Statistical data integration for single-cell RNA-sequencing - batch effect correction and lineage inference

Maren Büttner

February 2019

TECHNISCHE UNIVERSITÄT MÜNCHEN

Fakultät für Mathematik — Lehrstuhl M12 (Mathematische Modellierung biologischer Systeme)

Statistical data integration for single-cell RNA-sequencing - batch effect correction and lineage inference

Maren Büttner

Vollständiger Abdruck der von der Fakultät für Mathematik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzende: Prof. Dr. Silke Rolles

Prüfer der Dissertation:

1. Prof. Dr. Dr. Fabian J. Theis
2. Prof. Dr. Julien Gagneur
3. Prof. Peter V. Kharchenko, Ph.D.
   (schriftliche Beurteilung)

Die Dissertation wurde am 12.02.2019 bei der Technischen Universität München eingereicht und durch die Fakultät für Mathematik am 09.05.2019 angenommen.

Acknowledgment

Zuallererst möchte ich meinem Doktorvater Fabian J. Theis danken, dass er mir mein Promotionsvorhaben ermöglicht und mich auf dem durchaus steinigen Weg begleitet hat. Ich hatte die Gelegenheit, mich eigenen Projektideen zu widmen, und wurde immer wieder vor neue Herausforderungen gestellt. Das ICB ist ein wunderbarer Ort für einen Doktoranden, wenn auch seine Lage das gut zu verstecken weiß.

Mein Dank gilt auch meinen vielen tollen, sympathischen und empathischen Kollegen, die immer ein offenes Ohr für die Sorgen, Nöte und Zweifel im Doktoranden-Dasein hatten, ihre eigenen Geschichten geteilt haben und für eine angenehme Atmosphäre am ICB gesorgt haben.

Ich möchte meinem Graduiertenkolleg QBM und deren Koordinatoren für ihre kontinuierliche Hingabe danken, mit der sie viele meiner Vorhaben begleitet und wegweisende Anreize schufen. Ohne euch wäre hätte es nie die produktive Kooperation mit Sarah Teichmann gegeben.

Ein ganz besonderer Dank gilt meinen Karate-Vereinen und meinen Freunden vom Sport, die mir über viele frustrierende Momente hinweg geholfen haben, weil ich in den Stunden beim Training, egal ob als Trainierender oder Trainerin, alles Übel vergessen konnte und wieder zu mir selbst zurück fand. Ohne den sportlichen Ausgleich wäre dieses Vorhaben nicht zu einem glücklichen Ende gekommen.

Zu guter Letzt bleibt mir noch zu sagen, dass eine Danksagung nur einen Bruchteil meiner Dankbarkeit widergeben kann und all jenen, die um ihre Wichtigkeit für mich wissen, kann ich nur immer wieder auf Knien danken, sie umarmen und meine Freude mit ihnen teilen.

Abstract

A single living cell holds the potential to create a complete multicellular organism. On this route, cells meet molecular crossroads and undergo various changes. When single cells create tissues, the heterogeneity that drives these cell fate decisions is only accessible at the single-cell resolution. Recently, the development of single-cell transcriptomics was key to reveal this heterogeneity. This technology empowers sequencing of ten thousands of cells and machine learning algorithms allow dissecting such big data. Therefore, single-cell transcriptomics has great potential for personalized medicine.

In this thesis, I address the issue of reproducibility through replicated measurements. A systematic bias, called *batch effect*, reduces the interpretability of data, but it has been unclear how to best correct for batch effects in single-cell data. Furthermore, as part of the Human Cell Atlas, which charts all cells in the human body, we need systematic tests for batch effects to integrate data from various platforms and tissues. Thus, I implemented a sensitive and unbiased measure called *kBET*, which allows to quantify batch effects in single-cell data with replicates and large sample numbers. Furthermore, to establish best practices for single-cell transcriptome data analysis, I used kBET to evaluate common normalization and batch correction strategies.

The second goal of this thesis was to characterize cell fate decision making and lineage choice in the continuously self-renewing small intestinal epithelium of adult mice. Here, high-throughput single-cell transcriptomics allows profiling rare cell types and its progenitors. First, I characterized all major cell populations and their progenitors in novel detail. To derive a biologically sound connection between cell types, I contributed to the concept of diffusion pseudotime, which assigns cells by similarity of their transcriptomic profile to a one-dimensional metric. This metric is the basis to infer a single-cell graph that connects the cell type clusters as lineages. I validated connections in the lineage model using short-term lineage tracing experiments. Second, I investigated how a perturbation of molecular signaling affects cell composition and lineage choice in the small intestinal epithelium.

Zusammenfassung

Eine einzige lebende Zelle birgt das Potenzial, einen kompletten multizellulären Organismus zu bilden. Auf dem Weg dorthin treffen die Zellen auf molekulare Kreuzungen und durchlaufen unzählige Veränderungen. Wenn einzelne Zellen neue Gewebe bilden, dann ist die Heterogenität, die die zellulären Entscheidungen treibt, nur durch Einzelzellauflösung zugänglich. In jüngster Zeit war die Entwicklung der Einzelzell-Transkriptomik der Schlüssel, um diese Heterogenität aufzudecken. Diese Technologie ermöglicht die Sequenzierung von Zehntausenden Zellen und maschinelle Lernalgorithmen können so große Datenmengen handhaben. Daher hat die Einzelzell-Transkriptomik ein großes Potenzial für die personalisierte Medizin.

In dieser Arbeit beschäftige ich mich mit der Frage der Reproduzierbarkeit durch wiederholte Messungen. Eine systematische Verzerrung, genannt *Batch-Effekt*, über Replikate hinweg reduziert die Interpretationsfähigkeit der Daten. Es war jedoch unklar, wie man Batch-Effekte in Einzelzelldaten am besten korrigieren kann. Darüber hinaus benötigen wir im Rahmen des Human Cell Atlas, der alle Zellen im menschlichen Körper zu kartieren wird, systematische Tests für Batch-Effekte, um Daten von verschiedenen Plattformen und Geweben zu integrieren. In der vorliegenden Arbeit habe ich eine empfindliche und unverzerrte Methode namens *kBET* implementiert, die es ermöglicht, Batch-Effekte in Einzelzelldaten mit Replikaten und großen Probenzahlen zu quantifizieren. Darüber hinaus möchte ich einen Goldstandard für die Einzelzell-Transkriptomdatenanalyse etablieren. Für diesen Zweck habe ich kBET verwendet, um gängige Normalisierungs- und Batch-Korrektur-strategien zu bewerten.

Das zweite Ziel dieser Arbeit war es, die zelluläre Entscheidungsfindung und die Abstammungslinien im sich ständig selbst erneuernden Dünndarm-Epithel erwachsener Mäuse aufzuklären. Hier ermöglichte die Hochdurchsatz-Einzelzellen-Transkriptomik die Erfassung seltener Zelltypen und ihrer Voräufer, die bei Niedrigdurchsatz-Ansätzen unbemerkt blieben. Zuerst habe ich alle wichtigen Zellpopulationen und ihre Vorfahren in neuer Detailiertheit charakterisiert. Um eine biologisch sinnvolle Verbindung zwischen Zelltypen abzuleiten, habe ich zum Konzept der Diffusionspseudozeit beigetragen, das Zellen durch Ähnlichkeit ihres transkriptomischen Profils einer eindimensionalen Metrik zuordnet. Diese Metrik erlaubt die Konstruktion eines Einzelzell-Graphs, welcher die Zelltypen-Cluster miteinander verbindet. Weiterhin konnte ich die Verbindungen des Abstammungsmodells mittels Kurzzeit-Abstammungsverfolgungsexperimente validieren. Zweitens untersuchte ich, wie eine Störung der streng kontrollierten molekularen Signalgebung die Zellzusammensetzung und die Abstammungslinien im Dünndarm-Epithel beeinflusst.

Contents

1 Introduction **1**

1.1 The molecular organisation of the animal cell 2

1.2 The study of single cells at the molecular level 4

1.3 The technology of single-cell RNA-sequencing 7

1.4 Computational challenges for single-cell RNA-seq data integration 11

1.5 New insights to development and lineage specification 13

1.5.1 Mouse embryonic stem cells 14

1.5.2 Lineage specification in the small intestinal epithelium 15

1.6 Research questions and thesis overview 17

1.7 List of contributed articles 20

2 Methods for single-cell RNA-seq data analysis **21**

2.1 Cell filtering 23

2.2 Gene filtering 25

2.3 Normalization 25

2.4 Batch effect correction 31

2.5 Correcting for cell cycle effects 38

2.6 Identification of highly variable genes 39

2.7 Dimensionality reduction 41

2.8 Clustering 47

2.9 Pseudotime 49

2.10 Differential Expression Analysis 52

3 Batch effect quantification **57**

3.1 A new test metric for batch effect quantification 58
3.1.1 χ^2-based test 59
3.1.2 Exact test 59
3.1.3 Likelihood-ratio test 60
3.1.4 Optimal neighborhood size 60
3.1.5 Isolated cells and adaptive frequencies 61
3.1.6 A significance test of the batch effect estimate 63
3.2 Prior approaches of batch effect quantification 64
3.2.1 Silhouette coefficient 64
3.2.2 Principal Component Regression 65
3.3 Quantifying preservation of biological signals 70
3.4 Application to simulation studies 71
3.4.1 A zero-inflated negative binomial model 71
3.4.2 Batch effect strength is modeled as an increasing fraction affected genes 72
3.4.3 Simulating single-cell RNA-seq count data using Splat 73
3.4.4 Simulation results and comparison to alternative measures 74
3.5 Benchmark batch effect correction in scRNA-seq data 80
3.5.1 kBET outperforms visual inspection 81
3.5.2 Deriving a best practice in batch correction 83
3.5.3 Time and memory test of the batch regression methods 89
3.5.4 Dropout handling in batch effect correction of scRNA-seq data . . . 90
3.6 Going beyond batch effect correction 91
3.6.1 Going beyond replicates - dataset integration across multiple studies 92

3.6.2 Going beyond batch effects - kBET detects variability between individuals in PBMC data . . . 97

3.7 Discussion and outlook . . . 99

4 Decipher the intestinal lineage hierarchy 103

4.1 Data collection . . . 105

4.2 Computational analysis of single-cell RNA-seq data . . . 107

4.2.1 Batch correction using a modified version of ComBat . . . 110

4.2.2 Batch effect quantification . . . 111

4.3 Results . . . 112

4.3.1 *Fltp* is a new marker for Paneth and enteroendocrine cells . . . 113

4.3.2 Wnt/PCP activates a subset of Lgr5$^+$ ISCs . . . 114

4.3.3 Single-cell RNA-seq recovers all lineages in the small intestinal epithelium . . . 114

4.3.4 Enrichment for reporter expression captures Paneth cells and EECs 117

4.3.5 Each intestinal lineage has a distinct transcriptional profile . . . 123

4.3.6 Disturbed Wnt/PCP signaling reduces Paneth cell formation . . . 130

4.3.7 Disturbed Wnt/PCP signaling affects transcriptional activity . . . 132

4.4 Discussion . . . 136

5 Summary and Outlook 141

5.1 Summary . . . 141

5.2 Outlook . . . 142

Chapter 1

Introduction

> Je suis de ceux qui pensent que la science est d'une grande beauté.
>
> Marie Curie

The discovery of cells with a simple light microscope dates back to 1665 when Robert Hooke examined thin slices of bottle cork. He observed regular patterns of cell walls and termed them 'cells' as they resembled monastic cells, which are inhabited by monks [Hooke, 1665]. Still, the mystery of cells remained unsolved for another 150 years. The role of the cell as fundamental unit of life and the fact that cells compose all organisms is self-evident today, however, was vividly debated in the 19^{th} century. Eventually, the cell theory was formulated in 1839. It is mostly accredited to Matthias Schleiden and Theodor Schwann [Schleiden, 1838, Schwann, 1839]. In 1858, Rudolf Virchow added the fundamental aspect that every cell arises from a pre-existing cell ("*Omnis cellula e cellula*"), thereby ignoring seminal publications on cell division from Robert Remak in 1852 [Mazzarello, 1999].
Ever since, biologists have described and classified cells by their morphology, shape, function, location and interaction with other cells. As such, every technological advance in light microscopy revealed a more detailed picture. Sulston et al. [1983] even reconstructed the lineage of every single-cell in *C. elegans*, a worm-like organism consisting of $\sim 1,000$ cells, via direct observation in a light microscope. Such endeavors are practically infeasible for higher organisms.
The possibility to study organisms and their genetic background revealed astonishing relations. For example, the avian family tree, which is based traditionally on morphologic features, was fundamentally rewired when the phylogenetic relationship were revealed through genome sequencing [Jarvis et al., 2014]. Analogously, a description of cells based on obvious parameters such as appearance or specific molecular markers lacks the comprehensive picture, as the same shape and morphology do not necessarily imply common function, molecular configuration or ancestry. Moreover, cell classification based on molecular markers is not systematic, especially due to the lack of systematic screening for the most specific marker of the chosen cell type. Nowadays, in order to gain more insights to

the complexity and redundancy in biological systems, the joint effort of interdisciplinary teams is essential. The following chapter sheds light on biological methods to study molecular properties of single cells, and on subfields that were of particular interest in this work. It must be noted that the mathematical methods to analyze high-dimensional data from single cells are reviewed in Chapter 2.
We first introduce the structure of the animal cell and how a cell stores and utilizes information. Then, we review experimental methods to study single cells with respect to cell differentiation and specialization. The subsequent section reviews the experimental techniques for single-cell transcriptomics, which is the particular focus of this thesis. Furthermore, we shed light on biological systems, which we studied in detail using single-cell transcriptomics.

1.1 The molecular organisation of the animal cell

The cell is the smallest unit in a living organism and is equipped to fulfill various functions. Cells are highly structured and organized [Alberts et al., 2008]. The substructures of a cell are called *organelles*. Generally, cells are classified by the presence or absence of a nucleus into *Eukaryota* and *Prokaryota*. Animals, plants and fungi belong to the domain of Eukaryota, while bacteria and archaea are Prokaryota.

A vital characteristic of life is to procreate. In particular, a cell produces two identical daughter cells through *mitosis* [Alberts et al., 2008]. A series of events called *cell cycle* is required for successful cell division (see Fig. 1.1a). The cell cycle is divided up into four phases: In *gap phase 1* (G_1 phase), which follows the mitosis, the cell is resuming its biosynthetic activity and cell growth. In the *synthesis phase* (S phase), the cell duplicates its DNA. In the following *gap phase 2* (G_2 phase), the cell has high biosynthetic activity and prepares for cell division. In mitosis (M phase), the cell divides and distributes the duplicated genetic material on the daughter cells. During this phase, the nucleus breaks down and biosynthetic activity is halted. In addition, cells may enter a *quiescent* state called *G_0 phase* from G_1 phase, where they do not divide. Cells with active cell cycle are called *proliferating*. Most cells in an organism, however, stay in G_0 phase.

Next, we focus on the structure of animal cells (see Fig. 1.1b). An animal cell has an outer membrane, which separates the cell plasm from its environment and acts as filter for nutrients or recognizes external stimuli. Further, every organelle of the animal cell is surrounded by a plasma membrane or in case of nucleus and mitochondria, by two plasma membranes. The nucleus contains the genetic material of the cell, called *DNA*. In higher

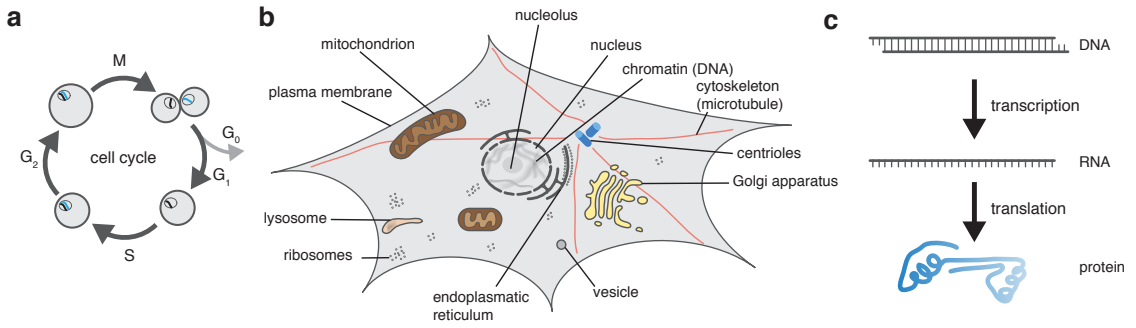

Figure 1.1: Cell structure and the molecular organization of genetic information. **a** The phases of the cell cycle. **b** The structure of an animal cell (adapted and simplified from Alberts et al. [2008]). **c** The central dogma of molecular biology describes the flow of information from DNA to protein.

organisms, the genetic material is organized as chromosomes. Furthermore, the DNA molecule is wrapped around histones, which enable the compaction of a genome to fit in the nucleus. The complex of DNA and histone proteins is called *chromatin*. Moreover, histones are modified to silence unneeded regions or to relax the compaction of active regions in the DNA.

Mitochondria provide chemically bound energy in the form of ATP[1], thereby consuming oxygen. The endoplasmatic reticulum (ER) is a complex network of interconnected tubes and sacs stretching through the whole cell plasm. Amongst others, the ER is involved in protein synthesis and modification, and lipid synthesis. Modified proteins and lipids are packaged into vesicles that bud from the ER and are transported to the Golgi apparatus. The Golgi apparatus is a series of stacked membrane disks called *cisternae*, which serves as sorting and dispatching station for the products from the ER. Furthermore, the Golgi apparatus is an integral part of the cellular vesicle transport. For example, it produces secretory vesicles and receives endocytotic vesicles from the outer membrane.

Lysosomes also take part in the vesicle transport system of the cell. These organelles encapsulate hydrolytic enzymes in an acidic environment and decompose macromolecules such as lipids and proteins. Vesicle transport requires a directed movement of various components, which is provided by the cytoskeleton. This cable-like structure starts from the centrioles and spans the complete cell plasm. Furthermore, the cytoskeleton determines cell morphology and enables cell movements.

Cells constantly produce and degrade proteins in order to maintain the cellular functions. Protein biosynthesis is a two-step process consisting of transcription and translation (see

[1] ATP stands for Adenosine-Triphosphate, the universal unit of energy in a living cell.

Fig. 1.1c). The blueprints for every protein, called *genes*, are encoded in the DNA, which serves as a long-term storage of genetic information in the cell. In eukaryotes, a gene consists of alternating segments of coding regions called *exons* and non-coding regions called *introns*. During transcription, the complete gene sequence is converted into a messenger RNA (mRNA) molecule. As a remark, DNA and RNA molecules have a head called *5' end* and a tail called *3' end*. Nucleotides[2] can only be added to the 3' end.
After transcription, a 5' cap protects the 5' end and the 3' end is extended by a poly-A-tail through *polyadenylation*[3]. In the process called *splicing*, the mRNA molecule is reduced to the exons. As a remark, splicing and polyadenylation are considered to be co-transcriptional, co-occurring processes [Misra and Green, 2016], but polyadenylation may be faster than splicing [Gray et al., 2014, Hendriks et al., 2014, La Manno et al., 2018].
Being processed, the mature mRNA molecule is transported to the cell plasm to the ribosomes. The ribosome is a protein-RNA-complex with several subunits that perform the translation of mRNA to protein. Here, the ribosome synthesizes proteins from 20 proteinogenic amino acids based on the mRNA template, where every amino acid is encoded by a nucleotide-triplet called *codon* [Alberts et al., 2008, Leder and Nirenberg, 1964]. As soon as the ribosome reaches a *stop codon*, both mRNA and protein are released.

The general transfer from DNA to protein via RNA is summarized in the *central dogma of molecular biology* [Crick, 1970], *i.e.* information cannot be transferred from protein into RNA or DNA, and the transfer of information from RNA to DNA was questioned. Shortly thereafter, the discovery of reverse transcriptase in retroviruses [Baltimore, 1970, Temin and Mizutani, 2004] effectively extended the central dogma of molecular biology and created the basis for RNA sequencing.

1.2 The study of single cells at the molecular level

To study cellular processes, researchers usually employ a targeted approach, *i.e.* few features are studied at once in thousands of cells. Experimental techniques that operate at single-cell resolution allow to study cellular heterogeneity as in response to a stimulus, gene regulation, cell fate decisions, and *de novo* discovery of new cell types. On the other hand, to study complex molecular traits in health and disease, so-called *'omics'* technolo-

[2] A nucleotide consists of a phosphate, a pentose sugar and a nucleobase. Four different nucleobases make up the DNA: adenine (A), guanine (G), thymine (T) and cytosine (C). In RNA, thymine is replaced by uracil (U). The pentose sugar is ribose in RNA and desoxyribose in DNA.
[3] The poly-A-tail is a long sequence of adenin nucleotides.

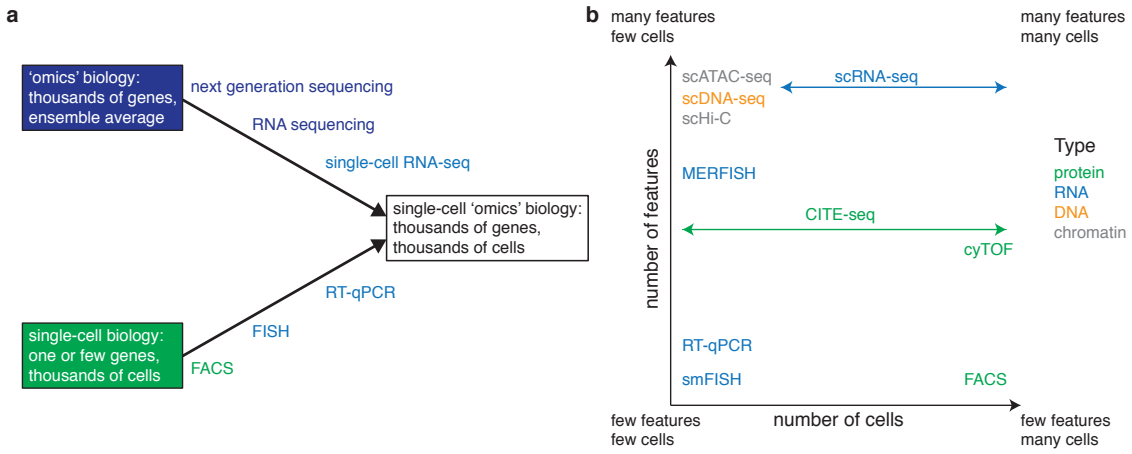

Figure 1.2: Development of single-cell 'omics' and experimental techniques for single-cell measurements. a Converging development of *'omics'* approaches from bulk to single-cell measurements and the upscaling in the number of features in single-cell biology. Figure adapted from Junker and van Oudenaarden [2014]. *'Omics'* approaches aim to cover the full feature space while targeted methods assess few features. **b** Method overview based on the number of measurable features and sample size per experiment. Methods are colored by the measured molecule type (see Marr et al. [2016] for comparison).

gies have been developed to measure all possible features such as all expressed genes in a cell ensemble (see Fig. 1.2a). The targeted approach of single-cell biology allows studying thousands of cells while 'omics' biology allows studying thousands of genes, proteins, etc. Combining 'omics' approaches with single-cell biology makes it possible to study all features of every single cell in complex tissues [Haber et al., 2017], organs [Baron et al., 2016, Byrnes et al., 2018], embryos [Wagner et al., 2018] or complete adult animals [Plass et al., 2018].
In this section, we review experimental techniques in single-cell biology that target the different layers of the molecular machinery (see Fig. 1.2b). It must be noted that experimental aspects of single-cell RNA-sequencing are discussed in detail in section 1.3.

In the past decades, the expression of a protein has been studied on the bulk level for example by Western blotting [Burnette, 1981, Towbin et al., 1979], mostly detached from its localization in the cell. In contrast, studying a gene or protein on the single-cell level returns a highly resolved and sophisticated picture. Finally, we embed cells in close context to its neighboring cells in a tissue [Eberwine et al., 2014]. For example, Elowitz et al. [2002] reported cellular heterogeneity on *E. coli* populations whose expression of a single gene varied highly due to stochastic expression bursts. In higher organisms, changes in the gene expression pattern are immediately accompanied by structural changes in chromatin

accessibility[4], histone modification and DNA methylation. These changes are of particular interest in the field of *epigenomics*.

One of the earliest single-cell techniques is fluorescently activated cell sorting (FACS) [Hulett et al., 1969, Julius et al., 1972]. Usually, antibodies coupled with a fluorescent label bind to specific surface markers of cells in suspension. Then, the suspension is applied to a flow cytometer, the fluorescent light signal is collected and the cells are sorted for subsequent use. Furthermore, a flow cytometer measures non-fluorescent features such as cell size, forward scatter and side scatter. Flow cytometry and cell sorting are well-established single-cell technologies with a huge palette of clinical and research applications [Han et al., 2016]. However, they rely on the principle of fluorescence, which is limited to ~ 12 colors due to spectral overlap [Autissier et al., 2010].
In contrast, single-cell mass cytometry (*cyTOF*) uses antibodies coupled with elemental isotopes and a mass spectrometer replaces the flow cytometer [Bendall et al., 2011]. Here, up to 100 different features can be distinguished in principle [Bandura et al., 2009]. Notably, cells are ionized and nebularized in a spray for measuring the antigen binding and thus subsequent experiments are impossible. More recently, Stoeckius et al. [2017] presented *CITE-seq*, where antibodies are labelled with short DNA-strands. Coupling this method with droplet-based scRNA-seq (see section 1.3), the number of measured features is only restricted by the antibody diversity and price, and the number of relevant surface molecules. Practically, CITE-seq provides as many features as cyTOF, but the antibody information can be directly read out via scRNA-seq, providing both protein and mRNA measurements in a single experiment. Nonetheless, mass cytometry has been shown to combine both mRNA and protein quantification recently [Schulz et al., 2018].

On the mRNA level, scRNA-seq quantifies the complete transcriptome of hundreds to thousands of single cells [Angerer et al., 2017] (see section 1.3). Before the advent of transcriptomic profiling, single-cell reverse-transcription quantitative PCR (RT-qPCR) enabled a sensitive and robust quantification of up to 96 genes in up to 96 cells [Bengtsson et al., 2005, Ståhlberg and Bengtsson, 2010]. In this method, mRNAs are first converted into cDNAs. Then, the cDNAs undergo the standard PCR amplification procedure with denaturation, annealing and elongation [Saiki et al., 1988]. After the elongation step, the number of cDNA molecules is quantified by the signal of a fluorescent dye, which binds to double-stranded DNA. As the number of DNA molecules doubles with every cycle, the fluorescent signal is increasing until it surpasses the background noise level. Thus, RT-qPCR returns a relative mRNA amount per gene and cell.
In contrast, fluorescence *in situ* hybridization (FISH) on the single molecule level (*i.e.*

[4]DNA regions without active transcription are compacted while regions with active transcription are open.

smFISH) allows counting mRNAs in a cell. FISH is a sensitive measure for a highly specific number of genes [Femino et al., 1998, Levsky et al., 2002]. Notably, cells are fixed before a fluorescently labelled DNA-probe hybridizes with present mRNAs, such that the spatial information of the cells within a tissue is preserved. Moreover, smFISH is the current gold standard to quantify gene expression in single-cells [Grün et al., 2014, Halpern et al., 2017, Svensson et al., 2017]. Similar to FACS, smFISH is restricted to ~ 12 colors due to spectral overlap. To overcome the limitation of fluorescence, Chen et al. [2015] recently developed the multiplexed error-robust FISH (*MERFISH*), where multiple hybridization and subsequent readout rounds combined with an error-robust labelling scheme allow to quantify thousands of mRNA species. Still, both smFISH and MERFISH depend on efficient imaging analysis to separate cells and quantify gene expression and in order to upscale these methods.

Next, we briefly illustrate scDNA-seq and chromatin structure analysis with scATAC-seq and scHi-C as representatives of the versatile epigenomics-toolbox to study cell diversity and plasticity [Schwartzman and Tanay, 2015]. Briefly, scDNA-seq plays a major role in the analysis of microbes to assess microbial diversity. Usually, this technique is applied to study non-cultivable species via whole-genome amplification [Marcy et al., 2007].
To study chromatin structure, the assay for chromatin accessibility combined with sequencing (ATAC-seq) reveals all open chromatin regions in the genome [Buenrostro et al., 2013, 2015a,b]. Open chromatin not only reveals sites with active gene expression, but also enhancer regions. Furthermore, ATAC-seq gives access to identify active enhancer regions. In complement the information of open chromatin, highly resolved chromatin conformation capture by scHi-C allows reconstructing the unique three dimensional chromosome architecture and within-chromosome connections [Nagano et al., 2013, Schwartzman and Tanay, 2015]. However, while ATAC-seq even resolves nucleosome positioning ranging $\sim 150\,bp$, Hi-C is coarse-grained with a resolution scale of mega basepairs [Dixon et al., 2012, Lieberman-Aiden et al., 2009].

1.3 The technology of single-cell RNA-sequencing

The analysis of single-cell transcriptomes was first demonstrated on seven single cells in 2009 [Tang et al., 2009] and increased to over 1 million cells measured by the commercial platform *Chromium*^{TM} (10X Genomics) (see Fig. 1.3). Early single-cell mRNA sequencing (scRNA-seq) studies focussed on cells from early embryonic development due to the very limited amount of biological input material [Tang et al., 2009, 2010]. Steady technological advances allowed to reduce sample volume from microliter to nanoliter volume thereby

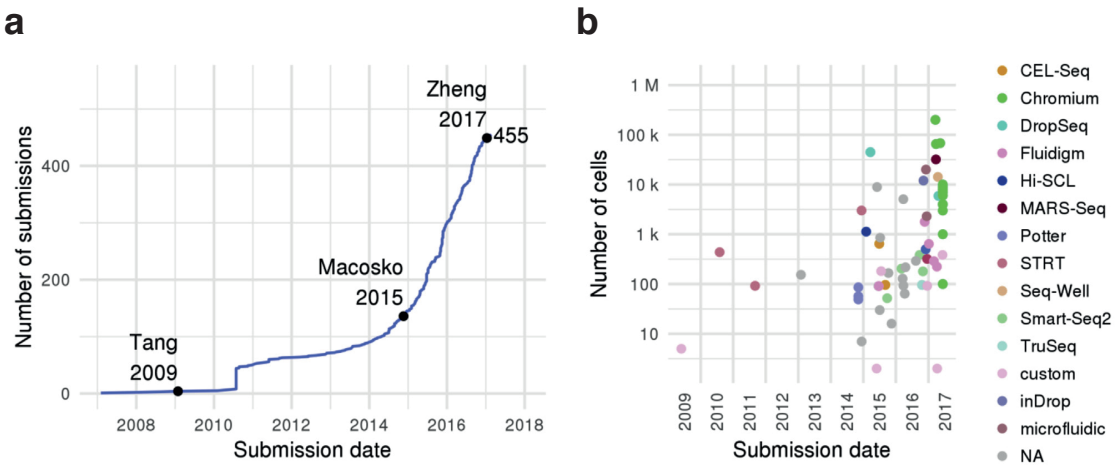

Figure 1.3: Increasing availability (**a**) and sample size (**b**) of single-cell RNA-seq data. Figure adopted with courtesy from Angerer et al. [2017].

increasing sensitivity [Wu et al., 2013], increasing sample size [Jaitin et al., 2014, Klein et al., 2015, Macosko et al., 2015] and reducing costs [Hashimshony et al., 2016, Ziegenhain et al., 2017].

Single-cell RNA-seq, is performed in the following steps. First, the cells are isolated and lysed. Then, as many mRNAs as possible are reverse transcribed into DNA (see Fig. 1.4). Specifically, the primer for reverse transcription has a poly-T sequence to select for the poly-A tail of mRNAs. Then, a second strand is synthesized and all DNA fragments are amplified. Finally, all fragments are sequenced. The resulting reads are aligned to the reference genome (*i.e.* to known exons).

In general, we distinguish scRNA-seq protocols by protocol platform and quantification strategy. As platform serves either a microwell plate, a microfluidics chip or a nanoliter droplet device (see. Fig.s 1.4 and 1.5). The second important distinction is the quantification strategy, which is closely connected to the gene body coverage. So-called *full-length* methods cover most of the gene body [Picelli et al., 2013, Ramsköld et al., 2012], while 3'-tag counting methods cover mainly the 3'-end of a transcript (*i.e.* its end). Even though we may observe a 3'-bias also in full-length methods [Ziegenhain et al., 2017], the coverage of the complete gene body allows to study e.g. splicing variants [Kolodziejczyk et al., 2015b].

Tag-counting methods become more popular with the idea of transcript counting using *unique molecular identifiers* (UMIs) [Islam et al., 2014]. During the amplification step, we create several copies of the same transcript and cannot distinguish a unique transcript

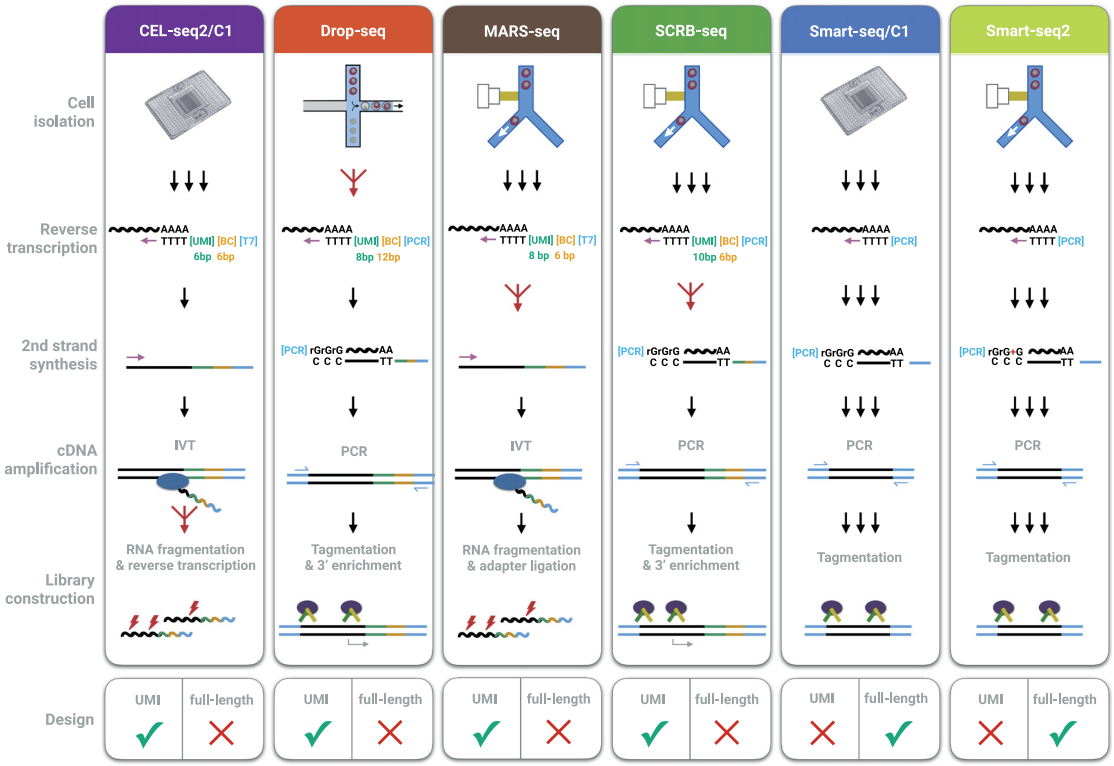

Figure 1.4: Schematic of library preparation in single-cell RNA-seq analysis for selected protocols. The general workflow is the same for all depicted methods, variations are described in the text. Figure adopted with courtesy from [Ziegenhain et al., 2017].

from its copies. Thus, amplification contributes to the noise for each gene. In contrast, a UMI acts as ID for a transcript. It is a short random sequence of $6-10$ bp that is directly ligated to the poly-T primer sequence. When all sequences are aligned to the respective gene, we count the number of different UMIs. Thus, we discard copies of the same mRNA transcript and quantify only unique transcripts. Based on the UMI sequence length, we can distinguish $2^{12}-2^{20} \approx 10^3 - 10^6$ transcripts.

We consider two different methods for amplification of DNA fragments. The 'classical' fashion is to use the polymerase chain reaction (PCR) with a thermostable enzyme [Saiki et al., 1988]. This enzyme approximately doubles the input DNA per cycle, where the efficiency is sequence-dependent [Chakrabarti and Schutt, 2001]. The second method is called *in vitro transcription* (IVT) [Eberwine et al., 1992]. IVT linearly increases the DNA fragments and has been described a few years after PCR, but needed some optimization to meet the low amount of mRNA input [Hashimshony et al., 2012].

Finally, a DNA library is created by adding sequencing-specific adapters (*tagmentation*,

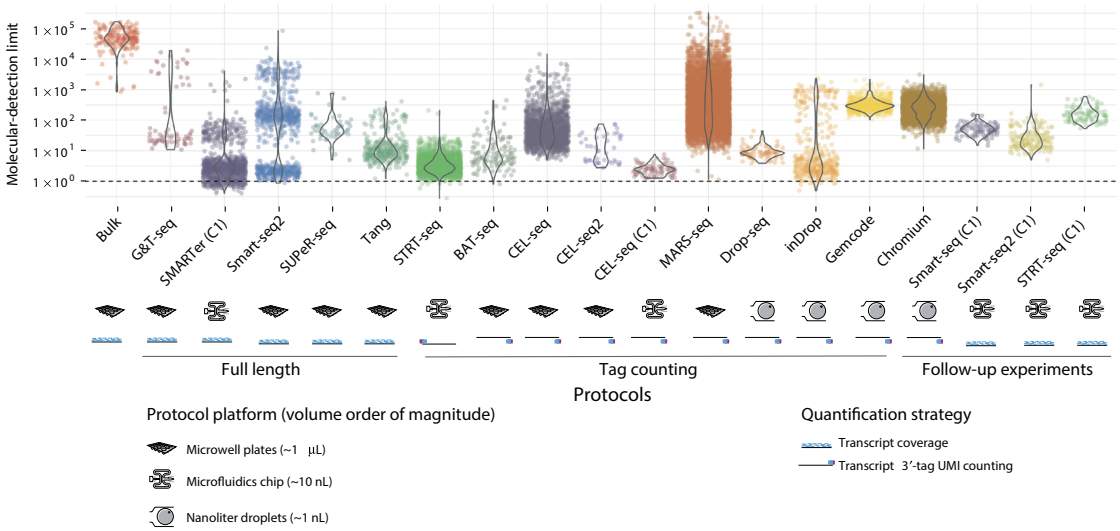

Figure 1.5: Sensitivity of single-cell RNA-seq protocols (*i.e.* the molecular detection limit). Sensitivity is defined by a logistic regression model based on the detection rate of technical (spike-in) genes. The more sensitive the method, the lower the molecular detection limit. Figure adopted with courtesy from Svensson et al. [2017].

e.g. Nextera, Illumina). Then, the libraries are sequenced. If the libraries were pooled before sequencing, the resulting read-out data file contains the sequences for all fragments per cell or per data set. Otherwise, we obtain one data file with sequences per cell.

Variations in the workflow, such as experimental platform, and reaction volume affect the mRNA capture in the reverse transcription step [Svensson et al., 2017, Wu et al., 2013]. An important parameter for single-cell RNA-seq is the sensitivity of the methods (see Fig. 1.5). The molecular detection limit is inverse proportional to the sensitivity of a method. Bulk RNA-seq has a comparatively low sensitivity, which is compensated by the high amount of input material of several thousands of cells. In contrast, single-cell methods require a high sensitivity as the input material is considerably lower compared to bulk RNA-seq. A sensitivity of 10 % corresponds to a molecular detection limit of 10. Hence, 1 out of 10 copies of the same gene transcript is recorded by the method — Svensson et al. [2017] have investigated the sensitivity of available single-cell RNA-seq methods to report few methods with a molecular detection limit between 1 and 10 (and corresponding sensitivity of $10-100\,\%$). On the one hand, these methods have reached a high sensitivity already. On the other hand, they also suffer from variability in the capture rate. This issue increases the loss of mRNAs and the so-called *dropout* rate of a gene.
The term *dropout* defines the phenomenon of observing a zero entry for a gene. These

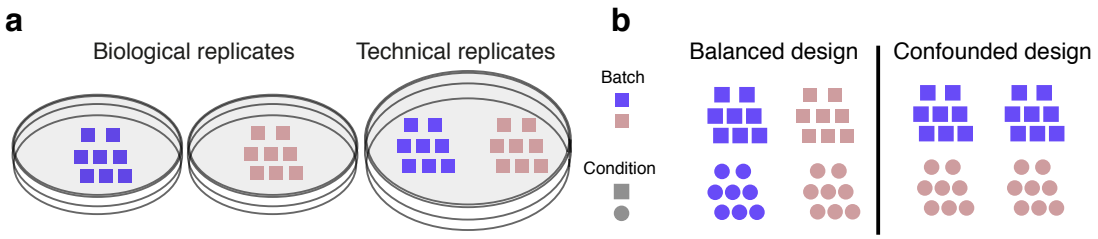

Figure 1.6: Batch types and experimental design. Estimating the batch effect in single-cell RNA sequencing data (Figure is adapted from Büttner et al. [2019]). **a** Biological and technical replicates have different origins. Technical replicates are derived from the same biological samples (in this case cell cultures), while biological replicates are independent samples. **b** Experimental designs: A balanced design allows one to separate technical and biological sources of variation, while a confounded design mixes both.

zeros have several possible origins: A gene might be not expressed at all, or it is not expressed at the moment of quantification due to stochastic or bursty gene expression as reported by Elowitz et al. [2002]. If the gene is expressed, but was not observed, it was not captured in the experiment. All these sources result in the same observation, the zero entry for a gene, but should be treated differently. However, we cannot distinguish non-expressed genes from non-captured genes *a priori*.

In order to capture the increase on noise in scRNA-seq, Stegle et al. [2015] have encouraged the use of spike-ins. Spike-ins are a mix of synthetic, polyadenylated RNAs designed by the External RNA Control Consortium (ERCC). These genes do not map to the reference genome of the organism and cover different lengths, GC-contents, and initial concentrations. The downside is that technical variance of ERCC spike-ins may differ from the variation of biological samples [Risso et al., 2014]. Further, some very popular methods as droplet-based sequencing (e.g. drop-seq [Macosko et al., 2015], inDrop [Klein et al., 2015] and the commercial platform *Chromium*^{TM} (10x Genomics)) do not allow the use of spike-ins.

1.4 Computational challenges for single-cell RNA-seq data integration

With the variation in sensitivity across protocols, we already glimpsed at another important aspect: we have to control for the systematic bias arising from repeated measurements

and consider those effects in the experimental design of our study [Büttner et al., 2019][5]. Usually, data collection by experimental measurements is distributed over several experimental runs, so-called *batches*. We term any variation that emerges through technical effects arising when samples are handled in distinct batches as *batch effects*. In general, batch effects occur if an experiment is repeated with biologically equivalent cells (e.g. different patients of the same disease) or technically equivalent cells (e.g. sequencing cells of the same culture condition on subsequent days) as depicted in Fig. 1.6a. Both biological and technical variations contribute substantially to the total variability in single-cell RNA-sequencing (scRNA-seq) data. In a balanced design for a sequencing experiment, we are able to identify and distinguish biological from technical variation (Fig. 1.6b).
In contrast, a confounded design groups cells of the same condition into the same sequencing runs, and separates biologically distinct cells into entirely distinct handling and sequencing runs. This confounds biological with technical variability. In such a setup, a worst case scenario might be that the technical variation swamps out the biological variation. Accounting for technical factors in a dataset is a key step in the pre-processing workflow after cell and gene filtering [Lun et al., 2016b]. Moreover, it impacts the selection of 'potentially interesting genes'. As such, the choice of 'interesting genes' removes noise from the data but also defines the potential outcome of data analysis. If balanced design is not possible e.g. by chip design, it is recommended to create several replicates per biological condition [Tung et al., 2017]. The need to study batch effects becomes more pressing when we aim to integrate data from different platforms, tissues and individuals as the Human Cell Atlas aspires to. We review various approaches for single-cell RNA-seq data analysis in Chapter 2.

Furthermore, upscaling sample size in scRNA-seq draws particular attention to the computational aspects of data analysis. With increasing number of cells and features, the computational requirements for data storage and processing is almost constant, however, the actual bottleneck is the data analysis [Angerer et al., 2017]. In particular, when relating all cells to one another, the complete data matrix is loaded into the computer memory in classical data analysis. This may become infeasible with even larger datasets. Moreover, due to global efforts such as the Human Cell Atlas [Regev et al., 2017], we are able to access huge amounts of data, stressing the need for scalable methods to data integration. Taken together, algorithmic complexity, parallelization and efficient implementation are key instruments to the critical step of data analysis [Angerer et al., 2017, Wolf et al., 2018a].

In total, single-cell RNA-seq has become a reliable method for transcriptional profiling

[5]The following section has high similarity to the introduction of Büttner et al. [2019]

of single cells. Still, the impact of sensitivity and dropout to the signal-to-noise ratio and systematic bias per measurement demands for distinctive statistical approaches that deal with high-dimensional data and effective feature extraction in order to exploit these data for knowledge discovery. Thus, the huge prospectives of single-cell RNA-sequencing technology is coupled to the advances and standardization in data analysis.

1.5 New insights to development and lineage specification through scRNA-seq

The technology of scRNA-seq allows transcriptomic profiling in thousands of cells and is widely applicable in cell biology research. Furthermore, scRNA-seq offers a number of advantages over bulk RNA-seq as it facilitates the study of the stochastic nature of gene expression, cellular heterogeneity, rare cell types, and cell lineage specifications during development. For example, embryonic development in the mouse involves only few cells and the gene expression pattern of oocytes to blastocysts is the key to understand pluripotency. The full scale of it is only accessible via scRNA-seq and has been studied extensively [Biase et al., 2014, Boroviak et al., 2015, Deng et al., 2014, Fan et al., 2015, Goolam et al., 2016, Liu et al., 2016, Tang et al., 2009, Wu et al., 2016, Xue et al., 2013]. Furthermore, during early stages of embryonic development, cells differentiate into three germ layers. Mouse embryonic development and scRNA-seq technology has been reviewed in detail in Pijuan-Sala et al. [2018]. In particular, the emergence of the distinct cell types was successfully characterized through computational modelling by Gaussian Process Latent Variable Models [Buettner and Theis, 2012] and diffusion maps [Haghverdi et al., 2015].

Detecting rare cell types is particularly important when it comes to distinguish healthy tissue samples from cancer [Sandberg, 2014]. For example, Grün et al. [2015] developed the tool *RaceID* to identify rare cell populations in scRNA-seq data. However, the number of cells per sample has to be increased accordingly to detect rare cell types according to their expected abundance[6]. To date, a clinical application of scRNA-seq is not available, but with decreasing sequencing costs per cell, increasing throughput per samples, automated pipelines for sequencing, and increasing standardization in analysis, translation into the clinics is quite imaginable.

The recent advances in upscaling scRNA-seq experiments and corresponding data analysis

[6]Rahul Satija's lab provide a simple web application to create a sample size estimation for scRNA-seq experiments (https://satijalab.org/howmanycells). Here, they use a negative binomial model to estimate the sample size in order to detect at least n cells with a 95 % confidence level.

methods allow to sequence and analyze full organisms such as fruit fly embryos [Karaiskos et al., 2017], zebrafish embryos [Wagner et al., 2018], and *C. elegans* [Cao et al., 2017]. One astonishing example for the capabilities of the sequencing technology presented [Plass et al., 2018]. The authors sequenced a complete adult planaria animal, annotated all cell types and derived a lineage tree for all populations using graph abstraction (PAGA) [Wolf et al., 2018b]. These advances combine large-scale experiments with potent computational models. Hence, the extensive implications of scRNA-seq for medicine and fundamental cell biology research have been recognized as the *Breakthrough of the Year 2018* by the *Science* magazine [Pennisi, 2018].
All in all, the experimental methods are ready to tackle a global project even larger than the Human Genome Project in the late 1990's — the Human Cell Atlas aims to chart all cell types in the human body [Regev et al., 2017]. However, data analysis is still a bottleneck, but key to interpret the biological data. In particular, the numerous challenges such as data integration from different labs and platforms, automated cell type annotation, lineage inference in human development and tissue regeneration demand an equally huge contribution from the computational side. An atlas with all cell types in all tissues including cell-to-cell communication and interaction, molecular signaling, developmental processes, and regulatory networks certainly reveals tremendous insights in healthy human tissue and furthers novel insights into physiological imbalance of a diseased state.

In the following, we introduce two model systems, which we studied in detail in this thesis. First, we briefly discuss mouse embryonic stem cells (mESCs), which are among the best characterized cell types and have propelled stem cell research to exceptional opportunities. Second, we review the composition and function of the small intestinal epithelium, which renews every 4-5 days throughout life and serves as astounding example in balanced proliferation, differentiation and regenerative abilities.

1.5.1 Mouse embryonic stem cells

Mouse embryonic stem cells (mESCs) are pluripotent cells derived from the epiblast of a mouse embryo in blastocyst stage E4.5 (*i.e.* 4.5 days after fertilization of the oocyte) [Evans and Kaufman, 1981, Martin, 1981] and cultured using various feeding conditions *in vitro* [Czechanski et al., 2014, Rathjen et al., 1999, Ying et al., 2008]. In particular, mESCs have the potential to form all tissues in the mouse given a close control on the differentiation process [Czechanski et al., 2014]. In contrast to cancer cell lines, they have a normal karyotype [Evans and Kaufman, 1981]. mESCs serve as model system to study differentiation and lineage commitment, cellular signaling and reprogramming, disease modeling, and potential therapeutic applications [Boroviak et al., 2015, Srivas-

tava and DeWitt, 2016, Wu and Belmonte, 2016, Young, 2011]. In the field of single-cell transcriptomics, mESCs have become a reference cell type on which single cell RNA-seq protocol comparisons and technological advances are demonstrated [Grün et al., 2014, Islam et al., 2014, Klein et al., 2015, Kolodziejczyk et al., 2015b, Semrau et al., 2017, Vallejos et al., 2017, Ziegenhain et al., 2017].
For example, single-cell RNA-seq enabled to study gene expression profiles and cell cycle activity in different culture conditions [Kolodziejczyk et al., 2015b]. Further, measuring several samples of a differentiating mESC culture allows to detect the emergence of new cell types and a direct access to cell fate decisions in early development [Klein et al., 2015, Semrau et al., 2017]. In this thesis, we study normalization and batch correction algorithms for scRNA-seq on the example of mESCs (see Chapter 3).

1.5.2 Lineage specification in the small intestinal epithelium

The intestinal epithelium forms an outer layer to the lumen of both small and large intestine. Its functions are to absorb nutrients, interact with microbes and orchestrate immune responses. The daily turnover rate is approx. 10^{11} cells, which corresponds to 200 g of cells all being replenished from a small stem cell population. The anatomic structure of the intestinal layer shows a distinctive adaptation to enlarge the surface that optimizes nutrient uptake. One of these adaptations is the separation into villi that point towards the intestinal lumen and crypts of Lieberkühn where intestinal stem cells (ISCs) reside. During the differentiation process, cells move upwards to the villus while passing a transit amplifying (TA) zone [van der Flier and Clevers, 2009]. We distinguish six different cell types originating in the intestinal crypt - enterocytes, Paneth cells, enterendocrine cells (EECs), goblet cells, tuft cells (or brush cells) and microfold cells (M cells) [Haber et al., 2017, van der Flier and Clevers, 2009] (see Fig. 1.7).

The most abundant cell type in the intestinal lining is the enterocyte, whose function is absorption of nutrients as sugars, fatty acids and amino acids, and their transport across the epithelium. These cells make up the vast majority of the epithelium and form the absorptive lineage.
The other main cell types belong to the secretory lineage. Goblet cells earned their name from the goblet-like shape of the cell - its apical part is shaped like a cup and is filled with secretory vesicles. Their main function is to secrete gel-forming proteins (mucins, especially Muc2) and trefoil factors (e.g. Tff3) into the lumen, which in turn alleviate movement of the gut contents.
Tuft cells are chemosensory cells and have been shown to play a role in the response to parasite infection [Haber et al., 2017, Harris, 2016, Howitt et al., 2016].

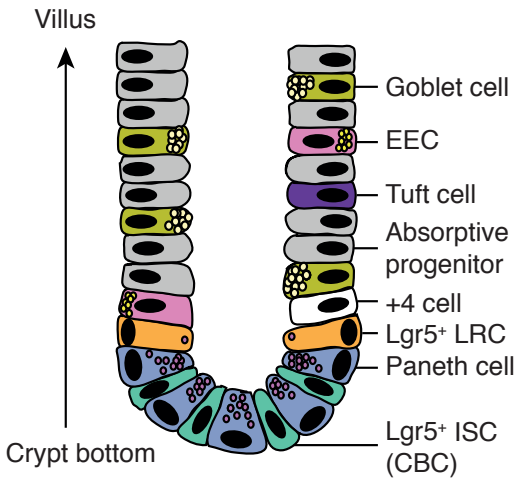

Figure 1.7: Scheme of a small intestinal crypt. Intestinal stem cells (ISCs) reside in the crypt bottom together with Paneth cells. Other cell types move up towards the villus while differentiating.

Microfold cells (M cells) cover the surface of lymphoid follicles and direct gut specific antigens to immune cells [Kucharzik et al., 2006]. They are involved in the development of chronic intestinal inflammation, as Crohn's disease and inflammatory bowel disease.
Enteroendocrine cells (EECs) produce a variety of hormones regulating e.g. blood glucose levels or satiety and make up approx. 1 % of the epithelial cells. The variety of subtypes in EECs is classically determined by principally secreted hormone [Gribble and Reimann, 2016]. Recently, Haber et al. [2017] have characterized subtypes of EECs based on gene expression profiling.
Paneth cells reside in the crypt bottom together with the intestinal stem cells. They produce antimicrobial peptides and enzymes such as defensins and lysozymes. In this way, Paneth cells protect the intestinal epithelium and, in particular, the crypt region from harmful bacteria but also shape the commensal gut microbiota. Paneth cells also secrete Wnt3 and express the Notch ligand Dll4 and thus might regulate stem cell maintenance [Sato et al., 2011]. Interestingly, Paneth cells move downward again after differentiation [Bjerknes and Cheng, 1981]. Paneth cells are long-lived (approx. $6-8$ weeks) in contrast to all other cell types, which have a life-span of $4-7$ days [Bjerknes and Cheng, 1981].
Intestinal stem cells (ISCs) reside in the intestinal crypt. ISCs are characterized by high expression of canonical Wnt target genes such as Lgr5, Axin2, Olfm4 and Ascl2 [Barker et al., 2007, Schuijers et al., 2015]. These cells fully replenish the intestinal epithelium.

The pool of ISCs consists of several subtypes. Most ISCs are described as rapidly cycling [Barker et al., 2007]. Apart from rapidly cycling ISCs, a quiescent *Lgr5*$^{+}$ stem cell resides in '+4' position [Buczacki et al., 2013, Muñoz et al., 2012].

To date, the transition and maturation stages of the known cell types in the gut remain elusive. We contribute to the debate with an extensive single cell study (see Chapter 4).

1.6 Research questions and thesis overview

The advent of single-cell RNA-sequencing revolutionizes cell biology research and biomedicine. In recent years, the availability of single-cell transcriptomics data, the commercialization of experimental platforms, cost reduction and tremendous increase in sample size make it possible to monitor almost every aspect of cellular development, cell identity and heterogeneity, and lineage commitment early on. For example, we are gaining insights to cellular processes in severe diseases such as diabetes and to the molecular changes caused by obesity at unprecedented scope and resolution. Computational method development is similarly accelerated by the speed and emerging opportunities of experimental advances. Especially, this technology demands for fast and accurate analysis of the large-scale and high-dimensional data.

This thesis addresses the question how to integrate scRNA-seq data in a robust and reliable way (see Fig. 1.8). At the same time, data integration has to preserve biological variability while removing as much technical noise as possible. Addressing these goals is essential to establish a Human Cell Atlas [Regev et al., 2017]. In particular, such an atlas relies on reproducible results over several labs and experimental platforms, and computational methods to establish a common reference from a plethora of datasets. However, existing test measures to test if data integration was successful are biased or lack sensitivity. In addition, batch effect correction models for data integration have not been investigated in a systematic analysis. Therefore, we developed an unbiased, robust and sensitive measure to quantify the performance of data integration models in general and batch effect correction in particular.
Furthermore, disease modelling requires a detailed understanding of the molecular mechanisms in healthy tissues. Here, we discuss lineage commitment and tissue regeneration in the small intestinal epithelium. These cellular processes are not fully understood and scRNA-seq is well suited to characterize lineage commitment and the underlying transcriptional changes. Understanding lineage commitment in the gut will help to develop new treatments for chronic diseases, to understand the changes of the epithelium in obesity

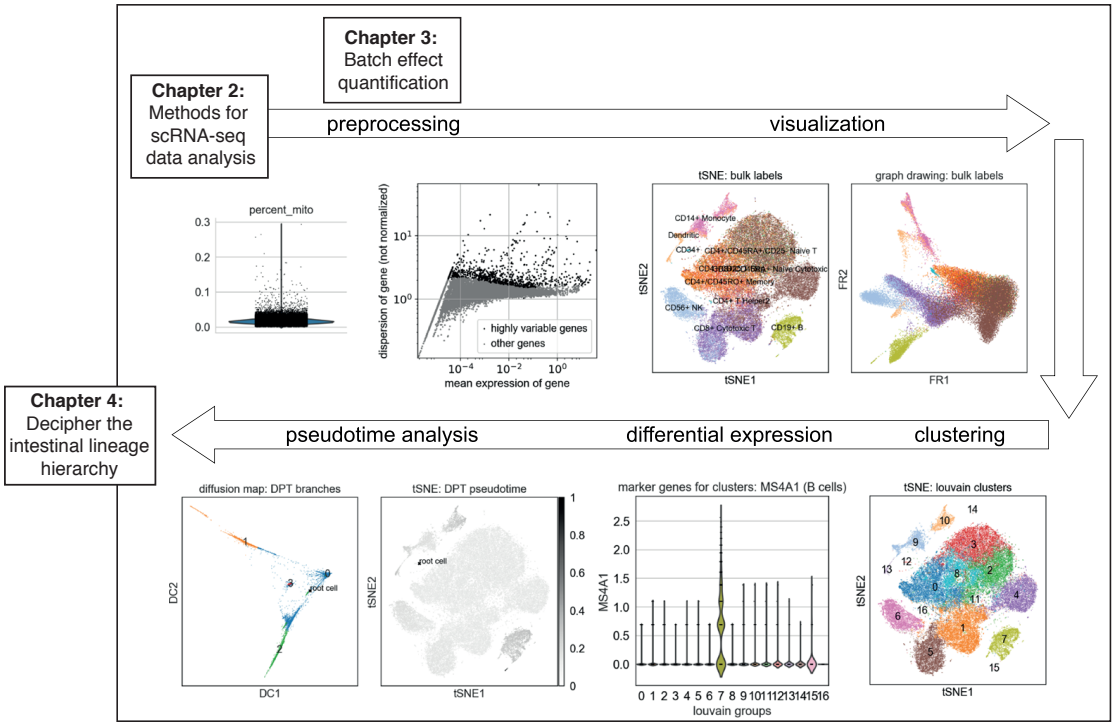

Figure 1.8: Organisation of this thesis illustrated on the basis of the `Scanpy` analysis pipeline. Chapter 2 introduces the important steps of scRNA-seq data analysis. Chapter 3 describes a novel approach to quantify batch effects. Chapter 4 presents evidence for direct lineage allocation in the small intestinal epithelium based on scRNA-seq data. Figure adopted with courtesy from Wolf et al. [2018a].

and shed light on the regulatory crosstalk of intestinal epithelium and pancreas. Gaining insight to such aspects is an important step to understanding the molecular mechanisms that connect obesity and diabetes. Therefore, we study lineage specification of the small intestinal epithelium in the mouse model with scRNA-seq.

In Chapter 2 we introduce general methods and approaches to single-cell RNA-seq data processing and analysis. Computational modelling has not been able to fully exploit the complexity of gene expression data [Marx, 2013]. Specifically, scRNA-seq analysis models are confronted with sparse matrices, which are ill-conditioned due to $\sim 10,000$ observed genes and only up to $\sim 1,000$ cells [Angerer et al., 2017]. In order to standardize scRNA-seq data analysis for large scale data integration, we also address the advantages and disadvantages of the methods.

All methods that play a role in the following chapters are discussed in detail. In particular, we focussed on data normalization and mathematical models for batch effect correction.

These steps are crucial to reduce cell specific and systematic bias. However, some high-level analysis do not apply for every data set. Instead, they are applied depending on the research question and data structure.

In Chapter 3 we aim to establish best practices for the analysis of scRNA-seq data. Briefly, we describe a new method to detect and quantify batch effects in scRNA-seq data: the k-nearest neighbor batch effect test (*kBET*). We compare its ability to performance against established measures for batch effects and use it to benchmark the a number of normalization and batch effect correction strategies. Furthermore, we explored the applicability of kBET to assess inter-individual variations. We established the project in collaboration with the group of Sarah A. Teichmann from the Wellcome Trust Sanger Institute, Hinxton, United Kingdom. A manuscript was recently published at *Nature Methods*: **M. Büttner**, Z. Miao, F. A. Wolf, S. A. Teichmann, and F. J. Theis. *A test metric for assessing single-cell RNA-seq batch correction.*

In Chapter 4 we aimed to understand stem cell lineage differentiation and lineage commitment of the small intestinal epithelium. Here, our experimental partners from the Institute of Diabetes Research (IDR) have developed a mouse model that expresses a Wnt/planar cell polarity (PCP) reporter. The reporter specifically labels the enteroendocrine and Paneth cell lineage. Together with short-term lineage tracing experiments and single-cell transcriptomic profiling of $\sim 75,000$ single cells, we conduct a robust study on the lineage hierarchy of the small intestinal epithelium. First, we characterized all major cell types and their progenitors of the small intestinal epithelium based on transcriptomic profiling. Moreover, we described a subset of intestinal stem cells being primed towards the Paneth cell lineage and disproved the existence of a bipotent Paneth-goblet cell progenitor. Second, we studied perturbed Wnt/PCP signaling in a heterozygous *Celsr1* deficient mouse strain to understand the role of non-canonical Wnt signaling in the formation of Paneth cells.
The author of this thesis conducted a complete scRNA-seq data analysis integrating data from several studies of the same lab. Furthermore, the existing computational models were adapted to the situation where appropriate. A manuscript is currently in revision: A. Böttcher, **M. Büttner**, S. Tritschler, M. Sterr, A. Aliluev, I. Burtscher, S. Sass, M. Irmler, J. Beckers, C. Ziegenhain, W. Enard, A. C. Schamberger, F. M. Verhamme, O. Eickelberg, F. J. Theis, and H. Lickert *Wnt/planar cell polarity primed intestinal stem cells directly differentiate into enteroendocrine or Paneth cells.*

In Chapter 5 we summarize the results and biological insights that emerged during the preparation of this dissertation. Furthermore, we shed light on future developments in the

field and potential extensions of the presented projects and methods.

1.7 List of contributed articles

The author of this dissertation has contributed to the preparation and publication of several articles. The contributions resulted in the following co-authorships:

- P. Angerer, L. Haghverdi, **M. Büttner**, F. J. Theis, C. Marr, and F. Buettner. *destiny: diffusion maps for large-scale single-cell data in R.* Bioinformatics, 32 (8):1241-1243 (2016)
- L. Haghverdi, **M. Büttner**, F. A. Wolf, F. Buettner, and F. J. Theis. *Diffusion pseudotime robustly reconstructs lineage branching.* Nature Methods, 13, 845-848 (2016)
- **M. Büttner***, Z. Miao*, F. A. Wolf, S. A. Teichmann, and F. J. Theis. *A test metric for assessing single-cell RNA-seq batch correction.* Nature Methods, 16, 43-49 (2019)
- A. Böttcher, **M. Büttner**, S. Tritschler, M. Sterr, A. Aliluev, I. Burtscher, S. Sass, M. Irmler, J. Beckers, C. Ziegenhain, W. Enard, A. C. Schamberger, F. M. Verhamme, O. Eickelberg, F. J. Theis, and H. Lickert *Wnt/planar cell polarity primed intestinal stem cells directly differentiate into enteroendocrine or Paneth cells.* Manuscript in revision
- E. Mameishvili, I. Serafimidis, M. Lesche, S. Iwaszkiewicz, S. Reinhardt, **M. Büttner**, D. Stellas, A. Papadimitropoulou, M. Szabolcs, K. Anastassiadis, A. Dahl, F. J. Theis, A. Efstratiadis, and A. Gavalas. *Aldh1b1 expression marks adult stem cells and is required for Kras-induced pancreatic cancer.* Manuscript submitted
- S. Tritschler*, **M. Büttner***, D. S. Fischer, M. Lange, V. Bergen, H. Lickert, and F. J. Theis. *Concepts and limitations for learning developmental trajectories from single-cell genomics.* Manuscript submitted

* indicates equal contribution.

Chapter 2

Methods for single-cell RNA-seq data analysis

> Mathematical science shows what is. It is the language of unseen relations between things. But to use and apply that language, we must be able fully to appreciate, to feel, to seize the unseen, the unconscious.
>
> Ada Lovelace

This thesis focuses on analysis methods for single-cell transcriptomic data. This data comprises quantitative reverse transcription polymerase chain reaction (RT-qPCR) and RNA sequencing (RNA-seq) data. These data types require different handling as qRT-PCR returns relative mRNA levels and RNA-seq returns total mRNA counts (see Chapter 1). Hence, RT-qPCR differs from RNA-seq in terms of normalization, but concepts such as batch effect correction, dimensionality reduction, clustering and pseudotime are applicable for both data types.
In this chapter, we introduce an analysis workflow and give an overview on common methods. In particular, we review common normalization and batch effect correction methods, as we studied their performance to remove batch effects from scRNA-seq data in Chapter 3. Whenever we employed a method in the subsequent chapters, we elaborate these methods in detail. There, we start the workflow from a count matrix and neglect the preceding sequence alignment. The analysis itself divides up into quality control, preprocessing and high-level analysis (see Fig. 2.1).

Currently, analysis gold standards for scRNA-seq do not exist. Instead, analysis workflows differ depending on the experimental setup, the single-cell measurement protocol (see section 1.3), and the cell type composition. Even within these categories analysis workflows vary as new tools become available. To organize the huge amount of new tools being

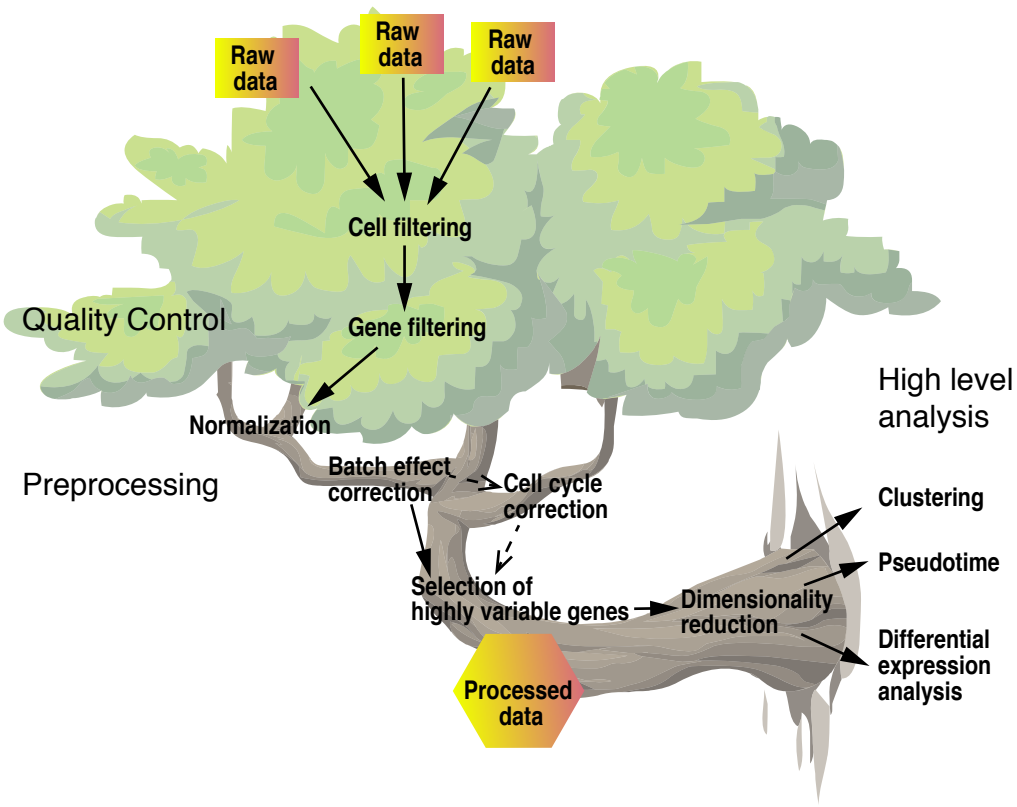

Figure 2.1: A generic analysis workflow for scRNA-seq. We distinguish three parts, depicted as leafage, trunk and root of a tree. First, raw data enter quality control steps. Then, data is further processed by normalization and correction for technical variations such as batch effects. Cell cycle correction is optional. After the selection of highly variable genes and dimensionality reduction (for visualization), the processed dataset is ready for high level analysis. Here, we present clustering, pseudotime and differential gene expression analysis. These tasks are performed depending on the dataset.

published, Zappia et al. [2018] provided a scRNA-seq tool database, which catalogues new and existing tools, and records the emergence of new approaches to scRNA-seq analysis. In fact, different methods lead to different results. In order to ensure reliability and reproducibility of results, we recommend to test different methods to check the robustness of a result. For example, the results of a differential gene expression analysis can be considered robust, if at least two different differential expression tests produce roughly the same result.

Several analysis platforms exist, which facilitate pipeline construction for scRNA-seq analysis. For example, `scater` [McCarthy et al., 2017], provides low-level pre-processing and visualization capabilities. Further, `PAGODA` [Fan et al., 2016] and `Seurat` [Butler

et al., 2018] cover the whole processing workflow including differential expression tests. `CellRanger` [Zheng et al., 2017] is specifically designed for large-scale data from the droplet-based platform *Chromium (10X Genomics)* and includes read alignment to the genome. While `PAGODA` and `Seurat` are written in R and are limited in performance, `Scanpy` [Wolf et al., 2018a] utilizes Python with a remarkable increase in speed and memory efficiency. Recently, Lopez et al. [2018] presented `scVI`, which combines statistical and deep learning models in a single cell analysis pipeline.

To describe the methods for scRNA-seq data analysis, we use the following notation unless stated otherwise.

Definition 2.1 (Notation)**.** *A scRNA-seq experiment with n cells and g genes (*i.e. *features) yields a full gene expression dataset $\mathbb{D} = \{x_1, \ldots, x_n\}$, where $x_i \in \mathbb{R}^g \; \forall i \in \{1, \ldots, n\}$ is the observation vector for cell i. Further, G denotes the set of genes. Then, $\mathbf{X} \in \mathbb{R}^{n \times g}$ denotes the corresponding gene expression data matrix. The entries in the data matrix $\mathbf{X}$ are count data $C_i g$ for cell i and gene g. Further, $\mu_g \in \mathbb{R}$ denotes the mean gene expression for gene $g \in G$.*

2.1 Cell filtering

The first step in quality control is to identify and remove low quality cells to reduce bias from technical sources. In particular, we define cells as 'low quality' when they are doublets (*i.e.* two cells instead of one), empty wells/droplets, or dying or damaged during cell isolation. Furthermore, failed reverse transcription or amplification leads to low quality cells as we expect few total counts from these cells. When RNA-seq is performed on microwell plates or microfluidics chips, cells can be imaged prior to cell lysis and mRNA quantification. These images can serve to classify cells into high or low quality. For example, Ilicic et al. [2016] trained a classifier based on images to identify specific features in the transcriptome of low quality cells. In some protocols, spike-ins are added to each cell. A spike-in is a mix of synthetic, polyadenylated RNAs as, for example, designed by the External RNA Control Consortium (ERCC) that do not map to the reference genome and that cover different lengths, GC-contents and initial concentrations. These shall provide a comparable scale for gene expression quantification when UMIs are inapplicable in an experiment. However, technical variance of ERCC spike-ins can differ from the variation of biological samples [Risso et al., 2014]. Hence, the standardization of data by spike-ins does not necessarily always reflect variation in endogenous mRNA content [Vallejos et al., 2017].

Current cell filtering methods rely on thresholding to identify empty wells and failed amplification as well as dying or damaged cells. In particular, the library size defined as the total number of reads in a cell helps to distinguish low and high quality cells. Only cells which exceed user determined thresholds for library size and number of expressed genes are considered 'high quality' and kept in the dataset. These thresholds are dependent on the scRNA-seq protocol and the cell type. For example, Haber et al. [2017] removed all cells with less than 800 expressed genes (UMIs), while excluding SMARTseq2 sequenced cells with less than $3,000$ genes. Furthermore, the fraction of reads assigned to mitochondrial genes and ERCCs help to filter out low quality cells [Ilicic et al., 2016, Islam et al., 2014, Lun et al., 2016b]. In general, cells with an active metabolism have many mitochondria and a higher fraction of mRNAs from mitochondrial genes. However, Ilicic et al. [2016] have connected high fractions of mitochondrial reads to ruptures of the cell membrane and suggest to remove these cells from further analysis.

Filtering for doublets is a delicate procedure that became relevant with the emergence of droplet-based methods [Klein et al., 2015, Macosko et al., 2015, Zheng et al., 2017]. There, the doublet rate is proportional to the sample size (*i.e.* number of cells per droplet-sequencing run). Sequencing a larger amount of cells in one experimental run requires a higher cell concentration, which leads to a higher chance to encapsulate several cells in one droplet [Macosko et al., 2015]. Kang et al. [2018] used the molecular phenotype variations across individuals to assign reads to the correct person with *demuxlet* and determined the doublet proportion to $1\,\%$ per $1,000$ cells. It must be noted, that *demuxlet* cannot detect doublets from the same individual and performs better if the number of individuals per experimental run is higher. A doublet rate of $1 - 5\,\%$ is a desirable trade-off for high-throughput methods [Ziegenhain et al., 2017]. Still, doublets may result in spurious, apparently novel cell-types, especially when we study rare cell types. Only thorough analysis can discriminate doublets from real results. For example, Haber et al. [2017] assumed that any relevant subset should consist of at least 50 cells and adapted the number of sequenced cells per experimental run using a negative binomial model. If we do not aim to mix several species samples or have a phenotype available, tools like *scrublet* [Wolock et al., 2018] simulate doublets based on a given dataset and estimate a doublet score. However, *scrublet* is only able to detect about $50\,\%$ of the doublets and it might mistake cells with diverse gene expression (e.g. differentiating cells) as doublets.

2.2 Gene filtering

Filtering genes for quality control removes unexpressed and low-abundance genes. Low-abundance genes are dominated by noise and dropout events [Brennecke et al., 2013] and do not contain enough information for robust statistical inference [Bourgon et al., 2010]. Therefore, we reduce computational effort and memory use with little loss of information. We distinguish between mean-based filtering [Lun et al., 2016b] and count-based thresholding (*e.g.* Klein et al. [2015]). Mean-based filtering expects lowly-expressed genes to be uninformative and therefore removes them. The disadvantage of the approach that it may complicate the detection of rare cell type when applied to complex datasets. For example, let us assume that a cell type with 1 % abundance and a gene is exclusively expressed in this population with mean μ_1 and in the rest with $\mu_2 = 0$. Then, the overall expression is $\mu = 0.01 \cdot \mu_1 + 0.99 \cdot \mu_2 = 0.01 \cdot \mu_1$. Thus, given a mean threshold of 1, μ_1 has to be higher than 100 in order to be included for further analysis.
In comparison, a count-based approach is morel liberal as it keeps a gene as soon as it was detected with a user threshold of reads. For example, Klein et al. [2015] kept all genes with at least 5 UMI-counts in at least one cell. Especially when performing differential gene expression tests (see section 2.10) in complex datasets with potentially rare cell types, a count-based thresholding is agnostic to the cell type abundance. The disadvantage is that we keep a lot of genes, which are dominated by noise.
Another important aspect is potential contamination in the sequencing of single cells. Especially in droplet-based sequencing protocols, where cells are encapsulated in nanoliter droplets, contamination by *ambient mRNA* is quite common [Macosko et al., 2015, Zheng et al., 2017]. The term 'ambient mRNA' describes a background contamination from cells which burst prior to library construction, thereby distributing their mRNAs in the cell suspension. To date, the issue of ambient RNA is recognized, but generally ignored in the analysis of scRNA-seq data [Young and Behjati, 2018]. Recently, Young and Behjati [2018] presented a method for droplet-based protocols, where they estimate the background noise from empty droplets and subtract it from the expression level in every single cell. Angelidis et al. [2018] presented a similar approach to filter ambient mRNAs for creating a lung cell atlas.

2.3 Normalization

Data normalization in single-cell RNA-seq aims to reduce cell- and gene-specific biases from the data in order to increase comparability and to regularize the data. For example,

Paneth cells in the small intestine produce high amounts of defensins [Haber et al., 2017] and before normalization the corresponding transcriptome is dominated by reads from defensins. An appropriate normalization reduces the influence of these highly expressed genes in favor for medium and lowly expressed genes.
In general, a suitable data normalization has to meet the characteristics of the experimental single-cell RNA-seq setup (see section 1.3). One aspect is the so-called *gene coverage*. Either, an experimental protocol provides reads from the complete gene body or solely from the last exon. When the reads cover the complete gene body, normalization has to account for the length of the gene as longer genes have more reads. Similarly, scRNA-seq protocols that cover only the last exon are less influenced by gene length.
Another aspect concerns the library size and the number of expressed genes. These terms are closely connected. The library size is defined as the total read count of a cell. Especially in UMI-based protocols, the number of expressed genes increases with increasing library size. Most scRNA-seq protocols have a sensitivity around 10 % and do not detect all expressed genes in a cell, otherwise the number of expressed genes would converge. Thus, increasing the sensitivity of the protocol increases both the library size and the number of expressed genes. However, the sensitivity also varies within an experiment from cell to cell (see Fig. 1.5) and should be accounted for by normalization.
Here, we provide an overview of commonly used normalization approaches. Most methods determine a cell-wise scaling factor, so-called *global scaling*, such as CPM, RPKM and TPM. In addition, methods like RLE, MRN and TMM consider gene expression across all cells to determine a cell-wise scaling. Thus, the global scaling factor also reflects the gene expression level in the complete sample. Single-cell specific normalization methods take into account dropout events, or changes of the noise level caused by higher sequencing depth. In the following, we present the aforementioned methods in detail and discuss advantages and disadvantages. In Chapter 3, we evaluate these methods on several datasets.

CPM

Normalizing by *counts per million* (CPM) is a method to correct for library size effects by global scaling. Here, the total number of reads per cell j, N_j, is used to scale the reads for every gene C_{ig} as follows:

$$R_{ig} = 10^6 \cdot \frac{C_{ig}}{N_i}. \tag{2.1}$$

The advantage of the *size factor* in eq. 2.1 is that we obtain relative proportions for each gene. However, there are several disadvantages. First, Wagner et al. [2012] have shown that the library size N_i does not reflect the number of mRNA copies of a single cell. Second, highly expressed genes, which dominate the transcriptome, will be scaled the

same as lowly expressed genes. Third, differences in library size may be caused by relevant factors that we should preserve. For example, Finak et al. [2015] have shown that library sizes correlate with biological effects such as the number of expressed genes. Further, Lun et al. [2016a] have demonstrated the influence of cell type specific properties on library size. Last, library sizes may be biased by a considerable number of differentially expressed genes [Anders and Huber, 2010].
However, CPM is a frequently used normalization method. Usually, CPM-normalized count data are log-transformed to reduce the influence of highly expressed genes, and to create approximately Normally distributed data.

RPKM

Every gene is unique in the number of exons and introns, and the gene length. The gene length varies due to the size of the encoded protein, and ranges from 500 bp to $2 \cdot 10^6$ bp [Grishkevich and Yanai, 2014].
In RNA-seq, we map reads to the full length of the gene body in mRNA quantification in the sequencing step. This implies that longer genes have more reads. Furthermore, reads are not unique because of the preceding amplification step as the library becomes automatically larger with more PCR cycles. There, biases are accounted for by a global scaling factor called *reads per kilobase per million reads (RKPM)*. RPKM is calculated by

$$R_{ig} = 10^9 \cdot \frac{C_{ig}}{N_i \cdot L_g}.$$

Here C_{ig} is the number of mappable reads for a gene g in cell i, N_i denotes the library size of the cell i, and L is the length of the mappable regions in a gene g, counted in base pairs of the exon [Mortazavi et al., 2008].
RPKM normalization was developed specifically for bulk RNA-seq samples with high amounts of input RNA and little dropout. However, in single-cell RNA-seq, we observe low amounts of input RNA and technical zeros due to low sensitivity. Moreover, the mapping coverage of the gene body is shifted towards the 3' end. Thus, there is no uniform coverage of the full gene body. In addition, as stated for CPM, the library size does not reflect the number of mRNA copies of a single cell. Hence, RPKM represents a different relative mRNA level for every cell, which renders the approach inappropriate for scRNA-seq.

TPM

In order to overcome the inconsistencies of RPKM, Li et al. [2009], Wagner et al. [2012] proposed *transcripts per million (TPM)* values. TPM is computed as

$$R_{ig} = 10^6 \cdot \frac{C_{ig} \cdot L_r}{\tilde{N}_i \cdot L_g},$$

where C_{ig} is the number of mappable reads for a gene g and cell i, L_r denotes the average read length, L_g is the length of the mappable regions in a gene g, and $\tilde{N}_i$ is the sum of length-normalized reads over all genes $g \in G$ in cell i, *i.e.* $\tilde{N}_i = \sum_{g \in G} \frac{C_{ig} \cdot L_r}{L_g}$.
Wagner et al. [2012] have demonstrated that TPM reflects relative molar concentrations of mRNA molecules more consistently than RPKM. However, the mapping coverage of a gene is neglected the same as in RPKM. To take into account the mapping bias, Patro et al. [2017] developed a probabilistic transcript quantification model and implemented it in the read alignment tool *Salmon*. In particular, they infer an *effective gene length*, which averages out the mapping bias. Furthermore, the model uses prior information such as the empirical fragment length distribution of the pseudoalignment model implemented in *kallisto* [Bray et al., 2016]. Similarly to CPM, TPM is often used for normalization, especially for SMARTseq data.

RLE and MRN

Anders and Huber [2010] introduced *relative log-expression* (RLE) and *median-ratio normalization* (MRN) representing global scaling factor normalizations. Both approaches are implemented in the *DESeq* package [Anders and Huber, 2010]. In contrast to the global scaling methods above, these methods use information across all cells. In particular, these methods assume that the expected gene expression level of two cells i and j reflects the difference in sequencing depths. The assumption is valid for all genes g that are not differentially expressed between the two cells. In order to apply the idea to several cells and genes, the authors compute a cell-specific size factor s_i over all genes $g \in G$:

$$s_i = \text{median}_{g \in G} \frac{C_{ig}}{(\prod_{i=1}^{n} C_{ig})^{1/n}}. \tag{2.2}$$

There, C_{ig} denotes the counts per gene g in cell i and the denominator denotes the geometric mean for gene g over all cells n. In RLE, the size factor s_i is additionally divided by library size before being used to correct count data. The RLE method is implemented in the *scater* package [McCarthy et al., 2017]. In MRN, the count data of cell i are normalized by dividing by the size factor s_i.

The geometric mean in eq. 2.2 can be considered as a pseudo-reference. Hence, the size factor s_i considers information across all cells. Taking the median over all genes ensures that the size factor is more robust to biases such as differentially expressed genes. For example, if less than 50 % of the genes are balanced up- and down-regulated, the median is unaffected from the gene expression shift. However, the geometric mean is 0 if $C_{ig} = 0$ for any $i \in \{1, \ldots, n\}$, which is often the case in scRNA-seq. Lun et al. [2016a] have shown that size factors were not designed for datasets with many zeros.

TMM

The global scaling approach of *trimmed mean of M-values* (TMM) was originally designed to compare expression levels across bulk RNA-seq samples with a particular focus on differential gene expression analysis [Robinson and Oshlack, 2010]. A trimmed mean is a mean, which is computed on a subset of the data, where the upper and lower p-percentile is removed ($p = 30$ in Robinson and Oshlack [2010]). In this approach, the samples are considered largely homogeneous. In fact, $40 - 70\,\%$ of the genes are assumed to be not-differentially expressed.
To make gene expression comparable, one cell is selected as reference cell r. Then, all other cells are scaled to the reference using M values, *i.e.* library-size adjusted expression ratios with respect to the reference cell. Let N_r and N_i be the library size for reference cell r and cell i. Further, C_{rg} and C_{ig} denote the counts for a gene g in reference cell r and cell i. Then, the log-fold changes M_{ig}^r are given by

$$M_{ig}^r = \frac{\log_2\left(\frac{C_{ig}}{N_i}\right)}{\log_2\left(\frac{C_{rg}}{N_r}\right)}.$$

Here, C_{rg} and C_{ig} have to be non-zero. Then, a set of M_{ig}^r-values for all cells i is computed and too extreme M_{ig}^r-values are trimmed away. Thus, let G^* represent the set of genes with valid M_{ig}^r-values. Next, the authors define weights w_{ig}^r for a cell i and gene g:

$$w_{ig}^r = \frac{N_i - C_{ig}}{N_i \cdot C_{ig}} + \frac{N_r - C_{rg}}{N_r \cdot C_{rg}}.$$

The weights are small for highly expressed genes and large for lowly expressed genes. Thus, the weights reduce the influence of highly-expressed genes.
Finally, TMM values of gene g for a cell i given reference r is the weighted mean of M_{ig}^r-values:

$$\log_2\left(\mathrm{TMM}_{ig}^r\right) = \frac{\sum_{g \in G^*} w_{ig}^r M_{ig}^r}{\sum_{g \in G^*} w_{ig}^r}. \tag{2.3}$$

The resulting factor in eq. 2.3 is then applied as cell-specific scaling factor. It must be noted, that the TMM values are similar to the size factors in eq. 2.2.
As stated above, a valid M-value requires non-zero count data. This limits its use for scRNA-seq data. Intuitively, one may remove all zero counts from the data or add a pseudo-count. However, this introduces a considerable bias to the data.

Pooling across cells and deconvolution of size factors

The pooling approach described by Lun et al. [2016a] is one of the first single-cell specific normalization approaches. It accounts for the sparsity of single-cell data by summing gene expression counts across a pool of cells. Pool-specific size factors can be deconvolved into cell-specific size factors by using different pools of cells. The main assumption here is that the majority of genes of the pooled cells are non-differentially expressed in the dataset. Clustering the cells by their expression profile helps to weaken this assumption as the pooling and deconvolution steps are applied cluster-wise. The method is implemented in the *scran* package.
In particular, the deconvolution method creates a system of linear equations from pools of cells with similar library size, such that the number of linear equations is equal to the number of cells. First, the cells are ranked by library size, then split into odd and even ranks, and arranged on a ring. Thus, the cells with odd rank and smallest library size lie next to cells with even rank and small library size. The same holds for cells with the largest library size. The pools are defined by a sliding window approach on the ring. An additional set of equations relates the size factor estimate against the pseudo-cell reference, which is computed as the summed gene expression counts of all cells.
The pooling approach reduces cell-specific biases from dropout events and differential gene expression. This is an advantage over the classical size factor approaches such as RLE, MRN and TMM.

Other normalization methods

We close this section with a brief overview on other approaches to normalization.
Several methods based on non-parametric statistics have been proposed. For example, Bullard et al. [2010] suggested an *upper quartile normalization (UQN)*, where they normalize the gene expression distribution to a reference distribution and round all values to integers. Recently, Cole et al. [2018] presented the scRNA-seq pipeline `scone`, which allows to compare several normalization approaches. Among others, they compare TMM, MRN and full-quantile normalization. Full-quantile normalization scales the gene expression distribution to a reference distribution. It must be noted that quantile normalization

is a gene-specific scaling, in contrast to cell-specific global scaling.
With increasing availability of single-cell RNA-seq data, we are able to study the properties of the data type in more detail. Bacher et al. [2017] demonstrated that sequencing depth increases the gene count for highly and medium expressed genes, but not for lowly expressed genes (shown for full-length single-cell RNA-seq without UMIs). Consequently, a global scaling factor introduces a bias for either lowly or highly expressed genes. Their method called *SCnorm* evaluates the so-called *count-depth relationship* for several groups of genes based on the expression level and computes an appropriate scaling factor for each group. While this model is not suited for batch effect correction, between-sample effects can be accounted for. The major disadvantage of the model is its computational cost rendering it unsuitable for datasets with more than 1000 cells.
We have already mentioned that gene expression is cell-type specific and varies largely across tissues. To account for cell-specific gene expression, Paulson et al. [2017] developed a method called *quantile shrinkage normalization (qsmooth)*. Here, the authors identify genes with tissue-independent expression level, which they call *housekeeping genes*[1]. This invariant gene set is then used as reference for a quantile normalization for all genes. However, the method is only applicable for data from several tissues.
In addition to scaling approaches, regression models can account for several covariates, such as experimental design, GC-content, gene length, and sequencing depth. For example, Hansen et al. [2012] used a Poisson mixture to model gene expression, and systematic biases such as gene length and GC-content. This method is called *conditional quantile normalization (CQN)*. Of note, regression models are also frequently used to model technical effects such as batch effects. However, we discuss the removal of systematic bias in the following section.

2.4 Batch effect correction

Batch effect detection is an important part of quality control for single-cell RNA-seq data. A batch effect is any variation caused by technical effects arising when samples are handled in distinct batches (see section 1.3 for details). Accurate handling of batch effects has important implications on downstream data analysis due to their property to mask the true biological variability and their adverse effects on reproducibility.
In the following, we describe commonly used methods for batch effect correction whose performance study is evaluated in Chapter 3. In particular, we distinguish linear models

[1] Housekeeping genes are constantly expressed genes, which are present in all cells. These genes are required to maintain a cell's metabolism, for example. In scRNA-seq, these genes show the same noise level as non-housekeeping genes.

such as limma and ComBat, factor analysis models such as PEER, f-scLVM, RUV, and CCA, and non-linear models such as MNN.

limma

The name *limma* is an acronym for 'Linear Models for Microarray Data' - the analysis framework `limma` was developed to study microarrays, protein arrays, and other data [Ritchie et al., 2015]. Thus, `limma` provides data pre-processing, batch correction, linear modelling, and differential expression tests. We discuss `limma`'s differential expression test in section 2.10. In order to correct for batch effects with the *removeBatchEffect* function, `limma` uses a linear regression model for C_{ijg} the gene expression for gene g for sample j from batch i:

$$C_{ijg} = \alpha_g + X\beta_g + \varepsilon_{ijg}$$

where α_g denotes the overall gene expression, X is a design matrix[2] with corresponding regression coefficients β_g, and $\varepsilon_{ijg} \sim \mathcal{N}(0, \sigma_g^2)$ is an expression noise term. The design matrix X contains all values of explanatory variables. For example, it contains an indicator variable to describe, which cell belongs to which batch. A corrected gene expression value C^*_{ijg} reads

$$C^*_{ijg} = C_{ijg} - X_b\hat{\beta}_g,$$

where X_b is the model matrix part for the batch covariate and $\hat{\beta}_g$ is the parameter estimate for β_g. This is the simplest approach to correct for batch effects and as mentioned in the previous section, it is not robust for small sample sizes.

ComBat

The empirical Bayesian (EB) method implemented in *ComBat* was originally designed for microarray data [Johnson et al., 2007, Leek et al., 2012]. The model basis is an adjustment of location L (mean) and scale S (variance) between the batches, *i.e.* we assume that the difference between batches can be corrected by scaling both mean and variance. Simple L/S methods are not robust to outliers in small sample sizes and hence require large batch sizes.

The advantage of ComBat is the preservation of the original first and second order moments of the data. The batch correction runs on standardized data, *i.e.* we assume that the gene expression is Normally distributed for each gene. Implicitly, ComBat assumes that all batches are (exact) replicates and thus, the cell type composition is similar in all batches. In Chapter 4, our data violate this assumption as certain cell types are enriched in a few

[2]The design matrix is also known as model matrix.

samples and we implemented a modified version of ComBat to address the issue. Thus, we introduce the model in detail.
If C_{ijg} is the gene expression for gene g for sample j from batch i, then, the L/S model is

$$C_{ijg} = \alpha_g + X\beta_g + \gamma_{ig} + \delta_{ig}\varepsilon_{ijg}.$$

Here, α_g denotes a gene-specific expression offset, X is a design matrix with corresponding regression coefficients β_g, γ_{ig} and δ_{ig} denote additive and multiplicative batch effects, respectively, and $\varepsilon_{ijg} \sim \mathcal{N}(0, \sigma_g^2)$ is the error term with variance σ_g^2. The additive batch effect term γ_{ig} models a systematic shift in the offset of gene expression. The multiplicative batch effect term δ_{ig} models a systematic increase or decrease of the variance across batches. We aim to compute batch-corrected data C^*_{ijg} using the equation:

$$C^*_{ijg} = \hat{\alpha}_g + X\hat{\beta}_g + \frac{1}{\hat{\delta}_{ig}}\left(C_{ijg} - \hat{\alpha}_g - X\hat{\beta}_g - \hat{\gamma}_{ig}\right), \tag{2.4}$$

where $\hat{\alpha}_g$, $\hat{\beta}_g$, $\hat{\gamma}_{ig}$, $\hat{\delta}_{ig}$ are estimated parameters based on the model. Here, the EB method uses a parametric empirical prior to estimate batch effect parameters. In particular, EB methods perform a shared variance estimation across all genes and cells based on a prior distribution of variances.
in the first step, the authors standardize the data matrix $\mathbf{X}$. For this purpose, they determine the empirical parameter values $\hat{\alpha}_g$, $\hat{\beta}_g$, $\hat{\gamma}_{ig}$ and $\hat{\sigma}_{ig}$ in eq. 2.4. Here, the authors introduce the following constraints for identifiability:

$$\sum_i n_i \hat{\gamma}_{ig} = 0$$

$$\hat{\sigma}_g^2 = \frac{1}{n}\sum_{ij}\left(C_{ijg} - \hat{\alpha}_g - X\hat{\beta}_g - \hat{\gamma}_{ig}\right)^2,$$

where n_i is the number of cells per batch and n is the total number of cells. Of note, the multiplicative batch effect term $\hat{\delta}_{ig}$ was not estimated. The expression matrix is standardized gene-wise to similar overall mean and variance:

$$Z_{ijg} = \frac{C_{ijg} - \hat{\alpha}_g - X\hat{\beta}_g}{\hat{\sigma}_g}.$$

The standardized data Z should be normally distributed with mean γ'_{ig} and variance δ^2_{ig}. In the next step, the authors employ the EB method to estimate mean γ'_{ig} and variance δ^2_{ig} of the standardized data Z. As prior distributions for the batch effect parameters, the

authors chose two conjugate priors for the normal distribution:

$$\gamma'_{ig} \sim \mathcal{N}\left(\gamma_i, \tau_i^2\right) \quad \text{and} \quad \delta_{ig}^2 \sim \text{Inverse Gamma}\left(\lambda_i, \theta_i\right),$$

with hyperparameters γ_i, τ_i^2, λ_i, θ_i. The hyperparameters are estimated separately for each batch i borrowing information across all genes. In ComBat, the authors use the empirical mean and variance estimates.
Then, conditional posterior means for γ'_{ig} and δ_{ig}^2 are given by

$$\gamma_{ig}^* = \frac{n_i \bar{\tau}_i^2 \hat{\gamma}_{ig} + \delta_{ig}^{2*} \bar{\gamma}_i}{n_i \bar{\tau}_i^2 + \delta_{ig}^{2*}} \quad \text{and} \quad \delta_{ig}^{2*} = \frac{\bar{\theta}_i + \frac{1}{2} \sum_j \left(Z_{ijg} - \gamma_{ig}^*\right)^2}{\frac{n_j}{2} + \bar{\lambda}_i - 1}.$$

Finally, the adjusted batch effect parameter estimates are used to compute batch-corrected data c_{ijg}^* similarly to eq. 2.4,

$$c_{ijg}^* = \hat{\alpha}_g + X\hat{\beta}_g + \frac{\hat{\sigma}_g}{\hat{\delta}_{ig}^*}\left(Z_{ijg} - \hat{\gamma}_{ig}^*\right).$$

A detailed description of the parameter estimation procedure can be found in the supplementary material of Johnson et al. [2007].

scLVM

The *single-cell latent variable model* (scLVM) [Buettner et al., 2015] is a random mixed-effect model, *i.e.* it combines fixed effects with random effects. Let $C \in \mathbb{R}^{n \times g}$ an expression matrix with n cells and g genes. Then, the fixed effect model of the gene expression vector C_g is

$$C_g = X\beta_g + Z\gamma_g + \varepsilon_g.$$

Here, $X, Z \in \mathbb{R}^{n \times g}$ denote design matrices for the fixed and random effects, respectively. A fixed effect is modeled as presented above in the linear regression case as parameter estimate, such as the covariate β_g. A random effect is modeled as a distribution parameter, *i.e.* its effect on the data follows a distribution in each data point. In scLVM, $\gamma \sim \mathcal{N}(0, ZZ^T)$, where Z is a so-called *hidden factor* estimated from the data, and ZZ^T is the covariance matrix of the random effect vector γ. The model was originally designed to account for cell cycle effects in scRNA-seq data.
We have used scLVM to correct for batch effects in Haghverdi et al. [2016], where we modeled the batch effect as random effect. However, scLVM scales with $\mathcal{O}(n^2)$ with the number of cells n and is therefore computationally inefficient to use for large datasets.

PEER

The basis for *probabilistic estimation of expression residuals* (PEER) is a factor analysis model [Stegle et al., 2010, 2012]. Originally, its purpose was to detect hidden confounding factors that interfere with expression quantitative trait loci (eQTL) studies. In contrast to known factors, a hidden factor is an unobserved factor, such as the cell culture conditions. These effects introduce a systematic shift in the data, that is different from the noise term. Noise terms usually follow the Normal distribution. If a dataset contains a hidden factor, which is not accounted for, a linear regression model estimates residuals, which do not follow the Normal distribution. In this case, the model has to be adapted to account for the unknown factor. Batch effects are also modeled as unknown factors in PEER. Stegle et al. [2010] have shown that accounting for hidden factors increases the increases the power of those eQTL studies. In contrast to the previously introduced linear models, PEER does not depend on a design matrix and an explicit batch covariate. Instead, experimental confounders can be included in the model.
Initially, the gene expression matrix $\mathbf{X} \in \mathbb{R}^{n \times g}$ with n cells and g genes is modeled using a general additive model with known and unknown sources of variation:

$$\mathbf{X} = \mathbf{X}^{(1)}(T) + \mathbf{X}^{(2)}(F) + \mathbf{X}^{(3)}(U) + \varepsilon.$$

The model includes additive contributions $\{\mathbf{X}^{(.)}\}$ based on the transcriptional state T, known factors F and hidden factors U, plus a noise term ε. The gene expression c_{ig} for each gene g and cell i is modeled using a joint Bayesian model. In particular, gene expression follows a Normal distribution with mean C_{ig} given the known and hidden factor contributions:

$$P\left(C_{ig}|C_{ig}^{(1)}, C_{ig}^{(2)}, C_{ig}^{(3)}, \tau_g\right) = \mathcal{N}\left(C_{ig}|C_{ig}^{(1)} + C_{ig}^{(2)} + C_{ig}^{(3)}, \frac{1}{\tau_g}\right),$$

where τ_g is the precision of Gaussian noise with a Gamma-distributed prior

$$P\left(\tau_g\right) = \Gamma\left(\tau_g|a_\tau, b_\tau\right).$$

Furthermore, the model contributions per gene $C_{ig}^{(.)}$ are modeled as random variables. A detailed description of the model contributions and parameter inference are given in Stegle et al. [2010].

f-scLVM

The *factorial single-cell latent variable model* (f-scLVM) follows a similar modelling scheme as PEER [Buettner et al., 2017]. If $\mathbf{X} \in \mathbb{R}^{n \times g}$ is a gene expression matrix, the contributing factors $\mathbf{X}^{(.)}$ are pre-annotated factors as gene sets for cell-cycle or tumor suppressor regulation, unannotated factors, and noise. Unannotated factors can be both sparse factors (e.g. gene sets) and dense factors (*i.e.* likely confounding factors). In contrast to PEER, batch effects can be included in the model as a known factor and regressed out in the process of model fitting.
As advancement of PEER f-scLVM includes an explicit model for dropout and thus represents a single-cell RNA-seq specific modelling approach. Details on the model and parameter inference can be found in Buettner et al. [2017].
It must be noted that f-scLVM and scLVM share a similar name, but are conceptually different — f-scLVM is based on factor analysis and scLVM is a random fixed effect model.

RUV

A central problem in batch effect removal is overcorrection, *i.e.* unintentionally removing biological signal. A possible solution for this problem is to use non-differentially expressed genes for factor analysis, as suggested by the authors of *Remove Unwanted Variation, 2-step* (RUV-2) [Gagnon-Bartsch and Speed, 2012]. [Risso et al., 2014] extended the RUV model of Gagnon-Bartsch and Speed [2012] to discrete RNA-seq data.
Let $\mathbf{X} \in \mathbb{R}^{n \times g}$ be the observed gene expression matrix for g genes and n cells. Then,

$$\log(\mathbf{X}) = X\beta + W\alpha + O + \varepsilon \tag{2.5}$$

is a linear decomposition of the observed gene expression matrix into factors of interest X, unobserved covariates W and an offset O with parameters α and β. The aim is to estimate β while W is unobserved. In order to ensure identifiability of β, Gagnon-Bartsch and Speed [2012] and Risso et al. [2014] proposed several model constraints for the unobserved covariates W. The first idea for RUV_g was introduced by Gagnon-Bartsch and Speed [2012]. Here, the estimation of W is restricted to control genes g_c, where $\beta_{g_c} = 0$ holds, *i.e.* these genes are invariant for the factors of interest X, such as control and disease conditions. In this case, eq. 2.5 simplifies to

$$\log(Z_{g_c}) = \log(\mathbf{X}_{g_c}) - O_{g_c} = W_{g_c}\alpha_{g_c} + \varepsilon, \tag{2.6}$$

which represents a simple factor analysis model and can be solved, *e.g.* by singular value decomposition (SVD).

The second approach is implemented in RUV_r. Here, residuals are computed from the count matrix $\mathbf{X}$ by regressing out the covariate of interest X. Then, W is estimated on the residual matrix. The third approach RUV_s uses replicates or negative control samples for which the factors of interest X are constant. A detailed description of the method and parameter inference can be found in Gagnon-Bartsch and Speed [2012] and Risso et al. [2014].

CCA and dynamic time warping

Canonical correlation analysis (CCA) is a linear latent variable model. It is used to extract interesting features from high-dimensional datasets or serves for dimensionality reduction [Murphy, 2012]. Conceptually, the basis vectors of CCA represent a projection that maximizes correlation between two datasets. In the context of single-cell RNA-seq, Butler et al. [2018] adopted CCA for batch effect correction. In particular, the authors identified the shared correlation structure between two datasets, which is the biological structure of interest. In order to correct for batch effects, the shared correlation structure is aligned using dynamic time warping. Furthermore, this method adjusts for differences in cell density between datasets.

MNN

Matching mutual nearest neighbors (MNN) defines a graph-based approach that is independent of cell type composition across batches [Haghverdi et al., 2018]. The method first detects pairs of nearest neighbors for different batches. Then, a batch-correction vector is computed to map batches onto the reference batch. Specifically, the projection vector is derived from the distance of the mutual nearest neighbors. Furthermore, the projection vectors are combined by a weighted average to obtain a single projection vector per batch. This projection ansatz requires that the batch effect is approximately orthogonal to the biological effect. In other words, we compute the first N principal components (see section 2.7) of the dataset and the principal components correlate either with the batch effect or with the biological signal. Given the high dimensionality of scRNA-seq data and the low intrinsic dimension of the data manifold, this assumption is likely fulfilled. Furthermore, MNN relies on the biological variation between cell types being larger than the variation caused by the batch effect. In general, this is a reasonable assumption, which assumes reproducibility of experimental results. A detailed description of the MNN algorithm is given in the supplement of Haghverdi et al. [2018].
It must be noted that MNN uses the cosine similarity to normalize data prior to batch correction. Hence, MNN does not work with an arbitrary data normalization (see sec-

tion 2.3). MNN also maps cells without mutual nearest neighbors to the reference plane, but separate from the matching pairs. Hence, when a batch contains a certain cell type that is absent in another batch, these cells appear in the embedding with a reasonable distance to the other cells.

Other methods

We close the section of batch effect correction by briefly mentioning a few more methods, which we do not evaluate in Chapter 3. For example, Chen and Zhou [2017] developed the single cell partial least squares *(scPLS)* model. This model uses partial least squares regression classify genes into invariant or influenced for the factor of interest. Thus, it is similar to the factor analysis models presented above.
Risso et al. [2018] created an extension of RUV called *ZINB-WaVE*, *i.e. zero-inflated negative Binomial-based Wanted Variation Extraction*. The authors account for scRNA-seq specific characteristics as dropout events, and overdispersion by using a zero-inflated negative Binomial model of the data.
Recently, Korsunsky et al. [2018] developed *Harmony*, an iterative algorithm that projects cells from distinct batches or studies into a shared embedding space and simultaneously corrects for various confounding factors.
Last, Eraslan et al. [2019] presented the Deep Count Autoencoder *(DCA)*. Originally, this is a neural network approach to denoise scRNA-seq data, *i.e.* to reduce dropout events. In addition, this model can be used to reduce technical variation, such as batch effects.

2.5 Correcting for cell cycle effects

The cell cycle describes an extensive cellular process that controls cell growth, DNA strand duplication, and division of a cell (see section 1.1 for details). Thus, a proliferating cell differs from a quiescent cell in its gene expression profile. These differences may lead to a bias in the scRNA-seq analysis, if, for example, a differentiation process overlaps with the cell cycle [Buettner et al., 2015, Lönnberg et al., 2017]. Thus, it is prudent to determine the cell cycle state of the cells.
Buettner et al. [2015] used gene annotations from Cyclebase and Gene Ontology to determine the variance explained from cell cycle genes. This work is one of the earliest methods to correct for cell cycle effects in single-cell RNA-seq data is the *single-cell latent variable model (scLVM)*. In principle, this model subtracts the variance imposed by cell cycle genes [Buettner et al., 2015]. Moreover, the actual cell cycle state is not inferred. However, the model does not scale well with the number of cells. A successor, the *factor*

analysis single-cell latent variable model (*f-scLVM*) provides a scalable approach to account for cell cycle effects [Buettner et al., 2017] (see section 2.4 for details).
In contrast, gene set scoring determines the cell cycle state based on the respective gene sets for S and G_2/M phase. Whitfield et al. [2002] have experimentally characterized these gene sets from human HeLa cells, and they have subsequently been used in both the human and mouse contexts for scRNA-seq [Macosko et al., 2015, Satija et al., 2015, Tirosh et al., 2016]. Here, we present two different approaches to determine the cell cycle state. Satija et al. [2015] presented a scoring algorithm, that compares the mean expression level of the given gene set with a random gene set. The score is the difference of these to groups. In this way, the score for S phase and G_2/M phase are determined. A cell is assigned to the phase with the highest score. If both scores are negative, the cell is assigned to the G_1 phase. Hence, Satija et al. [2015]'s approach relies heavily on the accuracy of these gene sets.
Scialdone et al. [2015] presented a different scoring approach with the tools *sandbag* and *cyclone*. It uses pre-defined gene sets to compute gene pairs as markers for the respective phase. Then, these marker pairs are used to predict the cell cycle state.
The cell cycle is only one type of biological variation that can be assessed from scRNA-seq data. Other types of biological variation are aging, inter-individual variability or gender effects. In contrast to technical variation such as batch effects, we do not necessarily have to correct for it. In addition, we already presented models such as *PEER* and *RUV* as batch correction models, but they are also designed to correct for biological sources of variation.

2.6 Identification of highly variable genes

The quantification of gene expression levels in a single cell creates both opportunities and pitfalls. While we are able to look up the expression level of over $20,000$ genes, we face the challenge of determining cell-type specific gene signatures, decision makers, and drivers of disease. Here, informative is translated into highly variable, *i.e.* the variance of a gene is above the intrinsic noise level due to stochastic gene expression. An appropriate selection of highly variable genes fully reflects the biological data manifold and reduces noise while an inappropriate selection may cloak cell types. For example, Cacchiarelli et al. [2018] inferred pseudotime and branching events of reprogrammed cells. Depending on the choice of highly variable genes, they uncovered a branching of partially reprogrammed und unreprogrammed cells, or they did not. This example illustrates how the selection of highly variable genes affects the downstream analysis.

Here, we review several models that identify highly variable genes. Variability in gene expression is closely connected to the intrinsic noise level of the data, *i.e.* the gene variability has to be above the noise level. In general, we focus on highly variable genes because they are considered to be more informative of the heterogeneity in the data, and to increase computational efficiency of downstream analysis methods.
One of the first noise models for single-cell RNA-seq was the Anders model [Brennecke et al., 2013]. The Anders model was developed to estimate noise based on the magnitude of expression of technical RNAs (for example ERCCs) in non-UMI based scRNA-seq protocols (e.g. SMARTseq). Briefly, during the amplification step in scRNA-seq, all reads are multiplied, but the noise level increases at the same time (see section 1.3 for details). In general, the coefficient of variation (CV^2) decreases with increasing mean expression level μ. Brennecke et al. [2013] modeled this relation with a Gamma regression model:

$$\mathrm{CV}^2 \sim \frac{\alpha_1}{\mu} + \alpha_0,$$

where α_1 is the slope parameter and α_0 denotes the offset, *i.e.* the noise level converges to a lower limit with increasing expression. Both parameters are estimated with a least squares fit. Genes with a higher CV^2 than expected by the model fit are considered 'highly variable'. This implies that the variability is caused by biological variation rather than technical noise. It must be noted that the model assumes that most genes are not differentially expressed. Thus, the model works well for technical genes such as ERCC spike-ins. However, if we use the model for the data itself, we may observe a bias from differentially expressed genes. This bias is particularly pronounced, when studying heterogeneous data.
With the invention of unique molecular identifiers (UMIs) [Islam et al., 2014], it became possible to remove a substantial level of noise from read amplification. UMIs specifically tag reads from the same mRNA molecule. Hence, UMIs allows us to distinguish copies of the same transcript from copies of the amplification. In this case, we may apply a Poisson noise model as demonstrated in Klein et al. [2015]. There the authors considered a gene highly variable, if it was above the Poisson model fit plus 10 % uncertainty

$$\mathrm{CV}^2 \sim \alpha \cdot \mu.$$

In this model, the coefficient of variation CV^2 and mean expression level μ are negatively correlated (on the log-scale, *i.e.* $\alpha < 0$).
The third approach to identify highly variable genes uses the dispersion δ instead of the CV^2 [Macosko et al., 2015]. This method is implemented in `CellRanger` [Zheng et al., 2017] and adapted to Python in `Scanpy` [Wolf et al., 2018a]. It must be noted that this model uses a log-normalized data matrix to compute mean and variance.

For any random variable X, dispersion δ and coefficient of variation CV^2 are closely related:

$$CV^2(X) = \frac{\mathrm{Var}^2(X)}{\mu(X)} = \mathrm{Var}(X) \cdot \delta(X).$$

In this approach, the genes are distributed based on their mean expression into a set of l equally sized bins. For each bin i, the mean expression $\mu_{\mathrm{bin},i}$ is calculated and the top 2,000 genes with highest normalized dispersion δ_{norm} are selected. Here, δ_{norm} reads

$$\delta_{\mathrm{norm}} = \frac{\delta - \mathrm{median}_{i\in\{1,\dots,l\}}\left(\mu_{\mathrm{bin},i}\right)}{\mathrm{mad}_{i\in\{1,\dots,l\}}\left(\mu_{\mathrm{bin},i}\right)},$$

where *mad* denotes the *median absolute deviation*. Grouping the genes by mean value averages out biases from outliers and thus increases robustness of the normalized dispersion. Furthermore, selecting highly variable genes based on dispersion does not rely on parametric noise model assumptions as the models presented before. In the data analysis presented in Chapter 4, we select highly variable genes based on the third approach.

2.7 Dimensionality reduction

The motivation for dimensionality reduction is to project the data to a low-dimensional space, which preserves all essential properties of the data. Here, we assume that several features are correlated such that the data manifold has a lower dimension. Therefore, dimensionality reduction aims to recover essential *latent factors* from the data, which sufficiently describe the data [Murphy, 2012]. This is a classical machine learning task and has been applied successfully in various fields such as face recognition. Furthermore, dimensionality reduction is largely used for preprocessing, noise-reduction and visualization. A different aspect is the application for data analysis. Hence, the advent of big data has spurred the development of novel linear and nonlinear dimensionality reduction techniques, which we review in this section.

ScRNA-seq is a novel member in the big data family. It quantifies the gene expression level for over 20,000 genes and several hundreds to thousands of cells. The inspection of every gene (feature) is infeasible. Instead, dimensionality reduction projects data into a low-dimensional space while preserving as much 'structure' as possible. Notably, the methods for dimensionality reduction differ largely in the interpretation of the term 'structure'. For example, linear dimensionality reduction such as Principal Component Analysis collapses the most important features into the leading components and orders them according to the variance explained by each component.

Furthermore, dimensionality reduction helps us to learn about the underlying manifold of the data — cells of the same cell type group together and continuous processes such as differentiation can be abstracted to a line in a low dimensional representation. Here, we present common methods to represent scRNA-seq data in a low dimensional space.

Principal Component Analysis (PCA)

Principal component analysis was first described by Karl Pearson in the early 20^{th} century [Pearson, 1901]. The method aims to represent as much variance in the leading components as possible. In principle, the empirical covariance matrix of the data and the corresponding eigenvalues and -vectors are computed. The eigenvectors are called principal components (PCs), while the eigenvalues correspond to the explained variance by each PC. Using PCA, a set of potentially correlated variables is transformed into a set of uncorrelated variables with respect to the new basis (*i.e.* principal components).
Let $\mathbf{X} \in \mathbb{R}^{n \times g}$ be a data matrix with n observations (cells) and g features (genes). Furthermore, let $\Sigma \in \mathbb{R}^{g \times g}$ be the corresponding covariance matrix, which is symmetric and positive semidefinite. The corresponding eigenvalue decomposition of Σ reads

$$\Lambda = \Gamma^T \Sigma \Gamma.$$

Here $\Lambda \in \mathbb{R}^{g \times g}$ is a diagonal matrix with the eigenvalues as diagonal elements (sorted in decreasing order) and $\Gamma \in \mathbb{R}^{g \times g}$ is an orthogonal matrix consisting of the eigenvectors of Σ. Finally, we transform the data matrix $\mathbf{X}$ as follows:

$$Y = \left(\Gamma^T \mathbf{X}^T\right)^T.$$

Here, Y is the data matrix projected into the orthogonal PC-space. The corresponding covariance matrix of $Y \in \mathbb{R}^{n \times g}$ is Λ. Λ has only entries on its diagonale and thus the features on Y are uncorrelated.
PCA assumes that data is normal distributed which does not hold true for single-cell RNA-seq. Moreover, PCA preserves large distances making it sensitive to outliers. However, PCA is a common preprocessing method to reduce the initial number of dimensions of a dataset. In Chapter 3, we derive a scoring method based on PCA.

Diffusion Maps

Diffusion maps [Coifman et al., 2005, Haghverdi et al., 2015] represent a nonlinear dimensionality reduction method with similarity to PCA. There the distance between two points is evaluated by a Markov jump process, *i.e.* the distance of two points corresponds

to the probability of jumping from one point to the other in n iterations. The resulting transition probability matrix is decomposed into its eigenvectors, which are called *diffusion components*. Diffusion maps preserve and resolve local distances, while outliers have little influence on the projection. Simultaneously, the intrinsic noise of the data is reduced in each of the diffusion components, because the noise term is marginalized out by the diffusion process.

Here, we recapitulate the symmetric, locally scaled version of diffusion maps (cf. Haghverdi et al. [2016]). Let $\mathbb{D} = \{x_1, \ldots, x_n\}$ be a gene expression dataset with n samples and g genes, where $x_i \in \mathbb{R}^g \ \forall i \in \{1, \ldots, n\}$. The probability of reaching cell x_j from x_i in a Markov jump process is given by a wave function:

$$p(x_j|x_i) = \left(\frac{2}{\pi\sigma_i^2}\right)^{\frac{1}{4}} \exp\left(-\frac{||x_j - x_i||^2}{\sigma_i^2}\right).$$

There, σ_i is the Gaussian kernel width for cell x_i. For two neighboring cells x_i and x_k we compute the locally scaled kernel K as overlap integral over all possible positions x as

$$K(x_i, x_j) = \int_{-\infty}^{\infty} p(x|x_i)p(x|x_j)\mathrm{d}x = \left(\frac{2\sigma_i\sigma_j}{\sigma_i^2 + \sigma_j^2}\right)^{\frac{1}{2}} \exp\left(-\frac{||x_i - x_j||^2}{2(\sigma_i^2 + \sigma_j^2)}\right).$$

By normalizing $K(x_i, x_j))$ for sampling density $Z(x)$, we obtain

$$W_{ij} = \frac{K(x_i, x_j)}{Z(x_i)Z(x_j)}, \text{ where}$$
$$Z(x) = \sum_{k=1}^{n} K(x, x_k).$$

Next, we construct a right-stochastic matrix T_{ij}^{asym} by applying a row-normalization $\tilde{Z}_i$ to the matrix W as follows:

$$T_{ij}^{\text{asym}} = \frac{1}{\tilde{Z}_i} W_{ij}, \text{ where}$$
$$\tilde{Z}_i = \sum_j W_{ij}.$$

In general, T^{asym} is a normalized Laplacian. Next, the T^{asym} is transformed into a symmetric matrix T^{sym} while preserving the spectrum of T^{asym}. The proof of the following lemma 2.1 can be found in von Luxburg [2007].

Lemma 2.1. *Let T^{asym} be a normalized Laplacian, λ an eigenvalue of T^{asym}, and u the corresponding right eigenvalue. Then, λ is an eigenvalue of T^{asym}, and u the corresponding*

right eigenvalue, if and only if λ is eigenvalue of T^{sym} with eigenvector w with entries $w_i = \tilde{Z}_i^{\frac{1}{2}} u_i$.

From the lemma follows directly a construction for T^{sym} via

$$T_{ik}^{\text{sym}} = \tilde{Z}_i^{-\frac{1}{2}} W_{ik} \tilde{Z}_i^{\frac{1}{2}}.$$

In the last step, the eigenvalue decomposition of $T^{\text{sym}} \in \mathbb{R}^{n \times n}$ reads

$$T^{\text{sym}} = \Gamma^T \Lambda \Gamma.$$

The first eigenvector ψ_0, which corresponds to the eigenvalue $\lambda_0 = 1$, is removed. Thus, the set of $\{\psi_1, \ldots, \psi_n\}$ are called *diffusion components*.
The theory on diffusion maps on scRNA-seq data will be extended by theory on diffusion pseudotime (see section 2.9).

t-distributed Stochastic Neighborhood Embedding (t-SNE)

t-SNE [van der Maaten and Hinton, 2008] is a nonlinear dimensionality reduction method that uses a similarity measure to preserve local structures in the data. In brief, it maximizes the probability that two points are adjacent in a low dimensional space if they are proximal in high dimensional space. After projecting the points into a low dimensional space, the resulting similarity of all points should correspond to the similarity in the original space. The optimal placement of all points is achieve by minimizing the Kullback-Leibler divergence [Kullback and Leibler, 1951] of the joint probability distributions in high-dimensional and low-dimensional space. t-SNE preserves local structures well and clusters data into closely related and distantly related parts. Distances between clusters in the low-dimensional space cannot be interpreted as large-scale structure is not accounted for in the model. In recent years, t-SNE has become a popular method to visualize scRNA-seq data.

Let $\mathbb{D} = \{x_1, \ldots, x_n\}$ be a gene expression dataset with n samples and g genes, where $x_i \in \mathbb{R}^g \ \forall i \in \{1, \ldots, n\}$. The low-dimensional representation is denoted by $\mathbb{Y} = \{y_1, \ldots, y_n\}$, where $y_i \in \mathbb{R}^m \ \forall i \in \{1, \ldots, n\}$ and $m \in \{2, 3\}$. The conditional probability of x_j given x_i reads

$$p_{j|i} = \frac{\exp\left(-||x_i - x_j||^2 / 2\sigma_i^2\right)}{\sum_{k \neq i} \exp\left(-||x_i - x_k||^2 / 2\sigma_i^2\right)}, \tag{2.7}$$

where σ_i is the variance of a Gaussian centered in x_i. The joint probability p_{ij} is symmetrized and scaled as $p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n}$. The scaling ensures that the sum over all joint

probabilities is not-vanishing as $\sum_j p_{ij} > \frac{1}{2n}$ for all $x_i \in \mathbb{D}$. Therefore, every data point x_i contributes to the probability distribution. The variance σ_i^2 is computed based on the *effective number of neighbors*, called *perplexity* $Perp(P_i)$ (see van der Maaten and Hinton [2008] for details). Thus, the scaling is different for every data point.
In the low dimensional space, the joint probability q_{ij} follows a Student's t-distribution:

$$q_{ij} = \frac{\left(1 + ||y_i - y_j||^2\right)^{-1}}{\sum_{k \neq l} \left(1 + ||y_k - y_l||^2\right)^{-1}}. \tag{2.8}$$

The Student's t-distribution has a 'heavier tail' than the Normal distribution, but it disfavors large-scale distances as well. The objective function of the t-SNE algorithm is the Kullback-Leibler divergence C:

$$C = \mathrm{KL}(P||Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}}.$$

Thus, t-SNE optimizes the distributional equality of the high dimensional probability density function p (see eq. 2.7) and the density function q (see eq. 2.8) of the low-dimensional representation. Finally, the t-SNE embedding is performed by an iterative numeric algorithm using stochastic gradient descent. The embedding $\mathbb{Y}^{(t)}$ at iteration $t \in \{1, \ldots, T\}$ is given by

$$\mathbb{Y}^{(t)} = \mathbb{Y}^{(t-1)} + \eta \frac{\partial C}{\partial \mathbb{Y}} + \alpha(t) \left(\mathbb{Y}^{(t-1)} - \mathbb{Y}^{(t-2)}\right), \tag{2.9}$$

with the optimization parameters η (learning rate) and $\alpha(t)$ (momentum).

Uniform Manifold Approximation and Projection (UMAP)

UMAP [McInnes and Healy, 2018] is a manifold learning technique that uses previous work on Laplacian eigenmaps [Belkin and Niyogi, 2002, 2003] and algebraic topology to derive a low-dimensional representation of a dataset. The authors assume that a dataset can be approximated as Riemannian manifold in the high-dimensional space or, in other words, as a single k-nearest neighbor graph. In contrast to t-SNE, UMAP better preserves the global structure and is not restricted to the the dimensions 2 and 3.
Let a dataset $\mathbb{D} = \{x_1, \ldots, x_n\}$ with n samples be uniformly distributed on the manifold $\mathcal{M} \in \mathbb{R}^d$ with d features. Furthermore, let g be a Riemannian metric. Then, $(\mathcal{M}, g)$ is a Riemannian manifold. Given the uniform sampling of the dataset $\mathbb{D}$ from the manifold $\mathcal{M}$, the authors derived a geodesic distance metric for each point $x_i \in \mathbb{D}$. In particular, a sphere of fixed volume centered at x_i contains approximately the same number of nearest neighbors and, conversely, a neighborhood of k-nearest neighbors to x_i determines an individual distance for each x_i that is approximately the same for all $x_i \in \mathbb{D}$. In order to

integrate this family of metric spaces into a global structure, they employed the theory of simplicial fuzzy sets [Friedman, 2012] and relied heavily on the self-published work of Spivak [2009] over metrics on simplicial fuzzy sets.

Briefly, a fuzzy set consists of a reference set A and a membership strength function $\mu : A \to [0, 1]$. For example, let $S_k^{d-1}(x_i)$ a sphere in R^d with fixed radius ρ_i, that contains approximately k-nearest neighbors of x_i. For UMAP, the membership strength for all points within the sphere equals 1, whereas the membership strength of points outside the sphere decreases exponentially. Moreover, the fuzzy set that contains $S_k^{d-1}(x_i)$ also contains neighboring points outside the radius ρ_i, but with decreasing membership strength. Thus, UMAP takes into account larger distances than t-SNE.
Two fuzzy sets (A, μ) and (A, ν) are comparable, if they have the same reference set. For UMAP, the authors optimize the cross-entropy C of two fuzzy sets (A, μ) and (A, ν) as follows:

$$C\left((A,\mu),(A,\nu)\right) \triangleq \sum_{a\in A} \mu(a)\log\left(\frac{\mu(a)}{\nu(a)}\right) + (1-\mu(a))\log\left(\frac{1-\mu(a)}{1-\nu(a)}\right). \tag{2.10}$$

Learning a low-dimensional manifold that represents the high-dimensional manifolds involves, similar to t-SNE, the optimization of the cross-entropy C in eq. 2.10 via stochastic gradient descent (see eq. 2.9). We will use the UMAP dimensionality reduction to visualize complex tissue data in Chapter 4.

Other methods

Given the numerous ways to visualize data and use dimension, we will mention some less common dimensionality reduction techniques. The spanning-tree progression analysis of density-normalized events *SPADE* [Qiu et al., 2011] includes both clustering and visualization. However, it is outdated due to its unstable behavior. *Isomap* [Tenenbaum et al., 2000] is a nonlinear dimensionality reduction method. It uses geodesic distances induced by a nearest neighbor graph. *Force-directed graph layouts* [Islam et al., 2011, Jacomy et al., 2014] are used to draw graphs in a visually pleasing way. In the case of scRNA-seq data we use a k-nearest neighbor graph. Attractive forces a used to attract the endpoints of the graph while repulsive forces are applied between all pairs of cells. The graph is then simulated as physical system and the result is visualized. Self-organizing maps (*SOM*) [Kohonen, 1982] is an artificial neural network approach to learn the two-dimensional representation of a high-dimensional manifold.

2.8 Clustering

The term clustering describes a grouping of similar observations in a dataset and is closely related to classification. As such, it is a classical machine learning task (cf. Murphy [2012] Chapter 25 for a comprehensive overview). In scRNA-seq, we use clustering to discover abundant and rare cell types, as well as low quality cells or outliers. It should be noted that discovering cell types is typically an unsupervised clustering task and validating these discoveries using a ground truth from complementary data can be difficult. Cell types are typically validated using literature derived marker genes. However, cell type specific marker genes are not always available and a stringent classification into discrete cell types does not apply for cells undergoing continuous changes as in developmental processes (cf. section 2.9). Therefore, we have to qualitatively assess the result of any employed clustering algorithm. Currently, the choice of the 'best' clustering algorithm is under debate and depends on both the task and data structure [Duò et al., 2018, Wiwie et al., 2015]. A more detailed review on clustering in scRNA-seq can be found in Andrews and Hemberg [2018b] and Kiselev et al. [2019].
In scRNA-seq, we are confronted with a plethora of traditional and new clustering tools especially in clustering [Zappia et al., 2018], such that describing them is infeasible. Instead, we focus on a few commonly used tools and review their advantages and disadvantages.

k-means

k-means [MacQueen, 1967] is a general purpose strategy in machine learning to partition data into a fixed number of groups. Its aim is to find a partition such that the distance of assigned cells to the respective cluster center is minimal.
Let $\mathbb{D} = \{x_1, \ldots, x_n\}$ a dataset with n observations and $x_i \in \mathbb{R}^g$ an observation in the g-dimensional feature space R^g. In this thesis, we have g genes as features, but we can apply *k-means* similarly on a low-dimensional embedding. Starting with k randomly chosen starting points (the 'centroids' μ_k) in the g-dimensional space, it assigns each observation x_i to the nearest center:

$$z_i = \arg\min_k ||x_i - \mu_k||^2.$$

Then, it updates the positions of all k centers by taking the mean of all points per assigned cluster:

$$\mu_k = \frac{1}{N_k} \sum_{i:z_i=k} x_i,$$

where N_k is the cluster size. Assignment and center update are repeated until convergence. *k-means* is a fast and simple algorithm, which works best if all clusters are $g-1$-spheres

$\mathcal{S}^{g-1}$ and equally-sized. However, the number of clusters is *a priori* unknown in most cases. Furthermore, the initialization of the centroids may influence the results. In order to alleviate this issue, *SC3* [Kiselev et al., 2017] averages over the results using several cluster numbers k to determine a robust result.

Hierarchical Clustering

In contrast to *k-means* clustering, the number of cluster is not predetermined in hierarchical clustering. Instead, hierarchical clustering creates a relationship between clusters with different *heterogeneity*. In general, we distinguish between *agglomerative* (bottom-up) and *divisive* (top-down) clustering algorithms (cf. Murphy [2012]). Both methods use a dissimilarity matrix as input. In the top-down approach, groups are split for example using the Ward's minimum variance method [Ward Jr, 1963]. Briefly, the Ward's minimum variance method merges two clusters if the increase of total within-cluster variance is minimal. In the bottom-up approach, one starts with k clusters and iteratively merge the most similar clusters. The result of *k-means* clustering is often used to define the starting clusters.
The clustering result is usually visualized as dendrogram and a 'cut' parameter determines the number of clusters after the hierarchical relationship is inferred. It should be noted that hierarchical clustering always produces a hierarchy, even if the input data is random. Thus, choosing the 'correct' number of clusters is not trivial.

Density-Based Spatial Clustering of Applications with Noise (DBSCAN)

DBSCAN [Ester et al., 1996] is a density-based clustering algorithm. In contrast to *k-means*, DBSCAN works without assumptions on the shape, size, or number of the clusters. As input parameters, DBSCAN uses an ε-neighborhood and an minimum number of neighbors *minPts*, which determine the density threshold of the algorithm. If a point p has more than *minPts* neighbors closer than ε, it is a *core point*. If a point q has less than *minPts* neighbors, but is within a ε-neighborhood of a core point, it is a *border point* and still belongs to the cluster. Points that fall into neither of the above categories are classified as noise.
DBSCAN scales linearly with the number of observations as it visits every point only once. Furthermore, DBSCAN determines the number of clusters automatically. Thus, is particularly useful in large-scale datasets with various cell types. It is applied to cluster complex datasets with various cell types as the mouse retina [Macosko et al., 2015].

Louvain

Representing a dataset as a graph with edges between data points (*i.e.* connecting nearest neighbors) allows us to extend the density-based clustering approach to 'community detection'. In our case, transcriptionally similar cells are neighbors. The louvain algorithm [Blondel et al., 2008] partitions a graph into densely-connected sub-graphs via a metric called *modularity* [Newman and Girvan, 2004]. Similar to DBSCAN, clusters may have different shapes, sizes, and number. The only optimization parameter for louvain clustering is *resolution*. As it scales favorably with the number of cells, it applies well to datasets of up to a million data points [Wolf et al., 2018a]. Further, single-cell methods like *PhenoGraph* [Levine et al., 2015], `Seurat` v.1.4 [Butler et al., 2018] and `Scanpy` [Wolf et al., 2018a] use the louvain algorithm for clustering. Hence, community detection algorithms are applied to solve various clustering problems in scRNA-seq. Recently, Traag et al. [2018] present the *Leiden* algorithm as improved version of the louvain algorithm. The authors showed that the louvain algorithm has connectivity issues and may yield arbitrarily badly connected communities.

Other methods

We close the section on clustering by mentioning a few other clustering methods. Most clustering methods are extensions of *k-means*, hierarchical clustering, or DBSCAN. For example, *RaceID* uses *k-means* clustering combined with outlier detection to find rare cell types [Grün et al., 2015]. Ho [2017] created Clustering through Imputation and Dimensionality Reduction (CIDR), which is based on hierarchical clustering. Further, *GiniClust* identifies genes with skewed gene expression patterns with the Gini index and applies DBSCAN on this preselection to identify rare cell types [Jiang et al., 2016].

2.9 Pseudotime

Developmental processes such as cell differentiation and cell fate decisions involve a gradual change in the gene expression pattern along the developmental trajectory. The most famous metaphor for cell differentiation is Waddington's landscape [Waddington, 1957]. There, a cell is rolling down a furrowed hill in which each channel corresponds to a different fate. The act of rolling corresponds to cell differentiation over time. Using single-cell transcriptomics, we can order cellular gene expression profiles based on similarity to one another. This ordering is often interpreted as the intrinsic temporal order of developing and differentiating cells. Pseudotime methods have created important insights to cell fate

decision making or lineage segregation [Trapnell et al., 2014].
Widely used approaches that construct pseudotemporal orders from gene expression profiles use for example minimum spanning trees (Monocle [Trapnell et al., 2014], Slingshot [Street et al., 2018]), average graph distance (Wanderlust/Wishbone [Bendall et al., 2014, Setty et al., 2016]) or a Markov-jump-process-based metric in diffusion space (diffusion pseudotime [Haghverdi et al., 2016]). Saelens et al. [2018] provide a comprehensive overview and benchmarking of pseudotime algorithms on simulated and real data, which consist of a multitude of different topologies.
Importantly, pseudotime methods are linked to trajectory inference methods that are used to learn the manifold of a lineage specification process. Such a manifold may reflect a transformation from one cell type to another, a cell-fate choice with two or more options (*i.e.* branching), a ring for recurrent processes such as cell cycle, or more complex topologies such as the hematopoietic lineage hierarchy.

Diffusion pseudotime (DPT)

In section 2.7, we introduced *diffusion maps* as a nonlinear dimensionality reduction method based on random walks. Here, we describe the concept of diffusion pseudotime [Haghverdi et al., 2016] that directly builds upon diffusion maps. In contrast to other pseudotime algorithms, which require a root cell to start from, DPT is a distance metric, where distances between arbitrary cells are computed and root or tip cells are defined afterwards. DPT corresponds to the sum of all possible random walks of any length between two cells and is robust to noise (see Haghverdi et al. [2016] for details).
Let T be a symmetric transition matrix as derived in lemma 2.1. Further, let $f(t) \in \mathbb{R}^n$ be a probability density with initial density $f(0)$ and $t \in \mathbb{N}$ be a discretized time step. The probability density $f(t)$ changes over t time steps as

$$f(t) = f(t-1)T = f(0)T^t.$$

By construction of T, its largest eigenvalue $\lambda_0 = 1$ with corresponding eigenvector ψ_0, whose entries $\psi_{0i} = \tilde{Z}_i$. Next, the authors exclude the possibility to stay in the current state by subtracting largest eigenvalue from the spectrum of T. This step also removes the largest eigenvalue λ_0 from the spectrum of T. The matrix M denotes the sum of all possible random walks for all cells:

$$\begin{aligned} M &= \sum_{t=1}^{\infty} (T - \psi_0 \psi_0^T)^t \\ &= \left(I - (T - \psi_0 \psi_0^T)\right)^{-1} - I, \end{aligned}$$

using the properties of the Neumann series. It must be noted that Neumann series only converges, if the absolute value of all eigenvalues is smaller than one. Then, the matrix entry M_{ik} denotes the transition of cell x_i to x_k:

$$\begin{aligned} M_{ik} &= \sum_{t=1}^{\infty} (T - \psi_{0i}\psi_{0k}^T)^t \\ &= \sum_{t=1}^{\infty} \sum_{j=1}^{n-1} \lambda_j^t \psi_{ji} \psi_{jk}^T \\ &= \sum_{j=1}^{n-1} \frac{\lambda_j}{1-\lambda_j} \psi_{ji} \psi_{jk}^T, \end{aligned}$$

where $\lambda_j \in \mathbb{R}$ and $\psi_j \in \mathbb{R}^n$ denote the j^{th} eigenvalue and the corresponding eigenvector of T, respectively.
In the final step, let a random walk start in cell x_i, *i.e.* $f(0) = \delta(x_i)$. Then, the *diffusion pseudotime* DPT is a distance metric defined by

$$\begin{aligned} \mathrm{DPT}(x_i, x_k) &= \sqrt[2]{\sum_j (M_{ij} - M_{kj})^2} \\ &= \sqrt[2]{\sum_{j=1}^{n-1} \left(\frac{\lambda_j}{1-\lambda_j} \right)^2 (\psi_{ji} - \psi_{jk})^2}. \end{aligned}$$

In the original version of the publication, DPT is used to detect a single branching event. Moreover, we can couple DPT with an arbitrary manifold inference method. In the following section, we describe a unifying approach that captures a multitude of complex topologies and applies DPT to draw gene expression trajectories.

Partition-based graph abstraction (PAGA)

Partition-based graph abstraction (PAGA) [Wolf et al., 2018b] uses a graph representation of single-cell data and its partition into clusters to learn the data manifold. Usually, gene expression data are represented as a k-nearest neighbor graph and the louvain clustering is the obvious choice of clustering method. Using these inputs, PAGA circumvents the problem of learning the data manifold based on a one-dimensional pseudotime metric. Moreover, PAGA is more robust to sampling bias and small populations as often encountered in more complex datasets.
In the PAGA algorithm, each cluster is represented by a node (*coarse-graining step*). Second, PAGA connects each node by weighted edges which correspond to a statistical

measure of connectivity between clusters. Last, spurious edges are discarded.
As a result, PAGA creates a denoised topology with both connected and disconnected regions: a PAGA tree of arbitrary complexity. A key factor for PAGA is the resolution of the clusters — too few clusters results in every cluster connecting to every other cluster. However, PAGA deals well with multiple-resolutions and is stable even for small clusters. As a remark, PAGA uses diffusion pseudotime (DPT) as standard to draw gene expression trajectories along specified paths.
In summary, PAGA uses an unsupervised manifold learning approach based on graphs, which is independent of a (mechanistic) assumption on the topology. Together with DPT, we will employ the PAGA model in Chapter 4 to infer a complex lineage hierarchy in the small intestine of the mouse.

2.10 Differential Expression Analysis

Differential expression analysis employs a statistical model for gene expression and allows us to identify significant changes across conditions, such as healthy controls and disease samples. Furthermore, differential expression is used to identify genes whose expression is characteristic to a single group or cluster of cells (*i.e. marker genes*). We start with the example of two-group comparisons. Here, we define two groups I and J and assume that we can use the same model parameters to describe the properties of the two groups. This assumption is a so-called *null hypothesis* $\mathcal{H}_0$. Example null hypotheses include:

$$\begin{aligned} &\mathcal{H}_0 : \mu_i = \mu_j, \text{ or} \\ &\mathcal{H}_0 : \mu_i = \mu_j \wedge \sigma_i^2 = \sigma_j^2, \text{ or} \\ &\mathcal{H}_0 : \mu_i = \mu_j \wedge p_{\text{dropout},i} = p_{\text{dropout},j}, \end{aligned} \tag{2.11}$$

for every gene. Here μ is the mean expression, σ^2 the (empirical) variance, and p_{dropout} is the dropout probability. In general, the null hypothesis $\mathcal{H}_0$ is induced by the respective differential expression model. Furthermore, the more elaborate a differential expression is, the more parameters are required for the model fit. In order to generalize from two groups to multiple conditions, a differential expression model uses either a Wald test [Wald, 1943] on the set of model parameters θ, or uses a likelihood-ratio test (LRT), where it compares the values of the likelihood function for the null model and alternative model. In the Wald test, we test if the set of model parameters θ is the same as the expected set of parameters θ_0, *i.e.* $\mathcal{H}_0 : \theta = \theta_0$. In the LRT, we first create a likelihood function for the null model and the alternative model [Neyman et al., 1933]. Then, we evaluate both

functions for our data and compute the ratio. In order to compute a p-value of the LRT, we have to determine the distribution of values from the LRT, which is non-trivial in most cases [Neyman et al., 1933].

A simple differential expression test as the t-test can be evaluated gene-wise as illustrated in eq.s 2.12, 2.13. When the number of parameters in the model increases, the number of samples can be too small to estimate both parameters, mean, and variance. Therefore, models frameworks such as `DESeq2` [Anders and Huber, 2010] and `SCDE` [Kharchenko et al., 2014] borrow information across genes. In these settings, the corresponding null model is not fitted gene-wise but on all genes. Importantly, the null hypothesis $\mathcal{H}_0$ holds for the majority of genes such that a model fit has an appropriate reference. As a consequence, differential expression analysis should be applied on all expressed genes rather than the subset of highly variable genes, because the latter are already selected to be more variable than expected from stochastic noise modelling.

Statistical modelling for differential expression requires a complete trust in a dataset or, in other words, the 'garbage in, garbage out' principle. Its meaning is simple: If the input data is flawed, your result will be too. First, the statistical model should reflect the distributional properties of the data. Most scRNA-seq data are best described by a negative binomial model fit rather than zero-inflated negative binomial model [Vieth et al., 2017][3]. However, using the negative binomial with or without zero-inflation for the stochastic nature of gene expression has been enthusiastically discussed [Elowitz et al., 2002, Fan et al., 2016, Finak et al., 2015, Grün et al., 2014, Pierson and Yau, 2015, Vallejos et al., 2016]. Second, we need to make sure that we use a meaningful grouping in the analysis, especially when we interpret a clustering as cell types. To date, there is no accepted measure to discriminate if a clustering reflects distinct cell types or simply an overestimation of the heterogeneity in a cell type, which might be driven by a separate process such as cell cycle [Andrews and Hemberg, 2018b, Buettner et al., 2015].
Furthermore, when we apply a differential expression test to discover differences between clusters, the null hypothesis $\mathcal{H}_0$ as in eq. 2.11 is violated as the clustering already separated the groups of cells to maximize a dissimilarity. Therefore, p-values obtained in this situation are inflated [Zhang et al., 2018a,b].

In extension of differential expression analysis is to test gene expression changes along a (pseudo-)temporal axis (see section 2.9 on pseudotime algorithms). For example, Trapnell et al. [2014] used generalized additive model fits to determine changes along the pseudotem-

[3] A negative binomial model is an overdispersed Poisson model, *i.e.* we have the two parameters mean μ and variance $\phi\mu$. A zero-inflated negative binomial model accounts for excessive zeros in the data by coupling the negative binomial model to an binomial model with dropout parameter p.

poral axis. In order to cluster the genes, the authors computed pairwise distances based on Pearson's correlation for all trajectories. Then, the *k-medoids* algorithm grouped the trajectories roughly into 'upregulation' and 'downregulation' clusters. Another approach is to group the cells again into 'initial', 'intermediate' and 'terminal' developmental groups and apply classical differential expression tests [Haghverdi et al., 2016, Wolf et al., 2018a]. Lastly, statistical testing may correct for complex study design using a model matrix. To illustrate, we might have to take into account biological variation across individuals, variation across experiments, and changes in the experimental protocol. Moreover, many experiments place their controls and cells of interest on different sequencing runs, such that the biological effect confounds with a systematic batch effect. Lun and Marioni [2017] have shown that failing to correct for confounding increases false positive results. However, creating an experimental design without confounding of several sources of variation remains a challenge in itself.
In the following, we outline a few methods for differential expression, which we use in this thesis.

Welch's *t*-test

The simplest approach to differential expression analysis is the Welch t-test as implemented among others for gene ranking in `Scanpy` [Wolf et al., 2018a]. A Welch t-test is a modification of the Student's t-test [Welch, 1947]. Both test types use $\mathcal{H}_0 : \mu_i = \mu_j$ as null hypothesis, *i.e.* two groups I and J are normally distributed with equal means. However, Welch's t-tests also corrects for inequal variance and sample sizes.
For two groups I and J, let μ_i, μ_j and σ_i^2, σ_j^2 be the respective means and empirical variance with sample size n_i and n_j, respectively. Then, the score t is defined as

$$t = \frac{\mu_i - \mu_j}{\sqrt{\frac{\sigma_i^2}{n_i} + \frac{\sigma_j^2}{n_j}}}. \tag{2.12}$$

The score t is t-distributed with ν degrees of freedom, where

$$\nu = \frac{\left(\frac{\sigma_i^2}{n_i} + \frac{\sigma_j^2}{n_j}\right)^2}{\frac{\sigma_i^4}{n_i(n_i-1)} + \frac{\sigma_j^4}{n_j(n_j-1)}}. \tag{2.13}$$

In `Scanpy`, differentially expressed genes are ranked by the score t. As we already pointed out earlier, the normal distribution does not describe scRNA-seq data well. Therefore, we have to handle results from a t-test carefully and check them with a different method.

However, this test provides a fast and simple approach to compute the top differentially expressed genes, which is viable as strong differences between groups will be picked up reliably by all methods. In the case of marker gene detection we usually define the groups as a single cluster with all others as reference.

limma

We have introduced *limma* [Ritchie et al., 2015] in the context of batch effect correction in section 2.4. limma was originally designed as analysis pipeline for microarray data. It readily supports several covariates in the form of a model matrix for the linear model fit for differential expression analysis. In particular, limma fits a linear model for each gene, and borrows information across genes by setting a prior for the estimates of the gene-variance (empirical Bayes method, see Smyth [2011] for details). Thus, the shrinkage of gene-variance is less sensitive to outliers and allows estimating the variance at all when the sample number is insufficient to estimate variance. However, scRNA-seq data usually provides sufficient samples. For each gene and covariate of the linear model limma provides an empirical Bayes moderated t-statistic, an associated p-value, and a log-fold change. Importantly, limma provides multiple testing correction for every chosen contrast in the model matrix. It should be noted that limma expects normally distributed data that are best approximated by log-transformed data rather than count data in scRNA-seq. Then, limma estimates the mean-variance relation on the log-space.

Other methods

To close the section on differential expression tests, we summarize some more differential expression analysis approaches for scRNA-seq. It must be noted that the approaches to model differential expression are quite diverse. `SCDE` is one of the first differential expression models for scRNA-seq data [Kharchenko et al., 2014]. It uses a complex Bayesian framework of a zero-inflated negative binomial model. However, the implementation of `SCDE` is computationally costly. In the `PAGODA` framework of Fan et al. [2016] uses the same error model as SCDE, but is computationally more efficient. Further, `PAGODA` identifies overdispersed genes and groups them based on a PCA score. In contrast to other differential expression methods `PAGODA` also provides the possibility for functional interpretation of gene sets.

`DESeq2` was originally designed for bulk data and uses a negative binomial model [Anders and Huber, 2010]. Further, the mean-variance relation is derived from all genes. Andrews and Hemberg [2018a] identify differentially expressed genes with Michaelis-Menten kinetics in the *M3Drop* model. Here, mean expression level μ and dropout probability P_{dropout}

correlate via

$$P_{\text{dropout}} = 1 - \frac{\mu}{K_M + \mu},$$

where K_M is Michaelis-Menten constant. In this context, K_M denotes the mean expression required to detect a gene in half of the cells.

The Model-based Analysis of Single-cell Transcriptomics (*MAST* [Finak et al., 2015]) models dropout and expression levels simultaneously. The approach is called *Hurdle* model. The dropout component follows a binomial model and the mean expression level is modeled with a normal distribution. Thus, *MAST* works well with log-normalized count data in the case of scRNA-seq. More generally than *MAST*, *scDD* is a Bayesian framework that identifies differential distributions in a single cell dataset [Korthauer et al., 2016]. In particular, *scDD* models the number of modes such as unimodal versus bimodal distributions and compares these distributions across conditions. Hence, *scDD* detects changes in modality and proportional shifts between modes.

Lastly, *BASiCS* is a Bayesian framework to model differences in both mean and variance [Vallejos et al., 2016]. The model is able to detect biologically relevant changes in variance, for example during aging [Martinez-Jimenez et al., 2017].

Chapter 3

Batch effect quantification

> The most dangerous phrase in the language is, "We've always done it this way."
>
> Grace Hopper

In this chapter, we describe a new method for batch effect quantification, which detects batch effects robustly and sensitively. Using this method, we gain new insight to batch effect quantification in single-cell RNA-seq and its application to simulated and real data. The chapter is based on and partly identical to the following publication at Nature Methods: **M. Büttner***, Z. Miao*, F. A. Wolf, S. A. Teichmann, and F. J. Theis. *A test metric for assessing single-cell RNA-seq batch correction.* Nature Methods, 43-49, 2019 (* indicates equal contribution)

The goal of this project is to create an unbiased, sensitive and robust batch effect quantification in scRNA-seq data that outperforms existing methods. Further, an unbiased measure allows investigating the properties of batch effects in scRNA-seq data. Moreover, a unbiased, robust and sensitive batch effect measure allows benchmarking which batch effect correction method performs best to remove batch effects.
For this purpose, we established a methodology to test for batch effects and to compare results from different batch effect metrics. Then, we studied the properties of the test metrics on several simulated datasets, which aim to mimic the properties of single-cell RNA-seq data. The results assured us to apply our test metric for several normalization and batch effect correction methods on single-cell RNA-seq data of mouse embryonic stem cells (mESCs). Here, we focussed first on a pair of technical replicates, where a single cell culture was split into two consecutive sequencing runs. Adding complexity, we investigated batches of mESCs grown in different cell cultures. Going further, we benchmarked batch effect correction on cells from several mouse embryonic development stages to demonstrate the applicability of our method in integrating numerous studies. Lastly, we shifted focus to use our method to study inter-individual variability in peripheral blood mononuclear cells (PBMCs).

3.1 A new test metric for batch effect quantification

Here, we introduce the *k-nearest neighbor batch effect test (kBET)*.
Briefly, we reintroduce some notation (see Chapter 2). Let the full gene expression dataset $\mathbb{D} = \{x_1, \ldots, x_n\}$, where $x_i \in \mathbb{R}^g$ and $\mathbf{X} \in \mathbb{R}^{n \times g}$ the corresponding gene expression data matrix with n samples and g genes. In a single-cell RNA-seq dataset $\mathbb{D}$, each sample has meta-information such as cell type, FACS gate or the batch i it was measured in.
The batch variable i has l categories such that n_i denotes the number of samples in batch i, $f_i = \frac{n_i}{n}$ the *global fraction* of samples in batch i, and $\nu = (n_1, \ldots, n_l)$ the batch configuration of all samples. Furthermore, we define $\tilde{f}_i$ as the *local fraction* of samples in batch i in some subset $N \subset \mathbb{D}$. In particular, we consider subsets of k nearest neighbors. Then, we formulate the null hypothesis $\mathcal{H}_0$ of *'data being well-mixed'*, *i.e.* the absence of a batch effect, as

$$\mathcal{H}_0 : \tilde{f}_i = f_i \quad \forall\, i \in \{1, \ldots, l\} \,\forall \text{ subsets } N \subset \mathbb{D}. \tag{3.1}$$

We illustrated the principal idea in Fig. 3.1. In order to statistically test the null hypothesis, let us define a neighborhood subset

$$N_j = x_j \cup \{x_s | s \text{ is among } k-1 \text{ nearest neighbors of } j\}\,.$$

Nearest neighbors are computed with the cover-tree algorithm (*FNN R package*). To optimize computation efficiency, we pre-compute the first 50 eigenvectors of the largest eigenvalues with the *svd* function and use the reduced dataset to find nearest neighbors. Let n^k_{ji} denote the number of cells in batch i that are in subset j of size k. Testing the null hypothesis involves two steps:

(i) Test the null hypothesis in each subset N_j of a given sequence of subsets. In each subset N_j, this amounts to testing whether the distribution of n^k_{ji} with respect to i matches the distribution under the null hypothesis.

(ii) Summarize the result of the sequence of tests by computing the average rejection rate S over all tests — a test statistic for the whole dataset. Hence, testing whether S exceeds a significance threshold allows to reject the null-hypothesis for the whole dataset.

Note that by performing these two steps, *kBET* goes beyond a standard test for homogeneity of subsets of a given dataset. In the following, we present three different ways to create a test statistic based on the null hypothesis (3.1).

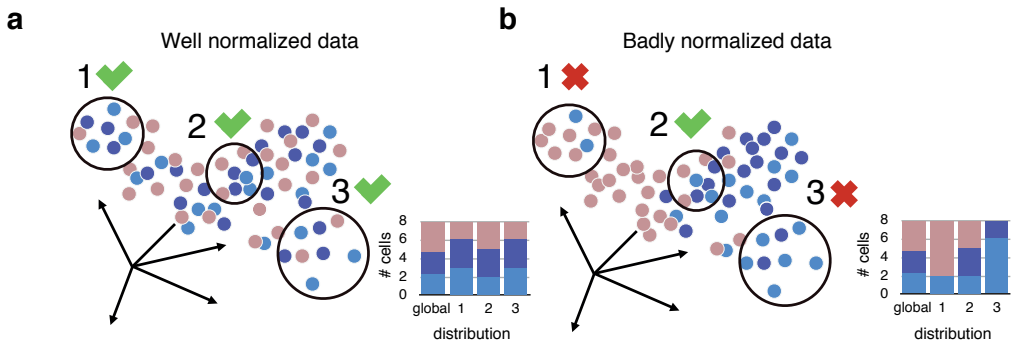

Figure 3.1: The concept of a k-nearest neighbor batch effect test (kBET). **a** In a dataset with replicates without batch effects, the fraction of the batch label does not differ from the global label distribution in any neighborhood. **b** If a dataset has a batch effect, data points from respective batches tend to cluster with their 'peers'. Then the fraction of batch labels differs considerably in arbitrarily chosen neighborhoods.

3.1.1 χ^2-based test

In the limit of high values of k, the residuals of the test statistic is normal distributed with respect to i. Then, we can use Pearson's χ^2 test, the test statistic of which reads

$$\kappa_j^k = \sum_{i=1}^{l} \frac{\left(n_{ji}^k - f_i \cdot k\right)^2}{f_i \cdot k} \sim \chi_{l-1}^2,$$

where χ_{l-1}^2 denotes the χ^2-distribution with $l-1$ degrees of freedom. The p-value for each κ_j^k is computed as

$$p_j^k = 1 - \mathrm{F}_{l-1}\left(\kappa_j^k\right),$$

where $F_{l-1}(x)$ denotes the cumulative distribution function of the χ^2-distribution with $l-1$ degrees of freedom.

3.1.2 Exact test

For small values of k, κ_j^k does not necessarily follow a χ^2-distribution. Hence, we also evaluate the performance of Pearson's χ^2 test with the exact multinomial test and its approximation, a likelihood ratio test [Cressie and Read, 1989]. In brief, the exact probability of an observed configuration $\mathbf{n}_j^k = \left(n_{j1}^k, \ldots, n_{jl}^k\right)$ under the null hypothesis $\mathcal{H}_0$ (see eq. 3.1) is

$$\mathrm{P}_0\left(\mathbf{n}_j^k\right) = k! \prod_{i=1}^{l} \frac{f_i^{n_{ji}^k}}{n_{ji}^k!}. \tag{3.2}$$

The probability in eq. (3.2) is used to compute the p-value as the sum over all possible configurations y whose probability is lower than $\mathrm{P}_0\left(\mathbf{n}_j^k\right)$:

$$\mathrm{P}_0\left(\mathrm{sig}\right) = \sum_{y:\, \mathrm{P}_0(y) \leq \mathrm{P}_0\left(\mathbf{n}_j^k\right)} \mathrm{P}_0(y) \tag{3.3}$$

The exact test is in-feasible to compute for larger samples as the number of computations to carry out in eq. (3.3) scales with $\binom{k+l-1}{l-1}$.

3.1.3 Likelihood-ratio test

To meet the need for a scalable solution, we also compared to the approximation of the exact probability in the form of a likelihood ratio test. Let us define an alternative hypothesis $\mathcal{H}_1$ with the adjusted empirical frequencies

$$g_i = \max\left(\frac{1}{k+d}, \frac{n_{ji}^k}{k+d}\right),$$

where d denotes the number of non-chosen categories (absent batches) in a particular subset $\mathcal{N}_j$, in contrast to the exact probabilities under the null hypothesis $\mathcal{H}_0$ given in eq. 3.2. The probability of the alternative hypothesis $\mathcal{H}_1$ reads

$$\mathrm{P}_A\left(\mathbf{n}_j^k\right) = k! \prod_{i=1}^{l} \frac{g_i^{n_{ji}^k}}{n_{ji}^k!},$$

i.e. the likelihood-ratio test statistics G^2 reads:

$$\begin{aligned} G^2 &= -2 \log\left(\frac{\mathrm{P}_0\left(\mathbf{n}_j^k\right)}{\mathrm{P}_A\left(\mathbf{n}_j^k\right)}\right) \\ &= -2 \sum_{i=1}^{l} n_{ji}^k \log\left(\frac{f_i}{g_i}\right) \longrightarrow \chi_{l-1}^2 \end{aligned}$$

as k increases [Wilks, 1938]. If G^2 gets large, the null hypothesis is rejected.

3.1.4 Optimal neighborhood size

The kBET rejection rate can be influenced by the number of data points for each subset tested. For very high and very low k (*i.e.* $< 5\,\%$ or $> 80\,\%$ sample size), we observe a decrease in the rejection rate of the null hypothesis, but there is a plateau of similar results

for 'intermediate' values. We conclude that there is an interval of neighborhood sizes k for which kBET returns a maximal rejection rate and that this interval is dependent on the data.

Data: a set of k_0-nearest neighbors, batch labels, bounds a and b
Result: optimal neighborhood size k
while $|a - b| > 10$ *or* $|f(a) - f(b)| > 0.01$ **do**
  Set $c = 0.5 \cdot (b - a)$;
  Compute $f(c)$;
  if $f(c) > f(a)$ **then**
    Set $a = c$;
    Repeat bisection;
  else
    Set $b = c$;
    Repeat bisection;
  end
end
Return a

Algorithm 1: Recursive version of the bisection method in kBET. The rejection rate computed by kBET is denoted as the general function *f*. Bisection is initialized with $a = 10$ and $b = k_0$.

We employ a recursive bisection to determine the most suitable neighborhood size (see algorithm 1): for an optimal neighborhood size k, the kBET rejection rate is highest. Further, we assume that there is an interval of neighborhood sizes k for which we obtain maximal rejection rates. The complexity of the bisection algorithm is dependent on the number of samples and ranges between $\mathcal{O}\left(n^2\right)$ and $\mathcal{O}(n)$ (Fig. 3.2).

3.1.5 Isolated cells and adaptive frequencies

When the algorithm determines the local structure of the data as neighborhoods, we implicitly assume that every cell has mutual nearest neighbors. That means we expect that every cell can be found in more than one neighborhood of size k. If a cell is more distant to its k-1 nearest neighbors than those neighbors among each other, we call this cell an isolated cell or outsider. Such a cell does not contribute to any other neighborhood and the composition of a tested neighborhood will not change if an isolated cell is removed from the data entirely. However, a considerable amount of isolated cells bias the observed outcome if their label composition is significantly different from the global batch label distribution. We introduce adaptive expected frequencies that are computed without the isolated cells to adjust for this effect.

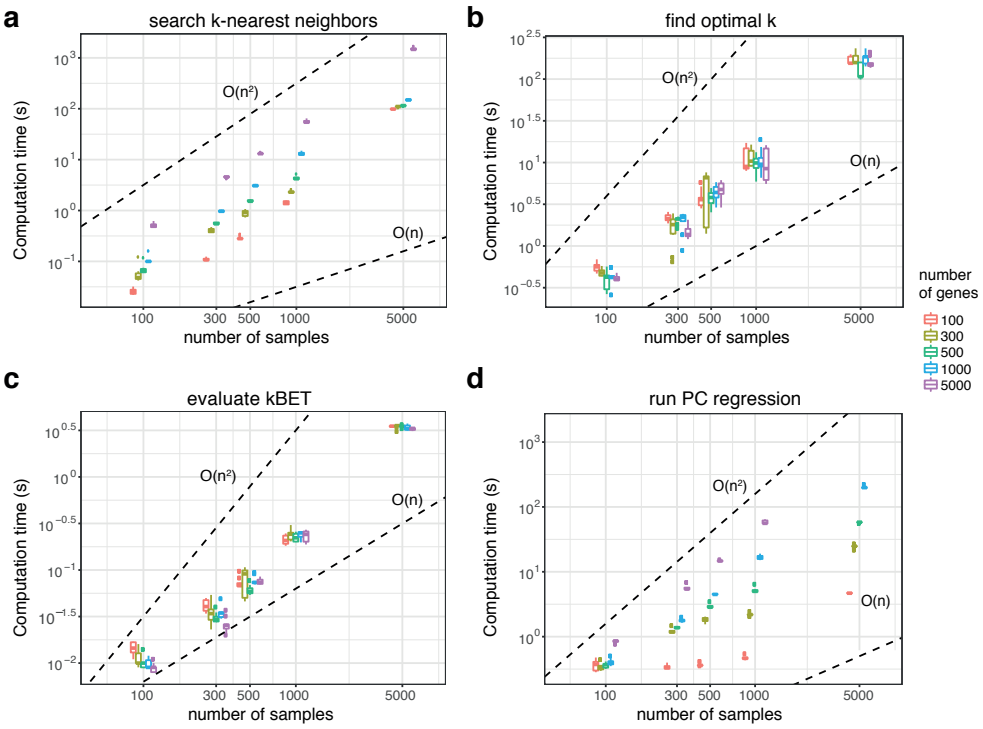

Figure 3.2: Complexity of kBET and PC regression Comparison of kBET and PC regression runtime, averaged over 10 runs. **a-c** components of kBET function, **d** PC regression. **a** k-nearest neighbor search has a complexity of approximately $\mathcal{O}\left(n^2\right)$ and is the slowest component of the kBET function. **b** Finding an optimal neighborhood size k has a complexity of approximately $\mathcal{O}(n)$. **c** Evaluation of the batch effect test is the fastest component. **d** PC regression involves a PCA first and has a complexity of approximately $\mathcal{O}\left(n^2\right)$.

Let $N_{iso} = \{x_j | j \text{ has no mutual nearest neighbor}\}$ and n_i^{iso} denotes the number of cells in batch i that are isolated. Then, we apply Pearson's χ^2 test to determine if a certain batch label class is overrepresented:

$$\kappa^{iso} = \sum_{i=1}^{l} \frac{\left(n_i^{iso} - f_i \cdot k\right)^2}{f_i \cdot k} \sim \chi^2_{l-1},$$

where the p-value reads $p^{iso} = 1 - F_{l-1}\left(\kappa^{iso}\right)$, where $F_{l-1}(x)$ denotes the cumulative distribution function of the χ^2-distribution with $l-1$ degrees of freedom. In case that $p^{iso} < \alpha = 0.05$, we compute adapted expected frequencies on the basis of $\nu = \left(n_1 - n_1^{iso}, \ldots, n_l - n_l^{iso}\right)$ the batch configuration excluding isolated cells.

3.1.6 A significance test of the batch effect estimate

Using the mean of rejected tests with spatial choice of subset and with stratified label sampling (*i.e.* the *expected* result if no batch effect is present), we construct a second test to decide whether there is a batch effect or not.

Definition 3.1. *Let $Y_j^k \sim Ber(p)$ the test result for some data point x_j,* i.e. *$Y_j^k \in \{0,1\}$ with p defined as the* success *probability. In our case, p is the probability for the rejection of the null hypothesis, which converges to the significance level α chose for the test of each subset: $p \longrightarrow_{m\to\infty} \alpha$ as k increases. Hence, we have $Var\left(Y_j^k\right) = p(1-p)$ for each subset $\mathcal{N}_j$. Let*

$$S^k = \frac{1}{m}\sum_{j=1}^{m} Y_j^k, \tag{3.4}$$

$$Var\left(S^k\right) = \frac{1}{m}p(1-p) \tag{3.5}$$

and $S^k \sim \mathcal{N}(\mu, \sigma^2 | 0 \leq S^k \leq 1)$ for $m \geq 30$.

We wish to assess whether $S^k \sim \mathcal{N}(\alpha, \frac{1}{m}\alpha(1-\alpha)|0 \leq S^k \leq 1)$ as imposed by the null hypothesis for *'well-mixed'* data of different batches.

Definition 3.2 (Truncated normal distribution density function)**.** *The cumulative density function of the truncated normal distribution is defined as follows:*

$$F\left(s|\mu,\sigma,\zeta,\eta\right) = \frac{\Phi(\xi)-\Phi(\zeta)}{\Phi(\eta)-\Phi(\zeta)}$$
$$\text{where } \xi = \frac{s-\mu}{\sigma}, \zeta = \frac{a-\mu}{\sigma}, \eta = \frac{b-\mu}{\sigma},$$

where Φ is the cumulative normal distribution function and a, b lower and upper bound of the support.

The cumulative density function F of the truncated normal distribution is a scaled version of the cumulative normal distribution Φ, such that at the lower and upper bound, $F(\zeta) = 0$ and $F(\eta) = 1$, respectively. We employ either the empirical mean and variance obtained by random assignment of labels to compute eqs. (3.4) and (3.5).

Lemma 3.1. *Let $\mathcal{H}_0$ be the null hypothesis as follows:*

$$S^k \sim \mathcal{N}(\alpha, \frac{1}{m}\alpha(1-\alpha)|0 \leq S^k \leq 1) \quad for \quad m \geq 30,$$

where α is the significance level. The p-value of the rejection rate S^k reads

$$p_{S^k} = 1 - F\left(S^k|0, 1, \zeta, \eta\right).$$

Then, the null hypothesis $\mathcal{H}_0$ is rejected, if $p_{S^k} < \alpha$.

Proof. S^k follows the distribution of the truncated normal distribution (definition 3.2) and the alternative hypothesis requires a one-sided test. □

We still observe the dependency of the rejection rate and neighborhood size k (see section 3.1.4). A criterion for the acceptance of the null hypothesis (*i.e.* there is no batch effect in the data) is that the null hypothesis is not rejected for all neighborhood sizes k. Nevertheless, it must be noted that there may be high rejection rates for small neighborhood sizes with no rejection of the null hypothesis for larger neighborhood sizes.

3.2 Prior approaches of batch effect quantification

3.2.1 Silhouette coefficient

The calculation of a silhouette aims to determine if a particular clustering has minimized within-cluster dissimilarity and maximized inter-cluster dissimilarity [Rousseeuw, 1987]. Let us assume there is a given clustering into more than one cluster. For each sample i, the silhouette width $s(i)$ is defined as follows:

Definition 3.3 (Silhouette width and silhouette coefficient)**.** *Let $a(i)$ be the average dissimilarity between cell i and all other data points of its cluster A. If cell i is the only observation in this cluster, set $s(i) := 0$. For all other clusters $C \neq A$, let $d(i, C)$ the average dissimilarity of cell i to all samples of C. There is some cluster B whose dissimilarity $d(i, B)$ is minimal: $b(i) := \min_C d(i, C)$, which is the 'neighboring' cluster to sample i. Then, the* silhouette width *$s(i)$ is the scaled difference of average dissimilarity within a cluster and the average dissimilarity to its 'neighboring' cluster:*

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

The mean of all silhouette widths $s(i)$ gives the silhouette coefficient *s.*

We adapted the calculation from *scone* R package [Risso et al., 2014]. We summarize the properties of the silhouette widths in the following lemma.

Lemma 3.2 (Properties of the silhouette width)**.** *Let clusters A and B contain more than one cell each. Then, the silhouette width*

(i) $s(i) \in [-1, 1]$,

(ii) $s(i) \longrightarrow 1 \, \forall i \in A$ *if two clusters A and B are separated,*

(iii) $s(i) \longrightarrow -1 \, \forall i \in A$ *if clusters A and B overlap, but cluster A is more dissimilar (and more dispersed) than cluster B,*

(iv) $s(i) \longrightarrow 0 \, \forall i \in A$*, if both clusters overlap and have roughly the same extend.*

Proof. For property (i), the lower bound is obtained, if $a(i) \gg b(i)$ and the upper bound is obtained, if $b(i) \gg a(i)$.
For property (ii), let $s(i) \longrightarrow 1$, if $b(i) \gg a(i) \, \forall i \in A$. As $b(i) := \min_C d(i, C)$ is already minimal per definition, all cells of cluster B are more dissimilar than the average cell in A. If cluster A and B are not separated, then there exists at least one cell j, where $a(j) \geq b(j) = \min_C d(j, C)$, but then $s(j) \leq 0$.
Property (iii) follows out of the proof for property (ii). Let $s(i) \longrightarrow -1 \, \forall i \in A$. Then, $a(i) \gg b(i)$, *i.e.* the average dissimilarity for cell i of cluster A is larger than the average dissimilarity of cell i to cluster B.
For property (iv), we have $a(i) \approx b(i) \, \forall i \in A$. If clusters A and B do not overlap, then there exists at least one cell j with $b(j) > a(j) \Rightarrow s(j) > 0$. Similarly, from property (iii) follows immediately, that clusters A and B have the same extend measured in terms of dissimilarity. □

Hence, we use the **absolute value of the silhouette coefficient** $|s|$ as indicator for presence or absence of batch effects.

3.2.2 Principal Component Regression

Principal component analysis (PCA) is an orthogonal transformation the data matrix $\mathbf{X} \in \mathbb{R}^{n \times g}$ into a set of linearly uncorrelated variables. The principal components (PCs) correspond to the eigenvectors of the covariance matrix $\mathrm{Cov}(\mathbf{X})$ of the data and are ordered by the explained variance of the data, *i.e.* by decreasing eigenvalues. Hence, PCA is a

popular linear dimension reduction technique aiming to maximally preserve variances in the data. We aim to construct a measure based on PCA that allows quantifying batch effects in a dataset.
If a batch effect is present in the data, it contributes to the variance.

Definition 3.4. *For each PC, Pearson's correlation coefficient with the batch covariate* B *(with l batch categories) reads*

$$r_i = corr(PC_i, B).$$

Lemma 3.3 (Equivalence of Pearson's correlation and coefficient of determination)**.** *For every linear model that explains a PC with batch covariate* B*, the coefficient of determination R^2 equals the squared Pearson's correlation coefficient:*

$$r_i^2 = R^2(PC_i, B),$$

and r_i is always positive.

Proof. In every linear model with a single dependent variable, the coefficient of determination R^2 is

$$r_i^2 = R^2(\mathrm{PC}_i, B).$$

Thus, the Pearson's correlation coefficient and coefficient of determination are equivalent. In the case of PCA, the eigenvector is well defined apart from its sign. Without loss of generality, we flip the direction of any PC such that it is positively correlated or uncorrelated to batch effects. □

Let us restrict our analysis to the top N PCs, where N represents a 'reasonable' trade-off between accuracy and computational cost and is typically much smaller than the total dimension i.e. number of genes.

Definition 3.5. *We define a set of indices* D *as the subset of $\{1, \ldots, N\}$, where the corresponding PCs correlate significantly with the batch covariate B.*

Here the most general approach to detect batch effects using PCs is to determine the cardinality of D. By construction, D is non-empty if there is at least one of the top N PCs that correlates significantly with the batch covariate. This implies a binary outcome that does not allow for a ranking of batch effect correction methods. However, the set D is essential to interpret the correlation of batch covariates and PCs and we will keep the notation for the following analyses.

A first intuition for a PCA based measure of batch effects is to determine the highest correlation and coefficient of determination with the batch covariate, respectively, in any of the PCs.

Definition 3.6 (Maximum correlation with batch effect)**.** *We define a batch effect measure by the maximum correlation and coefficient of determination, respectively, by*

$$r = \max_{i \in \{1,\dots,N\}} r_i \quad or$$
$$R = \max_{i \in \{1,\dots,N\}} R^2(PC_i, B).$$

More specifically, restricting the computation of the maximal correlation to the set of significantly batch-correlated PCs excludes non-significant results as

$$\tilde{r} = \max\left(\max_{i \in D} r_i, 0\right) \quad or$$
$$\tilde{R} = \max\left(\max_{i \in D} R^2(PC_i, B)\right).$$

However, both $\tilde{r}$ and $\tilde{R}$ neglect the order and thus importance of the PCs. For example, we have corrected our dataset with two different batch effect correction methods. Method A yields $\tilde{r}_A = 0.6$ (in PC_3) and method B yields $\tilde{r}_B = 0.8$ (in PC_8). If we only consider $\tilde{r}_A$ and $\tilde{r}_B$, method A would outperform method B, but this neglects the fact that PC_3 has greater importance in the dataset than PC_8.
Please note that the maximum-approach works well if the batch effect is linear and orthogonal to the biological signal. In this case, a single PC may reflect the batch effect completely. If the batch effect is non-linear, several PCs correlate significantly with the batch effect, and the correlation for the PCs is lower in general.

Apart from maximum-correlation or maximum-R^2, we also consider two index-based approaches. In the first approach, we use the rank of the PC with the highest correlation and coefficient of determination, respectively.
Another approach is to consider only the first PC that is significantly correlated with the batch effect — here we account more for the weights of the PCs.

Definition 3.7 (Index score from PC regression)**.** *The smallest element in set* D *reads* $d = \min_{i \in D} i$*. For index based batch quantification approaches, let us define a PC score* s

as follows:

$$
\begin{aligned}
&s = 1 - \frac{m-1}{N}, \textit{ where} \\
&m = \min\left(\arg\max_{i \in D} r_i, N+1\right), && \textit{(Approach 1)} \\
&m = \min\left(d, N+1\right). && \textit{(Approach 2)}
\end{aligned}
$$

Lemma 3.4. *The support of the PC score s given in definition 3.7 is $[0,1]$ and the function is linearly decreasing.*

Proof. s is a linear function and attains its extremal points for $m = 1$ and $m = N + 1$. If $m = 1 \Rightarrow\ s = 1$. If the set D is empty, none of the top N PCs is significantly correlated with the batch covariate B. Then, $m = N + 1$ and $s = 0$. □

Finally, we establish a variance-based approach.

Lemma 3.5 (Variance explained by batch effect)**.** *The total contribution of the batch effect to the variance in the transformed data matrix Y may be approximated by*

$$
Var(Y|B) = \sum_{i=1}^{g} Var\left(Y|PC_i\right) \cdot R^2\left(PC_i|B\right),
$$

where $Var\left(Y|PC_i\right)$ is the variance of Y explained by the i^{th} PC.

Proof. The set of PCs is orthogonal. Then, regressing the batch covariate B and the i^{th} PC returns the coefficient of determination $R^2\left(\mathrm{PC}_i|B\right)$. Following the law of total variance, this is an approximation of the variance explained by B in each PC (*principal component regression*, similar to McCarthy et al. [2017]). □

However, we can evaluate the significance of $R^2\left(\mathrm{PC}_i|B\right)$ using a linear regression model. For the case of two batches, the significance test equals a univariate t-test on the loadings of each PC split by batch covariate. A generalized model for several batches is the one-way analysis of variance (ANOVA) whose test statistic is F-distributed.
It must be noted that, as the number of features (genes) increases, the largest and smallest eigenvalue of the sample covariance matrix converge [Baik and Silverstein, 2006]. Consequently, $\mathrm{Var}\left(Y|B\right)$ decreases with the number of features as well and due to the high-dimensionality of scRNA-seq data, batch effects as defined above are difficult to interpret.

As alternative, we only use the explained variance with the most significant (*i.e.* highest) $R^2\left(\mathrm{PC}_i|B\right)$ as proxy for the batch effect.

Definition 3.8 (Approximation for variance explained by batch effect)**.** *Let $Var\left(Y|PC_i\right)$ for $i \in \{1,\ldots,N\}$ be the variance explained by the i^{th} PC. Then, we define the explained variance with the most significant (*i.e. *highest) $R^2\left(PC_i|B\right)$ as approximation for the variance explained by the batch effect:*

$$Var(Y|B) \approx Var\left(Y|PC_i\right), \quad i = \underset{j\in\{1,\ldots,N\}}{\arg\min}\; p\left(R^2\left(PC_j|B\right)\right)$$

Alternatively, in order to increase interpretability of the variance-based approach, we sum the variance of all significantly batch-correlated PCs and scale this sum by the total variance of the top N PCs (scaled variance explained by batch effect or 'scaled PC regression').

Lemma 3.6 (Scaled PC regression)**.** *Let $Var\left(Y|PC_i\right)$ for $i \in \{1,\ldots,N\}$ be the variance explained by the i^{th} PC, and let the set of significantly batch correlated components D be non-empty. Then, the scaled variance explained by batch effect,*

$$Var_{scaled}(Y|B) = \frac{\sum_{i\in D} Var(Y|PC_i)}{\sum_{i=1}^{N} Var(Y|PC_i)},$$

is as an upper bound for the batch effect in the data.

Proof.

$$\begin{aligned}
\mathrm{Var}(Y|B) &= \sum_{i=1}^{g} \mathrm{Var}\left(Y|\mathrm{PC}_i\right) \cdot R^2\left(\mathrm{PC}_i|B\right) \quad \text{(lemma 3.5)} \\
&< \sum_{i\in D} \mathrm{Var}\left(Y|\mathrm{PC}_i\right) \\
&< \frac{\sum_{i\in D} \mathrm{Var}(Y|\mathrm{PC}_i)}{\sum_{i=1}^{N} \mathrm{Var}(Y|\mathrm{PC}_i)},
\end{aligned}$$

since $\sum_{i=1}^{N} \mathrm{Var}(Y|\mathrm{PC}_i) < 1$ is the fraction of total variance explained by the first N PCs. □

Here, we massively over-estimate the variance explained by the batch effect. This scaling can be particularly useful to detect batch effects that are not orthogonal to the biological signal, *i.e.* when the batch effect is distributed over several components. However, it is still based on PCA and hence able to detect only linear batch effects.

3.3 Quantifying preservation of biological signals

The implicit constraint in batch correction is to preserve biological heterogeneity in the data, otherwise the optimal batch correction would remove all variance, setting each observation to the same constant. An important part in scRNA-seq data analysis is to select biologically informative genes (see section 2.6 for details). If the variance of the gene set is dominated by batch effects, we select an uninformative set of genes. On the other hand, if we remove variance induced by the batch effect and, in addition, biologically relevant variance, we impair the downstream analysis as well. Our goal is to remove variance in the data caused by the batch effect and to preserve as much of the biological variance as possible. Hence, we assess biologically relevant heterogeneity by computing highly variable genes (HVG) [Brennecke et al., 2013]) before and after correction.

Before correction, we compute HVG in each batch separately. Then, we only consider HVG present in all replicates. We find considerably more HVG in the whole dataset than replicate-wise, due to the batch effect. For example, let A, B and C represent three replicates, and a, b and c the corresponding sets of HVG. Then, the batch-free, conservative set of HVG is the intersection of a, b and c:

$$\mathrm{HVG}_{\text{batch-free}} = a \cap b \cap c.$$

When we check the set of HVG after batch correction ($\mathrm{HVG}_{\text{corr}}$), this set would ideally contain the complete set of $\mathrm{HVG}_{\text{batch-free}}$. In total, we evaluated the fraction of retained HVG after correction. To complement the concept of retained HVG, batch correction is not supposed to introduce additional variability to the data. Thus, we consider the difference set of HVG before and after correction, as false discoveries that we use to compute a false positive rate (FPR). In this example, the different set $\mathrm{HVG}_{\text{diff}}$ reads

$$\mathrm{HVG}_{\text{diff}} = \mathrm{HVG}_{\text{corr}} \setminus (a \cup b \cup c)\,.$$

To define a batch-free gene set before batch correction, we fit the Anders model to each batch separately and intersect the corresponding sets of HVG. Let l be the number of batches and a_i the set of HVG for batch i, then we denote

$$\mathrm{HVG}_{\text{batch-free}} = \bigcup_{i=1}^{l} a_i$$

as the set of HVG present in each of the batches in a dataset.
More specifically, we considered the fact that highly variable genes depend on the type of normalization [Vallejos et al., 2017]. Then, the reference set of highly variable genes consists of all genes that are highly variable in all batches with $\log(\text{counts}+1)$ normalization. After batch correction, we compute HVG for the whole corrected dataset (HVG_{corr}). Ideally, we would retain all $\text{HVG}_{\text{batch-free}}$ after batch correction. We define the fraction of retained batch-free HVG,

$$p_{\text{retained}} = \frac{|\text{HVG}_{\text{batch-free}} \cap \text{HVG}_{\text{corr}}|}{|\text{HVG}_{\text{batch-free}}|},$$

to determine if the biological signal in the data is preserved upon batch correction.
We quantify the number of HVG caused by the batch effect as a false positive rate (FPR). In contrast to the fraction of retained HVG, we define the FPR by the fraction of HVG that are found in the whole dataset but not in any of the batches. Let a be the set of HVG in the complete dataset and a_i the set of HVG in batch i, where $i \in \{1, \ldots, l\}$. Then, the false positive rate reads

$$FPR = 1 - \frac{\bigcup_{i=1}^{l}(a \cap a_i)}{|a|}.$$

3.4 kBET outperforms other methods for batch effect detection in simulation studies

We evaluated the performance of kBET on several simulated datasets where the degree of the batch effect is fully controlled. We employed two different models for simulating single-cell RNA-seq data and batch effects. The first model is based on a zero-inflated negative binomial distribution for count data similar to Lun et al. [2016a]. The second approach utilizes the *Splat* model of the R framework *Splatter* [Zappia et al., 2017]. We use *Splat* to explore the influence of differences in dropout and in the noise model.

3.4.1 A zero-inflated negative binomial model

We model the number of transcripts per gene and per cell as count data that follow the negative binomial distribution with zero-inflation (ZINB) to account for dispersion and sparsity caused by dropouts. Mean expression levels for each gene are sampled from the beta-distribution (with appropriate scaling):

$$\mu \sim \text{Beta}(a, b) \cdot c,$$

where we choose the parameters $a = 2$, $b = 5$ and $c = 100$. Then, the dropout probability for each simulated gene $j \in \{1, \ldots, G\}$ in batch $i \in \{1, 2\}$ is modeled by the logistic (sigmoid) function

$$p_{ij} = \text{sigm}\left(-\left(\beta_0 + \beta_{1,i} \cdot \mu_{ij}\right)\right),$$

where we chose $\beta_0 = -1.5$ and $\beta_{1,i} = 1/\text{median}(\mu_i)$. Hence, the higher the expression level, the lower the dropout probability. We designed the dropout model with the sigmoid function similar to the dropout model in Andrews and Hemberg [2018a].
In total, every sample is drawn from $s_{ij} \sim \text{NB}\left(\mu_{ij}, \theta | \text{Ber}(p_{ij})\right)$, where $\theta = 1$ and *Ber* is the Bernoulli distribution. However, we do not consider a correlation structure between genes, which would mimic a gene regulatory network. Instead, we use the same gene means μ_j for both batch simulations.

3.4.2 Batch effect strength is modeled as an increasing fraction affected genes

Having set up the parameters for the first batch, the mean expression levels of the second batch μ_2 are subject to different degrees of variation. We multiply a fraction of 1 %, 10 % and 20 % of the mean expression levels μ with a Gamma distributed random variable $\gamma \sim \Gamma(\alpha, \beta))$ and $\alpha = \beta = 1$[1]:

$$\mu_{2,j} = \begin{cases} \mu_{1,j} \cdot \gamma & : j \in \{1, \ldots, h \cdot G\} \\ \mu_{1,j} & : \text{else}, \end{cases}$$

where $h \in \{1\,\%, 10\,\%, 20\,\%\}$ and G is the number of genes in the dataset. The Gamma distribution with the chosen parameters has mean and variance equal to 1 such that the expected value of the sampled mean expression levels stays unchanged.

However, the rejection rate of kBET depends on the neighborhood size. We illustrate the phenomenon on two simulated datasets (Fig. 3.3). In a simulation with hardly any distinguishable feature between the two batches (Fig. 3.3a, c), the results for the expected and observed kBET tests are almost identical, and decrease with increasing neighborhood size. Both visual inspection of the PCA plot and kBET computation lead to the conclusion that no batch effect is detectable. In a simulation where 20 % of the mean expression levels are shifted, we find a separation of two batches in the PCA plot and the maximal kBET rejection rate (Fig. 3.3b, d).
In addition, we varied the sample size of the two batches to evaluate the effect of sample

[1]Let $X \sim \Gamma(\alpha, \beta)$ be a Gamma distributed random variable with shape α and rate β. The probability density function is $p(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\beta x)$ with $x \in (0, \infty)$. Then, $E[X] = \frac{\alpha}{\beta}$ and $\text{Var}[X] = \frac{\alpha}{\beta^2}$.

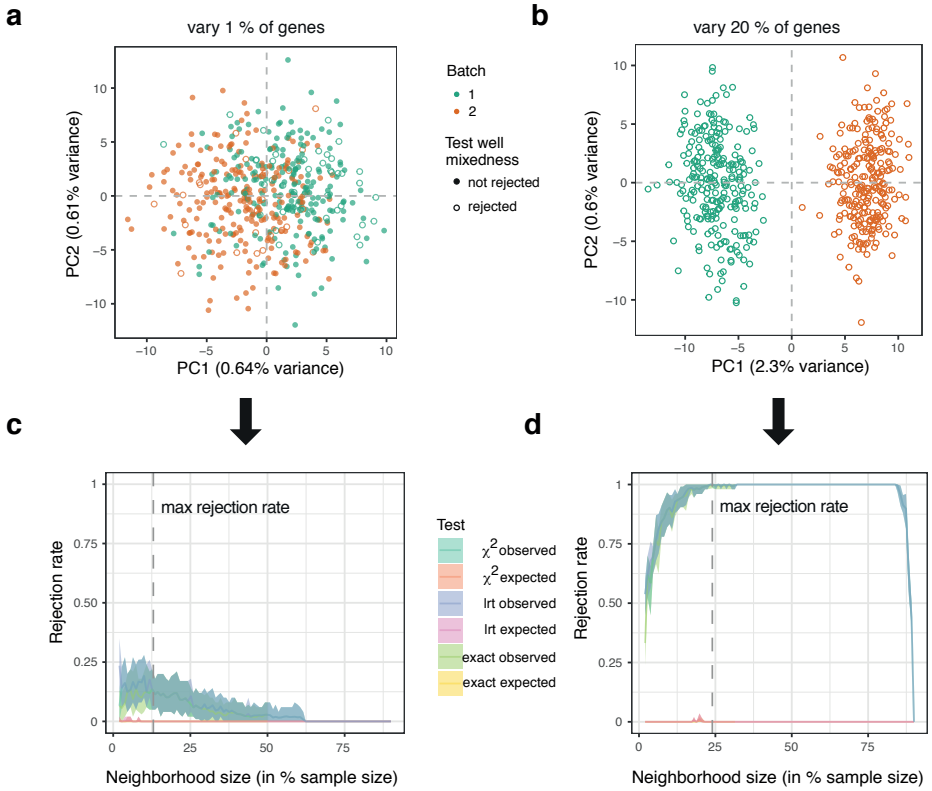

Figure 3.3: kBET result depends on neighborhood size. Simulation results for 1000 genes and 500 cells. Batch effect is modeled as a fraction of shifted gene means in the second batch. Two batches in a) and b) are equally sized with $1\,\%$ **(a)** and $20\,\%$ **(b)** varied gene mean expression levels. **a** and **b** show PCA plots of log-scaled data. **c** and **d** illustrate how kBET values depend on neighborhood size for the simulations shown in **a** and **b**. Dashed vertical line shows the optimal neighborhood size for batch effect detection, i.e. where the rejection rate is maximal. Shaded areas represent the 95-percentile of repeated kBET tests. In each run, the number of tested neighborhoods is $10\,\%$ of the sample size.

size imbalance. In each simulation, we sample 500 instances with 1000 genes each, with the size ratio of the batches being $r \in \left\{1, \frac{1}{3}, \frac{1}{4}, \frac{1}{9}, \frac{1}{19}\right\}$. This means equally sized batches contain 250 samples each, and batches with $r = \frac{1}{9}$ have 450 and 50 samples, respectively. In all simulations with the ZINB model, we compared kBET's ability to detect batch effects with prior quantification approaches (Fig. 3.4).

3.4.3 Simulating single-cell RNA-seq count data using Splat

We complemented our ZINB-based simulation of batch effects using the more complex model Splat as implemented in the R framework *Splatter* (version 1.3.2) [Zappia et al.,

2017]. We investigate the following two scenarios:

(i) the batch effect is promoted by differences in dropout rates

(ii) the batch effect is promoted by additional noise

Before we elaborate both scenarios, let us briefly describe the Splat model. Gene mean expression μ_i is modeled by a Gamma distribution with shape α and rate β: $\mu_i \sim \Gamma(\alpha, \beta)$. The gene mean is modified by the occurrence of outliers, scaled by observed and expected library size per cell and modified by simulated biological coefficients of variation (BCV) (see Fig. 1 in Zappia et al. [2017]). Then, counts are drawn from the Poisson distribution using the final means μ'_{ij}. Dropout can be employed in this final step as option: Dropout is modeled the non-success probability π_{ij}^D in a Bernoulli distribution: Ber $\left(1 - \pi_{ij}^D\right)$. The probability of observing non-dropout values rises as the final gene mean rises in a logistic (sigmoid) function:

$$\pi_{ij}^D = \frac{1}{1 + \exp\left(-k \cdot \left(\ln\left(\mu'_{ij}\right) - x_0\right)\right)},$$

where k denotes the dropout shape and x_0 the dropout midpoint. By default, $x_0 = 0$ and $k = -1$. For the batch simulations, we computed 1000 gene means in two equally sized batches, where we also varied the sample size. For the first batch, we always employed default parameters. The dropout rate simulations are carried out by shifting shape parameter k at several sample sizes, while we ensured the gene means being the same in both batches (Fig. 3.5). In order to simulate additional noise, Splatter provides a batch effect simulation of its own. In particular, batch effects are implemented as additional log-normal factor on gene means. The batch effect for a gene, b_i, is drawn from $LN(\mu, \sigma)$,where μ is the batch factor and σ is the batch scale, and finally multiplied with the final gene mean $\mu'_{i,j}$.
We have varied both batch factor and scale in order to determine the impact of the batch effect to simulated data (Fig.s 3.6 and 3.7). The major difference to the ZINB-based simulations is that any batch effect will now be applied to all genes instead of shifting a fraction of genes as described in the section above.

3.4.4 Simulation results and comparison to alternative measures

We compared the ability of kBET to detect batch effects with alternative measures: the absolute average silhouette width ('silhouette') and the scaled variance of top 50 principal components (PCs) that correlate significantly with the batch effect ($FDR < 0.05$)('scaled PC regression'). In addition, we compared the performance of different PCA based mea-

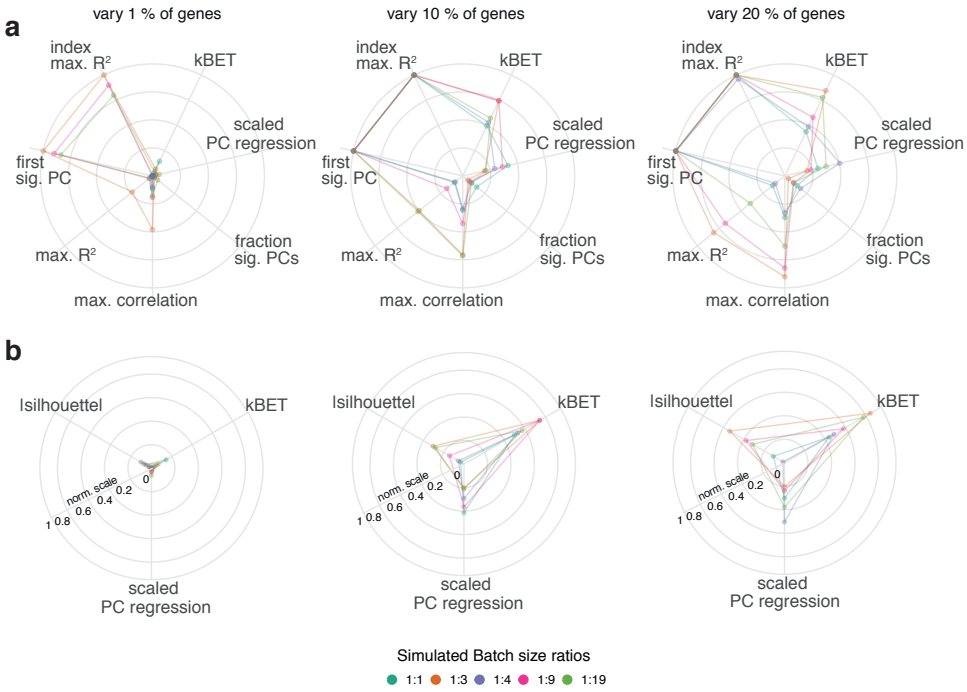

Figure 3.4: kBET is more sensitive than other batch tests on simulated data. Comparison of kBET to other batch effect tests on a normalized scale. **a** PCA based measures are compared to kBET. See text for details **b** Comparison of kBET to silhouette and scaled PC regression in different simulation study cases. kBET (rejection rate), scaled PC regression (*i.e.* variance explained by batch effect in the top *N* PCs) and absolute silhouette coefficient. Several batch sizes were simulated to assess the impact of unequal batch sizes. kBET detects most strongly (*i.e.* most sensitively) the degree of bias.

sures to kBET. All measures operate on a normalized scale (*i.e.* 0 - 'no effect', 1 - 'strong effect'). As initial test cases, we varied an increasing fraction of gene means (Fig. 3.4), simulated different dropouts (Fig. 3.5) or added noise to gene means (Fig.s 3.6 and 3.7). In addition to the fraction of varied gene means, we simulated different batch sizes ranging from equal size (1:1) to strong size imbalance (1:19).

The different PCA based approaches vary strongly in their performance on simulated data. Interestingly, we always find a significantly batch-correlated PC, even if the batch effect is very small (Fig. 3.4a, left image). The index-based scoring approaches turn out to be insensitive to the degree of batch effect, whereas scaled PC regression, maximal correlation and coefficient of determination, respectively, increase with the degree of batch effect. However, we observe a correlation of 0.5 in a simulated dataset with almost non-existent batch effect, which we consider a good example for false positive detection of a batch effect.

In contrast, kBET's rejection rate increases most strongly by the degree of bias compared with silhouette: silhouette shows little difference between 10 % and 20 % varied genes, where kBET clearly discriminates the effect (Fig. 3.4b). Scaled PC regression increases with the degree of batch effect, but also returns a significant result where only 1 % of genes is varied. kBET performs well when only a few data points are biased by batch, as it still reveals a substantial bias in size-imbalanced batches. Overall, kBET is clearly the most sensitive and robust measure of batch effects in this comparison.

When we complemented the initial batch effect simulation with the *Splat* model, we focussed at first on dropout rates. In particular, we restricted our simulations to shifting shape parameter k at several sample sizes, while we ensured the gene means being the same in both batches (Fig. 3.5). We observed that by varying the shape parameter k, the kBET starts to detect changes as little as $\Delta k = 0.05$, whereas both PC regression and silhouette coefficient start to rise for at least $\Delta k = 0.5$ (Fig. 3.5a). In contrast, if no batch effect is present, kBET reliably detects none. This indicates again that by investigating neighborhoods of cells, as kBET does, we are able to detect subtle changes in the dataset, that are invisible in a low-dimensional PCA plot (Fig. 3.5b).
However, the nature of dropout effect still allows for an almost perfect correlation of simulated means and dropout rates in both batches even for high deviation in the shape parameter k (Fig. 3.5c, d). Furthermore, when dropout rates are higher, we observe lower cellular detection rates (CDR), which still correlates strongly with the simulated library size. When sample sizes are smaller, simulated data are noisier with respect to dropout (Fig. 3.5d), which tends to be picked up by kBET more strongly than by PC regression or silhouette. In total, kBET is detecting already little deviations in dropout rates that are hard to detect via visual inspection. Thus, kBET is more sensitive to subtle changes in the data, which PC regression or silhouette do not detect.

The third group of simulations uses the batch effect simulation option that is provided by *Splatter*. For each scenario, we drew 1000 samples and created random subsets of the same simulated dataset. For both variation of batch factor (ranging from 0.01 to 0.1) and batch scale (ranging from 0.1 to 1), kBET decreases with the sample size, though we observe both high silhouette coefficients and high amount of variance explained by batch effect (Fig. 3.6a). Furthermore, we illustrated the dependence of kBET on the distance metric - computing kBET on the original data space yields maximal rejection rates while using a lower dimensional space removes noise from the data (Fig.s 3.6a and 3.7a, top panels). The dependence of kBET from the sample size, however, is expected: By reducing the number of samples, we also increase uncertainty in any evaluation of a neighborhood, leading to a decrease in rejecting the null hypothesis.

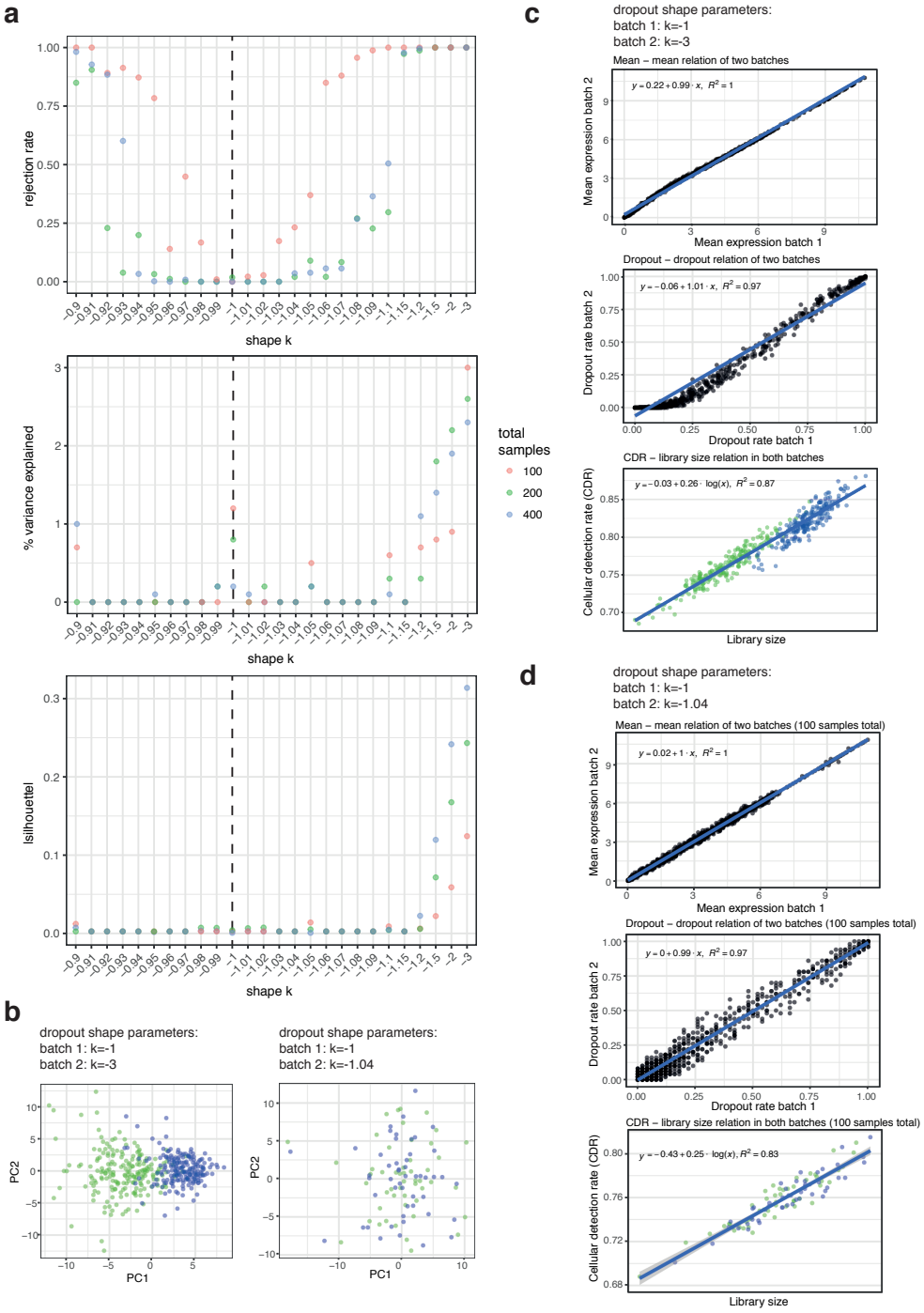

Figure 3.5: Simulation of different dropout effects in single-cell RNA-seq data Dropout is controlled by the shape parameter k, where batch 1 has $k_1 = -1$and k_2 in batch 2 ranges from -0.9 to -3. Batches are always equally sized, sample sizes refer to the total number of samples in the dataset. **a** kBET rejection rates, variance explained and silhouette coefficient for each simulation (from top panel to bottom panel). **b** PCA plots for a large difference (left, $\Delta k = 2$) and a small difference (right, $\Delta k = 0.04$). **c-d** Mean relation (top panel), dropout relation (center panel), CDR effect (bottom panel) for a large difference **(c)** and a small difference **(d)**, respectively. Blue lines indicate the linear fit with parameters depicted in the top-left corner. R^2 values denote the variance explained by the fit.

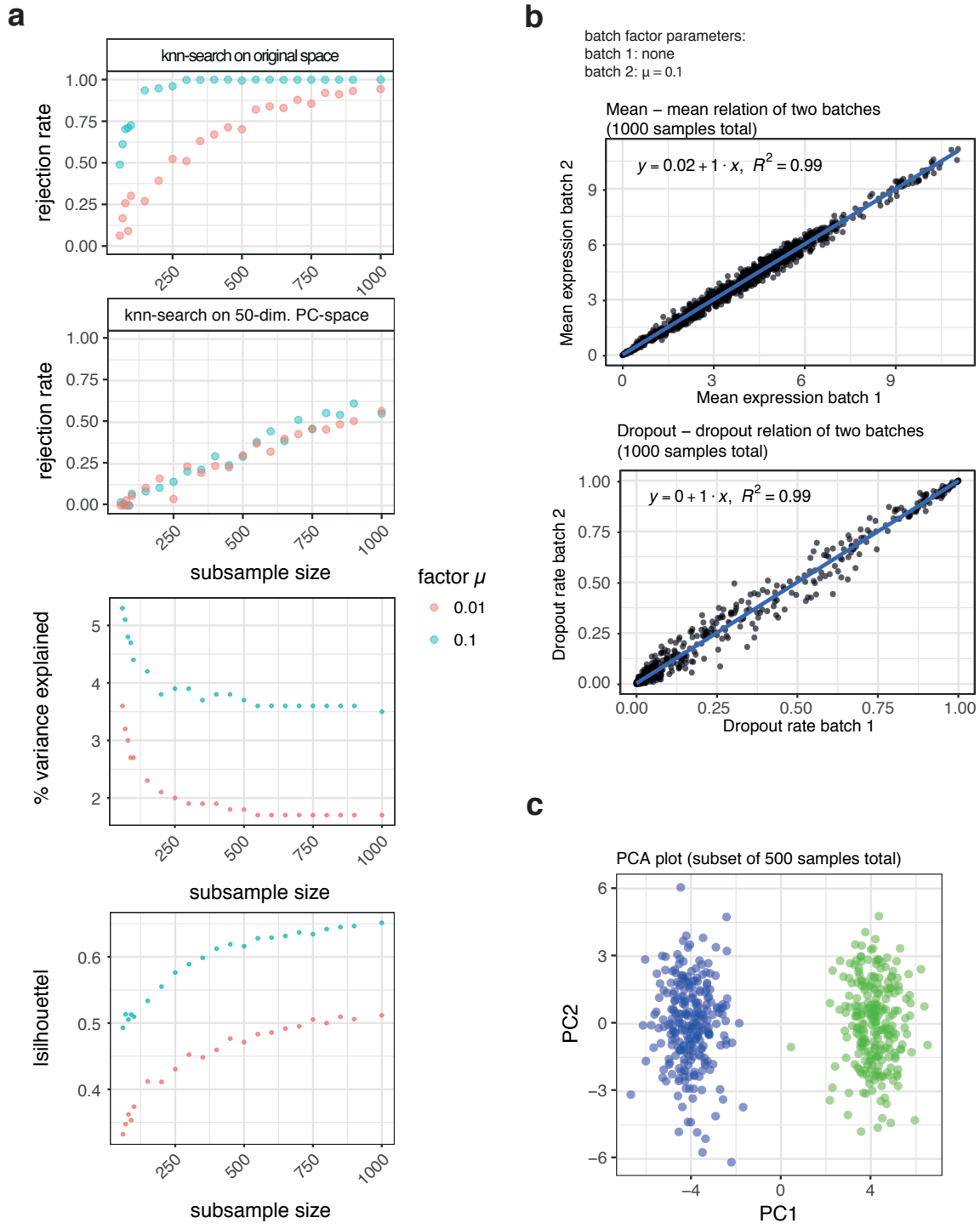

Figure 3.6: Simulation of batch effects as additional noise factor in one batch a Batch effect analysis for several batch factors using kBET, PC regression (variance explained by batch effect) and silhouette coefficients (from top to bottom panels). kBET rejection rates where computed for both the original data space and the default 50-dimensional PC-space (top two panels). **b** Mean relation (top panel) and dropout relation (center panel) for batch factor $\mu = 0.1$. Blue lines indicate the linear fit with parameters depicted in the top-left corner. R^2 values denote the variance explained by the fit and corresponds to high correlation of mean and dropout of both batches. **c** PCA plot for the simulation shown in **b**.

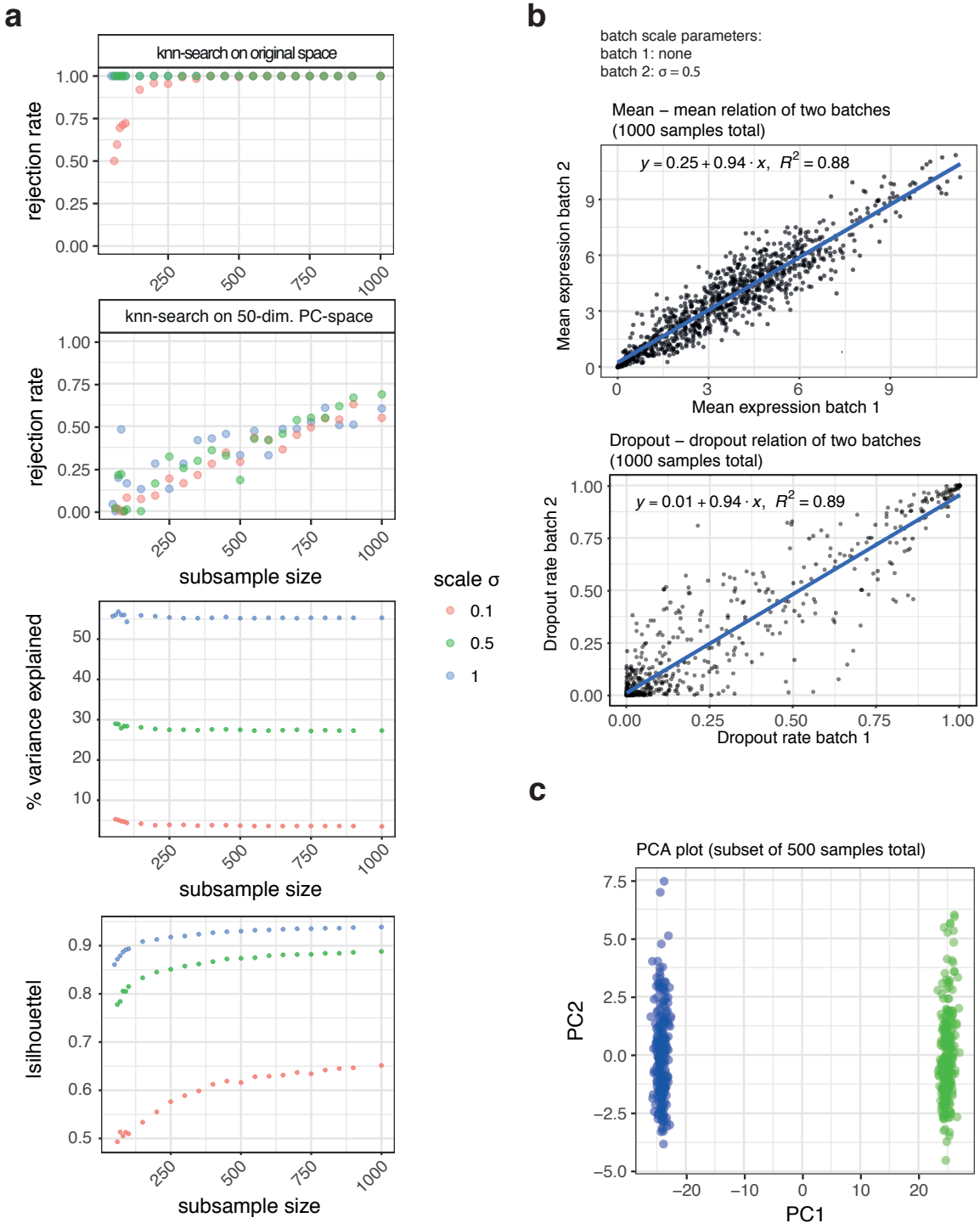

Figure 3.7: Simulation of batch effects as additional noise scale in one batch a Batch effect analysis for several batch factors using kBET, PC regression (variance explained by batch effect) and silhouette coefficients (from top to bottom panels). kBET rejection rates where computed for both the original data space and the default 50-dimensional PC-space (top two panels). **b** Mean relation (top panel) and dropout relation (center panel) for batch scale $\sigma = 0.5$. Blue lines indicate the linear fit with parameters depicted in the top-left corner. R^2 values denote the variance explained by the fit and corresponds to high correlation of mean and dropout of both batches. **c** PCA plot for the simulation shown in **b**.

Variation of the batch scale and batch factor leads to considerably more noise in both mean expression and dropout rates between batches (Fig.s 3.6b and 3.7b). Curiously, the batch effect simulated here is completely described by the first principal component, which allows to discriminate batches in a visual fashion, but the data is lacking considerable structure otherwise (Fig.s 3.6c and 3.7c). That is a more pathological situation as batch effects tend to distribute more over several principal components. By choosing a 50-dimensional subspace for the k-nearest neighbor search, we may obscure the obvious distance that is shown by the first principal component and kBET rejection rates may not reflect the clear distinct pattern that is in turn picked up by both silhouette and PC regression.

In summary, kBET is sensitive to the distance measure used for k-nearest neighbor search (see Fig.s3.6a and 3.7a). As a remark, we advice caution with the underlying distance metric. Furthermore, we found decreasing rejection rates if the corresponding dataset was too small. We observed a similar effect in the silhouette coefficient and the variance explained. Thus, batch effect quantification relies on sufficient sample numbers. Overall, kBET was the most sensitive tool the simulation study and outperformed alternative measures based on PCA or the silhouette coefficient.

3.5 Using kBET to benchmark batch effect correction in scRNA-seq data

In this section, we aim to determine best practices for scRNA-seq data normalization and batch effect correction. A best practice for scRNA-seq is the first step towards standardization of the scRNA-seq analysis pipeline and thus enhances comparability of different datasets and reproducibility across studies, which is of particular importance with regard to the Human Cell Atlas [Regev et al., 2017].

For this purpose, we analyzed six single-cell datasets derived from mice that cover both microwell plate-based and droplet-based methods with sample sizes ranging from 100 to 3000 cells per batch. We demonstrate the performance and accuracy of 11 normalization and 7 batch effect regression approaches (Fig. 3.8 and sections 2.3 and 2.4 for details). Furthermore, we illustrate that visual inspection of low-dimensional embeddings in complex cases may be ineffective to study the performance of batch effect removal. Hence, we recommend an additional quantitative step after exploratory data analysis. When performing tasks like clustering into cell types, which precedes differential testing, we have to ensure that the low-dimensional embedding reflects neither incomplete nor over-corrected data.

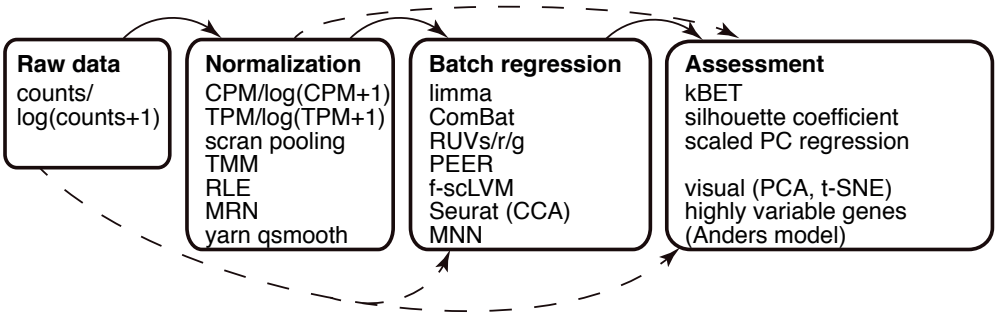

Figure 3.8: Batch correction assessment scheme. Overview of normalization and batch regression methods as well as assessment approaches. Raw data as counts are either directly batch corrected or normalized before correction.

dataset		normalization	batch correction
Klein et al		$\log(counts+1)$/scran pooling	ComBat
	2i	$\log(CPM+1)$	ComBat
Kolodziejczyk et al	a2i	scran pooling	ComBat
	LIF	TMM/$\log(CPM+1)$	limma/ComBat
mouse embryo		$\log(counts+1)$	ComBat

Table 3.1: Best overall normalization and batch correction methods The ranking of batch correction strategies is based on kBET, retained HVG and false positive rates for Klein et al and Kolodziejczyk et al data. For mouse early embryonic development data integration, the ranking is based on both kBET and silhouette coefficient.

Moving beyond, we used *kBET* to detect batch effects in complex datasets and how to address variability across individuals. In particular, we address the question of whether it is possible to integrate separate studies, and show with mouse development datasets that it is possible to correct for study-to-study effects. As a result, batch correction based on log(counts+1), log(CPM+1) or scran pooling, together with ComBat or limma regression, reduced the batch effect while preserving biological structure across all datasets (Table 3.1). Further, *kBET* allows studying inter-individual variability based on a droplet-based study with peripheral blood mononuclear cells (PBMCs).

3.5.1 kBET accurately captures batch effects in single-cell RNA-seq datasets and outperforms visual inspection

Batch effects originate from different sources, as is evident when comparing technical replicates. We investigated the mouse embryonic stem cell (mESC) LIF cultures of Klein

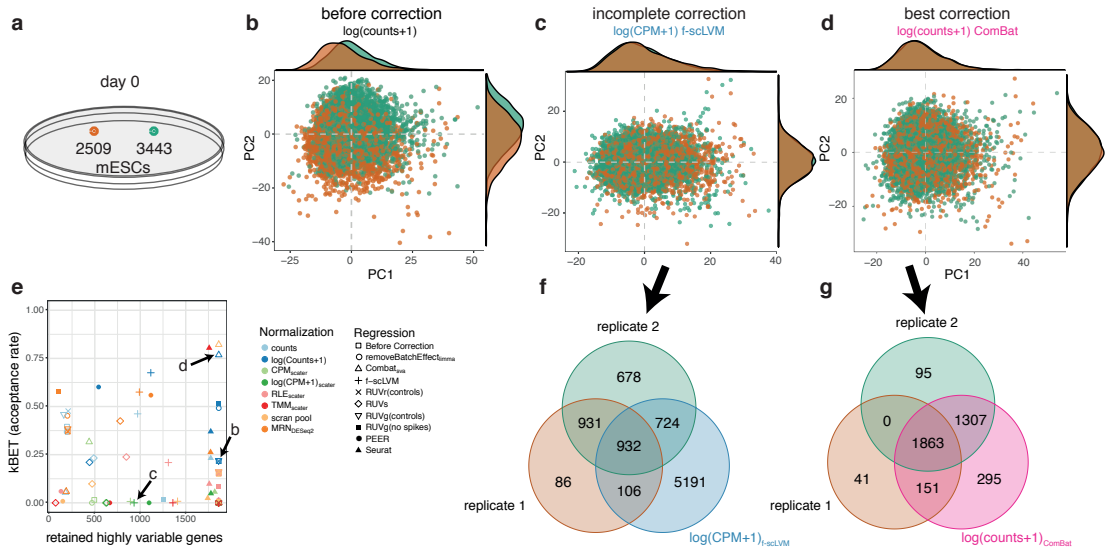

Figure 3.9: ComBat corrects best on mESC inDrop technical replicates. a The inDrop protocol provides a large UMI-count dataset with two technical replicates. PCA plots **(b-d)** display log-normalized counts, a biology-removing batch removal (f-scLVM on log-transformed CPM) and a biology-preserving batch removal (ComBat on log-transformed counts), respectively. Density plots depicted on the axes show the frequency of the replicates along the PCs. By visual inspection, both approaches shown in **c** and **d** appear equally well, but have very different kBET results **(e)**. **e** Percentage of retained highly variable genes vs. acceptance rate (equals 1 - rejection rate) for all combinations of normalizations and batch regression approaches. Arrows point towards the in **b-d** illustrated approaches and indicate strong performance differences between f-scLVM and ComBat. **f,g** Venn diagrams of highly variable genes per replicate before correction and for the whole dataset after batch correction. Highly variable genes in each replicate are computed on log(counts+1) values. The f-scLVM method retains 932 highly variable genes but has a high false positive rate, while ComBat captures all the highly variable genes with a low false positive rate.

et al [Klein et al., 2015], which were generated with the inDrop protocol. The authors provided two technical replicates in the samples of day 0 culture (Fig. 3.9a), which offers an ideal case for batch correction assessment.

Prior to batch correction, we visualized the data and observed a clear inter-batch difference as the shift of the technical replicates in the top 2 PCs (Fig. 3.9b). While we compared all combinations of normalization and batch correction strategies, we illustrate major performance differences on f-scLVM corrected log(CPM+1) values (Fig. 3.9c) and ComBat corrected log(counts+1) (Fig. 3.9d) in terms of the top 2 PCs. Both appear successful in removing batch effects according to visual inspection of the PCA. However, besides the top 2 PCs, we find 13 PCs have a significant correlation with the batch covariate in the f-scLVM

case ($FDR < 0.05$), which explain 1.2 % variance in the data. Thus, the batch effect has not been corrected and has become untraceable using data visualization. For ComBat corrected log(counts+1), none of the PCs correlates significantly with the batch covariate. The more sensitive kBET reveals that ComBat applied to log(counts+1) or scran pooling removes batch effects best (y-axis inFig. 3.9e, 'acceptance rate' is the reverted 'rejection rate'), in contrast to incomplete batch correction performance of f-scLVM. The PCA plot only shows the batch effect of the first two PCs, while kBET effectively quantifies more subtle batch effects.

The second challenge in batch correction is to preserve biological heterogeneity in the data, as described in section 3.3. Briefly, we determine the fraction of retained HVG after correction (Fig. 3.9e-g), which indicates the preservation of the true biological signal (see section 3.3 for details). The two technical replicates of mESCs share 1863 $\text{HVG}_{\text{batch-free}}$ and over 700 HVG reside in either of the replicates (Fig. 3.9f-g). After correction by f-scLVM, we retained half of $\text{HVG}_{\text{batch-free}}$ while we discovered over 5000 HVG in the whole dataset (Fig. 3.9f and Fig. 3.10a-b), which explains f-scLVM's minimal kBET acceptance rate (Fig. 3.9e). When we compute the FPR on the basis of log(CPM+1) normalized data, we find a FPR of 27 % (Fig. 3.10c). We obtained the best result for the combination of log-transformed Counts and ComBat (Fig. 3.9d) - all $\text{HVG}_{\text{batch-free}}$ were kept after batch correction and only 295 HVG were caused by batch correction (8 % FPR, see Fig. 3.9g).

In conclusion, batch correction may confound observations massively, masking the biological signal completely. In addition, even the best batch correction strategy leaves part of the batch effect in the data (Fig. 3.9e, g). This explains the increase of the total amount of HVG after correction (Fig. 3.10b) and FPR (Fig. 3.10c). Both the silhouette coefficient and PC regression show little discrimination between most of the correction strategies (Fig. 3.10d-e), whereas kBET resolves them in detail (Fig. 3.9e and Fig. 3.10d-e).

3.5.2 Deriving a best practice in batch correction

Next, we examined mouse embryonic stem cells cultured in three different conditions (2i, a2i and LIF) [Kolodziejczyk et al., 2015a] and sequenced with the SMARTseq2/C1 protocol (Fig. 3.11a). These datasets are rather similar in terms of heterogeneity, but the biological origin of the heterogeneity in each culture condition is different (compare Kolodziejczyk et al. [2015a] for details). Thus, applying the same batch correction and normalization strategies led to similar results across all culture conditions. We obtained well-mixed data for all datasets with log(CPM+1) normalization and batch correction with ComBat (Fig. 3.11b and black arrows in Fig. 3.11c).

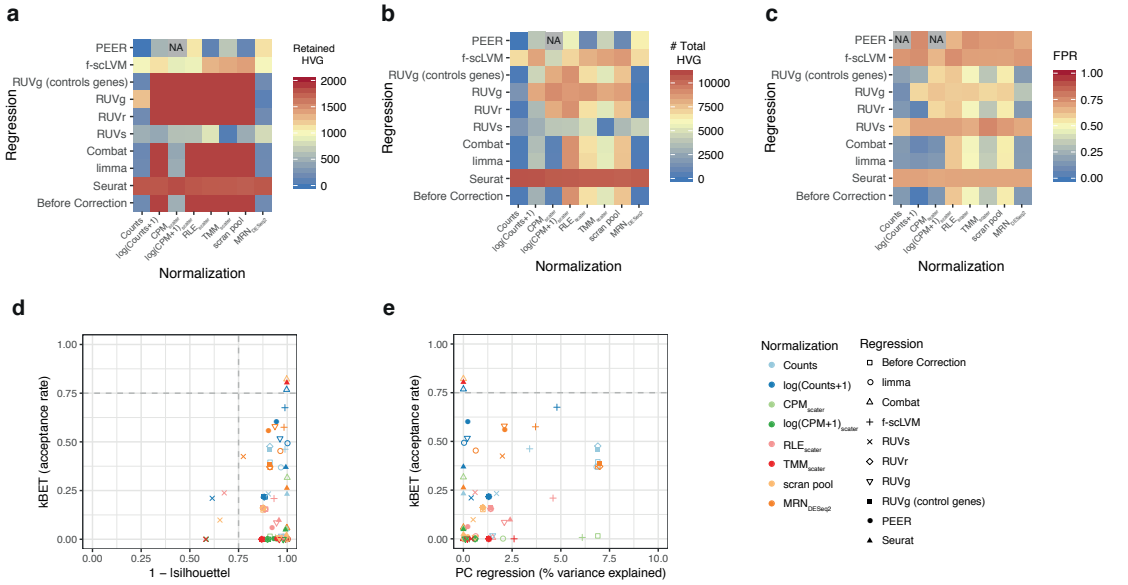

Figure 3.10: Highly variable genes and kBET results upon batch regression. **a** Number of retained highly variable genes before and after batch correction. Reference: Intersect of highly variable genes per batch with log(counts+1) normalization. **b** Total number of highly variable genes after batch correction. **c** False positive rates on highly variable genes for all combinations of normalization and batch correction methods. **d** Comparison of silhouette coefficient and kBET 'acceptance rate' (1-rejection rate). **e** Comparison of PC regression and kBET 'acceptance rate' (1-rejection rate).

In all datasets, log-transformed count data are locally well-mixed in over 50 % of cases (dark blue full squares in Fig. 3.11c and Fig. 3.12a). We notice a decrease in acceptance rate in 2i and LIF culture when we compared count data to other normalization methods. For batch effect correction, RUV controlled both acceptance rate and overlap in highly variable genes, but we also noticed an increase in the total number of HVG (Fig. 3.12b) as well as an increased false positive rate (Fig. 3.12c). Furthermore, increasing the number of factors used in RUV increases batch effect removal only slightly (Fig. 3.13). In contrast, ComBat increased the acceptance rate, preserved almost all $\text{HVG}_{\text{batch-free}}$ in all datasets and had a low false positive rates at the same time.

In a2i, we observed almost perfect mixing in log-transformed data, but we also found that a considerable amount of cells are isolated, *i.e.* they do not have a mutual nearest neighbor (Fig. 3.14a-c). These cells are implicitly removed from the local structure evaluation. If the isolated cells have a different label composition than the global dataset, we find strong differences in kBET's rejection rate if isolated cells remain unconsidered (Fig. 3.14d,e).

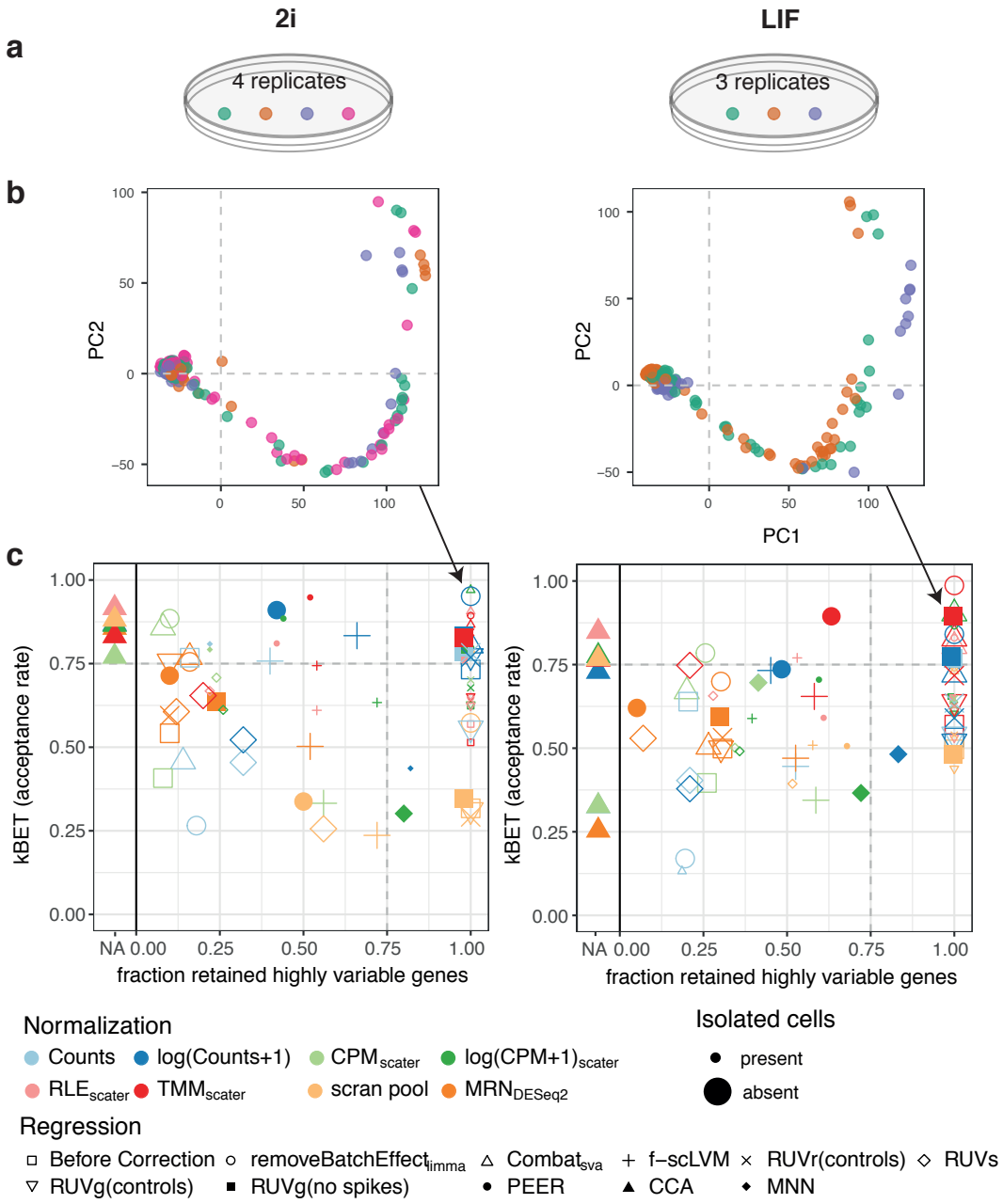

Figure 3.11: Deeply sequenced SMARTseq2/C1 mESC data have similar characteristics for batch correction. **a** Illustration of two full-length read datasets with replicates in 2i and LIF culture (219 and 207 cells, respectively). **b** PCA plots for log(CPM+1) ComBat corrected data. **c** Percentage of retained highly variable genes vs. kBET acceptance rate (equals 1 - rejection rate) for all combinations of normalization and batch correction approaches. Best performing normalization-regression strategies cluster in the top right corner, such as ComBat on log(CPM+1) data. Isolated cells do not have mutual nearest neighbors and appear in some correction models.

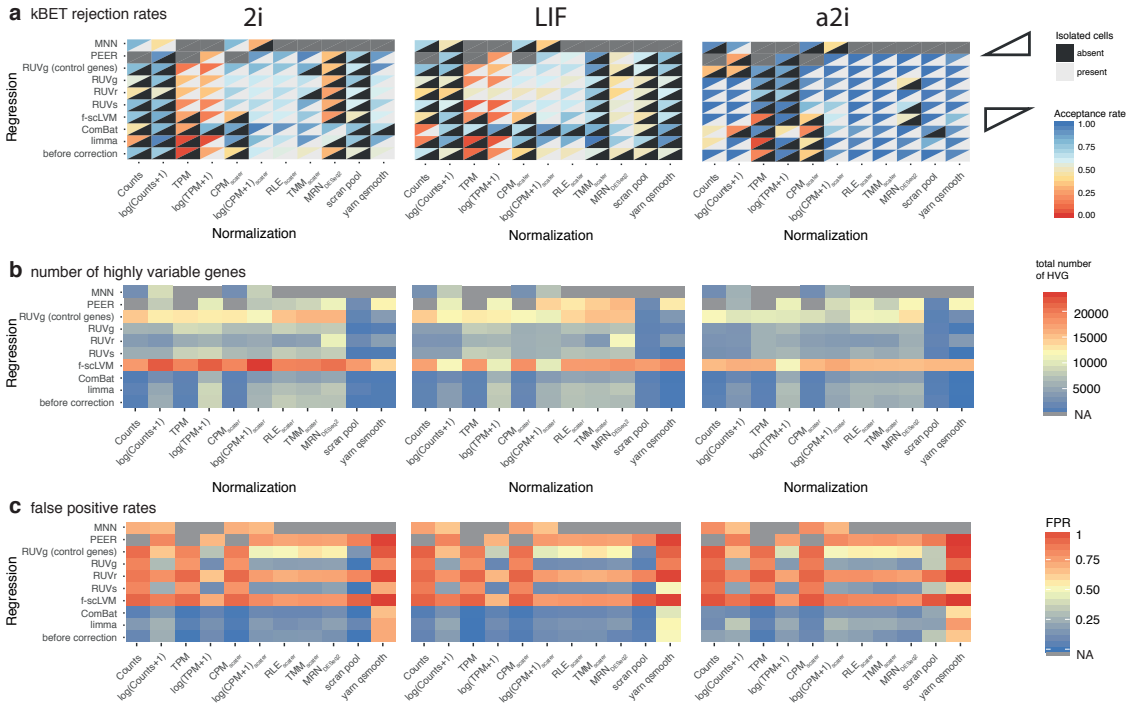

Figure 3.12: Batch regression assessment for mESC data in several culture media. **a** kBET results in 2i (left), LIF (center) and a2i (right) culture, respectively. **b** Total number of highly variable genes after batch correction in 2i, LIF and a2i culture. **c** False positive rate after batch correction in 2i, LIF and a2i culture.

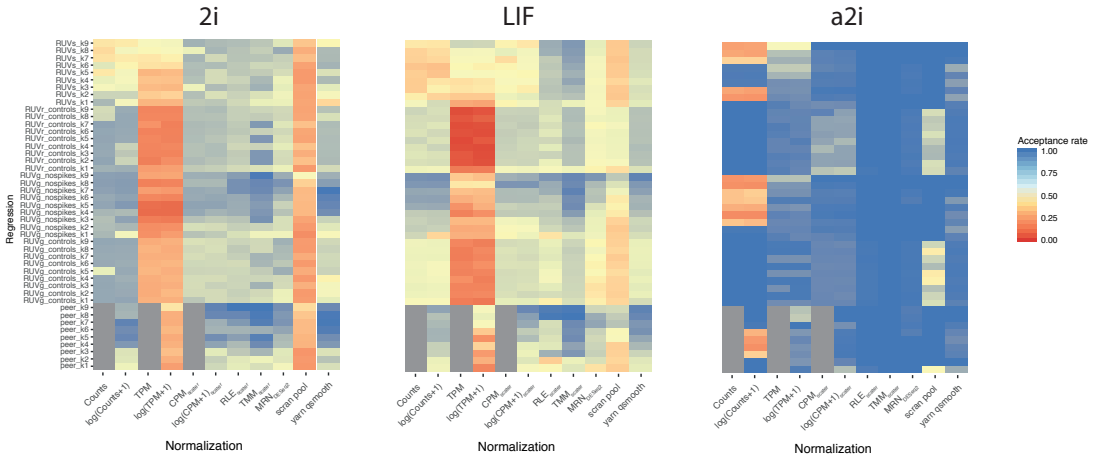

Figure 3.13: kBET results for several factors of RUV and PEER in 2i (left), LIF (center) and a2i (right) culture, respectively.

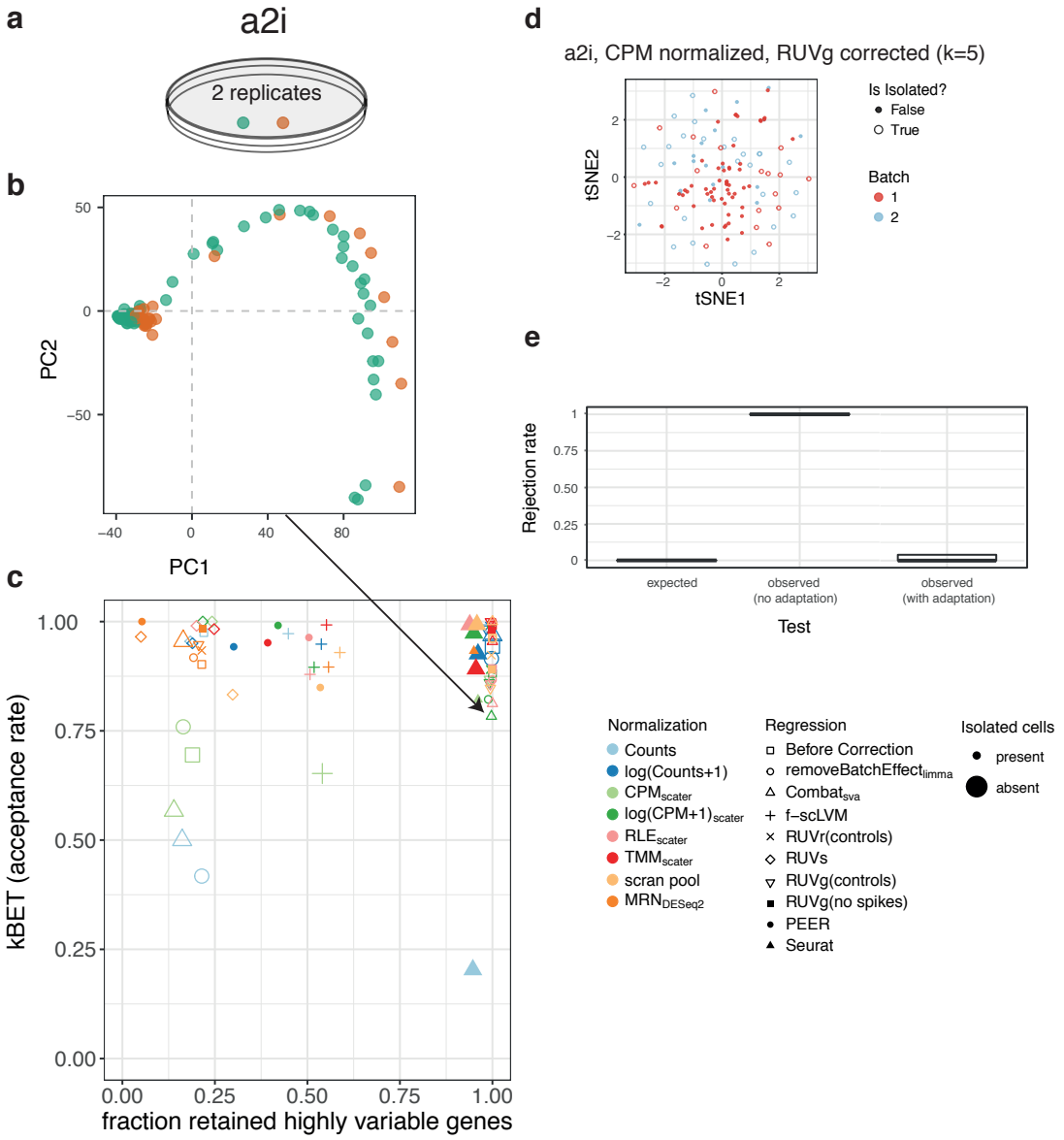

Figure 3.14: Batch regression assessment for mESC data in several culture media. a Illustration of a2i dataset (123 cells total), **b** PCA plot for log(CPM+1) ComBat corrected data and **c** scatter plot of highly variable genes vs. kBET acceptance rate. Symbol sizes denote the presence or absence of isolated cells. **d** tSNE plot of a2i data (CPM normalized and RUVg ($k = 5$) corrected. We observe $\sim 30\,\%$ of isolated cells (*i.e.* without mutual nearest neighbor). **e** kBET rejection rate plot with adapted frequencies for dataset shown in **d**. Without adapting for isolated cells, we observe maximal rejection rate, and with adapted expected frequencies the rejection rate is insignificant.

Therefore, we adapted the expected label composition such that we can neglect the isolated cells (see section 3.1.5). In general, we considered a correction strategy is ill-advised if it produces considerable amounts of isolated cells when removing batch effects (Fig. 3.12a).

The performance of batch correction methods varied slightly from dataset to dataset (Fig. 3.11c). In order to test the dependence of batch correction performance and the number of batches, we subset the 2i data to combinations of three and two batches (Fig. 3.15). The 2i dataset consists of four batches, all biological replicates. The comparison of t-SNE and PCA representation of the data (Fig. 3.15a) shows four clusters in the t-SNE plot that are indistinguishable in the PCA plot. Moreover, two of these clusters belong to separate batches, while the other two batches are in close proximity, which is in accordance with the low kBET acceptance rate (Fig. 3.12a). This example also illustrates how a PCA plot can be misleading, as the batches overlap to an extent in the PCA.
As the four batches appear qualitatively different, we created four subsets of the 2i dataset by leaving out each of the four batches one by one. Then, we performed batch correction with ComBat, limma and RUVg ($k = 3$) and compared the results with kBET before and after regression (Fig. 3.15b). Notably, we find fairly different acceptance rates for each of the methods depending on the test set composition. ComBat and limma performed best on a test set with batch 2 removed where we also find the highest acceptance rate before correction. In contrast, RUVg has the highest acceptance rate for a test set with batch 4 removed. Pairwise comparisons of the batches show that batch 2 separates from each of the other batches (Fig. 3.15c), with acceptance rates of 0 (upper right-hand corner in each plot of Fig. 3.15c). In contrast, cells in the other batches mingle to different degrees. A possible reason is that, in these batches, library size and number of detected genes do not correlate well (Fig. 3.15d). On one hand, this shows that kBET results are not dependent on the number of batches but on the degree of batch effects. This example shows that if the degree of batch effect is different across samples, the ability to reduce batch effects is lower in limma and ComBat. On the other hand, we saw that depending on the subset chosen, we would either chose ComBat, limma or RUVg as the best performing method (Fig. 3.15b).

In summary, depending on the combination of batches, we observed performance differences in batch effect correction that was independent on the number of batches and rather explicable by batch-specific dissimilarity. We demonstrated that classical batch correction tools, in particular ComBat, successfully remove batch effects and preserve the biological signal (Table 3.1).

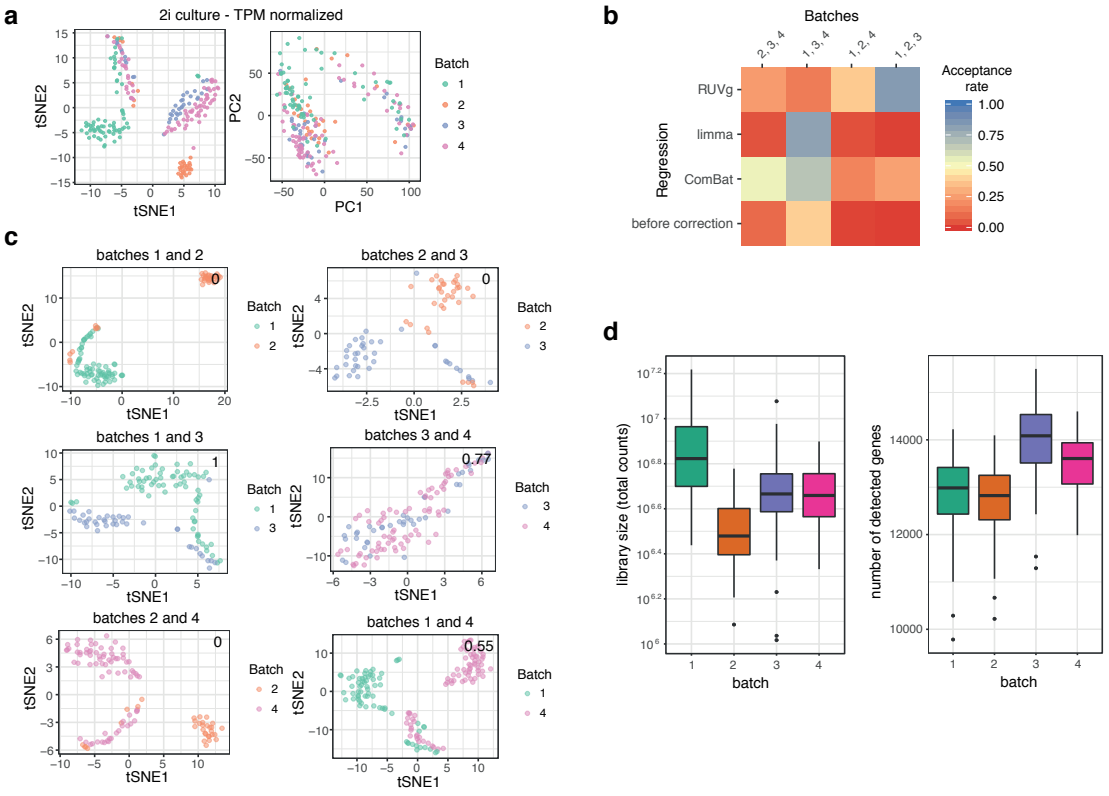

Figure 3.15: Batch regression assessment for subsets of mESCs data reveal performance differences. **a** Visualization of TPM normalized 2i culture. Left: tSNE, Right: PCA. **b** kBET acceptance rates for all subsets of 3 batches with $\log(counts+1)$ normalization. Batch correction success varied among test sets. **c** t-SNE plots for all combinations of 2 batches. kBET acceptance rate is depicted in upper right-hand corner in each plot. Batch 2 does not mix with either of the other batches. **d** Boxplot for library size (left) and number of detected genes (right) of the dataset.

3.5.3 Time and memory test of the batch regression methods

Apart from accuracy of a model, computational efficiency plays a key role in the feasibility of large-scale single-cell transcriptomic data analysis. Hence, a batch correction model needs to provide both accurate results and efficient implementation. In this context, we recorded the CPU time and maximal memory required to run each of the batch regression methods on the data from Kolodziejczyk et al. [2015a] (Fig. 3.16). The method with least CPU time and memory requirements is limma's *removeBatchEffect*, which is based on simple least-square-fit. ComBat is the second fastest program and requires within 2G memory. Next are the different flavours of RUV. The factor analysis based models f-scLVM and PEER are slowest, but PEER is quite memory efficient.

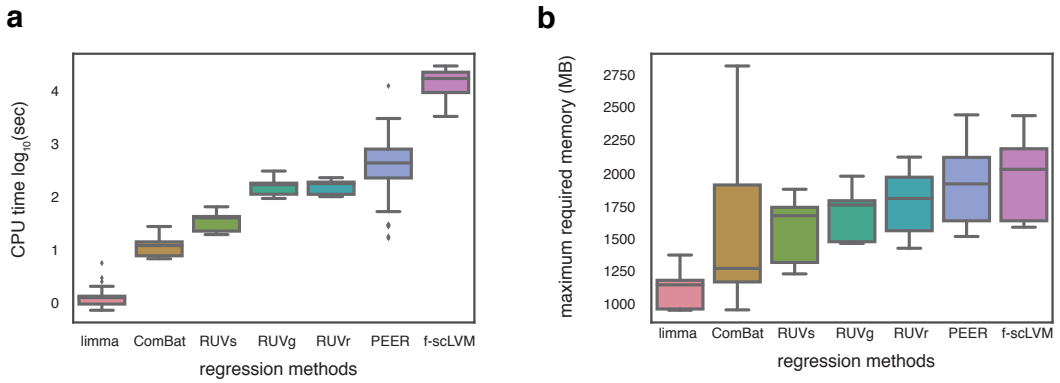

Figure 3.16: Time and memory tests on SMARTseq2 mESC data from Kolodziejczyk et al. [2015a]. **a** Time cost of the batch regression methods. **b** Maximal required memory of the batch regression methods. Factor analysis models as PEER and f-scLVM both consume the most CPU time, limma is the most efficient in CPU time and memory usage.

In general, regression based approaches outperform factor analysis based approaches in terms of both time and memory.

3.5.4 Dropout handling in batch effect correction of scRNA-seq data

The term *dropout* describes the observation of sparsity in scRNA-seq data. On the one hand, dropout is a false negative observation - an mRNA transcript from a certain gene was not captured in the experiment. On the other hand, dropout is a true negative observation - the gene was not expressed due to the stochastic nature of gene expression [Elowitz et al., 2002]. Eraslan et al. [2019] have shown that increasing dropout rates reduces the signal-to-noise ratio in a way that reliable clustering of the data into cell types becomes infeasible.

In the example of inDrop count data [Klein et al., 2015], we found $\sim$ 55 % dropout, where the first batch has a slightly larger dropout rate than the second batch (see Fig. 3.17a). Thus, batch effects can also manifest in different dropout rates. The scatter plots displays only those methods, where we can define dropout at all (*i.e.* TMM or MRN normalized data consist as well of negative values making it is difficult to track dropout in a general fashion).

Let us describe the dropout handling at the example of ComBat: First, it converts data internally into Z-score (*i.e.* set gene mean to zero and variance to one), then shifts gene expression for each according to its batch ID and converts back into the original space. Gene means are shifted regardless to the actual expression level, which means for count or

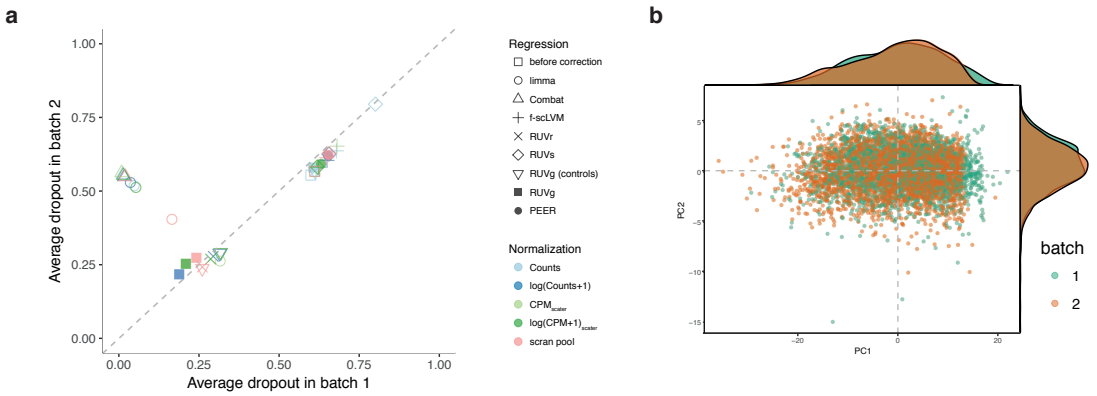

Figure 3.17: Dropout handling in batch correction is crucial, yet neglected. (Klein et al) **a** Scatter plot of average dropout rate per batch in UMI-count data. Dashed line depicts equal dropout rates. Points below the dashed line have a higher dropout rate in batch 1. Most methods reduce of increase dropout in both batches similarly. ComBat reduces the dropout rate in batch 1 dramatically. **b** PCA plot of $\log(counts+1)$ corrected with a modified ComBat version: ComBat was applied as usual but zero-values were again set to zero after the correction. Density plots on the sides of the plot indicate a shift between the batches, and kBET returns a rejection rate of 1, that is mostly driven by the dropout.

CPM data that the lower limit of gene expression is shifted as well. Hence, an originally zero valued entry might be shifted in one batch to positive values and in another batch to negative values. That leads to the observation, that the dropout rate sinks dramatically in one batch compared to the other (see Fig. 3.17a). However, if one wishes to keep the original lower limit (*i.e.* preserves all zero entries), the batch correction with ComBat (plus zero-value preservation) has a kBET rejection rate of 1, due to differences in the dropout rate (see Fig. 3.17b).
For latent variable methods as RUV or f-scLVM we found that dropout decreased in remarkably in both batches, but dropout rate differences where not alleviated. All in all, we think that dropout in scRNA-seq has to be addressed more thoroughly, for example, by imputation [Eraslan et al., 2019, Li and Li, 2018].

3.6 Going beyond batch effect correction

The sensitive and unbiased k-nearest neighbor batch effect test allows studying sophisticated systematic bias in scRNA-seq data. In particular, batch effects are prominent factors to address in data integration, where the cell type composition may vary across datasets. Here, we apply *kBET* cluster-wise to account for differences in cell type composition across

studies. Furthermore, *kBET* allows quantifying variability across individuals, if we can exclude batch effects experimentally by pooling all cells and reconstruct a cell's donor via phenotyping with tools such as *demuxlet* [Kang et al., 2018]. The following sections demonstrate the versatile applications of *kBET* for scRNA-seq data.

3.6.1 Going beyond replicates - dataset integration across multiple studies

With the explosion of scRNA-seq data in recent years [Angerer et al., 2017], we begin to realise the need for a comprehensive strategy of data integration. Of course, correcting batch effects between different studies is more challenging than within the same study, especially, if cell types vary between studies. In this work, we benchmark the batch correction performance on 8 different SMARTseq based datasets [Biase et al., 2014, Boroviak et al., 2015, Deng et al., 2014, Fan et al., 2015, Goolam et al., 2016, Liu et al., 2016, Wu et al., 2016, Xue et al., 2013], all profiling mouse early embryonic development from oocyte to blastocyst (Fig. 3.18a).
We observe a range of library sizes from 10^6 to 10^8 reads per cell, depending on the study (Fig. 3.18b). Batch effects before correction are quite obvious even in PCA (Fig. 3.19a-b): Biase et al data and Deng et al data deviate significantly from the others (average acceptance rate of 16 %). Consequently, the cells are more likely to cluster by study rather than embryonic developmental stage. However, we observe that it is possible to correct the batch effect computationally: the best results are with ComBat on log(counts+1) (Fig. 3.19c-d) with an average acceptance rate of 62 %.

When integrating developmental data, we expect that the same cell types from different studies mix, maintaining the correct trajectory of successive developmental stages. Hence, we assess the batch effect of each developmental stage based on averaged kBET results. In parallel, we monitor the developmental progression by silhouette(Fig. 3.19e). A high acceptance rate implies good mixing within each developmental stage, and a good separation of developmental stages translates into higher silhouette coefficients. Before correction the developmental stages separate weakly (silhouette 0.08 for log(counts+1)). After correction with linear methods as limma and ComBat, we observe distinct clustering according to developmental stages, but only ComBat achieves a good mixing of study batches. Notably, the PC1 corresponds to the real developmental time of the cells.
As desired, one of the normalization and regression methods captures both the developmental stages and mixes the data from different studies. ComBat with log(counts+1) yields high acceptance rates and clear clustering by developmental time. Surprisingly, Seurat's CCA alignment is among the top performing methods and yields the second best

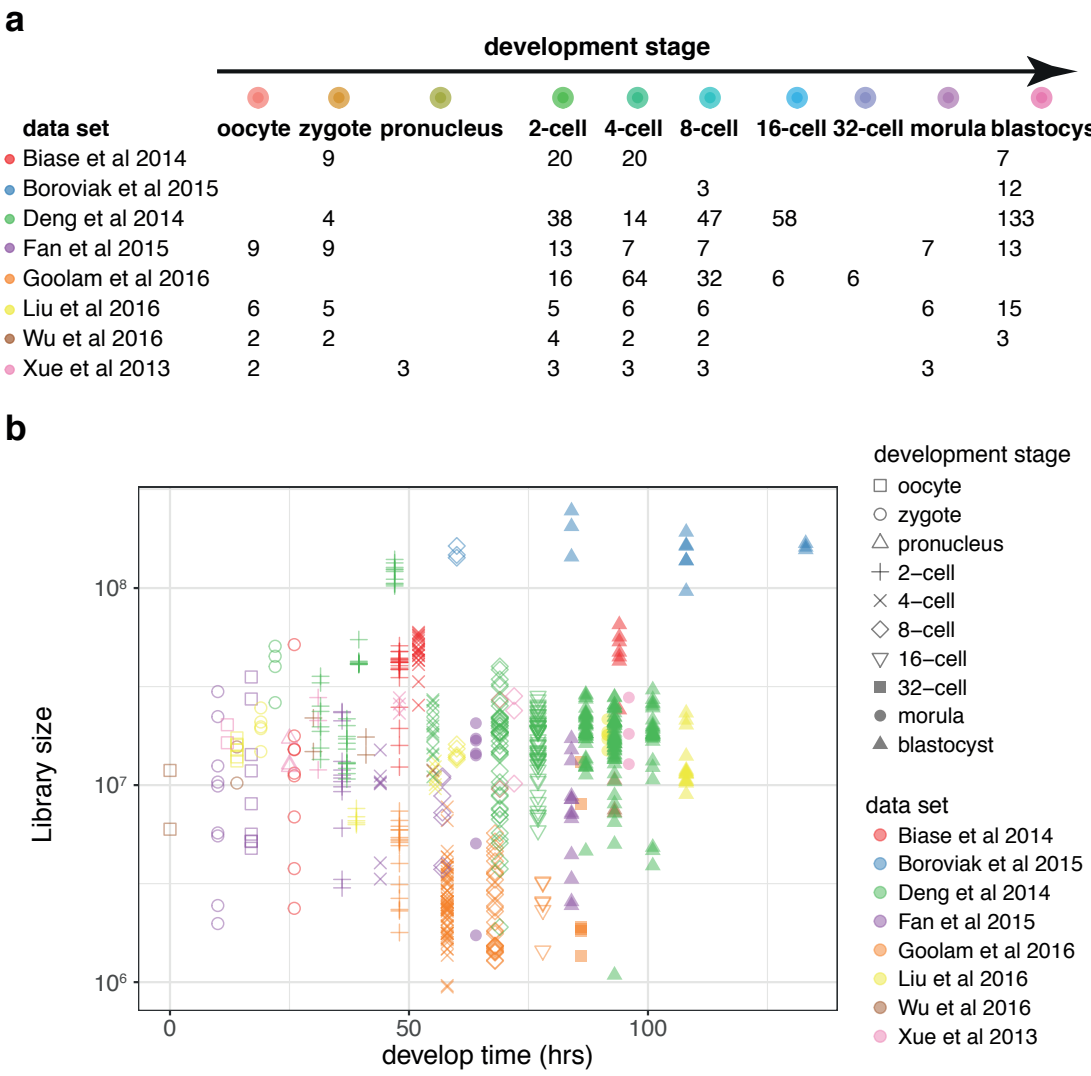

Figure 3.18: Overview of data from early mouse development from 8 different datasets.a Number of cells per development stage and study. **b** Sequencing depth for every cell distinguished by development stage (shapes) and study (colors). Library size varies by study rather than cell type.

kBET result for log(CPM+1), but a silhouette coefficient of approx. 0 indicates overcorrection. MNN also yielded a low acceptance rate, and improved clustering by cell types only for counts, possibly due to low sample numbers.

This example illustrates how batch effect correction tools can play a key role in data integration and provide an effective separation of the biological signal from complex technical variations. In this work, we considered each study as one batch and ignored any tech-

nical substructure within the studies. Furthermore, cell type distribution as well as cell collection time points differed slightly across studies, which made the task more difficult. Data quality plays an essential role for successful data integration: for instance, batch correction quality is very limited if one aims to integrate poor quality data. For future data integration efforts with more complex data structure and less prior knowledge on cell types, the community needs more sophisticated batch correction methods that model nested batch structures and several batch variables.

We remapped the reads to the same reference transcriptome with Salmon [Patro et al., 2017] to reduce quantification biases [Teng et al., 2016]. Interestingly, even different versions of Salmon resulted in different degrees of batch effect. We have used both Salmon v0.7.2 and v0.8.2 (with the same reference) to map the reads and quantify expression levels with TPM. Surprisingly, the two versions of Salmon produce different results in terms of batch effects. When we inspected PCA plots, Salmon v0.7.2 appears to produce much less of a batch effect, while v0.8.2 produces significant batch effects in the studies of Baise et al and Deng et al (Fig. 3.20). Hence, bioinformatic pre-processing has to be considered as source of technical bias.
However, we find almost no difference in the datasets after batch correction by limma (Fig. 3.20c-d and g-h). This result indicates another important aspect of batch correction: each study uses a specific set of bioinformatic tools for pre-processing and genomic references, and as these tools differ from study to study, we deal with the batch effect known as mapping bias. To reduce this source of variability in the data, it is advisable to map the reads of each study to the same reference genome using the same tool to control the mapping bias.

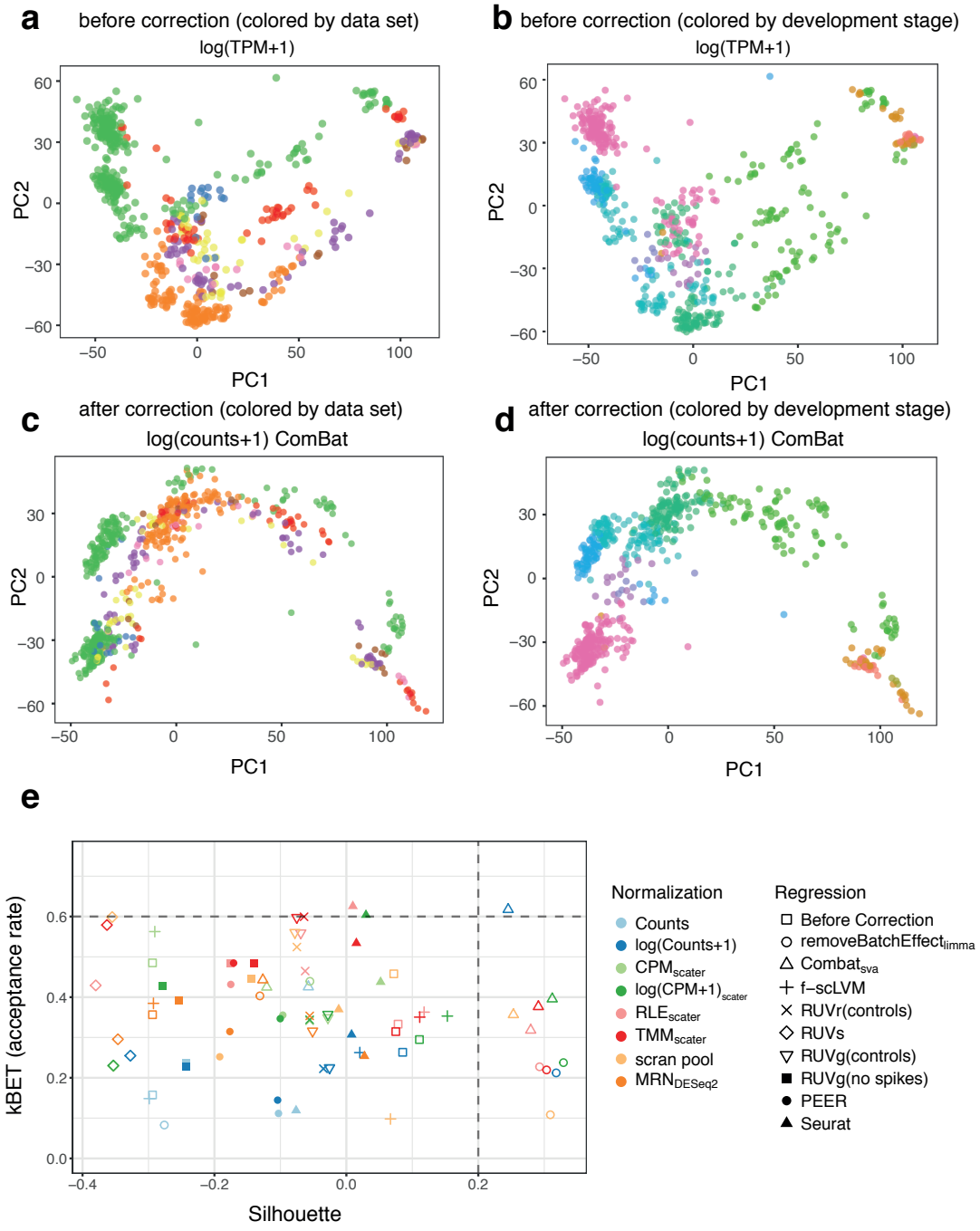

Figure 3.19: kBET assesses in agreement of cell stages in early mouse embryo data integration.a-b PCA plots of log(counts+1) normalized expression data colored by dataset **a** and by developmental stage **b**. Cells of the same developmental stage largely do not cluster together across different studies. **c-d** PCA plots of of log(counts+1) normalized expression data after batch correction by limma colored by dataset **c** and by developmental stage **d**. Cells of the same developmental stage are more aligned after correction. **e** Silhouette coefficient of embryonic development vs. average kBET acceptance rate (weighted per developmental stage) reveals that ComBat applied to log(counts+1) provides good mixing of cells from different studies in the same developmental stages. This is indicated by the high kBET acceptance rate and by high silhouette coefficient, indicating the best separation of developmental stages.

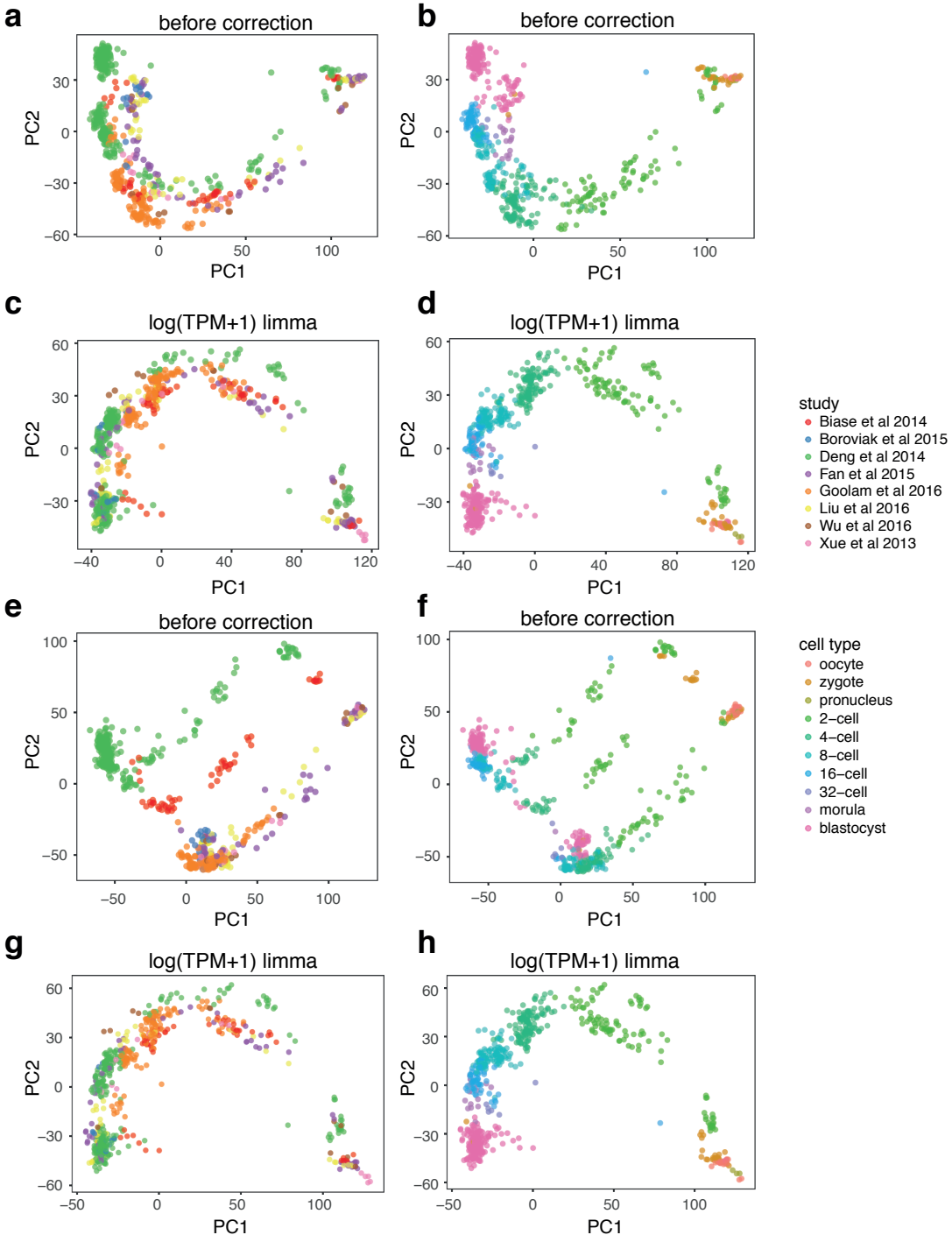

Figure 3.20: PCA plots of mouse early embryo development data based on mapping with Salmon v0.7.2 and v0.8.2. **a-b** PCA plots of log(TPM+1) normalized expression data mapped by Salmon v0.7.2, colored by study **a** and by developmental stage **b**. **c-d** PCA plots of of log(TPM+1) normalized expression data mapped by Salmon v0.7.2, after batch correction by limma, colored by dataset **c** and by developmental stage **d**. **e-f** PCA plots of log(TPM+1) normalized expression data mapped by Salmon v0.8.2, colored by dataset **e** and by developmental stage **f**. **g-h** PCA plots of of log(TPM+1) normalized expression data mapped by Salmon v0.8.2, after batch correction by limma, colored by dataset **g** and by developmental stage **h**.

3.6.2 Going beyond batch effects - kBET detects variability between individuals in PBMC data

As pointed out in the introduction, the distinction between biological and technical variability can sometimes be challenging. In order to estimate pure biological variability with kBET, we studied a human PBMC dataset from eight unrelated individuals where samples were pooled during scRNA-seq and the corresponding individual donor identity for each cell was then reconstructed using *demuxlet* [Kang et al., 2018]. Hence, technical variation is entirely absent when we compare cells from different individuals. Clustering and t-SNE visualization reveals several cell types (Fig. 3.21a). Furthermore, cell type frequencies vary significantly between individuals (Fig. 3.21b). Note that all samples were distributed across three independent experiments (batch A: individuals 1 - 4. batch B: individuals 5 - 8 and batch C: all individuals). Cell type frequencies are very similar between batches, thus excluding sampling bias.

We applied kBET to estimate inter-individual-variability in all these experimental batches. Surprisingly, we found much higher acceptance rates in some cell types than for the complete dataset, even though we downsampled all tested cell types to the same sample size (Fig. 3.21c-d). We interpret this as variation in bulk being driven not only by single cell expression differences but also by variation of cell population sizes. Furthermore, kBET offers a sensitive and unbiased way to estimate inter-individual-variability within the same cell type, which is a novel context to be explored. Furthermore, in the context of the Human Cell Atlas, the community is discussing whether to sequence more cells or more subjects. Based on the kBET result showing overall variation driven by differences in cluster proportions, we would argue for more cells from fewer donors for complex samples such as PBMCs (in contrast to some preprints such as Bhaduri et al. [2017]).

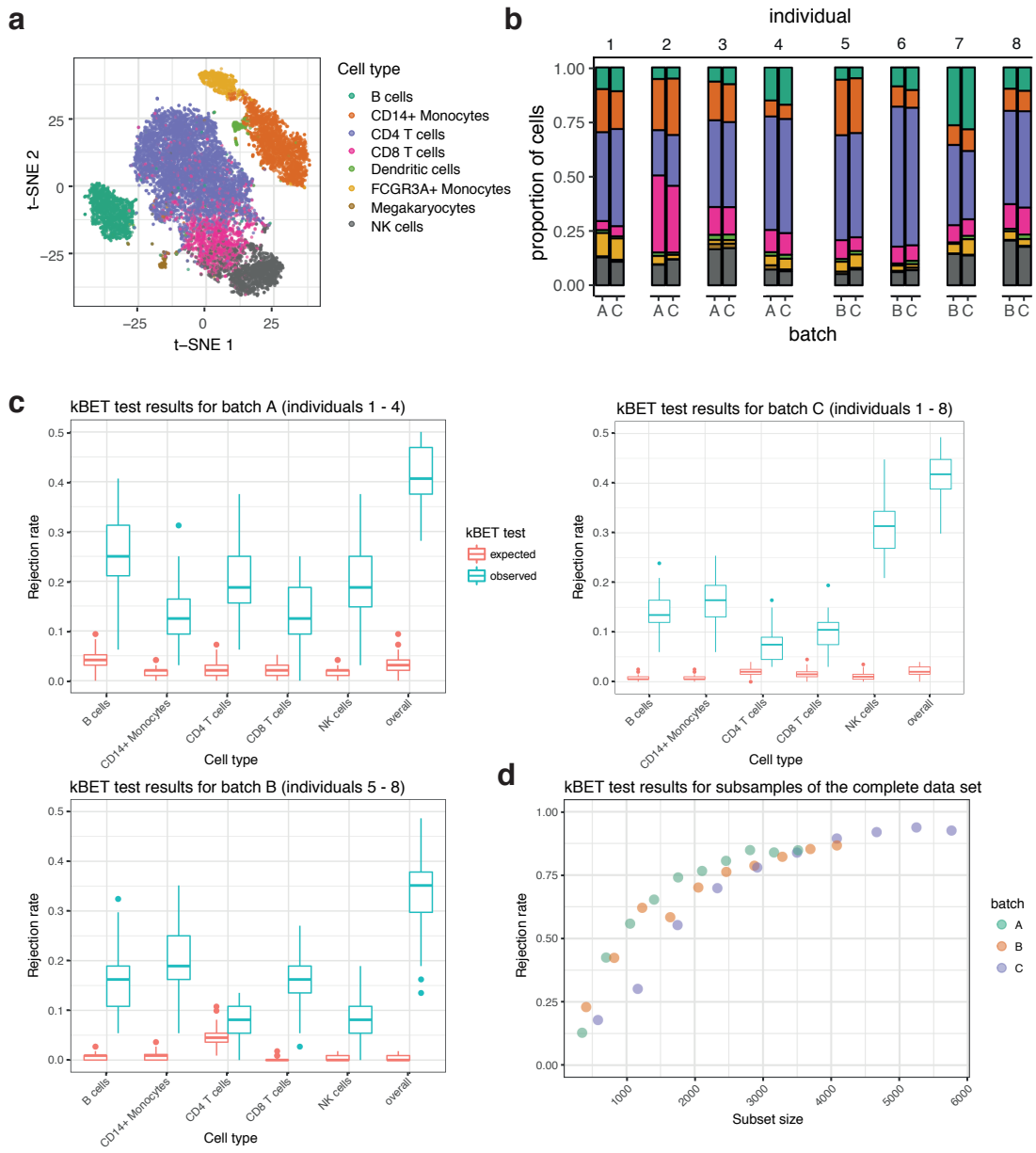

Figure 3.21: kBET detects inter-individual-variability in PBMC data. PBMC data from 8 unrelated individuals processed in three experiments (batches) (Chromium 10X Genomics device) and whose cell identity was reconstructed using *demuxlet* [Kang et al., 2018]. We apply kBET as sensitive indicator for inter-individual-variability. **a** tSNE plot of all data. Cell types are annotated as in original publication. **b** Cell type frequencies per individual and batch. Cells were prepared once and processed in two separate runs. Inter-individual-variability is stronger than preparation bias. **c** kBET rejection rates for each experiment using the individual ID as batch variable. For assessment, we downsampled abundant all cell types and the complete dataset, respectively, to the same number of cells. We observed that kBET yields higher rejection rates if we take the complete dataset (and neglect variation in cell type frequencies) while rejection rates are lower for the respective cell types. **d** kBET rejection rates for several subsets of the complete dataset. Subsample sizes were chosen from 10 to 100 % of the data sample size. Subsampling was repeated three-fold and kBET rejection rates were averaged across these replicates to reduce bias from subsampling. With decreasing sample size, we find decreasing rejection rates. This result is due to decreasing certainty for each tested neighborhood as it leads to enhanced failure to reject the null hypothesis.

3.7 Discussion and outlook

Batch effects in single-cell RNA-seq data can have a severe impact on downstream data analysis if they are not properly accounted for. Moreover, they have a substantial random noise component that stems mostly from technical factors of the experiment. In the simplest possible case, where we have technical replicates that are otherwise homogeneous, *ComBat* corrects the data and preserves the underlying biological properties. On biological replicates with greater batch-to-batch variability, such as two independent cultures of the same cell type, *ComBat* again performs well because of its regularization for low sample numbers. A study using ComBat on complex tissue data finds similar results [Liu et al., 2018].

Many batch correction methods have been designed to correct bulk RNA-seq or microarray data. With little or no adaptation, they can be and are applied to single-cell RNA-seq data. While single-cell RNA-seq data mirror cell-to-cell variability, they are much sparser because of stochastic gene expression and dropouts in the experiment. Yet, batch effect correction approaches from bulk era do not address the expression distribution of single-cell RNA-seq. A mere mean shift and variance stabilization would not take into account a batch-to-batch difference that is solely addressing dropout rates (see section 3.5.4). Furthermore, dropout and cellular detection rates (CDR) [Finak et al., 2015, Hicks et al., 2017] are closely correlated to library size (Fig. 3.22). The single-cell specific approaches model stochastic expression and dropout explicitly [Risso et al., 2018] or implicitly [Buettner et al., 2017]. As zeros in gene expression comprise both biological and technical variation, several approaches aim to impute dropout to rescue biological information [Eraslan et al., 2019, Huang et al., 2018, Li and Li, 2018, van Dijk et al., 2018]. For complex tissue data, matching similar populations with CCA [Butler et al., 2018] or MNN [Haghverdi et al., 2018] provide a generalized, nonlinear modelling approach compared to e.g. ComBat. The two approaches (CCA and MNN) are independent of population density variations in contrast to ComBat [Butler et al., 2018, Haghverdi et al., 2018]. Moreover, with thousands of measured cells per dataset, optimal memory usage and efficient implementation will be as important as accurate correction for confounders (section 3.5.3). Furthermore, batch effect correction is also approached with neural networks such as generative modeling (*scGen* [Lotfollahi et al., 2018] and *scVI* [Lopez et al., 2018]), and general adversarial networks [Ghahramani et al., 2018]. These novel approaches unfold huge potential to batch effect correction and to data integration across various conditions and species in particular.

In contrast to batch correction with regression models, normalization aims to reduce cell-to-cell bias within a batch. Previous studies have discussed appropriate scaling factors [Fi-

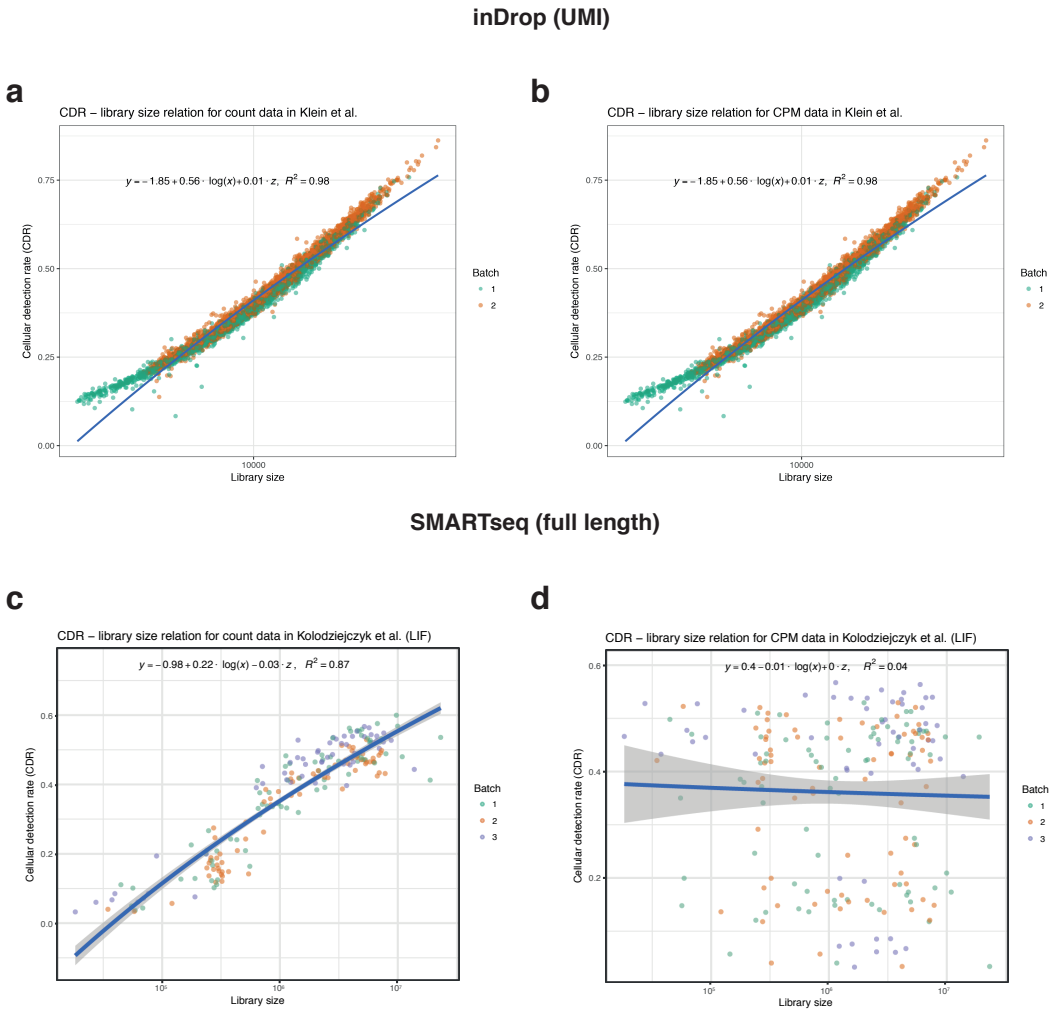

Figure 3.22: Normalization affects CDR - library size relation in SMARTseq data Comparison of cellular detection rate (CDR) effect and library size for inDrop UMI data (Klein et al) **a-b** and C1 SMARTseq data (Kolodziejczyk et al) **c-d**. **a** and **c** show CDR for count data, **b** and **d** show CDR for CPM normalized data where any gene was counted as detected if $CPM > 1$. Blue line shows a linear fit of CDR (y) to library size (x) and batch variable (z) and R^2 denotes the variance explained by the model. Grey shaded areas indicate standard error of the fit. In UMI data, the library size explains largely the CDR effect independent of normalization **a-b**. In SMARTseq data, we observe a significant contribution of library size to CDR effect **c** which is lost in CPM-normalized data **d**.

nak et al., 2015, Vallejos et al., 2017], but we found that normalizing for library size with CPM consistently increased batch effect compared to raw count data. Furthermore, the number of genes per cell and the library size may not correlate well across batches (Fig. 3.15d and Fig. 3.22). Nevertheless, CPM normalization and the more advanced scaling with scran in combination with ComBat worked very well in deeply sequenced SMARTseq2/C1 data (Fig. 3.11 and table 3.1). In addition, all batch effect correction methods require a certain statistical property in their model. For example, RUV requires count data with negative binomial distribution. If the data input in RUV violates the model assumption, RUV introduces additional variability to the data, which we described as isolated cells (section 3.1.5). In total, a simple linear tool such as ComBat showed the best performance in terms of batch correction in our test settings.

The k-nearest neighbor batch effect test (kBET) approach is a powerful tool for the comparison of batch effect correction schemes, as it allows the study of high-dimensional data without prior assumptions regarding statistical properties. Analysis tasks like clustering into cell types or ordering cells by pseudotime [Saelens et al., 2018] rely on batch-effect corrected data. Still, the underlying model assumption requires equivalent, interchangeable batches. While simple, the translation into balanced experimental design is challenging. For complex experimental setups as time series data collection, only collecting and sequencing all cells of all time points together prevents confounding with both technical factors and biological variation between samples. The *demuxlet* [Kang et al., 2018] approach allows quantitative assessment of inter-individual variability without technical confounding and kBET's heterogeneity statistics is a useful measure for biological variability across individuals. Interestingly, kBET detects significant variation within a cell type even after accounting for frequency shifts (acceptance $\approx 0.75 - 0.9$ for each cell type vs. $0.62 - 0.72$ for complete data).
In the worst case, batch effect correction may fail completely if data lack a minimum level of quality. By quantifying batch effects before and after correction with kBET, we are able to detect poor quality correction or poor quality data. We expect this discussion to have serious impact on experimental design decisions in emerging single-cell expression atlases such as the Human Cell Atlas [Regev et al., 2017] and the Mouse Cell Atlas [The Tabula Muris Consortium, 2018].

Chapter 4

Deciphering the lineage hierarchy in the small intestinal epithelium

> Ich möchte die Natur verstehen,
> ich möchte wissen, wie es dazu
> gekommen ist, dass etwas so ist,
> und zwar genau und
> nachprüfbar und nicht geglaubt
> oder kulturell eingeprägt.
>
> Christiane Nüsslein-Volhard

In this chapter, we aim to infer the lineage specification of all major cell populations in the small intestinal epithelium. While most cell types of gut epithelia are known, the lineage hierarchy is still under debate. In particular, little is known about the decision making process in intestinal stem cells and early progenitors. This chapter is based on the following manuscript: A. Böttcher, **M. Büttner**, S. Tritschler, M. Sterr, A. Aliluev, I. Burtscher, S. Sass, M. Irmler, J. Beckers, C. Ziegenhain, W. Enard, A. C. Schamberger, F. M. Verhamme, O. Eickelberg, F. J. Theis, and H. Lickert. *Wnt/planar cell polarity primed intestinal stem cells directly differentiate into enteroendocrine or Paneth cells.* Manuscript in revision

Here, we present the analysis of single-cell transcriptomic data of whole crypt cells as well as enriched samples from transgenic mice. This dataset consists of 14 batches with over $75,000$ single cells. Therefore, this is a prominent use-case to demonstrate the importance of batch effect correction in data integration and the potential of the computational models, which we reviewed in Chapter 2. In particular, computational models for data analysis should be developed in a data driven way to unfold its full potential. Here, different cell type compositions motivated the development of a modified version of ComBat for batch effect correction. It must be noted that the author of this thesis conducted the complete scRNA-seq data analysis, interpreted and discussed the results. The main focus of the scRNA-seq analysis is to robustly infer the cellular hierarchy in lineage commitment.

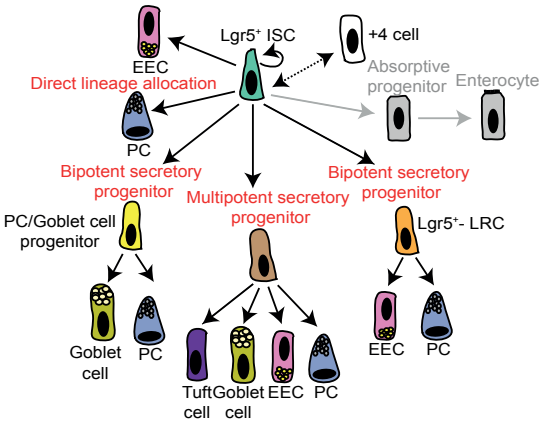

Figure 4.1: Competitive models in small intestine lineage determination. ISCs differentiate into an absorptive and secretory lineage. The lineage for secretory cells is largely debated. For example, Grün et al. [2016] inferred a bipotent progenitor for Paneth and goblet cells from single-cell RNA-seq, Kim et al. [2016] reported multilineage priming in ISCs and Buczacki et al. [2013] reported a *Lgr5+* label-retaining cell (LRC) being committed to Paneth and enteroendocrine lineage, respectively.

Experimental data collection such as immuno-staining, imaging, cell counting per crypt, and respective statistical analysis was performed by Anika Böttcher and co-authors.
The aim is to prove cell fate priming in intestinal stem cells and to disprove the concept of multipotent progenitors in the gut. First, we introduce the biological methods and data collection. Then, we outline the single-cell RNA-seq analysis steps that are fine-tuned towards the characteristics of the data set. In the results part, we present an overview of the data set and its cell type annotation based on transcriptomic profiling. Further, we analyzed the enrichment with FVR^+ and FVF^+ cells.

The small intestine has a constant self-renewal throughout adult life. A detailed understanding of intestinal stem cell (ISC) self-renewal and differentiation is required for better treatment options for a variety of chronic intestinal diseases. The lineage hierarchy of the intestinal epithelium is subject to a controversial debate [Buczacki et al., 2013, Grün et al., 2015, 2016, Kim et al., 2016, Muñoz et al., 2012, van der Flier and Clevers, 2009] (see Fig. 4.1).
Several signaling pathways are involved in the regulation of proliferation, stem cell maintenance, differentiation, and cell positioning within the crypt. Briefly, canonical Wnt/β-catenin signaling controls stem cell maintenance, but is strongly balanced by proliferation limiting signaling [van der Flier and Clevers, 2009]. Notch signaling supports stem cell maintenance and determines the differentiation into secretory and absorptive lineages, re-

spectively [Fre et al., 2005]. While Haber et al. [2017] present an extensive survey of cell types of the small intestinal epithelium, they do not speculate on lineage priming or the existence of multipotent or bipotent progenitors. Furthermore, a quiescent label-retaining cell (LRC) in the supra-Paneth cell position ('+4' position relative to the crypt bottom) has been reported [Basak et al., 2017, Buczacki et al., 2013, Muñoz et al., 2012]. Its progenies preferentially differentiate into Paneth and enteroendocrine cells. In contrast, Grün et al. [2015] reported a homogeneous stem cell population based on single-cell RNA sequencing.
However, Kim et al. [2016] report multilineage priming early on in the stem cell regime. Lastly, Grün et al. [2015] used transcriptomic profiling of single-cells and a clustering algorithm for rare cell types to postulate a common goblet/Paneth cell progenitor.

In this chapter, we report the identification of Lgr5$^+$ ISCs that express Flattop (*Fltp*), a Wnt/planar cell polarity (PCP) reporter and effector gene. Functional analysis and lineage tracing revealed that Wnt/PCP-activated Fltp$^+$ ISCs are primed towards either the enteroendocrine or Paneth cell lineage *in vivo*, while retaining self-renewal and multilineage capacity *in vitro*.
Surprisingly, canonical Wnt/β-catenin- and non-canonical Wnt/PCP-activated Lgr5$^+$ ISCs are indistinguishable by the expression of stem-cell signature or secretory lineage specifying genes. Hence, lineage priming and cell-cycle exit is triggered at the post-transcriptional level by polarity signaling, thereby repressing canonical Wnt signaling. Pseudotemporal ordering of targeted single-cell gene expression data allowed us to delineate the ISC differentiation path into enteroendocrine and Paneth cells. Thus, our data supports the direct lineage allocation model (see Fig. 4.1). To complement, a heterozygous knockout of Wnt/PCP signaling cascade (*Celsr1*) lead to a reduced number in Paneth cells as well as less proliferative Paneth progenitors. Therefore, active Wnt/PCP signaling is crucial for the normal development of Paneth cells.

4.1 Data collection

This section presents the experimental approach of this study. Experiments where carried out by Anika Böttcher and are described briefly.

Animal studies

Animal experiments were carried out in compliance with the German Animal Protection Act and with the approved guidelines of the Society of Laboratory Animals (GV-SOLAS)

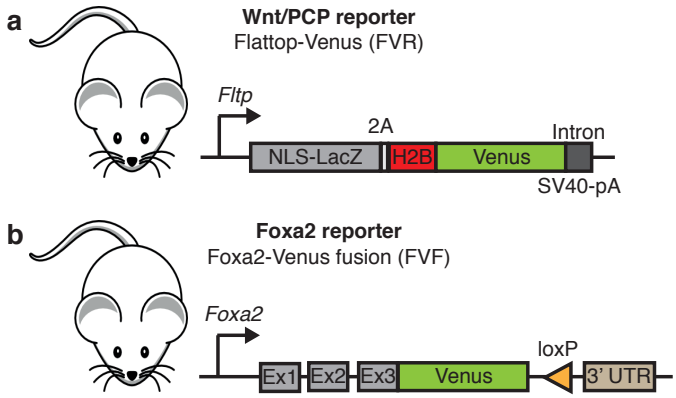

Figure 4.2: Transgenic mouse line reporters. **a** Structure of the Flattop-Venus reporter. **b** Structure of the Foxa2-Venus fusion reporter.

and of the Federation of Laboratory Animal Science Associations (FELASA). We used the following mouse models including transgenic mice (see Fig. 4.2):

- C57BL/6J strain (control mice)
- Fltp^{ZV} (C57BL/6J): knockout-knockin reporter — the open reading frame of *Fltp* was replaced by LacZ and H2B-Venus [Gegg et al., 2014],
- Foxa2-Venus fusion (C57BL/6J): open reading frame of *Foxa2* is fused to Venus
- Celsr1$^{crsh/+}$ (BALB/c): Celsr1$^{crsh/+}$ is a mouse line with a heterozygous missense mutation in the *Celsr1* coding region [Curtin et al., 2003].
- Celsr1$^{crsh/+}$; *Fltp*$^{ZV/ZV}$ (mixed BALB/c, C57BL/6J background). For single-cell RNA-seq, we used mice with *Fltp* knockout and Celsr1$^{crsh/+}$ heterozygous locus.

All experiments were performed using 3-6-month-old mice, unless indicated otherwise.

Crypt isolation, single cell preparation for FACS, sequencing

Isolation and culture of small intestinal crypts and organoid culture was performed as follows. Briefly, intestines were harvested and washed with PBS. Villi were scraped away using coverslips. The remaining tissue was cut into 2 cm pieces and incubated in 2 mM EDTA for 35 min at 4°C. Finally, crypts were collected by shaking.
For single cell preparation, the crypt pellet was resuspended in $1 - 1.5$ ml TrypLE, incubated on ice for 5 min, followed by 5 min incubation at 37°C in a water bath. Then, 6 ml of

crypt complete medium containing 10 % FCS and 10 μg/ml DNase were added, and cells were incubated for 5 min at 37°C in a water bath. The cells were gently resuspended by pipetting up and down 20 times, 10 ml FACS buffer (2 % FCS, 2 mM EDTA in PBS) was added and the cells were centrifuged at $300 \times g$, 5 min, 4°C. Cells were washed twice with FACS buffer and finally the cell pellet was re-suspended in $1 - 2$ ml FACS buffer containing 10 μM Rock-Inhibitor (Sigma, Y0503), and cells were passed through the 40 μm cell strainer caps of FACS tubes. Venus-positive cells were enriched by flow cytometry based on Venus fluorescence levels and dead cells were excluded using 7-AAD viability staining. Single cell libraries were generated using the *Chromium*^{TM} Single cell 3' library and gel bead kit v2 (PN # 120237) from *10X Genomics*. Briefly, to reach a target cell number of 10.000 cells per sample, 16.000 cells per sample were loaded onto a channel of the 10X chip to produce Gel Bead-in-Emulsions (GEMs). This underwent reverse transcription to barcode RNA before cleanup and cDNA amplification followed by enzymatic fragmentation and 5' adaptor and sample index attachment. Libraries were sequenced on the HiSeq4000 (Illumina) with 150 bp paired-end sequencing of read 2.

4.2 Computational analysis of single-cell RNA-seq data

This section presents the computational analysis of the scRNA-seq data in this study. The steps of the data analysis follow the general workflow described in Chapter 2. Sequence alignment and cell filtering was performed by Sophie Tritschler.

Pre-processing of droplet-based scRNA-seq data De-multiplexing and alignment to mm10 mouse genome, identification of unique molecular identifiers (UMI) and barcode filtering was performed using the 'CellRanger' toolkit (version 2.0.0) provided by 10X Genomics. We performed a further barcode (=cell) selection step and additionally included cells with more than 1,000 expressed genes, where a gene is counted as expressed if we found at least one UMI mapped to it. We further filtered cells with a fraction of counts from mitochondrial genes > 10 % indicative for stressed or dying cells.
We removed doublets by computing the doublet score from *scrublet* [Wolock et al., 2018] on the UMI count matrix separately for every sample. Using a threshold of 0.4, we removed 1651 cells from the analysis.
Cells from all samples were log-normalized and batch corrected using ComBat as Python implementation. It must be noted that we observed an overrepresentation of Paneth cells in the FVR-enriched samples (with 50 % and 90 % enrichment, respectively), where we subsampled the Paneth cell populations to 15 % to fit the cell type distribution of all

other samples before we corrected with ComBat. Then, we used the pre-computed regression coefficients to correct for batch effects in the filtered cells and merged them with the full data set. Further, we fixed all zero values to remain zero in order to preserve the support of the count data (see section 4.2.1). We evaluated the degree of the batch effect per cell type using kBET and the silhouette coefficient [Büttner et al., 2019] (see Chapter 3).
We computed the top 2,000 highly variable genes based on mean and dispersion (*i.e.* `pp.filter_genes_dispersion` in `Scanpy` [Wolf et al., 2018a] v. 1.3.1 in Python 3.6 with the flavor 'cell_ranger' to compute normalized dispersions [Zheng et al., 2017]). Further, we corrected for the library size by scaling the reads per cell to the factor 100,000 (`pp.normalize_per_cell` in `Scanpy` v. 1.3.1 in Python 3.6).

Dimension reduction We used a UMAP [McInnes and Healy, 2018] to represent the data in the two-dimensional embedding and data visualization (`tl.umap`). This was created based on a PCA-space with n=50 components, and the k-nearest neighbor graph on the PCA-space with k=30 (`tl.pca` and `pp.neighbors`). PCA etc. was computed on the scaled and normalized data with 2,000 highly variable genes.

Clustering and cell type annotation We determined the clustering and cell type annotation for the control samples as follows. We inspected first marker gene expression for respective major cell types and computed gene scores using known marker genes (see table 4.1, `tl.score_genes`, based on Satija et al. [2015]). Analogously, we computed a cell cycle score to determine the respective cell cycle phase state of the cells (`tl.score_genes_cell_cycle`, based on Satija et al. [2015], gene set specifying S and G2/M phase were taken from Tirosh et al. [2016]). Then, we performed clustering using the *louvain* algorithm [Blondel et al., 2008] (`tl.louvain` with default resolution parameter 1.0) and found 18 clusters. Here, we annotated and merged clusters again according to the gene scores and marker gene expression. Here, we also identified 1589 immune cells, which were extremely distinct from the remaining cells.
Subsequently, we inspected all main clusters for substructure and resolved it further based on marker gene expression. Here, we used louvain clustering with resolution parameters $\sim 1-2$ and merged the sub-clusters again according to marker gene expression (hierarchical 'split-and-merge' approach). Finally, we annotated 7 major cell types (ISC, enterocytes, goblet cells, Paneth cells, enteroendocrine cells and tuft cells), subdivided them into progenitor and mature cells. In the ISC population, we identified a Paneth primed ISC population that was more similar to the Paneth progenitor population.

Cell type	Gene names
ISC	*Lgr5, Ascl2, Olfm4, Prom1, Axin2, Fzd2, Fzd7, Lrp5, Lrp6, Notch1, Hes1, Smo, Yap1*
Enterocyte	*Elf3, Sis, Fabp1*
Goblet cell	*Muc2, Spdef, Foxa1*
Paneth cell	*Lyz1, Mmp7, Dll4*
EEC	*Gfi1, Neurog3, Neurod1, Chga, Chgb, Isl1, Arx, Pax6, Foxa2, Sst, Gck, Gcg, Tph1, Pyy*
Tuft cell	*Dclk1, Ptprc, Avil*

Table 4.1: Overview on marker genes used for cell type scoring.

For the mutant samples, we employed a k-nearest neighbor approach to match every cell to the corresponding cluster ($k = 30$), i.e. we derived the cell identity of the mutants from the cell identity of neighboring cells of the control samples, which we annotated beforehand. Again, we identified 2289 immune cells, which we removed from analysis.

Differential expression analysis We used the *limma* package [Ritchie et al., 2015] (version 3.34.9 in R 3.4.3) to study differential expression. First, we excluded the FVF enriched samples and mutant samples. In order to determine differentially expressed genes between ISCs and progenitor populations, we tested pairwise progenitor populations vs ISC (without Paneth primed ISCs) and Paneth progenitors vs goblet progenitors. It must be noted that differential expression was performed on the log-normalized (not batch-corrected) data, where we included the sample information (*i.e.* batch) as covariate. In addition, we excluded all genes with mean expression < 0.05. We considered only significant genes ($FDR < 0.05$) with $logFC > 0.05$. For the analysis of transcription factors, we filtered differentially expressed transcription factors (gene ontology ID GO:0003700 (transcription factor activity)) with the *biomartr* package (version 0.7.0 in R 3.4.3) [Drost and Paszkowski, 2017] with a p-value threshold $p_{adj} < 10^{-5}$.
Analogously, in order to determine differentially expressed genes between mutants and control samples (without FVF enriched samples), we tested every cluster separately with *limma*. It must be noted that differential expression was performed on the log-normalized (not batch-corrected) data, where we included the number of expressed genes as covariate, but not the sample information (*i.e.* batch) due to confounding of genetic condition and sample covariates. In addition, we excluded all genes with mean expression < 0.05. We considered only significant genes ($FDR < 0.05$) with $logFC > 0.05$. For the analysis of transcription factors, we filtered differentially expressed transcription factors (gene ontology ID GO:0003700 (transcription factor activity)) with the *biomartr* package (version

0.7.0 in R 3.4.3) and with an adjusted p-value threshold $p_{adj} < 10^{-5}$ (Benjamini-Hochberg correction).

Identifying cell differentiation trajectories via graph abstraction To derive cell trajectories, we computed a pseudotemporal ordering using diffusion pseudotime (DPT, `tl.dpt` in `Scanpy`) [Haghverdi et al., 2016]). As the topology of the data is complex, we used partition-based graph abstraction (PAGA) [Wolf et al., 2018b] to quantify the connections between the clusters (i.e. connections of ISCs to respective progenitors and mature cell types, `tl.paga` in `Scanpy`). We display all connections with a scaled connectivity of at least 0.05 ('threshold' parameter in `pl.paga` in `Scanpy`).

Gene Set Enrichment Analysis We performed a gene set enrichment analysis based on both GO [Ashburner et al., 2000, Consortium, 2016] terms and KEGG [Kanehisa et al., 2016] terms using g:profiler [Reimand et al., 2016]. In particular, we adapted the Python wrapper from V. Svensson (https://github.com/vals/python-gprofiler). We set the background to all expressed genes and removed all resulting gene sets with significance level $p \geq 0.05$. Further, we split the input data set by the sign of the log-fold change, such that we considered up-regulated and down-regulated gene sets separately in each set of differentially expressed genes. For visualization of gene set significance, we abridged p-values at 10^{-10}.

4.2.1 Batch correction using a modified version of ComBat

In the previous Chapter 3, we have found that the empirical Bayes method *ComBat* [Johnson et al., 2007] preserves vastly the biological signal while removing batch effects in case that all samples are similar. The method assumes that the gene mean expression is similar for all batches. The assumption holds true, if only one cell type is present or if the relative frequencies of the cell types is approximately the same. Our experimental setup aims to enrich for certain populations such as Paneth cells. In these cases, the gene mean expression of a Paneth cell enriched sample will reflect the characteristic Paneth cell gene expression profile, which differs largely from other cell types. Hence, the usual ComBat model will misjudge the difference in gene expression as batch effect. Our extended approach will now take the following steps:

(i) Identity batches with different cell composition by clustering

(ii) Downsample these batches to match the cell type composition of the other samples

(iii) Apply ComBat model using the downsampled data and save weights

(iv) Re-apply ComBat model with the pre-computed weights to remaining data

This model allows to batch correct data that has different cell compositions while preserving most of the biological variance.

4.2.2 Batch effect quantification

We evaluated the degree of batch effects across samples separately in each major cell type before and after correction (see Fig. 4.3). Quantifying batch effects with kBET resulted in a rejection rate of 1 before correction in all cell types. After batch correction with the modified ComBat, we still found rejection rates of $0.98 - 1$. Therefore, kBET is too sensitive to assess the batch effect removal of this method. However, we need batch correction primarily for unbiased clustering, *i.e.* to reduce clusters sampling only cells of a single batch. As alternative to kBET, we used the silhouette coefficient to quantify batch effects, as it is less sensitive than kBET (see Chapter 3).

When we applied ComBat without accounting for cell type composition, homogeneous cell types such as enterocytes became more similar (measured by decrease of the silhouette coefficient on the first 3 PCs), while Paneth cells are separated. Interestingly, even though samples 'Control_5_FVR' and 'Control_7_FVR_only' were experimentally handled together and overlapped largely before correction, they did not after correction due to their different cell type composition. This example highlights the importance to account for cell type composition in the samples. The separation is clearly visible in the UMAP plot (Fig. 4.3a), however, it is not reflected in the silhouette coefficients per cell type (Fig. 4.3b). Therefore, we tested how louvain clustering is affected by batch effects (Fig. 4.3c). In theory, the cluster should consist of a mixture of cells of different samples. In contrast, batch effect biased clustering yields clusters that are dominated by the cells of one sample. Here, we used louvain clustering with the same resolution parameter to assess the batch effect on clustering before and after correction.
Before correction, we observe several clusters consisting of a single sample (e.g. clusters 9 and 15). Furthermore, we observe a banded structure in the UMAP plot, which means that the cells cluster by sample rather than cell type. In the ComBat corrected dataset, we observe a more balanced clustering for most samples, but samples 'Control_5_FVR' and 'Control_7_FVR_only' separate now from each other and from the remaining samples. This issue is accounted by in the modified ComBat batch correction. There, we observe less clusters and 8 out of 9 clusters show a balanced contribution from all samples.

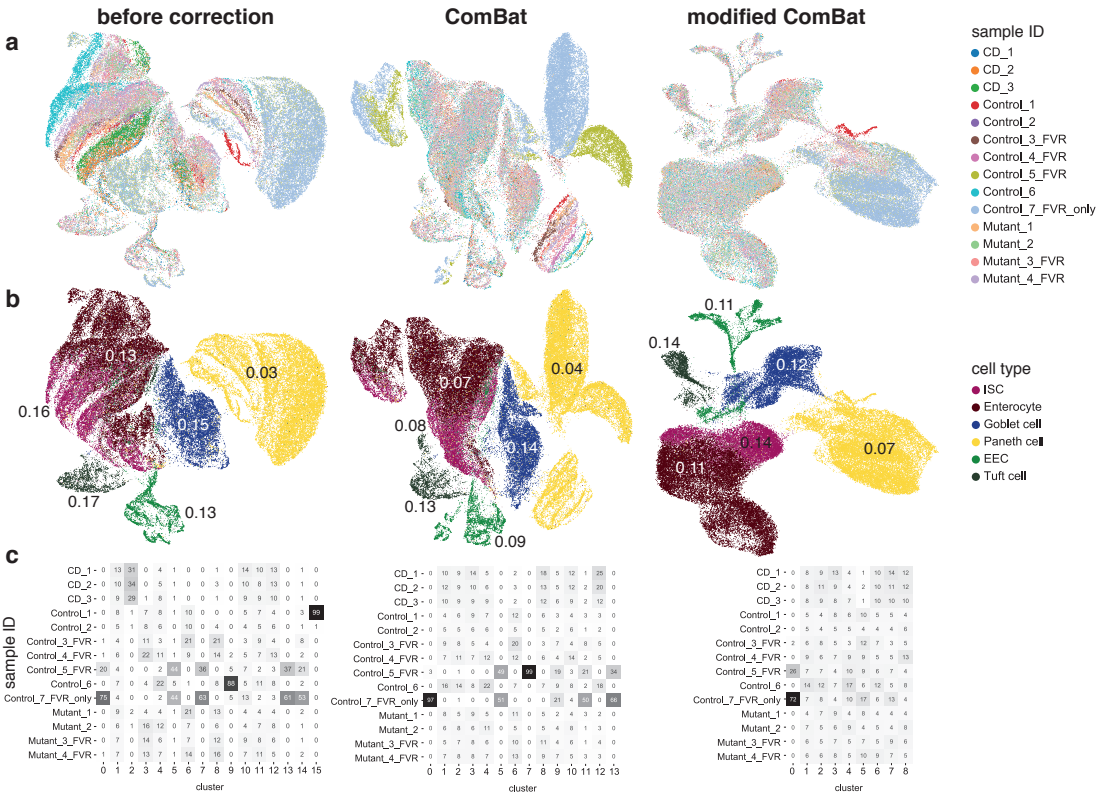

Figure 4.3: Modified ComBat corrects accurately for batch effects in data with compositional changes. Comparison of batch effects in the gut dataset. Columns show results for non-corrected data (left), ComBat corrected data (center) and modified ComBat corrected data (right). **a** UMAP plots colored by sample ID. Banded structure in non-corrected data indicates prominent batch effect. **b** UMAP colored by major cell type. Numbers denote the absolute silhouette per major cell type. Little variation in the silhouette is found across correction methods. **c** Confusion matrix of sample ID and the result of louvain clustering with resolution parameter 0.5, normalized column-wise. Grey scale and numbers denote the fraction of cells of sample i in cluster j. Before correction, louvain clustering yielded 16 clusters, where two clusters carry mostly cells from one sample. The number of clusters reduces with batch effect correction, and clustering on modified ComBat corrected data shows a balanced contribution of all samples to all clusters. The exceptions are samples 'Control_5_FVR' and 'Control_7_FVR_only', which are enriched for $50\,\%$ and $90\,\%$ Paneth cells, respectively.

4.3 Results

In this section, we present the results of scRNA-seq data analysis together with experimental results from laser scanning microscopy (LSM). The author of this thesis conducted scRNA-seq data analysis, while Anika Böttcher and co-authors performed LSM.

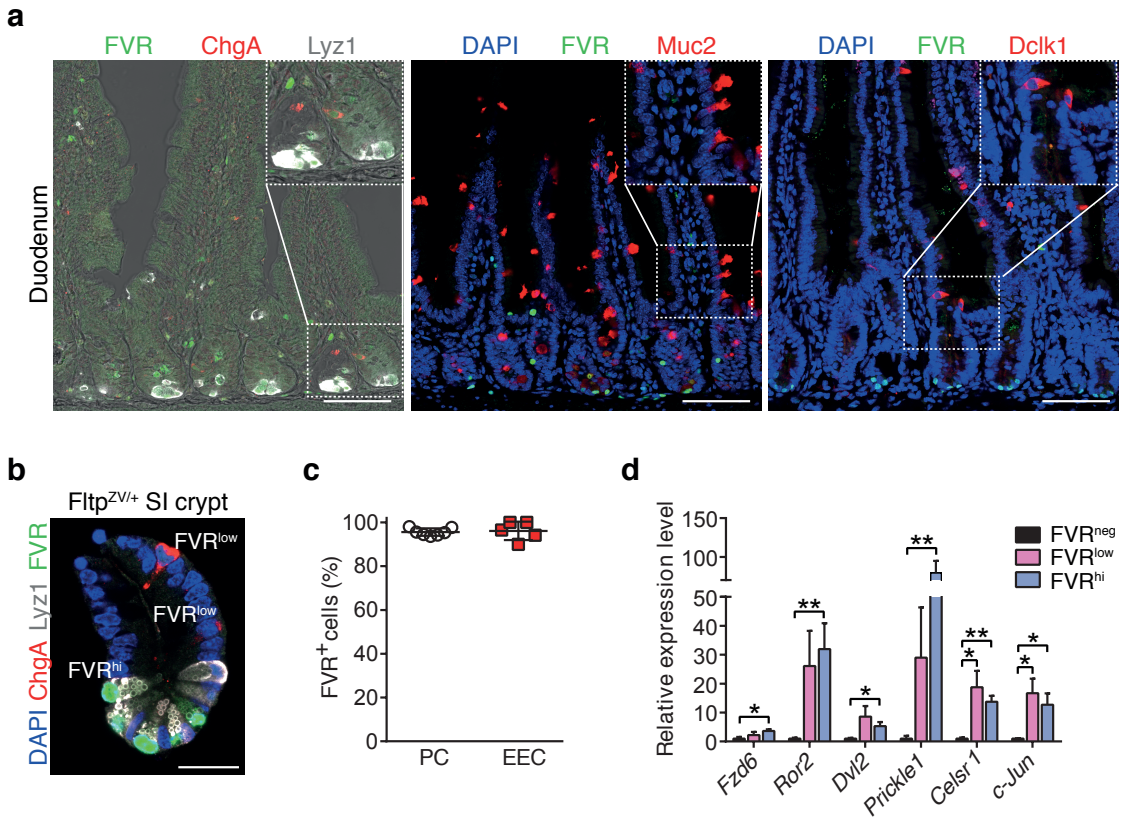

Figure 4.4: FVR labels active Wnt/PCP signaling in enteroendocrine and Paneth cells. a LSM images for FVR (Venus, green) expression in the secretory lineages in the adult intestine of Fltp$^{ZV/+}$ mice (4 months) co-stained against ChgA (red, enteroendocrine cells), Lyz1 (white, Paneth cells), Muc2 (red, goblet cells), and Dclk1 (red, tuft cells). DAPI (blue) stains the nucleus. Scale bars, $75\,\mu m$. **b-c** LSM image showing a representative SI crypt isolated from adult Fltp$^{ZV/+}$ reporter mice with indicated FVR$^{neg/low/hi}$ cells stained for DAPI (blue, nucleus), FVR (Venus, green), ChgA (red, enteroendocrine cells), and Lyz1 (white, Paneth cells) (**b**) and quantification (**c**) of Lyz1$^+$ FVR$^+$ Paneth cells (PCs) and ChgA$^+$ FVR$^+$ enteroendocrine cells (EECs). For Paneth cells: $n\,(\text{mice}) = 8$ with 99 analyzed crypts ($95.58\,\%$ of PCs = FVR$^+$). For enteroendocrine cells: $n\,(\text{mice}) = 5$ with 76 analyzed crypts ($96.13\,\%$ of EECs = FVR$^+$). Scale bar, $25\,\mu m$. **d** Expression analysis of Wnt/PCP genes in FVR$^{hi/low/neg}$ crypt cells obtained from Fltp$^{ZV/+}$ mice by qRT-PCR. n(independent experiments) $= 3$. Error bars represent SEM. Two-tailed Student's t-test, $*P < 0.05$, $**P < 0.01$.

4.3.1 *Fltp* is a new marker for Paneth and enteroendocrine cells

To analyse the impact of Wnt/PCP activation on ISC self-renewal and differentiation in the crypt stem cell niche we took advantage of the first reported sentinel for pathway activation [Gegg et al., 2014], our knock-out knock-in Fltp^{ZV} dual reporter mouse model (see Fig. 4.2). In this model Fltp-H2B Venus reporter (FVR) activity is restricted to

quiescent and terminal-differentiated cells that activated *Fltp* expression during Wnt/PCP acquisition. Co-staining with markers for EECs (ChgA), Paneth cells (Lyz1), goblet cells (Muc2) and tuft cells (Dclk1) reveals a distinctive restriction of FVR to Paneth cells and EECs, while none of the goblet or tuft cells carried an FVR label.
Closer investigation of the small intestinal (SI) crypt shows a clear subdivision into three populations according to FVR label intensity: FVR^{neg}, FVR^{low} and FVR^{hi} cells (see Fig. 4.4b). Vice versa, FVR labelled almost all $Lyz1^+$ PCs (95.58 %) and $ChgA^+$ crypt EECs (96.13 %)(see Fig. 4.4c). Thus, *Fltp* expression marks the cell fate decision towards Paneth or enteroendocrine cells.

To examine the activation of Wnt/PCP signaling in FVR^+ populations, we investigated the expression of pathway genes. The Wnt/PCP receptor Fzd6, the Wnt/PCP co-receptor Ror2, as well as core Wnt/PCP genes Prickle1, Celsr1, and c-Jun were highly expressed in both FVR^{low} and FVR^{hi} cells (Fig. 4.4d). In contrast, the same Wnt/PCP pathway genes are absent in FVR^{neg} cells. Hence, the absence of *Fltp* correlates with inactive non-canonical Wnt/PCP signaling.

4.3.2 Wnt/PCP activates a subset of $Lgr5^+$ ISCs

Notably, the presence of FVR in Paneth cells and EECs does not allow backtracing the exact cell with Wnt/PCP activation. We hypothesize an early cell fate priming in ISCs and screened for *Fltp* expression using a targeted single-cell RT-qPCR approach. Indeed we detected *Fltp* expressing $Lgr5^+$ ISCs (6.2 %), whereas expression was low or absent in FVR^{low} EECs and FVR^{hi} PCs, respectively (see Fig. 4.5a, b). Interestingly, more of the FVR^{low} than FVR^{hi} cells express *Fltp*. Furthermore, FVR^{low} cells expressed significantly higher levels of *Fltp* compared to FVR^{neg} cells. Thus, *Fltp* is transiently expressed in ISCs after Wnt/PCP activation and FVR (due to reporter protein stability) labels the immediate daughter cells of $Fltp^+$ ISCs, namely more than 95 % of the EECs and PCs.

4.3.3 Single-cell RNA-seq recovers all lineages in the small intestinal epithelium

We sequenced in total over $76,000$ cells to elucidate cell fate priming and to clarify the lineage determination from intestinal stem cells (ISCs) into Paneth cells and enteroendocrine cells. Using unsupervised clustering, followed by the study of marker gene expression and proliferation status (see methods 4.1), we identified ISCs, enterocytes, goblet cells, Paneth cell, EECs and tuft cells as distinct clusters as visualized in the UMAP plot (see Fig. 4.6a). We also found over $1,500$ immune cells in two distinct cluster, which we excluded from

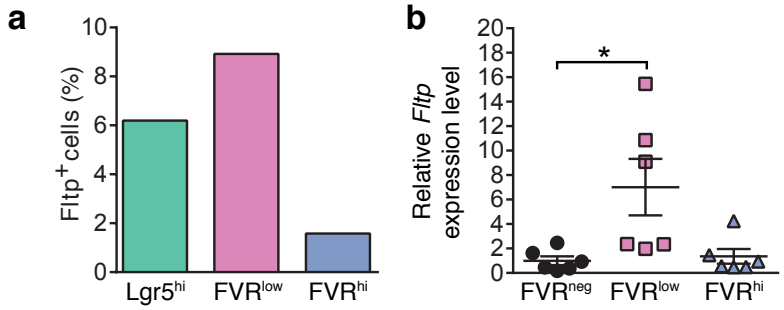

Figure 4.5: Wnt/PCP primes a subset of Lgr5$^+$ ISCs towards Paneth cells or EECs. **a** Relative abundance of Fltp$^+$ cells in Lgr5^{hi} cells, FVR^{low} cells and FVR^{hi} cells determined by single cell qPCR. $n = 145$ Lgr5^{hi} cells; $n = 112$ FVR^{low} cells; $n = 126$ FVR^{hi} cells. **b** qRT-PCR data showing relative Fltp expression in FVR$^{neg/low/hi}$ cells isolated from adult Fltp$^{ZV/+}$ reporter mice. n(independent experiments) $= 6$, two different *Fltp* probes. Error bars represent SEM. Two-tailed Students's t-test, $*P < 0.05$.

further analysis. Subsequently, we investigated every cell type to distinguish mature cells from the respective progenitors (see Fig. 4.6b). Hierarchical clustering based on highly variable genes places most progenitors next to the corresponding mature population (see Fig. 4.6c). Paneth cells and progenitors have no to negative correlation with all other cell populations demonstrating their particular differentiation process in contrast to all other cell types. ISCs are closely related to the enterocyte progenitor, also because we lack a position information in the scRNA-seq data, which is present otherwise with imaging. Interestingly, the EE progenitor is both highly correlated with ISCs and EECs, while being placed closer to ISCs by hierarchical clustering.

We used cell type specific signatures and a cell cycle signature [Tirosh et al., 2016] to support our clustering analysis (see table 4.1 and Fig.s 4.6d, e). The scores for tuft cells and EECs are very specific the respective cell types while not all enterocytes have high enterocyte score and vice versa, some terminally differentiated tuft cells and EECs also appear to express enterocyte genes. Thus, enterocytes appear more diverse and some of the markers less specific than expected. Furthermore, ISC scores also extend to Paneth, EE and tuft progenitors indicating a possible stem cell property in early differentiating cells. On the other hand, Paneth and goblet cell scores are both high in both populations, *i.e.* expression of similar genes. Thus, these cells have either common origin or similar function. We have effectively ruled out the former by FVR lineage tracing, while we inspect the latter in a subsequent section.
In terms of proliferation, ISCs and all progenitor populations are active proliferating, while terminally differentiated cells are quiescent (see Fig. 4.6e). In the case of goblet cells, we

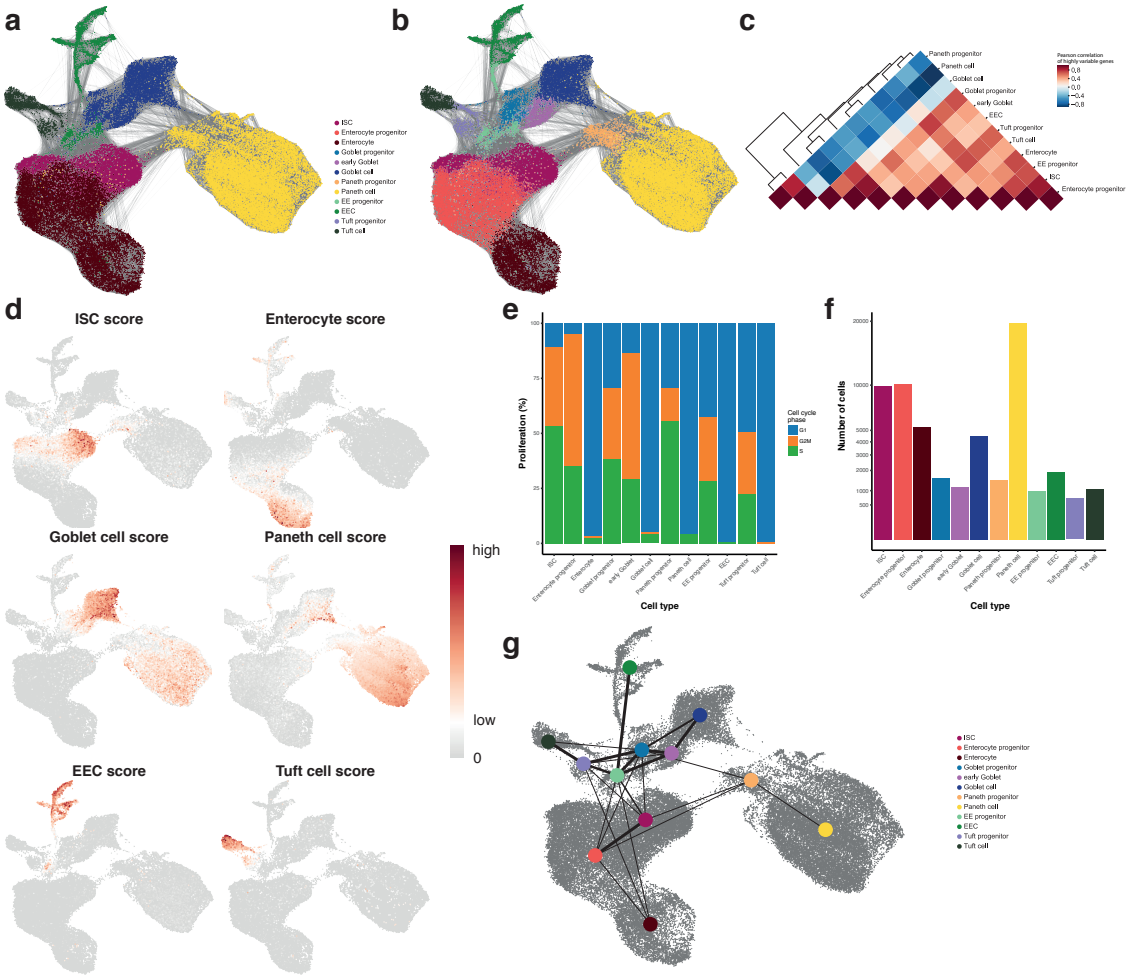

Figure 4.6: Single cell gene expression of major cell types in the intestinal epithelium. All plots refer to the control (wild type) samples. **a** UMAP plot of wild type samples with major cell types annotation. **b** UMAP plot of wild type samples with progenitor cell type annotation. **c** Correlation map of highly variable genes between progenitor cell types. **d** Gene score of major cell types. **e** Proliferation score of the progenitor and mature cell types. **f** Number of cells per cell type. **g** Abstracted graph of cell type connections. Thick lines denote a high confidence of the connection, thin lines correspond to connections with lower, but significant confidence. ISCs connect to all progenitor populations.

also distinguished between early and mature cells because early goblet cells are still actively cycling, but also show obvious and pronounced goblet marker gene expression, which is also different from the goblet progenitor.

Our data set of the small intestinal epithelium reflects all common cell types with sufficient coverage of rare cell types as EECs and Paneth cells as well as progenitor populations (see Fig. 4.6f). On the one hand, the cell type frequencies of the complete data set do not

correspond to the actual cell type frequencies in the small intestinal epithelium and we cannot use the scRNA-seq data to study differences in cell type frequency shifts. On the other hand, a sufficient coverage in all cell types allows us to derive a complete lineage tree via graph-based topology analysis (see method section 4.1 and Fig. 4.6g). Line thickness corresponds to higher confidence of the connection. Notably, ISCs connect directly to all progenitor populations indicating a direct allocation into the respective lineages. Moreover, all progenitors connect the most to the respective progenitor populations. In particular, EECs only connect to EE progenitors and Paneth cells only connect to Paneth progenitors.

4.3.4 Enrichment for reporter expression captures Paneth cells and EECs

The enrichment for specific cell types enables the study of lineage determining gene expression patterns, which are unnoticed otherwise due to low cell type abundance. Moreover, large cells as Paneth cells more rarely pass the cell preparation process [Haber et al., 2017]. In this study, we have sequenced cells with increasing FVR enrichment and compared them to FVF enrichment and un-biased sampling (*i.e.* no enrichment, see Fig. 4.7). Notably, the samples without enrichment mostly yield ISCs, enterocyte progenitors and enterocytes ($\sim$ 75 %). Further, 5.3 % Paneth cells without enrichment and 0.8 % Paneth progenitor cells. For comparison, FVF enrichment, which enriches for the rare secretory cell types and their progenitors, yields only 2.2 % Paneth cells and 0.2 % Paneth progenitors. Hence, studying Paneth progenitors and Paneth cell maturation, we have to enrich specifically for the Paneth lineage.
In order to increase the sampling from both Paneth cells and Paneth progenitors, we sequence the small intestinal crypt of $\text{Fltp}^{ZV/+}$ (C57BL/6J) mice. The FVR reporter system marks Paneth and enteroendocrine cells exclusively (see sections 4.3.1 and 4.3.2). Thus, enriching with FVR yielded exceedingly more Paneth cells (see Fig. 4.7). Furthermore, the fraction of goblet cells and progenitors is diminishing in the FVR enrichment. We observed *Fltp* expression already in ISCs such that we can trace the lineage of these stem cells in the mature cell types (see sections 4.3.1 and 4.3.2). Consequently, as we find an excessively large number Paneth cells and progenitors via FVR enrichment, we support the claim of *Fltp* determining the Paneth cell lineage. Furthermore, as we do not find more goblet cells with FVR enrichment, we exclude the existence of a bipotent goblet-Paneth cell progenitor, which was speculated on before [Grün et al., 2016].

Instead, FVR lineage tracing and the sample enrichment for scRNA-seq supports the existence of a Paneth primed ISC population, which we identified in our data set and distinguished it from all other ISCs (Fig. 4.8). In particular, ISCs express higher levels of

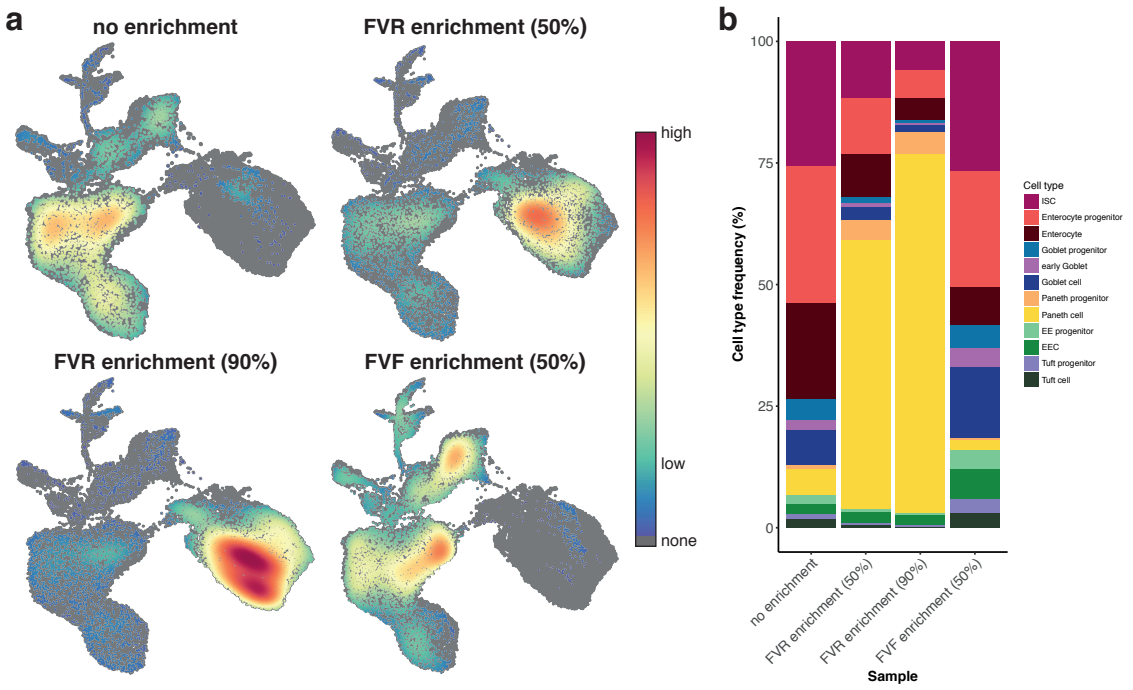

Figure 4.7: Subtype enrichment reveals distinct progenitor stages for each intestinal lineage. a Cell density plots of increasing FVR enrichment as well as FVF enrichment indicate the lineage determining potential of *Fltp* on the UMAP plot. **b** Relative abundance bar plot for mature and progenitor cell types. FVR enriched samples mainly consist of Paneth cells and its progenitors, which show low abundance in un-enriched and FVF enriched samples.

cell cycle genes and genes involved in the cellular response to radiation (Fig. 4.8a). Both GO terms support the stem cell identity of ISCs as they are actively proliferating and are more susceptible to radiation [Buczacki et al., 2013]. Furthermore, Paneth primed ISCs connect with higher confidence to the Paneth progenitor population in the abstracted graph and, in contrast to ISCs, do not connect to all progenitor populations (Fig. 4.8b). Hence, the identified subset of Paneth primed ISCs have slightly reduced cell cycle activity and are less susceptible to radiation compared to ISCs. Possibly, Paneth primed ISCs contain a quiescent subpopulation of label-retaining cells (LRCs) [Buczacki et al., 2013], whose existence is still under debate.

For enteroendocrine cells, we do not observe a particular increase in proportion during FVR enrichment, even though these compose the FVR^{low} population. Possibly, cell sorting biases towards higher fluorescence levels such that FVR^{low} cells are underrepresented.
In contrast, sequencing FVF enriched samples yielded increased proportions of goblet cells as well as enteroendocrine cells, but no Paneth cells. Again, *Foxa2* is also expressed in

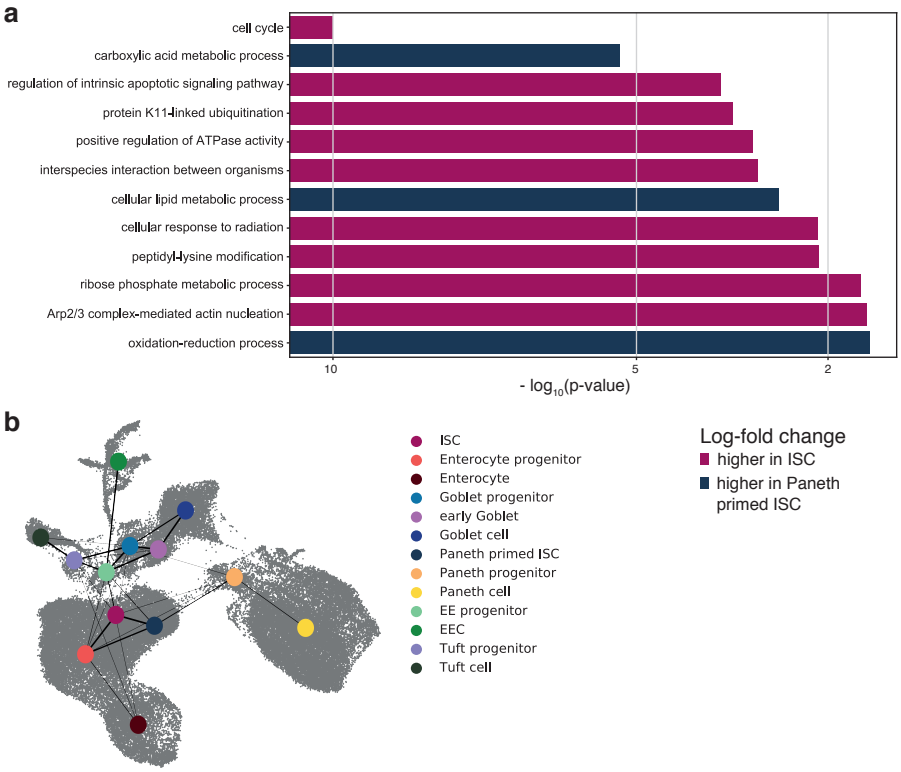

Figure 4.8: A subset of ISCs is primed towards Paneth cell lineage. a Gene set enrichment analysis with *g:profiler* based on GO terms on differentially expressed genes Paneth primed ISCs and ISCs. **b** Abstracted graph of cell type connections. Thick lines denote a high confidence of the connection, thin lines correspond to connections with lower, but significant confidence. Paneth primed ISCs connect with higher confidence to Paneth progenitors than remaining ISCs.

ISCs implying a lineage priming by *Foxa2* expressing ISCs.

Next, we investigated the transition of ISCs to Paneth cells (Fig. 4.9a). We determined several sub-groups in both ISCs and Paneth cell populations. In ISCs, we identified a Paneth primed ISC population that connects to the Paneth progenitor population. In mature Paneth cells, we have found two different types of Paneth cells which we distinguish by the expression of *Lyz2* — Paneth 1 population does not express *Lyz2*, but Paneth 2 population does. Pseudotemporal ordering of ISCs and Paneth cells clearly places the Paneth progenitor in between the mature Paneth cells and ISCs, while Paneth primed ISCs link to Paneth progenitors (see also Fig. 4.9b). The Paneth progenitor population starts to express the complete set of defensins and *Lyz1*, but it also expresses stem-cell associated genes as shown in the ISC score compared to the Paneth cell score (Fig. 4.9c). The common ISC marker gene *Lgr5* decreases simultaneously with Paneth cell maturation

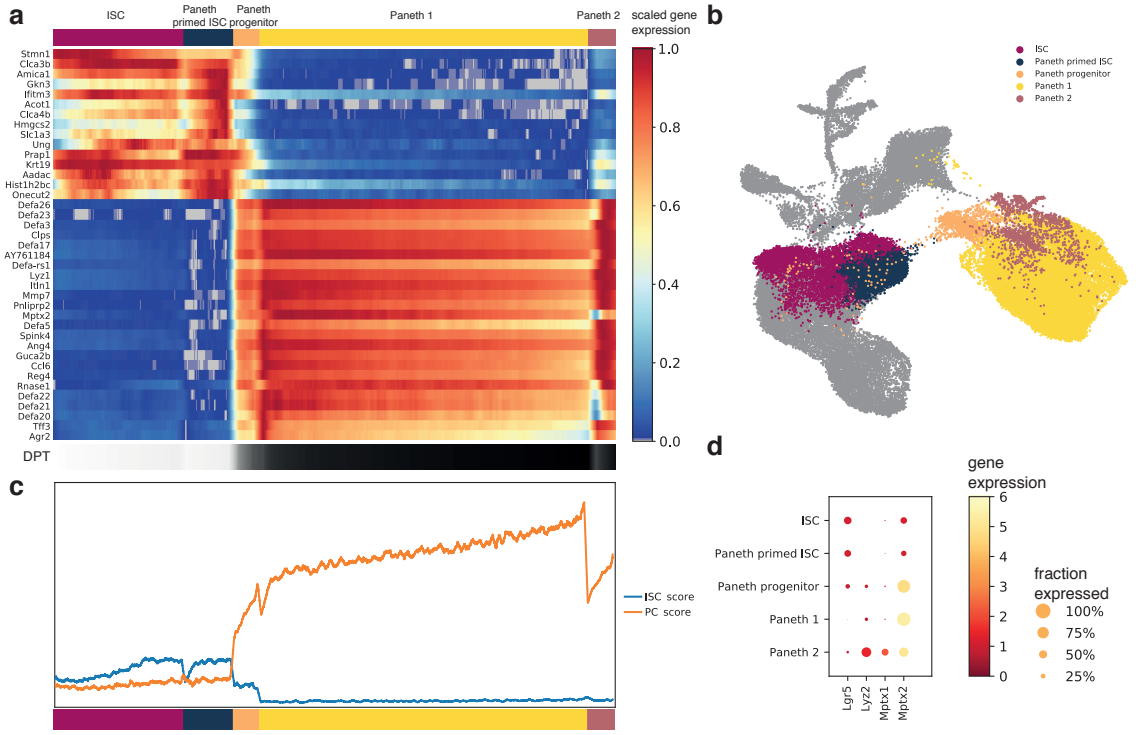

Figure 4.9: Pseudotemporal ordering for PCs reveals direct allocation from ISCs. a Smoothed gene expression heatmap of differentially expressed genes of the displayed groups plotted along diffusion pseudotime from ISCs to mature Paneth cells. We have identified an ISC subpopulation that is primed towards Paneth cell fate (*i.e.* Paneth primed ISCs) and a second *Lyz2*-positive Paneth cell population. **b** UMAP plot with populations contributing to the Paneth cell maturation. **c** ISC score and Paneth cell (PC) score along pseudotime shows high ISC score in Paneth primed cells. Notably, Paneth progenitors express both ISC and PC markers. **d** Dotplot of *Lgr5*, *Mptx1/2* and*Lyz1/2* expression in the ISC - Paneth cell lineage. Stem cells and progenitor cells show substantial *Lgr5* expression, which is mostly absent in mature Paneth cells.

— while both ISC and Paneth primed ISCs express *Lgr5*, a notable fraction of Paneth progenitors expresses *Lgr5* as well (Fig. 4.9d). Subsequently, *Lgr5* expression is absent in mature Paneth cells. Taken together, we have found considerate evidence for a direct lineage allocation of Paneth cells from ISCs.

In the enteroendocrine lineage, we observe a progenitor population with several sub-states (see Fig. 4.10a). In the enteroendocrine (EE) progenitor population, we distinguish a *Sox4* positive progenitor expressing lineage determining transcription factors (*Hes1*, *Dll1* and *Sox4*) and a *Neurog3-Sox4* double-positive progenitor (see Fig. 4.10b). Using diffusion

pseudotime to assess the sequence of these populations, the *Sox4* positive population resembles more the ISC population but it downregulates *Lgr5* while turning on *Sox4* — therefore we label this population '*Sox4+* early EE progenitor'. The late stage of the EE progenitors starts to express *Neurog3* alongside with *Neurod1* and *Neurod2*, hence we labelled it '*Ngn3* progenitor' (Fig. 4.10c). Finally, EE progenitors do not express the characteristic marker genes *Chga* and *Chgb*.

All EECs express both *Chga* and *Chgb*, but we also identified several progenitor populations expressing a distinct pattern of lineage determining markers (see Fig. 4.10c), similar to the *Ngn3* progenitor. We term these cells '*late EEC progenitors*'. Furthermore, we observe increasing expression levels of the the transcription regulating gene *Pax6*. A closer inspection of the late EEC progenitors reveals the *Isl1/Arx* and *Ghrl* progenitor populations, which express the factors *Isl1*, *Arx* and in case of the *Ghrl* progenitor, high amounts of the hunger hormone *Ghrl*. In contrast to the *Isl1/Arx* and *Ghrl* progenitors, we also identified a *Pax4* progenitor, which expresses high amounts of *Pax4* and lacks *Isl1*, *Arx* and *Ghrl* expression. The expression pattern in these late EEC progenitor populations indicates a cell fate decision towards different subclasses of EECs, which we describe in the following.

Using the hierarchical version of louvain clustering with manual merge of similar populations based on marker genes, we identified 8 different hormone secreting EEC subpopulations (see Method section). Traditionally, EECs were classified by hormone expression [Gribble and Reimann, 2016, Sjölund et al., 1983], assuming they secrete predominantly a single hormone (see table 4.2). However, evidence points towards considerable co-expression in single cells [Gribble and Reimann, 2016]. Recently, Haber et al. [2017] demonstrated via smFISH and scRNA-seq that EECs may well produce more than one hormone and defined new cell types according to the one-letter code in table 4.2.

The subclass of enterochromaffin (EC) cells is characterized by serotonin secretion (marker gene is *Tph1*). In accordance to Haber et al. [2017], we identified two EC subpopulations; one with high *Reg4* expression and the other lacking it. Presumably, these cells are derived from the *Pax4* progenitor as the EC (without *Reg4* expression) population shows both proximity in the UMAP plot and elevated levels of *Pax4*, that is absent in all other cell populations. Interestingly, all EEC subpopulations including the previously described progenitors express *secretin (Sct)*, the traditional marker for S cells.

Apart from EC cells, we observe six different EEC subpopulations, which express a combination of several hormones, which are described in the classical EEC classification scheme (see table 4.2). For example, we found a population expressing *Sct*, *Cc)* and *Gip*, which are markers for S cells, I cells and K cells, respectively. Thus, we follow the new taxonomy proposed by Haber et al. [2017] and term this population *SIK* cells. It must be noted

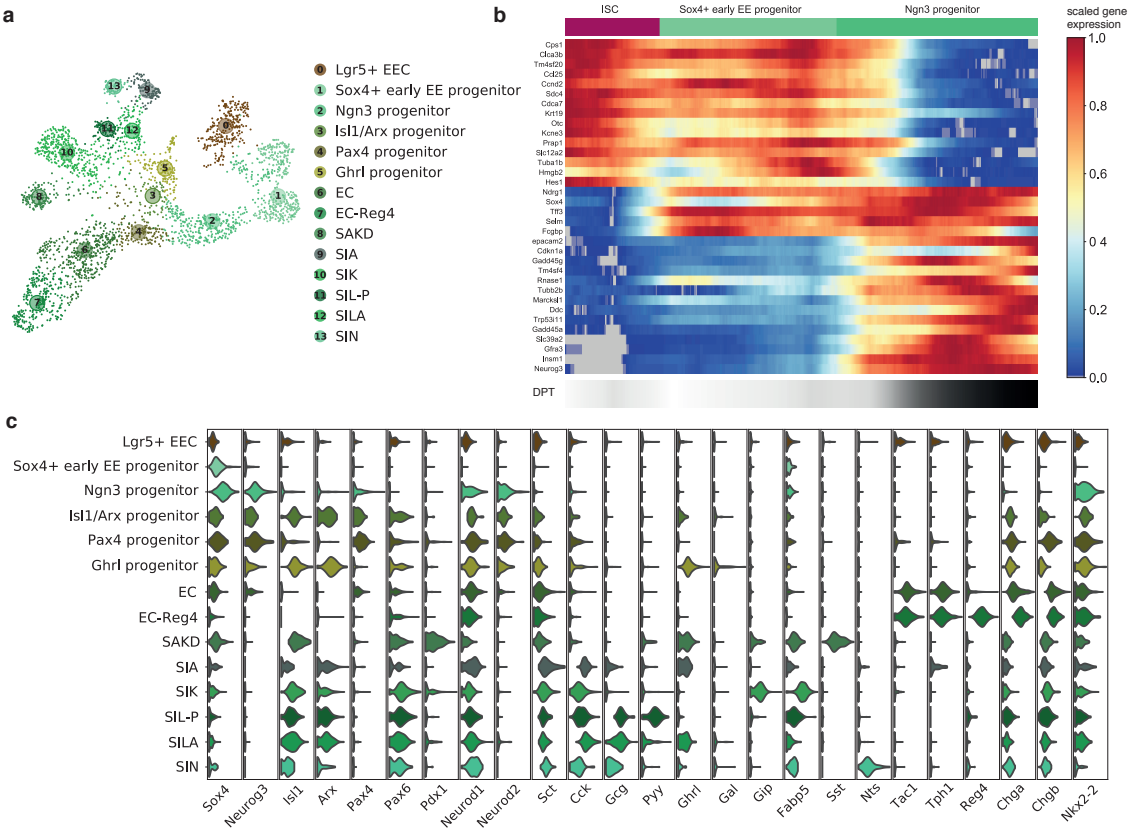

Figure 4.10: Pseudotemporal ordering for EECs reveals direct allocation from ISCs. a UMAP plot of EECs and progenitors with annotated subpopulations. **b** Smoothed heatmap of differentially expressed genes of the displayed groups plotted along diffusion pseudotime from ISCs to late enteroendocrine progenitors. **c** Marker gene expression in enteroendocrine subpopulations. Ngn3 and Sox4 positive progenitors are counted as enteroendocrine progenitors whereas late EEC progenitors (Isl1/Arx, Ghrl and Pax4) are already in the mature domain of EECs. Lgr5$^+$ EECs form an extra cluster as their gene expression profile reflects all other EEC sub-clusters.

that the SIK population is well separable by louvain clustering. Some of the hormones are exclusively expressed in only one of the subpopulations, for example *neurotensin (Nts)* in SIN cells and *somatostatin (Sst)* to SAKD cells. Notably, the SIA population expressing *Sct*, *Cck* and *Ghrl*, but not *Gcg* has not been reported before. In contrast to Haber et al. [2017], we can report a single subclass expressing *peptide YY (Pyy)*, SIL-P population) and *galanin (Gal)* expression is not restricted to SILA cells in our data. Further, we have found a single cell within the SIK population also expressing *gastrin (Gast)*.

Presumably, all these subpopulations arise from *Ghrl* and *Isl1/Arx* progenitors, respectively, as five of the six subpopulations express *Isl1* and *Arx*. For the sixth population,

hormone name	gene symbol	class
ghrelin	*Ghrl*	A
somatostatin	*Sst*	D
gastrin	*Gast*	G
cholestocystokinin	*Cck*	I
proglucagon	*Gcg*	L
glucose-dependent insulinotropic polypeptide	*Gip*	K
neurotensin	*Nts*	N
secretin	*Sct*	S
serotonin	*Tph1*	EC

Table 4.2: EEC classification scheme. The classical EEC classification defines a subclass by principal hormone secretion. Adapted from Haber et al. [2017] andGribble and Reimann [2016]. It must be noted that serotonin is a tryptophan derivate, in which the first and rate-limiting step in serotonin biosynthesis is performed by the tryptophan 5-hydroxylase 1 (Tph1) [Cooper and Melcer, 1961].

SAKD cells, turning off *Arx* allows expressing *Sst* [Du et al., 2012]. In addition, *Ghrl* expression was found in three out of six subpopulations (SIA, SILA and SAKD).
Interestingly, we also observe an *Lgr5* positive EEC population mirroring all EEC subclasses. They originate from all samples irrespective of the sample size, so they are not doublet artifacts, but their role in the gut and lineage forming potential is unknown.

4.3.5 Each intestinal lineage has a distinct transcriptional profile

In this section, we study the distinct transcriptional profiles of all mature cell populations in comparison to stem cells by differential gene expression analysis. We also inspect the same differential gene expression patterns in the respective progenitor populations to identify early transcriptional changes that define the respective lineages. Further, we can re-analyze the gene scoring approach that we used to identify mature cell types. In particular, we will propose novel marker gene sets to distinguish goblet and Paneth cells.

We performed pairwise differential expression tests of ISCs vs enterocytes, goblet cells, Paneth cells, EECs and tuft cells, respectively and display the top 15 differentially expressed genes (see Fig. 4.11, filtered from pseudogenes, ribosomal protein-coding genes and mitochondrial genes). In ISCs, we find mostly cell cycle related factors (*e.g. Cdca7* and *Smoc2*) or indicators for active transcription (*e.g. Impdh2*, *Hmgn1* and *Hnrnpa1*). In enterocytes, we find *Sis*, a known marker gene, while the expression of factors like *arginase 2 (Arg2)* and *galectin3 (Lgals3)* is equally well restricted to enterocytes. However, *Sis* is also considerably expressed in EECs and we found the *Microsomal Triglyceride Transfer*

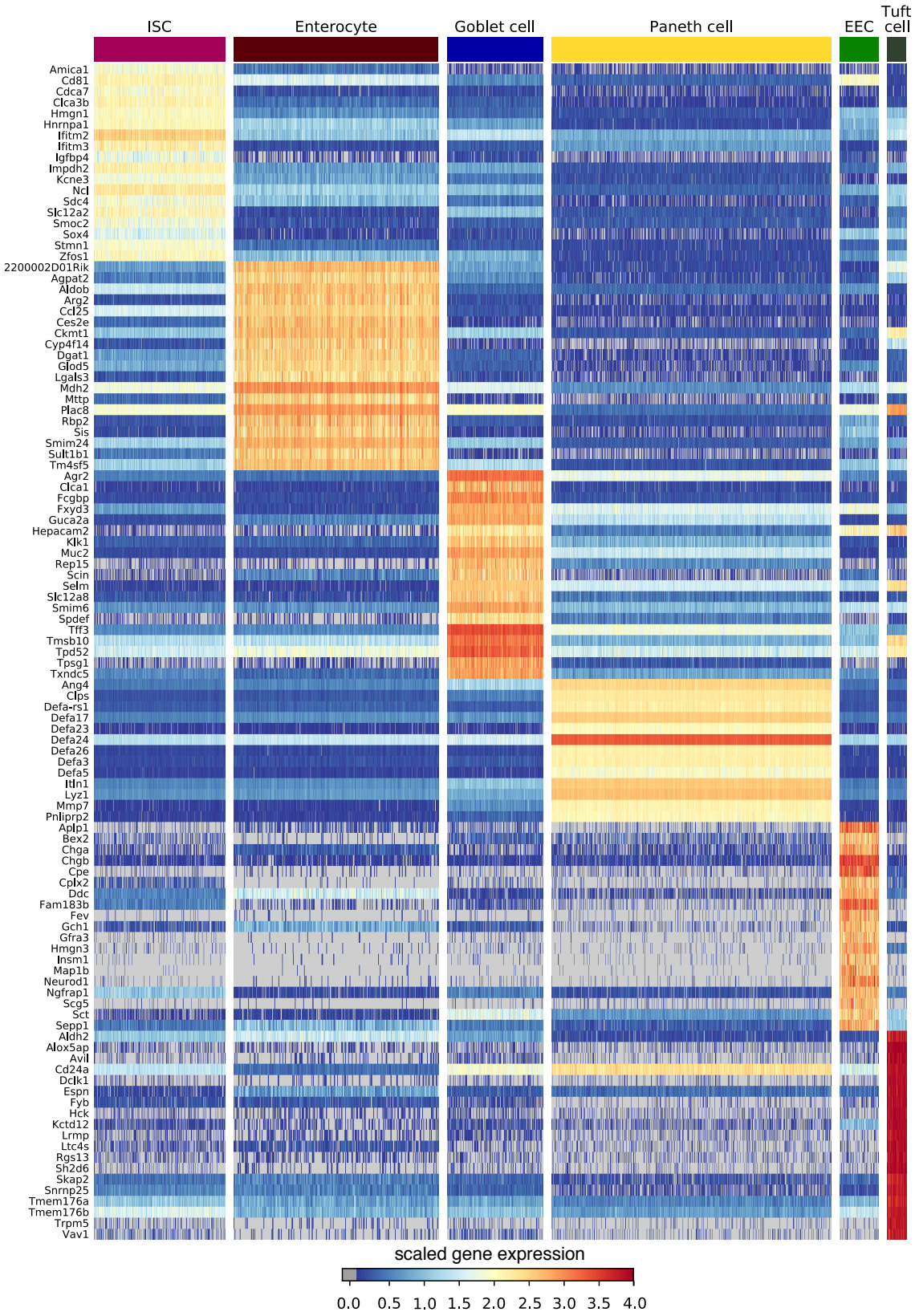

Figure 4.11: Differential expression analysis of mature cells and ISCs reveals new marker genes. We selected the top 15 differentially expressed genes in the comparison 'ISC vs mature cell type', which are sorted block-wise alphabetically. Among those are both known marker genes (see table 4.1) and new markers, *e.g.* *Espn* and *Avil* in tuft cells and *Arg2* as marker for enterocytes. Further, the goblet cell marker *Muc2* shows moderate expression levels in Paneth cells and explains the increased goblet cell score in Paneth cells earlier.

Cell type	Gene names
Enterocyte	*Arg2, Mttp, Lgals3*
Goblet cell	*Clca1, Scin, Tpsg1, Foxa1*
Paneth cell	*Defa3, Defa5, Defa17, Defa23, Defa26, AY761184*
Tuft cell	*Dclk1, Ptprc, Avil, Espn, Spib*

Table 4.3: Overview on more specific marker genes used for cell type scoring.

Protein (Mttp) more specific for enterocytes and a knockout of *Mttp* severely impairs the lipid transport in the intestine [Liang et al., 2014].

The differentially expressed genes of goblet cells vs ISCs are mostly also present in Paneth cells, but on a lower level. Both cell types produce a large amount of secretory vesicles and even though Paneth cells secrete mainly defensins rather than mucus forming components, Muc2 has been reported to coat Paneth cell vesicles and to facilitate their secretion [Stahl et al., 2018]. Possibly, the similarity of Paneth and goblet cells is rather explained by structural similarity (and the respective molecular machinery to constantly produce granules etc.) than common origin. Yet, we have used the goblet marker genes *Muc2*, *Spdef* and *Foxa1* to identify goblet cells by gene scoring (see table 4.1). However, we also found considerably high goblet scores in Paneth cells and vice versa. Both *Spdef* and *Muc2* are also expressed in Paneth cells. The differential expression plot further shows that *Tpsg1*, *Clca1* and *Scin* are more restricted to goblet cells and provide a clearer distinction between Paneth and goblet cells (see Fig. 4.12 and table 4.3). Similarly, we observe elevated *lysozyme 1(Lyz1)* expression levels in goblet cells, even though it is a classical Paneth cell marker. In contrast, the expression of numerous *defensins* such as *Defa3/5/17/23/26* and *AY671184* is restricted to Paneth cells. Thus, a marker gene set using *defensins* helps to discriminate Paneth and goblet cells further.
The top differentially expressed genes in EECs and tuft cells are more distinct from the other mature cell types in the small intestine. In EECs marker gene selection, we profit from the very specific hormone expression to classify this population. However, we identified novel markers such as the transcription factor *Fev*, the cell surface receptor *Gfra3* and the *insulinoma-associated 1 gene (Insm1)*. In tuft cells, known markers as *Dclk1*, *Lrmp* and *Trpm5* are among the top differentially expressed genes. This study contains more than 2,000 mature tuft cells and tuft progenitors, which allows us to propose new markers such as *advillin (Avil)*[1] and *espin (Espn)*, which are likely both involved in actin binding and the formation of the brush-like villi in a tuft cell.

[1]We included *Avil* already for gene scoring before we re-evaluated marker gene specificity.

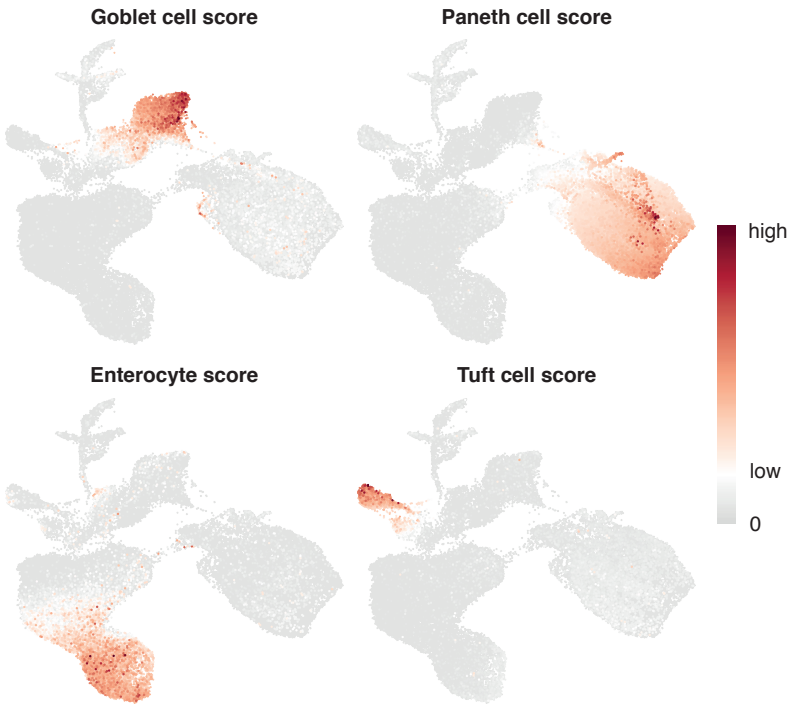

Figure 4.12: A more specific gene score improves the distinction in similar cell populations. Based on the differential expression analysis, we identified more specific gene sets to distinguish enterocytes, Paneth cells, goblet cells and tuft cells (see tab. 4.3). Especially goblet and Paneth cells are clearly distinct according to the chosen marker sets.

In the last step, we investigated the progenitor populations to identify genes that define a distinct progenitor population and thus might be involved in intestinal lineage determination (Fig. 4.13a). For ISC marker genes, we observe an intermediate expression level for all progenitor populations regardless of the lineage. One exception is *Sox4*, where we find the highest levels in the EE and tuft progenitors. In enterocyte progenitors, the expression of *Arg2* is already significantly higher compared to ISCs ($FDR < 0.05$). Further, *Arg2* is negligible in other cell types than enterocytes. The corresponding enzyme arginase 2 hydrolyses arginine to ornithine and urea — these metabolites play a role, amongst others, in the activation of both innate and adaptive immunity [Kim et al., 2018]. Thus, enterocyte may contribute to the communication of immune system and the small intestine.

The goblet cell populations are divided into three sub-populations, where we may identify increasing gene expression transitions from progenitors to mature cells. Notably, the lineage determining transcription factor *Spdef* has highest expression levels in early goblet cells. While *Muc2* is highest in mature goblet cells, progenitor and early goblet cells

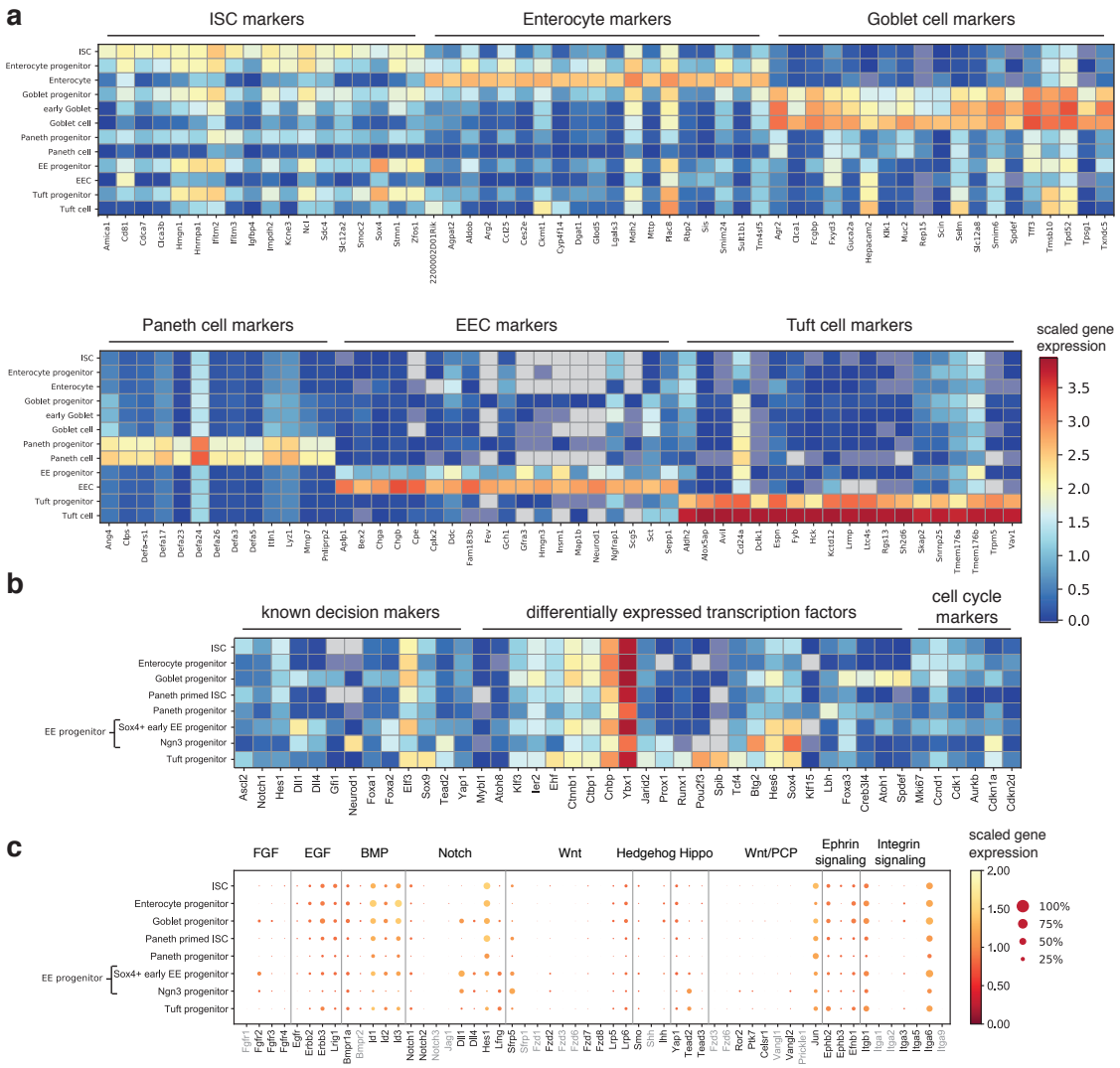

Figure 4.13: Identification of marker genes for intestinal progenitor populations. **a** Mean expression of top differentially expressed genes (displayed in Fig. 4.11) in progenitor and mature cell populations. Distinct marker genes are already expressed in progenitor cells. **b** Differentially expressed transcription factors which determine the respective lineages. **c** Signaling pathways involved in stem cell maintenance, differentiation and cell movement. Genes in grey show no significant expression differences in any of the tested groups.

express only little more compared to Paneth cells. The same holds true for other marker genes like *Guca2a* and *Tff3*. Moreover, *Fcgbp* is both highly expressed in goblet cells and increases during maturation, but we also observe lower levels of *Fcgbp* in the *Sox4+* early EE progenitor.

Paneth cell markers are almost equally high in the Paneth progenitor population. How-

ever, when we compare ISC marker expression in Paneth progenitors and mature cells, these factors are generally higher in the progenitor population. A specific on-switch remains to be determined.
The transcriptional profile of EECs shows little overlap with the other cell types. Exceptions are secretin *Sct* and *Ngfrap1*, which are both also lowly expressed in goblet cells. Comparing EE progenitors to EECs reveals a number of novel markers turned on already in EE progenitors like *Insm1*, *Gfra1* and *Ddc*. The latter catalyzes the decarboxylation of 5-hydroxytryptophan to serotonin. In addition, the transcription factor *Fev* increases during EEC maturation, but is completely absent in all other cell types.
The transcriptional pattern of tuft cells is as distinct from other cell types as the EEC one as 10 of the top 15 differentially expressed genes are practically absent in all other cell types. For all genes in the top 15 list, we find an increasing expression level from tuft progenitor to mature tuft cells. Largest differences can be found in the expression of *Dclk1*, *Fyb*, *Hck* or *Skap2*. Known tuft cell markers as *Dclk1*, *Lrmp* and *Trpm5* are specifically expressed in the tuft cell lineage.

In order to further investigate lineage determination in progenitor populations, we performed pairwise differential expression tests of ISCs vs respective progenitor populations (Fig. 4.13b). First, we studied the expression patterns of 13 known lineage determining markers in the progenitors. The factors *Hes1* followed by *Elf3* are responsible for enterocyte development, but *Elf3* is also involved in goblet cell differentiation together with *Klf4* [van der Flier and Clevers, 2009]. Astonishingly, the enterocyte marker *Elf3* is expressed in several progenitor populations, which effectively disproves the required lineage marker specificity on the gene level. In accordance with previous findings, *Gfi1* is only expressed in goblet, Paneth, EE and Tuft progenitors. However, the *Gfi1* level in both Paneth and Ngn3 progenitors is approximately the same, but *Gfi1* is antagonized by *Neurog3* and thus, *Gfi1* presumably plays a secondary role in Paneth cell maturation.

Using differential expression analysis, we found 362 differentially expressed transcription factors, from which we display 24 factors with lineage restricted binding pattern. Among the differentially expressed transcription factors, *Klf15* and *Lbh* potentially determine the Paneth cell lineage, where *Klf15* was already identified by Haber et al. [2017]. Though weakly expressed, the iron-regulated transcription factor *Atoh8* acts specifically in EE progenitor cells, but has only been reported in the liver [Patel et al., 2013]. The Notch pathway target gene *Hes6* was found exclusively expressed in the small intestinal crypt in adult mice [Schröder and Gossler, 2002], and is particularly highly expressed in goblet, EE and Tuft progenitors. Also in accordance with the findings of Haber et al. [2017], we have found goblet cell specific markers as *Foxa3*, *Creb3l4*, *Atoh1* and *Spdef*.

Cell cycle markers give us a notion for the proliferation and potential quiescence of the respective populations (Fig. 4.13b right side). Here, we show six different markers, four being involved in cell cycle progression (*Mki67*, *Ccnd1*, *Cdk1*, *Aurkb*) and two markers for cell cycle exit (*Cdkn1a*, *Cdkn2d*). In ISCs and enterocyte progenitors, we observe comparatively high expression of proliferation markers, with highest expression in enterocyte progenitors. Enterocytes contribute the most to the gut intestinal lining, thus, an active proliferation in the respective progenitors is required to replenish this cell type. In contrast, Paneth primed ISCs express less proliferation markers making these cells more likely a quiescent population. Cell cycle exit markers are more pronounced in the progenitor populations (tuft, enteroendocrine, Paneth and goblet progenitors) and mark the exit from proliferation into differentiation.

We investigated the expression patterns of several interacting and competing signaling pathways in the gut (Fig. 4.13c). Briefly, FGF (fibroblast growth factor) and EGF (epidermal growth factor) signaling promote proliferation [Danopoulos et al., 2017]. Additionally, EGF blocks apoptosis [Suzuki et al., 2010]. BMP signaling limits proliferation in *Lgr5*$^+$ stem cells independent from Wnt/β-catenin signaling [Qi et al., 2016]. Wnt/β-catenin signaling ensures stem cell maintenance in the crypt. Notch signaling creates spatial patterning and controls fate decisions of adjacent cells [Richmond and Breault, 2010]. Hedgehog signaling is involved in crypt spacing [Richmond and Breault, 2010]. Hippo signaling is implicated in the regulation of proliferation, stem cell maintenance and regeneration [Hong et al., 2016]. Wnt/PCP (planar cell polarity) is a non-canonical Wnt signaling pathway that inhibits canonical Wnt signaling, is independent from β-catenin and a candidate pathway regulating PC and EEC formation from ISCs. Ephrin signaling controls cell positioning along the crypt [Richmond and Breault, 2010], while integrin signaling controls surface adhesion and motility [Xu et al., 2014].
Several signaling pathways are known to regulate proliferation, stem cell maintenance, differentiation or cell positioning within the crypt. In general, all genes from signaling pathways are lowly expressed. With differential expression analysis, we observe high expression of BMP target genes *Id1* and *Id3* in ISCs, Paneth primed ISCs, enterocyte, goblet and Sox4+ early EE progenitors compared to Paneth, Ngn3 and tuft progenitors, which indicates active BMP regulation of proliferation. Further, FGF signaling (*i.e.* represented by *Fgfr2/3* expression) is highest in goblet and EE progenitors. EGF receptor expression (*i.e. Egfr*) and its co-factors *ErbB2/3* suggest higher susceptibility to EGF in ISCs, enterocytes and goblet progenitors. For Notch signaling, we found low expression levels for the receptor genes *Notch1-3*, while effectors like *Dll1/4*, *Hes1* and *Lfng* are comparatively high expressed. The Notch signaling effectors *Dll1/4* are mainly restricted to goblet and EE progenitors and increase Notch signaling by binding to the Notch1 receptor. Further-

more, *Hes1* and *Lfng* appear mutually exclusive in their gene expression pattern.
Regarding integrin signaling, we find *Itgb1* and *Itga6* expressed in all cells with lowest levels in Paneth progenitor cells. Integrin *Itgb1* promotes the Hedgehog pathway [Xu et al., 2014]. While Hedgehog is essential for regulation of proliferation and differentiation, we observe only scarce expression in all cell types. In addition, *Yap1* and *Tead2/3* are transcriptional activators in the Hippo pathway involved in proliferation and cell survival [Hong et al., 2016]. Single-cell RNA-seq shows a constitutive *Yap1* expression while *Tead2* is restricted to EE and tuft progenitors.
In accordance to prior studies, *EphB3* expression is high in ISCs and Paneth progenitors. Especially for Paneth cells, EphB3 allows for residing in the intestinal crypt [Richmond and Breault, 2010]. In contrast, mostly ISCs, enterocyte and goblet progenitors express high levels of *Ephb2*, which is directly controlled by canonical Wnt signaling. Tracing and distinguishing canonical and non-canonical Wnt signaling on the expression level turned out to be difficult due to expression levels on the verge of the sensitivity limit. For the frizzled receptor *Fzd1/2/3/5/7/8*, we find only none to little expression, while the co-receptors *Lrp5/6* are constitutively expressed in all cell populations. The secreted frizzled related protein 5 (*Sfrp5* gene) is known to regulate canonical Wnt signaling via interaction with frizzled receptors and Wnt binding and shows highest expression in EE progenitors. High levels of *Sfrp5* suggest an inhibition of canonical Wnt signaling in EE progenitors. Accordingly, we observe the highest levels of Wnt/PCP genes (*i.e. Ror2*, *Ptk7* and *Vangl*) in the Ngn3 positive EE progenitor indicating active Wnt/PCP signaling. In Paneth primed ISCs and Paneth progenitors, we observe higher levels of the Wnt/PCP effector *Jun*, which implies active Wnt/PCP signaling as well in the Paneth cell lineage.

4.3.6 Disturbed Wnt/PCP signaling reduces Paneth cell formation

In the previous sections, we have analyzed the relations of emerging cell populations in the small intestinal epithelium in great detail. Furthermore, we have studied the role of *Fltp* as lineage determining marker for Paneth cells and for enteroendocrine cells.
In order to elucidate the role of Wnt/PCP signaling in intestinal lineage allocation we compared control mice with the heterozygous Celsr1$^{\mathrm{crsh}/+}$ mutation strain and compound mutant mice: Celsr1$^{\mathrm{crsh}/+}$; *Fltp*$^{ZV/+}$ and Celsr1$^{\mathrm{crsh}/+}$; *Fltp*$^{ZV/ZV}$ (see Fig. 4.14 and method section 4.1). Celsr1 is a Wnt/PCP core protein und homozygous deletion or *crsh* mutations result in lethality in the neonatal state. In a qRT-PCR experiment on crypt cells, we determined *Celsr1* and *Fltp* relative gene expression and observed a significant decrease in both genes (Fig. 4.14a). Notably, the knock-out of *Fltp* is connected with a reduction of *Celsr1* expression. Furthermore, the expression levels of key Paneth cell genes as *Defa5* is halved in both Celsr1$^{\mathrm{crsh}/+}$ and Celsr1$^{\mathrm{crsh}/+}$; *Fltp*$^{ZV/+}$ compared to control mice

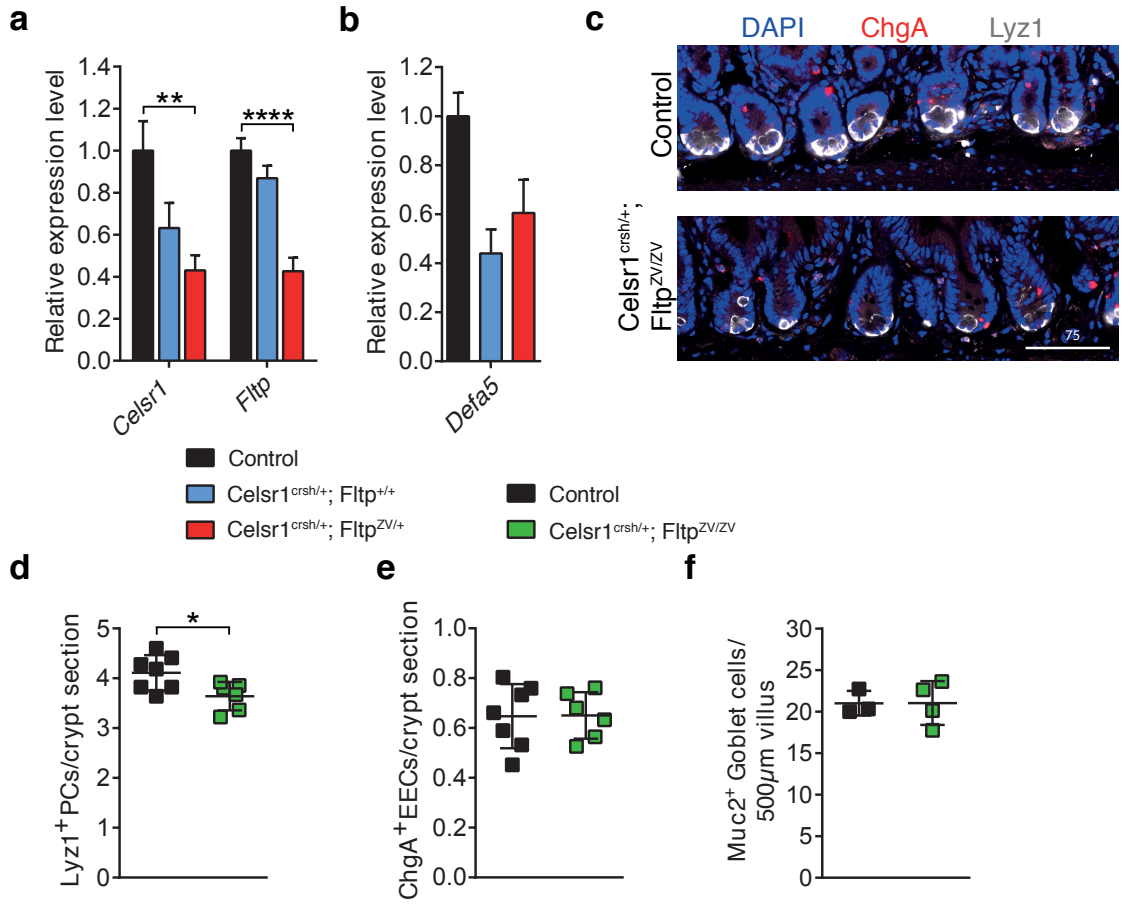

Figure 4.14: Disturbed Wnt/PCP signaling impacts Paneth cell formation. **a** Relative expression level of *Celsr1* and *Fltp* in crypt cells isolated from two month old control mice, Celsr1$^{crsh/+}$ single mutant and Celsr1$^{crsh/+}$; Fltp$^{ZV/+}$ compound mutant mice assessed by qRT-PCR. $n = 6$ control mice, $n = 5$ Celsr1$^{crsh/+}$ mice, and $n = 9$ Celsr1$^{crsh/+}$; Fltp$^{ZV/+}$ mice. Error bars represent SEM. Two-tailed Student's t-test, $^{**}P < 0.01$, $^{****}P < 0.0001$. **b** Relative expression levels of key Paneth cell gene *Defa5* is lower in crypt cells of Celsr1$^{crsh/+}$ mutant mice and compound Celsr1$^{crsh/+}$; Fltp$^{ZV/+}$ mutant mice. $n = 6$ control mice, $n = 5$ Celsr1$^{crsh/+}$ mice, and $n = 9$ Celsr1$^{crsh/+}$; Fltp$^{ZV/+}$ mice. Error bars represent SEM. Two-tailed Student's t-test, $^{*}P < 0.05$, $^{**}P < 0.01$, $^{****}P < 0.0001$. **c** Representative LSM images of duodenal sections from control and Celsr1$^{crsh/+}$; Fltp$^{ZV/ZV}$ mice stained for DAPI (blue), ChgA (red, EECs) and Lyz1 (white, Paneth cells). Scale bar, $75\,\mu m$. **d** Quantification of Lyz1-positive cells (Paneth cells) in the duodenum of control and Celsr1$^{crsh/+}$; Fltp$^{ZV/ZV}$ mice. $n = 7$ control mice and $n = 7$ Celsr1$^{crsh/+}$; Fltp$^{ZV/ZV}$ mice. Error bars represent SD. Two-tailed Student's t-test, $^{*}P < 0.05$. **e** Quantification of ChgA-positive cells (EECs) in the duodenum of control and Celsr1$^{crsh/+}$; Fltp$^{ZV/ZV}$ mice. $n = 7$ control mice and $n = 6$ Celsr1$^{crsh/+}$; Fltp$^{ZV/ZV}$ mice. Error bars represent SD. No significant difference. **f** Quantification of Muc2-positive cells (goblet cells) in the duodenum of control and Celsr1$^{crsh/+}$; Fltp$^{ZV/ZV}$ mice. $n = 3$ control mice and $n = 4$ Celsr1$^{crsh/+}$; Fltp$^{ZV/ZV}$ mice. Error bars represent SD. No significant difference.

(Fig. 4.14b). Thus, disturbed Wnt/PCP signaling induces a reduction in Paneth cell marker gene expression.
Next, we studied duodenal sections to study the effects of the *Celsr1* mutation on the cellular level (Figs. 4.14c-f). Intestinal crypts of Celsr1$^{crsh/+}$; *Fltp*$^{ZV/ZV}$ mutants have Paneth cells at the crypt bottom, but they express less Lyz1 compared to control mice (Fig. 4.14c). Image based quantification reveals a significant reduction in the number of Paneth cells in the Celsr1$^{crsh/+}$; *Fltp*$^{ZV/ZV}$, while we could not detect a change in Chga+ EECs and Muc2+ Goblet cells (Figs. 4.14d-f). Taken together, Paneth cell formation is strongly dependent on Wnt/PCP signaling while EECs and Goblet cells are not affected by disturbed Wnt/PCP signaling.

4.3.7 Disturbed Wnt/PCP signaling affects transcriptional activity

To study the molecular mechanisms and transcriptional changes caused by decreased Wnt/PCP signaling, we sequenced the small intestines of Celsr1$^{crsh/+}$; *Fltp*$^{ZV/ZV}$ mice without enrichment and another two samples with 10 % FVR enrichment yielding 14,527 cells in addition to 59,964 control sample cells.

First, we assigned all mutant cells to the respective cell type by a k-nearest neighbor approach (see methods section 4.1). We could assign all but 57 cells to the respective cell types in control samples (see Fig. 4.15a,b). Closer examination of the unassigned cells showed they were an outlying cluster with Paneth cell signature and we excluded them from further analysis. Hierarchical clustering of all cell types grouped by mutant and control condition clusters always the cell types together rather than the condition (see Fig. 4.15c). Furthermore, hierarchical clustering largely recovers the relations between cell types as shown for the control condition alone (see Fig. 4.6c). We only observe a slight decrease in correlation in the Paneth progenitor population. Moreover, mutant Paneth progenitor and Paneth cell population show a higher correlation in gene expression than Paneth progenitors in mutant and control samples. Further, Paneth cells in both control and mutant condition are highly correlated. Next, we studied the cell cycle state of all cell types (see Fig. 4.15d). We observe a similar proliferation status for all cell types in both mutant and control, except for Paneth progenitors which appear less proliferative in the mutant.
Taken together, we conclude that the disturbed Wnt/PCP signaling affects most Paneth progenitors, but we expect little changes in the gene expression profiles because we were able to cluster all mutant samples together with the control samples without further adjustments apart from batch correction.

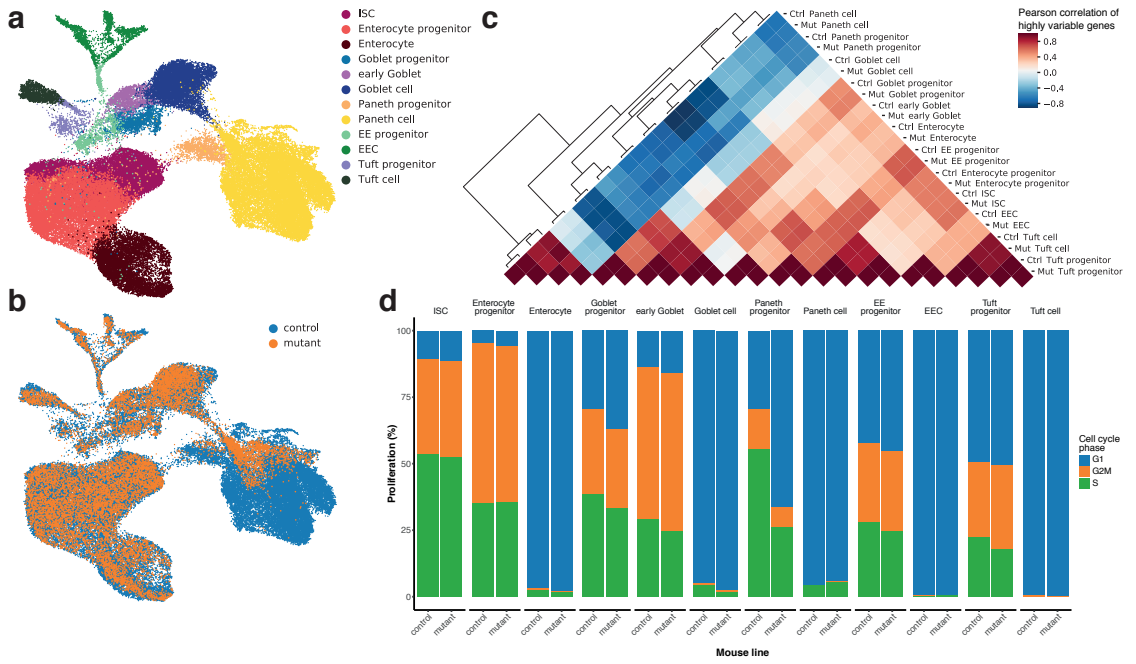

Figure 4.15: Mutant gene expression largely overlaps with control group. a-b UMAP plot annotated by cell type (**a**) integrating both control and mutant samples (**b**). **c** Pearson correlation of the highly variable genes in all cell types present in both mutant and control samples. In all cases, mutant and control samples of the respective cell type correlate the most. **d** Cell cycle annotation for all cell types into G1 (non-cycling), S and G2M (proliferating) phase. Paneth progenitors in mutant samples are less proliferative than in control samples. All other cell types are not affected in terms of proliferation.

In order to compare the gene expression patterns in control and mutant, we performed differential expression tests for each cell type on the raw data before batch correction and screened for enriched gene sets (see method section 4.1). We used both gene ontology (GO) and KEGG terms, but most terms attributed to metabolic processes or replication.

Hence, instead of interpreting gene sets, we focussed on transcription factors or genes with DNA binding ability (see method section 4.1) and signaling pathways (see Fig. 4.16). For simplicity, we refer to all genes with DNA binding activity as transcription factors. We tested ISCs and all progenitor populations in mutant and control. Among the differentially expressed genes, we found 114 transcription factors. Interestingly, all but one (*E2f1*) are also differential between ISCs and progenitors populations in the control condition. Thus, the altered Wnt/PCP signaling leads to distinctively shifted transcription factor activation. More specifically, 79 transcription factors are differentially expressed in Paneth primed ISCs or Paneth progenitors, while only one gene (*Hmga1*) is differential in the

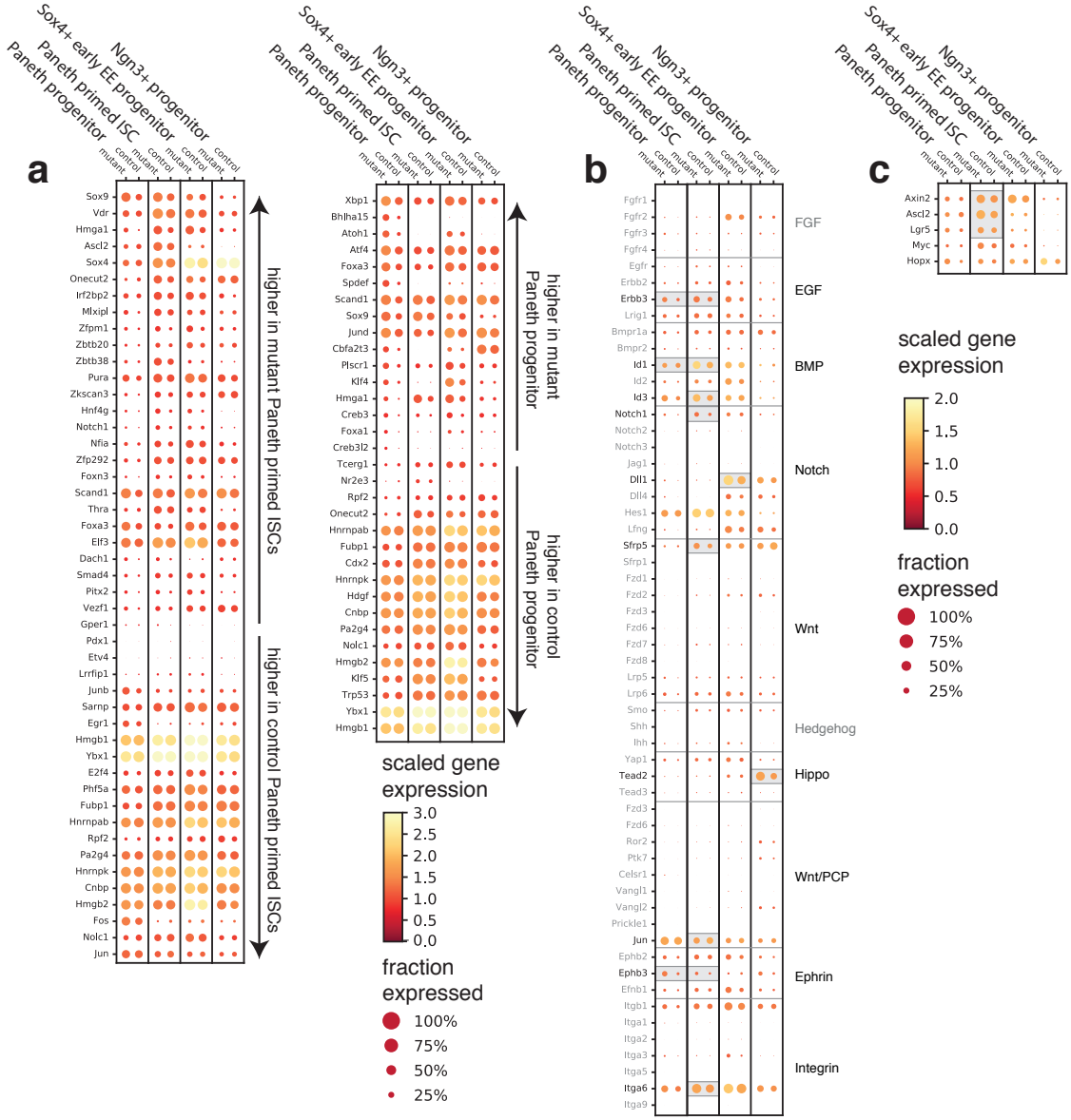

Figure 4.16: Disturbed Wnt/PCP signaling results in aberrant gene expression of canonical Wnt and Wnt/PCP target genes. a Transcription factors differential in mutant and between displayed progenitor groups, ordered by log-fold change in Paneth primed ISCs (left panel) and in Paneth progenitors (right panel). In the EE progenitor regime, only *Hmga1* was significantly different in Sox4+ early EE progenitors and none was significantly different in Ngn3 progenitors. *b* Differential gene expression in specific signaling pathways. Gene names in grey had no significant differences, grey boxes in the dotplot highlight significantly different genes. **c** Differential expression in canonical Wnt signaling targets. Grey boxes in the dotplot indicate significant changes.

Sox4+ early EE progenitor population and none in Ngn3 progenitors. We observe low expression levels of several Wnt/PCP signaling components such as *Ror2* and *Vangl2* in the EE progenitor lineage, which presumably compensate for the loss in *Celsr1*. In conclusion, Wnt/PCP signaling plays a critical role in determining the enteroendocrine and Paneth cell lineage. In particular, Celsr1 is more involved in the regulation of the Paneth cell fate.

In Paneth primed ISCs and Paneth progenitors, we found evidence for both altered canonical Wnt and non-canonical Wnt/PCP signaling (see Fig. 4.16a). In particular, mutant Paneth primed ISCs expressed higher levels of *Sox9*, a canonical Wnt target, which is required for differentiation in Paneth cells [Mori Akiyama et al., 2007] and partially represses *Muc2*, a typical goblet cell marker [Blache et al., 2004]. However, Sox9 and Klf4 together interact with β-catenin to reduce binding activity to Tfc4 and thus reducing canonical Wnt signaling activity [Sellak et al., 2012]. Moreover, the elevated expression of the vitamin D3 receptor gene *Vdr*, which reduces Wnt/β-catenin signaling [Larriba et al., 2011], indicate reduced canonical Wnt signaling in the mutant. In contrast, canonical Wnt activating genes like *Ascl2*, *Sox4* and *Elf3* show significantly elevated gene expression in mutants that apparently balanced the aforementioned reduction in canonical Wnt signaling. Moreover, the Notch receptor gene *Notch1* is a canonical Wnt target gene and increased in mutants, which indicates a reduction in EECs and goblet cells [Fre et al., 2005].
When we investigate the significantly highly expressed genes in the control samples of Paneth primed ISCs, we find *Jun*, *Junb* and *Fos*, members of the same transcription factor family and cell growth stimulating factors. What is more, *Jun* is a non-canonical Wnt/PCP target gene and its downregulation indicates impaired non-canonical Wnt/PCP signal transduction (see Fig. 4.16b). However, BMP signaling target genes *Id1/3* show higher expression levels in Paneth primed ISCs, and hint on decreased proliferation, but the EGF receptor gene *Erbb3* is also increased, blocking apoptosis. As a remark, canonical Wnt signaling targets (*Axin2*, *Ascl2* and *Lgr5*) are significantly increased in Paneth primed ISC mutants (see Fig. 4.16c).
Taken together, we disturbances of canonical Wnt signaling, both in activation and repression. Likewise, non-canonical Wnt/PCP signaling is likely reduced, which implies a reduced Paneth cell fate determination in Paneth primed ISCs. Concerning proliferation, 17 % of the control Paneth primed ISCs are in G1 phase compared to 21 % of mutant Paneth primed ISCs. Thus, the disturbed Wnt/PCP signaling creates a complex transcriptional pattern and indicates redundancy in the signaling pathways such that yet unidentified effectors jump in to balance the loss of *Celsr1*.

In Paneth progenitor cells, we also observe increased *Sox9* expression in the mutant condition, but in addition, we also observe significantly higher levels of *Klf4*, which is known to propel goblet cell fate. However, both factors together antagonize β-catenin [Sellak et al., 2012] and, moreover, *Klf4* activates non-canonical Wnt signaling by targeting Wnt5a [Tetreault et al., 2016]. Furthermore, *Atf4* as non-canonical Wnt target gene [Voloshanenko et al., 2018] is increased, which implies active non-canonical Wnt/PCP signaling. In contrast, the effector of the Notch signaling pathway *Atoh1* and its target *Spdef* (transcriptional activator and goblet cell marker) [Lo et al., 2017] are significantly increased in mutant Paneth progenitors. However, the expression of the *Atoh1* regulator *Hes1*, remains unchanged in mutant and control (see Fig. 4.16b). *Cbfa2t3* acts as transcriptional co-repressor and represses *Mmp7* (Paneth cell marker gene) expression. Interestingly, the intestine specific transcription factor *Cdx2* [Guo et al., 2009] is reduced in mutant Paneth progenitors. The transcription of *Cdx2* is activated by Foxa3 [Silberg et al., 2002], but repressed by Sox9 [Blache et al., 2004]. Previously, *Cdx2*-negative intestinal crypts were vacant of Paneth cells [Stringer et al., 2012]. Thus, as *Cdx2* is essential in the development and survival of Paneth progenitors and its decrease in mutant Paneth progenitors is in line with decreased proliferation.

In summary, the role of *Fltp* and *Celsr1* as part of the non-canonical Wnt/PCP signaling cascade in the small intestine is difficult to read from transcription patterns. Disturbed Wnt/PCP signaling affects strongly transcriptional activity and has an essential role on the formation of Paneth cells, possibly by balancing the effects of *Foxa3* and *Sox9* on *Cdx2*.

4.4 Discussion

The study of transcriptomic profiles of the intestinal epithelium at the single-cell level reveals the overwhelming complexity of this tissue. In order to account for complexity and compositional changes in the data, we modified the ComBat model to correct for batch effects. The modification allowed accurate clustering into cell types rather than clustering by samples. Furthermore, it reduced the potential for false positive discovery of novel cell types.

In principal, our study recovered all major cell types and annotated via clustering and marker gene expression. In contrast to earlier studies [Grün et al., 2015, Haber et al., 2017, Yan et al., 2017], we do not classify transit amplifying (TA) cells for several reasons. First, the term is misleading as all differentiating cells but Paneth cells migrate upwards

from the crypt towards the villus and pass through a 'transit amplifying' zone. Second, most studies describe TA cells actually as enterocyte progenitors and characterize them by high cell cycle activity and their adjacency to ISCs [Grün et al., 2015, Haber et al., 2017]. Third, we and others (*e.g.* van der Flier and Clevers [2009]) have found no evidence for multipotent transit amplifying cells as proposed in Yan et al. [2017]. Thus, closest correspondence for TA cells is the enterocyte progenitor.
Our sampling strategy with enrichment allowed describing all progenitor cell types, especially Paneth progenitors, which have not been characterized using transcriptomic profiling. Paneth progenitors have been shown to exhibit both stem cell character (e.g. *Lgr5* expression) and the onset of Paneth cell gene expression. In particular, we identified Paneth cell specific transcription factors in the early lineage specification process, which needs to be explored further. The disadvantage of the enrichment strategy is a potential distortion of the data manifold. Weinreb et al. [2018] reported unbiased sampling as a major limitation for lineage inference methods that rely on the cell density. Here, we massively enriched for Paneth cells to characterize progenitors, which would go unnoticed otherwise. Thus, enrichment for specific rare cell types comes at the cost of limited gene expression dynamics inference.

For EECs, we recovered clear subsets and described several progenitor populations, mainly in accordance with previously described results [Haber et al., 2017]. However, it has been speculated on the nature and plasticity of the respective subclasses of EECs. For example, it is not known, if EECs produce the same set of hormones throughout their short life time or if they can switch on and off various combinations of hormone related genes. To support the hypothesis of a cell fate decision into several subclasses of EECs, we would require a time-series experiment such as molecular lineage barcoding with *CRISPR/Cas9* [Spanjaard et al., 2018], which has not been established in a mouse model, yet.

A major advantage of scRNA-seq is the availability of gene expression information for almost all genes in every cell. This allows us to evaluate the specificity of marker genes and to identify more cell-type specific markers on the gene expression level. For example, we discovered that goblet cell specific markers as *SAM Pointed Domain Containing ETS Transcription Factor (Spdef)* and *mucin 2 (Muc2)* are also lowly expressed in Paneth cells. Thus, we proposed alternative markers such as *scinderin (Scin)*, which appears highly specific to goblet cells, but its role has only been studied in airway goblet cells [Ehre et al., 2005].

When it comes to study the impact of signaling pathways, scRNA-seq is only of limited use for several reasons. First, the expression level of almost all signaling associated genes

is low and almost untraceable with the drop-seq technology, which has high-throughput as advantage nonetheless — in turn, a highly sensitive scRNA-seq approach like SMARTseq2 can only measure a few hundred cells and thus, we require a thorough sorting strategy. Second, cellular signaling is carried out on the protein level, and the transcriptomic readout may correlate poorly with protein expression depending on e.g. mRNA half-life and protein localization [Paul et al., 2015, Schulz et al., 2018]. Third, we cannot derive protein-protein interactions from scRNA-seq. As such, we aim to reconstruct a mechanistic model of all signaling cascades and respective crosstalk from a snapshot of a different type of omics data. Thus, the insight on Wnt/β-catenin and Wnt/PCP signaling only helps to generate new hypotheses, but a final proof has to be administered with a different type of experiment. In terms of hypothesis generation, we observed that the Paneth cell formation is affected most by disturbed Wnt/PCP signaling based on transcriptomic profiling and histological images. Moreover, we observed changes in both canonical and non-canonical Wnt signaling, potentially as a result of complex interactions with redundant pathways, which balance the loss in *Celsr1*.

The lineage tree of the intestinal epithelium remains a complex issue. By combining short-term lineage tracing and scRNA-seq, we effectively ruled out a bipotent Paneth-goblet cell progenitor, which was hypothesized before [Grün et al., 2016]. Instead, we sought evidence for a bipotent Paneth-enteroendocrine progenitor as we observed transient Wnt/PCP signaling in both Paneth cell and EEC lineages with the Fltp-Venus reporter (FVR). With scRNA-seq, we found evidence for direct lineage allocation via primed stem cells. Moreover, EECs were not affected in the *Celsr1* mutants, indicating a different mechanism for either Wnt/PCP signaling or *Fltp* activation in this lineage.

Finally, data from scRNA-seq is not confounded with changes in cell proportions as we observe this in bulk samples. Therefore, scRNA-seq data allow testing for compositional changes across cell types. In order to model cell type compositional changes, we have to account for the inherent anti-correlation in such data, *i.e.* when one fraction rises, another fraction of cells decreases due to the unit-sum constraint [Aitchison, 1986]. While the field of scRNA-seq has not been confronted with differential testing of compositional data, microbiome research has proposed to use Dirichlet, Multinomial and Gamma-Poisson models to quantify changes in composition [Weiss et al., 2017].
When we aim to study changes in cell proportions due to genetic variation (*i.e.* as presented with *Celsr1* mutants and controls), we have have to control if the cells have been sampled in an unbiased manner prior to sequencing. In our case, enrichment for FVR and FVF via cell sorting is a classical bias, which we needed to sufficiently sample from certain (progenitor) populations. Furthermore, Haber et al. [2017] observed that large cells

as Paneth cells are less abundant in the data set due to sample preparation. Similarly, studying the small intestinal epithelium in mice involved sacrificing an individual for each measurement, such that cell population sizes are essentially confounded with individual variation. In order to deal with the inter-individual changes in cell population sizes, we have to establish an unbiased sampling approach for cell sorting, and obtain independent data from more individuals in both control and mutant (*i.e.* at least 6 mice per group; a detailed analysis of the effective sample size was not part of the project).

In conclusion, this study gave new insights into the lineage hierarchy of the small intestinal epithelium and presented a thorough characterization of all cell types on a highly-resolved level complementing pre-existing studies. Furthermore, we shed light on the contribution of canonical and non-canonical Wnt signaling in proliferation and differentiation of the secretory lineage and identified several regulatory candidates whose involvement in the cell fate of several cell types is to be determined in future experiments.

Chapter 5

Summary and Outlook

Ich habe das symbolische Rechnen mit Stumpf und Stil verlernt.

Emmy Noether

On ne fait jamais attention à ce qui a été fait; on ne voit que ce qui reste à faire.

Marie Curie

5.1 Summary

In this thesis, we presented the widespread biological applications and methodological implications of single-cell RNA-sequencing with particular focus on data analysis. As scRNA-seq data represent a novel data type, numerous contributions in the scientific community have discussed appropriate modelling and simulation approaches to account for all properties found in a dataset. We started with general approaches in single cell biology and outlined biological systems of our interest. The methods section (Chapter 2) described the typical workflow of a single-cell RNA-seq analysis including numerous options to address the variable properties of a dataset. In particular, we reviewed the most commonly used normalization methods and batch correction models to remove cell specific bias and unwanted experimental bias.

The scientific contribution of this thesis is twofold. First, we designed a new test metric called kBET to assess batch effects in scRNA-seq data in a quantitative manner (Chapter 3). Furthermore, we employed kBET to compare the most commonly used normalization methods and batch correction models, which we introduced in the methods section,

and we studied kBET's applicability to assess inter-individual variation in healthy donors. The second contribution is a thorough analysis to infer a lineage hierarchy in the small intestinal epithelium (Chapter 4). Here, we demonstrated an extensive scRNA-seq data integration and analysis combined with experimental results from microscopy. However, we had to admit that the we could not fully succeed to reconstruct a lineage hierarchy from transcriptional profiling alone, but we gained valuable insights into cell type characteristics from an exceptional amount of cells. Moreover, we described all progenitor cell types in the tissue, which was not done before in this detail.

5.2 Outlook

Extend batch effect quantification

We have studied batch effects, their quantification and ways for correction extensively in Chapter 3. As pointed out, kBET is a very sensitive tool which allows comparing several batch correction strategies. However, the downside of the high sensitivity is a fast saturation at the upper end of the scale. Especially in tissue data sets with many clusters, a cluster of one cell type appears to have small patches of cells from the same batch, but kBET assigns them high rejection rates due to the small scale patches within a cluster. The case of the patchwork cluster implies that the accuracy in batch correction does not necessarily meet the criteria of kBET in order to reduce the batch effect sufficiently. Thus, we have identified different classes of batch effect removal depending on the analysis task. One is accurate clustering such that the clustering reflects cell types and none of the clusters emerges due to batch effects (*i.e.* patchwork clusters). The characterization of the clusters is then performed against other clusters or the remaining data set via differential expression tests, where we recommend to use normalized data instead of batch corrected data, and to include the batch covariate in the model matrix. In such a case, the less sensitive measures based on PCA or the silhouette coefficient suffice as well to evaluate the batch correction. The second case is a trajectory inference from several replicates. In this case, we cannot define a cell type by clustering, as a trajectory describes the transition from one cell type to another. Thus, we have to ensure that the trajectory is inferred with as little batch effect as possible and here kBET is exactly the right tool to evaluate the contribution of the batch effect to the data.
In order to extend kBET, we have to scrutinize the null hypothesis. One of the underlying assumptions in kBET is that all tests are independent of each other, but a neighborhood where one population is overrepresented implies at least a second neighborhood where

exactly the same population is underrepresented. Those intrinsic frequency shift should be explored and be accounted for in an extended kBET model.

Reduce ambient mRNAs for better single-cell data analysis

In Chapter 4, we studied the transcriptomic profiles of the small intestinal epithelium. Even though we followed the most recent best practices for scRNA-seq analysis and evaluated several strategies ourselves, we found a few issues worth improving in the future. First of all, we observed ambient mRNA in several samples, which is essentially a background signal from burst cells with high expression levels of certain genes, especially when the cells are prepared in suspension as in droplet-based approaches. To date, we only examined its presence by visual inspection, and is difficult to be removed. One pre-print [Young and Behjati, 2018] estimated the background signal from empty droplets, which appears promising. Furthermore, ambient mRNA contributes to the batch effect. In the presented data, we have computed kBET to estimate the batch effect correction success with the modified ComBat version, but we found always the highest rejection rates for all major cell types after batch correction. In Chapter 3, we have shown that kBET is also sensitive to changes in the fraction of zeros and the presence or absence of ambient mRNA cannot be corrected for with the common batch effect correction schemes — especially after we decided to fix the number of zeros to preserve the non-negative support of the data set after batch correction (see section 4.1). In this case, we showed that the sample composition in every cluster is balanced. Therefore, a suitable clustering based on marker gene expression of the respective cell types is sufficient to exclude batch effect bias.

Estimate gene expression velocity

We have provided an overview on trajectory inference in differentiation processes and lineage decisions in the methods and data analysis part of this thesis (see Chapters 2 and 4). However, all trajectory inference and pseudotime methods cannot infer a developmental direction — defining a start and end point of a differentiation process is based on prior knowledge of the studied biological system. Recently, La Manno et al. [2018] distinguished unspliced (immature) and spliced (mature) mRNAs from RNA-sequencing data and developed a deterministic model to infer the onset or shutdown of gene expression in a population. Briefly, we suppose the rates of transcription and splicing are constant. Then, if we observe relatively many unspliced mRNAs, transcription of the respective gene has been turned on and if we observe relatively few unspliced mRNAs, transcription of the respective gene has been turned off. Based on this model, we can derive a complete vector field and an asymmetric Markov model, whose analysis immediately returns the starting

population (*i.e.* sources) and the end population (*i.e.* sinks of the vector field).
One of the drawbacks in estimating velocities is to determine a single direction out of all expressed genes. In particular, as in Chapter 4, we are interested in lineage inference and combining gene velocities with tree inference methods as PAGA [Wolf et al., 2018b] will indeed provide useful additional information on the true lineage hierarchy in the small intestinal epithelium. Furthermore, transient gene expression is of particular interest in differentiation processes and given an accurate trajectory inference based on the vector field could provide seminal insights to the regulatory process and drivers of cell fate decision. However, these analysis endeavors rely on data prone to technical perturbations and it is yet unclear if the estimation of velocities is impaired by batch effects and if we can adapt a batch correction method to the low read numbers of unspliced reads as well without introducing noise. Addressing these questions carries the promise to largely extend the impact and applications of gene expression velocity and is a novel context to be explored.

Test for changes in cell type composition

We briefly discussed changes in cell composition across conditions and individuals in Chapters 3 and 4. The advantage of scRNA-seq data over bulk measurements is that single-cell data are not confounded with proportional shifts across samples. Thus, scRNA-seq data allow testing for these compositional changes. As already mentioned, compositional changes have an intrinsic anti-correlation structure due to the unit-sum constraint [Aitchison, 1986]. Moreover, microbiome research already quantified changes in compositional data by e.g. Dirichlet, Multinomial and Gamma-Poisson models [Weiss et al., 2017]. However, the quantifying compositional changes in scRNA-seq is challenging due to confounding factors such as natural variability across individuals and sampling bias.
With the increasing availability of scRNA-seq data and the decreasing costs per sample, we are able to collect more samples per experiment, which is essential to account for additional variability through confounding factors. Here, the availability of scRNA-seq data was the limiting factor for testing compositional changes across conditions. Nonetheless, changes in cell type proportions describes relevant disease phenotypes such as diabetes, blood disorders and diseases of the immune system, and scRNA-seq offers a new window towards numerous diseases.

Perturb Wnt/PCP signalling in the small intestinal epithelium

The analysis of *Celsr1*$^{crsh/+}$ mutant mice revealed several disadvantages. First, homozygous missense mutation in the *Celsr1* coding region is lethal in the neonatal state, which is

why we used the heterozygous mouse line instead. Even though we observed reduced *Fltp* and *Celsr1* gene expression, it is unclear how the loss of *Celsr1* is compensated. Here, a conditional knock-out of *Celsr1* potentially provides informative insights on the role of Celsr1 in lineage commitment of Paneth cells.
The scRNA-seq mutant analysis also revealed a high number of affected genes, which are directly or indirectly associated with canonical Wnt signaling and Wnt/PCP signalling, respectively. The role and interaction partners of these genes is not fully understood and genes such as *Zfos1* are still uncharacterized. Due to incomplete annotations, we suspect that gene set enrichment analysis did not yield reliable results. Furthermore, the manual inspection of the differentially expressed genes between mutant and control condition showed that Wnt signaling is tightly regulated with various interacting partners. Therefore, testing all relevant genes in a (conditional) knock-out mouse line is practically infeasible and we have to explore new screening models to explore cause, interaction and effect of genetic perturbation in the small intestinal epithelium.

Integrate multi-omics data

The focus of this thesis was on single-cell RNA-sequencing data analysis. As pointed out in the introduction (see section 1.2), single cells experiments access different omics-layers, such as genome, proteome and chromatin organization. The development of other single-cell omics technologies is lagging behind the scRNA-seq field. However, obtaining so-called multi-omics data covering e.g. proteome and transcriptome, or transcriptome and chromatin organization of the same cell, will provide insights to gene regulation, cell fate decision making, and reorganization in de-/transdifferentiation processes.
Methods like the CITE-seq approach [Stoeckius et al., 2017] enables already simultaneous measurement of proteome and transcriptome in the same cell. Here, only few proteins can be measured through limited availability (and high costs) of antibodies. However, even with 50 proteins and a full transcriptome available per single cell, we are faced with new challenges for statistical data integration [Colomé-Tatché and Theis, 2018]. In particular, we have to connect two or more layers of information with different noise and expression characteristics [Bock et al., 2016]. Potential models to tackle multi-omics data integration are Gaussian graphical models [Do et al., 2015], which are already applied for bulk measurements. Moreover, Gaussian process latent variable models as in *MATCHER* [Welch et al., 2017] have been successfully applied for multi-omics data integration of single cells, but with omics measured in different cells. Similar potential has dynamic time warping on single-cell data as in *cellAlign* [Alpert et al., 2018] and `Seurat` [Butler et al., 2018].
Furthermore, the existing scRNA-seq analysis frameworks such as `Seurat` [Butler et al., 2018], `PAGODA` [Fan et al., 2016] and `Scanpy` [Wolf et al., 2018a] will have to extend to

integrate multi-omics data. This might be easy for proteome data, which is analogous to transcriptome data (except for post-translational modifications). In contrast, chromatin accessibility measured by scATAC-seq is organized by genomic position, where a considerable amount of reads come from distal intergenic regions such as *cis*-regulatory elements [Buenrostro et al., 2015b]. Thus, an annotation of scATAC-seq data into genes as in transcriptomics is not straightforward. All in all, leveraging the integrated information of single-cell multi-omics is a challenging future work.

Establish a standard analysis pipeline for scRNA-seq data

In Chapter 2 we presented several methods to filter, normalize, correct, cluster, order, and characterize scRNA-seq data. In Chapter 3 we compared normalization and batch correction methods with the goal to derive best practices for scRNA-seq data normalization and batch effect correction. However, we realized that different datasets required different handling, but we could identify well performing methods such as scran for normalization and ComBat for batch effect correction.
Working towards a standardized analysis pipeline, frameworks such as `Seurat` [Butler et al., 2018], `PAGODA` [Fan et al., 2016] and `Scanpy` [Wolf et al., 2018a] play an integral role. For example, the underlying data structure of `Scanpy` is similar to the *ExpressionSet* R object in Bioconductor [Huber et al., 2015]. Thus, `Scanpy` provides a customizable platform for scRNA-seq data analysis that allows the comparison of different analysis options. However, if we finally establish a standard pipeline, we have to consider our method choice. For example, it is unclear, if we should choose a best-performing method tailored specifically for a given dataset or if we are satisfied with a standard method that performs satisfactorily on a broad range of datasets. In the latter, we inevitably lose accuracy to increase comparability. In addition, every new method holds the potential to solve the issues of previous methods. Hence, establishing efficient tests for benchmarking new tools on a variety of datasets and methods will speed up method development for scRNA-seq.

Bibliography

John Aitchison. *The Statistical Analysis of Compositional Data*. Chapman and Hall Ltd, London, 1986.

Bruce Alberts, A Johnson, J Lewis, M Raff, and K Roberts Google Scholar. Molecular biology of the cell . Garland Science, 2008.

Ayelet Alpert, Lindsay S Moore, Tania Dubovik, and Shai S Shen-Orr. Alignment of single-cell trajectories to compare cellular expression dynamics. *Nature Methods*, 15: 267–270, 2018.

Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Genome Biology*, 11:R106, 2010.

Tallulah S Andrews and Martin Hemberg. Dropout-based feature selection for scRNASeq. *bioRxiv*, 2018a.

Tallulah S Andrews and Martin Hemberg. Identifying cell populations with scRNASeq. *Molecular Aspects of Medicine*, 59:114–122, 2018b.

Ilias Angelidis, Lukas M Simon, Isis E Fernandez, Maximilian Strunz, Christoph H Mayr, Flavia R Greiffo, George Tsitsiridis, Elisabeth Graf, Tim M Strom, Oliver Eickelberg, Matthias Mann, Fabian J Theis, and Herbert B Schiller. An atlas of the aging lung mapped by single cell transcriptomics and deep tissue proteomics. *bioRxiv*, 2018.

Philipp Angerer, Lukas Simon, Sophie Tritschler, F Alexander Wolf, David Fischer, and Fabian J Theis. Single cells make big data: New challenges and opportunities in transcriptomics. *Current Opinion in Systems Biology*, 4:85–91, 2017.

Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, Midori A Harris, David P Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C Matese, Joel E Richardson, Martin Ringwald, Gerald M Rubin, and Gavin Sherlock. Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25: 25–29, 2000.

Patrick Autissier, Caroline Soulas, Tricia H Burdo, and Kenneth C Williams. Evaluation of a 12-color flow cytometry panel to study lymphocyte, monocyte, and dendritic cell subsets in humans. *Cytometry A*, 77:410–419, 2010.

Rhonda Bacher, Li-Fang Chu, Ning Leng, Audrey P Gasch, James A Thomson, Ron M Stewart, Michael Newton, and Christina Kendziorski. SCnorm: robust normalization of single-cell RNA-seq data. *Nature Methods*, 14:584–586, 2017.

Jinho Baik and Jack W Silverstein. Eigenvalues of large sample covariance matrices of spiked population models. *Journal of Multivariate Analysis*, 97:1382–1408, 2006.

David Baltimore. RNA-dependent DNA Polymerase in Virions of RNA Tumour Viruses.

Nature, 226:1209–1211, 1970.

Dmitry R Bandura, Vladimir I Baranov, Olga I Ornatsky, Alexei Antonov, Robert Kinach, Xudong Lou, Serguei Pavlov, Sergey Vorobiev, John E Dick, and Scott D Tanner. Mass Cytometry: Technique for Real Time Single Cell Multitarget Immunoassay Based on Inductively Coupled Plasma Time-of-Flight Mass Spectrometry. *Analytical Chemistry*, 81:6813–6822, 2009.

Nick Barker, Johan H van Es, Jeroen Kuipers, Pekka Kujala, Maaike van den Born, Miranda Cozijnsen, Andrea Haegebarth, Jeroen Korving, Harry Begthel, Peter J Peters, and Hans Clevers. Identification of stem cells in small intestine and colon by marker gene Lgr5. *Nature*, 449:1003–1007, 2007.

Maayan Baron, Adrian Veres, Samuel L Wolock, Aubrey L Faust, Renaud Gaujoux, Amedeo Vetere, Jennifer Hyoje Ryu, Bridget K Wagner, Shai S Shen-Orr, Allon M Klein, Douglas A Melton, and Itai Yanai. A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure. *Cell Systems*, 3:1–15, 2016.

Onur Basak, Joep Beumer, Kay Wiebrands, Hiroshi Seno, Alexander van Oudenaarden, and Hans Clevers. Induced Quiescence of Lgr5+ Stem Cells in Intestinal Organoids Enables Differentiation of Hormone- Producing Enteroendocrine Cells. *Cell Stem Cell*, 20:177–190, 2017.

Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *Advances in neural information processing systems*, pages 585–591, 2002.

Mikhail Belkin and Partha Niyogi. Laplacian Eigenmaps for Dimensionality Reduction and Data Representation. *Neural Computation*, 15:1373–1396, 2003.

Sean C Bendall, Erin F Simonds, Peng Qiu, El-ad David Amir, Peter O Krutzik, Rachel Finck, Robert V Bruggner, Rachel Melamed, Angelica Trejo, Olga I Ornatsky, Robert S Balderas, Sylvia K Plevritis, Karen Sachs, Dana Peer, Scott D Tanner, and Garry P Nolan. Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. *Science*, 332(6030):687–696, 2011.

Sean C Bendall, Kara L Davis, El-ad David Amir, Michelle D Tadmor, Erin F Simonds, Tiffany J Chen, Daniel K Shenfeld, Garry P Nolan, and Dana Peer. Single-Cell Trajectory Detection Uncovers Progression and Regulatory Coordination in Human B Cell Development. *Cell*, 157:714–725, 2014.

Martin Bengtsson, Anders Ståhlberg, Patrik Rorsman, and Mikael Kubista. Gene expression profiling in single cells from the pancreatic islets of Langerhans reveals lognormal distribution of mRNA levels. *Genome Research*, 15(10):1388–1392, 2005.

Aparna Bhaduri, Tomasz J Nowakowski, Alex A Pollen, and Arnold R Kriegstein. Saturating Single-Cell Datasets. *bioRxiv*, 2017.

Fernando H Biase, Xiaoyi Cao, and Sheng Zhong. Cell fate inclination within 2-cell and 4-cell mouse embryos revealed by single-cell RNA sequencing. *Genome Research*, 24: 1787–1798, 2014.

Matthew Bjerknes and Hazel Cheng. The stem-cell zone of the small intestinal epithelium. I. Evidence from paneth cells in the adult mouse. *American Journal of Anatomy*, 160: 51–63, 1981.

Philippe Blache, Marc van de Wetering, Isabelle Duluc, Claire Domon, Philippe Berta, Jean-Noël Freund, Hans Clevers, and Philippe Jay. SOX9 is an intestine crypt transcription factor, is regulated by the Wnt pathway, and represses the CDX2 and MUC2 genes. *The Journal of Cell Biology*, 166:37–47, 2004.

Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, page P10008, 2008.

Christoph Bock, Matthias Farlik, and Nathan C Sheffield. Multi-Omics of Single Cells: Strategies and Applications. *Trends in Biotechnology*, 34(8):605–608, 2016.

Thorsten Boroviak, Remco Loos, Patrick Lombard, Junko Okahara, Rüdiger Behr, Erika Sasaki, Jennifer Nichols, Austin Smith, and Paul Bertone. Lineage-Specific Profiling Delineates the Emergence and Progression of Naive Pluripotency in Mammalian Embryogenesis. *Developmental cell*, 35:366–382, 2015.

Richard Bourgon, Robert Gentleman, and Wolfgang Huber. Independent filtering increases detection power for high-throughput experiments. *Proceedings of the National Academy of Sciences*, 107(21):9546–9551, 2010.

Nicolas L Bray, Harold Pimentel, Páll Melsted, and Lior Pachter. Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, 34:525–527, 2016.

Philip Brennecke, Simon Anders, Jong Kyoung Kim, Aleksandra A Kolodziejczyk, Xiuwei Zhang, Valentina Proserpio, Bianka Baying, Vladimir Benes, Sarah A Teichmann, John C Marioni, and Marcus G Heisler. Accounting for technical noise in single-cell RNA-seq experiments. *Nature Methods*, 10:1093–1095, 2013.

Simon J A Buczacki, Heather Ireland Zecchini, Anna M Nicholson, Roslin Russell, Louis Vermeulen, Richard Kemp, and Douglas J Winton. Intestinal label-retaining cells are secretory precursors expressing Lgr5. *Nature*, 495:65–69, 2013.

Jason D Buenrostro, P G Giresi, L C Zaba, HY Chang Nature, and William J Greenleaf. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature Methods*, 10:1213–

1218, 2013.

Jason D Buenrostro, Beijing Wu, Howard Y Chang, and William J Greenleaf. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. *Current Protocols in Molecular Biology*, 109:21.29.1–21.29.9, 2015a.

Jason D Buenrostro, Beijing Wu, Ulrike M Litzenburger, Dave Ruff, Michael L Gonzales, Michael P Snyder, Howard Y Chang, and William J Greenleaf. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature*, 523:485–490, 2015b.

Florian Buettner and Fabian J Theis. A novel approach for resolving differences in single-cell gene expression patterns from zygote to blastocyst. *Bioinformatics*, 28(18):i626–i632, 2012.

Florian Buettner, Kedar N Natarajan, F Paolo Casale, Valentina Proserpio, Antonio Scialdone, Fabian J Theis, Sarah A Teichmann, John C Marioni, and Oliver Stegle. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nature Biotechnology*, 33:155–160, 2015.

Florian Buettner, Naruemon Pratanwanich, Davis James McCarthy, John C Marioni, and Oliver Stegle. f-scLVM: scalable and versatile factor analysis for single-cell RNA-seq. *Genome Biology*, 18:212, 2017.

James H Bullard, Elizabeth Purdom, Kasper D Hansen, and Sandrine Dudoit. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics*, 11:94, 2010.

W Neal Burnette. "Western Blotting": Electrophoretic transfer of proteins from sodium dodecyl sulfate-polyacrylamide gels to unmodified nitrocellulose and radiographic detection with antibody and radioiodinated protein A. *Analytical Biochemistry*, 112(2): 195–203, 1981.

Andrew Butler, Paul Hoffman, Peter Smibert, Efthymia Papalexi, and Rahul Satija. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology*, 36(5):411–420, 2018.

Maren Büttner, Zhichao Miao, F Alexander Wolf, Sarah A Teichmann, and Fabian J Theis. A test metric for assessing single-cell RNA-seq batch correction. *Nature Methods*, 16: 43–49, 2019.

Lauren E Byrnes, Daniel M Wong, Meena Subramaniam, Nathaniel P Meyer, Caroline L Gilchrist, Sarah M Knox, Aaron D Tward, Chun J Ye, and Julie B Sneddon. Lineage dynamics of murine pancreatic development at single-cell resolution. *Nature Communications*, 9:3922, 2018.

Davide Cacchiarelli, Xiaojie Qiu, Sanjay Srivatsan, Anna Manfredi, Michael Ziller, Eliah Overbey, Antonio Grimaldi, Jonna Grimsby, Prapti Pokharel, Kenneth J Livak,

Shuqiang Li, Alexander Meissner, Tarjei S Mikkelsen, John L Rinn, and Cole Trapnell. Aligning Single-Cell Developmental and Reprogramming Trajectories Identifies Molecular Determinants of Myogenic Reprogramming Outcome. *Cell Systems*, 7(3):258–268, 2018.

Junyue Cao, Jonathan S Packer, Vijay Ramani, Darren A Cusanovich, Chau Huynh, Riza Daza, Xiaojie Qiu, Choli Lee, Scott N Furlan, Frank J Steemers, Andrew Adey, Robert H Waterston, Cole Trapnell, and Jay Shendure. Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science*, 357(6352):661–667, 2017.

Raj Chakrabarti and Clarence E Schutt. The enhancement of PCR amplification by low molecular-weight sulfones. *Gene*, 274:293–298, 2001.

Kok Hao Chen, Alistair N Boettiger, Jeffrey R Moffitt, Siyuan Wang, and Xiaowei Zhuang. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science*, 348(6233): 1–14, 2015.

Mengjie Chen and Xiang Zhou. Controlling for Confounding Effects in Single Cell RNA Sequencing Studies Using both Control and Target Genes. *Scientific Reports*, 7:13587, 2017.

R R Coifman, S Lafon, A B Lee, M Maggioni, B Nadler, F Warner, and S W Zucker. Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *Proceedings of the National Academy of Sciences*, 102(21):7426–7431, 2005.

Michael B Cole, Davide Risso, Allon Wagner, David DeTomaso, John Ngai, Elizabeth Purdom, Sandrine Dudoit, and Nir Yosef. Performance Assessment and Selection of Normalization Procedures for Single-Cell RNA-seq. *bioRxiv*, 2018.

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

The Gene Ontology Consortium. Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acid Research*, 45:D331–D338, 2016.

Jack R Cooper and Irving Melcer. The enzymatic oxidation of tryptophan to 3-hydroxytryptophan in the biosynthesis of serotonin. *Journal of Pharmacology and Experimental Therapeutics*, 132(3):265–268, 1961.

Noel Cressie and Timothy R C Read. Pearson's X 2 and the Loglikelihood Ratio Statistic G 2 : A Comparative Review. *International Statistical Review / Revue Internationale de Statistique*, 57(1):19–43, 1989.

Francis Crick. Central Dogma of Molecular Biology. *Nature*, 227:561–563, 1970.

John A Curtin, Elizabeth Quint, Vicky Tsipouri, Ruth M Arkell, Bruce Cattanach, Andrew J Copp, Deborah J Henderson, Nigel Spurr, Philip Stanier, Elizabeth M Fisher,

Patrick M Nolan, Karen P Steel, Steve D M Brown, Ian C Gray, and Jennifer N Murdoch. Mutation of Celsr1 Disrupts Planar Polarity of Inner Ear Hair Cells and Causes Severe Neural Tube Defects in the Mouse. *Current Biology*, 13:1129–1133, 2003.

Anne Czechanski, Candice Byers, Ian Greenstein, Nadine Schrode, Leah Rae Donahue, Anna-Katerina Hadjantonakis, and Laura G Reinholdt. Derivation and characterization of mouse embryonic stem cells from permissive and nonpermissive strains. *Nature protocols*, 9(3):559–574, 2014.

Soula Danopoulos, Christopher R Schlieve, Tracy C Grikscheit, and Denise Al Alam. Fibroblast Growth Factors in the Gastrointestinal Tract: Twists and Turns. *Developmental Dynamics*, 246:344–352, 2017.

Qiaolin Deng, Daniel Ramsköld, Björn Reinius, and Rickard Sandberg. Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science*, 343:193–196, 2014.

Jesse R Dixon, Siddarth Selvaraj, Feng Yue, Audrey Kim, Yan Li, Yin Shen, Ming Hu, Jun S Liu, and Bing Ren. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485:376–380, 2012.

Kieu Trinh Do, Gabi Kastenmüller, Dennis O Mook-Kanamori, Noha A Yousri, Fabian J Theis, Karsten Suhre, and Jan Krumsiek. Network-Based Approach for Analyzing Intra- and Interfluid Metabolite Associations in Human Blood, Urine, and Saliva. *Journal of Proteome Research*, 14(2):1183–1194, 2015.

Hajk-Georg Drost and Jerzy Paszkowski. Biomartr: genomic data retrieval with R. *Bioinformatics*, 33(8):1216–1217, 2017.

Aiping Du, Kyle W McCracken, Erik R Walp, Natalie A Terry, Thomas J Klein, Annie Han, James M Wells, and Catherine Lee May. Arx is required for normal enteroendocrine cell development in mice and humans. *Developmental Biology*, 365(1):175–188, 2012.

Angelo Duò, Mark D Robinson, and Charlotte Soneson. A systematic performance evaluation of clustering methods for single-cell RNA-seq data. *F1000Research*, 7:1141, 2018.

James Eberwine, H Yeh, K Miyashiro, Y Cao, S Nair, R Finnell, M Zettel, and P Coleman. Analysis of gene expression in single live neurons. *Proceedings of the National Academy of Sciences*, 89(7):3010–3014, 1992.

James Eberwine, Jai-Yoon Sul, Tamas Bartfai, and Junhyong Kim. The promise of single-cell sequencing. *Nature Methods*, 11:25–27, 2014.

Camille Ehre, Andrea H Rossi, Lubna H Abdullah, Kathleen De Pestel, Sandra Hill, John C Olsen, and C William Davis. Barrier role of actin filaments in regulated mucin secretion from airway goblet cells. *American Journal of Physiology-Cell Physiology*, 288: C46–C56, 2005.

Michael Elowitz, Arnold J Levine, Eric D Siggia, and Peter S Swain. Stochastic gene expression in a single cell. *Science*, 297:1183–1186, 2002.

Gökcen Eraslan, Lukas M Simon, Maria Mircea, Nikola S Mueller, and Fabian J Theis. Single-cell RNA-seq denoising using a deep count autoencoder. *Nature Communications*, 10:390, 2019.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD-96 Proceedings*, pages 226–231, 1996.

M J Evans and M H Kaufman. Establishment in culture of pluripotential cells from mouse embryos. *Nature*, 292:154–156, 1981.

Jean Fan, Neeraj Salathia, Rui Liu, Gwendolyn E Kaeser, Yun C Yung, Joseph L Herman, Fiona Kaper, Jian-Bing Fan, Kun Zhang, Jerold Chun, and Peter V Kharchenko. Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis. *Nature Methods*, 13:241–244, 2016.

Xiaoying Fan, Xiannian Zhang, Xinglong Wu, Hongshan Guo, Yuqiong Hu, Fuchou Tang, and Yanyi Huang. Single-cell RNA-seq transcriptome analysis of linear and circular RNAs in mouse preimplantation embryos. *Genome Biology*, 16:148, 2015.

Andrea M Femino, Fredric S Fay, Kevin Fogarty, and Robert H Singer. Visualization of Single RNA Transcripts in Situ. *Science*, 280:585–590, 1998.

Greg Finak, Andrew McDavid, Masanao Yajima, Jingyuan Deng, Vivian Gersuk, Alex K Shalek, Chloe K Slichter, Hannah W Miller, M Julianna McElrath, Martin Prlic, Peter S Linsley, and Raphael Gottardo. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biology*, 16:278, 2015.

Silvia Fre, Mathilde Huyghe, Philippos Mourikis, Sylvie Robine, Daniel Louvard, and Spyros Artavanis-Tsakonas. Notch signals control the fate of immature progenitor cells in the intestine. *Nature*, 435:964–968, 2005.

Greg Friedman. Survey Article: An elementary illustrated introduction to simplicial sets. *Rocky Mountain Journal of Mathematics*, 42(2):353–423, 2012.

Johann A Gagnon-Bartsch and Terence P Speed. Using control genes to correct for unwanted variation in microarray data. *Biostatistics*, 13(3):539–552, 2012.

Moritz Gegg, Anika Böttcher, Ingo Burtscher, Stefan Hasenoeder, Claude Van Campenhout, Michaela Aichler, Axel Walch, Seth G N Grant, and Heiko Lickert. Flattop regulates basal body docking and positioning in mono- and multiciliated cells. *eLife*, 3: 50–24, 2014.

Arsham Ghahramani, Fiona M Watt, and Nicholas M Luscombe. Generative adversarial

networks simulate gene expression and predict perturbations in single cells. *bioRxiv*, 2018.

Mubeen Goolam, Antonio Scialdone, Sarah J L Graham, Iain C Macaulay, Agnieszka Jedrusik, Anna Hupalowska, Thierry Voet, John C Marioni, and Magdalena Zernicka-Goetz. Heterogeneity in Oct4 and Sox2 Targets Biases Cell Fate in 4-Cell Mouse Embryos. *Cell*, 165:61–74, 2016.

Jesse M Gray, David A Harmin, Sarah A Boswell, Nicole Cloonan, Thomas E Mullen, Joseph J Ling, Nimrod Miller, Scott Kuersten, Yong-Chao Ma, Steven A McCarroll, Sean M Grimmond, and Michael Springer. SnapShot-Seq: A Method for Extracting Genome-Wide, In Vivo mRNA Dynamics from a Single Total RNA Sample. *PLoS ONE*, 9(2):e89673–17, 2014.

Fiona M Gribble and Frank Reimann. Enteroendocrine Cells: Chemosensors in the Intestinal Epithelium. *Annual Review of Physiology*, 78(1):277–299, 2016.

Vladislav Grishkevich and Itai Yanai. Gene length and expression level shape genomic novelties. *Genome Research*, 24:1497–1503, 2014.

Dominic Grün, Lennart Kester, and Alexander Van Oudenaarden. Validation of noise models for single-cell transcriptomics. *Nature Methods*, 11:637–640, 2014.

Dominic Grün, Anna Lyubimova, Lennart Kester, Kay Wiebrands, Onur Basak, Nobuo Sasaki, Hans Clevers, and Alexander van Oudenaarden. Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature*, 525:251–255, 2015.

Dominic Grün, Mauro J Muraro, Jean-Charles Boisset, Kay Wiebrands, Anna Lyubimova, Gitanjali Dharmadhikari, Maaike van den Born, Johan van Es, Erik Jansen, Hans Clevers, Eelco J P de Koning, and Alexander van Oudenaarden. De Novo Prediction of Stem Cell Identity using Single-Cell Transcriptome Data. *Cell Stem Cell*, 19: 266–277, 2016.

Rong-Jun Guo, Shinsuke Funakoshi, Hannah H Lee, Jianping Kong, and John P Lynch. The intestine-specific transcription factor Cdx2 inhibits β-catenin/TCF transcriptional activity by disrupting the β-catenin–TCF protein complex. *Carcinogenesis*, 31(2):159–166, 2009.

Adam L Haber, Moshe Biton, Noga Rogel, Rebecca H Herbst, Karthik Shekhar, Christopher Smillie, Grace Burgin, Toni M Delorey, Michael R Howitt, Yarden Katz, Itay Tirosh, Semir Beyaz, Danielle Dionne, Mei Zhang, Raktima Raychowdhury, Wendy S Garrett, Orit Rozenblatt-Rosen, Hai Ning Shi, Omer Yilmaz, Ramnik J Xavier, and Aviv Regev. A single-cell survey of the small intestinal epithelium. *Nature*, 551:333–339, 2017.

Laleh Haghverdi, Florian Buettner, and Fabian J Theis. Diffusion maps for high-

dimensional single-cell analysis of differentiation data. *Bioinformatics*, 31(18):2989–2998, 2015.

Laleh Haghverdi, Maren Büttner, F Alexander Wolf, and Florian Buettner. Diffusion pseudotime robustly reconstructs lineage branching. *Nature Methods*, 13:845–848, 2016.

Laleh Haghverdi, Aaron T L Lun, Michael D Morgan, and John C Marioni. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nature Biotechnology*, 36:421–427, 2018.

Keren Bahar Halpern, Rom Shenhav, Orit Matcovitch-Natan, Beata Toth, Doron Lemze, Matan Golan, Efi E Massasa, Shaked Baydatch, Shanie Landen, Andreas E Moor, Alexander Brandis, Amir Giladi, Avigail Stokar-Avihail, Eyal David, Ido Amit, and Shalev Itzkovitz. Single-cell spatial reconstruction reveals global division of labour in the mammalian liver. *Nature*, 542:352–356, 2017.

Yuanyuan Han, Yi Gu, Alex Ce Zhang, and Yu-Hwa Lo. Review: imaging technologies for flow cytometry. *Lab on a Chip*, 16(24):4639–4647, 2016.

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, 2012.

Nicola Harris. The enigmatic tuft cell in immunity. *Science*, 351(6279):1264–1265, 2016.

Tamar Hashimshony, Florian Wagner, Noa Sher, and Itai Yanai. CEL-Seq: Single-Cell RNA-Seq by Multiplexed Linear Amplification. *CellReports*, 2(3):666–673, 2012.

Tamar Hashimshony, Naftalie Senderovich, Gal Avital, Agnes Klochendler, Yaron de Leeuw, Leon Anavy, Dave Gennert, Shuqiang Li, Kenneth J Livak, Orit Rozenblatt-Rosen, Yuval Dor, Aviv Regev, and Itai Yanai. CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. *Genome Biology*, 17:77, 2016.

Gert-Jan Hendriks, Dimos Gaidatzis, Florian Aeschimann, and Helge Großhans. Extensive Oscillatory Gene Expression during C. elegans Larval Development. *Molecular Cell*, 53 (3):380–392, 2014.

Stephanie C Hicks, F Willliam Townes, Mingxiang Teng, and Rafael A Irizarry. Missing data and technical variability in single-cell RNA-sequencing experiments. *Biostatistics*, pages 1–17, 2017.

Joshua Ho. CIDR: Ultrafast and accurate clustering through imputation for single-cell RNA-seq data. *Genome Biology*, 18:59, 2017.

Audrey W Hong, Zhipeng Meng, and Kun-Liang Guan. The Hippo pathway in intestinal regeneration and disease. *Nature Methods*, 13:324–337, 2016.

Robert Hooke. *Micrographia*. London, 1665.

Michael R Howitt, Sydney Lavoie, Monia Michaud, Arthur M Blum, Sara V Tran, Joel V Weinstock, Carey Ann Gallini, Kevin Redding, Robert F Margolskee, Lisa C Osborne,

David Artis, and Wendy S Garrett. Tuft cells, taste-chemosensory cells, orchestrate parasite type 2 immunity in the gut. *Science*, 351(6279):1329–1333, 2016.

Mo Huang, Jingshu Wang, Eduardo Torre, Hannah Dueck, Sydney Shaffer, Roberto Bonasio, John I Murray, Arjun Raj, Mingyao Li, and Nancy R Zhang. SAVER: gene expression recovery for single-cell RNA sequencing. *Nature Methods*, 15:539–542, 2018.

Wolfgang Huber, Vincent J Carey, Robert Gentleman, Simon Anders, Marc Carlson, Benilton S Carvalho, Héctor Corrada Bravo, Sean Davis, Laurent Gatto, Thomas Girke, Raphael Gottardo, Florian Hahne, Kasper D Hansen, Rafael A Irizarry, Michael Lawrence, Michael I Love, James MacDonald, Valerie Obenchain, Andrzej K Oleś, Hervé Pagès, Alejandro Reyes, Paul Shannon, Gordon K Smyth, Dan Tenenbaum, Levi Waldron, and Martin Morgan. Orchestrating high-throughput genomic analysis with Bioconductor. *Nature Methods*, 12:115–121, 2015.

H R Hulett, William A Bonner, Janet Barrett, and Leonard A Herzenberg. Cell Sorting: Automated Separation of Mammalian Cells as a Function of Intracellular Fluorescence. *Science*, 166(3906):747–749, 1969.

Tomislav Ilicic, Jong Kyoung Kim, Aleksandra A Kolodziejczyk, Frederik Otzen Bagger, Davis James McCarthy, John C Marioni, and Sarah A Teichmann. Classification of low quality cells from single-cell RNA-seq data. *Genome Biology*, 17:29, 2016.

Saiful Islam, U Kjallquist, A Moliner, P Zajac, J B Fan, P Lonnerberg, and S Linnarsson. Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Research*, 21:1160–1167, 2011.

Saiful Islam, Amit Zeisel, Simon Joost, Gioele La Manno, Pawel Zajac, Maria Kasper, Peter Lönnerberg, and Sten Linnarsson. Quantitative single-cell RNA-seq with unique molecular identifiers. *Nature Methods*, 11(2):163–166, 2014.

Mathieu Jacomy, Tommaso Venturini, Sebastien Heymann, and Mathieu Bastian. ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software. *PLoS ONE*, 9(6):e98679–12, 2014.

Diego Adhemar Jaitin, Ephraim Kenigsberg, Hadas Keren-Shaul, Naama Elefant, Franziska Paul, Irina Zaretsky, Alexander Mildner, Nadav Cohen, Steffen Jung, Amos Tanay, and Ido Amit. Massively Parallel Single-Cell RNA-Seq for Marker-Free Decomposition of Tissues into Cell Types. *Science*, 343(6172):772–776, 2014.

Erich D Jarvis, Siavash Mirarab, Andre J Aberer, Bo Li, Peter Houde, Cai Li, Simon Y W Ho, Brant C Faircloth, Benoit Nabholz, Jason T Howard, Alexander Suh, Claudia C Weber, Rute R da Fonseca, Jianwen Li, Fang Zhang, Hui Li, Long Zhou, Nitish Narula, Liang Liu, Ganesh Ganapathy, Bastien Boussau, Md Shamsuzzoha Bayzid, Volodymyr Zavidovych, Sankar Subramanian, Toni Gabaldón, Salvador Capella-Gutiérrez, Jaime

Huerta-Cepas, Bhanu Rekepalli, Kasper Munch, Mikkel Schierup, Bent Lindow, Wesley C Warren, David Ray, Richard E Green, Michael W Bruford, Xiangjiang Zhan, Andrew Dixon, Shengbin Li, Ning Li, Yinhua Huang, Elizabeth P Derryberry, Mads Frost Bertelsen, Frederick H Sheldon, Robb T Brumfield, Claudio V Mello, Peter V Lovell, Morgan Wirthlin, Maria Paula Cruz Schneider, Francisco Prosdocimi, José Alfredo Samaniego, Amhed Missael Vargas Velazquez, Alonzo Alfaro-Núñez, Paula F Campos, Bent Petersen, Thomas Sicheritz-Ponten, An Pas, Tom Bailey, Paul Scofield, Michael Bunce, David M Lambert, Qi Zhou, Polina Perelman, Amy C Driskell, Beth Shapiro, Zijun Xiong, Yongli Zeng, Shiping Liu, Zhenyu Li, Binghang Liu, Kui Wu, Jin Xiao, Xiong Yinqi, Qiuemei Zheng, Yong Zhang, Huanming Yang, Jian Wang, Linnea Smeds, Frank E Rheindt, Michael Braun, Jon Fjeldsa, Ludovic Orlando, F Keith Barker, Knud Andreas Jønsson, Warren Johnson, Klaus-Peter Koepfli, Stephen O'Brien, David Haussler, Oliver A Ryder, Carsten Rahbek, Eske Willerslev, Gary R Graves, Travis C Glenn, John McCormack, Dave Burt, Hans Ellegren, Per Alström, Scott V Edwards, Alexandros Stamatakis, David P Mindell, Joel Cracraft, Edward L Braun, Tandy Warnow, Wang Jun, M Thomas P Gilbert, and Guojie Zhang. Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science*, 346:1320–1331, 2014.

Lan Jiang, Huidong Chen, Luca Pinello, and Guo-Cheng Yuan. GiniClust: detecting rare cell types from single-cell gene expression data with Gini index. *Genome Biology*, 17:144, 2016.

W Evan Johnson, Cheng Li, and Ariel Rabinovic. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, 8(1):118–127, 2007.

M H Julius, T Masuda, and Leonhard A Herzenberg. Demonstration That Antigen-Binding Cells Are Precursors of Antibody-Producing Cells After Purification with a Fluorescence-Activated Cell Sorter. *Proceedings of the National Academy of Sciences*, 69(7):1934–1938, 1972.

Jan Philipp Junker and Alexander van Oudenaarden. Every Cell Is Special: Genome-wide Studies Add a New Dimension to Single-Cell Biology. *Cell*, 157(1):8–11, 2014.

Minoru Kanehisa, Yoko Sato, Masayuki Kawashima, Miho Furumichi, and Mao Tanabe. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research*, 44(D1):D457–D462, 2016.

H M Kang, M Subramaniam, S Targ, M Nguyen, Lenka Maliskova, Elizabeth McCarthy, Eunice Wan, Simon Wong, Lauren Byrnes, Cristina M Lanata, Rachel E Gate, Sara Mostafavi, Alexander Marson, Noah Zaitlen, Lindsey A Criswell, and Chun Jimmie Ye. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nature Methods*, 36(1):89–94, 2018.

Nikos Karaiskos, Philipp Wahle, Jonathan Alles, Anastasiya Boltengagen, Salah Ayoub, Claudia Kipar, Christine Kocks, Nikolaus Rajewsky, and Robert P Zinzen. The Drosophila embryo at single-cell transcriptome resolution. *Science*, 358(6360):194–199, 2017.

Peter V Kharchenko, Lev Silberstein, and David T Scadden. Bayesian approach to single-cell differential expression analysis. *Nature Methods*, 11(7):740–742, 2014.

Sun-Hee Kim, Jason Roszik, Elizabeth A Grimm, and Suhendan Ekmekcioglu. Impact of l-Arginine Metabolism on Immune Response and Anticancer Immunotherapy. *Frontiers in Oncology*, 8:3016–5, 2018.

Tae-Hee Kim, Assieh Saadatpour, Guoji Guo, Madhurima Saxena, Alessia Cavazza, Niyati Desai, Unmesh Jadhav, Lan Jiang, Miguel N Rivera, Stuart H Orkin, Guo-Cheng Yuan, and Ramesh A Shivdasani. Single-Cell Transcript Profiles Reveal Multilineage Priming in Early Progenitors Derived from Lgr5+ Intestinal Stem Cells. *CellReports*, pages 1–21, 2016.

Vladimir Yu Kiselev, Kristina Kirschner, Michael T Schaub, Tallulah S Andrews, Andrew Yiu, Tamir Chandra, Kedar N Natarajan, Wolf Reik, Mauricio Barahona, Anthony R Green, and Martin Hemberg. SC3: consensus clustering of single-cell RNA-seq data. *Nature Methods*, 14:483–486, 2017.

Vladimir Yu Kiselev, Tallulah S Andrews, and Martin Hemberg. Challenges in unsupervised clustering of single-cell RNA-seq data. *Nature Reviews Genetics*, pages 1–10, 2019.

Allon M Klein, Linas Mazutis, Ilke Akartuna, Naren Tallapragada, Adrian Veres, Victor Li, Leonid Peshkin, David A Weitz, and Marc W Kirschner. Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells. *Cell*, 161(5):1187–1201, 2015.

Teuvo Kohonen. Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43(1):59–69, 1982.

Aleksandra A Kolodziejczyk, J K Kim, JCH Tsang, Tomislav Ilicic, Johan Henriksson, Kedar N Natarajan, Alex C Tuck, Xuefei Gao, Marc Bühler, Pentao Liu, John C Marioni, and Sarah A Teichmann. Single cell RNA-sequencing of pluripotent states unlocks modular transcriptional variation. *Cell Stem Cell*, 17:471–485, 2015a.

Aleksandra A Kolodziejczyk, Jong Kyoung Kim, Valentine Svensson, John C Marioni, and Sarah A Teichmann. The Technology and Biology of Single-Cell RNA Sequencing. *Molecular Cell*, 58(4):610–620, 2015b.

Ilya Korsunsky, Jean Fan, Kamil Slowikowski, Fan Zhang, Kevin Wei, Yuriy Baglaenko, Michael Brenner, Po-Ru Loh, and Soumya Raychaudhuri. Fast, sensitive, and accurate

integration of single cell data with Harmony. *bioRxiv*, 2018.

Keegan D Korthauer, Li-Fang Chu, Michael A Newton, Yuan Li, James Thomson, Ron Stewart, and Christina Kendziorski. A statistical approach for identifying differential distributions in single-cell RNA-seq experiments. *Genome Biology*, 17:222, 2016.

T Kucharzik, N Lügering, K Rautenberg, A Lügering, M A Schmidt, R Stoll, and W Domschke. Role of M Cells in Intestinal Barrier Function. *Annals of the New York Academy of Sciences*, 915(1):171–183, 2006.

S Kullback and R A Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

Gioele La Manno, Ruslan Soldatov, Amit Zeisel, Emelie Braun, Hannah Hochgerner, Viktor Petukhov, Katja Lidschreiber, Maria E Kastriti, Peter Lönnerberg, Alessandro Furlan, Jean Fan, Lars E Borm, Zehua Liu, David Bruggen, Jimin Guo, Xiaoling He, Roger Barker, Erik Sundstrom, Goncalo Castelo-Branco, Patrick Cramer, Igor Adameyko, Sten Linnarsson, and Peter V Kharchenko. RNA velocity of single cells. *Nature*, 560:494–498, 2018.

María Jesús Larriba, Paloma Ordóñez-Morán, Irene Chicote, Génesis Martín-Fernández, Isabel Puig, Alberto Muñoz, and Héctor G Pálmer. Vitamin D Receptor Deficiency Enhances Wnt/β-Catenin Signaling and Tumor Burden in Colon Cancer. *PLoS ONE*, 6(8):e23524–13, 2011.

Philip Leder and Marshall W Nirenberg. RNA Codewords and Protein Synthesis, III. On the nucleotide Sequence of a Cysteine and a Leucine RNA Codeword. *Proceedings of the National Academy of Sciences*, 52:1521–1529, 1964.

J T Leek, W Evan Johnson, H S Parker, and A E Jaffe. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Nucleic Acid Research*, 28(6):882–883, 2012.

Jacob H Levine, Erin F Simonds, Sean C Bendall, Kara L Davis, El-ad David Amir, Michelle D Tadmor, Oren Litvin, Harris G Fienberg, Astraea Jager, Eli R Zunder, Rachel Finck, Amanda L Gedman, Ina Radtke, James R Downing, Dana Peer, and Garry P Nolan. Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. *Cell*, 162(1):184–197, 2015.

Jeffrey M Levsky, Shailesh M Shenoy, Rossanna C Pezo, and Robert H Singer. Single-Cell Gene Expression Profiling. *Science*, 297:836–840, 2002.

Bo Li, Victor Ruotti, Ron M Stewart, James A Thomson, and Colin N Dewey. RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics*, 26(4): 493–500, 2009.

Wei Vivian Li and Jingyi Jessica Li. An accurate and robust imputation method scImpute

for single-cell RNA-seq data. *Nature Communications*, 9(1):997, 2018.

Zhe Liang, Yan Xie, Jessica A Dominguez, Elise R Breed, Benyam P Yoseph, Eileen M Burd, Alton B Farris, Nicholas O Davidson, and Craig M Coopersmith. Intestine-Specific Deletion of Microsomal Triglyceride Transfer Protein Increases Mortality in Aged Mice. *PLoS ONE*, 9(7):e101828–11, 2014.

Erez Lieberman-Aiden, Nynke L van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, Bryan R Lajoie, Peter J Sabo, Michael O Dorschner, Richard Sandstrom, Bradley Bernstein, M A Bender, Mark Groudine, Andreas Gnirke, John Stamatoyannopoulos, Leonid A Mirny, Eric S Lander, and Job Dekker. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326:289–293, 2009.

Qi Liu, Charles A Herring, Quanhu Sheng, Jie Ping, Alan J Simmons, Bob Chen, Amrita Banerjee, Guoqiang Gu, Robert J Coffey, Yu Shyr, and Ken S Lau. Quantitative assessment of cell population diversity in single-cell landscapes. *bioRxiv*, 2018.

Wenqiang Liu, Xiaoyu Liu, Chenfei Wang, Yawei Gao, Rui Gao, Xiaochen Kou, Yanhong Zhao, Jingyi Li, You Wu, Wenchao Xiu, Su Wang, Jiqing Yin, Wei Liu, Tao Cai, Hong Wang, Yong Zhang, and Shaorong Gao. Identification of key factors conquering developmental arrest of somatic cell cloned embryos by combining embryo biopsy and single-cell sequencing. *Cell Discovery*, 2(1):16010–16015, 2016.

Yuan-Hung Lo, Eunah Chung, Zhaohui Li, Ying-Wooi Wan, Maxime M Mahe, Min-Shan Chen, Taeko K Noah, Kristin N Bell, Hari Krishna Yalamanchili, Tiemo J Klisch, Zhandong Liu, Joo-Seop Park, and Noah F Shroyer. Transcriptional Regulation by ATOH1 and its Target SPDEF in the Intestine. *Cellular and Molecular Gastroenterology and Hepatology*, 3(1):51–71, 2017.

Tapio Lönnberg, Valentine Svensson, and K R James. Single-cell RNA-seq and computational analysis using temporal mixture modelling resolves Th1/Tfh fate bifurcation in malaria. *Science Immunology*, 2(9):eaal2192, 2017.

Romain Lopez, Jeffrey Regier, Michael B Cole, Michael I Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 15:1053–1058, 2018.

Mohammad Lotfollahi, F Alexander Wolf, and Fabian J Theis. Generative modeling and latent space arithmetics predict single-cell perturbation response across cell types, studies and species. *bioRxiv*, 2018.

Aaron T L Lun and John C Marioni. Overcoming confounding plate effects in differential expression analyses of single-cell RNA-seq data. *Biostatistics*, pages 1–14, 2017.

Aaron T L Lun, Karsten Bach, and John C Marioni. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biology*, 17:75, 2016a.

Aaron T L Lun, Davis James McCarthy, and John C Marioni. A step-by-step workflow for low-level analysis of single-cell RNA-seq data. *F1000Research*, 5:2122–2164, 2016b.

Evan Z Macosko, Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, Allison R Bialas, Nolan Kamitaki, Emily M Martersteck, John J Trombetta, David A Weitz, Joshua R Sanes, Alex K Shalek, Aviv Regev, and Steven A McCarroll. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell*, 161(5):1202–1214, 2015.

J MacQueen. Some methods for classification and analysis of multivariate observations. *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297, 1967.

Yann Marcy, Cleber Ouverney, Elisabeth M Bik, Tina Lösekann, Natalia Ivanova, Hector Garcia Martin, Ernest Szeto, Darren Platt, Philip Hugenholtz, David A Relman, and Stephen R Quake. Dissecting biological "dark matter" with single-cell genetic analysis of rare and uncultivated TM7 microbes from the human mouth. *Proceedings of the National Academy of Sciences*, 104(29):11889–11894, 2007.

Carsten Marr, Joseph X Zhou, and Sui Huang. Single-cell gene expression profiling and cell state dynamics: collecting data, correlating data points and connecting the dots. *Current Opinion in Biotechnology*, 39:207–214, 2016.

G R Martin. Isolation of a pluripotent cell line from early mouse embryos cultured in medium conditioned by teratocarcinoma stem cells. *Proceedings of the National Academy of Sciences*, 78(12):7634–7638, 1981.

Celia Pilar Martinez-Jimenez, Nils Eling, Hung-Chang Chen, Catalina A Vallejos, Aleksandra A Kolodziejczyk, Frances Connor, Lovorka Stojic, Timothy F Rayner, Michael J T Stubbington, Sarah A Teichmann, Maike de la Roche, John C Marioni, and Duncan T Odom. Aging increases cell-to-cell transcriptional variability upon immune stimulation. *Science*, 355(6332):1433–1436, 2017.

Vivien Marx. The big challenges of big data. *Nature*, 498:255–260, 2013.

Paolo Mazzarello. A unifying concept: the history of cell theory. *Nature Cell Biology*, 1: E13–E15, 1999.

Davis James McCarthy, Kieran R Campbell, Aaron T L Lun, and Quin F Wills. Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. *Bioinformatics*, pages 1–8, 2017.

Leland McInnes and John Healy. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv*, 2018.

Ashish Misra and Michael R Green. From polyadenylation to splicing: Dual role for mRNA 3' end formation factors. *RNA biology*, 13(3):259–264, 2016.

Yuko Mori Akiyama, Maaike van den Born, Johan H van Es, Stanley R Hamilton, Henry P Adams, Jiexin Zhang, Hans Clevers, and Benoit de Crombrugghe. SOX9 Is Required for the Differentiation of Paneth Cells in the Intestinal Epithelium. *Gastroenterology*, 133(2):539–546, 2007.

Ali Mortazavi, Brian A Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, 5 (7):621–628, 2008.

Javier Muñoz, Daniel E Stange, Arnout G Schepers, Marc van de Wetering, Bon-Kyoung Koo, Shalev Itzkovitz, Richard Volckmann, Kevin S Kung, Jan Koster, Sorina Radulescu, Kevin Myant, Rogier Versteeg, Owen J Sansom, Johan H van Es, Nick Barker, Alexander van Oudenaarden, Shabaz Mohammed, Albert J R Heck, and Hans Clevers. The Lgr5 intestinal stem cell signature: robust expression of proposed quiescent '+4' cell markers. *The EMBO Journal*, 31(14):3079–3091, 2012.

Kevin P Murphy. *Machine Learning: A probabilistic perpective*. MIT Press, 2012.

Takashi Nagano, Yaniv Lubling, Tim J Stevens, Stefan Schoenfelder, Eitan Yaffe, Wendy Dean, Ernest D Laue, Amos Tanay, and Peter Fraser. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature*, 502:59–64, 2013.

M E J Newman and M Girvan. Finding and evaluating community structure in networks. *Physical Review E*, 69:026113, 2004.

Jerzy Neyman, Egon S Pearson, and Karl Pearson. IX. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231 (694-706):289–337, 1933.

Neeta Patel, Joe Varghese, Patarabutr Masaratana, Gladys O Latunde-Dada, Molly Jacob, Robert J Simpson, and Andrew T McKie. The transcription factor ATOH8 is regulated by erythropoietic activity and regulates HAMPtranscription and cellular pSMAD1,5,8 levels. *British Journal of Haematology*, 164(4):586–596, 2013.

Rob Patro, Geet Duggal, Michael I Love, Rafael A Irizarry, and Carl Kingsford. Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*, pages 1–10, 2017.

Franziska Paul, Ya'ara Arkin, A Giladi, Diego Adhemar Jaitin, and Ephraim Kenigsberg. Transcriptional heterogeneity and lineage commitment in myeloid progenitors. *Cell*, 2015.

Joseph N Paulson, Cho-Yi Chen, Camila M Lopes-Ramos, Marieke L Kuijjer, John Platig, Abhijeet R Sonawane, Maud Fagny, and John Quackenbush. Tissue-aware RNA-Seq processing and normalization for heterogeneous and sparse data. *BMC Bioinformatics*,

18:437, 2017.

Karl Pearson. LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11): 559–572, 1901.

Elizabeth Pennisi. Development cell by cell. *Science*, 362(6421):1344–1345, 2018.

Simone Picelli, Åsa K Björklund, O R Faridani, S Sagasser Nature, Gösta Winberg, and Rickard Sandberg. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nature Methods*, 10:1096–1098, 2013.

E Pierson and Christopher Yau. ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome Biology*, 16:241, 2015.

Blanca Pijuan-Sala, Carolina Guibentif, and Berthold Göttgens. Single-cell transcriptional profiling: a window into embryonic cell-type specification. *Nature Reviews Molecular Cell Biology*, 19(6):399–412, 2018.

Mireya Plass, Jordi Solana, F Alexander Wolf, Salah Ayoub, Aristotelis Misios, Petar Glažar, Benedikt Obermayer, Fabian J Theis, Christine Kocks, and Nikolaus Rajewsky. Cell type atlas and lineage tree of a whole complex animal by single-cell transcriptomics. *Science*, 360(6391):eaaq1723, 2018.

Zhen Qi, Yehua Li, Bing Zhao, Chi Xu, Yuan Liu, Haonan Li, Bingjie Zhang, Xinquan Wang, Xiao Yang, Wei Xie, Baojie Li, Jing-Dong Jackie Han, and Ye-Guang Chen. BMP restricts stemness of intestinal Lgr5+ stem cells by directly suppressing their signature genes. *Nature Communications*, 7:1–14, 2016.

Peng Qiu, Erin F Simonds, Sean C Bendall, Kenneth D Gibbs, Robert V Bruggner, Michael D Linderman, Karen Sachs, Garry P Nolan, and Sylvia K Plevritis. Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE. *Nature Biotechnology*, 29(10):886–891, 2011.

Daniel Ramsköld, Shujun Luo, Yu-Chieh Wang, Robin Li, Qiaolin Deng, Omid R Faridani, Gregory A Daniels, Irina Khrebtukova, Jeanne F Loring, Louise C Laurent, Gary P Schroth, and Rickard Sandberg. Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nature Biotechnology*, 30:777–782, 2012.

J Rathjen, J A Lake, M D Bettess, J M Washington, G Chapman, and P D Rathjen. Formation of a primitive ectoderm like cell population, EPL cells, from ES cells in response to biologically derived factors. *Journal of Cell Science*, 112:601–612, 1999.

Aviv Regev, Sarah A Teichmann, Eric S Lander, Ido Amit, Christophe Benoist, Ewan Birney, Bernd Bodenmiller, Peter Campbell, Piero Carninci, Menna Clatworthy, Hans Clevers, Bart Deplancke, Ian Dunham, James Eberwine, Roland Eils, Wolfgang Enard, Andrew Farmer, Lars Fugger, Berthold Göttgens, Nir Hacohen, Muzlifah Haniffa, Mar-

tin Hemberg, Seung Kim, Paul Klenerman, Arnold Kriegstein, Ed Lein, Sten Linnarsson, Emma Lundberg, Joakim Lundeberg, Partha Majumder, John C Marioni, Miriam Merad, Musa Mhlanga, Martijn Nawijn, Mihai Netea, Garry Nolan, Dana Peer, Anthony Phillipakis, Chris P Ponting, Stephen Quake, Wolf Reik, Orit Rozenblatt-Rosen, Joshua Sanes, Rahul Satija, Ton N Schumacher, Alex Shalek, Ehud Shapiro, Padmanee Sharma, Jay W Shin, Oliver Stegle, Michael Stratton, Michael J T Stubbington, Fabian J Theis, Matthias Uhlen, Alexander van Oudenaarden, Allon Wagner, Fiona Watt, Jonathan Weissman, Barbara Wold, Ramnik Xavier, Nir Yosef, and Human Cell Atlas Meeting Participants. Science Forum: The Human Cell Atlas. *eLife*, 6:1–30, 2017.

Jüri Reimand, Tambet Arak, Priit Adler, Liis Kolberg, Sulev Reisberg, Hedi Peterson, and Jaak Vilo. g:Profiler—a web server for functional interpretation of gene lists (2016 update). *Nucleic Acids Research*, 44(W1):W83–W89, 2016.

Camilla A Richmond and David T Breault. Regulation of Gene Expression in the Intestinal Epithelium. *Progress in Molecular Biology and Translational Science*, 96:207–229, 2010.

Davide Risso, John Ngai, Terence P Speed, and Sandrine Dudoit. Normalization of RNA-seq data using factor analysis of control genes or samples. *Nature Biotechnology*, 32(9): 896–902, 2014.

Davide Risso, Fanny Perraudeau, Svetlana Gribkova, Sandrine Dudoit, and Jean-Philippe Vert. A general and flexible method for signal extraction from single-cell RNA-seq data. *Nature Communications*, 9:284, 2018.

M E Ritchie, B Phipson, D Wu, Y Hu, C W Law, W Shi, and Gordon K Smyth. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47–e47, 2015.

Mark D Robinson and Alicia Oshlack. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, 11:R25, 2010.

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.

Wouter Saelens, Robrecht Cannoodt, Helena Todorov, and Yvan Saeys. A comparison of single-cell trajectory inference methods: towards more accurate and robust tools. *bioRxiv*, 2018.

R K Saiki, D H Gelfand, S Stoffel, S J Scharf, R Higuchi, G T Horn, K B Mullis, and H A Erlich. Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science*, 239(4839):487–491, 1988.

Rickard Sandberg. Entering the era of single-cell transcriptomics in biology and medicine. *Nature Methods*, 11:22–24, 2014.

Rahul Satija, J A Farrell, D Gennert, and A F Schier. Spatial reconstruction of single-cell

gene expression data. *Nature Biotechnology*, 33(5), 2015.

Toshiro Sato, Johan H van Es, Hugo J Snippert, Daniel E Stange, Robert G Vries, Maaike van den Born, Nick Barker, Noah F Shroyer, Marc van de Wetering, and Hans Clevers. Paneth cells constitute the niche for Lgr5 stem cells in intestinal crypts. *Nature*, 469: 415–418, 2011.

Matthias J Schleiden. Beiträge zur Phytogenesis. *Archiv für Anatomie, Physiologie und Wissenschaftliche Medicin*, pages 137–176, 1838.

Natalia Schröder and Achim Gossler. Expression of Notch pathway components in fetal and adult mouse small intestine. *Gene Expression Patterns*, 2:247–250, 2002.

Jurian Schuijers, Jan Philipp Junker, Michal Mokry, Pantelis Hatzis, Bon-Kyoung Koo, Valentina Sasselli, Laurens G van der Flier, Edwin Cuppen, Alexander van Oudenaarden, and Hans Clevers. Ascl2 Acts as an R-spondin/Wnt-Responsive Switch to Control Stemness in Intestinal Crypts. *Stem Cell*, 16(2):158–170, 2015.

Daniel Schulz, Vito Riccardo Tomaso Zanotelli, Jana Raja Fischer, Denis Schapiro, Stefanie Engler, Xiao-Kang Lun, Hartland Warren Jackson, and Bernd Bodenmiller. Simultaneous Multiplexed Imaging of mRNA and Proteins with Subcellular Resolution in Breast Cancer Tissue Samples by Mass Cytometry. *Cell Systems*, 6:25–36, 2018.

Theodor Schwann. *Mikroskopische Untersuchungen über die Uebereinstimmung in der Struktur und dem Wachsthum der Thiere und Pflanzen.* Sander, Berlin, 1839.

Omer Schwartzman and Amos Tanay. Single-cell epigenomics: techniques and emerging applications. *Nature Methods*, 16:716–726, 2015.

Antonio Scialdone, K N Natarajan, L R Saraiva, and V Proserpio Methods. Computational assignment of cell-cycle stage from single-cell transcriptome data. *Cell Stem Cell*, 85: 54–61, 2015.

Hassan Sellak, Songwei Wu, and Thomas M Lincoln. KLF4 and SOX9 transcription factors antagonize β-catenin and inhibit TCF-activity in cancer cells. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research*, 1823(10):1666–1675, 2012.

Stefan Semrau, Johanna E Goldmann, Magali Soumillon, Tarjei S Mikkelsen, Rudolf Jaenisch, and Alexander van Oudenaarden. Dynamics of lineage commitment revealed by single-cell transcriptomics of differentiating embryonic stem cells. *Nature Communications*, pages 1–16, 2017.

Manu Setty, Michelle D Tadmor, Shlomit Reich-Zeliger, Omer Angel, Tomer Meir Salame, Pooja Kathail, Kristy Choi, Sean C Bendall, Nir Friedman, and Dana Peer. Wishbone identifies bifurcating developmental trajectories from single-cell data. *Nature Biotechnology*, pages 1–14, 2016.

Debra G Silberg, Jessica Sullivan, Eugene Kang, Gary P Swain, Jennifer Moffett, New-

man J Sund, Sara D Sackett, and Klaus H Kaestner. Cdx2 ectopic expression induces gastric intestinal metaplasia in transgenic mice. *Gastroenterology*, 122(3):689–696, 2002.

K Sjölund, G Sandén, R Håkanson, and F Sundler. Endocrine Cells in Human Intestine: An Immunocytochemical Study. *Gastroenterology*, 85(5):1120–1130, 1983.

Gordon K Smyth. Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments. *Statistical Applications in Genetics and Molecular Biology*, 3(1):1–25, 2011.

Bastiaan Spanjaard, Bo Hu, Nina Mitic, Pedro Olivares-Chauvet, Sharan Janjuha, Nikolay Ninov, and Jan Philipp Junker. Simultaneous lineage tracing and cell-type identification using CRISPR–Cas9-induced genetic scars. *Nature Biotechnology*, 36(5):469–473, 2018.

David I Spivak. *Metric realization of fuzzy simplicial sets*. Self published notes, 2009.

Deepak Srivastava and Natalie DeWitt. In Vivo Cellular Reprogramming: The Next Generation. *Cell*, 166(6):1386–1396, 2016.

Martin Stahl, Sarah Tremblay, Marinieve Montero, Wayne Vogl, Lijun Xia, Kevan Jacobson, Alfredo Menendez, and Bruce A Vallance. The Muc2 mucin coats murine Paneth cell granules and facilitates their content release and dispersion. *American Journal of Physiology-Gastrointestinal and Liver Physiology*, 315(2):G195–G205, 2018.

Anders Ståhlberg and Martin Bengtsson. Single-cell gene expression profiling using reverse transcription quantitative real-time PCR. *METHODS*, 50:282–288, 2010.

Oliver Stegle, Leopold Parts, Richard Durbin, and John Winn. A Bayesian Framework to Account for Complex Non-Genetic Factors in Gene Expression Levels Greatly Increases Power in eQTL Studies. *PloS Computational Biology*, 6(5):e1000770–11, 2010.

Oliver Stegle, Leopold Parts, Matias Piipari, John Winn, and Richard Durbin. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nature protocols*, 7(3):500–507, 2012.

Oliver Stegle, Sarah A Teichmann, and John C Marioni. Computational and analytical challenges in single-cell transcriptomics. *Nature Methods*, 16(3):133–145, 2015.

M Stoeckius, C Hafemeister, W Stephenson, Brian Houck-Loomis, Pratip K Chattopadhyay, Harold Swerdlow, Rahul Satija, and Peter Smibert. Simultaneous epitope and transcriptome measurement in single cells. *Nature Methods*, 14:865–868, 2017.

Kelly Street, Davide Risso, Russell B Fletcher, Diya Das, John Ngai, Nir Yosef, Elizabeth Purdom, and Sandrine Dudoit. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics*, 19(1):477, 2018.

Emma J Stringer, Isabelle Duluc, Thoueiba Saandi, Irwin Davidson, Monika Bialecka, Toshiro Sato, Nick Barker, Hans Clevers, Catrin A Pritchard, Doug J Winton, Nicholas A Wright, Jean-Noël Freund, Jacqueline Deschamps, and Felix Beck. Cdx2

determines the fate of postnatal intestinal endoderm. *Development*, 139(3):465–474, 2012.

John E Sulston, Einhard Schierenberg, John G White, and James N Thomson. The embryonic cell lineage of the nematode Caenorhabditis elegans. *Developmental Biology*, 100(1):64–119, 1983.

Atsushi Suzuki, Sayaka Sekiya, Eriko Gunshima, Setsuko Fujii, and Hideki Taniguchi. EGF signaling activates proliferation and blocks apoptosis of mouse and human intestinal stem/progenitor cells in long-term monolayer cell culture. *Laboratory Investigation*, 90 (10):1425–1436, 2010.

Valentine Svensson, K N Natarajan, L H Ly, and R J Miragaia. Power analysis of single-cell RNA-sequencing experiments. *Nature*, 2017.

Fuchou Tang, Catalin Barbacioru, Yangzhou Wang, Ellen Nordman, Clarence Lee, Nanlan Xu, Xiaohui Wang, John Bodeau, Brian B Tuch, Asim Siddiqui, Kaiqin Lao, and M Azim Surani. mRNA-Seq whole-transcriptome analysis of a single cell. *Nature Methods*, 6(5):377–382, 2009.

Fuchou Tang, Catalin Barbacioru, Siqin Bao, Caroline Lee, Ellen Nordman, Xiaohui Wang, Kaiqin Lao, and M Azim Surani. Tracing the Derivation of Embryonic Stem Cells from the Inner Cell Mass by Single-Cell RNA-Seq Analysis. *Stem Cell*, 6(5):468–478, 2010.

Howard M Temin and Satoshi Mizutani. RNA-dependent DNA Polymerase in Virions of Rous Sarcoma Virus. *Nature*, 226:1211–1213, 2004.

Joshua B Tenenbaum, Vin de Silva, and John C Langford. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science*, 290(5500):2319–2323, 2000.

Mingxiang Teng, Michael I Love, Carrie A Davis, Sarah Djebali, Alexander Dobin, Brenton R Graveley, Sheng Li, Christopher E Mason, Sara Olson, Dmitri Pervouchine, Cricket A Sloan, Xintao Wei, Lijun Zhan, and Rafael A Irizarry. A benchmark for RNA-seq quantification pipelines. *Genome Biology*, 17:74, 2016.

Marie-Pier Tetreault, Daniel Weinblatt, Khvaramze Shaverdashvili, Yizeng Yang, and Jonathan P Katz. KLF4 transcriptionally activates non-canonical WNT5A to control epithelial stratification. *Scientific Reports*, 6:26130, 2016.

The Tabula Muris Consortium. Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature*, 562:367–372, 2018.

Itay Tirosh, Benjamin Izar, Sanjay M Prakadan, Marc H Wadsworth, Daniel Treacy, John J Trombetta, Asaf Rotem, Christopher Rodman, Christine Lian, George Murphy, Mohammad Fallahi-Sichani, Ken Dutton-Regester, Jia-Ren Lin, Ofir Cohen, Parin Shah, Diana Lu, Alex S Genshaft, Travis K Hughes, Carly G K Ziegler, Samuel W Kazer, Aleth Gaillard, Kellie E Kolb, Alexandra-Chloé Villani, Cory M Johannessen,

Aleksandr Y Andreev, Eliezer M Van Allen, Monica Bertagnolli, Peter K Sorger, Ryan J Sullivan, Keith T Flaherty, Dennie T Frederick, Judit Jané-Valbuena, Charles H Yoon, Orit Rozenblatt-Rosen, Alex K Shalek, Aviv Regev, and Levi A Garraway. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science*, 352(6282):189–196, 2016.

Harry Towbin, Theophil Staehelin, and Julian Gordon. Electrophoretic transfer of proteins from polyacrylamide gels to nitrocellulose sheets:P rocedure and some applications. *Proceedings of the National Academy of Sciences*, 76(9):4350–4354, 1979.

V A Traag, L Waltman, and N J van Eck. From Louvain to Leiden: guaranteeing well-connected communities. *arXiv*, 2018.

Cole Trapnell, Davide Cacchiarelli, Jonna Grimsby, Prapti Pokharel, Shuqiang Li, Michael Morse, Niall J Lennon, Kenneth J Livak, Tarjei S Mikkelsen, and John L Rinn. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature Biotechnology*, 32(4):381–386, 2014.

P Y Tung, J D Blischak, C J Hsiao, and D A Knowles. Batch effects and the effective design of single-cell gene expression studies. *Scientific Reports*, 2017.

Catalina A Vallejos, Sylvia Richardson, and John C Marioni. Beyond comparisons of means: understanding changes in gene expression at the single-cell level. *Genome Biology*, 17:70, 2016.

Catalina A Vallejos, Davide Risso, Antonio Scialdone, Sandrine Dudoit, and John C Marioni. Normalizing single-cell RNA sequencing data: challenges and opportunities. *Nature Methods*, 14(6):565–571, 2017.

Laurens G van der Flier and Hans Clevers. Stem cells, self-renewal, and differentiation in the intestinal epithelium. *Annual Review of Physiology*, 71(1):241–260, 2009.

Laurens van der Maaten and Geoffrey Hinton. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9:1–27, 2008.

David van Dijk, Roshan Sharma, Juozas Nainys, Kristina Yim, Pooja Kathail, Ambrose J Carr, Cassandra Burdziak, Kevin R Moon, Christine L Chaffer, Diwakar Pattabiraman, Brian Bierie, Linas Mazutis, Guy Wolf, Smita Krishnaswamy, and Dana Peer. Recovering Gene Interactions from Single-Cell Data Using Data Diffusion. *Cell*, pages 1–42, 2018.

Beate Vieth, Christoph Ziegenhain, Swati Parekh, Wolfgang Enard, and Ines Hellmann. powsimR: power analysis for bulk and single cell RNA-seq experiments. *Bioinformatics*, 33(21):3486–3488, 2017.

Oksana Voloshanenko, Uwe Schwartz, Dominique Kranz, Benedikt Rauscher, Michael Linnebacher, Iris Augustin, and Michael Boutros. β-catenin-independent regulation of

Wnt target genes by RoR2 and ATF2/ATF4 in colon cancer cells. *Scientific Reports*, pages 1–14, 2018.

Ulrike von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4): 395–416, 2007.

Conrad Hal Waddington. *The strategy of the genes*. George Allen and Unwin, London, 1957.

Daniel E Wagner, Caleb Weinreb, Zach M Collins, James A Briggs, Sean G Megason, and Allon M Klein. Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. *Science*, 360(6392):981–987, 2018.

Günter P Wagner, K Kin, and Vincent J Lynch. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory in Biosciences*, 131(4):281–285, 2012.

Abraham Wald. Tests of Statistical Hypotheses Concerning Several Parameters When the Number of Observations is Large. *Transactions of the American Mathematical Society*, 54(3):426–482, 1943.

Joe H Ward Jr. Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, 58(301):236–244, 1963.

Caleb Weinreb, Samuel Wolock, Betsabeh K Tusi, Merav Socolovsky, and Allon M Klein. 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, 2018.

Sophie Weiss, Zhenjiang Zech Xu, Shyamal Peddada, Amnon Amir, Kyle Bittinger, Antonio Gonzalez, Catherine Lozupone, Jesse R Zaneveld, Yoshiki Vázquez-Baeza, Amanda Birmingham, Embriette R Hyde, and Rob Knight. Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome*, 5:27, 2017.

B L Welch. The Generalization of 'Student's' Problem when Several Different Population Variances are Involved. *Biometrika*, 34(1-2):28–35, 1947.

Joshua D Welch, Alexander J Hartemink, and Jan F Prins. MATCHER: manifold alignment reveals correspondence between single cell transcriptome and epigenome dynamics. *Genome Biology*, 18:138, 2017.

Michael L Whitfield, Gavin Sherlock, Alok J Saldanha, John I Murray, Catherine A Ball, Karen E Alexander, John C Matese, Charles M Perou, Myra M Hurt, Patrick O Brown, and David Botstein. Identification of Genes Periodically Expressed in the Human Cell Cycle and Their Expression in Tumors. *Molecular Biology of the Cell*, 13(6):1977–2000, 2002.

S S Wilks. The Large-Sample Distribution of the Likelihood Ratio for Testing Composite

Hypotheses. *The Annals of Mathematical Statistics*, 9(1):60–62, 1938.

Christian Wiwie, Jan Baumbach, and Richard Röttger. Comparing the performance of biomedical clustering methods. *Nature Methods*, 12(11):1033–1038, 2015.

F Alexander Wolf, Philipp Angerer, and Fabian J Theis. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biology*, 19:15, 2018a.

F Alexander Wolf, Fiona Hamey, Mireya Plass, Jordi Solana, Joakim S Dahlin, Berthold Göttgens, Nikolaus Rajewsky, Lukas Simon, and Fabian J Theis. Graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *bioRxiv*, 2018b.

Samuel L Wolock, Romain Lopez, and Allon M Klein. Scrublet: computational identification of cell doublets in single-cell transcriptomic data. *bioRxiv*, 2018.

Angela R Wu, Norma F Neff, Tomer Kalisky, Piero Dalerba, Barbara Treutlein, Michael E Rothenberg, Francis M Mburu, Gary L Mantalas, Sopheak Sim, Michael F Clarke, and Stephen R Quake. Quantitative assessment of single-cell RNA-sequencing methods. *Nature Methods*, 11:41–46, 2013.

Jingyi Wu, Bo Huang, He Chen, Qiangzong Yin, Yang Liu, Yunlong Xiang, Bingjie Zhang, Bofeng Liu, Qiujun Wang, Weikun Xia, Wenzhi Li, Yuanyuan Li, Jing Ma, Xu Peng, Hui Zheng, Jia Ming, Wenhao Zhang, Jing Zhang, Geng Tian, Feng Xu, Zai Chang, Jie Na, Xuerui Yang, and Wei Xie. The landscape of accessible chromatin in mammalian preimplantation embryos. *Nature*, 534(7609):652–657, 2016.

Jun Wu and Juan Carlos Izpisua Belmonte. Stem Cells: A Renaissance in Human Biology Research. *Cell*, 165(7):1572–1585, 2016.

Changxin Xu, Xiufen Li, Matthew K Topham, and Scott K Kuwada. Regulation of sonic hedgehog expression by integrin $\beta 1$ and epidermal growth factor receptor in intestinal epithelium. *IUBMB Life*, 66(10):694–703, 2014.

Zhigang Xue, Kevin Huang, Chaochao Cai, Lingbo Cai, Chun-yan Jiang, Yun Feng, Zhenshan Liu, Qiao Zeng, Liming Cheng, Yi E Sun, Jia-yin Liu, Steve Horvath, and Guoping Fan. Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing. *Nature*, 500(7464):593–597, 2013.

Kelley S Yan, Olivier Gevaert, Grace X Y Zheng, Benedict Anchang, Christopher S Probert, Kathryn A Larkin, Paige S Davies, Zhuan-fen Cheng, John S Kaddis, Arnold Han, Kelly Roelf, Ruben I Calderon, Esther Cynn, Xiaoyi Hu, Komal Mandleywala, Julie Wilhelmy, Sue M Grimes, David C Corney, Stéphane C Boutet, Jessica M Terry, Phillip Belgrader, Solongo B Ziraldo, Tarjei S Mikkelsen, Fengchao Wang, Richard J von Furstenberg, Nicholas R Smith, Parthasarathy Chandrakesan, Randal May, Mary Ann S Chrissy, Rajan Jain, Christine A Cartwright, Joyce C Niland, Young-Kwon Hong, Jill

Carrington, David T Breault, Jonathan Epstein, Courtney W Houchen, John P Lynch, Martin G Martin, Sylvia K Plevritis, Christina Curtis, Hanlee P Ji, Linheng Li, Susan J Henning, Melissa H Wong, and Calvin J Kuo. Intestinal Enteroendocrine Lineage Cells Possess Homeostatic and Injury-Inducible Stem Cell Activity. *Stem Cell*, 21(1):78–90, 2017.

Qi-Long Ying, Jason Wray, Jennifer Nichols, Laura Batlle-Morera, Bradley Doble, James Woodgett, Philip Cohen, and Austin Smith. The ground state of embryonic stem cell self-renewal. *Nature*, 453:519–523, 2008.

Matthew D Young and Sam Behjati. SoupX removes ambient RNA contamination from droplet based single cell RNA sequencing data. *bioRxiv*, 2018.

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

Luke Zappia, Belinda Phipson, and Alicia Oshlack. Splatter: simulation of single-cell RNA sequencing data. *Genome Biology*, 18:174, 2017.

Luke Zappia, Belinda Phipson, and Alicia Oshlack. Exploring the single-cell RNA-seq analysis landscape with the scRNA-tools database. *PloS Computational Biology*, 14(6): e1006245, 2018.

Jesse Min Zhang, Govinda M Kamath, and David N Tse. Towards a post-clustering test for differential expression. *bioRxiv*, 2018a.

Martin J Zhang, Vasilis Ntranos, and David Tse. One read per cell per gene is optimal for single-cell RNA-Seq. *bioRxiv*, 2018b.

Grace X Y Zheng, Jessica M Terry, Phillip Belgrader, Paul Ryvkin, Zachary W Bent, Ryan Wilson, Solongo B Ziraldo, Tobias D Wheeler, Geoff P McDermott, Junjie Zhu, Mark T Gregory, Joe Shuga, Luz Montesclaros, Jason G Underwood, Donald A Masquelier, Stefanie Y Nishimura, Michael Schnall-Levin, Paul W Wyatt, Christopher M Hindson, Rajiv Bharadwaj, Alexander Wong, Kevin D Ness, Lan W Beppu, H Joachim Deeg, Christopher McFarland, Keith R Loeb, William J Valente, Nolan G Ericson, Emily A Stevens, Jerald P Radich, Tarjei S Mikkelsen, Benjamin J Hindson, and Jason H Bielas. Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, 8:1–12, 2017.

Christoph Ziegenhain, Beate Vieth, Swati Parekh, Björn Reinius, Amy Guillaumet-Adkins, Martha Smets, Heinrich Leonhardt, Holger Heyn, Ines Hellmann, and Wolfgang Enard. Comparative Analysis of Single-Cell RNA Sequencing Methods. *Molecular Cell*, 65:631–643, 2017.