern kann.
master TF:
TF, die die Aktivität vieler Gene steuern können.
Genregulationsnetzwerk:
Ein Netzwerk von Proteinen, die sich gegenseitig aktivieren und unterdrücken.
single cell RNA sequencing:
Eine Methode zur Messung der gesamten RNA in jeder einzelnen Zelle.

Entwicklung einer embryonalen Zelle gegeben. Zelltypen werden nun auf der Grundlage der Anzahl der RNS definiert, und viele neue Zelltypen wurden in den letzten Jahrzehnten auf diese Weise entdeckt.

Kapitel 1 dieser Arbeit gibt einen genaueren Überblick über die Zellidentität und erklärt die Veränderungen, die während der Entwicklung einer Zelle auftreten. Es werden zahlreiche Einzelzelltechniken vorgestellt, die derzeit zur Messung von DNS, RNS, Proteinen und chemischen Veränderungen in jeder einzelnen Zelle verwendet werden. Schließlich werden mathematische Modelle für die Zellentwicklung vorgestellt, die dabei helfen die Mechanismen und kausalen Beziehungen zwischen Genen besser zu verstehen.

Kapitel 2 leitet eine Methode namens phiclust her, um die Reinheit eines Zelltyps in scRNA-seq-Daten zu bewerten. Da wir alle RNS-Moleküle in einer Zelle messen können, ist es wichtig zu bewerten, wann zwei Zellen vom gleichen Typ sind. Jede Zelle hat aufgrund von zufälligen Fluktuationen und Prozessen, die nichts mit der Zellidentität zu tun haben, unterschiedliche RNS-Moleküle. Mit unserer neuen Methode lässt sich feststellen, ob die Unterschiede in den RNS-Molekülen zwischen den Zellen von zufälligen Fluktuationen abweichen. Wenn die Unterschiede größer sind, handelt es sich um unterschiedliche Zelltypen. Wir haben phiclust auf scRNA-seq-Daten einer sich entwickelnden Niere angewendet und dabei bisher übersehene Zelltypen entdeckt. Auf diese Weise kann phiclust helfen, die vielen Zelltypen in unserem Körper zu klassifizieren und zu identifizieren.

in vivo:
Ein Prozess, bei dem sich Zellen in ihrem natürlichen biologischen Kontext befinden.
in vitro:
Ein Prozess, bei dem sich Zellen außerhalb ihres biologischen Kontextes in einer Schale befinden.
induzierte pluripotente Stammzellen: Ausgereifte Zellen, die in Stammzellen umgewandelt werden.

Während der Entwicklung werden die Zellen von vielen Faktoren beeinflusst, die die Entscheidung der Zellen verändern können. So wie bei uns, wenn wir uns für einen beruflichen Weg entscheiden, spielt es eine Rolle, welchen Umwelteinflüssen wir ausgesetzt waren. Das Gleiche gilt für Zellen: Ihre Entscheidung hängt von den Entscheidungen der anderen Zellen in ihrer Umgebung ab. Eine embryonale Zelle befindet sich normal im Mutterleib. Das Studium von Zellen im natürlichen Kontext wird als **in vivo** bezeichnet. Aus ethischen Gründen ist es nicht möglich, die menschliche Entwicklung in vivo zu untersuchen. Außerdem laufen in vivo viele Prozesse gleichzeitig ab, die wir nicht kontrollieren können und deren Einfluss wir auch nicht kennen. Daher haben wir beschlossen, uns mit **In-vitro-Systemen** zu befassen. In vitro bezieht sich auf Studien außerhalb des normalen biologischen Kontextes in einer kontrollierten Laborumgebung. Um embryonale Zellen außerhalb des Körpers zu untersuchen, müssen sie entnommen und in einer Schale kultiviert werden. Sobald sie in vitro stabil sind, nennen wir sie embryonale Stammzellen. Embryonale Stammzellen haben immer noch das Potenzial, sich in jeden Zelltyp des Körpers zu verwandeln, sind aber aus ihrer ursprünglichen Umgebung herausgelöst worden. Auch aus ethischen Gründen verwenden Biologen häufig **induzierte pluripotente Stammzellen** (iPSCs). Diese Zellen werden aus reifen Zellen des Körpers, z.B. der Haut, entnommen. Dann werden sie zu stammzellenartigen Zellen zurückverwandelt und erhalten wieder das Potenzial, sich in jeden Zelltyp des Körpers zu verwandeln. Zusammengefasst: *Embryonale Zellen in vivo sind embryonale Stammzellen in vitro.*

Kapitel 3 untersucht den Einfluss des Master-TF ETV2 in Blutgefäßzellen in vitro. Zellen, die die Gefäße in unserem Körper bilden, werden als Endothelzellen bezeichnet. Ihre Aufgabe besteht im Allgemeinen darin, eine Barriere zwischen einer Flüssigkeit, z.B. Blut, und dem umgebenden Gewebe zu bilden. Insbesondere die blutgefäßbildenden **Endothelzellen** sind wichtig für das reibungslose Funktionieren des Herzens. Um den Entscheidungsprozess einer Endothelzelle zu verstehen, wurden iPSCs in eine Schale gegeben und durch bestimmte chemische Reize dazu angeregt, Endothelzellen zu werden. Wie oben beschrieben, gibt es jedoch Zwischenstufen auf der Reise der embryonalen Zelle, die in vitro beachtet werden müssen. In Anlehnung an die Metapher des Karrierewegs bedeutet dies, dass sie zunächst einen der drei grundlegenden Karrierewege wählen muss, die Zellen haben. Für eine Endothelzelle ist dies das Mesoderm. Um ein Blutgefäß zu werden, muss sie sich, wie im Masterstudium, auf das Herzmesoderm spezialisieren. Nach ihrem Masterstudium kann sie nun den Beruf einer Endothelzelle wählen. Überraschenderweise haben wir festgestellt, dass bei den Experimenten zwei Zelltypen entstehen: Endothelzellen und Herzmuskelzellen, die für die Kontraktionen im Herzen verantwortlich sind. Das bedeutet, dass die Zellen nach den Masterstudien immer noch zwischen zwei verschiedenen Aufgaben wählen können. Mithilfe von scRNA-seq konnten wir zeigen, dass die Entscheidung von dem Master-TF ETV2 abhängt. Wenn weniger ETV2 vorhanden ist, werden die iPSCs zu Herzmuskelzellen, wenn mehr ETV2 vorhanden ist, werden sie zu Endothelzellen. Hier können wir sehen, wie wichtig Master-TF für die Entscheidung der Zelle sind. Als Master-TF ist ETV2 in der Lage, die Aktivität anderer Gene zu steuern. Wenn also genügend ETV2 in der Zelle vorhanden ist, aktiviert es Gene, die für die Identität der Endothelzelle verantwortlich sind, und verleiht der Zelle auf diese Weise ein neues Image. Nach dem Markenwechsel wird ETV2 nicht mehr benötigt und wird aus seiner Position als Direktor entlassen. Dies ist ein Beispiel für ein genregulatorisches Netzwerk und wie es sich verändert, wenn die Zelle Entscheidungen trifft.

endotheliale Zellen:
Zellen, die Gefäße im Körper bilden.
Liganden:
Proteine, die von Zellen zur Kommunikation verwendet werden.

Wie bereits erwähnt, können Umweltfaktoren die Entscheidung einer Zelle beeinflussen. Dies geschieht unter anderem durch die Kommunikation untereinander. Zellen kommunizieren zum Beispiel über bestimmte Proteine, die **Liganden** genannt werden. Diese Art von Protein hat die Fähigkeit, die Zelle zu verlassen und zur nächsten Zelle zu wandern. Sobald es in einer anderen Zelle ankommt, verändert es das Genregulationsnetzwerk der empfangenden Zelle. Mit anderen Worten: Dieses Nachrichtenprotein kann die Entscheidungen der Manager eines anderen Unternehmens beeinflussen. In der ersten Hälfte dieser Arbeit lag der Schwerpunkt auf den Veränderungen in der internen Zusammensetzung der Zellen. In der zweiten Hälfte werden wir dies erweitern, indem wir den Einfluss der zellulären Kommunikation auf die Zusammensetzung der Zelle untersuchen. Wir werden in vitro kombinierte Zelltypen untersuchen, die in vivo natürlich zusammen vorkommen, um zu verstehen, wie sie sich gegenseitig beeinflussen.

Kapitel 4 vergleicht den Einfluss des Entwicklungsursprungs auf die zelluläre Kommunikation in Endothelzellen. Dieses Kapitel ist eine Fortsetzung von *Kapitel 3*, wo Herzmuskel- und Endothelzellen abgeleitet wurden. In vivo besteht das menschliche Herz aus mehreren Zelltypen, darunter Herzmuskelzellen, Blutgefäßzellen (Endothelzellen) und Bindegewebe.

Daher wurden diese drei Zelltypen in einer Schale in vitro zusammengebracht, um eine Herzgewebeumgebung zu schaffen. Endothelzellen bilden jedoch nicht nur Blutgefäße, sondern können zum Beispiel auch Lymphgefäße bilden, die Lymphe statt Blut im Körper transportieren. Wir wollten verstehen, wie Endothelzellen entscheiden, welche Art von Gefäß sie bilden: Werden sie entweder durch die Kommunikation mit benachbarten Zellen oder durch die Entwicklungsherkunft der Zelle (frühere Berufswahl) beeinflusst? Aus diesem Grund wurde ein zweites Experiment durchgeführt, bei dem sich die Endothelzellen in ihren Masterstudien auf das **paraxial** Mesoderm statt auf das kardiale Mesoderm spezialisierten. Endothelzellen, die aus dem paraxialen Mesoderm stammen, bilden normalerweise eher Lymphgefäße als Blutgefäße. Anschließend wurden beide Endothelzellen, unabhängig von ihrem Hintergrund, in dieselbe Herzgewebeumgebung gebracht. Mittels scRNA-seq konnten wir feststellen, dass die RNS-Profile beider Endothelzellen nach der Integration in die Herzgewebeumgebung sehr ähnlich waren. Dieses Ergebnis zeigt, dass die zelluläre Kommunikation für Endothelzellen wichtiger ist als der Entwicklungshintergrund. Es zeigt auch allgemein, dass Umwelteinflüsse einen großen Einfluss auf die Entscheidungen von Zellen haben.

In einem zweiten Beispiel in *Kapitel 5* zeigen wir, wie zelluläre Kommunikation die Bildung eines Neuralrohrs in vitro verursachen kann. Wir untersuchten die zelluläre Kommunikation in einem bekannten In-vitro-Modellsystem, den Gastruloiden. Dieses System ist ein Modell für die Gastrulation, die der Moment der Wahl zwischen den drei universitären Studien der Zelle ist (siehe Abb. 6.9): Entoderm, Ektoderm und Mesoderm. Wir haben embryonale Stammzellen der Maus verwendet, um dieses wichtige Entscheidungsereignis zu reproduzieren. Die embryonalen Stammzellen sind in der Lage, alle drei Zelltypen in einer Schale selbst zu bilden. Aus In-vivo-Studien wissen wir jedoch, dass der Embryo von einer Schicht von Zellen umgeben ist, die als **extraembryonale Entodermzellen** (XEN) bezeichnet werden. Als embryonale Stammzellen in vitro mit XEN-Zellen kombiniert wurden, bildeten die XEN-Zellen eine äußere Schicht um die Stammzellen, genau wie in vivo. Noch auffälliger war, dass die Stammzellen anfingen, Strukturen zu bilden, die wie eine Röhre aussahen. Diese Strukturen sind ohne die Zugabe von XEN-Zellen nicht beobachtet worden. Sie müssen also ein direktes Ergebnis der Kommunikation zwischen den beiden Zelltypen sein. Mit scRNA-seq und Bildgebung haben wir diese Strukturen charakterisiert. Wir fanden heraus, dass sie sich aus dem Ektoderm entwickeln, das für die Bildung des Nervensystems verantwortlich ist. Diese Ergebnisse deuten auf die Beteiligung von XEN-Zellen an der Bildung von Neuralrohr-ähnlichen Strukturen hin. Darüber hinaus haben wir festgestellt, dass auch die embryonalen Stammzellen der Maus Veränderungen in der RNS der XEN-Zellen verursachen. Dieses Experiment zeigt, dass die Kommunikation zwischen den Zellen bereits in den frühen Stadien des Weges einer Zelle sehr wichtig ist.

paraxial:
Zellen, die entlang einer Achse liegen.
extraembryonale Endodermzellen:
Zellen, die die Außenseite des Embryos umgeben.

In den vorangegangenen Kapiteln haben wir Elemente, die an den Entscheidungen einer Zelle beteiligt sind, mit scRNA-seq und Bildgebung charakterisiert. Um jedoch die kausalen Beziehungen in einem genregulatorischen Netzwerk zu verstehen, müssen wir mathe-

matische Modelle formulieren. Deshalb haben wir uns vorgenommen, diese Daten zu nutzen, um mathematische Gleichungen über Geninteraktionen während der Entwicklung in *Kapitel 6* aufzustellen. Diese Gleichungen beschreiben die zeitliche Entwicklung jedes Gens im Genregulationsnetzwerk und werden als **dynamische Systeme** bezeichnet. In einem dynamischen System bestimmen die Parameter die Stärke der Wechselwirkungen zwischen Genen oder Molekülen. Wir haben physikinformierte tiefe neuronale Netze (PINN) eingesetzt, um diese Parameter anhand von Messdaten zu bestimmen. Ein neuronales Netz ist eine Technik des maschinellen Lernens, die normalerweise erfolgreich zur Mustererkennung eingesetzt wird. Die PINN kombinieren maschinelles Lernen mit physikalischen Gesetzen. Auf diese Weise können wir Muster aus den Daten lernen und dabei vordefinierte mathematische Gleichungen einhalten. Wir haben zwei relevante experimentelle Szenarien untersucht: Eines, in dem Zellen kommunizieren und Messungen von einzelnen Zellen verfügbar sind. Diese Daten können z. B. durch die Abbildung von Proteinen im Zeitverlauf gewonnen werden. In dem anderen Szenario kommunizieren die Zellen nicht, und es sind nur Momentaufnahmen möglich. Dieses Modell könnte dann z.B. scRNA-seq-Messungen einbeziehen. Diese Analyse bietet einen Ausgangspunkt für die Gewinnung mechanistischer Erkenntnisse über die Entwicklung einer embryonalen Zelle durch die Nutzung verschiedener Datensätze.

dynamische Systeme: Gleichungen, die sich im Laufe der Zeit weiterentwickeln.

Ich hoffe, diese Zusammenfassung hat Ihnen einen Einblick in die sehr komplexe Reise einer embryonalen Zelle gegeben. Wir haben gezeigt, wie die Zellidentität auf der Grundlage von scRNA-seq-Daten definiert werden kann, und haben uns den Einfluss von Master-TF genauer angesehen. Dann fügten wir den Experimenten die zelluläre Kommunikation hinzu und stellten fest, dass sie den Entwicklungsweg einer Zelle entscheidend beeinflusst. Schließlich wollten wir all diese Informationen in einem mathematischen Modell zusammenfassen, das es uns ermöglicht, die Mechanismen zu verstehen, die an den Entscheidungen der Zelle beteiligt sind. Es gibt noch viel zu erforschen, bevor wir vollständig verstehen können, wie eine Zelle ihre Entscheidungen trifft.
Möge die embryonale Zelle immer die richtige Entscheidung treffen!

List of Publications

(1) **Maria Mircea**, Mazène Hochane, Xueying Fan, Susana M. Chuva De Sousa Lopes, Diego Garlaschelli, Stefan Semrau. Phiclust: a clusterability measure for single-cell transcriptomics reveals phenotypic subpopulations. Genome Biology (2021). https://doi.org/10.1186/s13059-021-02590-x.

(2) **Maria Mircea** and Stefan Semrau. How a cell decides its own fate: A single-cell view of molecular mechanisms and dynamics of cell type specification. Biochemical Society Transactions (2021). https://doi.org/10.1042/BST20210135.

(3) Elisa Giacomelli, Viviana Meraviglia, Giulia Campostrini, Amy Cochrane, Xu Cao, Ruben W.J. van Helden, Ana Krotenberg Garcia, **Maria Mircea** et al. Human-iPSC-Derived Cardiac Stromal Cells Enhance Maturation in 3D Cardiac Microtissues and Reveal Non-cardiomyocyte Contributions to Heart Disease. Cell Stem Cell 26, 862-879 (2020). https://doi.org/10.1016/j.stem.2020.05.004.

(4) Noémie M. L. P. Bérenger-Currias, **Maria Mircea** et al. Early neurulation recapitulated in assemblies of embryonic and extraembryonic cells. bioRxiv (2020). https://doi.org/10.1101/2020.02.13.947655.

(5) Gökcen Eraslan, Lukas M. Simon, **Maria Mircea**, Nikola S. Mueller and Fabian J. Theis. Single-cell RNA-seq denoising using a deep count autoencoder. Nat Commun 10, 390 (2019). https://doi.org/10.1038/s41467-018-07931-2.

(6) **Maria Mircea** and Jürgen Pfeffer. Galois Lattice and Positional Dominance. Italian Journal of Applied Statistics. https://doi.org/10.26398/IJAS.0030-001 .

(7) Xu Cao*, **Maria Mircea*** et al. Tissue microenvironment partially removes signatures of developmental origin in a 3D in vitro model of cardiac endothelial cell differentiation. Manuscript in preparation.

(8) Xu Cao, **Maria Mircea**. ETV2 upregulation marks lineage-restricted cardiomyocyte and endothelial cell precursors during their co-differentiation from hiPSCs. Manuscript submitted.

Included in this thesis.
* equal contribution

Curriculum Vitæ

Maria Mircea

17.07.1993	Born in Bucharest, Romania

Education

2000 - 2012	High school Gymansium Dorfen, Germany
2012 - 2015	Bachelor in Mathematics Technical University Munich, Germany *Thesis:* Comparison of dimensionality reduction methods. *Supervisor:* Prof. Dr. Frank Filbir.
2015	Master Exchange Program Mathematics National University of Colombia, Colombia
2015 - 2018	Master in Mathematics Technical University Munich, Germany *Thesis:* Single-cell RNA-seq denoising using a deep count autoencoder. *Supervisors:* Dr. Lukas Simon and Prof. Dr. Fabian Theis.
2018 - 2022	Ph.D. Biophysics Leiden University, Netherlands *Thesis:* Unravelling cell fate decisions through single cell high throughput methods and mathematical models *Supervisors:* Dr. Stefan Semrau and Dr. Diego Garlaschelli.
2022 - present	PostDoc Mathematics in Life Science University Bonn, Germany *Topic:* Parameters of dynamical systems estimated by Deep Neural Networks to understand the effects of meta-inflammation. *Supervisor:* Prof. Dr. Jan Hasenauer.