# Cutting edge tools to reveal lineages and epigenetic modifications in single cells

Maria Florescu

The work described in this thesis was performed at the Hubrecht Institute for Developmental Biology and Stem Cell Research (the Royal Netherlands Academy of Arts and Sciences, KNAW) within the framework of the research school Cancer Stem cells & Developmental biology (CS&D), which is part of the Utrecht Graduate School of Life Sciences (Utrecht University).

ISBN-13: 978-94-6375-998-4


Illustrations by Sarah Bronder | mail@sarahbronder.com

Printed by Ridderprint | www.ridderprint.nl

# Cutting edge tools to reveal lineages and epigenetic modifications in single cells

**Innovatieve technologien om cell historie en epigenetische modificaties in enkele cellen te onthullen**
(met een samenvatting in het Nederlands)

**Proefschrift**

ter verkrijging van de graad van doctor aan de Universiteit Utrecht op gezag van de rector magnificus, prof. dr. H. R. B. M. Kummeling, ingevolge het besluit van het college voor promoties in het openbaar te verdedigen op

dinsdag 29 september 2020 des ochtends te 11.00 uur

door

Maria Florescu

geboren op 31 oktober 1989
te Buzău, Roemenië

Promotor: Prof. dr. ir. A. van Oudenaarden

# Contents

Chapter 1

# Introduction

Centuries ago, single units we now know as cells were first described by Hooke and Leeuwenhoek in multi-cellular organisms. Fast forward a few centuries, in the 1950s the molecular structure of DNA was identified establishing the link between heredity and genetic material. Since then, scientists have been decoding genetic information to better understand a cell's past, present and future. Single cell biology aims at measuring and quantifying multiple molecular layers in snapshots of time in order to establish not only how different single cells are at a current timepoint, but read out their molecular past and eventually predict their future. In order to achieve this goal, different biomolecules have to be measured in the same single cells with highly sensitive technologies. The technology tool box needs to provide accurate measurements, for example when detecting different amount of RNA molecules. In addition, measurements need to be precise, for example when detecting absence or presence of DNA loci, since these are usually encoded in two copies. Therefore technologies are the limiting factor for how accurately and precisely we can describe a single cell within a complex context such as development or disease. In this thesis I will present and discuss recent technological advancements in the field of single cell biology. I am to illustrate how these advances in single cell sequencing technologies can help shaping a more complete picture of a cell's inner mechanics.

## Beyond single cell transcriptomics

The goal of single cell sequencing analysis it to quantify at a genome-wide scale the relationships between biomolecules in the context of tissue composition (cellular heterogeneity) as well as display cellular dynamics, for example during embryonic development or during an immune response[60,123]. Even though single cell sequencing relies on snapshots of biomolecules in time, studies have shown reliable recovery of cellular differentiation processes or build pseudo-time trajectories[133]. Over the last decade, single cell methods were able to identify with increasing sensitivity a single modality, such as RNA expression, chromatin accessibility or whole genome sequencing. The most prominent method so far is measuring gene expression profiles with single cell RNA-sequencing (scRNA-seq) which has been widely adopted and optimised for sensitive, highly multiplexed or combinatorially barcoded transcriptome profiling[20,42,43,89,98]. With increasing technological ease came the profiling of millions of cells capturing heterogeneity and intermediate states along differentiation trajectories in an unbiased way leading to single cell atlases of tissues[89] and developmental trajecto-

ries[21,34,99]. Quantifying gene expression by measuring RNA molecules with high-throughput scRNA-seq technologies has helped to characterize new (sub-) cell types, providing insight into their potential fate[67] and function. scRNA-seq advances have been closely followed by complementary single-cell measurements related to DNA such as sequence, 3D structure, methylation and histone modifications. Usually, these DNA feature are studied in extracted cells or in disease context, for example through studying genomic instability and variation in cancer and lack high-throughput cell profiling and unbiased cell types identification.

Multimodal profiling (multi-omics) aims at integrating multiple single cell read-outs and has been on the horizon for years[29,30], but recent advances have made the simultaneous profiling of multiple biomolecules more accessible and prominent. Most popular is combining scRNA-seq profiling to identify cell types, together with, for example, protein abundance or DNA accessibility[124]. This illustrates that currently a cell's transcriptome profile is used as the primary link and reference for integrating and interpreting other modalities. Technically, four broad strategies are associated with multimodal data generation in single cells[123]:

1. Non-destructive assays before sequencing (e.g. FACS parameters)
2. Physical separation of different cellular fractions and molecules (e.g. scNMT-seq[29])
3. Conversion of different readouts into one common molecular readout (e.g. a genomic readout which is being transcribed as in GESTALT[103])
4. Analytical decoding of multiple layers of information from nucleotide sequences (e.g. somatic mutations, splicing dynamics)

Here we will focus on methods following the last strategy through integrating multiple biochemical approaches within the same single cell in order to extract different layers of information. A technical distinction for such an analytical decoding strategy is whether the different biochemical readouts are separately barcoded such as in ScarTrace (**chapter 3**) or not, such as in scChIX (**chapter 4**). For a successful readout of multiple modalities

within a single cell, one can leave a barcode marking the individual layers or computationally disentangle a signal of multiple modalities.

Combining, for example, genome sequencing together with transcriptome sequencing can relate genetic heterogeneity to cell type information, thereby linking genetic variation such as DNA copy number variation to transcriptional variation. Ideally, these advances in multi-omics will be able to decipher a cell's past, present and future including the positional information of the measured biomolecules and environmental interactions.

## Lineage tracing

Lineage tracing technologies enable to identify a cell's progeny during a defined time frame. Starting off more than a century ago as a technique for studying embryonic development, lineage tracing technologies are nowadays additionally used for stem cell homeostasis and differentiation, regenerative capacities, cancer progression, immune responses, ageing processes and phylogeny[63]. To cover a wide variety of biological applications different lineage tracing technologies have been developed, which vary in their resolution regarding the traceable number of unique clones and time of operation. From a practical point of view there are two branches of lineage tracing technologies, one being microscopy and the other being sequencing based lineage tracing technologies.

Microscopy based lineage tracing technologies have been applied and improved for decades leading to the first full organism lineage tree of *C. elegans*[126], a milestone in developmental biology. As imaging technologies as well as dye labelling and fluorescent reporters advanced, other popular and more complex vertebrates systems such as frog, chicken and zebrafish were studied. The development of these systems are all *ex-utero* and therefore optically easier to image. Advances in microscopy technologies in combination with dye labelling and culturing techniques have enabled difficult studies such as cellular tracing of neural progenitors in the forebrain during late mouse gastrulation[19]. Recent breakthroughs have enabled to use adaptive light-sheet imaging ( i.e. *in toto* imaging) during two days of mammalian gastrulation which not only requires sophisticated *ex-utero* culturing techniques but also advanced computational frameworks and optical solutions to cope with the dynamics of space and time[82].

Advantages of imaging over sequencing methods is evidently the maintenance of spatial information and the tracking of almost every cell division

of the system under study. Additionally cell death at certain time or location as well as migration patterns during dynamic developmental periods and cell-cell interactions can be tracked[126].
However, the relatively short imaging timeframe as well as technical difficulties of scaling to millions of single cells, is a disadvantage. Fluorescent gene reporter systems have been used *in vivo* to track specific cell types, however it remains challenging to up-scale to multiple reporters in order to distinguish between all cell types within an embryo. Cell types are defined by co-expression of several (marker) genes which change dynamically over the course of a few cell division during embryogenesis and can be therefore more completely described through scRNA-seq.

With the possibility for genetic readout and manipulation, sequencing based approaches for lineage tracing have been developed. A genetic lineage tracing technology needs to meet 5 basic requirements in order to provide clonal information for building a developmental lineage tree:

1. The labels need to be unique in order to mark one clone. This can be achieved through multiple barcodes with a combination uniquely labelling one clone or by one very complex label.
2. The labels need to be permanent.
3. The clonal information needs to be inheritable over cell divisions.
4. The labels need to be uniformly introduced over time.
5. The labels need to be cumulative or display a hierarchial order.

While engineered lineage recorders are designed to not have a reversible barcode and to label cells uniquely, endogenous lineage tracing marks are assumed to have a back-mutation rate and can occur in multiple clones. However, in human or cancer studies endogenous lineage tracing markers such as single nucleotide variants (SNV) or copy number variants (CNV) can be useful for lineage recording (Figure 1.1). In mouse pre-implantation embryos it has been shown that derivatives of 5 methyl cytosine (5mC) can be used to trace sister cells back over a few cell divisions[86]. Here, oxidation of 5mC contributes to the active de-methylation of the paternal chromosomes[86], introducing a strand bias in sister cells.

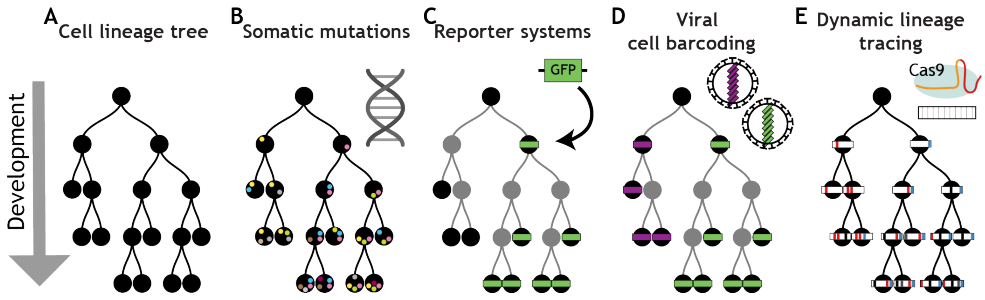



Figure 1.1: **Overview of developmental lineage trees based on genetic lineage tracing (adapted from [83]).** (A) Schematic of ground truth developmental lineage tree. (B) Schematic of lineage tree from somatic mutations (SNV) as endogenous lineage label. (C) Schematic of lineage tree which can be obtained through a GFP-reporter in a specific tree branch. (D) Schematic of lineage tree from viral barcoding at one specific time-point during development. (E) Dynamics lineage tree with CRISPR/Cas9 recorders.

In systems that can be manipulated lineage recorders have been introduced which vary in their recording time and resolution (Figure 1.1). Using viral barcoding strategies can result in a very high complexity of tens of thousands of uniquely labeled cells. Viral labelling is mostly used for one timepoint and it is rarely being used for multiple timepoints. DNA modifications can also be artificially introduced to follow certain cell lineages. Through introduction of normally not maintained DNA modifications (5hmC, 5fC, 5caC) into a particular chromosome or DNA strand one can trace its segregation into a particular cell type. CRISPR/Cas9 based recorders allow for a dynamic cell labelling over a longer period of time (Figure 1.1). Ultimately, the goal of lineage tracing technologies is to provide enough clonal resolution to eventually build a cellular lineage tree from the labeled clones. As genetic recorders and microscopy technologies have complementary advantages, it is best to validate biological questions with both approaches.

## Lineage tracing with CRISPR/Cas9

The CRISPR (Clustered Regularly Interspaced Short Palindromic Repeat) system has been proven to be a versatile and revolutionising tool beyond genome editing. Since the first description of the CRISPR locus 15 years ago[13,85,102] as an adaptive immune system in bacteria against bacteriophages, the CRISPR system has been an intense focus of research and engineering efforts. So far twenty different Cas proteins have been described and classified into three major classes. In bacteria, the most compact CRISPR system

pathogen response is complete when combining Cas9 protein together with a CRISPR RNA (crRNA) and a trans-activating RNA (tracrRNA). However, the system has been further simplified[50] to a two-component system which is now widely known as a single synthetic guide RNA (sgRNA = crRNA + tracrRNA) and Cas9 protein which is part of the class II CRISPR nuclease family. Cas9 protein consists of a single, large multi-functional protein, in contrast to other classes, where multiple large protein complexes form a functional unit.

In eurkaryotic cells, the Cas9 + sgRNA complex is used to target a genomic nucleotide sequence of choice which is complementary to the sequence encoded in the crRNA. At the targeted DNA location Cas9 induces a double strand break (DSB). DSB can be repaired through multiple pathways : canonical NHEJ (c-NHEJ), alternative NHEJ (alt-NHEJ) and homologous recombination (HR). Depending on what induced the break, its location and cell cycle stage a repair pathway is selected. HR is activated when a homologous region is available. Also, heterochromatic DSBs, for example, are primarily repaired by HR in *drosophila melanogaster*[106]. When the NHEJ repair pathway is activated short mutations in the form of indels (small insertions/deletions) are introduced during the repair (Figure 1.2a).

The CRISPR system has been widely adopted in various organisms and Cas proteins have been engineered to be light inducible[93], have nickase activity as well as for gene activation and repression. For prospective lineage tracing purposes, we make use of the CRISPR/Cas9 system. The core idea around all CRISPR/Cas9 methods, from now on referred to as CRISPR recorders, for lineage tracing is that the NHEJ repair-dependent pathway introduces random indels at specified locus which can be read out as a unique barcode through sequencing.

For lineage tracing, clones are labeled through multiple of these generated barcodes which increases the timeframe and resolution of the labeled clones. CRISPR recorders are popular for usage in *in vivo* systems as artificial cassettes or non-coding regions can be targeted without introducing developmental effects. Compared to other sequencing based technologies for lineage tracing such as Cre-LoxP lineage tracing systems, CRISPR/Cas9-induced barcodes are short sequences which can either be expressed through an endogenous promoter[103] or directly PCR amplified[3,54,84]. This technical detail makes it straight forward to combine CRISPR recorders with scRNA-seq (or other molecular single cell readouts).

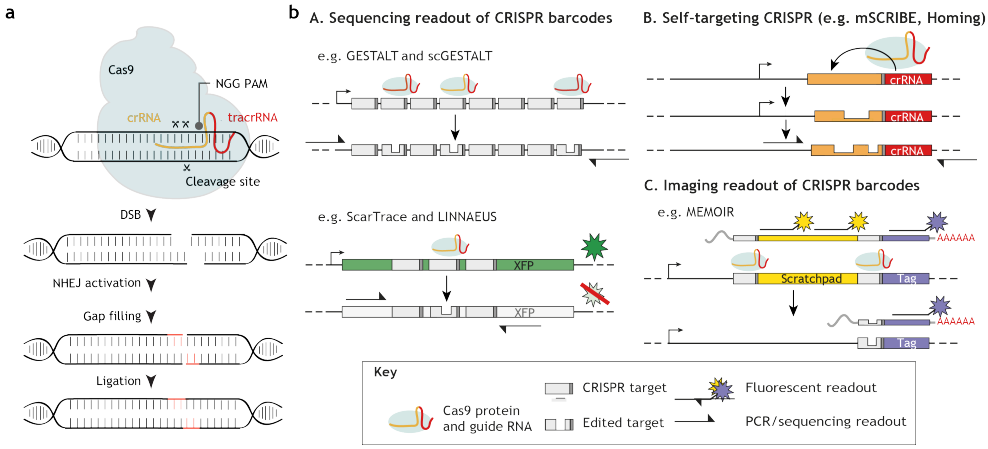



Figure 1.2: **Overview of lineage tracing methods with CRISPR/Cas9. a** Schematics of CRISPR/Cas9 induced DBS resulting in one barcode through indels. **b** CRISPR recorders (adapted from [83]). (A) Initial design, whereby multiple sequences are edited with a single or multiple gRNAs. Edited regions are read out by sequencing. (B) Multiple rounds of editing the same genomic sequence through self-targeting CRISPR guide RNAs. Edited regions are read out by sequencing (C) So called 'scratchpads' regions are inserted and collapsed into the genome by CRISPR editing. Subsequently these 'scratchpads' are expressed and detected through in situ hybridization probes.

A variety of CRIPSR/Cas9 recorders have been developed for mouse, zebrafish or cell lines (Figure 1.2b). Systems such as GESTALT [84,103], MEMOIRE [35] and homing-CRISPR [54] are stably integrated in the organism, whereas LINNEUS [119] and ScarTrace [3] deliver CRISPR/Cas9 through injection (Figure 1.2b). Further, droplet-based approaches such as GESTALT and LINNEUS rely on expression of the barcode and its amplification with the transcriptome. While this results in a higher drop-out rate of the barcode, droplet-based methods are high throughput methods which can handle tens of thousands of cells. While ScarTrace is a less high-throughput plate-based approach of FACS sorted cells it separately amplifies evolving barcodes and transcriptome resulting in a lower drop-out rate of barcodes.

The wide usage of CRISPR/Cas9 for lineage tracing has increased knowledge on the system and as a consequence, it became more easy to manipulate. We know now that the time resolution of barcode introduction can not only be manipulated by injection of Cas9 protein or Cas9 RNA, but also by the amount of gRNA as well as the length of the gRNA [54].

Generally, the strength of CRISPR/Cas9 recorders is in the vast possibilities of combining cellular lineage tracing *in vivo* with not only scRNA-seq, but

also other single cell genomics read-outs.

## Chromatin profiling in single cells

Chromatin modifications play a vital role in regulating and maintaining cellular identity through recruitment of proteins to the genome and therefore directly connect the genome with it functional output. Changes in chromatin regulators or chromatin components have been shown to result in a strong phenotype. Evolving from prokaryotic chromatin, eukaryotic chromatin has nucleosome complexes, with histone subunits, and has diverse functions. Besides their major association to nucleosomes which help to densely pack eukaryotic genomes, histones are regulators of gene expression in development, differentiation, metabolism and environmental stimuli. In eukaryotic cells DNA is densely packed around two copies of four histone proteins, H2A, H2B, H3 and H4, and hold together by a linker histone H1/H5. Histones can be modulated through a wide range of post-translational modifications at their four tails. These post-translational histone modifications are linked to gene regulation, in either a repressive or activating way. These include (de-)acetylation, (de-)methylation, (de-)phosphorylation, (de-)ubiquitination and (de-)sumoylation of the histone tail (Figure 1.3a).

Histone modifications are crucial for establishing a cell identity during development and disease as well as for maintaining cell identity.

The best studied post-translational histone modifications are acetylation and methylation. These modifications are often mutually exclusive since they have many overlapping binding sites at Arginine and Lysine residues[115]. While acetylated residues are mainly associated with open chromatin and transcription, methylated residues are associated with gene activation as well as repression dependent on their location. Lysine residues have been found to be monomethylated, dimethylated, or trimethylated.

For years, histone modifications have been studied in population of cells with chromatin immuno precipitation (ChIP) or at the genome-wide level with ChIP-seq. Results from these extensive studies categorised and linked various histone modifications to their function. Further, genome-wide studies with ChIP-seq have helped to associate histone modifications with occupancy of certain genomic regions, for example, genebodies or promoters (Figure 1.3b). The well-described histone modifications which are also described in this thesis are all marks associated with methylation (me) of the

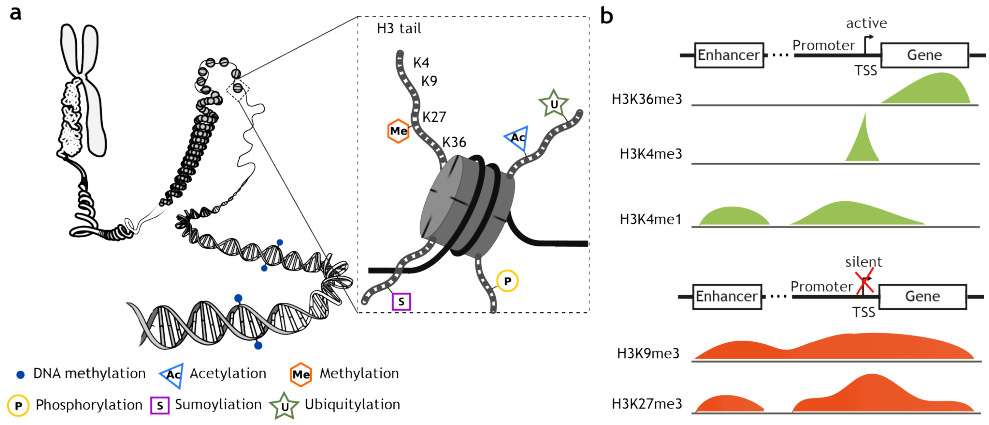



Figure 1.3: **Overview of histone modifications and their genomic distribution.a** Schematics of chromatin on a chromosome. Box highlights, as an example, one histone and commonly studied Lysine residues on the tail of histone 3. **b** Schematics of the genomic distribution of a few commonly studied histone modifications. Some genome features such as enhancers, promoters, transcription start site (TSS) and gene bodies are sketched above the examples of the genome distribution which are associated with active chromatin (green) or with inactive chromatin (red).

lysine-rich (K) tail of histone H3 (Figure 1.3b). H3K9me3 is associated primarily with closed heterochromatin, which are large genomic regions inaccessible for transcription. H3K27me3 is mainly found on transcriptionally inactive regions. In short, H3K9me3 and H3K27me3 can be called repressive histone modifications. In contrast, H3K4me1 is associated with accessible chromatin and can be linked to active transcription. H3K36me3 is a histone modification deposited by a subunit of the RNA polymerase II complex during gene transcription. H3K4me1 and H3K36me3 can hence be called active histone modifications.

ChIP-seq has been a reliable and popular method to profile populations of cells for certain chromatin modifications, but the crucial dependence on antibody binding to pull down modified histone together with associated DNA is inefficient in low-input material, such as in single cells. Recently droplet based or plate based methods are labelling the location of a histone modification of choice by making use of transposase tagging (Cut & Tag) or micrococcal nuclease digestion with subsequent barcoding (Cut & Run, scChIC-seq, CoBATCH[140]) (Figure 1.4).

Pushing the boundaries in chromatin profiling at the single cell level with multiple histone profiling or integrate with scRNA-seq is crucial to identify

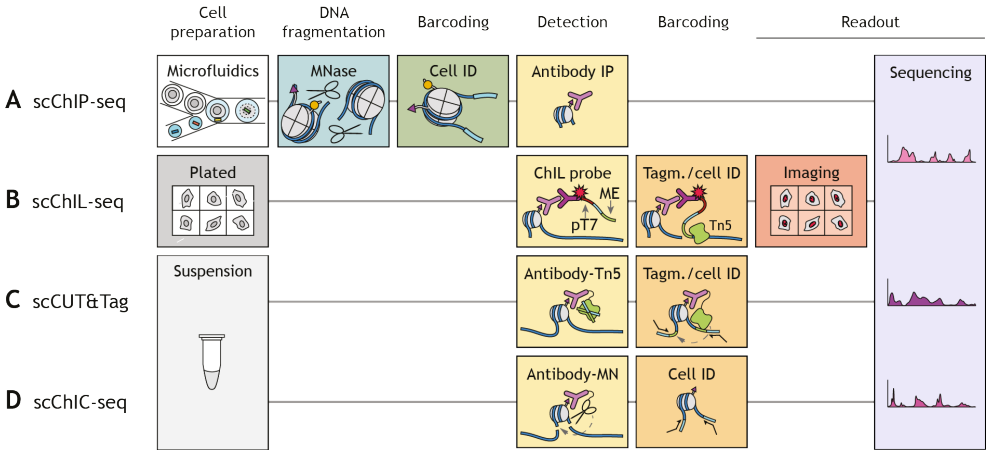



Figure 1.4: **Sequencing-based chromatin profiling in single cells starting from a pool of dissociated cells (adapted from [75]).** (A) Individual cells are encapsulate in droplet reaction chambers for DNA fragmentation by micrococcal nuclease (MNase) and ligation of cell-specific identifiers. Barcoded cells are pooled for immunoprecipitation (IP). Subsequentlly library preparation and sequencing are performed. (B) Fixed single cells are deposited in individual micro-wells. A primary antibody against a histone modification of choice is detected by a chromatin integration labelling (ChIL) probe which consists of a secondary antibody conjugated to fluorophore-labeled double-stranded DNA. This probe is imaged and integrated into the sequence adjacent to the modified nucleosome via Tn5 transposase-binding. Well-specific ( = cell-specific) barcodes are introduced during library preparation (not shown). (C) An antibody-Tn5 or a protein A-Tn5 fusion associates with a pre-incubated antibody to insert pre-complexed adapters near modified nucleosomes upon magnesium addition. The cell suspension is then sorted into nanowells. In each nanowell, DNA is fragmented upon Tn5 association and library preparation is performed whereby cell-specific barcodes are introduced. (D) Cell suspension of fixed or unfixed cells are incubated with an antibody-MNase or protein A-MNase fusion that associates with a pre-incubated antibody against a histone modification of choice. Single cells are sorted into nanowells or tubes and calcium is added to trigger MNase-mediated fragmentation. Subsequent library preparation incorporates cell-specific barcodes.

how heterogenous chromatin states are related to transcriptionally heterogeneous cells.

## Multi-modal profiling during embryonic development

Embryonic development must be the most exciting time in the life of a cell. Continuous changes in molecules and cellular movement lead to cell lineage specification, establishment of the the body axis and ultimately to complex organ development and a grown organism. Seemingly identical

looking cells during early embryonic stages are propagating and specifying into a particular cell type over time.

Even though vertebrates have a defined embryonic developmental plan, the timing and order of events vary across organisms. Popular embryonic model systems to study crucial events such as axis formation and gastrulation are zebrafish (*danio rerio*) and mouse (*mus musculus*). Great scientific efforts have lead to a profound understanding of embryonic development in these two model systems. Studying embryonic development brings a few pitfalls such as, for example, seemingly similar cells. This could be resolved through of multi-modal profiling within single cells. Multi-modal profiling can help to define the properties of a single cell more clearly at early embryonic stages where cells are seemingly alike, thereby providing crucial information about its cellular past or future. As an example, organizers have been described early on during transplantation experiments, as a group of cells inducing gradients of signalling molecules leading to movements of convergence and extension[81]. Organizers are to date still widely studied and better understood, but the underlying mechanisms of their formation remains to be described. Profiling multiple chromatin layers could help to find out how these organizers are formed. Further, embryonic cells have been shown to have characteristic molecular states, which are rarely found in adult cells, such as dual histone tail modification, termed bivalency. Profiling multiple histone modifications, across histone or on the same histone, might be a key aspect in understanding how and how long cells retain their plasticity.

## Scope and outline of the thesis

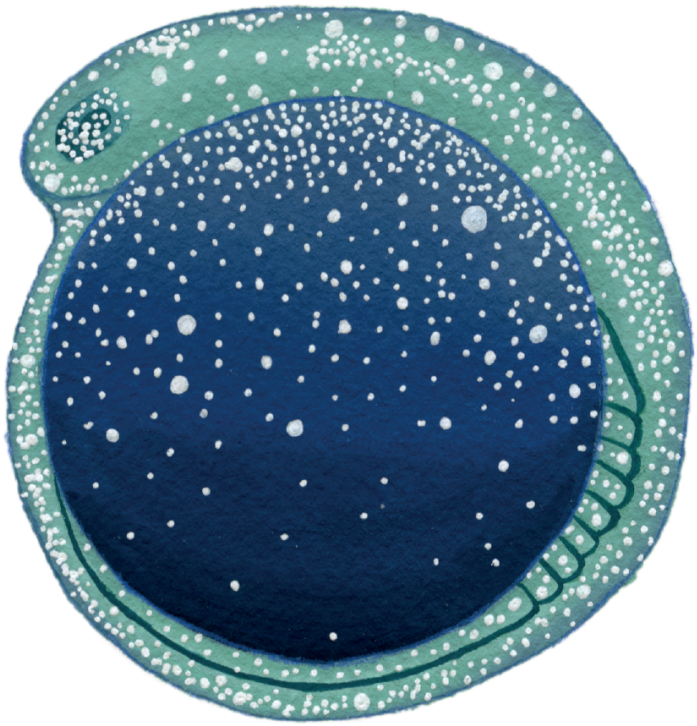
The tools presented in this thesis are primarily, but not exclusively, developed for revealing lineages and histone modifications during embryonic development. In **chapter 2** and **3**, ScarTrace is applied during zebrafish development. Our dual histone modification profiling through ChIX in **chapter 4** is showcased in adult bone marrow, but has been applied during mouse gastrulation (data not shown).

In *Chapter 2* we show a combined study of lineage tracing with imaging and a CRISPR recorder (ScarTrace) at tissue level during spinal cord formation in zebrafish. Our study suggests that there are two sub-populations of a bi-fated celltype called NMPs. A first population showing an early and limited contribution to spinal cord and muscle tissue and a second population, showing contribution to caudal-most regions of the tail.

In *Chapter 3* the development of ScarTrace in single cells is presented, whereby gene expression as well as clonal barcode information from the same cell are provided. Clonal tracing is starting within the first hours of zebrafish development and the combined information is read out in the adult zebrafish. Our study suggests that cells contributing to the left and right eye are separated before zygotic gene activation. Further we find a sub-population of tissue-resident macrophages.

Lastly, in *Chapter 4*, we introduce scChIX, a novel chromatin profiling framework for multiplexed histone modification readouts. We demonstrate that scChIX can separate a mixed signal of two histone modifications. Further, we use scChIX to transfer celltype labels from one (active) histone modification to single cells from another (repressive) histone modification.

I conclude this thesis in *Chapter 5* where I provide a general discussion of opportunities and challenges when using the presented technologies. I discuss a future outlook on how these technologies could be useful for studying embryonic development. Finally, I discuss the impact of multi-omics technologies determining celltype identity.



Chapter 2

# Neuromesodermal progenitors are a conserved source of spinal cord with divergent growth dynamics

A. Attardi[1,2,3,*], T. Fulton[1,*], M. Florescu[4], G. Shah[2,5], L. Muresan[6], M.O. Lenz[6], C. Lancaster[1], J. Huisken[2,7], A. van Oudenaarden[4] and B. Steventon[1]

published in *Development* 2018

*equal contribution
[1] Department of Genetics, University of Cambridge, Cambridge CB2 3EH, UK.
[2] Max Planck Institute of Molecular Cell Biology and Genetics, Dresden 01307, Germany.
[3] STEBICEF Department, Universita degli Studi di Palermo, Palermo 90133, Italy.
[4] Hubrecht Institute-KNAW (Royal Netherlands Academy of Arts and Sciences) and University Medical Center Utrecht, 3584 CT, Utrecht, The Netherlands.
[5] European Molecular Biology Laboratory, 08003 Barcelona, Spain.
[6] Cambridge Advanced Imaging Centre, Cambridge CB2 3EH, UK.
[7] Morgridge Institute for Research, Madison, WI 53715, USA.

During gastrulation, embryonic cells become specified into distinct germ layers. In mouse, this continues throughout somitogenesis from a population of bipotent stem cells called neuromesodermal progenitors (NMps). However, the degree of self-renewal associated with NMps in the fast-developing zebrafish embryo is unclear. Using a genetic clone-tracing method, we labelled early embryonic progenitors and found a strong clonal similarity between spinal cord and mesoderm tissues. We followed individual cell lineages using light-sheet imaging, revealing a common neuromesodermal lineage contribution to a subset of spinal cord tissue across the anterior-posterior body axis. An initial population subdivides at mid-gastrula stages and is directly allocated to neural and mesodermal compartments during gastrulation. A second population in the tailbud undergoes delayed allocation to contribute to the neural and mesodermal compartment only at late somitogenesis. Cell tracking and retrospective cell fate assignment at late somitogenesis stages reveal these cells to be a collection of mono-fated progenitors. Our results suggest that NMps are a conserved population of bipotential progenitors, the lineage of which varies in a species-specific manner due to vastly different rates of differentiation and growth.

## Introduction

In amniotes, a bipotent population of posterior progenitors continually allocate cells to the posterior pre-somitic mesoderm (PSM) and spinal cord[17,113,136]. Although no specific molecular marker has been identified for this population, they have been shown to express a combination of both the early neural and mesodermal markers Sox2 and brachyury[143] and have been named neuromesodermal progenitors (NMps;[45]). The observation of a NMp population is important as it suggests that germ layer specification continues throughout somitogenesis stages and is not restricted to primary gastrulation. Understanding when NMps allocate cells to spinal cord and paraxial mesoderm is an essential first step in exploring the underlying molecular processes that determine the timing of germ layer allocation during vertebrate embryonic development. The continuous allocation of cells from an NMp pool fits well with the cellular basis of axial elongation in mouse embryos, in which primary gastrulation generates only head structures and the rest of the body axis is generated by posterior growth[121]. However, externally developing embryos such as the zebrafish elongate their body axis in the absence of posterior volumetric growth, with elongation being a consequence of convergence and extension of mesodermal progenitors, and volumetric growth of tissue within the already segmented region of the body axis[122]. Despite these differences, genetic studies have revealed a marked conservation in the signals and gene regulatory networks that act to drive posterior body elongation across bilaterians[79]. Furthermore, experi-

ments in the zebrafish embryo have confirmed the presence of Sox2/Tbxta-positive cells in the tailbud, and transplantation of single cells into the marginal zone can be directed to either neural or mesodermal cell fates, depending on the level of canonical Wnt signalling[80]. An additional progenitor pool exists within the tailbud that generates cells of the notochord and floorplate in a Wnt- and Notch-dependent manner[105]. Although these studies demonstrate that a neuromesodermal competent population exists within the zebrafish tailbud, lineage analysis using intracellular injection of high molecular weight fluorescent dextran[55] argues against a stem cell-like population that is homologous to the mouse NMp pool[136]. These seemingly conflicting results can be resolved by a complete lineage analysis that determines the timing of neural and mesodermal lineage restriction in zebrafish, and answer in a clear manner whether these progenitors arise from a stem cell pool as they do in mouse embryos. We define neuromesodermal progenitors/NMps to be a population of cells that are competent to produce either spinal cord or paraxial mesoderm fates, and that have developed beyond stages at which embryonic cells display pluripotency. Although it is clear from cell transplant studies in the zebrafish that a population of bipotent progenitors exist[80], the degree to which individual cells will divide and give rise to both cell fates is unknown. One possibility is that a neuromesodermal competent state is a conserved feature of vertebrate embryonic development, but how potential is realized is dependent on the degree of posterior growth associated with the species in question; an interesting hypothesis that suggests a dissociation between those molecular processes driving multipotency and those that lead to a clonal expansion of progenitor populations. Zebrafish offer an extreme system with which to explore this hypothesis owing to their rapid development and the minimal volumetric growth that is associated with their body axis elongation[122]. We therefore distinguish between the term *bipotential* (reflecting a state of competence) from *mono-fated* or *bi-fated* cells (reflecting a retrospective assignment of fates during normal development by lineage tracing). Here, we have used a genetic clone-tracing method to address whether zebrafish NMps are a conserved source of spinal cord tissue, and the degree to which they populate neural and mesodermal structures during normal development. We find a closer clonal relationship between spinal cord and muscle when compared with spinal cord and anterior neural regions, which can be explained by a model of NM lineage decision at the basis of spinal cord generation in zebrafish. Tracing this lineage restriction with the combined use of photolabelling and the single cell tracking of lineages from an in toto light-sheet imaging dataset demonstrate that this restriction occurs during an early and direct segregation event with little or no amplification of the cellular pool. We observe a second population of NMps that remains resident in the tailbud and contributes to the caudal-most region of the tail, which matches a previously described tailbud NMp population[80]. Taken together with recent studies, this suggests that an NMp population is a conserved source of spinal cord and

paraxial mesoderm, but with large differences in their potential for self-renewal in vivo.

## Results

### Spinal cord scars are more closely associated to mesodermal than to anterior neural derivatives

To determine whether there are shared progenitors between the spinal cord and mesodermal derivatives, such as muscle, at the whole-organism level, we used ScarTrace, a CRISPR/Cas9-based genetic clone-tracing method that labels clones uniquely in the developing embryo and allows for the reconstruction of clonal relationships in a retrospective manner[3,52]. When Cas9 is injected together with a single-guide RNA (sgRNA) targeting a tandem array of histone- GFP transgenes in a zygote, a series of insertions and deletions of different lengths at different positions (scars) is introduced after double-stranded DNA breaks. These scars are inherited by all daughters of each labelled progenitor cell. If progenitors are shared between the spinal cord lineage and paraxial mesoderm tissues in zebrafish, it would be expected that descendants of the paraxial mesoderm tissues would have similar scars to those within the spinal cord. Alternatively, if spinal cord were generated from an ectodermal territory that is segregated from the mesoderm prior to anterior/posterior neuronal lineage segregation, as predicted by the activation/transformation model[59,92], it would be expected that it would share a common set of scars with brain regions. To distinguish between these two possibilities, we injected either Cas9 RNA (fish R1 to R3) or Cas9 protein (fish P1 to P3) together with a sgRNA against GFP at the one-cell stage into embryos transgenic for eight copies of H2A-GFP (Fig. 2.1A). Cas9 RNA injection has been shown to generate scars up until 10 hpf (hours post fertilization) and therefore would label any neural and mesodermal progenitors throughout gastrulation, whereas Cas9 protein scarring ends at around 3 hpf[3]. We isolated and sequenced scars of a total of 140 regions of six adult fish (~1 year post fertilization) and determined the amount of scars per fish and per organ (Supplementary Fig. 2.1). We then computed the distance between organs using a weighted distance measure of binarized rare scars (see Materials and Methods), and displayed them as heat map and bootstrapped trees (Fig. 2.1B,C; Supplementary Fig. 2.2 - 2.3). For Cas9 RNA-injected fish, we observe a smaller distance between spinal cord and muscle tissues at middle and tail regions of the body axis than to head regions (Fig. 2.1B; Supplementary Fig. 2.2; Fig 2.3). This results in brain, anterior and skin grouping consistently together as a distinct group from the more posterior spinal cord and muscle populations (green and blue boxes, respectively in Fig. 2.1C and Supplementary Fig.s 2.2; 2.3). Intestine, liver, kidney, blood and spleen all group together in a more distant third grouping (red box; Fig. 2.1C).

For Cas9 protein-injected fish we observe a weaker grouping, with only a close association of spinal cord tissue across the anterior to posterior axis to muscle tissue at only P1 (Supplementary Fig. 2.3). This is consistent with the early scarring of cells up until 3 hpf[3]. This observation is compatible with two models for the segregation of spinal cord and paraxial mesoderm fates during zebrafish embryogenesis.

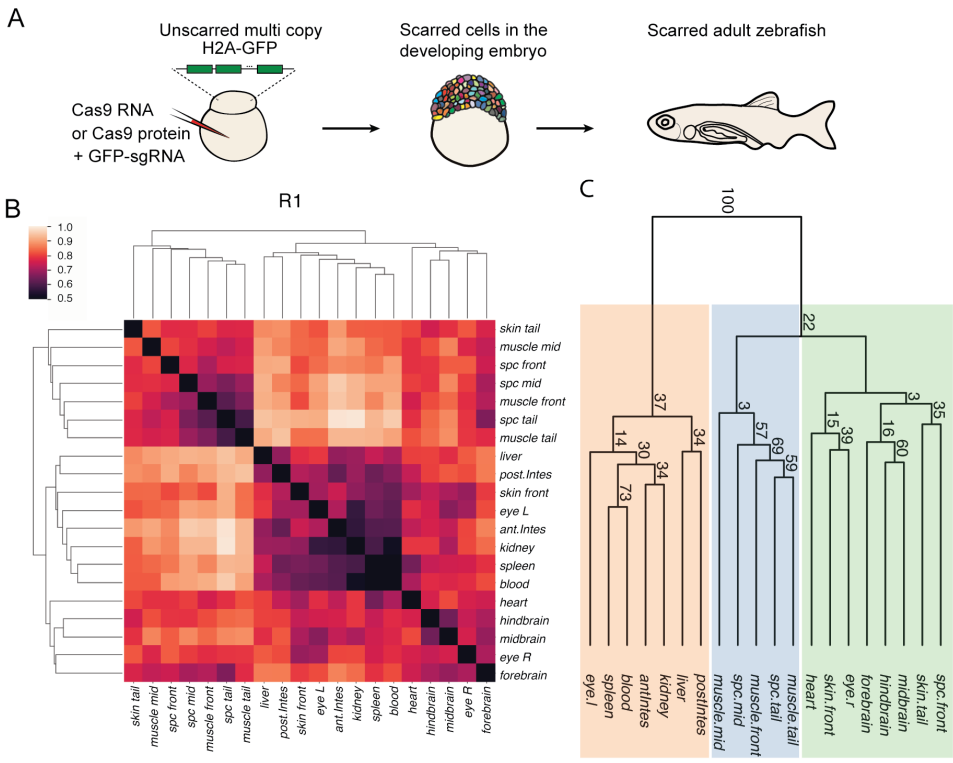



Figure 2.1: **Spinal cord clones are closely associated to mesodermal, but not anterior, neural derivatives.** (A) Experimental workflow for Scartrace lineage tracing of the zebrafish larvae. (B) Clustering of weighted distances (see Materials and Methods) of scars in dissected body structures of fish R1. The smaller the distance the closer the association between body structures. (C) Bootstrapped tree of weighted distances of scars in dissected body structures of fish R1. The tree was bootstrapped 100 times and the final proportion of clades (clade support values) are shown at each node of a clade. Coloured boxes highlight three groups: the ectodermal group (green), the mesodermal/endodermal group (orange), and the mixed ectodermal and mesodermal group (blue) containing muscle and spinal cord. There is close association of spinal cord and mesodermal cells across the anterior to posterior axis, that is distinct from brain regions. spc, spinal cord; antIntes, anterior intestine; postIntes, posterior intestine; l, left; r, right.

The first model, continuous allocation, follows the interpretation of retrospective

lineage analysis in the mouse (Fig. 2.2A) and assumes that both spinal cord and paraxial mesoderm cells are continually produced from a posteriorly localized neuromesodermal stem cell pool. The alternative is the early segregation model, according to which neural cell lineages diverge during early gastrulation stages, concomitant with the early specification of mesoderm tissue (Fig. 2.2B). To determine the relative contribution of these two segregation models of zebrafish NMps, we turned to imaging-based lineage-tracing methods.

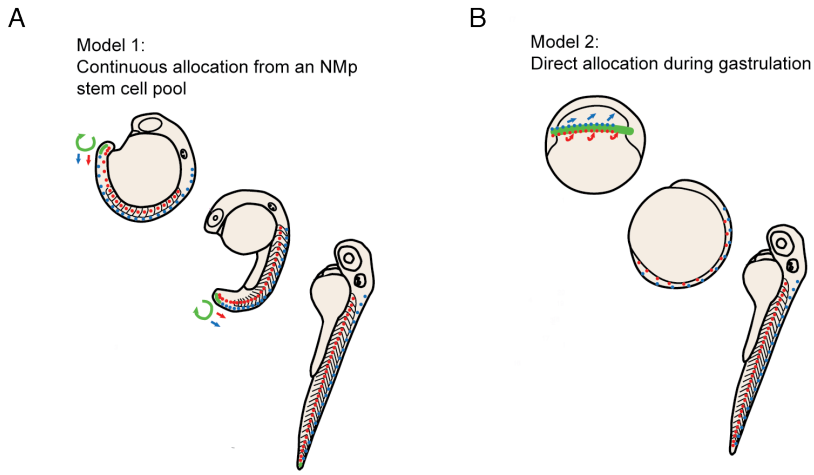



Figure 2.2: **Schematics of the continuous allocation model and the direct allocation model.** (A) Model 1. The continuous allocation model is based on lineage analysis in mouse embryos and proposes that a group of self-renewing neuromesodermal stem cells continually generates spinal cord and paraxial mesoderm throughout axial elongation. (B) Model 2. The direct allocation model proposes that an early segregation of neural and mesodermal lineages occurs during gastrulation, the derivatives of which then expand rapidly by convergence and extension.

## Tissue-level segregation of spinal cord and mesoderm populations occurs by 50% epiboly

To obtain an estimate of when spinal cord and paraxial mesoderm spatially segregate during zebrafish gastrulation, we performed a series of fate-mapping experiments using photolabelling. Embryos were injected at the one-cell stage with mRNA encoding for the photoconvertible protein, Kikume, targeted to the nucleus (nls-Kikume). Upon exposure to UV light at 30% or 50% epiboly, circular patches of blastodermal cells close to the marginal zone were labelled, enabling the direct visualization of mesoderm and ectoderm segregation, migration and tissue contribution by confocal microscopy (Fig. 2.3; Movie 1). Although labels in the prospective

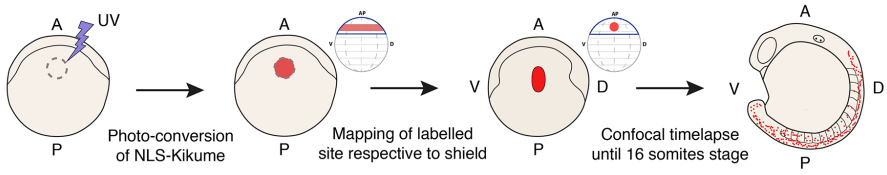



Figure 2.3: **Description of large-scale fate-mapping experiments.** NLS-Kik-injected embryos are mounted for confocal microscopy at 30 or 50% epiboly, and a circular region of the blastoderm is photoconverted. Time-lapse imaging is carried out until the 16-somite stage. Position of the label is retrospectively assigned based on its location on the anterior-posterior axis (determined at the moment of labelling) and the position of the prospective embryonic shield (appearing right after 50% epiboly), which marks the dorsal side of the embryo. AP, animal pole; V, prospective ventral side; D, prospective dorsal side (shield). Dorsal and ventral only indicate 3D orientation of the embryo and not future dorsoventral position of cells. *This figure continues in Fig. 2.4*

anterior-dorsal region of the 30% epiboly embryo contributed to large regions of the brain and notochord (Fig. 2.4A), these labels contributed little to the spinal cord and paraxial mesoderm territories, supporting the distinct lineage for brain regions found by Scartrace (Fig. 2.1). The co-labelling of anterior neural- and notochord-fated cells reflects the fact that the labelled ectodermal domain partially overlaps with the prospective shield region at 30% epiboly. A region of cells adjacent to the anterior neural domain contributes to midbrain, hindbrain and the most anterior part of the spinal cord (Fig. 2.4B; Fig. 2.6C), again consistent with the grouping of anterior spinal cord scars with the brain regions (Fig. 2.1C; green box). Labels along the medial regions of the marginal zone generate significant contributions to not only the spinal cord, but also paraxial mesoderm compartments (Fig. 2.4C, Movie 1). Labels spanning the margins of this prospective spinal cord domain performed at 50% epiboly result in significantly fewer mesodermal contributions, in line with the continued invagination of mesodermal progenitors at these stages (Fig. 2.4D; Fig. 2.6B,E). However, not even small (36 cells) labels could mark solely mesoderm-fated cells, suggesting that a degree of mixing between ectodermal and mesodermal lineages persists until 50% epiboly (Fig. 2.4E; Fig. 2.6D). A summary of all labels in terms of their contribution to neural tissues only (blue spots), mesodermal only (red spots) or both neural and mesoderm (green spots) can be observed after mapping the position of each label with respect to the embryonic shield and animal pole for labels at 30% (Fig. 2.5A) and 50% (Fig. 2.5B) epiboly. Following the 50% spinal cord/mesoderm-fated populations by time-lapse microscopy reveals a rapid convergence and extension of spinal cord progenitors that leads to a widespread contribution across a large proportion of the anterior-posterior axis (Movies 2 and 3). Contributions of each label were counted for somite and corresponding neural segments at the 16-somite stage (Fig. 2.6A), and displayed as histograms with the

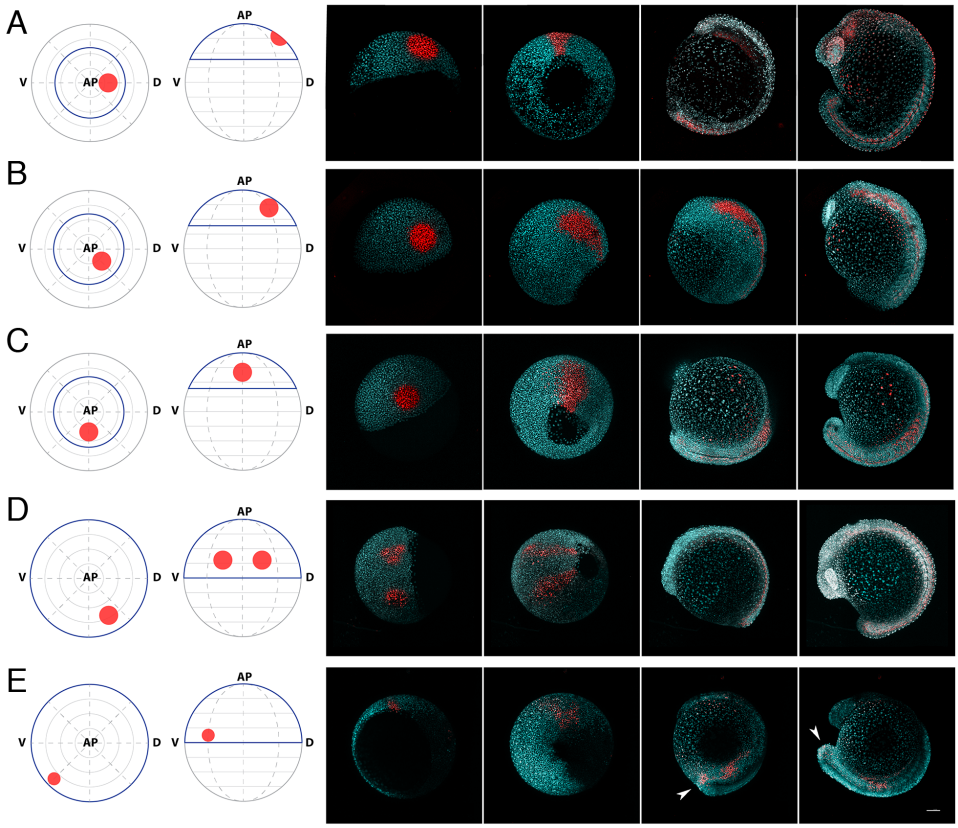



Figure 2.4: **Tissue-level segregation of spinal cord and mesoderm populations occurs by 50% epiboly.** (A-E) Embryos were photolabelled in the territories shown in the animal and lateral view diagrams (left-most image) and cells were followed until the 16-somite stage (right-most image panels) by time-lapse microscopy. Consecutive time points are displayed in the right-hand four panels for each example label. Labels were placed in the prospective anterior neural (A), marginal zone (B) and prospective spinal cord (C) territories at 30% epiboly. At 50% epiboly, the boundaries of the prospective spinal cord region (D) and a smaller, more ventro-marginal, mesodermal region (E) were mapped. Arrowheads in E indicate that a subset of labelled cells specifically migrates and contributes to the tailbud mesenchyme. Native NLS-KikGR is shown in cyan, photoconverted NLS-Kik in red. AP, animal pole; V, prospective ventral side; D, prospective dorsal side (shield). Dorsal and ventral only indicate 3D orientation of the embryo and not future dorsoventral position of cells. *This figure continues in Fig. 2.5.*

most anterior segment to the left of each plot (Fig. 2.6B-E). This shows how cells around the centre of the dorsal-to-ventral axis will contribute to neural tissue from the base of the hindbrain to the tailbud at the 16-somite stage (Fig. 2.6E).

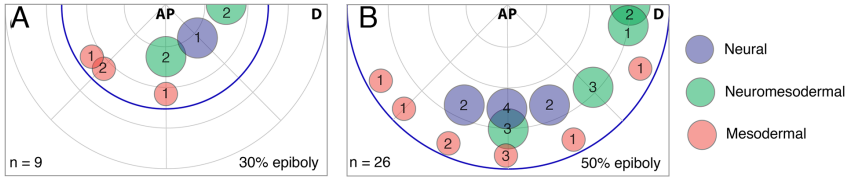



Figure 2.5: **Continuation of Tissue-level segregation of spinal cord and mesoderm populations occurs by 50% epiboly.** (A,B) Plots showing the location and diameter of labels in relation to the number of embryos imaged per label at 30% (A) and 50% (B) epiboly. Labels are colour-coded according to the ratio of cells allocated to neural tissues or paraxial mesoderm. Blue indicates that over 90% of labelled cells contribute to neural tissues; red indicates that over 90% of labelled cells contribute to paraxial mesoderm; green indicates that labelled cells contribute both to neural and paraxial mesoderm. n indicates total number of embryos fate mapped. AP, animal pole; V, prospective ventral side; D, prospective dorsal side (shield). Dorsal and ventral only indicate 3D orientation of the embryo and not future dorsoventral position of cells.

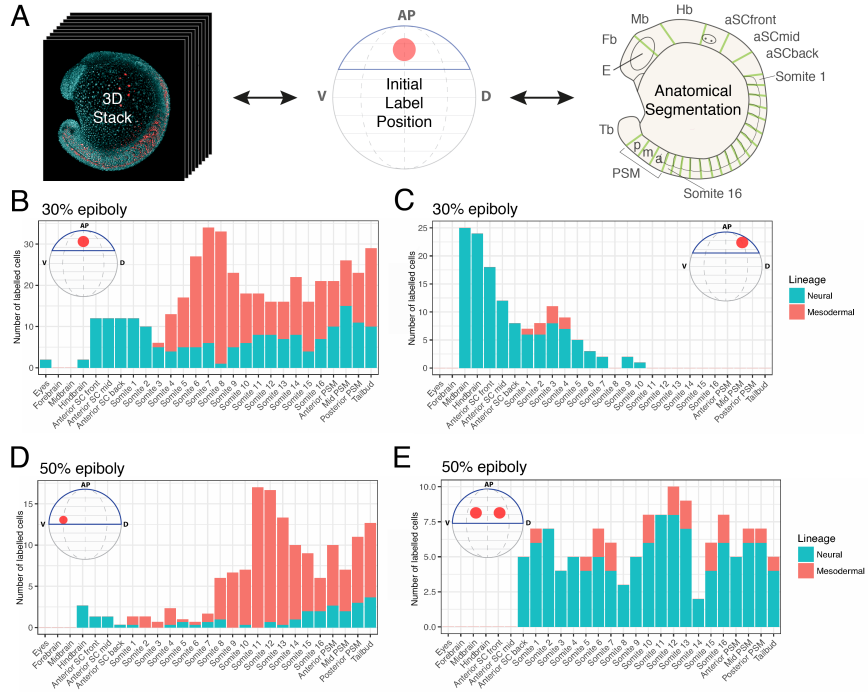



Figure 2.6: **Axial dispersion and neuro-mesodermal contribution of labelled cells.** (A) 3D confocal stacks of photolabelled embryos were analysed to relate the initial label position with the contribution of cells along the anterior-posterior axis. (B-E) The contributions of labelled populations from individual examples are plotted against the anterior-posterior axis with the number of cells in each tissue compartment shown in red for the somitic mesoderm or blue for the neural tube. There is a significant degree of overlap between spinal cord- and mesoderm-fated cells within the marginal zone at both 30% (B,C) and 50% (D,E) epiboly.

Cells that remain ectodermal upon invagination of the mesoderm become displaced posteriorly by the continued convergence and extension of cells in the animal pole (Movie 4). Thus, it appears that a large proportion of the spinal cord is allocated during gastrulation stages, and that this arises from a domain close to or overlapping with paraxial mesoderm-fated cells. However, in absence of single cell resolution, it is not possible to conclude whether these cells are a mixed population of mono-fated progenitors, or arise from a bi-fated neuromesodermal population.

## A mixed population of mono-fated and bi-fated neuromesodermal cells segregates rapidly during mid to late gastrulation

To assess whether single cells contribute to both spinal cord and mesoderm, we made use of an existing light-sheet dataset in which the onset of mesoderm specification can be observed with the use of a live reporter for mezzo[114]. In this dataset, germ layer segregation can be assessed live by detecting the increase in mezzo:eGFP fluorescence levels in the nuclei of mesendodermally specified cells (Fig. 2.7A). In the dataset used for tracking, a red channel is obtained to make mesodermal cells by taking all cells that are mezzo:eGFP positive and subtracting $Sox17^{+}$ cells that are fated towards endoderm. Similarly, a blue channel is created for ectodermal cells that results from cells expressing the ubiquitous h2b-rfp and subtracting from all those that are mezzo:eGFP+[114]. After segmentation and automated tracking of all nuclei within the gastrulating embryo, a custom MATLAB script was used to isolate tracks of cells in a user-defined region of the embryo at a chosen timepoint, thus allowing us to perform labelling experiments in silico. Automated cell tracking performed using TGMM[5] was validated by manually inspecting each track using the Fiji plug-in Mamut, by following each track and its associated cells through both time and z-slices[141]. A 'track' is defined as the sequence of connected x,y,z,t values associated with a cell and all of its progeny. To characterize the germ layer allocation of individual cells within the medial marginal zone at 30% epiboly, we focused on a region comparable with the clone of cells fated towards both spinal cord and paraxial mesoderm (Fig. 2.4C). We hypothesized that the more-animal part of this region could consist of bipotent progenitors. In fact, separation of the tracks of marginal cells from the rest of cells present in the medial region shows that overall track lengths of marginal cells are shorter than those of more-animal cells, all ending before shield stage. Shorter track lengths of these cells is a consequence of their rapid mesodermal segregation though marginal involution (Movie 5). We therefore focused on the more-animal cells of the clone, that continued up to the end of epiboly (Fig. 2.7A). Out of a total of 102 manually validated tracks from this region, 75% were assessed as being accurate without a requirement for manual correction. Twenty-three percent required manual correction, and the remaining tracks were discarded from the analysis due to gross inaccuracies in tracking.

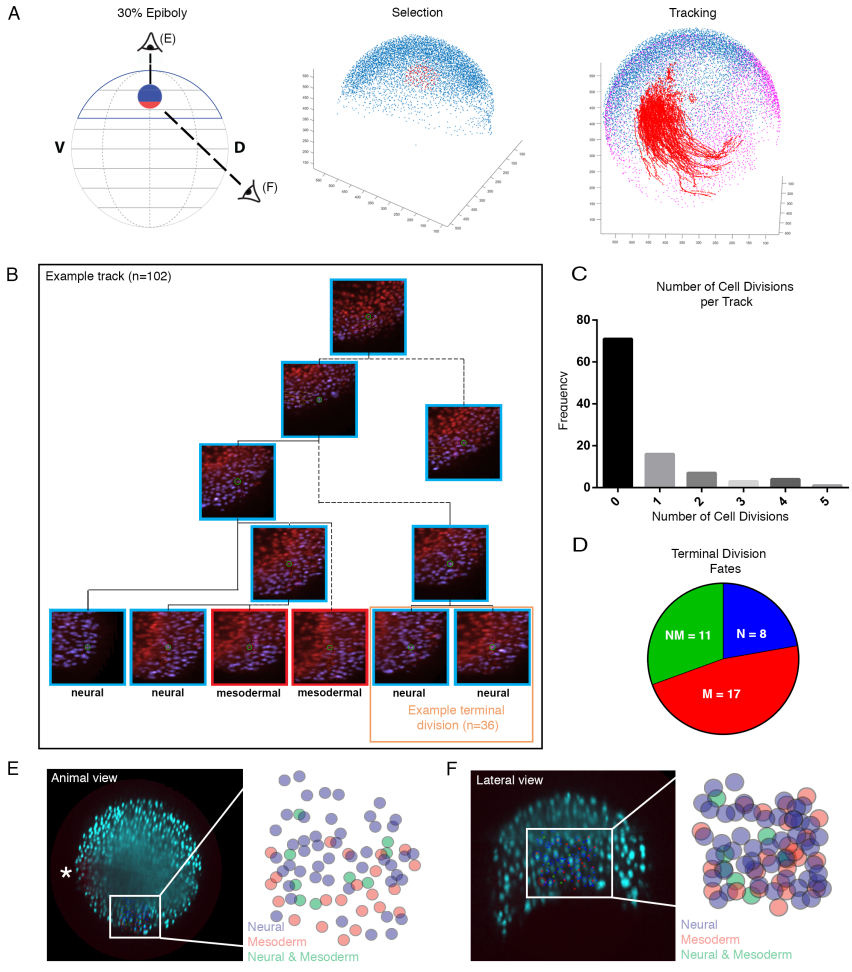



Figure 2.7: **Bipotential neuromesodermal cells segregate rapidly during mid to late gastrulation.** (A) In silico tracking of cells within the multi-view reconstruction was performed to follow lineage segregation from 30% epiboly. A custom MATLAB-based tool was developed to allow for the extraction of tracking data. Tracks were selected separately (n=100) in marginal and animal subgroups and visualized on top of 75% epiboly stage nuclei as a reference (purple). Cells were selected and tracked forwards in time, then scored based on expression or absence of mezzo:eGFP marking mesoderm. (B) An example track demonstrating a number of cell divisions and the two terminal divisions, one scored as both neural and one as a bi-fated NMp. In total 102 tracks were validated, with 36 terminal divisions scored. (C) Frequencies of cell divisions per track. (D) Of the 36 terminal divisions, 11 demonstrated contributions to both the neural and mesodermal compartment, with a remaining eight contributing to only neural and 17 to only the mesoderm. (E-F) From the animal and lateral views, the bi-fated cells are clearly intermixed with mono-fated cells. The fated cells are shown overlaid on the animal view image and then expanded in the schematic. Asterisk in E indicates the first mezzo:eGFP-positive cells and therefore the dorsal-most aspect of the marginal zone.

Cells that retained ectodermal lineage at track termination were re-selected; their final positions within the spinal cord could be observed at later stages (Fig. 2.4A) and are shown by a blue box in Fig. 2.7B. Those that started expressing mezzo:eGFP (but not Sox17) were determined as being of mesoderm fate (Movie 5) and are indicated by a red box (Fig. 2.7B). Over the duration of the movie analysis (from 30% epiboly to shield stage), 70% of tracks did not undergo a division and were thus directly determined as either being mono-fated spinal cord or paraxial mesoderm progenitors. The remaining tracks divided between one and five times over the duration of the analysis (Fig. 2.7C). To assign their fates, we focused on the terminal division progeny (Fig. 2.7B, orange box). Out of a total of 36 terminal divisions, 11 gave rise to both neural and mesodermal derivatives, with the remaining eight mono-fated as neural and 17 as mesodermal (Fig. 2.7D). Based on this, we conclude that a small population of bi-fated neuromesodermal progenitors exist close to the marginal zone of the zebrafish embryo, but rapidly segregate between the 70-90% epiboly stage. To assess the degree of mixing between either mono-fated or bi-fated cells, we colour-coded the tracks according to their retrospectively defined fates and plotted them at 30% epiboly according to either an animal view (Fig. 2.7A,E) or lateral view (Fig. 2.7A,F). Enlargement of these fate maps reveal a significant degree of mixing between neural-, mesoderm- and neuromesodermal-fated cells along both the lateral-medial (Fig. 2.7E) and animal-vegetal (Fig. 2.7F) axes.

## Tailbud neuromesodermal populations are restricted to the posterior-most aspect of the body axis

Upon completion of gastrulation, the tailbud forms via a continued convergence and extension of cells from the anterior ectoderm and a concomitant subduction of posterior cells to form the paraxial mesoderm[55]. The posterior-most subset of the labelled gastrula-stage NMp population becomes located within the dorsal tailbud upon the completion of primary gastrulation (Movie 3), suggesting that there may be a spatial continuity with the Sox2/Tbxta-positive population in this region[80]. We used photo-labelling to assess the contribution of this population. Using nls-Kik-injected embryos, multiple regions of the tailbud were photolabelled at the 6-somite stage and then re-imaged at the end of somitogenesis; an intermediate image was taken at the 22-somite stage. A progenitor region that gives rise to both neural and mesodermal tissue was identified dorsal to the posterior end of the notochord (Fig. 2.8A). In tracking these cells, it was shown that between the 6- and 22-somite stages, the progenitor cells track the dorso-posterior end of the notochord and show no contribution to the extending axis. The population was also observed to undergo a dorsal-to-ventral rotational movement along the posterior wall, aligning with the caudal neural hinge. The NM-fated populations (n=9) were observed contributing to only the final seven somites (25th to 32nd) and nine neural segments

(23rd to 32nd). Directly adjacent anteriorly to the NM-fated region, a progenitor pool with only neural fate was identified. This population contributed to earlier forming spinal cord and no mesodermal tissue. Labelled neural populations were shown to contribute to neural segments between the 14th and 27th somites (n=7). No mesoderm was labelled in these labelled populations. Directly posteriorly adjacent to the neural mesoderm progenitors, a third population of progenitors was identified that contributed to only mesodermal tissues, in particular the earlier forming somites between the 19th and 31st (n=3). To check whether our labels are cover-

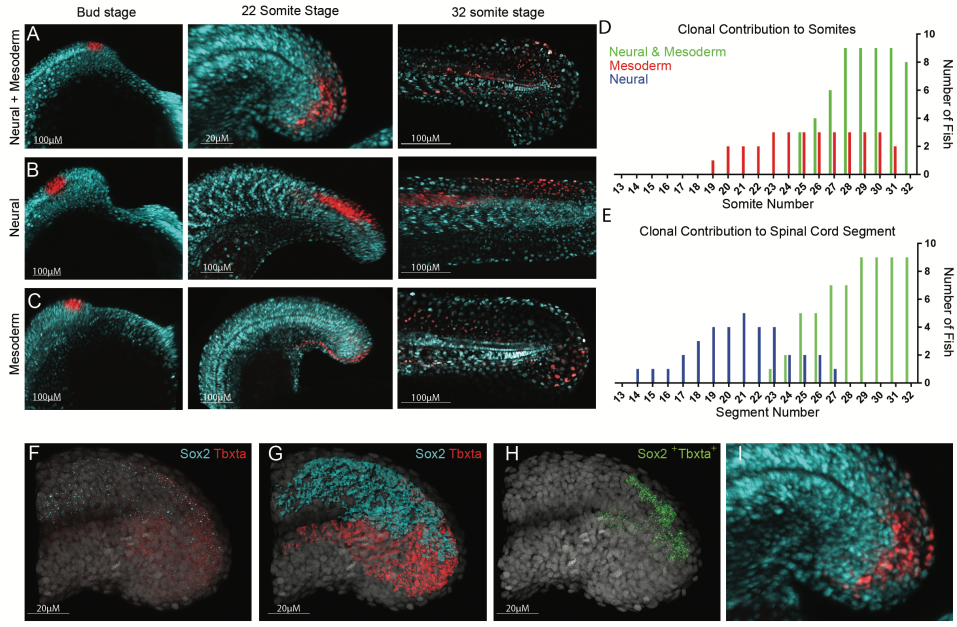



Figure 2.8: **Tailbud neuromesodermal populations are restricted to the posterior-most aspect of the body axis.** (A-C) Regions of the tailbud were photolabelled at bud stage and followed until the completion of somitogenesis. Populations on the dorsolateral wall of the tailbud gave rise to both spinal cord and mesoderm derivatives (A), those more anterior generated only spinal cord (B) and those more ventral gave rise to only mesoderm (C). All labels had an additional contribution to non-neural ectoderm. (D) The contribution of labelled populations to the anterior-posterior axis are plotted separately for mesodermal (D) and neural (E) fates. In red are labels with mesodermal contribution only (n=7), in blue are spinal cord-specific populations (n=5) and in green are those that gave rise to both germ layers (n=9). (F-I) Hybridization chain reaction (HCR) for sox2 and tbxta was used to locate double-positive cells within 3D confocal datasets. (F) Original dataset with sox2 in cyan and tbxta in red; DAPI is in grey. (G) Surface segmentation of the HCR stain was performed to mask sox2 expression within the tbxta channel. (H) Masking reveals only those cells that are co-expressing both genes. (I) Magnified image of that shown in A to compare photolabel with co-expressing cells.

ing the Sox2+Tbxta+ (previously ntl)+ domain previously described (Martin and Kimelman, 2012), we compared the labels with HCR stains for these genes (Choi et al., 2014; Fig. 2.8F). By segmenting each expression domain, the region of overlap can be visible (Fig. 2.8G; Movie 5). The tbxta expression domain was used to mask all sox2 expression outside of this region of interest, thereby allowing only regions with co-expression to be observed in 3D (Fig. 2.8H; Movie 6). This region overlaps with the example NM-labelled population shown as an enlarged version in Fig. 2.8A,I. By labelling the early progenitor populations and identifying their final fates at the end of somitogenesis (Fig. 2.8A-E), the number of cellular divisions was calculated (Fig. 2.8A-C). Neural- and mesoderm-specific progenitor populations were shown to both have a significantly ($P<0.001$, $n=7$) higher rate of proliferation than the neuromesodermal populations, which were shown to replicate infrequently. The mono-fated populations (MPs and NPs) underwent a much higher frequency of cell divisions, with NPs doubling in cell number between bud stage and the end of somitogenesis. These differences are matched by phopho-histone H3 antibody staining to quantify nuclei undergoing mitosis (Fig. 2.9D-F) and comparing division rates between tbxta-positive notochord (Fig. 2.9G), sox2-positive neural tissue (Fig. 2.9H) and co-expressing cells (Fig. 2.9I).

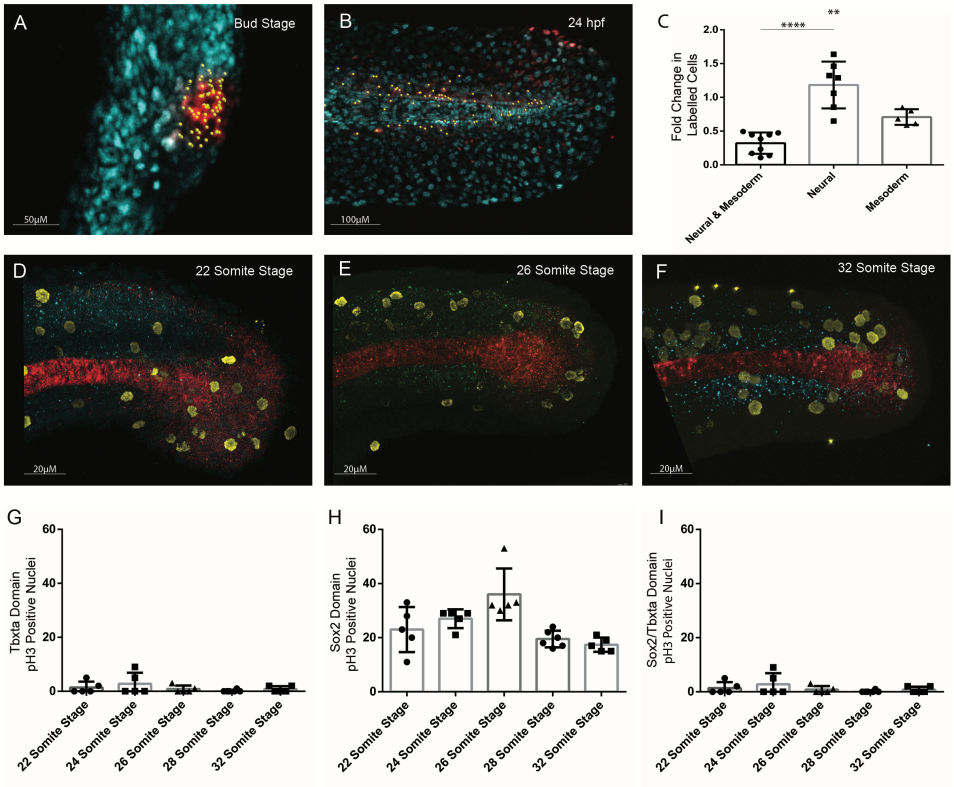



Figure 2.9: **Quantification of cell division in tailbud NMps.** Quantification of increase in number of cells photolabelled using nls-kikume from (A) bud stage through to (B) 24 hpf. (C) Fold-change increase in labelled clone number changes depending on the labelled progenitor type, with regions contributing only to neural tissue undergoing most clonal expansion and bipotent progenitors undergoing the least clonal expansion. (D-F) Replicating cells stained using phospho-histone H3 (pH3) as a marker of mitotic cells (yellow) with bipotent NMps identified through co-expression of Sox2 (blue) and Ntl (red) at the (D) 22-somite stage, (E) 26-somite stage and (F) 32-somite stage. (G-I) The frequency of Ph3-positive nuclei in the (G) Ntl-positive expression domain, including notochord, (H) Sox2-positive expression domain, and (I) Ntl and Sox2-positive NMp domain. Two-tailed Students t-test. $^{**}P<0.01$, $^{****}P<0.0001$.

## Zebrafish tailbud neuromesodermal progenitors are a population of mono-fated cells

Following zebrafish tailbud at single-cell resolution is a difficult task due to the rapid displacement of the tailbud, which continually moves the region of interest outside the field of view. To compensate for this, we adapted a vertical light-sheet set-up to allow for a continuous correction of the microscope stage position such that the tailbud is continually kept between the excitation and detection objectives (Fig. 2.10A). As neuromesodermal contributions are restricted to the last seven to nine somite segments (Fig. 2.8), we focused on generating time-lapse data from the 21 somites through to the completion of somitogenesis. Embryos transgenic for H2B-GFP were mounted for light-sheet imaging as previously described[47]. Image stacks were taken every 2.5 min for 8h. At every fifth time-point, the image stacks are downscaled and an image-based registration was performed with the image stack from five stacks before, with the x, y, z translation values stored. The values are then fed back to the microscope stage to update its position (Fig. 2.10A). The resulting datasets enable longterm imaging of the tailbud, with continuous displacement of the imaging volume in all directions (Movie 7, right-hand panel). Subsequent 3D registration of these imaging volumes removes the sudden displacements and is suitable for downstream analysis (Movie 7, left-hand panel). The entire dataset can then be observed as either maximally projected image volumes (Fig. 2.10B; Movie 8) or as slices through the dataset to observe single-cell level resolution (Fig. 2.10C; Movie 9). A boxed region of shaded background signal reflects the progressive displacement of the microscope stage during imaging (Movies 8 and 9). At late stages of somitogenesis, the tailbud has segregated into distinct territories that can be defined based on both morphology or gene expression. HCR was performed for sox2 and tbxta (Fig. 2.11D; yellow and cyan, respectively) and tbxta and tbx16 (Fig. 2.11E; cyan and red, respectively) at the 28-somite stage. This can be compared with a section through a light-sheet dataset at the same timepoint to reveal how paraxial mesoderm (Fig. 2.11F; red) and spinal cord (Fig. 2.11F; blue) can be reliably determined, with a small region of cells still co-expressing sox2 and tbxta, and therefore not assignable in terms of cell fate. Three independent datasets were tracked using TGMM, with subsequent manual tracking validation and correction performed using Mamut, as described for gastrula stages (Fig. 2.7). A summary of fates for each movie is shown in Table 1, with the regions tracked for each dataset marked in Fig. 2.11G. As expected from the proliferation analysis (Fig. 2.9), no cells in the sox2/tbxta-positive region divided in any of the three movies analysed; thus, there is zero chance of an NMp being a bi-fated progenitor during the delayed allocation of sox2/tbxta-positive cells to the body axis. We therefore focused our analysis on asking whether mono-fated NMps are arising from a mixed or a sorted progenitor population within the tailbud by mapping retrospectively defined cell fates back to

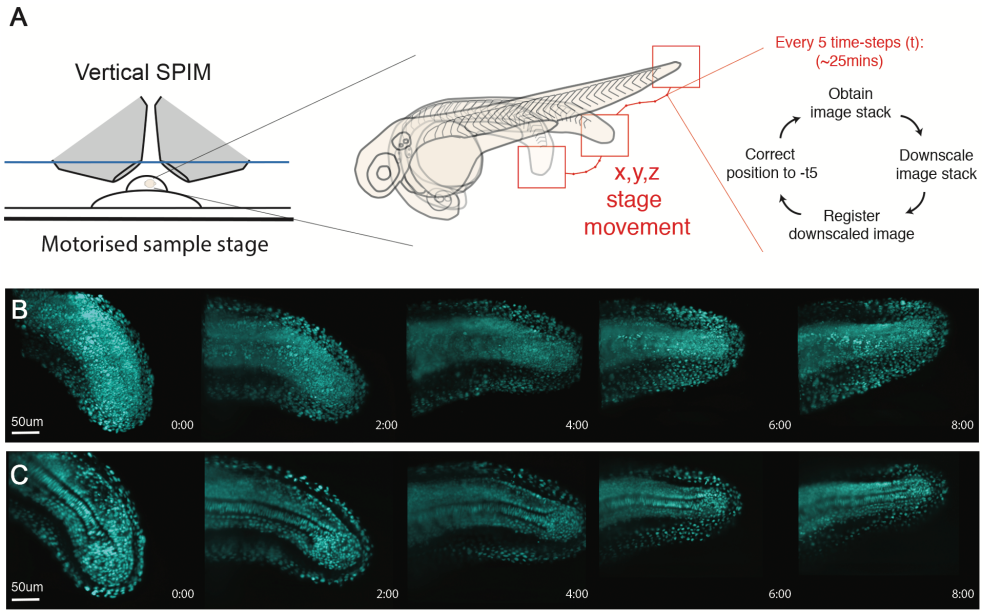



Figure 2.10: **Single cell tracking of tailbud progenitors during late somitogenesis demonstrates an absence of bi-fated cells and cellular mixing within the tailbud.** (A) Experimental design to allow long-term light-sheet imaging of the growing zebrafish tailbud. Embryos at the 21-somite stage were mounted as described by Hirsinger and Steventon (2017). Z-stacks were captured every 2.5 min with image-based registration on downsampled images conducted every five frames. The shift of this registration is then fed back into the stage position in all three directions to re-centre the tailbud within the image. (B,C) Individual frames are shown every 2 h across the 8 h movie shown as both maximum projections (B) and as a medial slice (C). *This figure continues in Fig. 2.11.*

the 21-somite stage (Fig. 2.11H,I). The Mamut tracking software displays cells close to the z-plane of the image as open circles, with coloured dots representing tracks above or below the z-plane shown (Fig. 2.11H). For clarity, we traced all positions onto a diagrammatic outline of the tailbud that shows little cell-mixing of neural (in blue) and mesodermal (in red) mono-fated progenitors (Fig. 2.11I). Those that were unassignable in terms of fate are coloured yellow. Although some degree of mixing can be seen above the notochord, these mesoderm fated cells are sitting laterally to the spinal cord, as can be seen by the open blue circles and dotted red tracks in this region (Fig. 2.11H). Together, this shows that although a large $Sox2^{+}Tbxta^{+}$ population is maintained in the zebrafish until at least the 21-somite stage (Fig. 2.8F-H; Movie 6), it is composed of entirely mono-fated progenitor cells that are largely spatially segregated.

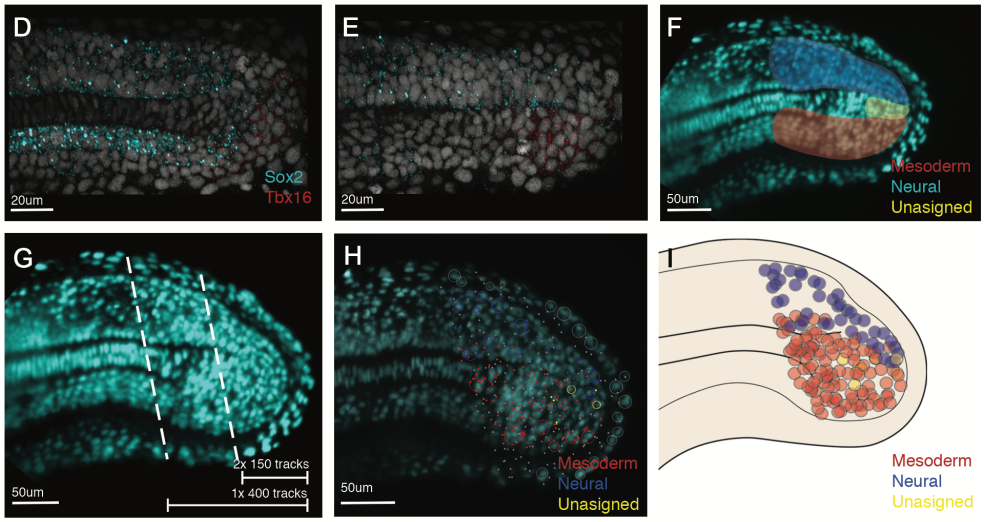



Figure 2.11: **Continuation of Single cell tracking of tailbud progenitors during late somitogenesis demonstrates an absence of bi-fated cells and cellular mixing within the tailbud.** (D-G) Fates of cells were assigned by the termination point of the track by scoring-based position within the anatomy of the tailbud and the associated expression of neural markers (Sox2) and mesdoermal markers (Tbx16 and Tbxta). Medial slices of hybridization chain reaction are shown in D and E, and used to zone the tailbud in F. Automatic tracks were selected using a custom MATLAB script by selecting all cells posteriorly to the lines shown in G. Two movies were subsetted using the most-posterior lines generating two movies of 150 tracks each. A third movie was subset using the more-anterior line, generating 400 tracks. (H,I) Final fates were assigned and have been overlaid over the starting timepoint (H) with open circles representing cells close to the viewed plane and dots representing cells on a different z-section. These different z-planes have been collapsed into a single 2D image (I) and demonstrate an absence of a mixed population of unfated cells. The mixing around the notochord is an artifact of collapsing the z-axis into a 2D image. No cells were observed dividing with progeny entering both lineages.

## Discussion

Taken together, these results show that a neuromesodermal progenitor pool is a source for at least a subset of spinal cord cells at all regions of the body axis. Crucially, however, this is not achieved by a bipotent stem cell population but rather by the rapid segregation of neural and mesodermal cell fates during gastrulation. Subsequently, a second population of neuromesodermal progenitors arises along the dorsal wall of the tailbud, co-expressing both sox2 and tbxta (shown in this study and [80]), and is competent to switch to either neural or mesoderm states in response to changes in Wnt signalling activity [80]. However, these are not continually added to the elongating body axis, as previously supposed, but are rather a largely quiescent population of cells that contributes to the last region of the body axis. The low proliferation levels of these cells during somitogenesis are in line with a previous study [15]. Although it is possible that a proportion of cells are bipotent within this population, the low rates of division make it unlikely that their derivatives will have sufficient time prior to the completion of somitogenesis to generate a stem-cell mode of growth as has been observed in mouse embryos [136]. Indeed, single cell tracking and retrospective cell fate assignment reveals that the tailbud NMp population consists of a population of fate-restricted and spatially segregated cells. Therefore, we surmize a composite model for spinal cord generation in zebrafish, with spatially and temporally separate pools. The first is directly allocated during gastrulation and is a population of NMps mixed in with a large proportion on monopotent progenitors. The second matches with a previously described tailbud NMp population [80] and has a delayed allocation to only the final region of the larval tail (Fig. 2.12). Why is it that zebrafish embryos maintain a tailbud NMp popula-

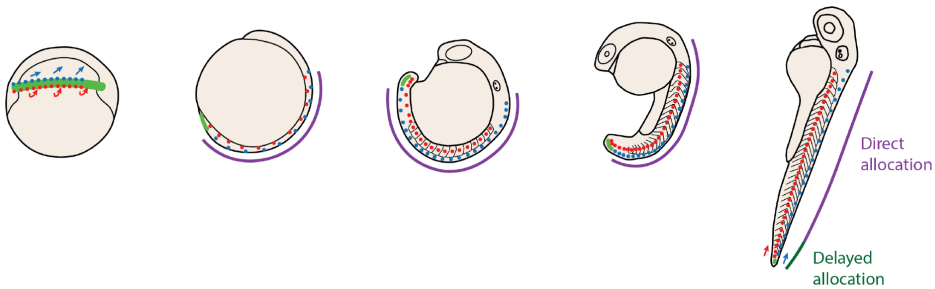
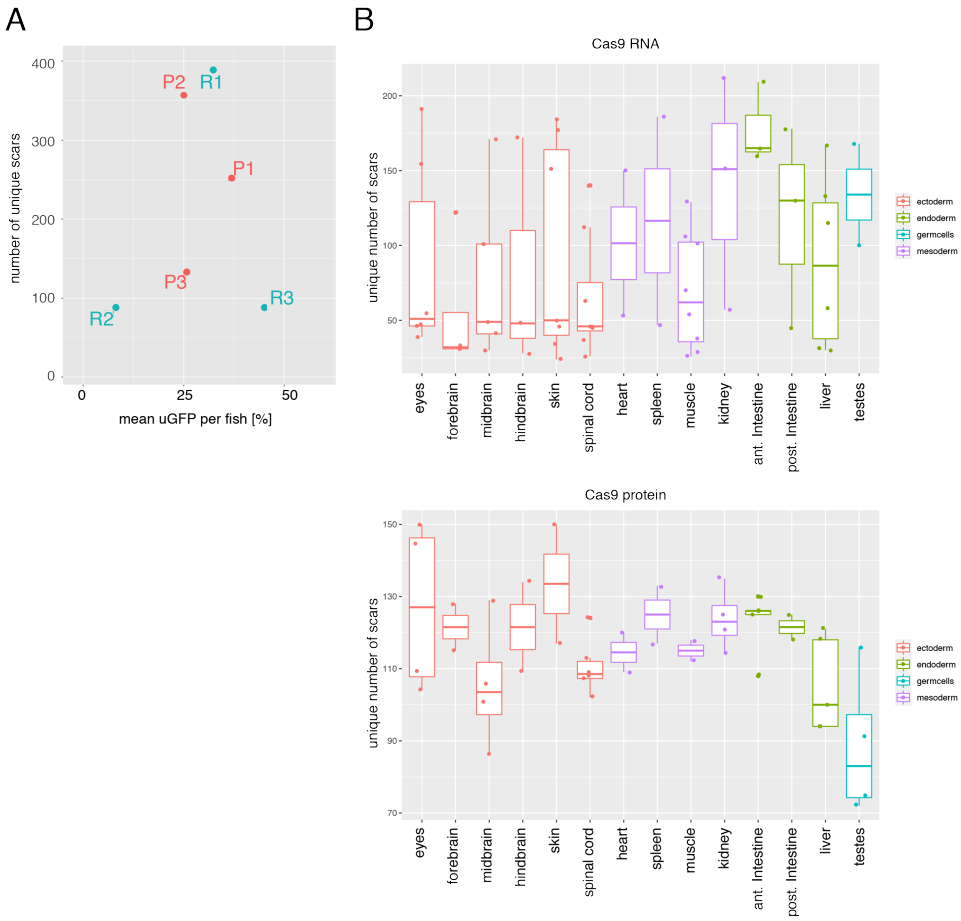


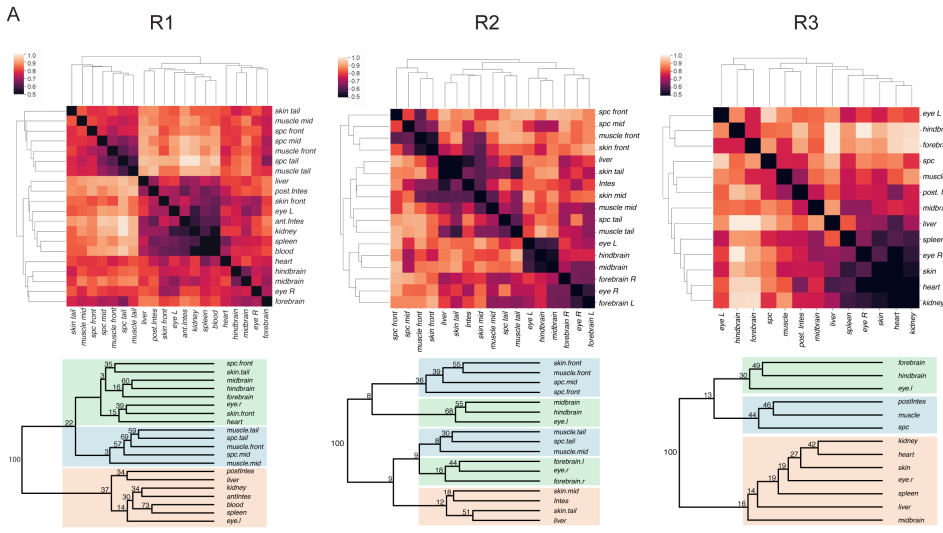

Figure 2.12: **Zebrafish allocated spinal cord and mesoderm progenitors through a combination of direct and delayed allocation modes.** Diagrams depict zebrafish embryos at progressive stages of development with neuromesodermal competent regions shown in green. Cells either become restricted to spinal cord or paraxial mesoderm fates at early gastrulation (direct allocation; purple) or remain within the tailbud until late somitogenesis (delayed allocation; dark green).

tion whose bipotentiality is never realized during normal embryonic development? One possibility is that developmental constraint is acting to conserve a bipotential NMp population across the amniote-anamniote transition, despite large variations in the cellular behaviours that are driving body axis elongation[121,122]. Developmental constraint in this case could be acting to maintain the regulatory interactions between transcription factors such as sox2 and tbxta, which maintain a bipotential state, thus leading to a conserved cellular trajectory for NMps (i.e. the series of gene expression states associated with a specific cell fate decision). In contrast, alterations in the multicellular dynamics driving the elongation process may lead to vast alterations in the way this is realized at the cellular level, leading to a divergence in cell lineage (i.e. the series of motherdaughter relationships associated with a specific cell fate decision). Further comparative analyses of NMp populations between chordate model organisms is likely to reveal some important insights into the degree to which cell trajectory and lineage can be separable during evolution and development. Despite a strong developmental constraint acting upon gastrulation in vertebrates[1], there exists a large degree of morphological variation acting at these stages of development[33]. This variation is largely a consequence of the different strategies of maternal-embryo energetic trade-off that have been adopted during chordate evolution. In the context of mouse, viviparity has led to an internal mode of development within which embryos increase in volume to a large degree, concomitant with establishing their body plan[122]. This is in contrast to macrolecithal embryos such as fish, whose energy supplies are contained within an external yolk sac. Such transitions have a great impact on the morphology of gastrulae, which must adopt their shape according to the physical constraints of yolk size and extra-embryonic structures. Furthermore, external modes of development provide a selective advantage for developmental strategies that favour a Fig.rapid development to swimming larval stages in order that they escape predators and find food. Zebrafish undergo a highly rapid mode of development, and develop their full complement of somites, prior to the development of a vascular system that is efficient at accessing nutritional supplies and allowing them to increase in mass. We propose that by shifting the allocation of spinal cord and mesodermal progenitors to early gastrulation stages, this has facilitated a rapid convergence and extension-based mode of axial elongation. Furthermore, the pool of NMps within the tailbud demonstrates a conservation of a tailbud progenitor pool that could allow for increased flexibility according to organism level heterochronies in the rates of growth. This hints at a novel evolutionary developmental mechanism that we term *growth-mode adaptability*. This proposes that specific cellular trajectories (such as the emergence of spinal cord from a neuromesodermal cell state) are conserved in evolution, yet highly adaptable in terms of their timing of cellular decision making and modes of growth.

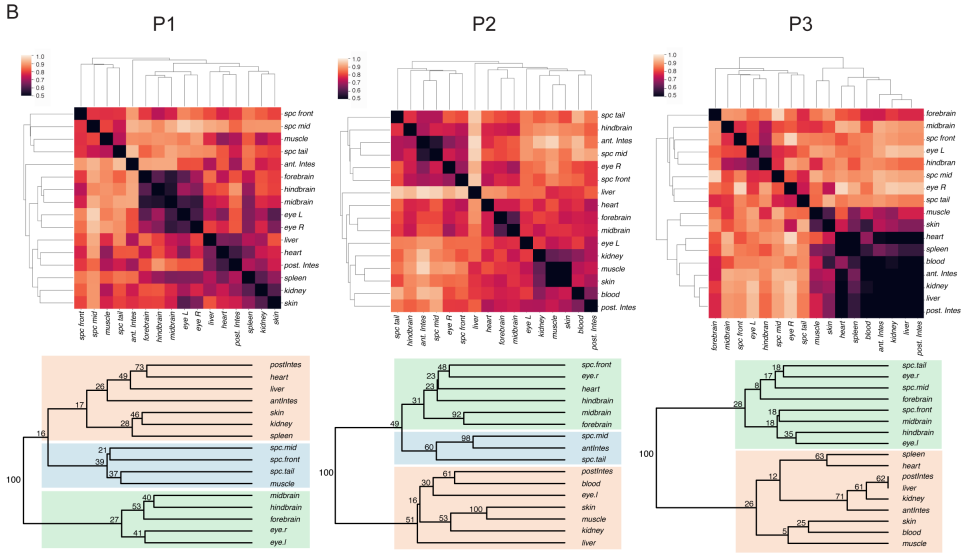
# Supplementary Material



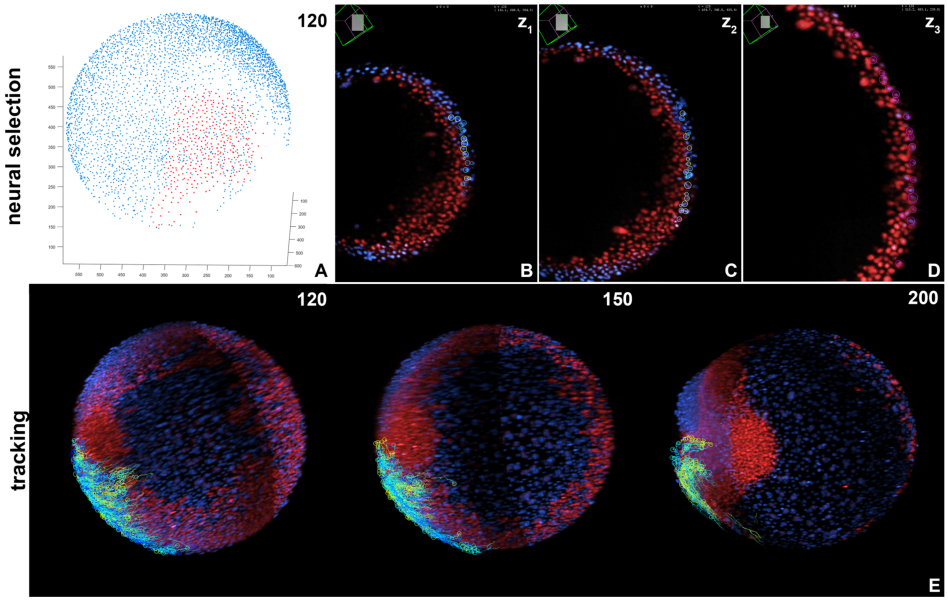

Supplementary Figure 2.1: **ScarTrace dynamics and efficiency of clonal labelling.** (A) Scarring dynamics for Cas9 RNA or protein shown as decrease in uGFP over time. (B) Scarring efficiency for Cas9 RNA injected fish R1-R3 and Cas9 protein injected fish P1-P3. (C) Scarring efficiency shown as number of clones for all Cas9 RNA injected fish.




Supplementary Figure 2.2: **ScarTrace Clustering per fish Cas9 RNA injected fish.** ScarTrace Clustering and trees of weighted distances between body structures of Cas9 RNA injected fish R1 (as shown in Fig. 1), R2 and R3. Each tree was bootstrapped 100 times and the final proportion of clades (clade support values) are shown at each node of a clade. Color in the background are a guide to the eye for following the close association of spinal cord and mesodermal body structures across the anterior to posterior axis, that is distinct from brain regions in R2, R3. antIntes = anterior intestine, postIntes = posterior intestine, l. = left, r. = right, Spc. = Spinal cord




Supplementary Figure 2.3: **ScarTrace Clustering per fish Cas9 protein injected fish.** ScarTrace Clustering and trees of weighted distances between body structures of Cas9 protein injected fish P1, P2 and P3. Legend same as in 2.2




Supplementary Figure 2.4: **Spinal cord allocation of upper marginal cells.** (A) Cells of the upper marginal label shown in Fig. 3 are selected (red spots) and followed from timepoint 120. (B-D) Mamut Viewer visualisation of the selection at three different optical sections shows that cells lack mezzo expression and are still localised in the outer layer of the blastoderm. (E) Further tracking of these ectodermal cells shows their migration towards the embryonic anteroposterior axis and final arrangement in one of the two arcs of cells that converge to form the spinal cord. Images in E depict partial maximum intensity projections to exclude the head region at the indicated timepoints.

**Supplementary Tables 1-6. All dissected body structures and barcodes.** For each Cas9 RNA injected fish (R1-R3) and Cas9 protein injected fish (P1-P3) we show a separate table listing the barcodes ("BC" column) needed to match reads from a sequenced library (one library per fish, can be found under GSE121114) to the corresponding dissected body structure ("Organ" column). Tables S1-S6 can be downloaded here: www.biologists.com/DEV_Movies/DEV166728/TableS1-S6.xlsx

**Supplementary Tables 7-12. Binarized scars tables.** For each scarred fish we are providing the binarized scars tables we are obtaining after mapping, filtering and binarizing as described in Materials and Methods. These tables were used as input to calculate the IWSS distances between the dissected body structures shown in form of heatmaps and trees in Fig. 2.1 and Supplementary Fig.s 2.1, 2.2. The rows of the tables show each scar as a CIGAR string and the columns the disstected body structure. Tables S7-S12 can be downloaded here: www.biologists.com/DEV_Movies/DEV166728/TableS7-S12.xlsx

**Movie 1. Photolabel at 50% epiboly with large contributions to both spinal cord and paraxial mesoderm.** http://movie.biologists.com/video/10.1242/dev.166728/video-1

**Movie 2. Convergence and extension of spinal cord fated photo-labeled cells at 50% epiboly.** http://movie.biologists.com/video/10.1242/dev.166728/video-2

**Movie 3. Photolabels across the marginal zone contribute to large portions of the anterior-posterior axis.** http://movie.biologists.com/video/10.1242/dev.166728/video-3

**Movie 4. Light-sheet movie of a 50% neuromesodermal fated photolabel.** http://movie.biologists.com/video/10.1242/dev.166728/video-4

**Movie 5. Manual verification of a mono-fated mesodermal progenitor.** http://movie.biologists.com/video/10.1242/dev.166728/video-5

**Movie 6. 3D segmentation of Sox2 and Tbxta positive cells within the zebrafish tailbud.** http://movie.biologists.com/video/10.1242/dev.166728/video-6

**Movie 7. Sections through a tailbud-stage light-sheet movie before and after image registration.** http://movie.biologists.com/video/10.1242/dev.166728/video-7

**Movie 8. Maximum projection of a registered tailbud movie.** http://movie.biologists.com/video/10.1242/dev.166728/video-8

**Movie 9. Midline section of a registered tailbud movie.** http://movie.biologists.com/video/10.1242/dev.166728/video-9

# Methods

## Animal lines and husbandry

This research was regulated under the Animals (Scientific Procedures) Act 1986 Amendment Regulations 2012 following ethical review by the University of Cambridge Animal Welfare and Ethical Review Body (AWERB).

## Scartrace

Scarred zebrafish (injected with RNA Cas9 or protein Cas9 and gRNA against GFP) were made as described by[3,52]. Scarred zebrafish (around 1 year old) were anaesthetized, sacrificed and small parts of body structures were dissected and collected individually (a full list of all dissected body structures and barcodes (BC) are available in Tables S1-S6). Genomic DNA was isolated from the body structures using DNeasy Blood and Tissue Kits (Qiagen, 69506) and scars were amplified and sequenced using the ScarTrace bulk protocol as described by[52]. Scars were mapped with the Burrows- Wheeler Aligner[3]. Raw sequencing data (as fastq

files) and mapped scars tables have been deposited in GEO under accession number GSE121114. After mapping, to filter out potential sequencing errors, the scar percentage was computed in normalised histogram with 100 bins. Scars were only kept when their fraction was at least 10  higher than the minimum detected scar fraction. To filter out fish with inefficient scarring, we determined the scarring efficiency as the number of scars after filtering and the mean unscarred GFP percentage across all organs per fish. Only fish with less than 50% unscarred GFP and more than 100 scars were kept (these are fish R1-R3 and fish P1-P3, Fig. 2.1B). Afterwards, to ensure we kept scars that were made once in the developing embryo, we filtered out all scars that we detected in other fish and in the dynamics experiments from[3]. This filtering step ensured that we kept only rare scars, with a probability of $P=10^{-5}$ of being made. After filtering, we summed scars from technical replicas of the same dissected body structure. As a last processing step, we binarised the filtered scars (Tables S7-S12) and computed the distances between body structures using a weighted distance measure for sparse samples (information weighted sparse sample distance, IWSS). IWSS incorporates both the matches and mismatches in an information-weighted manner to incorporate all available information in sparsely sampled vectors. The IWSS distances between the organs are shown as log distances in heatmaps (Fig. 2.1B; Supplementary Fig. 2.1 - 2.2) and bootstrapped trees (Fig. 2.1C; Fig. 2.2). Heatmaps were plotted using the python seaborn package with the settings method='average' and metric='euclidean'. Trees were bootstrapped 100 times using the R function boot.phylo and prop.clades from the *ape* package.

## Photolabelling and time-lapse microscopy fate mapping

Zebrafish wild-type embryos were injected at the one-cell stage with 200pg of nuclear-targeted Kikume (a kind gift from Ben Martin, Stony Brook University, NY, USA). Photolabelling was performed as described previously[122]. For fate-mapping experiments at gastrulation stages, six to ten 30% or 50% epiboly stage embryos were embedded in a drop of low gelling point agarose (1% w/v in E3 media) at the bottom of a MatTek petri dish (1.5 glass coverslip bottom) and imaged using a 10 air objective (NA=0.45). Time lapse imaging of labelled embryos was performed using a Zeiss LSM700 confocal microscope equipped with a temperature-controlled chamber set at 28°C. In our setup, nearly complete photoconversion of NLS-KikGR was obtained by scanning the 405 nm laser at 10-15% laser power on a circular user-defined region of the embryo. Photolabelling efficiency was directly visualised by simultaneously capturing both NLS-KikGR emissions on two different PMTs. After labelling of all embryos, overnight multidimensional acquisition of low resolution stacks for each embryo was set up with an interval of 10 min. The experiment was stopped the morning after embryos had reached the 16-somite stage, at which they were dissected out of the agarose, individually de-chorionated, re-embedded

and imaged at higher z-resolution. The excitation power of 488 and 561 nm lasers was kept below 12% throughout time-lapse acquisition and further lowered after de-chorionation. To quantify label contribution to neural or paraxial mesoderm tissues, as well as their axial spreading, high-resolution datasets of 16-somite stage embryos were segmented and cells counted using the surface and spot functions in Imaris v8.3 (bitplane.com). For post-gastrulation labelling, embryos were mounted in 2% methycellulose in E3. Mounting in low gelling point agarose in E3, as described by[47], was performed for live imaging of post-gastrulation embryos.

### Scanning light-sheet imaging microscope with online stage adjustment

The additional light-sheet imaging was performed using a home-build scanning light-sheet microscope in an upright geometry, where the excitation and emission lenses were each tilted 45° to the vertical axis to allow for horizontal sample placement on a xyz-translation stage. The sheet was generated using 2-4% laser output power at 488 nm (200 mW maximum output power), delivered to the sample using a Nikon 10x, 0.3NA water-immersion objective and a galvo scanning system (Cambridge Technology). The sheet thickness determined before the experiment in dye solution was $\sim 2\mu$m at the waist. Fluorescence was collected with a Nikon 40x, 0.8 NA water-immersion objective, using a 525/50 nm (Chroma Technologies) emission filter in front of the sCMOS camera (Hamamatsu Flash 4). The position of the detection lens was synchronized with the position of the excitation sheet using a long travel piezo objective positioner (Physik Instrumente). To allow automated long-term imaging of the structurally changing sample, an automated feature of the tracking method was implemented in the bespoke control software written in Labview. Data handling was performed by calling MATLAB scripts from within the software. The goal was to keep the area of interest in the field of view so that further registration could happen offline. During time-lapse acquisition, all image stacks are downsampled online to $1x1x1\mu m^3$ voxel size to allow for efficient data handling. For every 5th timepoint, a three-dimensional phase correlation is calculated between the current downsampled stack and a reference stack. From this, the necessary xyz-stage shift is determined and sample moved accordingly. The next image stack replaces the previous reference stack to take into account overall changes. All necessary computation steps were optimized so that they could be completed in between timepoints in order not to delay image acquisition. After the completed time-lapse image acquisition, consecutive image stacks were registered using custom MATLAB scripts employing the same 3d-phase-correlation algorithm.

## Hybridization chain reaction and immunohistochemistry

Five antisense DNA probes were designed against the full-length zebrafish *sox2* and *ntla* mRNA sequence as described by Choi et al. (2014). Embryos incubated in either embryo medium were fixed 4% PFA at 4°C and then stained according to Choi et al. (2014). HCR was combined with immunohistochemistry for phospho-histone H3 for identification of mitotic cells at the point of fixation. Embryos underwent HCR followed by blocking in 2% Roche blocking reagent, 5% donkey serum in maleic acid buffer. A mouse anti-pH3 antibody (abcam, ab14955) was incubated overnight at 4°C followed by a secondary anti-mouse conjugated to an Alexa 488 fluorophore (ThermoFisher Scientific, A32723). Imaging was performed on a Zeiss LSM700 confocal with identical imaging parameters: z-step 0.5683 $\mu m$; 1024x1024 resolution; 63x oil objective NA 1.4; 35% laser power at 488 nm; Gain 661; Digital offset –2; Pixel dwell time 3.12 $\mu s$.

## MATLAB script for trajectory selection

The custom-written MATLAB script imports trajectories saved in TGMM software format[5] and visualizes the position of the detected cells at two user-defined points in time. Through GUI the user selects the cells of interest and subsequently the trajectories of the respective cells are plotted in red. The trajectories retained also include those of the daughter cells originating from the originally selected cells at any timepoint. An (optional) post-processing step eliminates the cells on the opposite side of the embryo from the user's viewpoint, in case they were inadvertently selected. The selected trajectories can be saved in the TGMM software format or as a MATLAB ".mat" file.

## In silico fate mapping

The light-sheet dataset used for the lineage analysis, as well as the TGMM automated tracking[5], has been described by[114]. Trajectories selected at timepoint 30 (corresponding to 30% epiboly) were exported as TGMM output .xml file using our MATLAB script. Position of cells analogous to those analysed in fate-mapping experiments were obtained by comparing the 30% epiboly stage with the 75% epiboly stage, at which the embryonic shield is visible. TGMMdatawas imported into MaMuT[141] for visualization and validation of tracks. Lineageswere followed using the Track Scheme module and inspected by overlaying the tracking data on top of the original images in Mamut viewer. Germ layer segregation of cells along tracks was determined based on the expression of the mezzo:eGFP mesodermal reporter. Endodermal cells, expressing sox17::YFP, were excluded. Spinal cord contribution of nonmesodermal cells was manually verified using Mamut.

## Acknowledgements

We thank Alfonso Martinez-Arias, Adrian Friday, Ben Martin, Bertrand Benazeraf, Estelle Hirsinger, Octavian Voiculescu and all members of the Steventon lab for comments on the manuscript. We also kindly thank J. M. Pedraza Leal and B. de Barbanson for providing the IWSS distance, and A. Alemany for input on the ScarTrace analysis and the manuscript.

## Author contributions

Conceptualization: B.S.; Methodology: A.A., L.M., M.O.L.; Software: L.M.; Investigation: A.A., T.F., M.F., G.S., M.O.L., C.L., B.S.; Resources: G.S., J.H.; Data curation: A.A., T.F., C.L.; Writing - original draft: B.S.; Writing - review & editing: T.F., M.F., B.S.; Supervision: J.H., A.v.O., B.S.; Project administration: B.S.; Funding acquisition: J.H., A.v.O., B.S.

## Funding

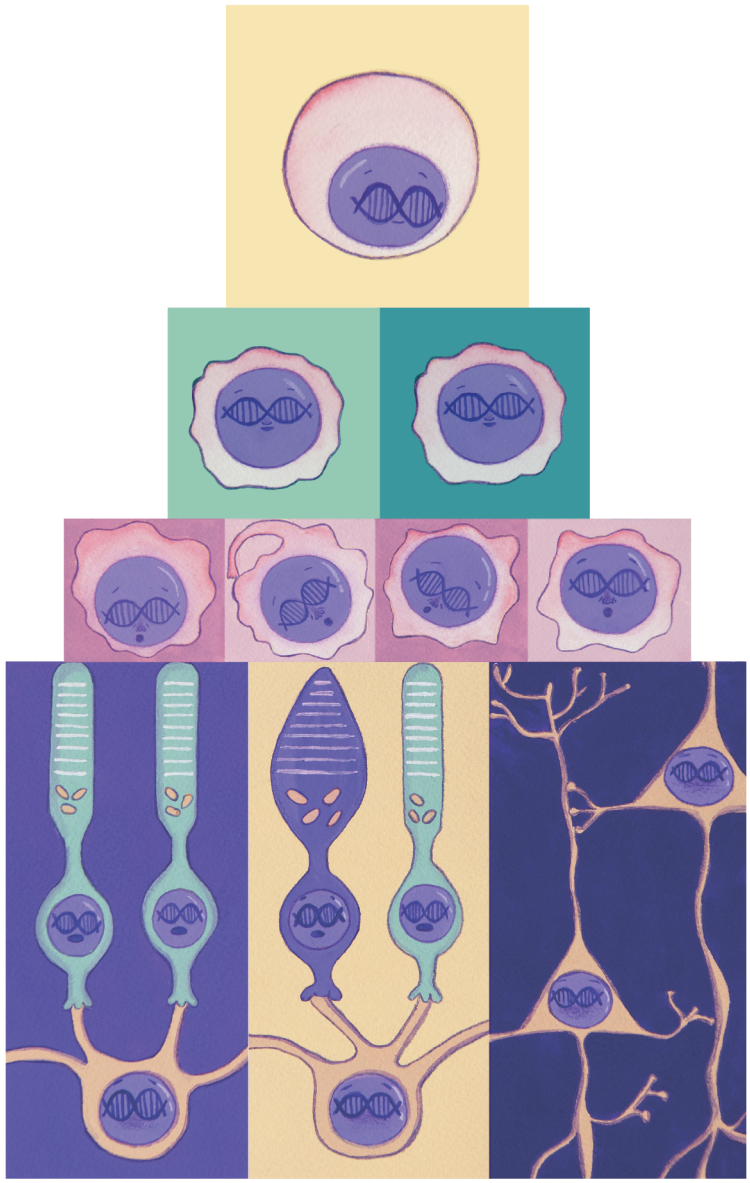
B.S. and T.F. are supported by a Henry Dale Fellowship jointly funded by the Wellcome Trust and the Royal Society (109408/Z/15/Z). A.A. was supported by the Erasmus+ Traineeship scheme of the European Commission. L.M. is supported by an Engineering and Physical Sciences Research Council grant EP/R025398/1. C.L. is supported by a British Society of Developmental Biology/The Company of Biologists/Gurdon Summer Studentship. Work in the lab of J.H. was supported by the Max-Planck-Gesellschaft and European Research Council (CoG SmartMic, 647885). The work of M.F. and A.v.O. was supported by a European Research Council Advanced grant (ERC-AdG 742225-IntScOmics), by a Nederlandse Organisatie voor Wetenschappelijk Onderzoek TOP award (NWO-CW 714.016.001) and by the Foundation for Fundamental Research on Matter, which is financially supported by Nederlandse Organisatie voor Wetenschappelijk Onderzoek (FOM-14NOISE01). This work is part of the Oncode Institute, which is partly financed by the KWF Kankerbestrijding. Deposited in PMC for immediate release.

## Data availability

Sequencing data associated with the ScarTrace lineage tracing study have been deposited in GEO under accession number GSE121114. The Matlab script allowing for user-defined selection of tracking data can be found here: doi.org/10.5281/zenodo.1475146.

## Supplementary information

Supplementary information available online at :
http://dev.biologists.org/lookup/doi/10.1242/dev.166728.supplemental



Chapter 3

# Whole-organism clone-tracing using single-cell sequencing

A. Alemany[1,*], M. Florescu[1,*], C.S. Baron[1,*], J. Peterson-Maduro[1,*] and A. van Oudenaarden[1]

published in *Nature* 2018

*equal contribution
[1] Hubrecht Institute-KNAW (Royal Netherlands Academy of Arts and Sciences) and University Medical Center Utrecht, 3584 CT, Utrecht, The Netherlands.

Embryonic development is a crucial period in the life of a multicellular organism, during which limited sets of embryonic progenitors produce all cells in the adult body. Determining which fate these progenitors acquire in adult tissues requires the simultaneous measurement of clonal history and cell identity at single-cell resolution, which has been a major challenge. Clonal history has traditionally been investigated by microscopically tracking cells during development[57,125] monitoring the heritable expression of genetically encoded fluorescent proteins[73] and, more recently, using next-generation sequencing technologies that exploit somatic mutations[10], microsatellite instability[104], transposon tagging[127], viral barcoding[90], CRISPR–Cas9 genome editing[35,40,52,53,84,111] and Cre- loxP recombination[97]. Single-cell transcriptomics[130] provides a powerful platform for unbiased cell-type classification. Here we present ScarTrace, a single-cell sequencing strategy that enables the simultaneous quantification of clonal history and cell type for thousands of cells obtained from different organs of the adult zebrafish. Using ScarTrace, we show that a small set of multipotent embryonic progenitors generate all haematopoietic cells in the kidney marrow, and that many progenitors produce specific cell types in the eyes and brain. In addition, we study when embryonic progenitors commit to the left or right eye. ScarTrace reveals that epidermal and mesenchymal cells in the caudal fin arise from the same progenitors, and that osteoblast-restricted precursors can produce mesenchymal cells during regeneration. Furthermore, we identify resident immune cells in the fin with a distinct clonal origin from other blood cell types. We envision that similar approaches will have major applications in other experimental systems, in which the matching of embryonic clonal origin to adult cell type will ultimately allow reconstruction of how the adult body is built from a single cell.

## Results

The goal of our experiment is twofold: first, to link cells in the embryo to their corresponding clones in adult tissue (Fig. 3.1a); second, to quantify cell-type composition of these clones to determine the multipotency of embryonic progenitors. To reach the first goal, we need to uniquely label the cells in an embryo with permanent and heritable labels. For this, we use CRISPR/Cas9 technology, which induces a double-stranded break at the targeted genomic site that is repaired as insertions or deletions of different lengths at different positions (scars)[49,96]. To allow for multiple scarring in the same cell, we use a zebrafish line with eight in-tandem copies of a histone-green fluorescent protein (GFP) transgene[96] (Methods). Scarring starts after injecting the yolk or cell of the zygote with Cas9 RNA or protein, and a single-guide RNA (sgRNA) that targets GFP (Fig. 3.1b). We quantified the scarring rate by measuring the fraction of unscarred GFP in zebrafish embryos at different times after Cas9 delivery (Fig. 3.1c), which is five times faster in Cas9 pro-

tein than in RNA injections. Cas9 activity ceases at around 3 h for protein and at 10 h for RNA injections, when zebrafish embryos have about 1,000 and 8,000 cells, respectively[57]. We detect more than 1,000 distinct scars, the abundances and probabilities of which span several orders of magnitude[52] (Supplementary Information sections 1 and 2). To detect scars and transcriptome from single cells, we developed ScarTrace, which integrates a nested PCR step after transcriptome conversion to cDNA into the sorting and robot-assisted transcriptome sequencing (SORT-seq) protocol[89] (Fig. 3.1d). Because the histone-GFP transgene is transcribed, scars can be detected from mRNA and genomic DNA (gDNA). Detection from gDNA is preferred because GFP expression might be tissue specific, vulnerable to silencing and scars might affect the half-life of the mRNA. We assessed the efficiency of scar detection from mRNA and gDNA by comparing scar patterns of single cells from the caudal fin obtained using ScarTrace with and without reverse transcription (Fig. 3.1d, step 1). We detected $3.3 \pm 0.3$ (mean $\pm$ s.e.m.) scars per clone on average, and approximately 25% of the cells remained unscarred and therefore do not contain clonal information (Supplementary Fig. 3.1a). Clone sizes from gDNA and gDNA mRNA detection are very similar (Supplementary Fig. 3.1a-d), indicating that ScarTrace reliably detects scars from gDNA in single cells.

## Revealing clonal composition of haematopoietic cells with ScarTrace

We next used ScarTrace to explore the clonal composition of haematopoietic cells isolated from the whole kidney marrow (WKM) of two protein-injected (P1 and P2) and two RNA-injected (R1 and R2) zebrafish. We found one and two major clones in P1 and P2, and eight and six in R1 and R2, respectively (Fig. 3.2 a, b, Supplementary Fig. 3.2a, b, Supplementary Table 1). This is a direct result of the time window of Cas9 activity (Fig. 3.1c). The number of observed clones agrees with previous findings using GESTALT[84], in which a similar Cas9-mediated approach is used to label embryonic clones in zebrafish, and with the number of clones (between 10.4 and 15.4) found for haematopoietic stem and progenitor cells at 10-14 hours post fertilization (hpf) using Zebrabow[44]. The average number of scars per clone equals $3.3 \pm 0.3$ for P1, $1.02 \pm 0.01$ for P2, $3.5 \pm 0.3$ for R1 and $3.0 \pm 0.3$ for R2, with a minimum of 1 scar and a maximum of 5 scars per clone, revealing that both Cas9 protein and RNA efficiently cause scarring. We determined the copy number for each scar in a clone by modelling the amplification and sequencing noise of ScarTrace as a branching process (Supplementary Information section 3). Typically, the resulting number of scars per clone is smaller than eight, as a consequence of two or more simultaneously Cas9-induced cuts in the same multi-copy tandem histone-GFP gene[111]. We computed the P value of a combination of scars to occur in a cell (Fig. 3.2a, b). Values obtained are commonly below $10^{-6}$, emphasizing that although identical scars might be independently introduced in different clones

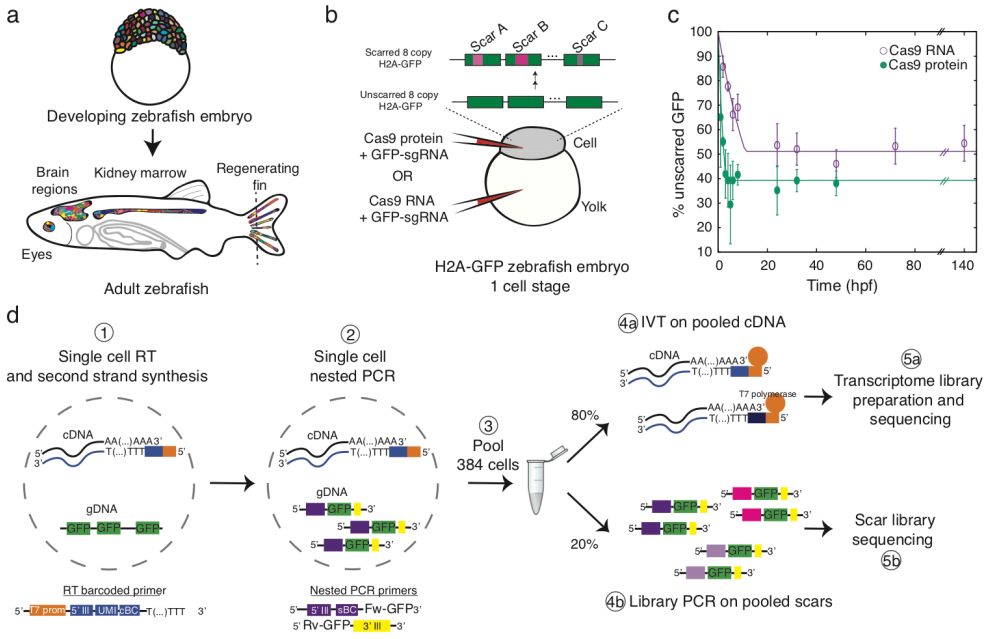



Figure 3.1: **Single-cell clonal tracing in zebrafish. a**, Embryonic cells get permanent and unique labels that are transmitted to the clones in the adult. **b**, Zygote injection with Cas9 RNA or protein and sgRNA that targets GFP (GFP-sgRNA). H2A, histone 2A. **c**, Mean fraction of unscarred GFP as a function of time, computed over ten independently injected embryos ($t > 6h$), and over three pools of ten embryos ($t \leq 6h$), in which GFP was PCR-amplified from gDNA. Error bars denote s.e.m. Unscarred GFP exponentially decreases at $0.064 \pm 0.002h^{-1}$ ($0.294 \pm 0.008h^{-1}$) and is constant after $10 \pm 1.0h$ ($3.1 \pm 0.1$h) for RNA (protein) injections (solid lines, Supplementary Information section 1). **d**, ScarTrace protocol (Methods). IVT, in vitro transcription; RT, reverse transcription. cBC, cell-specific barcode for transcriptome; sBC, cell-specific barcode for scars.

(for example, the yellow scar is present in one clone from fish R1 and four clones from fish R2), the chance of introducing the same combination of scars in independent clones is very small. Consistently, we do not find overlapping clones between different zebrafish. Using cell-to-cell variation in scar composition, we estimate a 90% scar detection efficiency (including unscarred GFP; Supplementary Fig. 1e, f). In addition, by assuming maximum parsimony for sequential scarring events, we build lineage trees for clones (Fig. 3.2c, d, Supplementary Fig. 3.2c, d, Supplementary Information section 4). Using RaceID[38] (Methods), we identify eight haematopoietic cell types in fish R1 and R2 (Fig. 3.2e). Gene expression profiles in the different cell types found for both fish are identical with the exception of erythrocytes, which show slight differences in the expression of characteristic markers (Supplementary Fig. 3.2e-h). After combining cell type and clonal information for

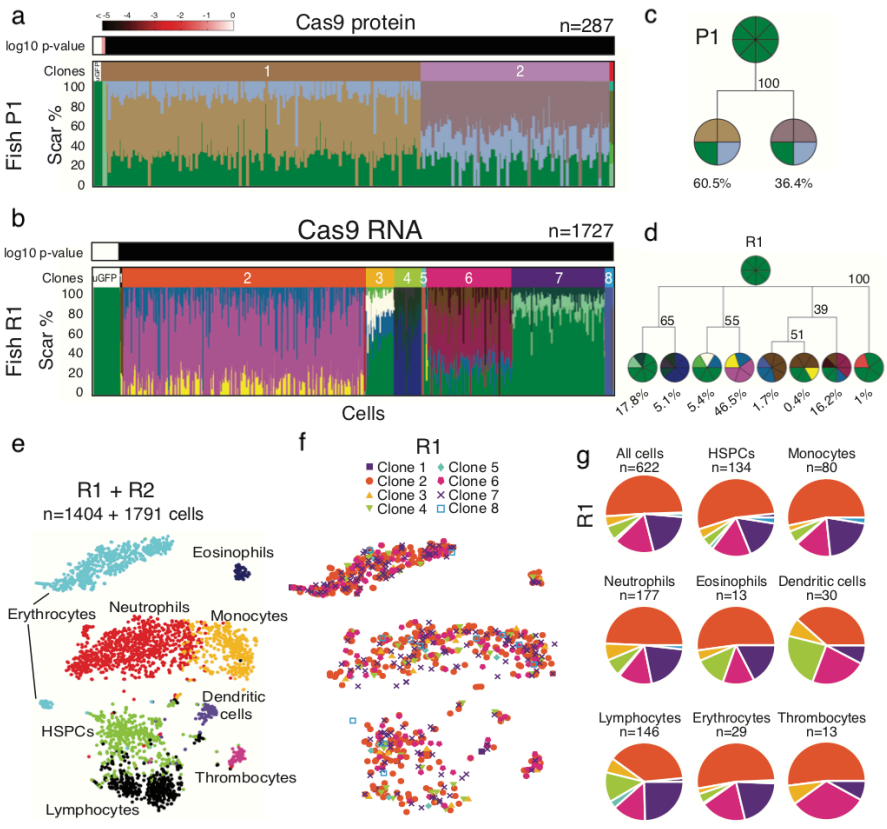



Figure 3.2: **Few clones produce haematopoietic cells.** **a**, **b**, Scar percentage per cell for fish P1 (a) and R1 (b) (for key, see Supplementary Fig. 3.2a). The bar above each panel indicates clones and corresponding P values. uGFP, unscarred GFP. **c**, **d**, Lineage trees for clones in P1 (c) and R1 (d). The root is an unscarred clone. Clones with corresponding cell fraction and scar copy number (Supplementary Information section 3) are at the tips. The statistical confidence of each branch is computed as its proportion among 10,000 bootstrapped tree replicates. Only clones with more than 2 cells are taken into account. **e**, t-distributed stochastic neighbour embedding (t-SNE) map of cells from fish R1 and R2 obtained with transcriptome data. Colours indicate the cell type (Supplementary Fig.3.2). HSPCs, haematopoietic stem and progenitor cells. **f**, t-SNE map of cells in fish R1. Colours indicate the clone. **g**, Clonal cell fraction per cell type for fish R1.

single cells, we observe all clones in all cell types with similar proportions (Fig. 3.2f, g, Supplementary Fig. 3.2i, j), indicating that all clones contribute to the production of all blood cells. This is consistent with haematopoietic stem and progenitor cells specification (around 28 hpf), when scarring is already completed[51].

## ScarTrace identifies early progenitors which later establish the left and right eye

Next, we used ScarTrace in the adult brain and eyes of two RNA injected fish (R2 and R3), in which we identified different neuronal, glia and immune cells (Fig. 3.3a, Supplementary Fig. 3.3). To determine clonal enrichment or depletion in certain cell types quantitatively, we used Fisher's exact test (Fig. 3.3b, Supplementary Fig. 3.4a). Here, several clones only generate neurons or retinal interneurons (Supplementary Fig. 3.4b, c). We observed that microglia share clones with the WKM, confirming that they originate from the WKM[145]. Upon the exclusion of WKM clones, we found that clones are not only cell-type specific, but also brain-region and eye specific (Supplementary Fig.s 4d-g, 5a-d). Although R2 and R3 left and right midbrains share a small fraction of clones, left and right eyes share none (Fig. 3.3c, Supplementary Fig. 3.4h). However, for fish P1, both midbrains share almost all clones whereas eyes share only one. To explore when this segregation is established, we injected one cell at the two-cell stage with Cas9-eScarlet fusion protein and sgRNA. We found Cas9-eScarlet protein present in only half of the embryo at dome stage (Supplementary Fig. 4i-l), approximately 3 h after Cas9 protein stops scarring. Therefore, scars only occur in one side of the embryo. However, ScarTrace on the left and right eyes of a 3-week-old-injected embryo (S1) reveals scars in both eyes. Upon removal of the clones found in immune cells and erythrocytes, the rest of the clones are specific to each eye (Fig. 3.3c). This indicates that both eyes get cellular contributions from both sides of the dome-stage embryo. To determine further when lateral commitment arises in eye progenitors, we built lineage trees for clones detected in the left and right eyes or midbrain for fish P1 and R2 (Supplementary Fig. 3.5e-h). In P1, no significant co-evolution is found among clones from the right (left) eye. By contrast, in R2 we observe a significant depletion of right eye clones evolving with left eye clones. This suggests that progenitors commit to the left or right eye shortly before the end of scarring with Cas9 protein. No significant co-evolution enhancement or depletion is found for clones detected in the left and right midbrain, indicating that cell mixing is important at 10 hpf. This is consistent with the processes of neurulation and neurogenesis[110].

## ScarTrace identifies tissue-resident macrophages upon fin regeneration

Next, we focused on zebrafish caudal fin ontogeny and regeneration. We performed ScarTrace on the primary, secondary and tertiary fins of fish R4, R5 and R6 (Fig. 3.4a). We identified four major cell types (osteoblasts, mesenchymal, epidermal and immune cells) and observed cell-type-restricted clones in all fish (Fig. 3.4b, c, Supplementary Fig. 6, 7a-e). We found that mesenchymal and epidermal cells share clones, revealing a common developmental origin that is maintained during regeneration. Together with previous imaging-based studies[71], this suggests that

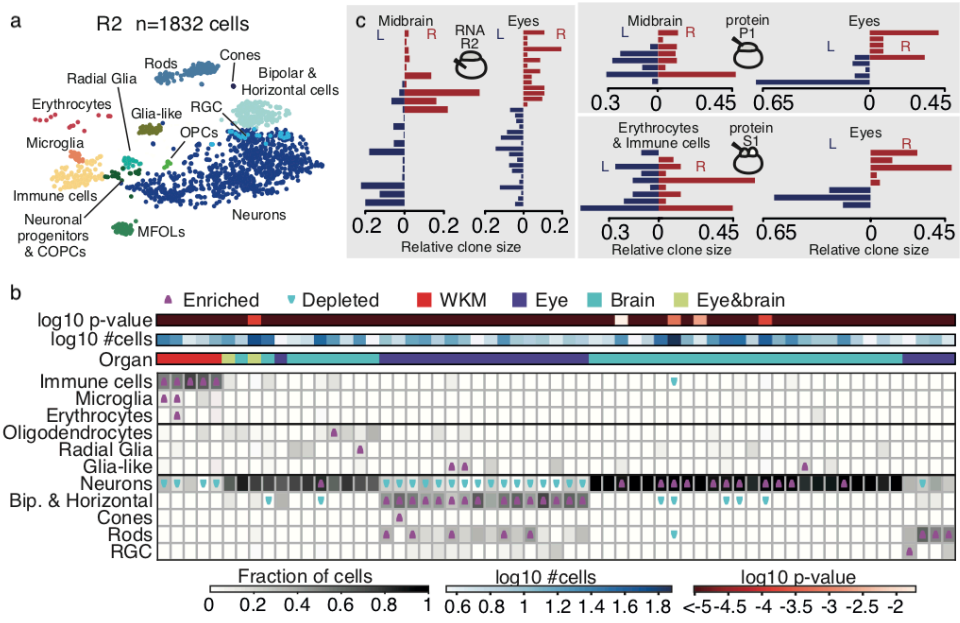



Figure 3.3: **Clonality in the brain and eyes.** **a**, t-SNE map of cells in fish R2. Colours indicate cell type (Supplementary Fig. 3.3). COPCs, committed oligodendrocyte progenitor cells; MFOLs, myelin-forming oligodendrocytes; OPCs, oligodendrocytes progenitors; RGC, retinal ganglion cell. **b**, Heat map of clonal cell fraction for cell types in fish R2 (COPC, OPC and MFOL clones merged as oligodendrocytes), and two-sided Fisher's exact test for enriched and depleted clones per cell type with $P < 0.05$. The bars at the top depict organ, total number of cells, and P value for each clone. Bip., bipolar; RGC, retinal ganglion cell. **c**, Relative clone frequency in the left (L) and right (R) eye and midbrain for fish R2 and P1, and left and right eye and erythrocytes and immune cells for fish S1.

epidermal ancestors undergo epithelial-to-mesenchymal transition during gastrulation to generate mesenchymal cells in the caudal fin. Osteoblasts did not share clones with any other cell type in the primary fin and showed dorsal-ventral segregation, confirming their early lineage commitment during development[52,61,116,131,135]. We found lineage restriction of the different cell types as the main mechanism of fin regeneration, consistent with previous results[131,135]. However, after regeneration, we observed osteoblast-committed clones that generate a fraction (approximately 21% in R4, 44% in R6) of mesenchymal cells (Fig. 3.4d, Supplementary Fig. 3.7f). This suggests a certain degree of plasticity after injury, in which progenitors that produce osteoblasts during development can also give rise to mesenchymal cells during fin regeneration[132].

Finally, we investigated the clonal overlap of single cells from the WKM of fish R4, R5 and R6 with immune cells found in the fin. Clones detected in the WKM are

enriched in the fin immune cells and depleted in the remaining cell types (Fig. 3.4c, e, Supplementary Fig. 3.7g). We found sub-populations of lymphoid and myeloid cells in all fins with different proportions of fin-specific clones, which we identify as resident immune cells (RICs). Differential gene expression analysis in myeloid cells revealed that subpopulation 4 expressed macrophage markers together with the epithelial marker epcam (Fig. 3.4f), which has been reported in resident macrophages in mice[36]. All RICs in the primary fin share clonality with epidermal and mesenchymal cells (Fig. 3.4g, Supplementary Fig. 3.7b, c). Therefore, our data indicate that RICs have a distinct origin from haematopoietic stem cells (Supplementary Fig. 3.7h), and arise either from epidermal and mesenchymal transdifferentiation, or from ectodermal ancestors similarly to mesenchymal cells. We developed ScarTrace as a new method to quantify clonal origin and cell type simultaneously at single-cell resolution. This enabled us to investigate the embryonic origin of clones found in different organs of the adult zebrafish and their cell-type commitment during development and regeneration. CRISPR-Cas9 genome editing technology for lineage tracing purposes at the single cell level has recently also been used in zebrafish to investigate lineages and cell types in the vertebrate brain, and to unravel developmental lineages[103,119]. We anticipate many applications of ScarTrace in developmental and stem-cell biology, and similar approaches to study clonal selection in cancer models. Because ScarTrace provides a glimpse of the cellular past, it will be interesting to explore how this history is predictive of the current epigenetic and expression state.

**Clonality during caudal fin regeneration. a**, Primary (first week), secondary (fully regenerated after 3 weeks) and tertiary (fully regenerated after 6 weeks) fins are amputated. Tertiary fins are split dorsally and ventrally. textbfb, t-SNE map of cells in fish R4, R5 and R6. Colours indicate cell type (Supplementary Fig. 3.6). **c**, Heat map of clonal cell fraction per cell type and fin in fish R4, and two-sided Fisher's exact test for clones that are enriched and depleted with $P < 0.05$. The bars at the top show clones in the WKM, number of cells and P value per clone. **d**, t-SNE map of cells from fish R4, with cells from the primary (top) or regenerated (bottom) fin detected in osteoblast clones or remaining cell types. Percentages are mesenchymal cell fractions that share clones with osteoblasts. **e**, Magnified view of t-SNE map indicating immune cells with fin-specific clones, immune cells that share clones with the WKM and cells with no clone information (grey). Subpopulations of lymphoid cells and myeloid cells (numbered 1-4), and the percentage of RICs are indicated. **f**, Differential gene expression between the four subpopulations of myeloid cells (Supplementary Table 6). **g**, Cell type fraction for fin-specific clones detected in RICs in the primary fin for R4.

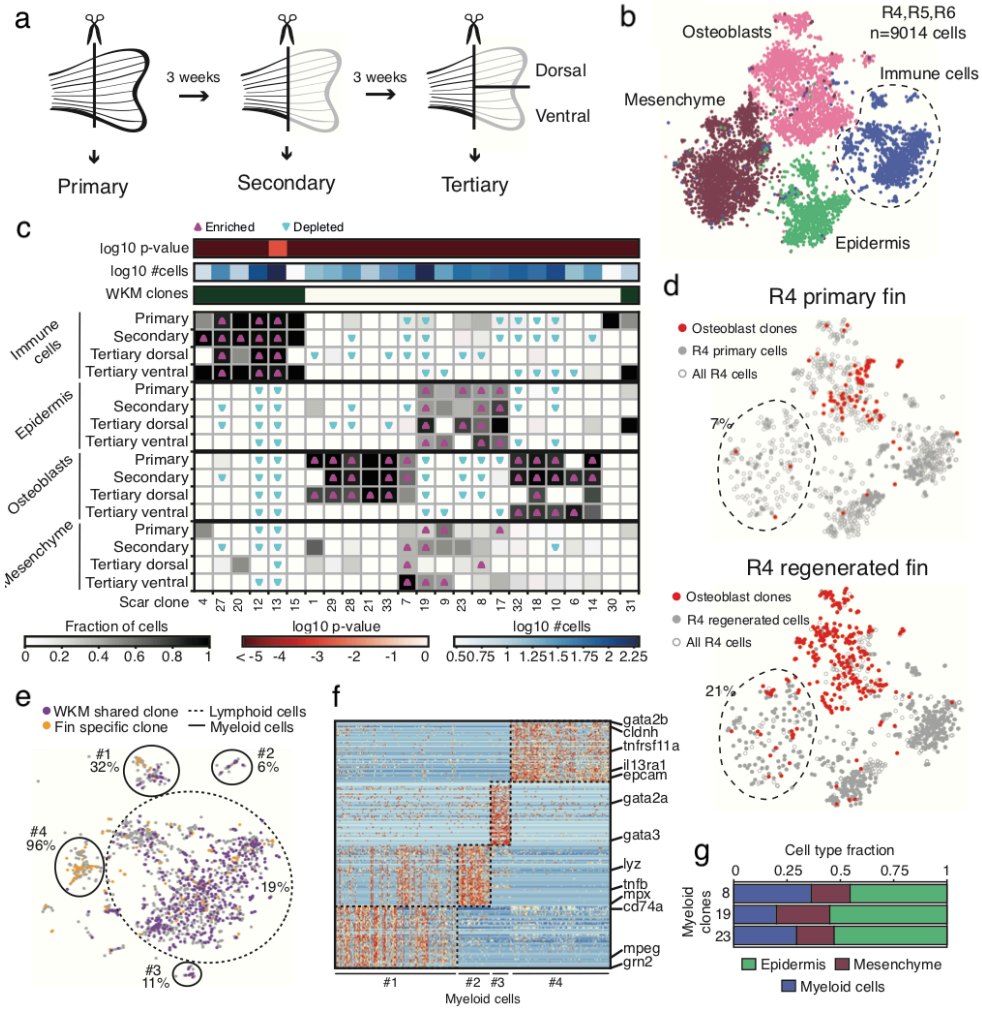



Figure 3.4: **Figure caption on previous page.**

# Methods

## Zebrafish Cas9 and sgRNA injections

Heterozygous zygotes of the transgenic zebrafish line Tg(h2afva:GFP)$^{kca66}$:(h2afva:GFP)$^{kca66\,96}$ were injected at the cell with 1 nl Cas9 protein (NEB; final concentration 1,590 $ng\mu l^{-1}$) or at the yolk with 1 nl Cas9 RNA (300 $ng\mu l^{-1}$) in combination with an sgRNA that targets GFP (25 $ng\mu l^{-1}$), sequence: GGTGTTCTGCTGGTAGTGGT) (Fig. 3.1b). Cas9 RNA was in vitro transcribed from a linearized pCS2-nCas9n vector (Addgene plasmid 47929)[49] using the mMESSAGE mMACHINE SP6 Transcription Kit (Thermo Scientific).

The sgRNA was in vitro transcribed from a template using the MEGAscript. T7 Transcription Kit (Thermo Scientific). The sgRNA template was synthesized with T4 DNA polymerase (New England Biolabs) by partially annealing two single-stranded DNA oligonucleotides containing the T7 promotor and the GFP-binding sequence, and the tracrRNA sequence, respectively[142]. Male and female zebrafish were used, no randomization was done, no blinding was done and no animals were excluded from the analysis. No statistical methods were used to predetermine sample size. The age of the fish used in isolated organ spans 3-18 months. For sample sizes, see Supplementary Table 3.2. All animal experiments were performed in accordance with institutional and governmental regulations, and were approved by the Dier Experimenten Commissie of the Royal Netherlands Academy of Arts and Sciences and performed according to the guidelines.

## Transgene copy number

To determine the number of integrations of the transgene, we performed whole-genome sequencing (NEBNext Ultra library preparation kit for Illumina (E7370S) and the NEB Multiplex Oligos for Illumina (E7500L)) on an homozygous Tg(h2afva:GFP)$^{kca66}$:(h2afva:GFP)$^{kca66}$ fish. Paired-end data were trimmed (TrimGalore-0.4.3) and mapped (bwa-0.7.10 mem) to the zebrafish reference genome (danRer10 from UCSC Genome Browser), and PCR and optical duplicates were removed (Picard-2.0.1) (Supplementary Fig. 3.8a, b). The copy number was extracted using FREEC-11.0[12] with default parameters. With a 1-kb window size, we find 19 copies of the transgene fragment, whereas with a 500-bp window size, we find 18 (Supplementary Fig. 3.8b). After correcting for reads due to endogenous copies, we estimate the number of copies of the transgene in a heterozygous fish to be $8 \pm 1$. This number agrees with single-cell data, because although we detected a maximum of 7 scars per clone (Fig. 3.8c, d), we see that sometimes 6 of the scars in those clones represent approximately 12.5% of the scar content per cell, and one represents approximately 25% (Supplementary Fig. Fig. 8e). This again suggests that the number of integrations of the histone-GFP transgene is 1/0.125 = 8.

## ScarTrace protocol

Live single cells (based on DAPI exclusion and scatter properties) were sorted into 384-well plates (Sigma-Aldrich) containing 5 $\mu l$ of mineral oil (Sigma-Aldrich), 50 nl of uniquely barcoded reverse transcription primers (Supplementary Table 3.3), dNTPs (Promega), Spike-in controls (Thermo Fisher) and RNase inhibitor (SUPERaseIn, Thermo Fisher). Plates were immediately spun down and stored at -80°C. Cells were lysed at 65°C for 5 min. Reverse transcription and second-strand

synthesis mixes were dispensed into each well using the Nanodrop II and reactions were performed at 42 and 16°C degrees, respectively (Fig. 3.1d, step 1). Genomic DNA was access by proteinase K treatment followed by a nested PCR strategy to amplify the scarred GFP region (Fig. 3.1d, step 2). In the second PCR, unique scar barcodes were introduced in each well (Supplementary Table 2). All cells were pooled and the aqueous phase was separated from the oil phase (Fig. 3.1d, step 3). The collected material was split for scar library and transcriptome library preparation (Fig. 3.1d, step 4). For transcriptome library preparation, the SORT-seq protocol[89] was used (Fig. 3.1d, step 5a). For scar library preparation, a PCR introducing only Illumina TruSeq adapters was performed (Fig. 3.1d, step 5b). All libraries were sequenced paired-end at 75 bp read-length on the Illumina NextSeq platform. A detailed description of the protocol is available in Protocol Exchange[4].

### WKM isolation

The WKM was isolated as previously described[120]. A ventral midline incision was made to open the adult zebrafish body cavity. All internal organs were carefully removed to access the kidney. The WKM was collected in PBS supplemented with FCS. The tissue was aspirated through a 1 ml pipet tip several times to mechanically dissociate haematopoietic cells. After two consecutive filtering steps (using 70-$\mu m$ and 40-$\mu m$ cell strainers (VWR), cells were centrifuged and washed. The pellet of haematopoietic cells was resuspended in PBS and FCS supplemented with DAPI (Thermo Fisher) to assess cell viability

### Brain parts and eye isolation

Brain and eyes were isolated from the zebrafish head and dissected in PBS. Optic nerves were removed. The forebrain (olfactory bulb and telencephalon) was isolated from the midbrain, followed by dissection of the hindbrain (rhombencephalon). The midbrain (mesencephalon) was dissected into left and right midbrain. The eyes lens was carefully removed. Brain parts and eyes were dissociated into single cells using a papain-based solution (Thermo Fisher, 88285) and washing solutions as previously described[74]. The washed cell pellet was resuspended in DMEM/F12 medium (Thermo Fisher, 11320033) and supplemented with DAPI (Thermo Fisher) to assess cell viability for FACS.

### Fin amputation

Caudal fin amputations were performed as previously described[101], after which fish were returned to 28°C aquarium water. Once isolated, this tissue was immediately dissociated by moderately shaking at 30°C for 1 h, with gentle trituration performed

every 10 min with a p200 pipet, in a solution of 2 mg $ml^{-1}$ collagenase A (Sigma-Aldrich) and 0.3 mg $ml^{-1}$ protease (type XIV, Sigma-Aldrich) in Hanks solution. After 1 h, the solution was incubated for 5 min in 0.05% trypsin in PBS. The solution was strained using 70-$\mu m$ and 40-$\mu m$ cell strainers (Corning) and cells were washed in 2% FBS in Hanks solution. Before flow cytometry, cells were centrifuged and resuspended in PBS and FBS supplemented with DAPI (Thermo Fisher) to assess cell viability.

### Transcriptome analysis

In transcriptome libraries, the first read contains cell barcode (Supplementary Table 3.3) and unique molecular identifier (UMI) information, and the second read contains biological information. Second reads with a valid cell barcode extracted from corresponding first reads are mapped using bwa mem-0.7.10 with default parameters to the reference zebrafish transcriptome (Danio rerio assembly Zv9, ensemble 74, extended with ERCC92). For each cell, the number of transcripts per gene was obtained as previously described[38]. We refer to transcripts as unique molecules based on UMI correction. We ran RaceID3 with different parameters for each organ under study (Supplementary Data 1) for cell filtering, normalization, gene filtering, cell clustering and differential gene expression analysis (in which P values are calculated using negative binomial distribution and corrected for multiple testing by the Benjamini-Hochberg method). The choice of filtering parameters was made to include the maximum number of cells in our analysis without losing cell type information. Supplementary Tables 3.8, 3.9 and 3.10 provide results for the differentially expressed genes for each cell type compared with all other cells in the organ: WKM[23,62,76,87] (90 dendritic cells, 76 eosinophils, 641 erytrhocytes, 516 haematopoietic stem and progenitor cells, 446 lymphocytes, 409 monocytes, 927 neutrophils and 76 thrombocytes), brain and eyes[18,26,31,46,66,78,91,94,128,147] (250 bipolar and horizontal cells, 45 COPCs, 9 cones, 290 erythrocytes, 254 immune cells, 88 glia-like cells, 89 MFOLs, 66 microglia, 1,427 neurons, 10 OPCs, 31 RCL, 53 radial glia and 202 rods), caudal fin[2,77,118] (144 epidermal cells, 2,834 fibroblasts, 1,784 immune cells, and 2,951 osteoblasts), and resident myeloid cell types in the fin (118 cells in subpopulation 1, 45 in subpopulation 2, 27 in subpopulation 3 and 133 in subpopulation 4).

### Scar analysis

In scar libraries, the first read contains the cell barcode (Supplementary Table 2) and the forward primer used in the nested PCR and second read contains the sequence for the scar and the reversed primer. Scripts to extract scars and detect clones are provided as Supplementary Data 2, together with a reference manual

(Supplementary Data 3). Bug fixes and updates of the scripts can be downloaded from https://github.com/anna-alemany/scScarTrace. Cells sharing an identical scar pattern are assumed to come from the same clone, independently of scar percentage. Cells with a detected scar pattern that can be assigned to another single clone by assuming that some scar was not sampled were pooled with that clone. Cells that according to their scar pattern can be ambiguously assigned to two or more other clones were removed from subsequent analysis. Clones with less than three cells were also removed.

## Code availability

Transcriptome analysis was performed using RaceID3 available at https://github.com/dgrun/RaceID3_StemID2, with parameters summarized in Supplementary Data 1. Scripts for scar extraction and clone detection are provided in Supplementary Data 2, together with a reference manual (Supplementary Data 3). Bug fixes and updates of the scripts can be downloaded from https://github.com/anna-alemany/scScarTrace.

## Data availability

The accession numbers for the RNA sequencing datasets reported in this paper have been deposited with the Gene Expression Omnibus (GEO) under accession GSE102990.

## Acknowledgements

This work was supported by a European Research Council Advanced grant (ERC-AdG 742225-IntScOmics), Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO) TOP award (NWO-CW 714.016.001), and the Foundation for Fundamental Research on Matter, financially supported by NWO (FOM-14NOISE01). This work is part of the Oncode Institute which is partly financed by the Dutch Cancer Society. We thank M. Sen for help with sequencing, R. der Linden for cell sorting, and B. de Barbanson for help with programming, and all the other members of the A.v.O. laboratory for discussions and input. In addition, we thank B. Artegiani and J. Bakkers for discussions, P. Shang and N. Geijsen for sharing the Cas9-eScarlet fusion protein, A. Klaus, the Hubrecht Sorting Facility, and the Utrecht Sequencing Facility, subsidized by the University Medical Center Utrecht, Hubrecht Institute and Utrecht University.

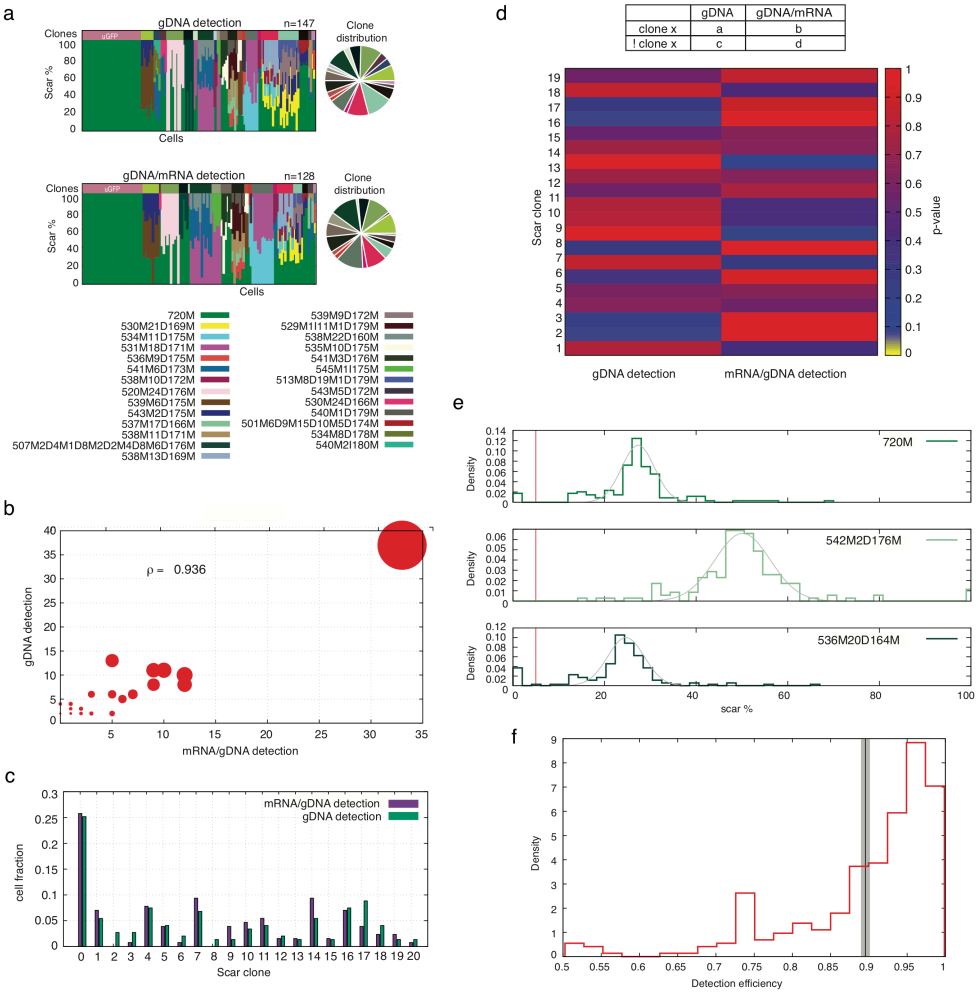
## Author Contributions

A.v.O. conceived and designed the project. J.P.-M. developed the experimental protocol, with support from A.A., M.F. and C.S.B. C.S.B. performed WKM-related ex-

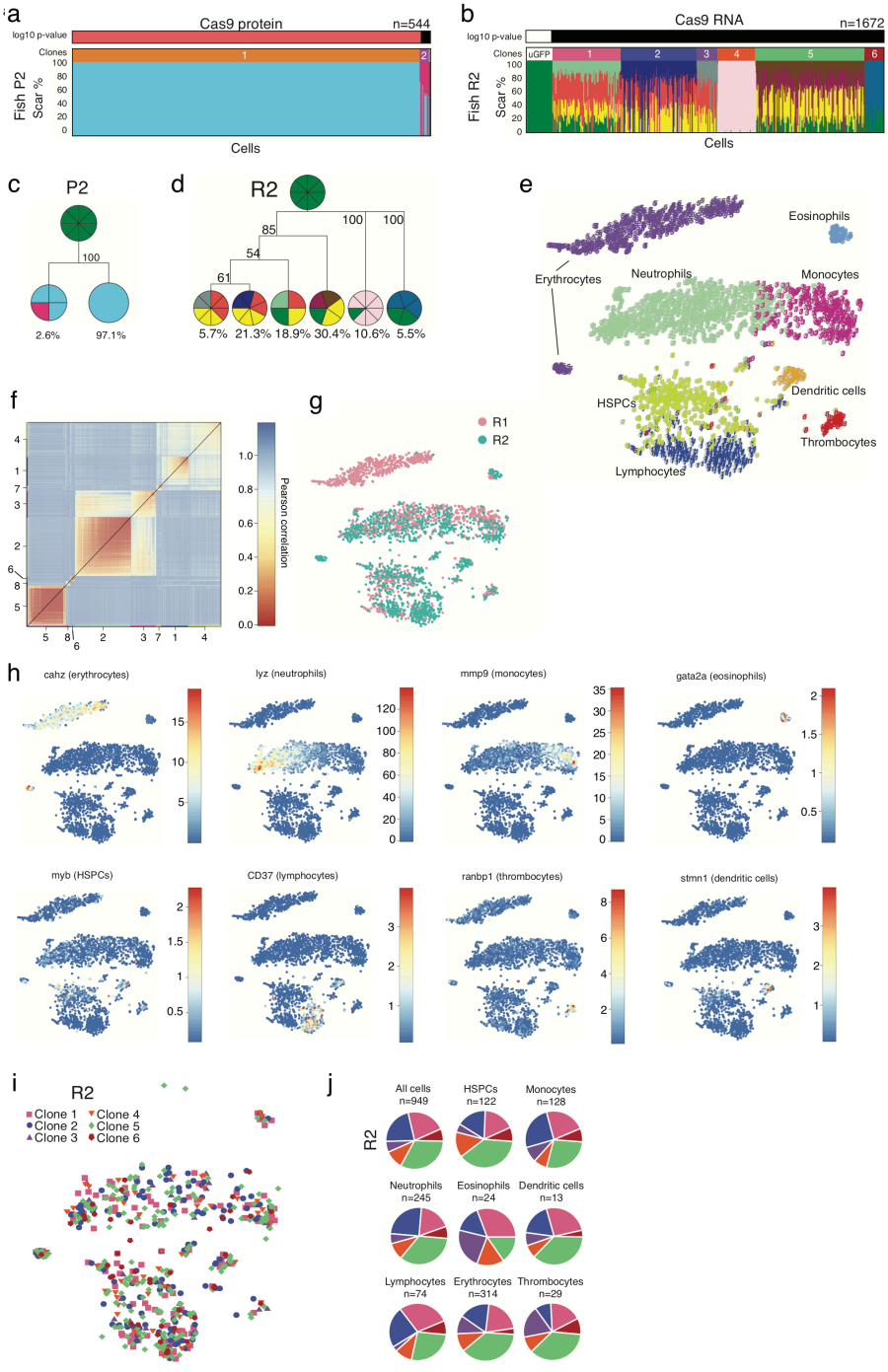
periments; M.F. performed brain- and eye-related experiments; and C.S.B. and J.P.-M. performed fin-related experiments. A.A. developed the computational methods and modelling. A.A., C.S.B. and A.v.O. analysed WKM-related data; A.A. and M.F. analysed brain- and eye-related data; A.A., C.S.B. and J.P.-M. analysed fin-related data. All authors discussed and interpreted results, and wrote the manuscript. A.A., M.F., C.S.B. and J.P.-M. contributed equally to this work.

## Supplementary Material

**gDNA versus gDNA-mRNA detection of scars. a**, Scar percentage per cell (top bar indicates clones), and pie chart of fraction of cells per clone (colours matching histograms' top bar) detected via gDNA (ScarTrace without step 1) and gDNA-mRNA detection (full protocol). **b**, Number of detected cells per clone in gDNA versus gDNA-mRNA detection and Pearson's correlation coefficient computed using the 20 different clones identified. Dot sizes are proportional to the total number of cells found taking together the two detection strategies. **c**, Fraction of cells detected per clone in gDNA (green) and gDNA-mRNA (purple) detection, in which clone '0' represents unscarred cells. **d**, Top, one-sided Fisher's exact test on a contingency table made of the number of cells detected for the given scar clone x for each detection strategy (a and b, respectively), and the number of cells taking together all other clones (c and d, respectively) found in gDNA detection (n = 147 cells in total) and gDNA-mRNA detection (n = 128 cells in total). Bottom, heat map shows the one-sided P value of each scar clone to be enriched in the gDNA or mRNA/gDNA detection protocol. No enrichment is found with $P < 0.05$, therefore results found for the two detection protocols are compatible. **e**, Normalized histograms and corresponding fit noise model function (grey line; Supplementary Information section 4) for the scar percentage detected for 1 clone found in fish P1. Scar detection efficiency is defined as the area above 5% scar content (vertical red line). Efficiency of detection of unscarred molecules is assumed to be the same as for scarred molecules. **f**, Normalized histogram of the scar detection efficiencies found after pooling all clones from all organs for all fish (in total, n = 371 detected clones; Supplementary Table 3.1). The vertical black line and the grey area indicate the mean scar detection efficiencies and s.e.m., respectively.



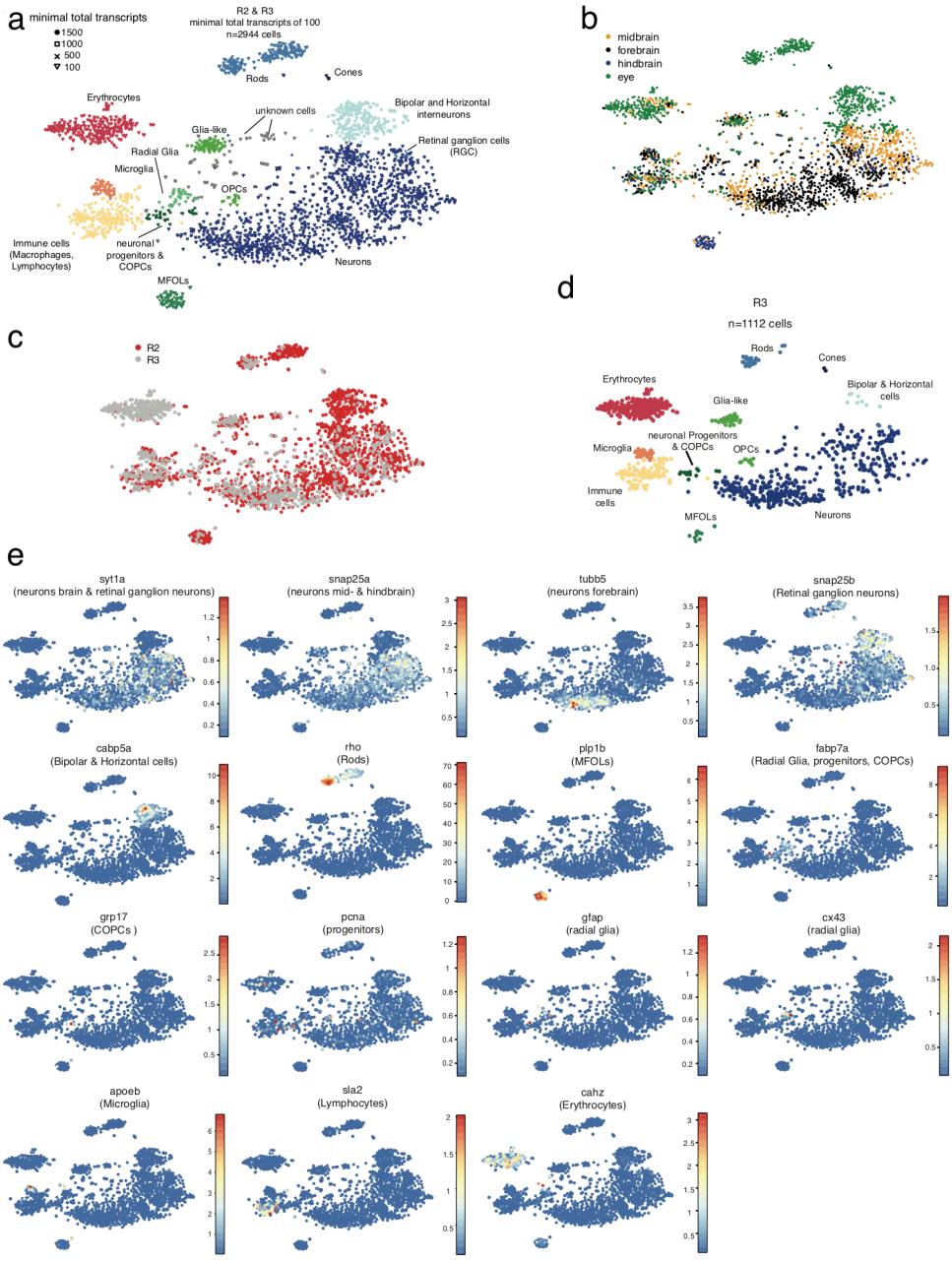

Supplementary Figure 3.1: Figure caption on previous page.



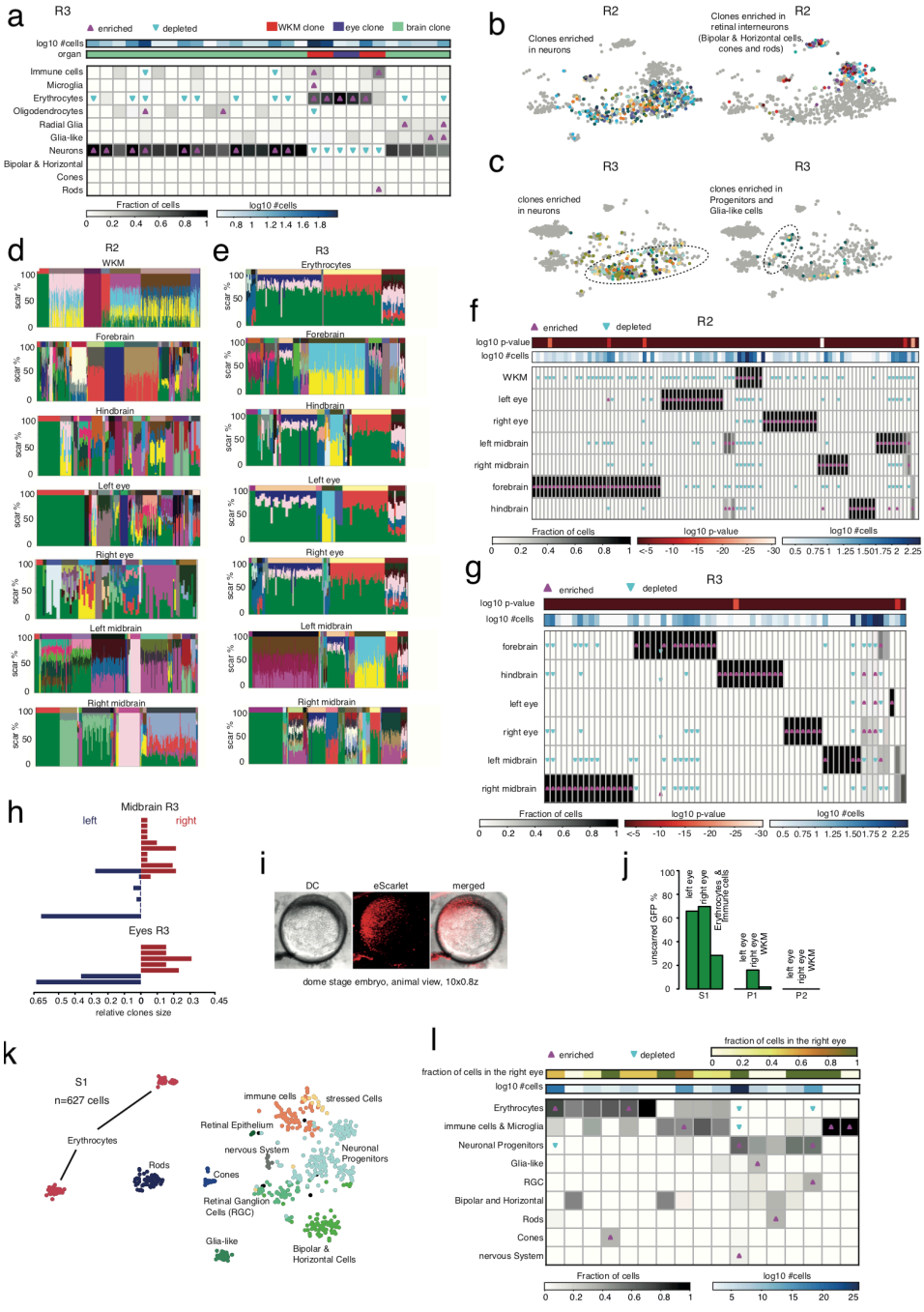

Supplementary Figure 3.2: Figure caption on next page.

**Transcriptome analysis of the zebrafish WKM. a, b,** Scar percentage per cell for fish P2 (a) and R2 (b). The bar above each panel indicates clones and corresponding P values. **c, d,** Lineage trees for clones detected in P2 (c) and R2 (d) obtained as described in Fig. 2. **e**, t-SNE map of cells from fish R1 and R2. Colours and numbers indicate RaceID clusters. **f**, Heat map of the Pearson correlation between cells sorted according to RaceID clusters. Cluster numbers are indicated on the x and y axes. **g**, t-SNE map for the WKM of R1 and R2 coloured according to fish of origin (R1 in pink and R2 in green). All cell types intermingle well, except erythrocytes. Even though erythrocytes appear separately on the t-SNE space, they belong to the same RaceID cluster. textbfh, t-SNE maps for R1 and R2, coloured according to the number of unique transcripts per single cell for each marker[23,62,76,87]. A full list of marker genes for each cell type is available in Supplementary Table 3. **i**, t-SNE map for fish R2 with cells coloured according to clone. j, Clonal cell fraction per cell type for fish R2.

**Cell types and batch effects in the brain and eyes for fish R2 and R3.** **a**, t-SNE map obtained with RaceID of pooled cells with a minimum of 100 total transcripts from fish R2 and R3 (isolated from the right and left eyes, right and left midbrain, forebrain and hindbrain). Different symbols indicate cells with different minimal total transcript counts. Cells are coloured according to the assigned cell type using the lowest cut off (that is, taking into account cells with at least 100 transcripts). We do not lose any cell type cluster when applying higher transcript cut offs, nor do we generate new clusters of low transcript cells when applying lower cut offs. The fraction of cells that would be termed a different cell type with a higher cut off is very low (< 1%). Low transcript cells cluster mainly around the clusters formed by high transcript cells. **b, c**, t-SNE maps as in a, in which cells are coloured according to organ (b) and fish (c) of origin. **d**, t-SNE map as in a, but showing only cells from fish R3, with corresponding cell types indicated. **e**, t-SNE maps for fish R2 and R3 coloured according to the number of unique transcripts per single cell [18,26,31,46,66,78,91,94,128,147]. A full list of marker genes for each cell type is available in Supplementary Table 4.



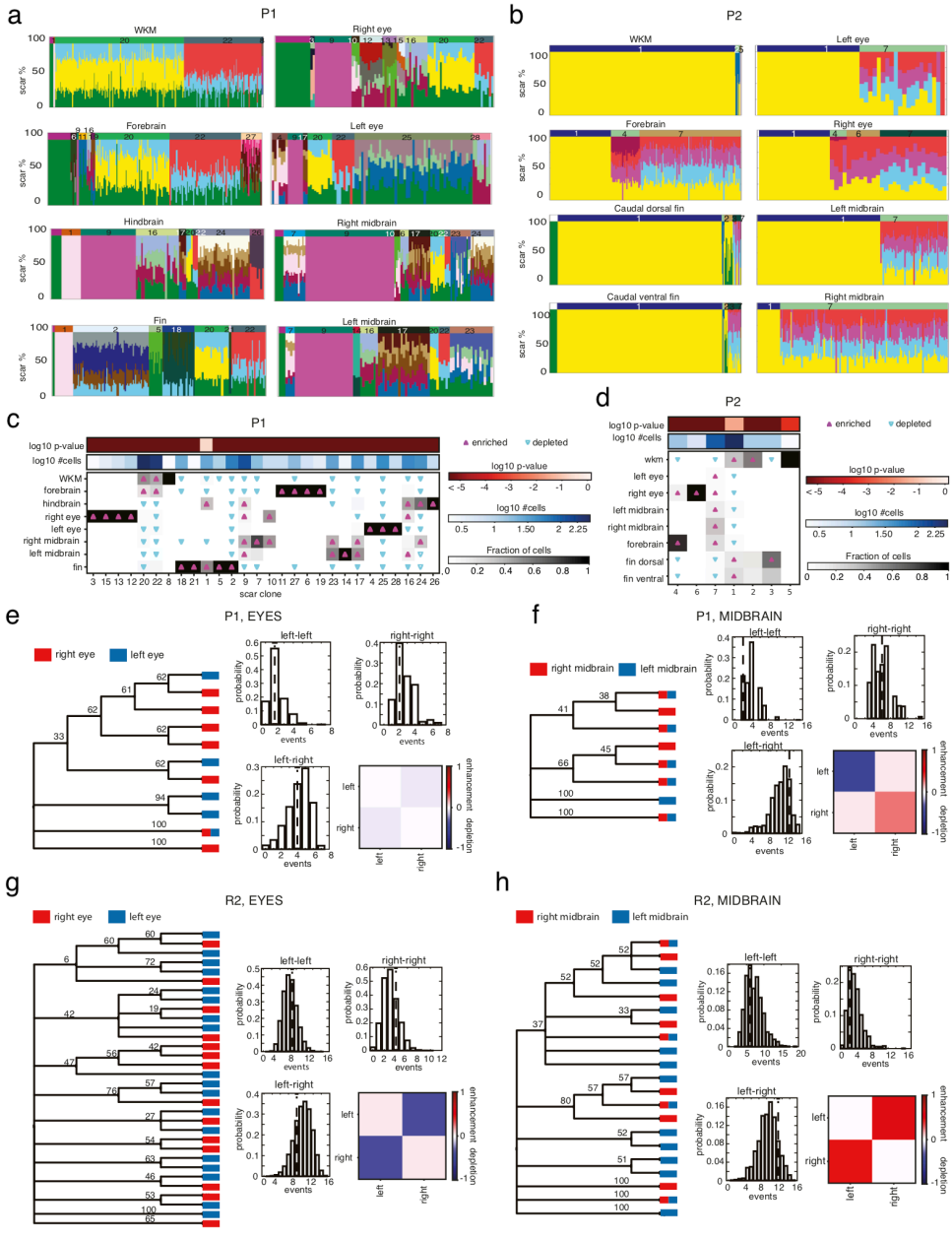

Supplementary Figure 3.3: Figure caption on previous page.



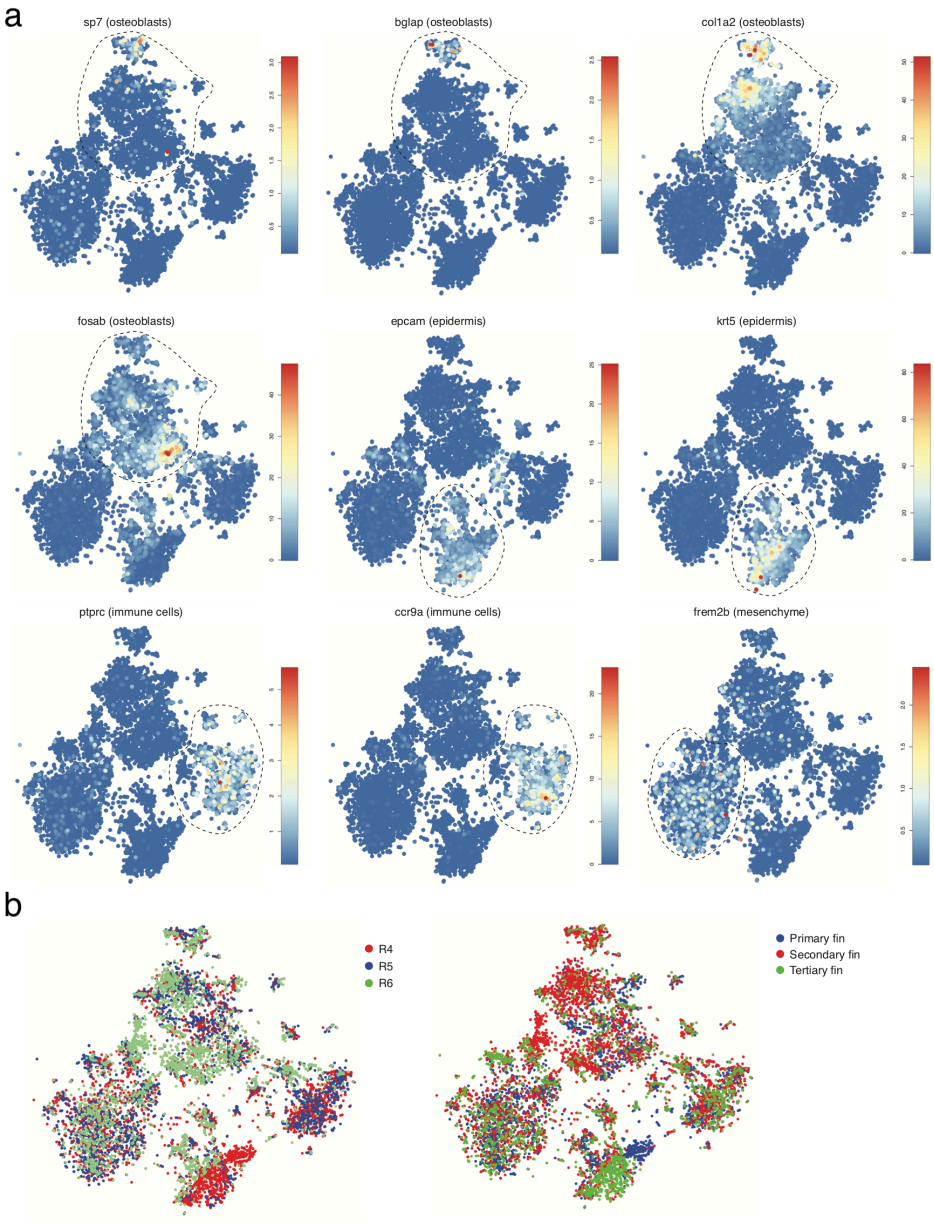

Supplementary Figure 3.4: Figure caption on next page.

**ScarTrace in the zebrafish brain and eyes. a**, Heat map of the fraction of cells per clones for cell types in fish R3 (COP, OPC and MFOL clones merged as oligodendrocytes), and two-sided Fisher's exact test for enriched (magenta upwards triangle) and depleted (blue downwards triangle) clones per cell type with $P < 0.05$. The bars at the top depict organ and corresponding total number of cells. All clones have P values $< 10^{-5}$. **b, c**, t-SNE map of fish R2 and R3 cells showing different colours for enriched clones detected in glia cells, neurons or retinal interneurons. Other cells are shown in grey. **d, e**, Scar percentage per cell for clones found in the WKM, forebrain, hindbrain, left and right eyes, and left and right midbrain for R2 and R3. In all panels, each colour represents the same scar (for example, the yellow scar is the same for R2 and R3), and unscarred GFP is shown in green. **f, g**, Heat maps of the fraction of cells per clones for each organ for R2 (f) and R3 (g). Enriched (magenta upwards triangles) and depleted (blue downwards triangles) scar clones per organ are determined by a two-sided Fisher's exact test with $P < 0.05$. The bar above each panel depicts the number of cells and P value for each clone. **h**, Histograms of the relative clone frequency in the left (blue) and right (red) midbrain and eye for R3. **i**, Image of domestage embryo injected with Cas9-eScarlet in one cell at the two-cell stage ($n > 10$ embryos showed similar patterns). BF, bright-field. **j**, Scarring efficiency shown as the percentage of unscarred GFP in S1, P1 and P2 for the left and right eyes, and the WKM. **k**, t-SNE map of cells isolated from the left and right eyes of S1, in which cells are coloured according to their cell type. **l**, Heat map of the fraction of cells per clones for each cell type in S1. Enriched (magenta upwards triangle) and depleted (blue downwards triangle) scar clones per cell type are determined from a two-sided Fisher's exact test with $P < 0.05$. The bars at the top depict the total number of cells and the fraction of cells found in the right eye for each clone. All P values are below $10^{-5}$.

**Clones for fish P1 and P2 and lineage tree of the eyes and midbrain. a, b**, Scar percentage per cell for clones found in the WKM, forebrain, hindbrain, left and right eyes, and left and right midbrain for P1 (a) and P2 (b). Each colour represents a different scar, in which unscarred GFP is always shown in green. Colour legend per scars is different between panels (yellow scar in a is not yellow scar in b). **c, d**, Heat maps of the fraction of cells per clones for each organ for P1 (c) and P2 (d). Enriched (magenta upwards triangles) and depleted (blue downwards triangles) scar clones per organ are determined from a two-sided Fisher's exact test with $P < 0.05$. The bar above each panel depicts the number of cells and P value for each clone. **e-h**, Lineage trees obtained assuming the principle of maximum parsimony as described in Supplementary Information section 5 for clones detected in the right and the left eyes (e, g) and right and left midbrain (f, h) of fish P1 (e, f) and R2 (g, h). The root of the trees is set as an unscarred clone, with eight copies of the GFP transgene. In the tips there are the detected clones. The statistical confidence of each branch is computed as the proportion of each branch among 10,000 tree replicates constructed by bootstrapping scars present in all clones. To assess statistically whether clones from the left or the right side co-evolve together, we randomized the clones at the tips of the tree and checked how many times, randomly, clones from the right or the left were found to be sisters with other clones from the right or the left. This allowed us to build a distribution of co-evolution (histograms in each tree) of clones for the null hypothesis and check whether the number of times we saw clones from one side together was statistically significant or not. The vertical dashed line in each histogram indicates the number of times we see clones from one side together as sisters in the reference tree. When such line is found at the right-hand (left-hand) side of the maximum, we assume that the co-evolution of the clones is enhanced (depleted). In the heat maps, we indicate the degree of co-evolution of clones in the right or the left eye or midbrain, computed as the fraction of the area of the histogram at the right- or the left-hand side (that is, enhanced or depleted co-evolution, respectively) of the vertical line divided by the corresponding area of the histogram at the right- or left-hand side of maximum of the distribution.



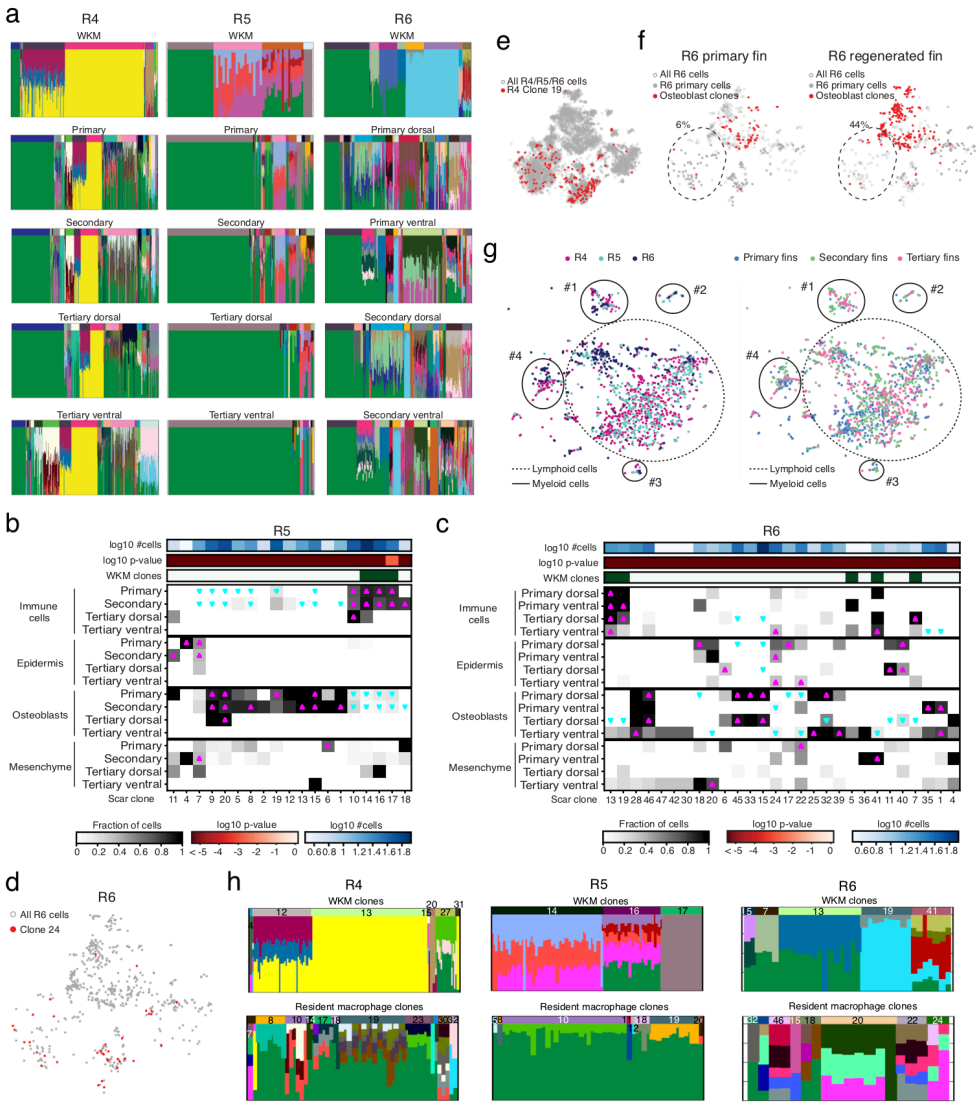

Supplementary Figure 3.5: Figure caption on previous page.



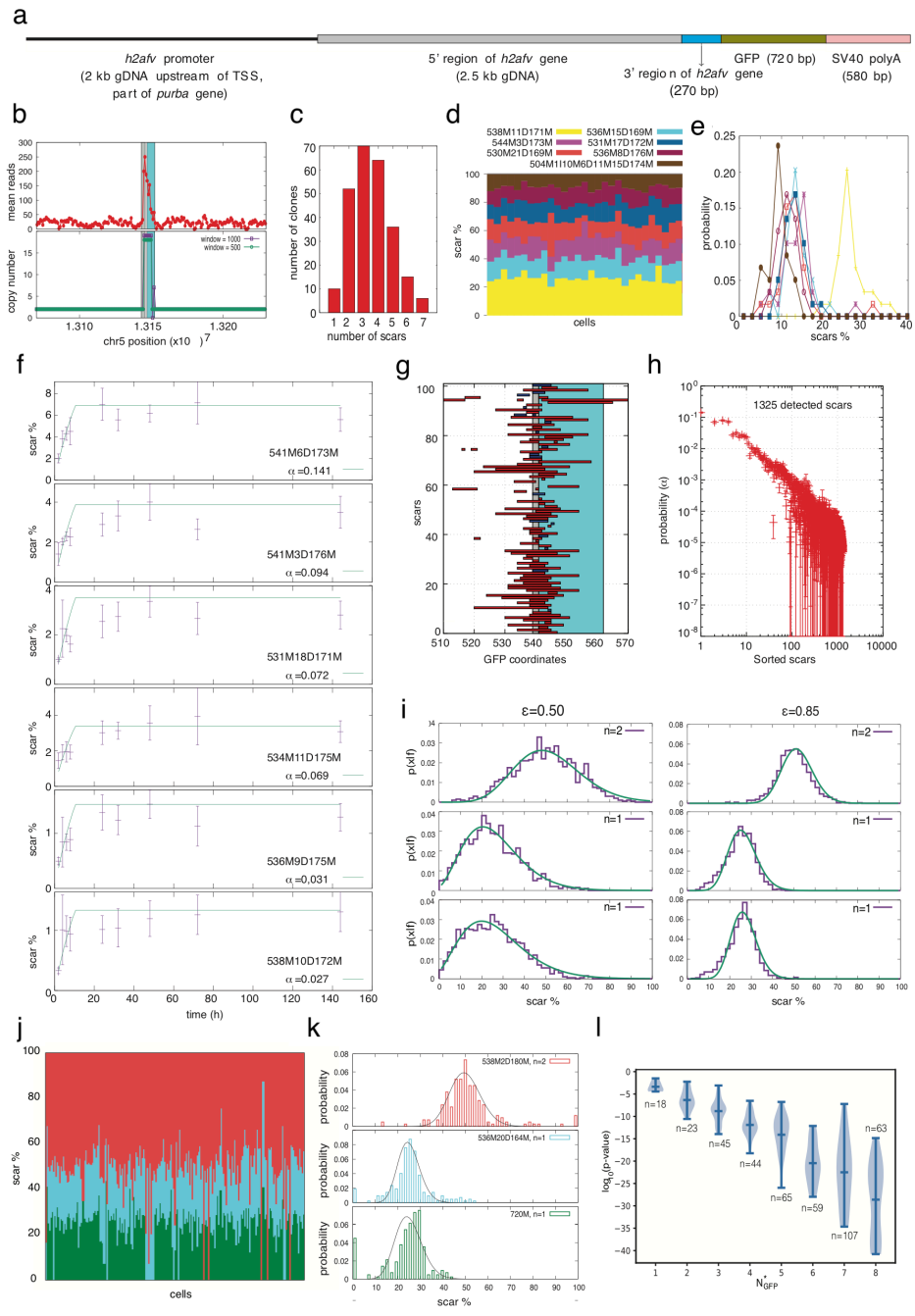

Supplementary Figure 3.6: Figure caption on next page.

**Transcriptome analysis of the zebrafish caudal fin.** **a**, t-SNE maps obtained by pooling together cells from primary and regenerated fins from fish R4, R5 and R6. In each panel, single cells are coloured according to the number of unique transcripts observed for a given gene. The corresponding cell type is indicated in parenthesis [2,77,118]. A complete list of marker genes used for each cell type is available in Supplementary Table 5. **b**, t-SNE maps for the caudal fin of R4, R5 and R6 coloured based on fish (left) and fin version (right) of origin. All cells are present in all fins and fin version. No batch effects are observed.

**Clonal analysis in the caudal fin.** **a**, Scar percentage per cell in clones detected in fish R4 (left), R5 (middle) and R6 (right). The corresponding organ is indicated above each barplot (WKM or fin version). Spatial information (dorsal or ventral) is indicated when available. The bars at the top indicate clones. Each colour represent a scar, the same colour scheme is used for all panels. **b, c**, Heat maps of the fraction of cells per clones for each cell type and fin in fish R5 (b) and R6 (c). Enriched (magenta upwards triangle) and depleted (blue downwards triangle) scar clones per cell type per primary, secondary and tertiary fin obtained from two-sided Fisher's exact test with $P < 0.05$. Top bars depict clones found in the WKM of the same fish, the corresponding number of cells, and the P value for each clone. **d**, t-SNE map of R6, in which cells with clone 24 (as a representative example of clones shared between mesenchymal and epidermal cells) are highlighted in red. **e**, t-SNE map of all cells detected in the caudal fin, in which cells from clone 19 (as a representative example of clones shared between mesenchymal and epidermal cells) in R4 are highlighted in red. **f**, t-SNE map of R6 primary (left) and regenerated (right) caudal fin cells (grey circles), in which cells from osteoblast clones are highlighted in red. Dashed lines represent mesenchymal cells (Fig. 3.4b, Supplementary Figure 3.6). The percentages indicate the fraction of mesenchymal cells that share clones with osteoblasts. **g**, Magnified view of the t-SNE maps of R4, R5 and R6 for immune cells (dashed line on Fig. 4b). Cells are coloured based on fish (left) and fin version (right) of origin. Subpopulations of lymphoid (dashed circles) and myeloid (solid circles) are found in all fish and fin versions. **h**, Scar percentage for cells detected in the WKM (top) and RICs (bottom) for R4 (left), R5 (middle) and R6 (right) in the primary fins reveals the absence of common scars between the two. The bar above each panel indicates the different clones.




Supplementary Figure 3.7: Figure caption on previous page.




Supplementary Figure 3.8: Figure caption on next page.

**The histone-GFP transgene and scar characterization.** **a**, Scheme of one copy of the h2afva:GFP transgene as previously described18. **b**, Copy number of the transgene. Top, average number of reads in bin sizes of 1 kb and sliding window of 200 bp obtained in whole-genome sequencing data. Bottom, copy number extracted using FREEC-11.0 with default parameters (Methods). **c**, Number of clones detected with a given number of scars, obtained by pooling all data from all fish used in this study. **d**, Scar percentage per cell in a clone in which seven different scars are detected. **e**, Probability density function (normalized histogram) of the fraction of scars detected in the clone depicted in d (colour code as in d). **f**, Average fraction of a scar per embryo computed over ten independently injected embryos (for times larger than 6 h) and over three pools of ten embryos (for times lower or equal to 6 h) detected from gDNA as a function of time for Cas9 RNA injections, for the six most observed scars (described with CIGAR strings). Error bars denote s.e.m. Solid green lines are the fit to equation (5) in Supplementary Information section 2. **g**, Top 100 observed scars, sorted according to their probability, and corresponding position of deletions (red) and insertions (blue) along the GFP coordinate. **h**, Scar probabilities as a function of sorted scars. Error bars denote s.e.m. from the fit (in f). **i**, Probabilities of measuring the percentages $x_i$ for three scars with copy numbers 2 (top), 1 (middle) and 1 (bottom), in which the expected percentages are $f_i$ = 50%, 25% and 25%, respectively, present in the same cell with four surviving integrations of GFP. The probability has been obtained by independently simulating 1,000 times the ScarTrace protocol for $\epsilon$ = 0.50 (left) and $\epsilon$ = 0.85 (right). Solid green lines are the fit to equation (7) in Supplementary Information section 3. **j**, Scar percentage for cells from the same clone made of three scars, in which each scar is represented with a different colour. **k**, Corresponding probability density functions (normalized histogram) for the fraction of each scar per cell (colour code as in j). Black lines denote the best fit for each scar to equation (7) (see Supplementary Information). **l**, Violin plot showing the distribution of measured P values obtained using a Gaussian kernel density with bandwidth determined using the Scott method[112], for all clones with a given estimated NGFP. Labels indicate the number of clones observed for $N_{GFP}$.

# Supplementary Tables

| | | WKM | Left eye | Right eye | Forebrain | Hindbrain | Left midbrain | Right midbran | Fin |
|---|---|---|---|---|---|---|---|---|---|
| RNA | R1 | 14 | - | - | - | - | - | - | - |
| | R2 | 7 | 18 | 14 | 34 | 17 | 15 | 10 | - |
| | R3 | - | 6 | 11 | 20 | 16 | 14 | 22 | - |
| | R4 | 10 | - | - | - | - | - | - | 25 |
| | R5 | 3 | - | - | - | - | - | - | 17 |
| | R6 | 8 | - | - | - | - | - | - | 39 |
| prot. | P1 | 4 | 5 | 5 | 2 | 5 | 6 | 7 | 4 |
| | P2 | 3 | 1 | 3 | 2 | - | 1 | 1 | 2 |

Supplementary Table 3.1: **Clone number in different organs.** Number of clones in different organs for different fish, for the two different Cas9 delivery strategies (RNA or protein). Unscarred clones are excluded and only clones with two or more cells are considered. Whenever clones from the WKM are known for a fish (R1, R2, R4, R5, R6, P1 or P2), they are excluded from the clones detected in the other organs.

| fish | Cas9 | age (months) | organ | cells | | |
|---|---|---|---|---|---|---|
| | | | | Scars | Transcripts | Combined |
| P1 | protein | 18 | WKM | 286 | | |
| | | | forebrain | 209 | | |
| | | | hindbrain | 123 | | |
| | | | left eye | 98 | | |
| | | | right eye | 106 | | |
| | | | left midbrain | 95 | | |
| | | | right midbrain | 152 | | |
| | | | c. fin (1) | 186 | | |
| P2 | protein | 18 | WKM | 544 | | |
| | | | forebrain | 175 | | |
| | | | left eye | 44 | | |
| | | | right eye | 45 | | |
| | | | left midbrain | 167 | | |
| | | | right midbrain | 151 | | |
| | | | c. fin (1) | 339 | | |
| R1 | RNA | 18 | WKM | 1738 | 1784 | 951 |
| R2 | RNA | 3 | WKM | 1673 | 1397 | 783 |
| | | | forebrain | 545 | 563 | 409 |
| | | | hindbrain | 133 | 179 | 64 |
| | | | left eye | 303 | 403 | 208 |
| | | | right eye | 95 | 250 | 78 |
| | | | left midbrain | 429 | 302 | 190 |
| | | | right midbrain | 238 | 135 | 78 |
| R3 | RNA | 3 | forebrain | 408 | 363 | 206 |
| | | | hindbrain | 152 | | |
| | | | left eye | 82 | | |
| | | | right eye | 204 | 336 | 171 |
| | | | left midbrain | 463 | 126 | 77 |
| | | | right midbrain | 458 | 287 | 186 |
| R4 | RNA | 18 | WKM | 625 | | |
| | | | c. fin (1) | 979 | 926 | 479 |
| | | | c. fin (2) | 893 | 1143 | 538 |
| | | | dorsal c. fin (3) | 504 | 694 | 329 |
| | | | ventral c. fin (3) | 687 | 582 | 292 |
| R5 | RNA | 18 | WKM | 66 | | |
| | | | c. fin (1) | 1090 | 757 | 393 |
| | | | c. fin (2) | 1235 | 1022 | 673 |
| | | | dorsal c. fin (3) | 103 | 348 | 63 |
| | | | ventral c. fin (3) | 96 | 169 | 24 |
| R6 | RNA | 18 | WKM | 432 | | |
| | | | dorsal c.fin (1) | 318 | 576 | 135 |
| | | | ventral c.fin (1) | 205 | 538 | 66 |
| | | | dorsal c.fin (2) | 378 | 697 | 195 |
| | | | ventral c.fin (2) | 391 | 714 | 199 |

Supplementary Table 3.2: **Overview of scarred fish and organs dissected.** Number of cells sequenced per organ that survive filtering thresholds for transcriptome libraries, scars libraries and combined per scarred fish. The total number of sorted plates is indicated in parenthesis. In the caudal fin (c. fin), primary (1), secondary (2) or tertiary (3) are also indicated in parenthesis.

5'- ATG GTG AGC AAG GGC GAG GAG CTG TTC ACC GGG GTG GTG CCC
ATC CTG GTC GAG CTG GAC GGC GAC GTA AAC GGC CAC AAG TTC AGC
GTG TCC GGC GAG GGC GAG GGC GAT GCC ACC TAC GGC AAG CTG ACC
CTG AAG TTC ATC TGC ACC ACC GGC AAG CTG CCC GTG CCC TGG CCC
ACC CTC GTG ACC ACC CTG ACC TAC GGC GTG CAG TGC TTC AGC CGC
TAC CCC GAC CAC ATG AAG CAG CAC GAC TT→C TTC AAG TCC GCC ATG
CCC GAA GGC TAC GTC CAG GAG CGC ACC ATC TTC TTC AAG GAC GAC
GGC AAC TAC AAG→ ACC CGC GCC GAG GTG AAG TTC GAG GGC GAC ACC
CTG GTG AAC CGC ATC GAG CTG AAG GGC ATC GAC TTC AAG GAG GAC
GGC AAC ATC CTG GGG CAC AAG CTG GAG TAC AAC TAC AAC AGC CAC
AAC GTC TAT ATC ATG GCC GAC AAG CAG AAG AAC GGC ATC AAG GTG
AAC TTC AAG ATC CGC CAC AAC ATC GAG GAC GGC AGC GTG CAG CTC
G CC G AC CAC TAC CAG CAG AAC ACC CCC ATC GGC GAC ←GGC CCC
GTG CTG CTG CCC GAC AA←C CAC TAC CTG AGC ACC CAG TCC GCC CTG
AGC AAA GAC CCC AAC GAG AAG CGC GAT CAC ATG GTC CTG CTG GAG
TTC GTG ACC GCC GCC GGG ATC ACT CTC GGC ATG GAC GAG CTG TAC
AAG TAA -3'

Supplementary Table 3.3: Wild type GFP reference sequence. The PAM and the sgRNA in used are indicated in red and cyan respectively. Additionally, primers for the nested PCR are shown.

720M
530M21D169M
534M11D175M
531M18D171M
536M9D175M
541M6D173M
538M10D172M
520M24D176M
539M6D175M
543M2D175M
537M17D166M
538M11D171M
507M2D4M1D8M2D2M4D8M6D176M
538M13D169M
539M9D172M
529M1I11M1D179M
538M22D160M
535M10D175M
541M3D176M
545M1I175M
513M8D19M1D179M
543M5D172M
530M24D166M
540M1D179M
501M6D9M15D10M5D174M
534M8D178M
540M2I180M

Supplementary Table 3.4: **Legend for Fig 3.1d.** Each scar has a color. The code for each scar is given a CIGAR code that indicates the size and position of indels and matches along the 720-nucleotide long GFP sequence. Accordingly, WT is 720M; and 530M21D169M indicates that the first 530 nucleotides match the reference GFP sequence, are then followed by a 21-long deletion, and the last 169 nucleotides match again the reference GFP.

720M
531M18D171M
533M11D176M
529M17D174M
543M2D175M
541M3D176M
543M4D173M
534M11D175M
536M9D175M
541M6D173M
542M2D176M
544M1D175M
536M20D164M
538M2D180M
523M28D169M
510M11D5M6D13M1I175M
541M5D174M
541M3D9M6D161M
544M1I176M
537M14D169M
542M8D170M
539M6D175M
530M21D169M
531M2D12M2I175M
520M2D14M1D183M
534M5D181M

Supplementary Table 3.5: **Legend for Fig. 3.2a.** Each scar, named using a CIGAR code, has a different color.

| | | WKM | Left eye | Right eye | Forebrain | Hindbrain | Left midbrain | Right midbran | Fin |
|---|---|---|---|---|---|---|---|---|---|
| mRNA | R1 | 7 | 18 | 14 | 34 | 17 | 15 | 10 | - |
| | R2 | 14 | - | - | - | - | - | - | - |
| | R3 | - | 5 | 11 | 20 | 16 | 12 | 19 | - |
| | R4 | 10 | - | - | - | - | - | - | 25 |
| | R5 | 3 | - | - | - | - | - | - | 17 |
| | R6 | - | - | - | - | - | - | - | - |
| prot. | P1 | 4 | 5 | 5 | 2 | 5 | 6 | 7 | 4 |
| | P2 | | | | | | | | |

Supplementary Table 3.6: Number of clones in different organs from different fish, for the two different Cas9 delivery strategies (mRNA or protein). Only clones with two cells or more are considered. Unscarred clones have been excluded. Whenever clones from the WKM were known for a fish (R1, R2, R4, R5, R6, P1, P2), they were excluded from the clones detected in the other organs.

| | Parameter | WKM | Brain & eyes | Caudal fin |
|---|---|---|---|---|
| filterdata | mintotal | 500 | 100 | 500 |
| | minexpr | 5 | 1 | 3 |
| | minnumber | 2 | 2 | 2 |
| | maxexpr | Inf | Inf | Inf |
| | downsample | True | False | True |
| | rseed | 1700 | 16000 | 16000 |
| | snf | – | – | False |
| clustexp | clusternr | 20 | 20 | 20 |
| | bootnr | 50 | 50 | 50 |
| | metric | Pearson | Pearson | Pearson |
| | sat | True | True | True |
| | rseed | 17000 | 17000 | 17000 |
| | FUNcluster | kmedoids | kmedoids | kmedoids |
| | Fselect | True | – | True |
| comptsne | rseed | 1 | 15550 | 1 |
| | sammonmap | False | False | False |
| | initial_cmd | True | True | True |
| | fast | True | True | True |
| | perplexity | 30 | 30 | 30 |
| findoutliers | outmic | 5 | 2 | 8 |
| | outlg | 5 | 2 | 8 |
| | probthr | $10^{-8}$ | $10^{-9}$ | $10^{-8}$ |
| | thr | $2^{-1:40}$ | $2^{-1:40}$ | $2^{-1:40}$ |
| | outdistquant | 0.95 | 0.95 | 0.95 |

Supplementary Table 3.7: RaceID parameters used to filter, cluster and find cell types in the WKM, brain and eyes, and caudal fin transcriptome data when pooling together all cells sequenced from every fish whose organ has been isolated. When a parameter is not specified, the default value was used.

| | | |
|---|---|---|
| Lymphocytes | known<br>unkwnon | bhlhe41 CD37 cxcr4 sla2 zfp36l1a cd74a cd74b<br>ENSDARG00000074014 |
| Monocytes | known<br>unknown | mmp9 mmp13 socs3a socs3b tnfb adam8 mpx<br>ENSDARG00000090730 ENSDARG00000068784 |
| Neutrophils | known | lyz nspn lect2l cpa5 |
| Eosinophils | known<br>unknown | hp itln1 gata2<br>ENSDARG00000094434 ENSDARG00000091736<br>ENSDARG00000079043 ENSDARG00000071806<br>ENSDARG00000093877 ENSDARG00000094408<br>ENSDARG00000076534 ENSDARG00000091236 |
| HSPCs | known<br>unknown | rpl\|rps _eef runx1 myb<br>ENSDARG00000030585 ENSDARG00000013012<br>ENSDARG00000023298 |
| Erythrocytes | known<br>unknown | ba1 alas2 cahz tspo<br>ENSDARG00000069734 ENSDARG00000079078<br>ENSDARG00000077504<br>ENSDARG00000069735 |
| Thrombocytes | known<br>unknown | ranbp10 GLDC<br>ENSDARG00000071657 ENSDARG00000070956<br>ENSDARG00000071052 |
| Dendritic cells | known<br>unknown | stmn1a hmga1a<br>ENSDARG00000093817 ENSDARG00000035423 |

Supplementary Table 3.8: **Gene markers for cell type calling in WKM.** For cell type calling in the WKM we used[7,23,62,76,87]. In addition, we used our dataset to study further the differential gene expression in the hematopoietic cell and therefore also found unknown genes specifically expressed in each cell type. Additionally, we clustered cd4low sorted HSPCs to validate HSPC cell type calling.

| | |
|---|---|
| MFOLs | olig2 plp1b serine |
| COPCs | fabp7a gpr17 |
| Astrocytes | fabp7a gfap slc1a4 |
| Glia-like | slc1a4 acsbg1 gldc c1qa |
| Microglia | apoeb cd74a cd74b |
| Bipolar and Horizontal cells | cabp5a cabp2a gnao1b sik2a prkca rs1a slc17a7a |
| Cone cells | opn1lw1 opn1sw1 |
| Rod cells | rho prom1a prom1b gnat1 gnat2 pde6a pde6b |
| Neurons | syt1a snap25a snap25b tuba1a slc17a6a slc17a6b slc32a1 gabra1 |

Supplementary Table 3.9: **Gene markers for cell type calling in the head.** WKM cells were called based on markers from Table 3.8.

| | |
|---|---|
| Neuronal cells | sox2 rgs5 |
| Epidermis | krt91 krt5 dspa jupa perp krt97 cfl1l cldna |
| Osteoblasts | krt8 sparc krt18 col6a col12a col5a |
| Fibroblasts | col9a2 |

Supplementary Table 3.10: **Gene markers for cell type calling in the fin.** WKM cells were called based on markers from Table 3.8.

Chapter 4

# Deconvolving multiplexed histone modifications in single cells

J. Yeung[1,*], M. Florescu[1,*], P. Zeller[1], B.A. de Barbanson[1] and A. van Oudenaarden[1]

*in preparation*

*equal contribution

[1] Oncode Institute, Hubrecht Institute-KNAW (Royal Netherlands Academy of Arts and Sciences) and University Medical Center Utrecht, 3584 CT, Utrecht, The Netherlands.

The regulation of chromatin relies on the interplay of different histone modifications to modulate the packaging of the genome. Quantifying this interplay is ideally performed at the single-cell level, because of the epigenetic heterogeneity between individual cells. While recent advances have enabled mapping of histone modifications in single cells, current methods are constrained to profile only one histone modification per cell. Here we present an integrated experimental and computational framework, scChIX (single-cell immunocleavage and unmixing), to multiplex histone modifications in single cells. We incubate cells with two antibodies so that adding pA-MNase enzyme makes targeted cuts in the genome associated with either of two histone modifications (double-incubated). Our statistical framework then uses single-incubated histone modification data as training to deconvolve the double-incubated data into their respective origins. We apply scChIX to mouse bone marrow to demonstrate that the method accurately maps two histone modifications in single cells from a heterogeneous sample. Although polycomb-repressed (marked by H3K27me3) and heterochromatin (marked by H3K9me3) regions are generally mutually exclusive, we find that the transitions between the two regions can vary between celltypes. scChIX also links together single-cleaved data, allowing joint analysis of functionally distinct histone modifications and enable information from one histone modification to transfer to another. We transfer celltype labels derived from scChIC-seq of active modifications (H3K4me1) to systematically annotate celltypes in repressive histone modifications (H3K27me3). Linked analysis reveals heterogeneity in chromatin regulation within celltypes. Genomic loci that code for immunoglobulin genes are in a repressed chromatin state in naive B cells, but turn active in activated B cells. H3K4me1 histone modification levels along neutrophil maturation pseudotime increases at mature neutrophil marker genes. Of note, we find a transcriptionally silent module consisting of *Hox* and other developmental genes where H3K27me3 levels progressively increase along pseudotime, highlighting that regulation of transcriptionally silent genes can be dynamic. scChIX expands the epigenetic landscape that can be measured in single cells. We anticipate that scChIX will be widely applicable to reveal single-cell regulation of activating and silencing gene expression programs.

## Introduction

Histone modifications form an adaptable epigenetic regulatory layer that controls dynamics and homeostasis of transcriptional programs[75]. Understanding the regulation of chromatin requires not only studying how a constellation of histone modifications is distributed across the genome in different cellular contexts but also how different histone modifications interact to regulate gene expression (Fig. 4.1a). Large-scale efforts have catalogued different histone modifications in a variety of cell populations[65,68]. These studies profile multiple histone modifica-

tions across tissues by pooling millions of cells that are then split into different technical batches and then performing chromatin immunoprecipitation and sequencing (ChIP-seq) in each batch. ChIP-seq resources are valuable, but the large starting material required limits the resolution to bulk tissue [146], cell lines [65], or purified celltypes from known cell surface markers [69]. Alternative strategies to ChIP-seq based on enzyme tethering (chromatin immunocleavage, ChIC) have reduced the background signal in profiling the epigenome [109]. Advances in ChIC enabled chromatin profiling in single-cells (e.g., scChIC-seq and variants ChIL-seq, CoBATCH, CUT&RUN, CUT&TAG [56,64,117,139]). In scChIC-seq, antibodies targeting a specific histone modification tethers proteinA-MNase (pA-MNase), a modified DNA cutter. Adding calcium activates pA-MNase which cuts at targeted DNA regions of the genome (Fig. 4.1b). Genomic fragments are then ligated to barcoded adapters containing unique molecular identifiers (UMI) and cell-specific barcodes, and are then sequenced. scChIC-seq holds promise in uncovering the epigenome at single-cell resolution, but is currently limited to only one histone modification per cell.

We present an integrated experimental and computational framework for multimodal analysis of two histone modifications in single cells. The strategy combines double immunocleavage followed by computational unmixing. We perform double immunocleavage by incubating cells with two antibodies each targeting a distinct histone modification. Activating pA-MNase on double-incubated cells generates targeted cuts at DNA regions associated with either of the two histone modifications. We developed a computational framework to probabilistically assign each cut to their respective histone modification by using training data from standard single-incubated scChIC-seq targeting each histone modification separately.

To profile two histone modifications in single cells (Fig. 4.1a), our experimental method first generates three genome-wide scChIC-seq datasets: two containing single-cell profiles of different histone modifications separately (single-incubated Fig. 4.1b), and the third containing single-cell profiles of both histone modifications (double-incubated, Fig. 4.1b). We use our two single-incubated datasets as training data to generate the possible pairs of genomic profiles that, when added together at a specific ratio, fit to a single-cell in the double-incubated dataset (Fig. 4.1c). After determining for each cell in the double-incubated data the most likely pair of genomic profiles and the mixing ratio, each cut from the double immunocleavage can then be probabilistically assigned to their respective histone modification. The unmixed signal enables multimodal analysis of histone modifications in single cells.

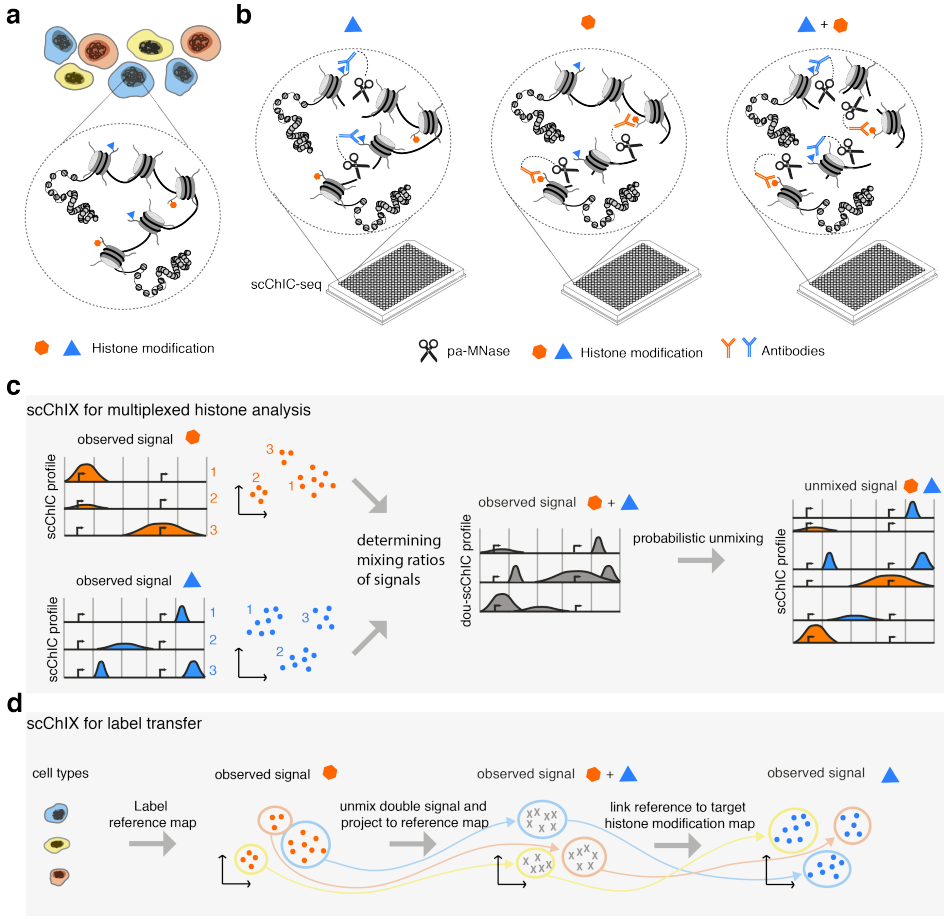



Figure 4.1: **Schematics of scChIX method.** **a** Cellular states in a heterogenous cellular context (different colored cells) are regulated through multiple histone modifications (two histone modifications shown as an example). **b** Input for scChIX are three scChIC-seq samples: Two samples with each a single histone modification targeted and one sample with the combined histone modification targeted. **c** Schematics of scChIX umixing of multiplexed histone datasets. Two single antibody targeted scChIC-seq can be assigned to cluster defining a particular cell state. We use the single-stained data to model how the two histone modifications can be combined across the possible combinations of cell states to generate a double-stained cell. Given a double-stained single cell, we probabilistically assign the cell to a pair of cell states (one state for each histone modification). Finally, for each double-stained single cell, we unmix the signal into each histone modification separately. **d** Label transfer links together two single-targeted scChIC-seq datasets from different histone modifications. We use an active mark to identify celltype label for observed cluster. We use the double-targeted dataset to determine the link between the single-incubated datasets. Thus labels, here celltype information, from one single-incubated dataset are transferred to another histone modification dataset.

Our method also enables linked analysis of the two scChIC-seq datasets by transferring cell labels from one dataset to the other (Fig. 4.1d). This strategy is useful when one of the histone modifications have clear associations with gene expression (e.g. activating histone modifications) and therefore celltype information (Fig. 4.1d) while the other modification may not be correlated with gene expression (Fig. 4.1d). In conventional genome-wide studies of repressive histone modifications, the celltype or tissue is known *a priori*. By contrast, single-cell studies annotate celltypes after the data is generated, often by using known celltype-specific marker genes. But these marker genes may not necessarily be informative in determining celltype for certain histone modifications. For example, although studies have established the repressive role of H3K27me3 on transcription [22], genome-wide of H3K27me3 ChIP-seq data is overall less predictive of gene expression than activating modifications such as H3K4me1, H3K4me3, and H3K36me3 [28]. Therefore, rather than annotating celltypes from H3K27me3 scChIC-seq datasets directly, we can first use a double-incubated data targeting both an active and repressive mark to link the active with the repressive single-incubated scChIC-seq datasets (Fig. 4.1d). Once linked, we can then transfer labels from the active to the repressive (Methods). Our method thus provides a systematic framework for linking together scChIC-seq datasets targeting distinct histone modifications.

Here we demonstrate that our method can infer two histone modifications in single cells as well as transfer labels from one scChIC-seq dataset to another.

To validate the accuracy of unmixing, we generate a ground truth scChIC-seq dataset of H3K27me3 (polycomb repressive mark), H3K9me3 (heterochromatin mark), and H3K27me3+H3K9me3. We purify known celltypes of mouse bone marrow using fluorescence-activated cell sorting (FACS, Methods). We use this ground truth dataset to show that our framework chooses the correct two underlying genomic profiles (one for H3K27me3, and one for H3K9me3) that explain each H3K27me3+H3K9me3 single-cell profile. We unmix the H3K27me3+H3K9me3 signal to their respective histone modifications and validate that the genomic profiles agree with the ground truth data.

Next, to validate label transfer and demonstrate a typical use case, we generate scChIC-seq data from whole bone marrow targeting H3K4me1 (active and primed promoters and enhancers), H3K27me3, and H3K4me1+H3K27me3. We label the celltypes for the H3K4me1 dataset and then transfer these labels to the H3K27me3 dataset. Overall, our method deconvolves multiplexed histone modifications in single cells.

## Results

### scChIX deconvolves double immunocleavage signal by unmixing

To validate our method, we generate a ground truth dataset by purifying three known celltypes from bone marrow using FACS (Methods). We incubate mouse bone marrow cells with two repressive histone modifications with mutually exclusive genomic profiles, H3K27me3 (polycomb repressed, gene rich regions) and H3K9me3 (heterochromatin, gene poor regions) as well as a combined genomic profile of H3K27me3+H3K9me3 (double-incubated). We stain each sample with antibodies against cell surface markers to purify granulocytes, B cells, and natural killer (NK) cells (Methods). We sort a total of 720 cells from each of the three celltypes into 384-well plates, each plate containing all three celltypes (Supplementary Fig. 4.1, Methods). Since our input is a complex sample containing multiple celltypes, we develop a statistical method that infers the pair-wise combination of celltype-specific H3K27me3 and H3K9me3 profiles that, when mixed together, fits the single-cell H3K27me3+H3K9me3 data. We then use the inferred model to probabilistically assign each double-incubated pA-MNase cuts to their respective modification (either H3K27me3 or H3K9me3) in single cells. We reduce the dimensions of the scChIC-seq data while taking into account the sparse count matrix using latent dirichlet allocation (LDA)[11]. Briefly, the LDA is a probabilistic matrix factorization method that uses multinomials to model the sparse single-cell cuts across genomic regions (we use nonoverlapping genomic bins of 50kb, Methods). LDA infers cell-specific relative frequencies of generating a cut across genomic regions. Correlations between cells and across regions (e.g. cells of the same celltype) are captured by assuming the data come from a lower-dimensional space (in LDA terminology, latent dimensions are called "topics", Methods). We apply LDA independently on each of the three scChIC-seq datasets: H3K27me3, H3K9me3, and H3K27me3+H3K9me3 (Supplementary Fig. 4.2a). The scChIC-seq datasets show separation between the FACS-identified celltypes, confirming celltype-specific distributions of repressive histone modifications. Downstream analysis of LDA finds clusters defined by the topic weights (Supplementary Fig.s 4.2b,c). We use these topics to cluster cells in the H3K27me3 and H3K9me3 scChIC-seq (Supplementary Fig. 4.2a). The training data takes possible pairs of clusters defined by LDA (one from H3K27me3, one from H3K9me3) and mixes their region-specific relative frequencies together at a specific ratio, which defines the multinomial likelihood of generating the single-cell double-cut data given the mixed relative frequencies. For each H3K27me3+H3K9me3 cell, the statistical model then selects the cluster pair that best explains the single-cell H3K27me3+H3K9me3 data.

After selecting the most likely H3K27me3-H3K9me3 cluster pair for each double-incubated cell, we probabistically assign each pA-MNase double-incubated cut

to either H3K27me3 or H3K9me3 (Methods). Each unmixed double-incubated cell thus generates a H3K27me3 and a H3K9me3 epigenomic profile. The unmixed H3K27me3 and H3K9me3 profiles are projected onto the H3K27me3 and H3K9me3 LDA embeddings trained on single-incubated cells (visualized as UMAPs in Fig. 4.2a). Therefore, each double-incubated cell has a location in each of the three LDA embeddings, allowing the three UMAPs to link together (lines connecting cells in Fig. 4.2a).

The double- and single-incubated cells in the H3K27me3 and H3K9me3 UMAPs intermingle, suggesting that the model accurately assigns pA-MNase cuts to their respective histone modification (Supplementary Fig. 4.3a, b). Comparing the H3K27me3 unmixed pseudobulk signal with the granulocytes, B cells, and NK cells pseudobulk show high correlation for the expected celltype, and lower for the other two celltypes (Supplementary Fig. 4.3c). The H3K9me3 unmixed pseudobulk signal also showed highest correlation with the expected celltype, although the other two celltypes also show high correlation (Supplementary Fig. 4.3d), suggesting that genome-wide H3K9me3 profiles contain both celltype-specific regions as well as constitutive heterochromatin domains[9]. Overall, we find that scChIX is accurate in assigning pA-MNase cuts to their respective histone modification.

We quantify the accuracy of scChIX to select the correct H3K27me3-H3K9me3 cluster pair to mix together. Plotting the celltype (defined by FACS) selected from H3K27me3 and H3K9me3 for each double-incubated cells shows that scChIX chooses the same celltype from both H3K27me3 and H3K9me3 (Fig. 4.2b, left, double-incubated cells fall mostly on the diagonal). The false discovery rates of sChIX predicting B cells, granulocytes, or NK cells are 10%, 3%, and 1%, respectively (Fig. 4.2b, middle). Similarly, scChIX has high specificity and sensitivity (Fig. 4.2b, right).

scChIX assigns each genomic cut to either H3K27me3 or H3K9me3 (Fig. 4.2c, Methods, Supplementary Material). Genome browser plots of unmixed H3K27me3+H3K9me3 together with the ground truth single signals of H3K27me3 or H3K9me3 across the genome confirms that our probabilistic assignment of reads recapitulates the mutual exclusive relationship between H3K27me3 and H3K9me3, exemplified in the *Bcl2* locus (Fig. 4.2d) and other loci (Supplementary Fig. 4.4) in B cells. Furthermore, we find celltype-specific signal that distinguishes H3K27me3 across tissues (Fig. 4.5). We also find celltype-specific H3K9me3 regions (Fig. 4.6b), although the differences are attenuated compared to celltype-specific H3K27me3 differences.

To characterize the transitions between H3K27me3 and H3K9me3 at single-cell resolution, we summarize the cell-specific assignment probabilities across the genome in a heatmap (Fig. 4.2e). At the *Bcl2* locus, we find the transitions to be celltype-independent. These transitions, however, can be celltype-specific, as

exemplified by the *Crim1* locus. Taken together, although the organization of the genome into heterochromatic and euchromatic alternating patterns have generally been known to be mutually exlusive[25], our scChIX method reveals that these spatial patterns of heterochromatic-euchromatic transitions can also be celltype-specific.

**Fig. 4.2: scChIX accurately unmixes multiplexed histone modifications in single cells.** **a** Low-dimensional representation (UMAP) of scChIC-seq data in sorted cell types, labeled by ground truth labels from FACS sorting for double-incubated (H3K27me3+H3K9me3, left) as well as H3K27me3 and H3K9me3 scChIC-seq data (right). H3K27me3 and H3K9me3 contain both single-incubated cells as well as double-incubated cells that have been unmixed to their respective histone modifications. Lines connecting across datasets come from double-incubated cells, which contain both H3K27me3 and H3K9me3 signal. Their unmixed H3K27me3- and H3K9me3-portion are projected onto the top-right and bottom-right datasets, respectively. **b** Matrix summarizing the cluster pair that scChIX selected for each double-incubated cell. Cells along the diagonal are predicted to be B cells, Granulocytes, and NK cells, respectively. Cells in the off-diagonal are cases where scChIX failed to assign double-incubated cells to a coherent H3K27me3-H3K9me3 pair (false negatives). Barplots summarizing false discovery rate (FDR), sensitivity, and specificity of assigning each celltype (right). **c** Coverage plot and single cell cuts in B cells of mixed (K27me3+K9me3, grey bars), unmixed (K27me3 and K9me3, orange and blue bars). Ground truth coverage are single-incubated scChIC-seq data targeting K27me3 (orange) and K9me3 (blue). Unique reads are shown for four single cells (A, B, C, and D) for H3K27me3+H3K9me3 signal (grey ticks) as well as their unmixed outputs (orange and blue ticks). **d** Zoom-in of the *Serpinb5* locus. Unique reads from K27me3+K9me3 are colored based on whether they have been assigned to K27me3 (orange) or K9me3 (blue). Circled reads and arrow highlight example of reads being assigned to either K9me3 or K27me3. **e** Heatmap of probabilities $p$ of assigning reads to K27me3 ($p = 1$, red) or K9me3 ($p = 0$, blue) of the zoomed out of the *Serpinb5* locus. Rows are single cells (ordered by predicted celltype), columns are genomic regions (50 kb). **f** Same as (e) but at the *Crim1* locus, where transitions from K9me3 to K27me3 (blue to red) appear celltype-specific.

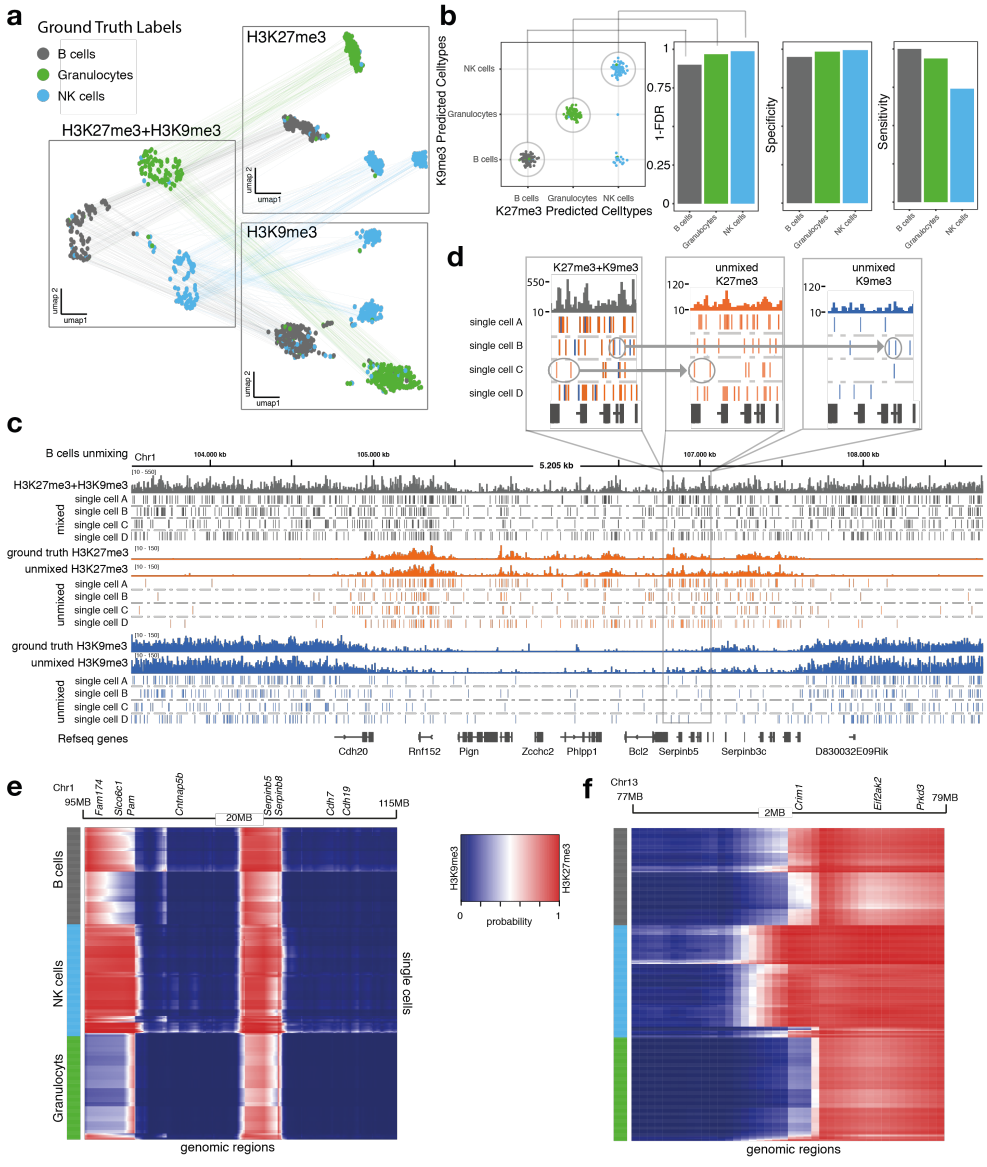



Figure 4.2: Figure caption on previous page.

## scChIX transfers celltype labels from active to repressive histone modifications

After validating our method on a ground truth dataset, we apply our method to an activating and a repressive histone modification in a heterogeneous sample. We target H3K4me1, H3K27me3, and H3K4me1+H3K27me3 in cells from mouse bone marrow. Applying LDA to the three datasets reveals cells clustered by topics, each topic being a distinct genomic distribution of histone modifications (Fig. 4.3a-c, Supplementary Fig. 4.7, 4.8). We compare H3K4me1 scChIC-seq data with publicly available scRNA-seq data of bone marrow[37]. We assign each genomic region to their nearest gene within 50 kb. Comparing mRNA expression levels across celltypes of genes near regions defined by high H3K4me1 topic weights assigns a celltype to each topic (Supplementary Fig. 4.9).

We model the H3K4me1+H3K27me3 scChIC-seq signal as a mixture of cluster pairs, one from a H3K4me1 cluster and the other from a H3K27md3 cluster (Supplementary Fig. 4.10a). Plotting the assignment of cluster pairs onto the H3K4me1+H3K27me3 UMAP confirms that the single-cell assignment is precise (Supplementary Fig. 4.10b, c). Neighboring cells in H3K4me1+H3K27me3 are mostly assigned to the same cluster pairs, with the exception of topic 26 in H3K27me3 (Supplementary Fig. 4.10c), which are assigned to the double-incubated cells near the center of the UMAP. We interpret this ambiguity in the center as cells that may have lower confidence in assigning to the correct cluster pair. We did not associate a basophils to a specific H3K27me3 cluster, possibly because basophils are rare[70] and were not captured in the H3K27me3 data (Supplementary Fig. 4.10a)).

Using the inferred cluster pairs, we then unmix the H3K4me1+H3K27me3 single-cell signal based on the pair assignments and project the respective H3K4me1 and H3K27me3 unmixed data onto their respective LDA embeddings. We find that the unmixed signal intermingle with their respective single-incubated histone modification, suggesting that our method deconvolves double-incubated scChIC-seq data accurately (Supplementary Fig. 4.11a, b).

With the H3K4me1 and H3K27me3 UMAPs both now integrated with the double-incubated cells in their respective LDA embeddings, we can now link the H3K4me1 and H3K27me3 embeddings together. We first transfer celltype labels from H3K4me1 single-incubated cells onto the double-incubated cells through the H3K4me1 embedding (Supplementary Fig. 4.12, Methods). Analogously, the labels from the double-incubated cells are then transferred to the single-incubated H3K27me3 cells through the H3K27me3 embedding (Supplementary Fig. 4.13). Linking the deconvolved H3K4me1+H3K27me3 signal with their respective single-incubated signal reveals the joint embedding of the H3K4me1 and H3K27me3 histone modification space in bone marrow, with celltype labels shared across embed-

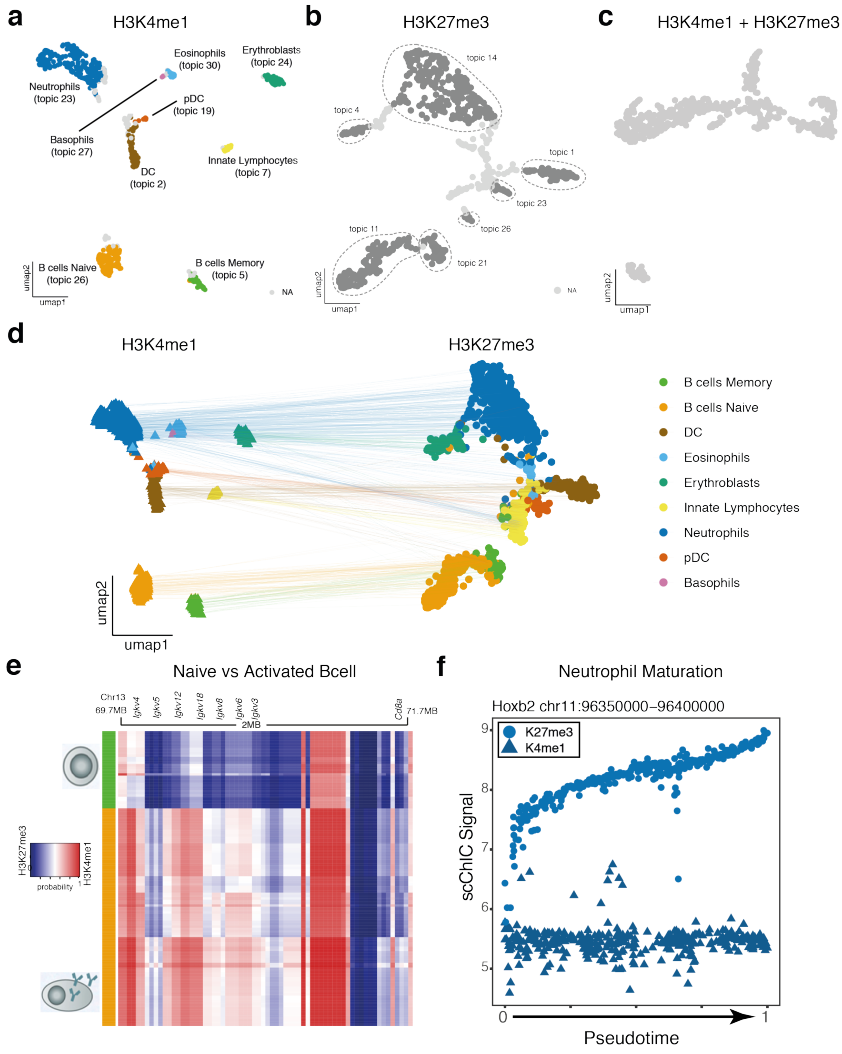



Figure 4.3: **scChIX for transferring labels from active to repressive histone modifications. a** UMAP of scChIC-seq signal of H3K4me1 in whole bone marrow (cell clusters are colored by known cell types from Giladi et al. 2018). **b** UMAP of scChIC-seq signal of H3K27me3 in whole bone marrow (cell clusters associated to relevant LDA topics are delineated by dashed lines, see Methods). **d** UMAP representation of double-targeted scChIC-seq signal of H3K4me1+H3K27me3. **d** Unmixing of H3K4me1+H3K27me3 double-targeted signal in single cells infers the matching between the cell clusters in H3K4me1 and H3K27me3 (the topic in H3K27me3 with the most cells in the corresponding H3K4me1 cluster is colored with the cell type identified from H3K4me1 signal). **e** Probability heatmaps of the *Igk* locus for naive versus activated B cells. Activated B cells show an active chromatin state while naive B cells show a repressive chromatin state. **f** K27me3 and K4me1 levels at the *Hoxb2* locus along pseudotime. Cells are ordered along increasing maturation of neutrophils.

dings (Fig. 4.3d).

To validate the predicted celltypes in both the single- and double-incubated scChIC-seq datasets, we compare with data from celltypes purified by FACS. We compare the H3K4me1 scChIC-seq data with publicly available ChIP-seq[69]. Pearson correlation between ChIP-seq of B cells, erythroids, granulocytes, and NK cells versus scChIC-seq from single- and double-incubated cells is highest for the predicted celltype (Supplementary Fig. 4.14. Although single-incubated cells had higher correlation with ChIP-seq reference data than double-incubated for the matched celltype, the double-incubated cells of the matched celltype consistently had higher correlation with ChIP-seq than single-incubated cells of unmatched celltype. We use our ground truth H3K27me3 ChIC-seq data purified from FACS. Pearson correlation between sorted scChIC-seq of B cells, granulocytes, and NK cells versus unsorted bone marrow scChIC-seq is highest for the predicted celltype (Supplementary Fig. 4.15).

We use the integrated dataset to investigate histone modification dynamics within bone marrow celltypes. Focusing on naive versus activated B cells (Supplementary Fig. 4.17a, b), we filter for B cells and re-cluster by running LDA on H3K4me1 and H3K27me3 separately. UMAPs from this re-clustering confirms the identity of the B cells. Overlaying H3K4me1 active histone modification signal at subtype-specific loci such as *Pax5* and *H2-Aa* for naive B cell-specific and *Irf4*, *Cd93*, and *Igkv3-2* for activated B cells (Supplementary Fig. 4.17c, d). We hypothesize that immunoglobulin production may differ between B cell subtypes[149], so we focus on the chromatin regulation of the *IgK* region in single cells. Plotting the heatmap of H3K4me1-H3K27me3 assignment probabilities reveals that the chromatin state in the *IgK* locus is active for activated B cells but repressed for naive B cells (Fig.4.3e), consistent with activated B cells expressing and secreting immunoglobulins while naive do not[14]. Next, we re-cluster neutrophils revealing a neutrophil maturation trajectory (Supplementary Fig. 4.18a, b). H3K4me1 levels at the *Retnlg* locus, a marker gene for mature neutrophils[39] increases along pseudotime, while H3K27me3 levels slightly decreases (Supplementary Fig. 4.18c). Our LDA analysis of H3K27me3 dynamics reveals a gene module consisting of *Hox* and other developmental genes at which the repressive mark increases along pseudotime (Supplementary Fig. 4.18d, e). This *Hox* module for H3K27me3 corroborates with previous results mapping restrictive chromatin in neutrophils[148].

In sum, we demonstrate our method in a heterogeneous sample, revealing diverse blood celltypes in both H3K4me1 and H3K27me3. By transferring celltype labels from H3K4me1 to H3K27me3, we enable a joint analysis of active and repressive chromatin dynamics at the single-cell level. We show that our label transfer is general, allowing active H3K4me1 levels to be overlaid onto H3K27me3 embeddings as well as aligning the H3K4me1 and H3K27me3 embeddings into a common pseu-

dotime. Our integrated analysis reveals a repressed chromatin signature in immunoglobulin genes in naive B cells, but become active in activated B cells. In neutrophil maturation, we show that transcriptionally silent regions can have dynamics in repressive histone modification levels.

## Discussion

The interplay between a variety of histone modifications regulate chromatin states, but profiling multiple histone modifications genome-wide in single cells is relatively unexplored. Genome-wide analysis of multiple histone modifications has so far been limited to bulk tissue, cell lines, or purified celltypes from known cell surface markers; while current scChIC-seq protocols have so far been limited to one histone modification per cell. Here we introduce an experimental and computational framework that allows multimodal analysis of histone modifications in single cells. scChIX opens up new analyses and expands the dimensions of chromatin regulation that can be profiled in single cells. Our framework provides systematic tools to study cell-to-cell heterogeneity in chromatin state dynamics.

Currently, data integration in single-cell analysis often relies on the assumption that the measurements have a shared correlation structure, such as mRNA abundance and open chromatin[123]. scChIX does not assume a relationship between the two datasets, but rather infers through the joint experimental design and statistical framework. This framework allows systematic integration of histone modifications with complex relationships, which can augment integrated datasets by adding distinct and orthogonal information. Profiling both an active and repressive histone modification in single cells can reveal regulation of transcriptional activation as well as silencing during cellular decision making. We anticipate that scChIX will be widely applicable as single-cell histone modification datasets become increasingly available and efforts expand to multimodal measurements of epigenetic features in single cells.

## Acknowledgements

We thank Marijn van Loenhout for experimental advice and Reinier van den Linden for FAC-sorting. This work was supported by a European Research Council Advanced grant (ERC-AdG 742225-IntScOmics), Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO) TOP award (NWO-CW 714.016.001), the Swiss National Science Foundation, and the Human Frontiers for Science Program. This work is part of the Oncode Institute which is partly financed by the Dutch Cancer Society.

## Author contributions

JY, MF, BAdB, and AvO conceived the project. MF developed technique and performed experiments, with help from PZ. JY developed and applied statistical methods. BAdB wrote the scChIC-seq demultiplexing and preprocessing pipeline, with help from MF and JY. JY, MF, and AvO analyzed the data. JY, MF, AvO wrote the manuscript, with input from PZ and BAdB.

## Methods

### Animal experiments

Male 13-week-old C57BL/6 mice were used to extract bone marrow cells. Experimental procedures were approved by the Dier Experimenten Commissie of the Royal Netherlands Academy of Arts and Sciences and performed according to the guidelines. Femur and Tibia were extracted, the bones ends were cut away to access the bone marrow which was flushed out using a 22G syringe with HBSS (-ca, -mag, -phenol red, Gibco 14175053) supplemented with Pen-Strep and 1% FCS. The bone marrow was dissociated and debris was removed by passing it through a 70$\mu$m cell strainer (Corning, 431751). Cells were washed with 25ml supplemented HBSS before depleting the sample of unnucleated cells using IOTest 3 Lysing solution (Beckman Coulter) following provider instructions. Cells were washed additional 2 times with PBS before processing them by the scChIC protocol for the histone modifications H3K27me3, H3K4me1 and H3K27me3 + H3K4me1. In case of the fixed cells experiment, isolated cells were resuspended in 70% ethanol and frozen down at -80°C prior to processing by the scChIC protocol for the histone modifications H3K27me3, H3K9me3 and H3K27me3 + H3K9me3.

### scChIC-seq experiments

scChIC-seq experiments were performed as described by Zeller et al. (*in preparation*):
Cells were processed in 0.5ml protein low-binding tubes. Following steps are performed on ice. Cells were resuspended in 500$\mu$l Wash Buffer (47.5ml $H_2O$ RNAse free, 1ml 1M HEPES pH 7.5 (Invitrogen), 1.5ml 5M NaCl, 3.6$\mu$l pure spermidine solution (Sigma Aldrich)). Cells were pelleted at 600g for 3min and resuspended in 400$\mu$l Wash Buffer 1 (wash buffer with 0.05% saponin (Sigma Aldrich), protease inhibitor cocktail (Sigma Aldrich), 4$\mu$l/ml 0.5M EDTA) containing the primary antibody (1:100 dilution for H3K4me1, H3K27me3 and H3K4me1+H3K27me3 Saponin has to be prepared fresh every time as a 10% solution in PBS). Cells were incubated overnight at 4°C on a roller, before they were washed once with 500$\mu$l Wash Buffer 2 (wash buffer with 0.05% saponin, protease inhibitor). Afterwards cells were

resuspended in 500$\mu$l Wash Buffer 2 containing PaMN (3ng/ml) and incubated for 1h at 4°C on a roller. Finally, cells were washed an additional 2 times with 500$\mu$l Wash Buffer 2 before passing it through a 70$\mu$m cell strainer (Corning, 431751) and sorting G1 cells based on Hoechst staining on a JAZZ FACS machine into 384 well plates containing 50nl Wash buffer 3 (Wash buffer containing 0.05% saponin) and 5$\mu$l sterile filtered mineral oil (Sigma Aldrich) per well. Small volumes were distributed using a Nanodrop II system (Innovadyme).

**Protein A-MN activation**

50nl of Wash Buffer 3, containing 4$\mu$M CaCl2, were added per cell to induce Protein A-MN mediated chromatin digestion that was performed for 30min on ice. Afterwards the reaction was stopped by adding 100nl of a stop solution containing 40mM EGTA (chelates Ca2+ and stops MN), 1.5% NP40 and 10$\mu$l 2mg/ml proteinase K (Ambion). Chromatin is subsequently released and PaMN permanently destroyed by proteinase K digestion at 65°C for 6h followed by 80°C for 20min to heat inactivate proteinase K. Plates can be stored at -20°C until further processing.

**Library preparation**

DNA fragments are further blunt ended using Klenow large fragment (NEB) and T4 PNK (NEB) in the presence of dNTPs. Afterwards nonintegrated dNTPs are removed using shrimp alkaline phosphatase (NEB) to increase A-tail efficiency. Subsequently blunt fragments are A-Tailed using AmpliTaq 360 DNA Polymerase (Thermofisher) in the presence of fresh dATP and T4 PNK (NEB). Next fragments are ligated to T-tail adaptors containing 3 bp UMIs and an 8bp cell barcode (final concentration 9nM from IDT) for 16h at 16°C using T4-ligase (NEB). Adaptors were added with a Mosquito HTS (ttp labtech). Ligation products were pooled by centrifugation and cleaned using Ampure XP beads at a bead to sample ratio of 0.8 (Beckman Coulter). The cleaned DNA is then linear amplified by invitro transcription with components of MessageAmp II (Invitrogen) for 12h at 37°C. The produced aRNA is purified using RNA Clean XP beads (Beckman Coulter) at 0.8 beads to sample ratio, followed by DNAse1 digestion. aRNA is cleaned up again, followed by aRNA fragmentation for 1.5 min at 94°C. After another bead cleanup, 40% of the aRNA is reverse transcribed with Superscript II (NEB). Single stranded DNA is purified through RNAseA incubation and one DNA bead cleanup followed by PCR amplification to add the Illumina smallRNA barcodes and handles with NEBNext Ultra II Q5 Master Mix. PCR cycles depended on the abundance of the histone modification assayed and the number of aRNA cleanup cycles after IVT needed

(8-10 for H3K9me3 and H3K27me3, 10-12 for H3K4me1).

## Processing of ethanol fixed cells for scChIC protocol and surface antibody incubation

For the ethanol fixed cells the above described scChIC-seq protocol was adapted. Wash Buffers were used as described above, except that 0.05% saponin was exchanged for 0.05% Tween. Ethanol fixed cells were thawed on ice. Cells were spun at 400g for 5min and washed once with 400$\mu$l Wash Buffer 1. Cells were spun again at 400g and resuspended in 400$\mu$l Wash Buffer 1. Cell suspension was split into 3 samples each having a volume of 400$\mu$l and incubated with one or two antibodies (1:100 dilution for H3K27me3, H2K9me3 and H3K27me+H3K9me3) overnight on a roller at 4 deg C. The next day cells were spun at 400g, washed once with 400$\mu$l Wash Buffer 2 and resuspended in 500$\mu$l Wash Buffer 2 containing PaMN (3ng/ml) and incubated for 1h on a rotator at 4°C. Next, cells were spun at 400g and resuspended in 400$\mu$l Wash Buffer 2 (with addition of 5% blocking rat serum). Surface antibodies were added according to this concentrations and were incubated for 30min on ice:

| antibody | info | working concentration |
|---|---|---|
| GR1 | A647 0.5mg/ml dilute 2x | 1:8000 |
| NK1 | A488 0.5mg/ml <0.25$\mu$g/100$\mu$l | 1:400 |
| CD19 | BV421 | 1:200 |
| CD3 | APC/cy7 0.2mg/ml <0.25$\mu$g/100$\mu$l | 1:100 |

Finally, samples were washed once with 500$\mu$l Wash Buffer 2 before passing them through a 70$\mu$m cell strainer (Corning, 431751) and sorting on a JAZZ FACS machine, with surface antibody specific gating 4.1, into 384 well plates containing 50nl Wash buffer 3 (Wash buffer containing 0.05% Tween) and 5$\mu$l sterile filtered mineral oil (Sigma Aldrich) per well. Small volumes were distributed using a Nanodrop II system (Innovadyme).

## Data processing

All DNA libraries were sequenced on a Illumina NextSeq500 with 2x75bp. Standard workflow of scCHIC data processing from *singlecellmultiomics* (github.com/BuysDB/SingleCellMultiOmics) was followed including fastq files demultiplexing,reads mapping with bwa and reads binning into 50kb bins for generating count tables. After filtering count tables for low counts cells and bins, LDA topic anaylsis was performed and subsequent multinomial analysis for unmixing https://github.com/jakeyeung/dblchic.

### Unmixing scChIX signal

We model the double-incubated scChIC-seq data as a mixture of two multinomial distributions, one coming from a cluster of histone modification $a$, the other from another cluster of the second histone modification $b$ (Methods, Supplementary Material). The log-likelihoood for a mixture of two multinomials is:

$$\mathrm{LL}_{(c,d)} = \log(P\left(\vec{y}|\vec{\alpha}^a, \vec{\alpha}^b, w\right)) \propto \sum_{k=1}^{K} y_k \log\left(w\alpha_k^a + (1-w)\,\alpha_k^b\right)$$

where $\vec{y}$ is the number of cuts across the genome for a double-incubated cell, $\vec{\alpha}^a$ and $\vec{\alpha}^b$ are relative frequencies for histone modifications $a$ and $b$ across $K$ genomic regions, $w$ is the fraction of histone modification $a$ relative to $b$ in the double-incubated cell. We estimate $w$ by maximizing the log-likelihood given $\vec{y}, \vec{\alpha}^a, and \vec{\alpha^b}$

For each double-incubated cell, we infer which pair of multinomials $\vec{\alpha}^a$ and $\vec{\alpha}^b$ most likely generated the genomic distribution of cuts by calculating the probability that a double-incubated cell comes from a particular pair $(c, d)$:

$$p_{(c,d)} = \frac{e^{L_{(c,d)}}}{\sum\limits_{(f,g)\in pairs} e^{L_{f,g}}},$$

where $L$ is the maximum likelihood of the fit.

For a double-incubated cell, we calculate the $p_i$ for each pair of $\vec{\alpha}^a$ and $\vec{\alpha}^b$ and select the pair with the highest probability.

We use the single-incubated scChIC-seq data to generate $C$ and $D$ clusters for histone modifications $a$ and $b$, respectively:

$$\vec{\alpha_{1,j}}^a, ..., \vec{\alpha_{C,j}}^a$$

$$\vec{\alpha_{1,j}}^b, ..., \vec{\alpha_{D,j}}^b$$

which we can then use to calculate the log-likelihood for each pair of multinomials $\left(alpha_c^a, alpha_d^b\right)$.

The relative frequencies from the training dataset are inferred using the Latent Dirichlet Allocation model and a Gibbs sampling algorithm for inferring the relative frequencies (LDA, Methods). Briefly, we cluster cells after running LDA to determine the number of clusters for each histone modification independently. The

relative frequencies for a cluster are taken as the arithmetic mean of the relative frequencies of the cells within the cluster.

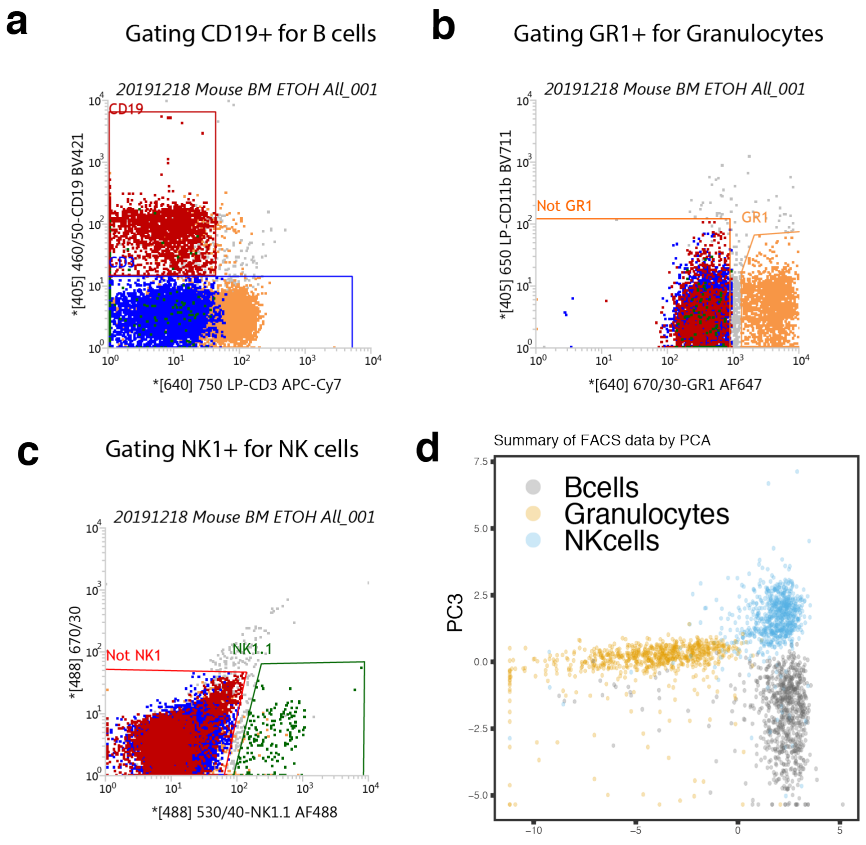
After inferring the cluster pairs $\left(alpha_c^a, alpha_d^b\right)$ and mixing fraction $w$, we can assign each genomic cut $y_{i,j}$ of a double-incubated cell $i$ in genomic region $j$ to histone mark $\alpha^a$ with probability $q_{i,j}$:

$$q_{i,j} = \frac{w\alpha_c^a}{w\alpha_c^a + (1-w)\,\alpha_d^b}$$

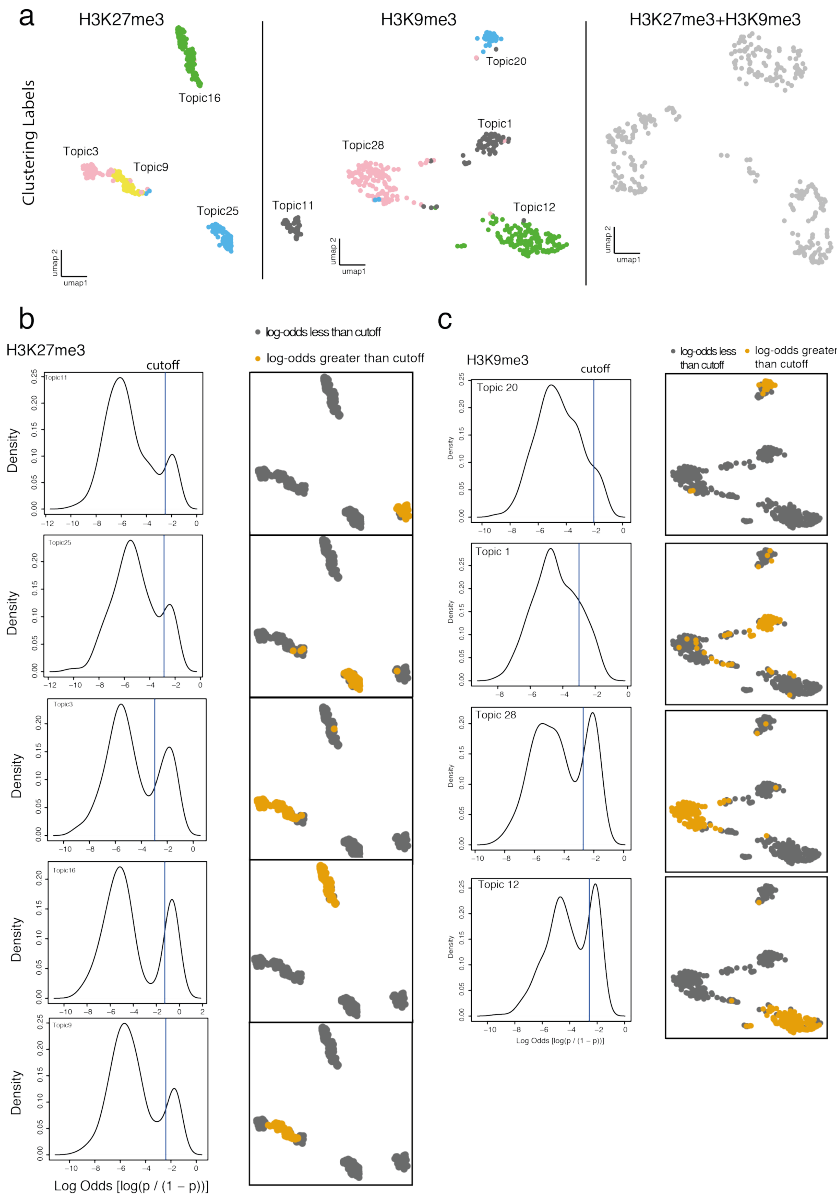
To transfer labels from one scChIC-seq data targeting histone modification $a$ to another scChIC-seq data targeting modification $b$, we use the double-incubated scChIC-seq to link together the two single-incubated datasets. After unmixing the double-incubated scChIC-seq data into their respective histone modifications, we project the $a$ part of the double-incubated scChIC-seq data (denoted as $a_{unmixed}$) onto LDA model trained by the scChIC-seq data $a$. We assign a cell from $a_{unmixed}$ to cluster $c$ if the cell to topic weight in $a_{unmixed}$ is greater than a threshold fit from a Gaussian mixture model (Methods, Section 1). After having transferred labels from $a$ to the double-incubated cells, we project the $b_{unmixed}$ onto $b$. Similarly, we assign cells in $b_{umixed}$ to a cluster in $b$. Each cluster in $b$ will have a number of cells from $b_{unmixed}$ assigned to it, some possibly with different labels. We assign the cluster in $b$ to a label defined in $b_{unmixed}$ by summing the weights across cells for each label and choosing the label which has the highest summed weight.

In the final analysis, there may be some cells that were not assigned a label because the topic to cell weight did not pass the threshold for any of the topics. We impute the labels for these cells by using the labels from the 50-nearest neighbors and taking the label with the most cells. The kNN graph is defined from the cell to topic matrix using Euclidean distance.

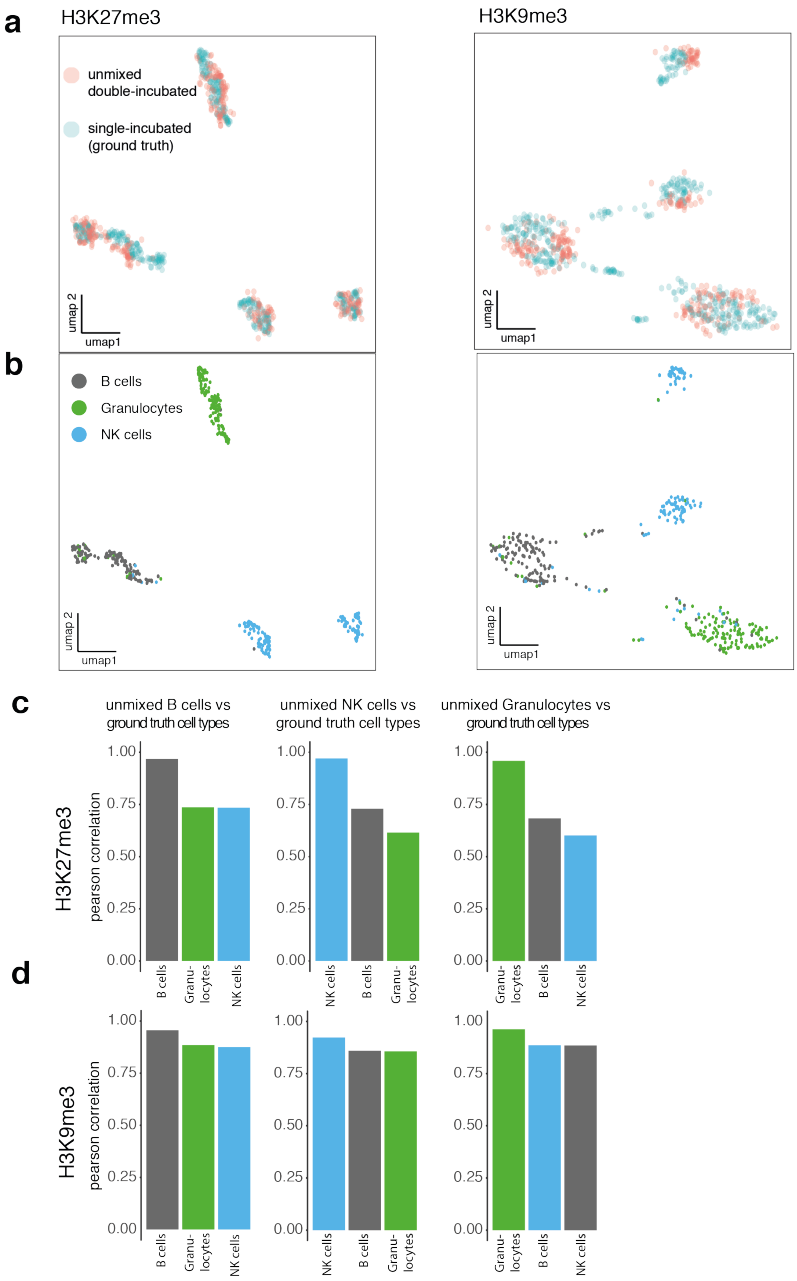
# Supplementary Material



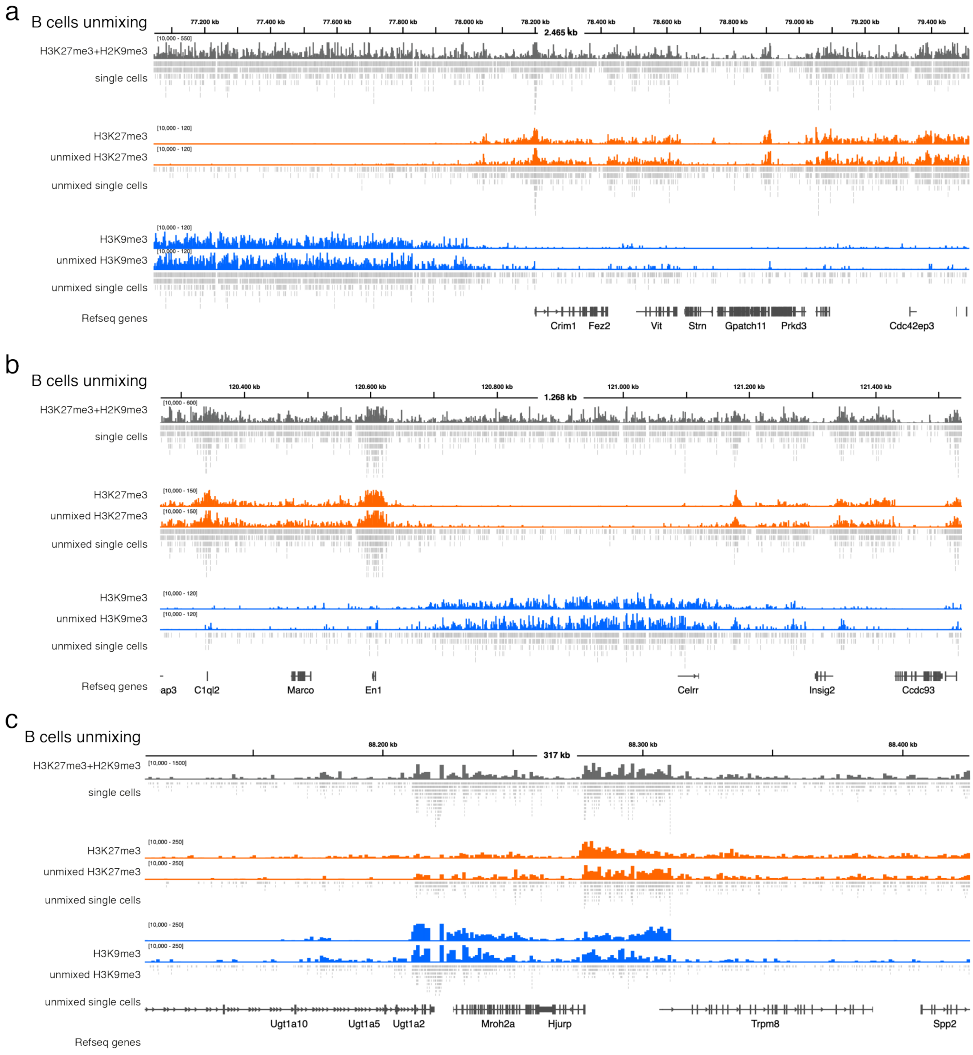

Supplementary Figure 4.1: **FACS gating strategies of fixed cells with surface antibodies.** **a** Gating strategy of CD19$^+$ cells for B cells sorting. **b** Gating strategy of GR1$^+$ cells for Granulocytes sorting. **c** Gating strategy of NK1$^+$ cells for NK cells sorting. **d** PCA representation of FACS parameters for the three sorted celltypes.



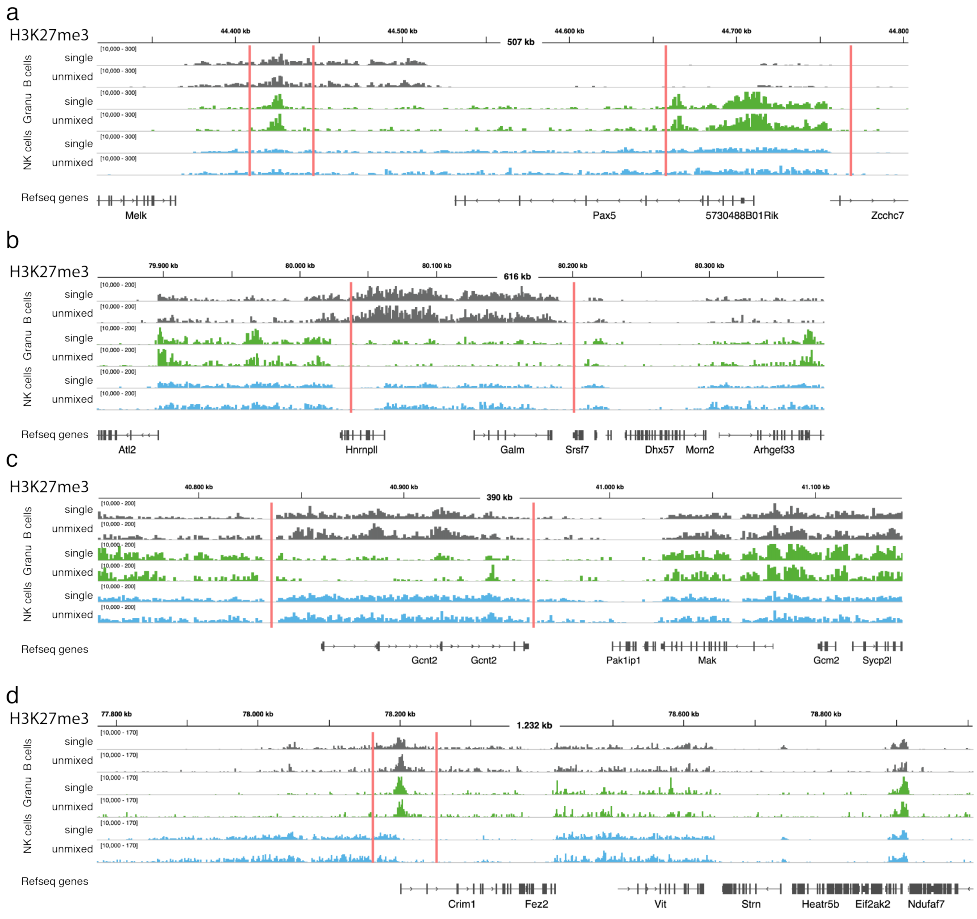

Supplementary Figure 4.2: **Celltype-specific clusters in scChIC-seq data.** **a** UMAP of H3K27me3 (left), H3K9em3 (middle), and H3K27me3+H3K9me3 (right) single-cell histone modification levels. For the single-incubated datasets, the clusters are colored and labeled as topics. Double-incubated cells are not colored to emphasize that scChIX deconvolves each double-incubated cell individually. **b** Density plot (left) of topic weights for different celltype-specific topics across cells. Vertical line shows the cutoff above which cells are clustered into the topic. UMAP visualization of cells (right) assigned to each cluster based on cutoffs from the corresponding topics on the left. **c** analogous to (b) but showing clusters in the H3K9me3 signal.



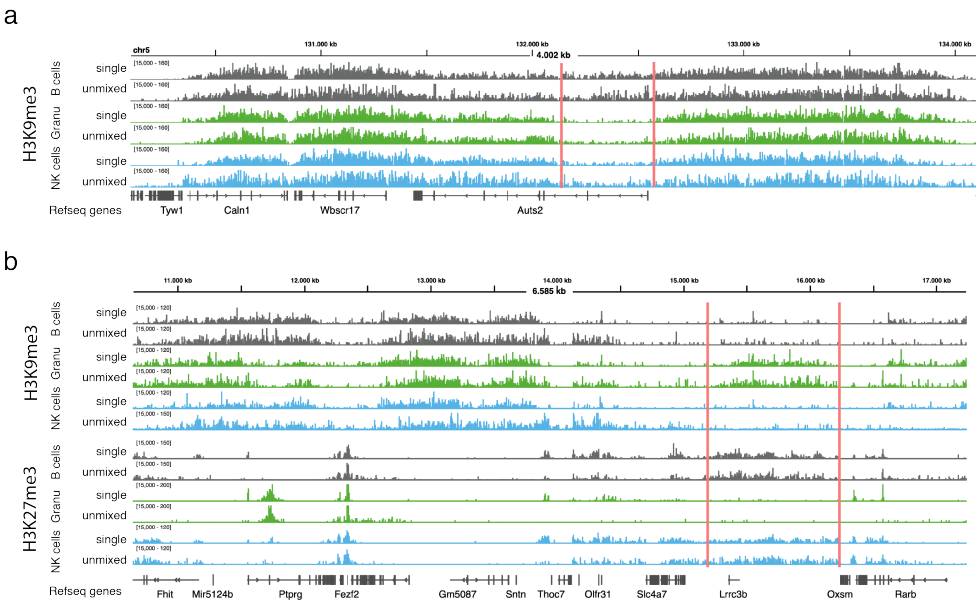

Supplementary Figure 4.3: **Unmixed double-incubated scChIC-seq intermingles with ground truth dataset and correlates with ground truth dataset. a, b** UMAP representation of H3K27me3 (left) and H3K9me3 (right) data colored by antibody incubation condition (a) or ground truth celltype labels defined by FACS (b). **c, d** Genome-wide pearson correlation between H3K27me3 (c) and H3K9me3 (d) H3K27me3+H3K9me3 signal versus ground truth scChIC-seq purified by FACS. Colors are from FACS-defined celltypes.



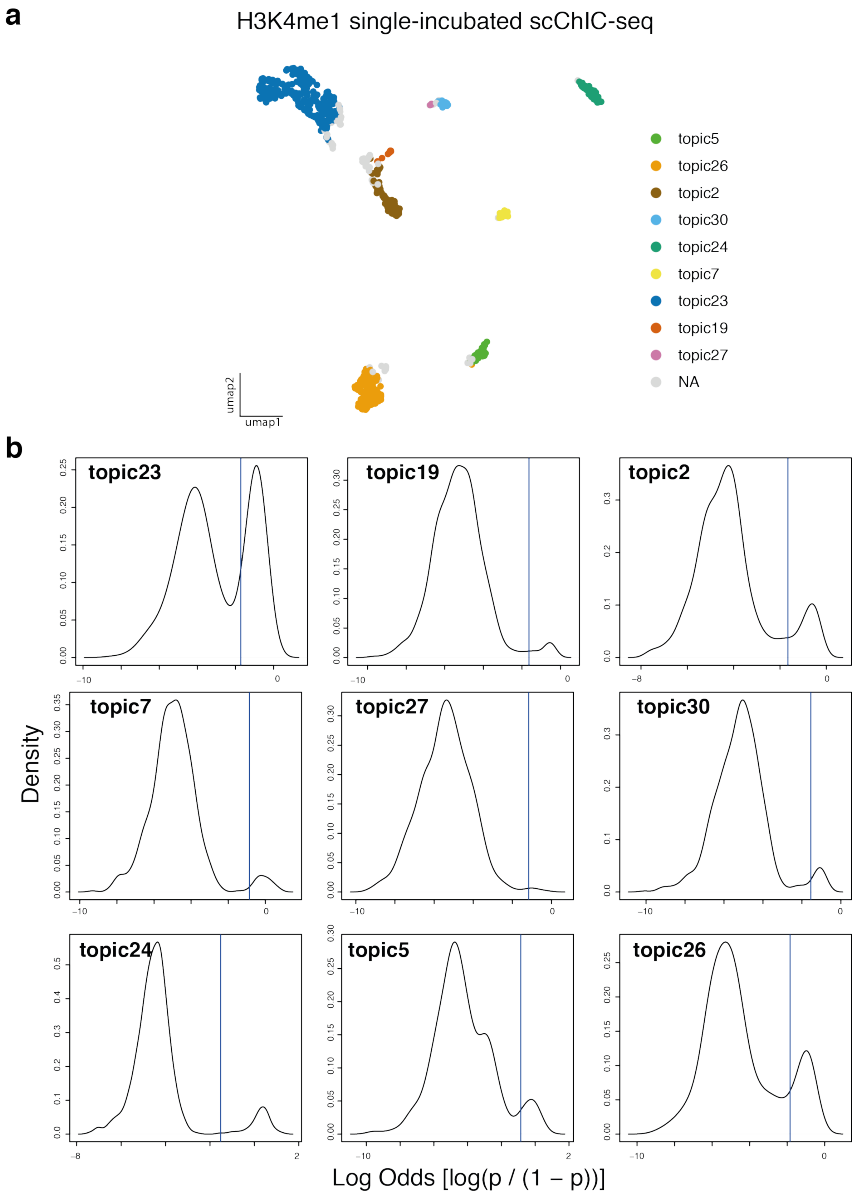

Supplementary Figure 4.4: **Coverage tracks visualisation of H3K27me3+H3K9me3, unmixed H3K27me3 or H3K9me3 and ground truth H3K27me3 or H3K9me3 for B cells in different regions. a, b, c.** Double targeted signal in grey, H3K27me3 single and unmixed signal in orange and H3K9me3 single and unmixed signal in blue. Under each H3K27me3+H3K9me3 as well as unmixed H3K27me3 or H3K9me3 the raw single cells molecules are shown (a total of 100 B cells).



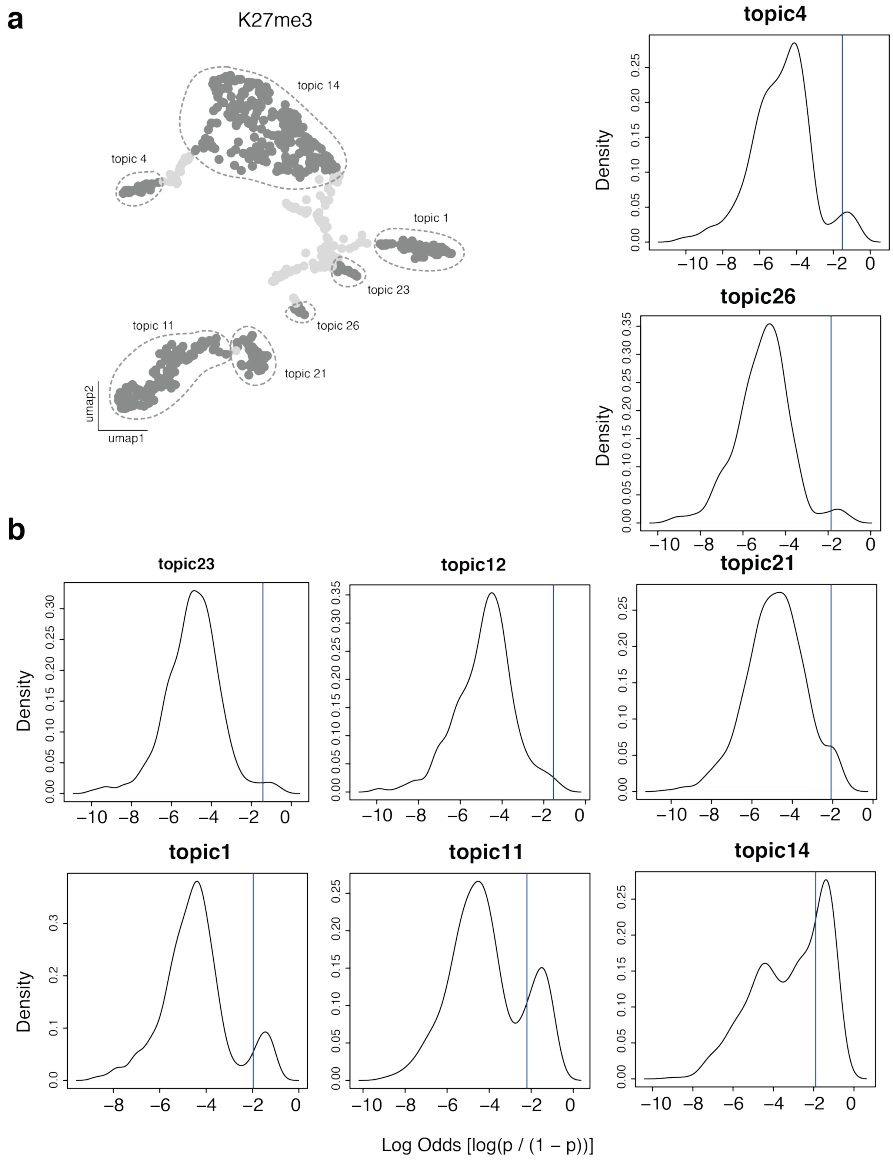

Supplementary Figure 4.5: **Coverage tracks of H3K27me3 celltype specific regions.a** H3K27me3 coverage tracks showing the region around *Pax5* for H3K27me3 (single) and for the unmixed H2K27me3 (unmixed) for B cells (grey), Granulocytes (green) and NK cells (blue), respectively. The region around *Pax5* is shown to be Granulocytes specific. **b** As described in (a), but for region around *Hnrnpll* and *Galm* which are B cells specific. **c** As described in (a) but for region around *Gcnt2* which appear to be specific in B cells and NK cells. **d** As described in (a) but for region around *Crim1* which appears to be B cells and Granulocytes specific.



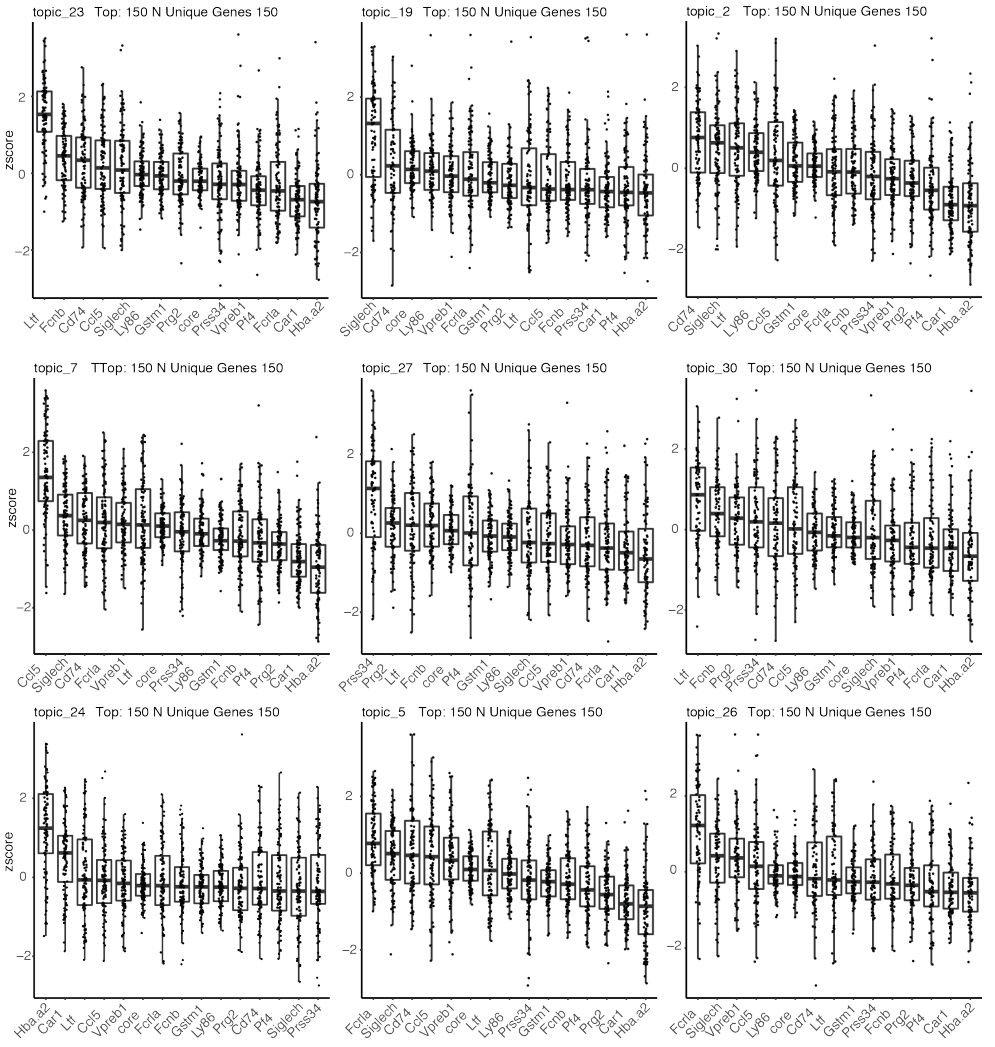

Supplementary Figure 4.6: **Coverage tracks of H3K9me3 celltype specific regions. a** H3K9me3 coverage tracks showing the region around *Auts2* for H3K9me3 (single) and for the unmixed H2K9me3 (unmixed) for B cells (grey), Granulocytes (green) and NK cells (blue), respectively. The region around *Auts2* appears to be B cell specific. **b** As described in a, but for regions around *Lrrc3b* which appears to be Granulocyte specific.



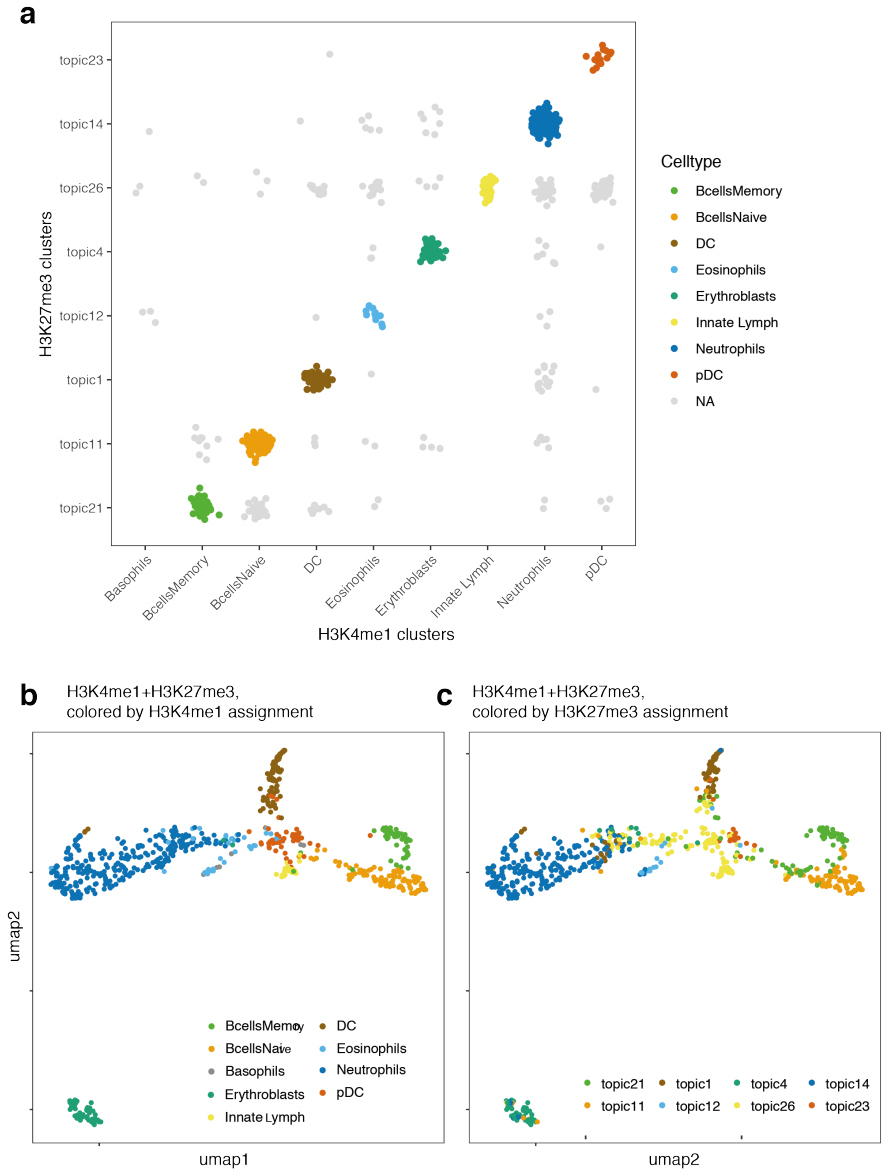

Supplementary Figure 4.7: **Topic weights from LDA define clustering of cells in H3K4me1 scChIC-seq.** Nine topics assign cells to a cluster. UMAP shows relationships between cells in low-dimensional space, colored by cells assigned to a celltype-specific topic. Cells are assigned to a topic by fitting a Gaussian mixture model to the distribution of topic weights, $p \in [0, 1]$, across cells. Vertical line indicates the threshold at which the value of the two Gaussian mixtures are equal. Cells that did not have topic weights exceeding the threshold were assigned as NA (grey cells in UMAP, 75/714 cells). NA cells were not used to define the training data but were nevertheless used for the final plot in (Fig. 4.3d ), where the NAs were assigned to the nearest cluster based on K-nearest neighbours (K=50).



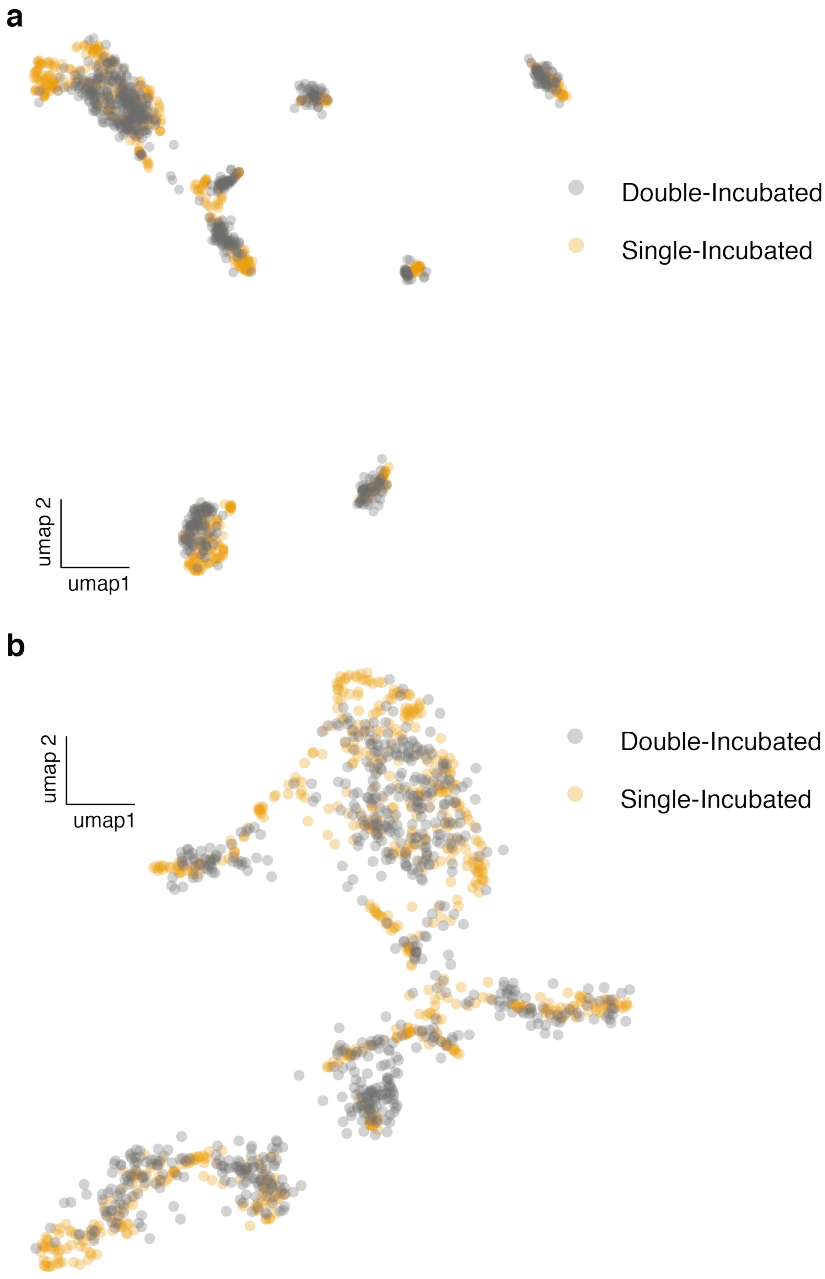

Supplementary Figure 4.8: **Topic weights from LDA define clustering of cells in H3K27me3 scChIC-seq.** Eight topics assign cells to a cluster. UMAP shows relationships between cells in low-dimensional space, colored by cells assigned to a celltype-specific topic. Cells are assigned to a topic by fitting a Gaussian mixture model to the distribution of topic weights, $p \in [0, 1]$, across cells (density plots are plotted in log odds of topic weight. Vertical line indicates the topic weight at which the value of the two Gaussian mixtures are equal. Cells that did not have topic weights exceeding the threshold were assigned as NA (grey cells in UMAP, 158/689 cells). NA cells were not used to define the training data but were nevertheless used for the final plot in (Fig. 4.3d ), where the NAs were assigned to the nearest cluster based on K-nearest neighbors (K=50).



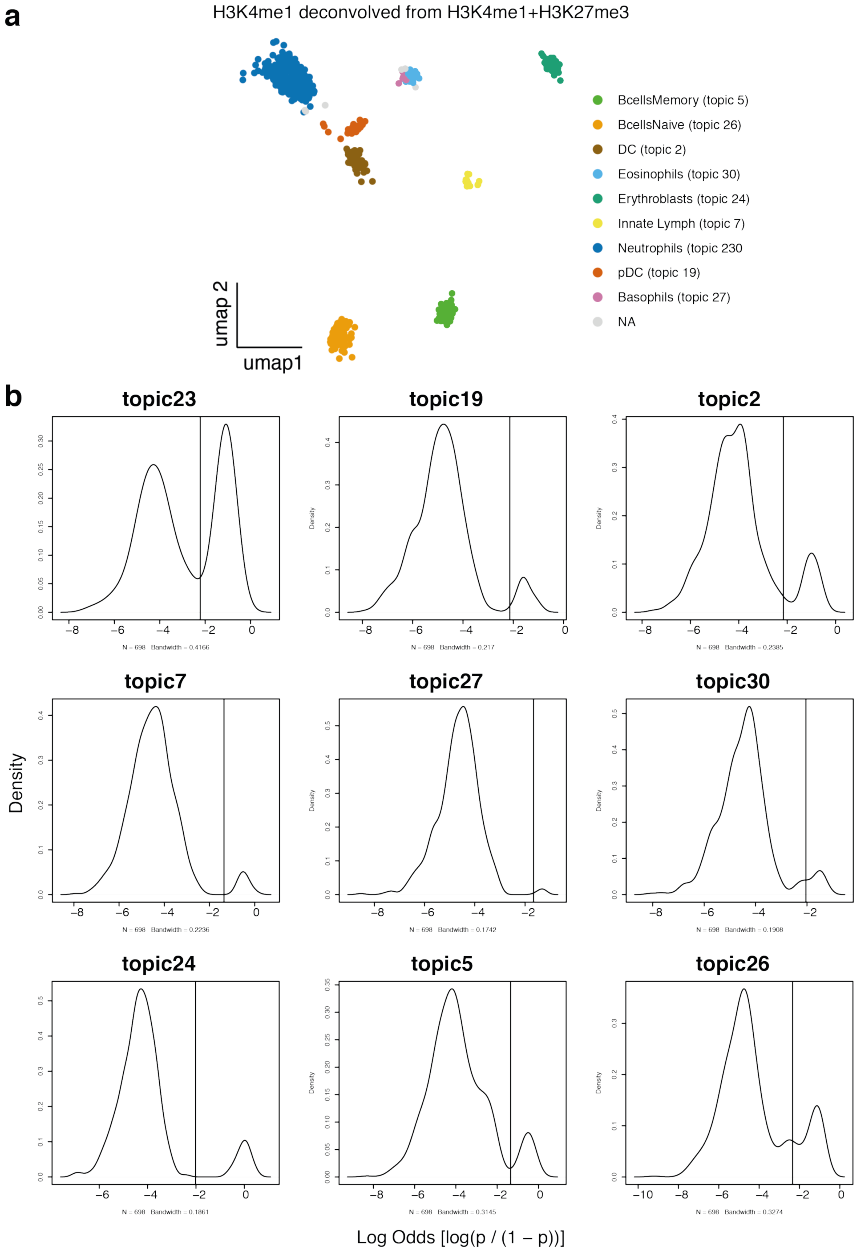

Supplementary Figure 4.9: **Genes associated to topics show celltype-specific expression.** UMAP shows relationships between cells in low-dimensional space, colored by cells assigned to a celltype-specific topic. Genomic regions are assigned to the nearest gene up to a distance of 50kb. For each topic, the top 150 regions are selected and their associated mRNA levels are compared by calculating the z-score of log expression levels across bone marrow celltypes using scRNA-seq data[37]. Each topic shows a specific enrichment for a celltype, which is used to define a H3K4me1 topic to a celltype. Pseudobulks(x-axis of boxplots) are defined by their marker gene expression. Ltf=neutrophil, Siglech=plasmacytoid dendritic cells, Cd74=conventional dendritic cells, Ccl5=innate lymphocyte, Hba-a2=erythroblasts, Fcrla=Bcells.



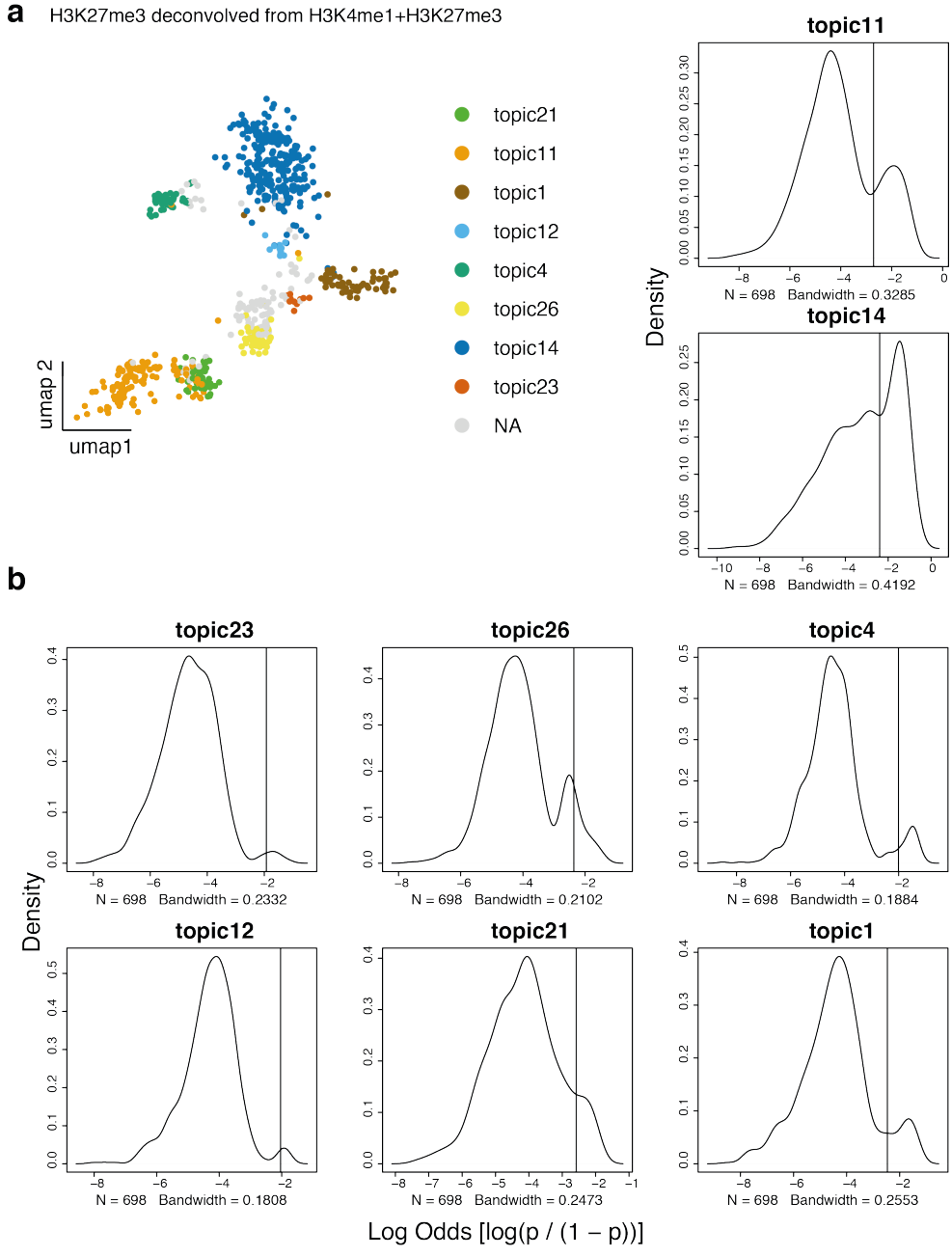

Supplementary Figure 4.10: **H3K4me1+H3K27me4 deconvolves into specific pairs of H3K4me1 and H3K27me3 clusters.** **a** Assignment plot showing individual H3K4me1+H3K27me3 cells (represented as dots) assigned to a pair of topics (x-axis are H3K4me1 topics, named by their celltype, while y-axis are H3K27me3 clusters). Colored diagonals represent the inferred matching between H3K4me1 and H3K27me3 topics. **b,c** UMAP representation of H3K4me1+H3K27me3 scChIC-seq, each cell colored by their assigned H3K4me1 (b) or H3K27me3 cluster(c). The inferred matching between H3K4me1 and H3K27me3 corresponds largely match between (b) and (c). Discrepancies occur with a minority of cells near the center of the UMAP.



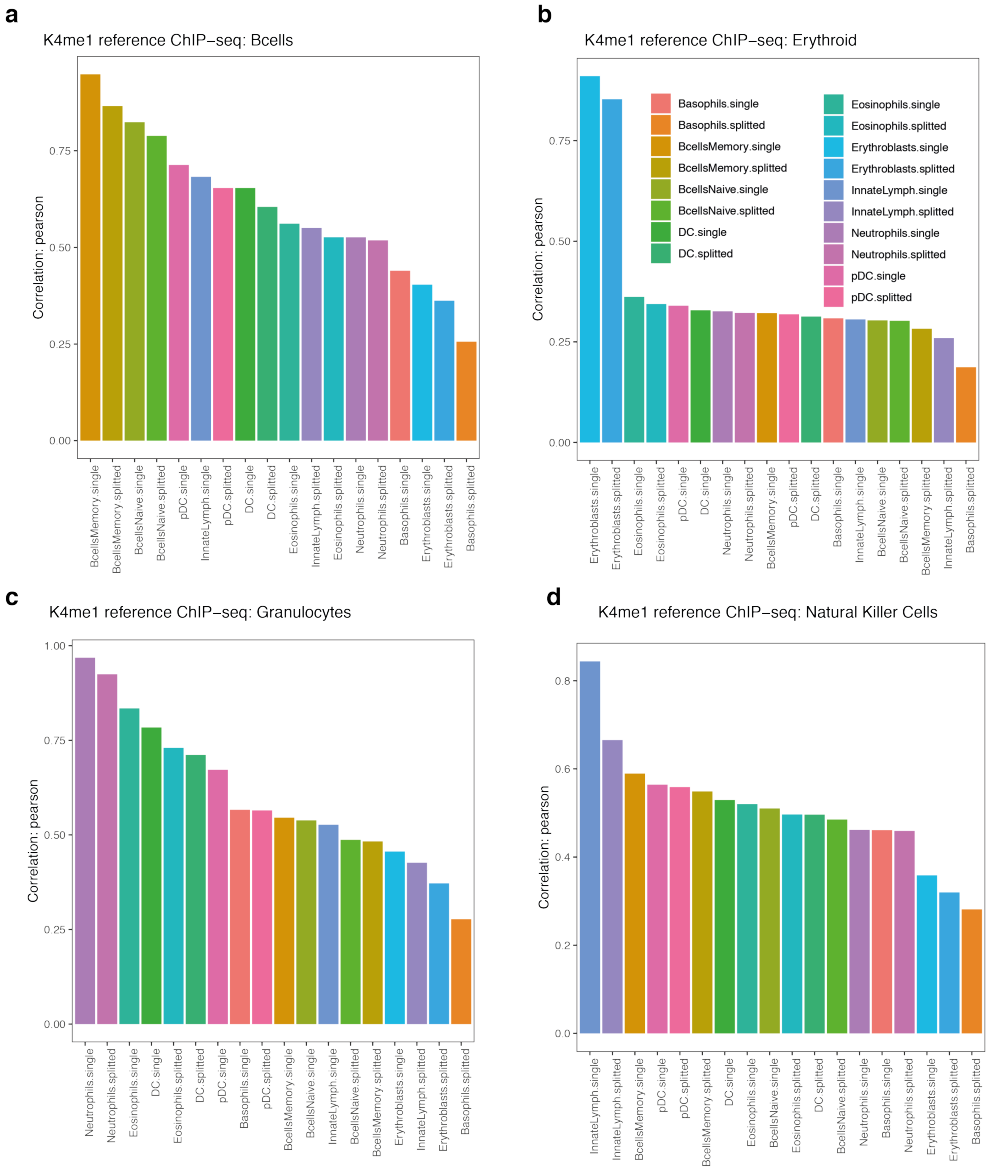

Supplementary Figure 4.11: **Deconvolved H3K4me1+H3K27me4 scChIC-seq signal projects accurately onto their respective histone modification space. a,b** UMAP representation of H3K4me1 (a) and H3K27me3 (b) single- and double- incubated scChIC-seq. Single- and double-incubated scChIC-seq datasets intermingle with each other, suggesting that our model accurately deconvolves the H3K4me1+H3K27me3 signal.



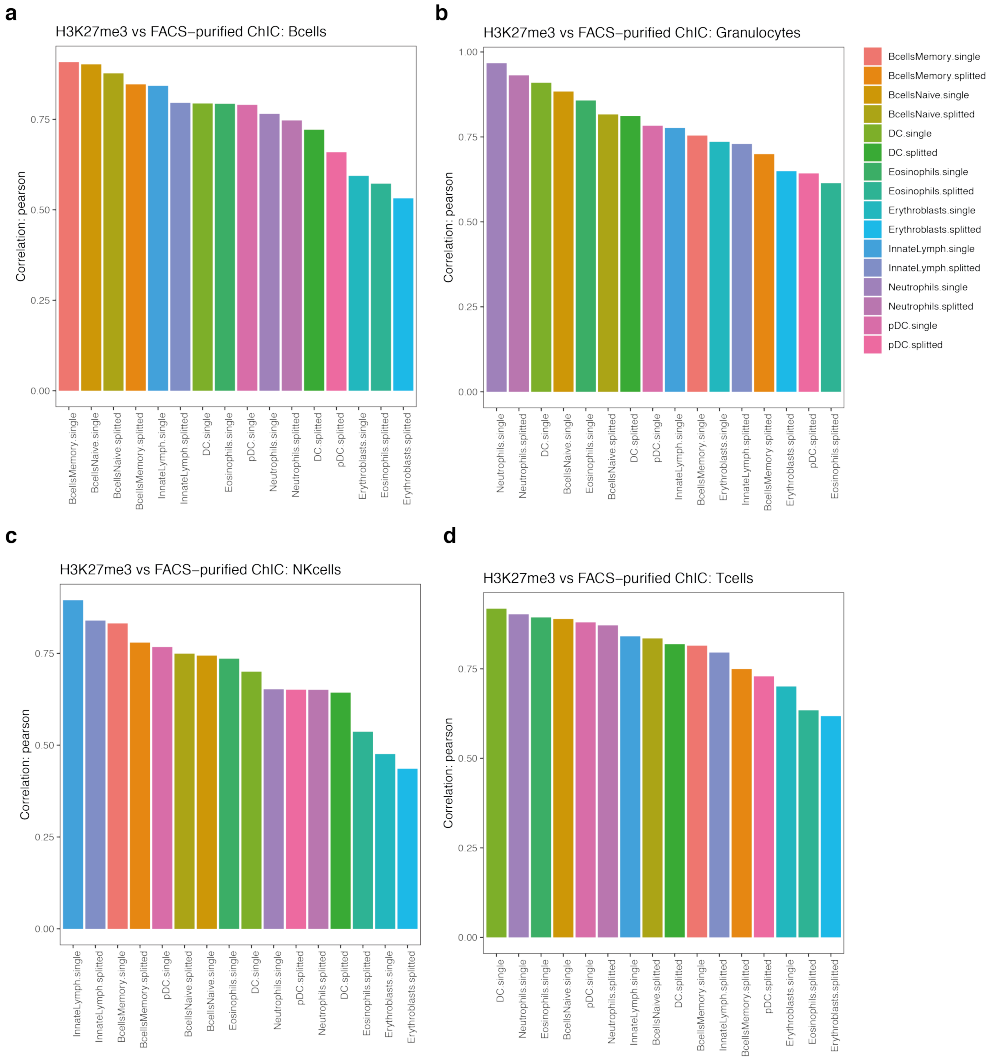

Supplementary Figure 4.12: **Topic weights from LDA transfers labels from H3K4me1 to H3K4me1+H3K27me3.** **a** H3K4me1 part of H3K4me1+H3K27me3 scChIC-seq signal projected onto the H3K4me1 LDA embedding, visualized on a UMAP. **b** Nine topic weights (same as in Supplementary Fig. 4.7 are quantified in the H3K4me1 part of the H3K4me1+H3K27me3 scChIC-seq signal. For each topic, the log odds of the topic weights $p$ are plotted as a density plot. Vertical lines indicate threshold above which cells are assigned to the topic (Methods). The cells in UMAP (a) are colored if they their topic weight is higher than the threshold.



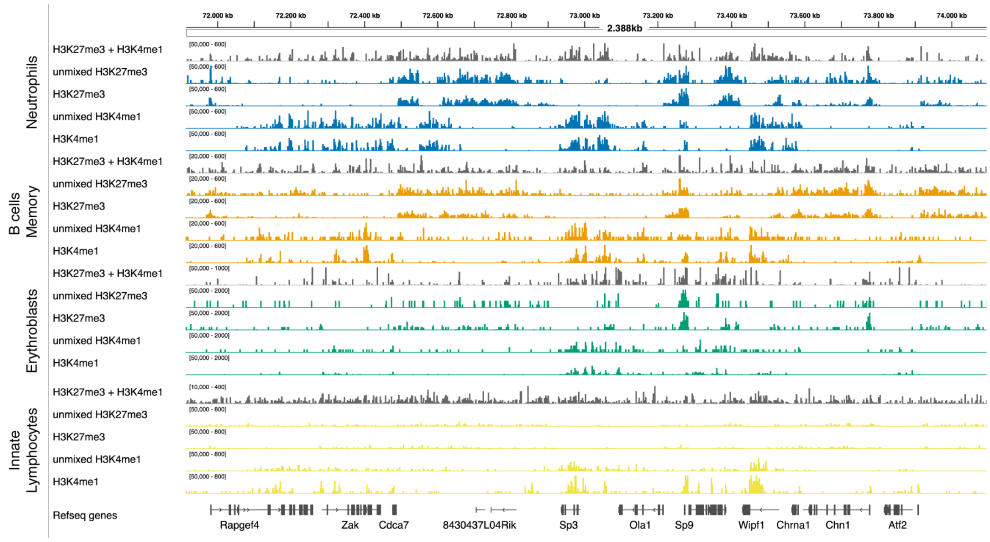

Supplementary Figure 4.13: **Topic weights from LDA transfers labels from H3K27me3 to H3K4me1+H3k27me3.** **a** H3K4me1 part of H3K4me1+H3K27me3 scChIC-seq signal projected onto the H3K27me3 LDA embedding, visualized on a UMAP. **b** Eight topic weights (same as in Supplementary Fig. 4.8 are quantified in the H3K4me1 part of the H3K4me1+H3K27me3 scChIC-seq signal. For each topic, the log odds of the topic weights $p$ are plotted as a density plot. Vertical lines indicate threshold above which cells are assigned to the topic (Methods). The cells in UMAP (a) are colored if they their topic weight is higher than the threshold.



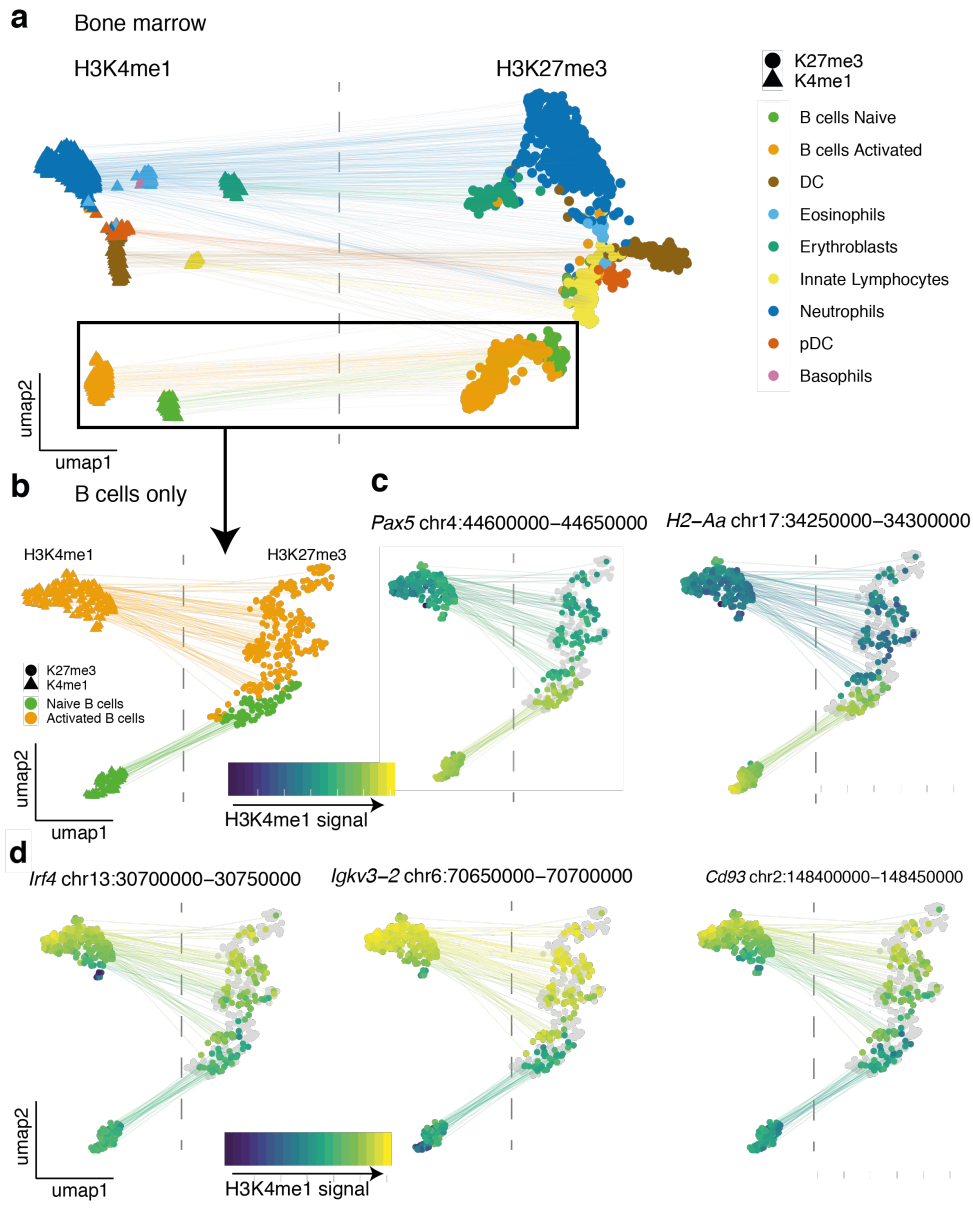

Supplementary Figure 4.14: **Histone modification signal correlates with publicly available H3K4me1 ChIP-seq data.** X-axis represent deconvolved H3K4me1, y-axis is the pearson correlation between the deconvolved H3K4me1 signal at transcription start sites versus ChIP-seq signal of H3K4me1 from sorted blood from Lara-Astiaso et al.



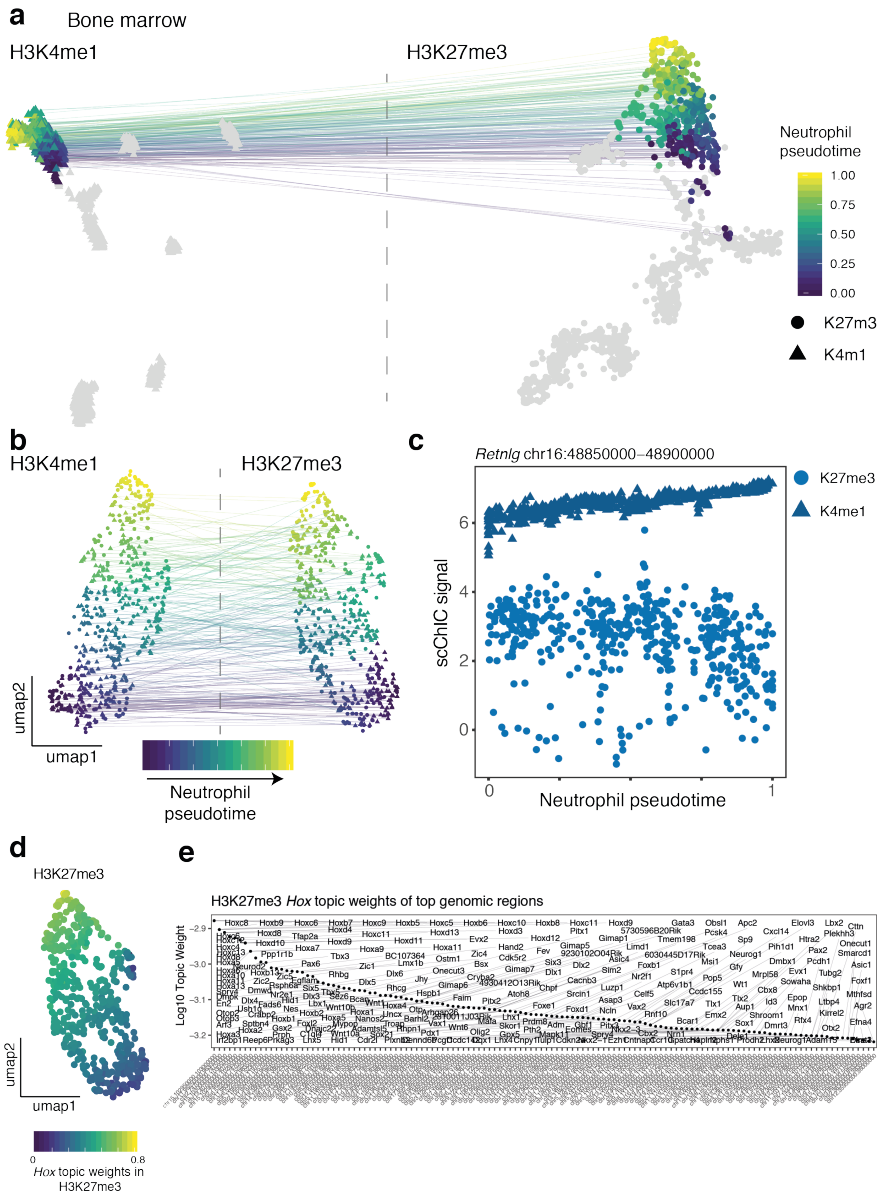

Supplementary Figure 4.15: **Histone modification signal correlates with H3K27me3 scCHiC-seq data from sorted celltypes.** X-axis represent deconvolved H3K4me1, y-axis is the pearson correlation between the deconvolved H3K4me1 signal at transcription start sites versus scChIC-seq signal of H3K27me3 from sorted celltypes from Fig. 2.




Supplementary Figure 4.16: **Coverage tracks visualization of mixed and unmixed H3K4me1 and H3K27me3** Coverage profiles for the 4 largest identified celltype groups (Neutrophils=blue, B cells memory=orange, Erythroblasts=green, Innate Lymphocytes=yellow). For each celltype we show a coverage tracks of the mixed signal of H3K4me1+H2K27me3 on top (grey), followed by the unmixed and the single-targeted histone modification signal for H3K27me3 or H3K4me1, respectively.




Supplementary Figure 4.17: **Re-clustering on B cells reveals heterogeneity within B cells.** **a** UMAP visualization of H3K4me1 and H3K27me3 (single signal and unmixed signal), colored by celltypes derived from H3K4me1 and transferred to H3K27me3. Black rectangle indicates the B cell population used to re-cluster in (b,c,d). **b** UMAP of naive and activated B cells after subsetting B cells. **c,d** Projection of H3K4me1 signal of marker genes for naive B cells (c) or for activated B cells (d). H3K4me1 signal is measured in all cells of the H3K4me1 UMAP (i.e. both single- and double-incubated have H3K4me1 signal in the H3K4me1 UMAP). Double- (colored) but not single-incubated (grey) cells have H3K4me1 signal in the H3K27me3 UMAP.




Supplementary Figure 4.18: **H3K4me1 and H3K27me3 signal of during neutrophil maturation.** **a** UMAP visualization of H3K4me1 and H3K27me3, lines join H3K4me1 and H3K27me3 embeddings of double-incubated neutrophils. Heterogeneity within neutrophils are colored as neutrophil pseudotime. **b** Joint visualization of H3K4me1 and H3K27me3 in neutrophils only, colored by pseudotime same as in (a). **c** H3K4me1 and H3K27me3 modification levels at the *Retnlg* (a mature neutrophil marker gene) locus along neutrophil pseudotime. **d** UMAP of H3K27me3 signal across single cells colored by weights of a topic containing high H3K27me3 levels at many *Hox* and developmental gene loci (*Hox* topic). **e** Topic weights of the top 150 genes associated with loci in the *Hox* topic for H3K27me3.

## Summary of the unmixing strategy for scChIX

scChIX is an integrated experimental and computational framework to multiplex histone modifications in single cells. Our strategy unlocks two novel features in single-cell epigenetics: (1) Multimodal analysis of histone modifications in single cells. (2) Transfer of labels from one single-cell histone modification dataset to another (e.g. from an active mark, where identifying celltypes is straightforward, to a repressive mark, where cell type information is less available). We combine single-cell double immunocleavage followed by computational unmixing to profile two histone modifications in single cells.

Conventional scChIC-seq technique (single-cell chromatin immunocleavage sequencing) is an enzyme-tethering strategy whereby antibodies targeting a specific histone modification tethers MNase-proteinA fusion proteins[109]. This tethering allows MNase-proteinA to make specific cuts at genomic regions associated with the histone modification of interest in single cells. To profile two histone modifications in single cells, we combine double immunocleavage followed by computational unmixing. By staining cells with two antibodies each targeting a different histone modification, the MNase-proteinA cuts at genomic regions associated with either histone modification. We probabilistically assign each cut to their respective histone modification by using training data from standard scChIC-seq experiments targeting each histone modification separately (i.e., scChIX requires one double-scChIC-seq and two single-scChIC-seq experiments all from the same biological system).

We develop a statistical framework based on multinomial models to computational unmix the double-scChIC-seq experiments into their respective histone modifications. We assume a generative model where double-scChIC-seq data comes from a mixture of two multnomials, one for each histone modification. For each histone modification, many possible genomic profiles are possible, depending on the complexity of the biological system (e.g. each genomic profile could correspond to a cell-type specific genomic profile). To constrain the space of genomic profiles, we use the training data to generate a limited number of possible genomic profiles which a particular biological system can generate. A genomic profile is parameterized as a multinomial model where each region (genomic bins of fixed size) has a relative frequency of generating a cut. Finally, for each double-cut cell, we infer which pair of multinomials (one genomic profile for each histone modification) are mixed together and at what mixing ratios to generate the observed cuts across the genome in the double-scChIC-seq data (Methods).

After having inferred the origins (both the specific pair of genomic profiles as well as the mixing ratio) from which each double-cut cell originated, each cut can then be probabilistically assigned back to its respective histone modification. These

region- and cell-specific assignment probabilities reflect the degree of overlap between the two histone modifications across the genome. Regions that are mutually exclusive for two histone modifications will have probabilities close to 0 or 1, while regions that overlap equally will have probabilities around 0.5. Thus, our strategy profiles single-cell multimodal histone modifications by linking each double-cut cell with their respective single-cut pairs from the training data.

The ability to link a double-cut cell with their respective single-cut pairs provides a framework to transfer labels from one single-scChIC-seq data to another. Transferring labels is useful when studying histone modifications that lack obvious cell type-identifying markers, such as repressive histone modifications. By coupling with active histone modifications, where cell types can readily be inferred by gene activity, we can study histone modifications that are not necessarily correlated with mRNA abundances while retaining our ability to assign cell types (which are traditionally identified from mRNA abundances of marker genes) to the data.

To transfer information, we use the double-cut cells as links from the labeled to unlabeled dataset. First, unmixed cells from double-scChIC-seq are projected onto their respective training datasets (Methods). The cells in labeled dataset transfers its labels to the unmixed cells based on the location of the unmixed cells in the projected space. The unmixed cells then transfer its labels to the cells in the unlabeled dataset, based on the location of the unlabeled cells in the projected space.

Overall, we combine an experimental and computational strategy into a statistically principled framework to multiplex histone modifications in single cells. Our framework allows multimodal analysis of histone modifications in single cells as well as transfer of labels from one single-cell histone modification dataset to another. We envision this strategy to enable systematic analysis of multimodal histone modifications as well as integration of distinct single-cell histone modification datasets.

## Overview of the statistical model

Our goal is to unmix duo-scChIC-seq data from cells stained with two different antibodies into its separate histone modifications. scChIC-seq data can be represented as a sparse count matrix with genomic regions in the rows (features) and individual cells in the columns (samples). A natural way to model scChIC-seq data is to assume the unique cuts in a cell ($\vec{y}$) are generated from a multinomial process. From this framework, each genomic region $k$ has a relative frequency $f_k$ to generate a cut, where $\sum_k f_k = 1$. Each cell generates a total of $\sum_k y_k = N$ cuts.

$$\vec{y} \sim \text{Multinomial}(\vec{f}, N)$$

In a double immunocleavage experiment, we assume that $\vec{f}$ come from a mixture of two multinomials $\vec{\alpha^a}$ and $\vec{\alpha^b}$ (one from each histone modification). For example, the relative frequency for genomic region $k$ can come from mixing $\alpha_k^a$ and $\alpha_k^b$ with mixing fraction $w$:

$$f_k = w\alpha_k^a + (1-w)\alpha_k^b$$

The multinomial log-likelihood of a cell with cuts across the genome $\vec{y}$ mixed from multinomials $\vec{\alpha^a}$ and $\vec{\alpha^b}$ with mixing fraction $w$ is:

$$\text{LL} = \log(P\left(\vec{y}|\alpha^a, \alpha^b, w\right)) \propto \sum_{k=1}^{K} y_k \log\left(w\alpha_k^a + (1-w)\,\alpha_k^b\right)$$

In general, $\alpha^a$ and $\alpha^b$ are cell type specific, we therefore use single-scChIC-seq training data to infer the possible pairs of multinomials. In whole bone marrow, the training data generates about 5-7 cell clusters (see ), which corresponds to 25-49 possible pairs of $\alpha^a$'s and $\alpha^b$'s. We do not know *a priori* which $\alpha^a$ pairs with which $\alpha^b$, but we know the possible pairs. We use the duo-cut cells to infer the most likely pairs. The probability that a duo-cut cell comes from a particular pair $(c, d)$:

$$p_{(c,d)} = \frac{e^{L_{(c,d)}}}{\sum\limits_{(f,g)\in pairs} e^{L_{f,g}}}$$

For a duo-cut cell, we calculate the $p_i$ for each pair of $\vec{\alpha}^a$ and $\vec{\alpha}^b$ and select the pair with the highest probability.

### Generating the possible pairs of genomic profiles

We use Latent Dirichlet Allocation (LDA), a matrix decomposition model that uses the multinomial generative process, to estimate the relative frequencies of each cell in each genomic region. LDA decomposes the count matrix into a cell-to-topics matrix and a topic-to-regions matrix. A topic can be thought of as loadings in the lower-dimensional space. Briefly, LDA can be seen as a hierarchical multinomial model, where a genomic cut in a cell is generated by sampling a topic from a cell-specific multinomial distrbution of topic probabilities, then sampling a genomic region from a topic-specific multinomial distribution of genomic region probabilities.

To generate the $j$th cut for cell $i$:

First choose a topic:

$$\vec{w}_i \sim \text{Dirichlet}(\gamma)$$
$$k_{i,j} \sim \text{Multinomial}(\vec{w}_i)$$

After choosing a topic, choose a genomic region:

$$\vec{p}_k \sim \text{Dirichlet}(\delta)$$
$$k_{i,j} \sim \text{Multinomial}(\vec{p}_{k_{i,j}})$$

The Dirichlet distributions are priors to prevent overfitting in the presence of low number of cuts. We used the LDA model implemented by the topicmodels R package, using the Gibbs sampling implementation with hyperparameters $\gamma = 1.67$, $\delta = 0.1$, where K is the number of topics[48].

We cluster cells into one group by applying Louvain clustering onto the cell-to-topics matrix. Finally, we define for each cluster $c$, a genomic profile used for training as a vector of relative frequencies across genomic regions by averaging the estimated relative frequencies across cells assigned to the cluster $N_c$.

We calculate the relative frequencies in a genomic bin $j$ of histone mark $a$ for cell cluster $c$ by:

$$\alpha^a_{c,j} = \frac{1}{N_c} \sum_{i \in cells} \sum_{k=1}^{K} w_{i,k} p_{k,j}$$

There will be $C$ clusters associated with histone mark $a$ and $D$ clusters associated with the other histone mark $b$, giving two sets of relative frequencies:

$$\vec{\alpha_{1,j}}^a, ..., \vec{\alpha_{C,j}}^a$$

$$\vec{\alpha_{1,j}}^b, ..., \vec{\alpha_{D,j}}^b$$

We can iterate each pair of multinomials $\left(\alpha^a_c, \alpha^b_d\right), c \in C, d \in D$, to calculate a log-likelihood for the double-cleaved cell count vector.

## Fitting mixing fraction $w$ in the log-likelihood

For a vector $\vec{y}$ of double-cleaved cuts across genomic regions, we calculate the log-likelihood of $\vec{y}$ given that it came from mixing a pair $(c, d)$ of two multinomials $\vec{\alpha}^a_c$ and $\vec{\alpha}^b_d$:

$$\mathrm{LL}_{(c,d)} = \log(P\left(\vec{y}|\vec{\alpha}^a, \vec{\alpha}^b, w\right)) \propto \sum_{k=1}^{K} y_k \log\left(w\alpha_k^a + (1-w)\,\alpha_k^b\right)$$

The mixing fraction $w$ is unknown, and we infer $w$ by maximizing the log-likelihood. We use the Brent method implemented in R (*optim*) for minimizing the negative log-likelihood with respect to $w$.

### Assigning double-immunocleaved DNA fragments in single cells back to their histone modification of origin

After having selected the best pair (c,d) with the highest probability $p_{(c,d)}$, we assign each genomic cut $y_{i,j}$ for cell $i$ in genomic region $j$ to histone mark $\alpha^a$ with probability $q_{i,j}$:

$$q_{i,j} = \frac{w\alpha_c^a}{w\alpha_c^a + (1-w)\,\alpha_d^b}$$

Reads are therefore assigned to the other histone mark $\alpha^b$ with probability $1 - q_{i,j}$

The probabilities $\vec{q}$ are used to split each read in the BAM file of the double-incubated datasets into two BAM files of the respective histone modification. When splitting reads at the BAM level, there may be genomic regions that were not considered in the model (e.g., regions with few reads in the training data). Probabilities for reads in a missing region $q'_x$ at position $x$ are imputed by linear interpolation between the nearest left $q_{x_l}$ at $x_l$ and right genomic bin $q_{x_r}$ at $x_r$:

$$q'_x = q_{x_l} + (x - x_l)\frac{q_{x_r} - q_{x_l}}{x_r - x_l}$$

### Inferring topic weights in the LDA model

To infer topic weights in the LDA framework, we use a collapsed Gibbs sampling algorithm implemented in the topicmodels R package[48]. Briefly, the posterior distribution of the LDA model is inferred from the data by sampling from the update equation[16]:

$$p(z_{d,n} = k|\vec{z}_{-d,n}, \vec{w}, \alpha, \delta) \propto \frac{n_{d,k} + \alpha_k}{\sum_i^K n_{d,i} + \alpha_i} \frac{v_{k,w_{d,n}} + \delta_{w_{d,n}}}{\sum_{j\in\{J\}} v_{k,j} + \delta_j}$$

Where:

- $z_{d,n}$ is the topic assignment for cell $d$ used to generate cut $n$. $\vec{z}_{-d,n}$ denotes $z_{d,n}$ with the assignment $z_{d,n}$ removed (i.e., the $n$th cut from cell $d$ that will have its topic assignment updated is removed),
- $n_{d,k}$ is the number of times cell $d$ uses topic $k$ across all genomic regions,
- $v_{k,w_{d,n}}$ is the number of times topic $k$ uses genomic region $w_{d,n}$ across all cells,
- $\vec{w}$ is the vector of cuts in the genome across all cells,
- $\alpha$ is the Dirichlet parameter for the prior distribution of cell to topic,
- $\delta$ is the Dirichlet parameter for the prior distribution of topic to genomic region.

### Projecting deconvolved scChIC-seq data onto a previously-defined embedding

After fitting the LDA model using Gibbs sampling, we can project new cells onto the existing model using the Gibbs sampling equation by updating only the cell to topic parameters while holding the topic to genomic region parameters constant[48].

### Defining cell clusters using LDA topic weights

LDA decomposes the count matrix into two matrices: a cell to topic and a topic to genomic region matrix. The cell to topic matrix assigns each cell to a vector of topic weights indicating a grade of membership of a cell across topics.

For each celltype-specific topic, plotting the distribution of topic weights across cells often reveals a bimodal distribution, suggesting a natural cutoff for assigning a cell to a cluster. We fit a Gaussian mixture model to each distribution to estimate the threshold at which the value of the two fitted Gaussian disributions are equal. We assign cells to clusters if the topic weights for the cell is greater than the threshold.

### Transferring celltype labels from one scChIC-seq to another

To transfer labels from one scChIC-seq data targeting histone modification $a$ to another scChIC-seq data targeting modification $b$, we use the double-incubated scChIC-seq to link together the two single-incubated datasets. After unmixing the double-incubated scChIC-seq data into their respective histone modifications, we project the $a$ part of the double-incubated scChIC-seq data (denoted as $a_{unmixed}$) onto LDA model trained by the scChIC-seq data $a$. We assign a cell from $a_{unmixed}$

to cluster $c$ if the cell to topic weight in $a_{unmixed}$ is greater than a threshold $t$. This threshold $t$ comes from previously fitting a Gaussian mixture model of only cells in $a$ (see Section 1). After having transferred labels from $a$ to the double-incubated cells, we project the $b_{unmixed}$ onto $b$. Similarly, we assign cells in $b_{umixed}$ to a cluster in $b$. Each cluster in $b$ will have a number of cells from $b_{unmixed}$ assigned to it, some possibly with different labels. We assign the cluster in $b$ to a label defined in $b_{unmixed}$ by summing the weights across cells for each label and choosing the label which has the highest summed weight.

We impute the labels for cells that were not assigned a label because the topic to cell weight was not greater than $t$ for any of the possible topics. We use the labels from the 50-nearest neighbors and taking the label with the most cells. The kNN graph is defined from the cell to topic matrix using Euclidean distance.

Chapter 5

# Discussion

This PhD thesis describes recent advances in single-cell sequencing technologies, focusing on measuring multiple modalities within a single cell. The (epi-)genomics methods used here are all modified DNA cutters which are employed to mark the region of interest for later readout. These marked regions can be combined with additional readouts such as, for example, scRNA-seq. Here I first discuss the potential of recent advances in genomics and genome editing in the context of lineage tracing. Secondly, I will discuss challenges arising from sparse (epi-)genomic data. Finally, I will conclude how multimodal profiling in single cells is changing the concept of cell identity and how to adapt these changes in an accessible way.

## Validating results from CRISPR/Cas9 recorders

Recently, barcoding systems using CRISPR/Cas9, such as shown in **Chapter 2** and **3**, have been popular and have gone through many adaptations. Pitfalls of CRISPR/Cas9 recorders, such as barcode drop-outs through simultaneous cleavage (mostly happening at close-by target sites as in GESTALT) or the amount of single cells for plate-based methods such as in ScarTrace, call for validation of (novel) clonal relationships through other methods. Microscopy-based lineage tracing methods, which can track cells over multiple cell divisions as well as providing cellular localization, can provide complementary results to evolving barcodes, such as shown in **Chapter 2**. Results from **Chapter 3** can also be validated with zebrafish embryogenesis microscopy lineage tracing (such as in [57]) or with flourescent reporters. Indeed, recently a study confirmed the results presented in **Chapter 3** that a sub-population of tissue-resident macrophages have been shown to contribute to zebrafish epidermis formation upon fin injury [72]. As ScarTrace and other CRISPR/Cas9 recorders are popular because of the possibility of integration with other -omics technologies, it will be possible to read out clonal information from other modalities alongside scars within the same cells. Other modalities encoding clonal information can be, for example, methylation sites or single nucleotide variants (SNVs). Such additional measurements might not only validate clonal lineage relations within the same embryo, but can also add more resolution if the time of operation is shifted from the scars operational timeframe or if it is measured in cells where scars were not recorded (drop-outs). Additionally, ScarTrace can be combined with imaging based techniques (e.g. *brainbow* mouse) to get multiple sets of information from within the same cell.

## Tweaking CRISPR/Cas9 recorders for lineage tracing

The first generation of CRISPR/Cas9 recorders, with at most 10 CRISPR targets, were primarily developed for multi-omics purposes, i.e. integration with gene expression. Therefore it is expected that tuneable parameters which contribute to

maximising successful accumulation of barcodes or the timeframe of clonal recording are going to be massively improved. However, Cas9 activity is a stochastic molecular process and it remains to be described how well the activity and barcode introduction can be controlled. In **Chapter 2** and **3** we discussed, how Cas9 protein or RNA can be used to limit or enlarge the operational time window of Cas9. Other options of time control for Cas9 are, for example, tagging Cas9 with a degradation flag from the FUCCI system[52] as well as a photoactivable Cas9[93], which has so far only been used in 2D cell culture systems and not *in vivo*. Another possibility for extending the timeframe for barcode introduction is to have two uncoupled Cas versions for targeting different locations such as, for example, Cas12 and Cas9. Cas9 and Cas12 can be induced independently of each other and also activate alternate repair mechanisms which can be explored for lineage tracing. For building refined lineage trees it can be desirable to have a new barcode with every cell division. Having multiple locations to target results on the one hand in more uniquely labeled clones but on the other can induce deadly stress in *in vivo* experiments. Continuous expression of CRISPR/Cas9 has been shown to induce cell death in human cell lines[41] and result in developmental delay in mouse embryos[24]. Recently developed CRISPR/Cas9 recorder systems with Cas9 nickase instead of Cas9 nuclease[129] might be a great alternative to accumulate more frequent barcodes without cellular stress responses. In mouse, the timeframe of barcode introduction has been shown to vary with the gRNA length, tested for 21, 25, 30 and 35bp length[54], which can also be explored for increasing resolution in lineage branching events.
For this first generation of CRISPR/Cas9 recorders Cas9 barcoding simulations have demonstrated that the overall accuracy and lineage resolution is limited[107] and technologies with more target sites or more appropriate sequence context are expected to be developed. Since many studies have revealed that DNA repair through NHEJ is biased through sequence contexts[27,138], barcode prediction tools have become more accessible allowing for simulating the complexity of barcodes introduced through NHEJ[27]. How cells react to induced DSB as well as the NHEJ repair mechanism is species dependent but predictive tools have been developed for zebrafish and mouse.

## Analysis challenges for sparse epigenomic data

DNA measurements from single cells are sparse given the rarity of epigenetic modifications within a huge genomics space, as well as the molecular limitations of barcoding and subsequent measuring of every modification. Single cell studies of histone modification profiles from DNA cutters, such as micrococcal nuclease used in **chapter 4**, have shown that is is difficult to decipher single cell heterogeneity from technical artefacts[58]. Technical artefacts can arise through, for example, random DNA fragmentation or antibody efficiency. Within such sparse datasets it is

difficult to calculate biological meaningful distances between cells, hence it is difficult to cluster single cells. This becomes even more of a problem if one wants, for example, to calculate single cell distances on individual epigenetic modified sites. In **Chapter 4** however, we show that we can measure two histone modifications from the same cell. With every additional measured DNA modality we increase the almost empty space by another dimension. We work through this problem by binning sites into larger genomic regions as well as employing an unsupervised machine learning algorithm, called Latent Dirichlet Allocation (LDA), which allows for grouping cells with similar signal in similar genomic regions. LDA has been shown to not only be applicable for scChIC-seq data (Zeller et al., *in preparation*), but also for accessibility data from scATAC-seq[100]. While LDA is showing great results with sparse datasets and helps in identifying cellular heterogeneity the underlying *topics* which define groups of genomic bins can of course be topics based on artefacts. These *topics* still need, therefore, a manual validation.

## Multi-omics in embryonic model systems easy to perturb

Cell fate decisions as discussed in **chapter 2** and **3** are crucial during embryonic development. As it is of great importance to better understand how cell fate decisions are made at multiple regulatory levels it is crucial to describe embryonic *in vivo* systems, but also to perturb embryonic systems. Here we demonstrated how novel technologies can be used to study embryonic development in wild type *in vivo* conditions. Making large-scale perturbations and genetic mutants is however cumbersome, no matter if in zebrafish or mouse models. In recent years *in vitro* systems of both, mouse[8] and zebrafish[108,134] have been established and named *gastruloids* and *pescoids*, respectively. These model systems show axis formation, *Hox* genes dependent patterning and somite formation[137]. However, while *gastruloids* develop tissues from all three germ layers, they do not form brain or extra-embryonic tissues. While it has to be kept in mind that these *embryoid* systems are limited to model only certain differentiation processes ocurring in embryos, they provide ideal conditions for large-scale perturbation and screens. Large-scale perturbations such as knock-outs are especially interesting for multi-omics technologies to draw direct connections between the measured molecular layers. *Embryoid* systems have clearly defined cell types[137] and a reproducible developmental window which makes them a great starting point for multi-omics studies and a feasible study subject of the goal of understandin cell fate decisions and underlying molecular regulatory networks. These *embryoid* systems can be for example combined with microscopy and ScarTrace studies for studying lineages in a background of, for example, transcription factors or histone modifier knockouts.

## Evolving concept of a cell type through multi-omics

While currently large efforts are put into building successful cellular atlases from transcriptomics data, more and more datatypes will be integrated to draw a complete multi-layered picture of different cellular states. This is pushing the boundaries of what the term *cell type* means, which is now widely associated to marker genes from surface markers, stainings or scRNA-seq data. A *cellular state* encompasses, beyond transcriptional profile, multiple cellular measurements and places a cell in the context of past, present and future [88,144]. The past of a certain cell type can be viewed through evolutionary phylogeny [6] or embryonic development, which are not necessarily the same and might obscure cell type comparisons across organisms. In this thesis we will only explore cellular past in reference to embryonic development. However, for a complete picture of cellular states across multiple organisms it is essential to take into account conserved regulatory complexes giving rise to homologous cell types [6]. For this it has been proposed that independent cell types should be distinguished based on at least one different transcription factor in conjunction with its corresponding protein-protein interactions [6]. The *present* cell type information would ideally integrate multiple omics measurements for defining a certain genomic state (e.g. chromatin, 3D organisation) along with its gene expression to draw a cell type specific phenotype and additionally contain also functional information. For the future of a cell, it is important to measure the spectrum of molecular or morphological phenotypes upon environmental stimulation (e.g. T-cell activation). RNA measurements, for example, quickly change in response to stimuli, while some epigenetic signatures remain as a more stable cell identity defining measure. Here transcription factors from scRNA-seq would be the cell type defining measure which would decipher a cell state (continuous due to environment, stimulation, activation) from a cell type. Most concrete and data-driven implementations can be found in the Cell ontology system [32,95], which is incorporating images, classical phenotypic cell types, genomic as well as gene expression profiles in a similar system as Gene Ontology. The most complete branch is the well studied hematopoietic branch. A considerable effort would be the incorporation of not only naturally occurring cell types, but also related cell lines and experimentally modified cells.

In terms of visualisation it is difficult to incorporate some key aspects, such as lineage and the wide variety of phenotypic states of the same cell type. Lineage information of a cell is related to phylogeny and development. These two definitions are not always compatible hierarchies within a cell type. Visualising the dynamic landscape, e.g. gene expression, within a cell type which is not reflected on the chromatin level also remains a data handling challenge. One could, for example, imagine a graph-based approach for a cell atlas (Figure 5.1), where the core cell identity is

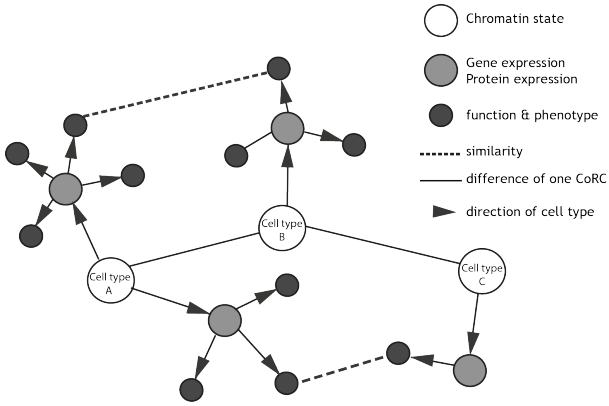



Figure 5.1: Schematics of potential graph-based single cell atlas with multi-layered information.

focused on single cell genomic studies such as chromatin, transcription factors and genome 3D structure. Small differences in the genome build-up would result in a different cell type. Connected to the cell identity cores are directed edges showing a variety of cellular states represented by, for example, scRNA-seq measurements. The last layer consists of phenotypic and functional measurements which can be connected to other cell types. Such a graph-based cellular state atlas additionally takes into account the linear flow of information in the cell, DNA $>$ RNA $>$ Protein. Practically, the information flow in a cell is of course circular, as proteins function on DNA. However, when building a cellular atlas with multiple modalities and the goal to identify and compare cell types, a hierarchy is needed.

The biggest challenges seem to come from the fact that cells are dynamic and continuous units responding to developmental and environmental changes while sequencing approaches provide discrete snapshots along the way. Even with 'big data' it is unlikely we will ever cover all cellular dynamics to a desirable detail (ranging from chromatin state to RNA structure to protein modifications and localisation) we might start to rely on molecular predictions and builds of virtual cell states. Functional assays and multi-omics data could be integrated in cellular simulations and missing data gaps could be predicted. These simulations could help to show nuanced *facettes* of cells and it might help to picture a refined cellular identity within its past, present and future.

# References


[1] A. Abzhanov. von Baer's law for the ages: lost and found principles of developmental evolution. *Trends in Genetics*, 29:712–722, 2013.

[2] A. A. Akerberg, S. Stewart, and K. Stankunas. Spatial and Temporal Control of Transgene Expression in Zebrafish. *PloS one*, 9, 2014.

[3] A. Alemany, M. Florescu, C. S. Baron, J. Peterson-Maduro, and A. van Oudenaarden. Whole-organism clone tracing using single-cell sequencing. *Nature*, 556:108–112, 2018.

[4] A. Alemany, M. Florescu, C. S. Baron, J. Peterson-Maduro, and A. van Oudenaarden. Single-cell ScarTrace. *Protocol Exchange*, pages 1–7, 2018.

[5] F. Amat, W. Lemon, D. P. Mossing, K. McDole, Y. Wan, K. Branson, E. W. Myers, and P. J. Keller. Fast, accurate reconstruction of cell lineages from large-scale fluorescence microscopy data. *Nature Methods*, 11:951–958, 2014.

[6] D. Arendt, J. M. Musser, C. V. Baker, A. Bergman, C. Cepko, D. H. Erwin, M. Pavlicev, G. Schlosser, S. Widder, M. D. Laubichler, and G. P. Wagner. The origin and evolution of cell types. *Nature Reviews Genetics*, 17:744–757, 2016.

[7] E. I. Athanasiadis, J. G. Botthof, H. Andres, L. Ferreira, P. Lio, and A. Cvejic. Single-cell RNA-Sequencing uncovers transcriptional states and fate decisions in haematopoiesis. *bioRxiv*, 2017.

[8] L. Beccari, N. Moris, M. Girgin, D. A. Turner, P. Baillie-Johnson, A.-C. Cossy, M. P. Lutolf, D. Duboule, and A. M. Arias. Multi-axial self-organization properties of mouse embryonic stem cells into gastruloids. *Nature*, 562:272–276, 2018.

[9] J. S. Becker, D. Nicetto, and K. S. Zaret. H3k9me3-dependent heterochromatin: barrier to cell fate changes. *Trends in Genetics*, 32:29–41, 2016.

[10] S. Behjati, M. Huch, R. van Boxtel, W. Karthaus, D. C. Wedge, A. U. Tamuri, I. Martincorena, M. Petljak, L. B. Alexandrov, G. Gundem, P. S. Tarpey, S. Roerink, J. Blokker, M. Maddison, L. Mudie, B. Robinson, S. Nik-Zainal, P. Campbell, N. Goldman, M. van de Wetering, E. Cuppen, H. Clevers, and M. R. Stratton. Genome sequencing of normal cells reveals developmental lineages and mutational processes. *Nature*, 513:422–425, 2014.

[11] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3:993–1022, 2003.

[12] V. Boeva, A. Zinovyev, K. Bleakley, J.-P. Vert, I. Janoueix-Lerosey, O. Delattre, and E. Barillot. Control-free calling of copy number alterations in deep-sequencing data using GC-content normalization. *Bioinformatics*, 27:268–269, 2010.

[13] A. Bolotin, B. Quinquis, A. Sorokin, and S. Dusko Ehrlich. Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. *Microbiology*, 151:2551–2561, 2005.

[14] A. Bortnick and C. Murre. Cellular and chromatin dynamics of antibody-secreting plasma cells. *Wiley Interdisciplinary Reviews: Developmental Biology*, 5:136–149, 2016.

[15] C. M. Bouldin, C. D. Snelson, G. H. Farr 3rd, and D. Kimelman. Restricted expression of cdc25a in the tailbud is essential for formation of the zebrafish posterior body. *Genes & development*, 28:384–395, 2014.

[16] J. Boyd-Graber, Y. Hu, D. Mimno, et al. Applications of topic models. *Foundations and Trends in Information Retrieval*, 11:143–296, 2017.

[17] J. M. Brown and K. G. Storey. A region of the vertebrate neural plate in which neighbouring cells can adopt neural or epidermal fates. *Current Biology*, 10: 869–872, 2000.

[18] J. D. Cahoy, B. Emery, A. Kaushal, L. C. Foo, J. L. Zamanian, K. S. Christopherson, Y. Xing, J. L. Lubischer, P. A. Krieg, S. A. Krupenko, W. J. Thompson, and B. A. Barres. A Transcriptome Database for Astrocytes, Neurons, and Oligodendrocytes: A New Resource for Understanding Brain Development and Function. *The Journal of Neuroscience*, 28:264–278, 2008.

[19] M. Cajal, K. A. Lawson, B. Hill, A. Moreau, J. Rao, A. Ross, J. Collignon, and A. Camus. Clonal and molecular analysis of the prospective anterior neural boundary in the mouse embryo. *Development*, 139:423–436, 2012.

[20] J. Cao, J. S. Packer, V. Ramani, D. A. Cusanovich, C. Huynh, R. Daza, X. Qiu, C. Lee, S. N. Furlan, F. J. Steemers, A. Adey, R. H. Waterston, C. Trapnell, and J. Shendure. Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science*, 357:661–667, 2017.

[21] J. Cao, M. Spielmann, X. Qiu, X. Huang, D. M. Ibrahim, A. J. Hill, F. Zhang, S. Mundlos, L. Christiansen, F. J. Steemers, C. Trapnell, and J. Shendure. The single-cell transcriptional landscape of mammalian organogenesis. *Nature*, 566:496–502, 2019.

[22] R. Cao, L. Wang, H. Wang, L. Xia, H. Erdjument-Bromage, P. Tempst, R. S. Jones, and Y. Zhang. Role of histone h3 lysine 27 methylation in polycomb-group silencing. *Science*, 298:1039–1043, 2002.

[23] S. J. Carmona, S. A. Teichmann, L. Ferreira, I. C. Macaulay, M. J. T. Stubbington, A. Cvejic, and D. Gfeller. Single-cell transcriptome analysis of fish immune cells provides insight into the evolution of vertebrate immune cell types. *Genome Research*, 27:451–461, 2017.

[24] M. M. Chan, Z. D. Smith, S. Grosswendt, H. Kretzmer, T. M. Norman, B. Adamson, M. Jost, J. J. Quinn, D. Yang, M. G. Jones, A. Khodaverdian, N. Yosef, A. Meissner, and J. S. Weissman. Molecular recording of mammalian embryogenesis. *Nature*, 570:77–82, 2019.

[25] T. Chandra, K. Kirschner, J.-Y. Thuret, B. D. Pope, T. Ryba, S. Newman, K. Ahmed, S. A. Samarajiwa, R. Salama, T. Carroll, et al. Independence of repressive histone marks and chromatin compaction during senescent heterochromatic layer formation. *Molecular cell*, 47:203–214, 2012.

[26] R. Chen, X. Wu, L. Jiang, and Y. Zhang. Single-Cell RNA-Seq Reveals Hypothalamic Cell Diversity. *Cell Reports*, 18:3227–3241, 2017.

[27] W. Chen, A. McKenna, J. Schreiber, M. Haeussler, Y. Yin, V. Agarwal, W. S. Noble, and J. Shendure. Massively parallel profiling and predictive modeling of the outcomes of CRISPR/Cas9-mediated double-strand break repair. *Nucleic Acids Research*, 47:7989–8003, 2019.

[28] C. Cheng and M. Gerstein. Modeling the relative relationship of transcription factor binding and histone modifications to gene expression levels in mouse embryonic stem cells. *Nucleic acids research*, 40:553–568, 2012.

[29] S. J. Clark, R. Argelaguet, C. A. Kapourani, T. M. Stubbs, H. J. Lee, C. Alda-Catalinas, F. Krueger, G. Sanguinetti, G. Kelsey, J. C. Marioni, O. Stegle, and W. Reik. ScNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells e. *Nature Communications*, 9:1–9, 2018.

[30] S. S. Dey, L. Kester, B. Spanjaard, M. Bienko, and A. van Oudenaarden. Integrated genome and transcriptome sequencing of the same cell. *Nature Biotechnology*, 33:285–289, 2015.

[31] V. Di Donato, T. O. Auer, K. Duroure, and F. Del Bene. Characterization of the Calcium Binding Protein Family in Zebrafish. *PloS one*, 8, 2013.

[32] A. D. Diehl, T. F. Meehan, Y. M. Bradford, M. H. Brush, W. M. Dahdul, D. S. Dougall, Y. He, D. Osumi-Sutherland, A. Ruttenberg, S. Sarntivijai, C. E. Van Slyke, N. A. Vasilevsky, M. A. Haendel, J. A. Blake, and C. J. Mungall. The cell ontology 2016: Enhanced content, modularization, and ontology interoperability. *Journal of Biomedical Semantics*, 7, 2016.

[33] D. Duboule. Temporal colinearity and the phylotypic progression: A basis for the stability of a vertebrate Bauplan and the evolution of morphologies through heterochrony. *Development*, 120:135–142, 1994.

[34] J. A. Farrell, Y. Wang, S. J. Riesenfeld, K. Shekhar, A. Regev, and A. F. Schier. Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis. *Science*, 3131, 2018.

[35] K. L. Frieda, J. M. Linton, S. Hormoz, J. Choi, K.-h. K. Chow, and Z. S. Singer. Synthetic recording and in situ readout of lineage information in single cells. *Nature Publishing Group*, 541:107–111, 2017.

[36] E. L. Gautier, T. Shay, J. Miller, M. Greter, C. Jakubzick, S. Ivanov, J. Helft, A. Chow, K. G. Elpek, S. Gordonov, A. R. Mazloom, A. Ma'ayan, W.-J. Chua, T. H. Hansen, S. J. Turley, M. Merad, G. J. Randolph, E. L. Gautier, C. Jakubzick, G. J. Randolph, A. J. Best, J. Knell, A. Goldrath, J. Miller, B. Brown, M. Merad, V. Jojic, D. Koller, N. Cohen, P. Brennan, M. Brenner, T. Shay, A. Regev, A. Fletcher, K. Elpek, A. Bellemare-Pelletier, D. Malhotra, S. Turley, R. Jianu, D. Laidlaw, J. Collins, K. Narayan, K. Sylvia, J. Kang, R. Gazit, B. S. Garrison, D. J. Rossi, F. Kim, T. N. Rao, A. Wagers, S. A. Shinton, R. R. Hardy, P. Monach, N. A. Bezman, J. C. Sun, C. C. Kim, L. L. Lanier, T. Heng, T. Kreslavsky, M. Painter, J. Ericson, S. Davis, D. Mathis, C. Benoist, and t. I. G. Consortium. Gene-expression profiles and transcriptional regulatory pathways that underlie the identity and diversity of mouse tissue macrophages. *Nature Immunology*, 13:1118–1128, 2012.

[37] A. Giladi and I. Amit. Single-Cell Genomics: A Stepping Stone for Future Immunology Discoveries. *Cell*, 172:14–21, 2018.

[38] D. Grün, A. Lyubimova, L. Kester, K. Wiebrands, O. Basak, N. Sasaki, H. Clevers, and A. van Oudenaarden. Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature*, 525:251–255, 2015.

[39] D. Grün, M. J. Muraro, J.-C. Boisset, K. Wiebrands, A. Lyubimova, G. Dharmadhikari, M. van den Born, J. Van Es, E. Jansen, H. Clevers, et al. De novo prediction of stem cell identity using single-cell transcriptome data. *Cell stem cell*, 19:266–277, 2016.

[40] A. Guernet, S. Mungamuri, D. Cartier, R. Sachidanandam, A. Jayaprakash, S. Adriouch, M. Vezain, F. Charbonnier, G. Rohkin, S. Coutant, S. Yao, H. Ainani, D. Alexandre, I. Tournier, O. Boyer, S. Aaronson, Y. Anouar, and L. Grumolato. CRISPR-Barcoding for Intratumor Genetic Heterogeneity Modeling and Functional Analysis of Oncogenic Driver Mutations. *Molecular Cell*, 63:526–538, 2016.

[41] E. Haapaniemi, S. Botla, J. Persson, B. Schmierer, and J. Taipale. CRISPR-Cas9 genome editing induces a p53-mediated DNA damage response. *Nature Medicine*, 24:927–930, 2018.

[42] T. Hashimshony, F. Wagner, N. Sher, and I. Yanai. CEL-Seq: Single-Cell RNA-Seq by Multiplexed Linear Amplification. *Cell Rep*, 2:666–673, 2012.

[43] T. Hashimshony, N. Senderovich, G. Avital, A. Klochendler, Y. de Leeuw, L. Anavy, D. Gennert, S. Li, K. J. Livak, O. Rozenblatt-Rosen, Y. Dor, A. Regev, and I. Yanai. CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. *Genome biology*, 17:77, 2016.

[44] J. Henninger, B. Santoso, S. Hans, E. Durand, J. Moore, C. Mosimann, M. Brand, D. Traver, and L. Zon. Clonal fate mapping quantifies the number of haematopoietic stem cells that arise during development. *Nature Cell Biology*, 19:17–27, 2017.

[45] D. Henrique, E. Abranches, L. Verrier, and K. G. Storey. Neuromesodermal progenitors and the making of the spinal cord. *Development*, 142:2864–2875, 2015.

[46] S. E. Hickman, N. D. Kingery, T. K. Ohsumi, M. L. Borowsky, L.-c. Wang, T. K. Means, and J. El Khoury. The microglial sensome revealed by direct RNA sequencing. *Nature neuroscience*, 16:1896–1905, 2013.

[47] E. Hirsinger and B. Steventon. A Versatile Mounting Method for Long Term Imaging of Zebrafish Development. *JoVE*, 119, 2017.

[48] K. Hornik and B. Grün. topicmodels: An r package for fitting topic models. *Journal of statistical software*, 40:1–30, 2011.

[49] L.-E. Jao, S. R. Wente, and W. Chen. Efficient multiplex biallelic zebrafish genome editing using a CRISPR nuclease system. *Proceedings of the National Academy of Sciences*, 110:13904–13909, 2013.

[50] M. Jinek, K. Chylinski, I. Fonfara, M. Hauer, J. A. Doudna, and E. Charpentier. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science*, 337:816–821, 2012.

[51] L. Jing and L. I. Zon. Zebrafish as a model for normal and malignant hematopoiesis. *Disease Models & Mechanisms*, 4:433–438, 2011.

[52] J. P. Junker, B. Spanjaard, J. Peterson-Maduro, A. Alemany, B. Hu, M. Florescu, and A. van Oudenaarden. Massively parallel whole-organism lineage tracing using CRISPR / Cas9 induced genetic scars. *bioRxiv*, 2016.

[53] R. Kalhor and G. M. Church. Rapidly evolving homing CRISPR barcodes. *Nature Methods*, 14:1–9, 2016.

[54] R. Kalhor, K. Kalhor, L. Mejia, K. Leeper, A. Graveline, P. Mali, and G. M. Church. Developmental barcoding of whole mouse via homing CRISPR. *Science*, 361(6405), 2018.

[55] J. P. Kanki and R. K. Ho. The development of the posterior body in zebrafish. *Development*, 124:881–893, 1997.

[56] H. S. Kaya-Okur, S. J. Wu, C. A. Codomo, E. S. Pledger, T. D. Bryson, J. G. Henikoff, K. Ahmad, and S. Henikoff. CUT&Tag for efficient epigenomic profiling of small samples and single cells. *Nature Communications*, 10:1–10, 2019.

[57] P. J. Keller, A. D. Schmidt, J. Wittbrodt, and E. H. K. Stelzer. Reconstruction of Zebrafish Early Light Sheet Microscopy. *Science*, 322:1065–1069, 2008.

[58] G. Kelsey. the Past and Predicting the Future. *Biochemistry and Molecular Biology Education*, 75:69–75, 2017.

[59] C. Kiecker and C. Niehrs. A morphogen gradient of Wnt/$\beta$-catenin signalling regulates anteroposterior neural patterning in *Xenopus*. *Development*, 128: 4189–4201, 2001.

[60] A. M. Klein and B. Treutlein. Single cell analyses of development in the modern era. *Development*, 146:1–3, 2019.

[61] F. Knopf, C. Hammond, A. Chekuru, T. Kurth, S. Hans, C. Weber, G. Mahatma, S. Fisher, M. Brand, S. Schulte-Merker, and G. Weidinger. Bone Regenerates via Dedifferentiation of Osteoblasts in the Zebrafish Fin. *Developmental Cell*, 20:713–724, 2011.

[62] I. Kobayashi, H. Ono, T. Moritomo, K. Kano, T. Nakanishi, and T. Suda. Comparative gene expression analysis of zebrafish and mammals identifies common regulators in hematopoietic stem cells. *Blood*, 115:1–9, 2010.

[63] K. Kretzschmar and F. M. Watt. Lineage tracing. *Cell*, 148:33–45, 2012.

[64] W. L. Ku, K. Nakamura, W. Gao, K. Cui, G. Hu, Q. Tang, B. Ni, and K. Zhao. Single-cell chromatin immunocleavage sequencing (scChIC-seq) to profile histone modification. *Nature Methods*, 16:323–325, 2019.

[65] A. Kundaje, W. Meuleman, J. Ernst, M. Bilenky, A. Yen, A. Heravi-Moussavi, P. Kheradpour, Z. Zhang, J. Wang, M. J. Ziller, et al. Integrative analysis of 111 reference human epigenomes. *Nature*, 518:317–330, 2015.

[66] G. La Manno, D. Gyllborg, S. Codeluppi, K. Nishimura, C. Salto, A. Zeisel, L. E. Borm, S. R. W. Stott, E. M. Toledo, J. C. Villaescusa, P. Lönnerberg, J. Ryge, R. A. Barker, E. Arenas, and S. Linnarsson. Molecular Diversity of Midbrain Development in Mouse, Human, and Stem Cells. *Cell*, 167:566–580, 2016.

[67] G. La Manno, R. Soldatov, A. Zeisel, E. Braun, H. Hochgerner, V. Petukhov, K. Lidschreiber, M. E. Kastriti, P. Lönnerberg, A. Furlan, J. Fan, L. E. Borm, Z. Liu, D. van Bruggen, J. Guo, X. He, R. Barker, E. Sundström, G. Castelo-Branco, P. Cramer, I. Adameyko, S. Linnarsson, and P. V. Kharchenko. RNA velocity of single cells. *Nature*, 560:494–498, 2018.

[68] S. G. Landt, G. K. Marinov, A. Kundaje, P. Kheradpour, F. Pauli, S. Batzoglou, B. E. Bernstein, P. Bickel, J. B. Brown, P. Cayting, et al. Chip-seq guidelines and practices of the encode and modencode consortia. *Genome research*, 22: 1813–1831, 2012.

[69] D. Lara-Astiaso, A. Weiner, E. Lorenzo-Vivas, I. Zaretsky, D. A. Jaitin, E. David, H. Keren-Shaul, A. Mildner, D. Winter, S. Jung, et al. Chromatin state dynamics during blood formation. *science*, 345:943–949, 2014.

[70] J. J. Lee and M. P. McGarry. When is a mouse basophil not a basophil? *Blood*, 109:859–861, 2007.

[71] R. T. H. Lee, E. W. Knapik, J. P. Thiery, and T. J. Carney. An exclusively mesodermal origin of fin mesenchyme demonstrates that zebrafish trunk neural crest does not generate ectomesenchyme. *Development*, 140:2923–2932, 2013.

[72] X. Lin, Q. Zhou, C. Zhao, G. Lin, J. Xu, and Z. Wen. An Ectoderm-Derived Myeloid-like Cell Population Functions as Antigen Transporters for Langerhans Cells in Zebrafish Epidermis. *Developmental Cell*, 49:605–617.e5, 2019.

[73] J. Livet, T. A. Weissman, H. Kang, R. W. Draft, J. Lu, R. A. Bennis, J. R. Sanes, and J. W. Lichtman. Transgenic strategies for combinatorial expression of fluorescent proteins in the nervous system. *Nature*, 450:56–62, 2007.

[74] M. A. Lopez-Ramirez, C.-F. Calvo, E. Ristori, J.-L. Thomas, and S. Nicoli. Isolation and Culture of Adult Zebrafish Brain-derived Neurospheres. *JoVE*, 2016.

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

[76] I. C. Macaulay, V. Svensson, C. Labalette, L. Ferreira, F. Hamey, T. Voet, S. A. Teichmann, and A. Cvejic. Single-Cell RNA-Sequencing Reveals a Continuous Spectrum of Differentiation in Hematopoietic Cells. *Cell Reports*, 14: 966–977, 2016.

[77] P. J. Marie. Transcription factors controlling osteoblastogenesis. *Archives of Biochemistry and Biophysics*, 473:98–105, 2008.

[78] S. Marques, A. Zeisel, S. Codeluppi, D. V. Bruggen, A. M. Falcão, L. Xiao, H. Li, M. Häring, H. Hochgerner, R. A. Romanov, D. Gyllborg, A. B. Muñozmanchado, G. L. Manno, P. Lönnerberg, E. M. Floriddia, F. Rezayee, P. Ernfors, E. Arenas, J. Hjerling-leffler, T. Harkany, W. D. Richardson, S. Linnarsson, and G. Castelo-branco. Oligodendrocyte heterogeneity in the mouse juvenile and adult central nervous system. *Science*, 352, 2016.

[79] B. L. Martin and D. Kimelman. Wnt Signaling and the Evolution of Embryonic Posterior Development. *Current Biology*, 19:215–219, 2009.

[80] B. L. Martin and D. Kimelman. Canonical Wnt Signaling Dynamically Controls Multiple Stem Cell Fate Decisions during Vertebrate Body Formation. *Developmental Cell*, 22:223–232, 2012.

[81] A. Martinez-Arias and B. Steventon. On the Nature and Organization of organizers. *Development*, 145, 2018.

[82] K. McDole, L. Guignard, F. Amat, A. Berger, G. Malandain, L. A. Royer, S. C. Turaga, K. Branson, and P. J. Keller. *In Toto* Imaging and Reconstruction of Post-Implantation Mouse Development at the Single-Cell Level. *Cell*, 175:859–876, 2018.

[83] A. McKenna and J. A. Gagnon. Recording development with single cell dynamic lineage tracing. *Development*, 146:1–10, 2019.

[84] A. McKenna, G. Findlay, J. a. Gagnon, M. Horwitz, A. F. F. Schier, and J. Shendure. Whole organism lineage tracing by combinatorial and cumulative genome editing. *science*, 7907, 2016.

[85] F. J. Mojica, C. Díez-Villaseñor, J. García-Martínez, and E. Soria. Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. *Journal of Molecular Evolution*, 60:174–182, 2005.

[86] D. Mooijman, S. S. Dey, J. C. Boisset, N. Crosetto, and A. Van Oudenaarden. Single-cell 5hmC sequencing reveals chromosome-wide cell-to-cell variability and enables lineage reconstruction. *Nature Biotechnology*, 34:852–856, 2016.

[87] F. E. Moore, E. G. Garcia, R. Lobbardi, E. Jain, Q. Tang, J. C. Moore, M. Cortes, A. Molodtsov, M. Kasheta, C. C. Luo, A. J. Garcia, R. Mylvaganam, J. A. Yoder, J. S. Blackburn, R. I. Sadreyev, C. J. Ceol, T. E. North, and D. M. Langenau. Single-cell transcriptional analysis of normal, aberrant, and malignant hematopoiesis in zebrafish. *The Journal of experimental medicine*, 213:979–992, 2016.

[88] S. A. Morris. The evolving concept of cell identity in the single cell era. *Development*, 146, 2019.

[89] M. J. Muraro, G. Dharmadhikari, D. Grün, N. Groen, T. Dielen, E. Jansen, L. van Gurp, M. A. Engelse, F. Carlotti, E. J. de Koning, and A. van Oudenaarden. A Single-Cell Transcriptome Atlas of the Human Pancreas. *Cell Systems*, 3:385–394, 2016.

[90] S. H. Naik, L. Perié, E. Swart, C. Gerlach, N. van Rooij, R. J. de Boer, and T. N. Schumacher. Diverse and heritable lineage imprinting of early haematopoietic progenitors. *Nature*, 496:229–232, 2013.

[91] S. M. Nelson, R. A. Frey, S. L. Wardwell, and D. L. Stenkamp. The developmental sequence of gene expression within the rod photoreceptor lineage in embryonic zebrafish. *Developmental dynamics : an official publication of the American Association of Anatomists*, 237:2903–2917, 2008.

[92] P. D. Nieuwkoop and G. V. Nigtevecht. Neural Activation and Transformation in Explants of Competent Ectoderm under the influence of Fragments of Anterior Notochord in Urodeles. *Development*, 1954.

[93] Y. Nihongaki, F. Kawano, T. Nakajima, and M. Sato. Photoactivatable CRISPR-Cas9 for optogenetic genome editing. *Nature Biotechnology*, 33, 2015.

[94] N. Oosterhof, I. R. Holtman, L. E. Kuil, H. C. van der Linde, E. W. G. M. Boddeke, B. J. L. Eggen, and T. J. van Ham. Identification of a conserved and acute neurodegeneration-specific microglial transcriptome in the zebrafish. *Glia*, 65:138–149, 2017.

[95] D. Osumi-Sutherland. Cell ontology in an age of data-driven cell classification. *BMC Bioinformatics*, 18, 2017.

[96] S. Pauls, B. Geldmacher-Voss, and J. A. Campos-Ortega. A zebrafish histone variant H2A.F/Z and a transgenic H2A.F/Z:GFP fusion protein for in vivo studies of embryonic development. *Development Genes and Evolution*, 211:603–610, 2001.

[97] W. Pei, T. B. Feyerabend, J. Rössler, X. Wang, D. Postrach, K. Busch, I. Rode, K. Klapproth, N. Dietlein, C. Quedenau, W. Chen, S. Sauer, S. Wolf, T. Höfer, and H.-r. Rodewald. Polylox barcoding reveals haematopoietic stem cell fates realized in vivo. *Nature*, pages 1–19, 2017.

[98] S. Picelli, O. R. Faridani, A. K. Björklund, G. Winberg, S. Sagasser, and R. Sandberg. Full-length RNA-seq from single cells using Smart-seq2. *Nat Protoc*, 9:171–181, 2014.

[99] B. Pijuan-Sala, J. A. Griffiths, C. Guibentif, T. W. Hiscock, W. Jawaid, F. J. Calero-Nieto, C. Mulas, X. Ibarra-Soria, R. C. Tyser, D. L. L. Ho, W. Reik, S. Srinivas, B. D. Simons, J. Nichols, J. C. Marioni, and B. Göttgens. A single-cell molecular map of mouse gastrulation and early organogenesis. *Nature*, 566:490–495, 2019.

[100] G. B. Pijuan-Sala B., Wilson N.K., Xia J., Hou X., Hannah R.L., Kinston S., Calero-Nieto F.J., Poirion O., Preissl S., Liu F. Single-Cell Chromatin Accessibility Maps Reveal Regulatory Programs Driving Early Mouse Organogenesis. *Nature Cell Biology*, 2020.

[101] K. D. Poss, J. Shen, A. Nechiporuk, G. McMahon, B. Thisse, C. Thisse, and M. T. Keating. Roles for Fgf Signaling during Zebrafish Fin Regeneration. *Developmental Biology*, 222:347–358, 2000.

[102] C. Pourcel, G. Salvignol, and G. Vergnaud. CRISPR elements in Yersinia pestis acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies. *Microbiology*, 151:653–663, 2005.

[103] B. Raj, D. E. Wagner, A. McKenna, S. Pandey, A. M. Klein, J. Shendure, J. A. Gagnon, and A. F. Schier. Simultaneous single-cell profiling of lineages and cell types in the vertebrate brain. *Nature Biotechnology*, 36:442–450, 2018.

[104] Y. Reizel, N. Chapal-Ilani, R. Adar, S. Itzkovitz, J. Elbaz, Y. E. Maruvka, E. Segev, L. I. Shlush, N. Dekel, and E. Shapiro. Colon Stem Cell and Crypt Dynamics Exposed by Cell Lineage Reconstruction. *PLOS Genetics*, 7, 2011.

[105] R. H. Row, S. R. Tsotras, H. Goto, and B. L. Martin. The zebrafish tailbud contains two independent populations of midline progenitor cells that maintain long-term germ layer plasticity and differentiate in response to local signaling cues. *Development*, 143:244–254, 2016.

[106] T. Ryu, B. Spatola, L. Delabaere, K. Bowlin, H. Hopp, R. Kunitake, G. H. Karpen, and I. Chiolo. Heterochromatic breaks move to the nuclear periphery to continue recombinational repair. *Nature Cell Biology*, 17:1401–1411, 2015.

[107] I. Salvador-Martínez, M. Grillo, M. Averof, and M. J. Telford. Is it possible to reconstruct an accurate cell lineage using CRISPR recorders? *eLife*, 8, 2019.

[108] A. Schauer, D. Pinheiro, R. Hauschild, and C.-P. Heisenberg. Zebrafish embryonic explants undergo genetically encoded self-assembly. *eLife*, 9, 2020.

[109] M. Schmid, T. Durussel, and U. K. Laemmli. Chic and chec: genomic mapping of chromatin proteins. *Molecular cell*, 16:147–157, 2004.

[110] R. Schmidt, U. Strähle, and S. Scholpp. Neurogenesis in zebrafish – from embryo to adult. *Neural Development*, 8, 2013.

[111] S. T. Schmidt, S. M. Zimmerman, J. Wang, S. K. Kim, and S. R. Quake. Quantitative Analysis of Synthetic Cell Lineage Tracing Using Nuclease Barcoding. *ACS Synth. Bio*, 2017.

[112] D. W. Scott. On optimal and data-based histograms. *Biometrika*, 66, 1979.

[113] M. A. Selleck and C. D. Stern. Fate mapping and cell lineage analysis of Hensenś node in the chick embryo. *Development*, 112:615–626, 1991.

[114] G. Shah, K. Thierbach, B. Schmid, A. Reade, I. Roeder, N. Scherf, and J. Huisken. Pan-embryo cell dynamics of germlayer formation in zebrafish. *bioRxiv*, 2017.

[115] A. Shilatifard. Molecular implementation and physiological roles for histone H3 lysine 4 (H3K4) methylation. *Current Opinion in Cell Biology*, 20:341–348, 2008.

[116] S. Singh, J. Holdway, and K. Poss. Regeneration of Amputated Zebrafish Fin Rays from De Novo Osteoblasts. *Developmental Cell*, 22:879–886, 2012.

[117] P. J. Skene, J. G. Henikoff, and S. Henikoff. Targeted in situ genome-wide profiling with high efficiency for low cell numbers. *Nature Protocols*, 13:1006–1019, 2018.

[118] I. Smyth, X. Du, M. S. Taylor, M. J. Justice, B. Beutler, and I. J. Jackson. The extracellular matrix gene Frem1 is essential for the normal adhesion of the embryonic epidermis. *Proceedings of the National Academy of Sciences of the United States of America*, 101:13560–13565, 2004.

[119] B. Spanjaard, B. Hu, N. Mitic, P. Olivares-Chauvet, S. Janjuha, N. Ninov, and J. P. Junker. Simultaneous lineage tracing and cell-type identification using CrIsPr-Cas9-induced genetic scars. *Nature Biotechnology*, 36:469–473, 2018.

[120] D. L. Stachura and D. Traver. *Cellular dissection of zebrafish hematopoiesis*, volume 133. Elsevier Ltd, 2016.

[121] B. Steventon and A. Martinez Arias. Evo-engineering and the cellular and molecular origins of the vertebrate spinal cord. *Developmental Biology*, 432: 3–13, 2017.

[122] B. Steventon, F. Duarte, R. Lagadec, S. Mazan, J. F. Nicolas, and E. Hirsinger. Species-specific contribution of volumetric growth and tissue convergence to posterior body elongation in vertebrates. *Development (Cambridge)*, 143:1732–1741, 2016.

[123] T. Stuart and R. Satija. Integrative single-cell analysis, 2019. ISSN 14710064.

[124] T. Stuart, A. Butler, P. Hoffman, C. Hafemeister, E. Papalexi, W. M. Mauck, Y. Hao, M. Stoeckius, P. Smibert, and R. Satija. Comprehensive Integration of Single-Cell Data. *Cell*, 177:1888–1902, 2019.

[125] J. E. Sulston and H. R. Horvitz. Post-embryonic cell lineages of the nematode, Caenorhabditis elegans. *Developmental Biology*, 56, 1977.

[126] J. E. Sulston, E. Schierenberg, J. G. White, and J. N. Thomson. The embryonic cell lineage of the nematode Caenorhabditis elegans. *Developmental Biology*, 100:64–119, 1983.

[127] J. Sun, A. Ramos, B. Chapman, J. B. Johnnidis, L. Le, Y.-J. Ho, A. Klein, O. Hofmann, and F. D. Camargo. Clonal dynamics of native haematopoiesis. *Nature*, 514:322–327, 2014.

[128] S. M. Sunkin, L. Ng, C. Lau, T. Dolbeare, T. L. Gilbert, C. L. Thompson, M. Hawrylycz, and C. Dang. Allen Brain Atlas: an integrated spatio-temporal portal for exploring the central nervous system. *Nucleic acids research*, 41:996–1008, 2013.

[129] S. K. Tabatabaei, B. Wang, N. Bala, M. Athreya, B. Enghiad, A. G. Hernandez, C. J. Fields, J.-p. Leburton, D. Soloveichik, H. Zhao, and O. Milenkovic. DNA punch cards for storing data on native. *Nature Communications*, 11, 2020.

[130] A. Tanay and A. Regev. Scaling single-cell genomics from phenomenology to mechanism. *Nature*, 541:331–338, 2017.

[131] V. A. Tornini, A. Puliafito, L. A. Slota, J. D. Thompson, G. Nachtrab, A.-L. Kaushik, M. Kapsimali, L. Primo, S. Di Talia, and K. D. Poss. Live Monitoring of Blastemal Cell Contributions during Appendage Regeneration. *Current Biology*, 26:2981–2991, 2016.

[132] V. A. Tornini, J. D. Thompson, R. L. Allen, and K. D. Poss. Live fate-mapping of joint-associated fibroblasts visualizes expansion of cell contributions during zebrafish fin regeneration. *Development*, 144:2889–2895, 2017.

[133] 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. *Development (Cambridge, England)*, 146, 2019.

[134] V. Trivedi, T. Fulton, A. Attardi, K. Anlas, C. Dingare, A. Martinez-Arias, and B. Steventon. Self-organised symmetry breaking in zebrafish reveals feedback from morphogenesis to pattern formation. *bioRxiv*, 2019.

[135] S. Tu and S. Johnson. Fate Restriction in the Growing and Regenerating Zebrafish Fin. *Developmental Cell*, 20:725–732, 2011.

[136] E. Tzouanacou, A. Wegener, F. J. Wymeersch, V. Wilson, and J.-F. Nicolas. Redefining the Progression of Lineage Segregations during Mammalian Embryogenesis by Clonal Analysis. *Developmental Cell*, 17:365–376, 2009.

[137] S. C. van den Brink, A. Alemany, V. van Batenburg, N. Moris, M. Blotenburg, J. Vivi , P. Baillie-Johnson, J. Nichols, K. F. Sonnen, A. Martinez Arias, and A. van Oudenaarden. Single-cell and spatial transcriptomics reveal somitogenesis in gastruloids. *Nature*, 583:405–409, 2020.

[138] M. van Overbeek, D. Capurso, M. M. Carter, M. S. Thompson, E. Frias, C. Russ, J. S. Reece-Hoyes, C. Nye, S. Gradia, B. Vidal, J. Zheng, G. R. Hoffman, C. K. Fuller, and A. P. May. DNA Repair Profiling Reveals Nonrandom Outcomes at Cas9-Mediated Breaks. *Molecular Cell*, 63:633–646, 2016.

[139] Q. Wang, H. Xiong, S. Ai, X. Yu, Y. Liu, J. Zhang, and A. He. CoBATCH for High-Throughput Single-Cell Epigenomic Profiling. *Molecular Cell*, 76:206–216.e7, 2019.

[140] Q. Wang, H. Xiong, S. Ai, X. Yu, Y. Liu, J. Zhang, and A. He. CoBATCH for High-Throughput Single-Cell Epigenomic Profiling. *Molecular Cell*, 76:206–216, 2019.

[141] C. Wolff, J.-Y. Tinevez, T. Pietzsch, E. Stamataki, B. Harich, L. Guignard, S. Preibisch, S. Shorte, P. J. Keller, P. Tomancak, and A. Pavlopoulos. Multi-view light-sheet imaging and tracking with the MaMuT software reveals the cell lineage of a direct developing arthropod limb. *eLife*, 7, 2018.

[142] K. Woo and S. E. Fraser. Order and coherence in the fate map of the zebrafish nervous system. *Development*, 121:2595–2609, 1995.

[143] F. J. Wymeersch, Y. Huang, G. Blin, N. Cambray, R. Wilkie, F. C. K. Wong, and V. Wilson. Position-dependent plasticity of distinct progenitor types in the primitive streak. *eLife*, 5:e10042, 2016. ISSN 2050-084X. doi:10.7554/eLife.10042.

[144] B. Xia and I. Yanai. A periodic table of cell types. *Development (Cambridge)*, 146:1–9, 2019.

[145] J. Xu, L. Zhu, S. He, Y. Wu, W. Jin, T. Yu, J. Y. Qu, and Z. Wen. Temporal-Spatial Resolution Fate Mapping Reveals Distinct Origins for Embryonic and Adult Microglia in Zebrafish. *Developmental Cell*, 34:632–641, 2015.

[146] F. Yue, Y. Cheng, A. Breschi, J. Vierstra, W. Wu, T. Ryba, R. Sandstrom, Z. Ma, C. Davis, B. D. Pope, et al. A comparative encyclopedia of dna elements in the mouse genome. *Nature*, 515:355–364, 2014.

[147] H. Zhang, M. Copara, and A. D. Ekstrom. Differential recruitment of brain networks following route and cartographic map learning of spatial environments. *PloS one*, 7:44886–44886, 2012.

[148] Y. Zhu, K. Gong, M. Denholtz, V. Chandra, M. P. Kamps, F. Alber, and C. Murre. Comprehensive characterization of neutrophil genome topology. *Genes & development*, 31:141–153, 2017.

[149] F. Zou, X. Wang, X. Han, G. Rothschild, S. G. Zheng, U. Basu, and J. Sun. expression and function of tetraspanins and their interacting partners in b cells. *Frontiers in immunology*, 9:1606, 2018.

# Addenda

# Nederlandse Samenvatting

Elke cel in ons lichaam heeft een verleden, heden en toekomst die is gecodeerd in biomoleculen. De geschiedenis van een cel kan bijvoorbeeld gecodeerd worden in kleine veranderingen in het DNA. Wanneer het DNA in een cel eenmaal is veranderd, draagt het deze veranderingen over naar zijn toekomstige nakomelingen. Het bestuderen van het verleden en het heden van een cel kan helpen te voorspellen wat deze cel in de toekomst gaat worden. Om van één enkele cel tot een complex van miljard cellen te komen, ondernemen cellen actie om zich te specialiseren in verschillende celtypen. Het meten van meerdere soorten biomoleculen, bijvoorbeeld DNA en RNA uit dezelfde enkele cel, is echter een uitdaging. Afzonderlijke cellen hebben slechts een klein aantal kopie n van deze biomoleculen en biomoleculen zijn kwetsbaar voor celverstoring, wat een noodzakelijke stap is om de cellulaire inhoud te meten. Met behulp van sequencing-technologie n kunnen RNA of DNA in afzonderlijke cellen betrouwbaar worden uitlezen.

In het eerste hoofdstuk geef ik een overzicht van tools die de hier gepresenteerde technologie en volgen. Ik introduceer hoe CRISPR/Cas9, een groot biomolecuulcomplex, het DNA kan doorknippen. Wanneer het DNA zichzelf herstelt, verwijdert het enkele nucleotiden (letters) die beschadigd zijn en kan het nieuwe introduceren. Dit laat een litteken of streepjescode achter in het DNA, dat wordt gepropageerd naar alle nakomelingen van die cel en dat we kunnen uitlezen. We gebruiken dit om gemeenschappelijke cellulaire voorouders te vinden in de vroege embryonale ontwikkeling waar cellen nog niet dezelfde functie hebben dan later in het volwassen organisme. Ten tweede introduceer ik de methode die we gebruiken om epigenetische modificaties uit te lezen. DNA is gewikkeld rond histoncomplexen die aan hun staart kunnen worden veranderd door histonmodificaties. De histonmodificaties hebben invloed op welke genen op het DNA kunnen worden uitgelezen en welke niet. Ze zijn daarom een cruciale regulator voor de huidige toestand van een cel en kunnen misschien zelfs voorspellen welke genen in de toekomst kunnen worden uitgelezen. We kunnen deze veranderingen aan de histonstaarten uitlezen met specifieke antilichamen die specifieke histonmodificaties herkennen. Door het specifieke antilichaam voor een histonmodificatie te binden aan een ander eiwitcomplex, kunnen we het DNA knippen op locaties die dichtbij de gebonden histonmodificaties liggen. Vervolgens kunnen we handvatten aan de afgebroken DNA einden binden om de posities van de histonmodificaties uit te lezen.

In het tweede hoofdstuk volgen we een specifieke groep stamcellen in het vroege zebravisembryo. Van soortgelijke stamcellen in de muis is het bekend dat ze zich in verschillende celtypen kunnen specificeren zoals ruggenmerg en spieren. Hoe deze stamcellen zich gedragen bij de zebravis is onduidelijk en niet goed gekarak-

teriseerd. Met CRISPR/Cas9 introduceren we in een zeer vroeg stadium een barcode in een grote hoeveelheid cellen en onderzoeken we verschillende organen op hun barcode-overeenkomst in de volwassen zebravis. We volgen ook cellen met microscopie in het jonge zebravisembryo. Onze resultaten suggereren dat deze stamcellen inderdaad een vergelijkbare functie hebben als in de ontwikkeling van muizen.

In het derde hoofdstuk presenteren we een nieuwe technologie, ScarTrace, die het uitlezen van de streepjescodes van CRISPR/Cas9 combineert met RNA-sequencing. Met deze twee soorten metingen van een enkele cel kunnen we zowel de afstamming van een cel als de huidige identiteit van een cel bepalen. We labelen cellen opnieuw in het vroege zebravisembryo en we bestuderen vier hoofdorganen bij de volwassen zebravis: hele niermerg, hersenen, ogen en staartvin. Onze studie toont aan dat de hele niercelklonen in alle celtypen te vinden zijn, wat de hypothese dat een paar stamcelklonen alle belangrijke bloedceltypes maken ondersteund. Verder suggereert onze studie dat cellen die bijdragen aan het linker- en rechteroog worden gescheiden voordat zygotische genactivering plaatsvindt. Ten slotte vinden we een subpopulatie van macrofagen die specifiek zijn voor de staartvin en bijdragen aan de regeneratie ervan na een verwonding.

In het vierde hoofdstuk introduceren we een nieuwe technologie, scChIX, voor het uitlezen van twee histonmodificaties binnen dezelfde enkele cel. Epigenetische regulatie berust op het samenspel van verschillende typen histonmodificaties om het genoom te reguleren. We demonstreren dat scChIX een gemengd signaal van twee typen histonmodificaties in gezuiverde celtypen kan scheiden voor cellen verkregen uit het beenmerg van een muis. Verder gebruiken we scChIX om de celtype bepaling gebaseerd van de ene (actieve) histonmodificatie naar afzonderlijke cellen waar een andere (repressieve) histonmodificatie gemeten is.

Ik sluit dit proefschrift af in het vijfde hoofdstuk, waar ik een algemene discussie geef over kansen en uitdagingen bij het gebruik van de gepresenteerde technologie en. Ik blik vooruit op een toekomst waarin deze technologie  n nuttig kunnen zijn voor het bestuderen van embryonale ontwikkeling. Ten slotte bespreek ik de impact van multi-omics-technologie en die de celtype-identiteit bepalen.

# Rezumat în română

Fiecare celulă din corpul nostru are trecutul, prezentul și viitorul, codate în biomolecule. Prin urmare, istoria unei celule poate fi, de exemplu, codificată cu mici modificări în ADN. Odată modificat, ADN-ul dintr-o celulă propagă aceste modificări urmașilor săi. Studierea trecutului și prezentului unei celule ar putea ajuta la prezicerea a ceea va deveni această celulă în viitor. Pornind de la a fi o singură celulă până la un organism complex cu miliarde de celule, se acționează specializarea celulară în distincte tipuri.

Măsurarea mai multor tipuri de biomolecule, de exemplu ADN și ARN din aceeași celulă, este un proces complicat. O singură celulă are doar o cantitate mică de copii de ADN și ARN, iar biomoleculele sunt sensibile la perturbarea celulelor, care este un pas necesar pentru măsurarea conținutului celular. Tehnologiile de secvențiere pot citi ARN-ul sau ADN-ul în celule unice.

În primul capitol ofer o imagine de ansamblu asupra tehnologiilor utilizate în procesele prezentate aici. Prezint modul în care CRISPR/Cas9, un complex mare de biomolecule, poate tăia ADN-ul. Când ADN-ul se repară, șterge unele nucleotide care sunt deteriorate și introduce unele noi. Această procedură lasă o cicatrice sau un cod de bare în ADN, care este propagat la toate descendențele acelei celule și pe care îl putem citi. Folosim acest cod de bare pentru a găsi strămoși celulari comuni în dezvoltarea embrionară timpurie atunci când celulele respective nu au încă aceeași funcție ca în organismul adult. În al doilea rând, prezint metoda folosită pentru citirea modificărilor epigenetice. ADN-ul este înfășurat în jurul complexelor histonice care pot fi modificate la coada lor prin modificări ale histonelor. Modificările histonice influențează care gene pot fi citite de pe ADN și care nu. Prin urmare, acestea sunt un regulator crucial pentru starea actuală a unei celule și pot chiar prezice care gene pot fi citite în viitor. Putem citi aceste modificări la cozile histonice cu anticorpi specifici care recunosc modificări specifice ale histonelor. Prin legarea anticorpului specific de modificare a histonelor de un alt complex de proteine, putem tăia ADN-ul în apropiera modificărilor histonice. Putem genera anumite legături pentru a citi poziția acestor modificări ale histonelor.

În cel de-al doilea capitol, urmărim un grup specific de celule stem din embrionul de pește zebra timpuriu. Se știe că aceste celule stem sunt capabile să se specializeze în diferite tipuri de celule din măduva spinării și mușchi la șoarece. Comportamentul acestor celule stem în peștele zebră nu este clar și nu este bine caracterizat. Cu CRISPR/Cas9, introducem un cod de bare în aceste celule într-un stadiu foarte timpuriu și studiem diferite organe pentru determinarea asemănării acestora cu codul de bare la peștii adulți. De asemena, urmărim la microscop celulele în embrionul de pește tânăr. Rezultatele noastre sugerează că aceste celule stem prezintă într-adevăr două tipuri de celule cu dezvoltare similară cu cea observată la a șoarece.

În cel de-al treilea capitol prezentăm o nouă tehnologie, ScarTrace, care combină citirea codurilor de bare de la CRISPR/Cas9 cu secvenţierea ARN. Cu aceste două măsurători dintr-o singură celulă, putem determina descendenţa unei celule, precum şi identitatea actuală a acesteia. Precum în celde-al doilea capitol, etichetăm din nou celulele din embrionul de peşte zebra timpuriu şi studiem patru organe majore ale peştelui adult: măduva renală întreagă, creierul, ochii şi aripioarele caudale. Studiul nostru arată că toate clonele de celule ale măduvei renale pot fi găsite în toate tipurile de celule ale măduvei renale, confirmând ipoteza conform căreia câteva clone de celule stem formează toate tipurile majore de celule sanguine. În continuare, studiul nostru sugerează că celulele care contribuie la ochiul stâng şi cel drept sunt separate înainte de activarea genei zigotice. În cele din urmă, găsim o sub-populaţie de macrofage care sunt specifice aripioarei caudal şi contribuie la regenerarea ei după rănire.

În al patrulea capitol, introducem o tehnologie nouă, scChIX, folosită pentru citirea a două modificări de histonă în cadrul aceleiași celule. Reglarea epigenetică se bazează pe interacţiunea diferitelor modificări ale histonelor pentru a regla genomul. Demonstrăm că scChIX poate separa un semnal mixt de două modificări de histonă în tipurile de celule din măduva osoasă a şoarecilor. Mai mult, folosim scChIX pentru a transfera etichete de tipul celulelor de la o modificare histonică (activă) la celule singure dintr-o altă modificare a histonelor (represive).

Închei această teză în capitolul al cincilea în care ofer o discuţie generală despre oportunităţile şi provocările folosirii tehnologiilor prezentate. Discut despre perspectivele viitoare cu privire la modul în care aceste tehnologii ar putea fi utile pentru studierea dezvoltării embrionare. În cele din urmă, discut despre impactul tehnologiilor multi-omice, adică tehnologii care pot să citească mai multe măsurători din aceeași celulă, şi influenza lor în definiţia noastră actuală de diferite tipurii de celule.

# Zusammenfassung auf Deutsch

Jede Zelle in unserem Körper hat eine Vergangenheit, Gegenwart und Zukunft, die in Biomolekülen kodiert ist. Die Geschichte einer Zelle kann beispielsweise in kleinen Veränderungen in der DNA kodiert sein. Wenn die DNA einer Zelle einmal verändert wird, überträgt sie diese Veränderungen auf ihre zukünftigen Nachkommen. Untersuchungen an der Vergangenheit und Gegenwart einer Zelle können helfen, vorherzusagen, was in Zukunft mit der Zelle passieren kann. In der embryonalen Entwicklung kann man dieses Phänomen anschaulich beschreiben. Wenn wir von einer befruchteten Eizelle zu einem komplexen Organisums aus Milliarden Zellen werden, müssen sich Zellen im Laufe der Zeit zu verschiedene Zelltypen zu spezialisieren.

Die Messung mehrerer Arten von Biomolekülen, beispielsweise DNA und RNA aus derselben Einzelzelle, ist jedoch eine Herausforderung. Einzelne Zellen haben nur eine geringe Menge an Kopien und Biomoleküle sind bei Zellzerstörung zerbrechlich, was ein notwendiger Schritt zur Messung des Zellinhalts ist. Sequenzierungstechnologien können RNA oder DNA in einzelnen Zellen zuverlässig auslesen.

Im ersten Kapitel gebe ich einen Überblick über Techniken, die die hier vorgestellten Technologien vorausgegangen sind oder vergleichbare Techniken mit den hier vorgestellten. Ich stelle vor, wie CRISPR / Cas9, ein gro  er Biomolekülkomplex, die DNA schneiden kann. Wenn sich die DNA selbst repariert, löscht sie einige Nukleotide, die beschädigt sind und führt neue hinzu. Dies hinterlässt eine Narbe oder ein *Barcode* in der DNA, die an alle Nachkommen dieser Zelle weitergegeben wird und die wir auslesen können. Wir nutzen diese *Barcodes*, um gemeinsame zelluläre Vorfahren in der frühen Embryonalentwicklung zu finden, wenn Zellen noch nicht die gleiche Funktion wie im erwachsenen Organismus haben. Zweitens stelle ich die Methode vor, mit der wir epigenetische Modifikationen auslesen. DNA ist um Histonkomplexe gewickelt, die an ihrem Schwanz durch Histonmodifikationen verändert werden können. Die Histonmodifikationen haben Einfluss darauf, welche Gene auf der DNA ausgelesen werden können und welche nicht. Sie sind daher ein entscheidender Regulator f  r den gegenwärtigen Zustand einer Zelle und können möglicherweise sogar vorhersagen, welche Gene in Zukunft ausgelesen werden können. Wir können diese Veränderungen an den Histonschwänzen mit spezifischen Antikörpern auslesen, die spezifische Histonmodifikationen erkennen. Durch die Bindung des Histonmodifikations-spezifischen Antikörpers an einen anderen Proteinkomplex können wir die DNA an Stellen in der Nähe dieser Histonmodifikationen schneiden. Wir können bestimmte Griffe ligieren, um die Position dieser Histonmodifikationen durch sequenzieren ablesen zu können.

Im zweiten Kapitel verfolgen wir eine bestimmte Gruppe von Stammzellen im

frühen Zebrafischembryo. Es ist bekannt, dass diese Stammzellen in der Lage sind, verschiedene Zelltypen von Rückenmark und Muskel in der Maus zu bestimmen. Wie sich diese Stammzellen im Zebrafisch verhalten, ist unklar und nicht gut charakterisiert. Mit CRISPR/Cas9 führen wir sehr früh einen Barcode in diese Zellen ein und untersuchen verschiedene Organe auf ihre Barcode-Ähnlichkeit im erwachsenen Zebrafisch. Darüber hinaus verfolgen auch diese bestimmte Gruppe von Stamzellen über mehrere Zellteilungen mit Mikroskopie im jungen Zebrafischembryo. Unsere Ergebnisse legen nahe, dass diese Stammzellen tatsächlich zwei Zellschicksale haben, die der Mausentwicklung ähnlich sind.

Im dritten Kapitel stellen wir eine neuartige Technologie namens ScarTrace vor, die das Auslesen von Barcodes aus CRISPR/Cas9 mit RNA-Sequenzierung kombiniert. Mit diesen beiden Messungen an einer einzelnen Zelle können wir die Abstammungslinie einer Zelle sowie die aktuelle Identität einer Zelle bestimmen. Wie im vorherigen Kapitel, markieren wir erneut Zellen im frühen Zebrafischembryo und untersuchen vier Hauptorgane im adulten Zebrafisch: Nierenmark, Gehirn, Augen und Schwanzflosse. Unsere Studie zeigt, erstens, dass alle Nierenmarkzellklone in allen Nierenmarkzelltypen gefunden werden können, was die Hypothese bestätigt, dass eine bestimmte Anzahl Stammzellklone alle wichtigen Blutzelltypen bilden. Ferner legt unsere Studie nahe, dass Zellen, die zum linken und rechten Auge beitragen, vor der Aktivierung zygotischen Genen schon getrennt sind. Schliesslich finden wir eine Teilpopulation von Makrophagen, die spezifisch für die Schwanzflosse sind und zu ihrer Regeneration nach einer Verletzung beitragen.

Im vierten Kapitel stellen wir eine neuartige Technologie vor, scChIX, mit der zwei Histonmodifikationen aus derselben Zelle ausgelesen werden können. Die epigenetische Regulation beruht auf dem Zusammenspiel verschiedener Histonmodifikationen, um das Genom zu regulieren. Wir zeigen, dass scChIX das duale Signal zweier Histonmodifikationen in Zelltypen vom Knochenmark der Maus trennen kann. Weiterhin verwenden wir scChIX, um Celltyp-Markierungen von einer (aktiven) Histonmodifikation auf einzelne Zellen von einer anderen (repressiven) Histonmodifikation zu übertragen.

Ich schliesse diese These im fünften Kapitel ab, in dem ich eine allgemeine Diskussion über Chancen und Herausforderungen beim Einsatz der vorgestellten Technologien einleite. Ich diskutiere einen zukünftigen Ausblick darauf, wie diese Technologien f r die Untersuchung der Embryonalentwicklung nützlich sein könnten. Abschliessend diskutiere ich die Auswirkungen von Multi-Omics-Technologien, also Technologien die mehrere Biomoleküle aus der gleichen Zelle messen können, auf unser Verständnis von der Identität von Zelltypen haben können.

# Acknowledgments/ Mulţumiri/ Danksagungen

I want to take the opportunity to thank the people which inspired, supported and made my PhD journey smooth and pleasant.

First of all, **Alexander**, thank you for the opportunity to do my doctoral research in your group. I have been constantly learning in this innovative and dynamic environment. It was great to follow my curiosity and thank you for all your feedback and enthusiasm over the years.

I want to thank **Jeroen Bakkers** and **Rob de Boer** for their feedback throughout the years as part of my PhD committee meetings.

I also want to thank all the professors in my assessment committee, **Nadine Vastenhouw**,**Wouter de Laat**, **Niels Geijsen**, **Jeroen Bakkers** and **Rob die Boer** who provided feedback about this thesis prior to publication.

Meine ehemaligen Kollegen aus Basel: **Katharina**, **Robert**, **Dimos**, **Michael** und **Lukas**, es war ganz schön schwer das schöne Basel zu verlassen und ich danke euch ganz herzlich für die tolle Zeit. Ich habe viel von euch gelernt. Ihr habt mich super vorbereitet auf die 5 PhD Jahre, die ich hinter mir habe. Auch wenn ich nicht besonders gut darin bin, Kontakt über die Ferne zu halten, bin ich froh, euch doch noch ab und zu gesehen oder von euch gelesen zu haben.

My collaborators in Cambridge: Thank you, **Ben**, **Andrea** and **Tim** for the fruitful collaboration and biological discussions. I learned so much from you which really deepened my interest for developmental biology.

**Sarah**, besonders Dank für all deine schönen Illustrationen und brainstorming Treffen, die diese Thesis bereichern.

At the Hubrecht, I want to thank some people who created a great atmosphere:

**Thea** en **Elsa**, hartelijk bedankt dat jullie waarschijnlijk bijna iedere week een pakketje voor mij hebt ontvangen. Het was altijd leuk om bij het afhalen nog een korte praatje met jullie te hebben.
I want to thank the animal caretaker team in the fish as well as the mouse facility for their great help throughout my projects as well as **Jeroen Korving** for introducing me to proper mouse embryos dissection.

Our female-powered group neighbours: the whole Kind group! I want to thank especially **Kim**, **Koos**, **Franka**, **Sandra**, **Samy**, **Tess** and **Silke**, for the nice scientific and non-scientific chats at the water dispenser, bathroom or lab.
**Annemiek**, thank you for your caring interest in the overall well-being of the group and for your support with organising this thesis and any other meetings.
**Annabel**, thank you for my first Sinterklaas experience, your awesome dinner parties and your cheerful and good *humeur* which is always great to experience.
**Jens** and **Maria**, I always had a good evening when we went to see you two. Thank you for that. I wish you all the best with finishing your PhDs and afterwards.

Thank you to all the van Oudenaarden group members for their support throughout the years. First of all, the people I spent my first years with and welcomed me greatly into the group. It was quite some culinary experience involving more cake than I could handle. **Judith**, thank you for your constant good mood. **Jean-Charles** you were a great zen oase in the office. **Susanne**, I hope you'll get some well-deserved rest after your PhD. **Anna L.**, thank you for your honest comments to any discussion and for keeping the lab up to date with the latest trends. **Lennart**, you almost made me wonder whether eating an apple every workday at 4pm is the key to a successful PhD. **Maya**, thank you for so often convincing me of going out for dinner. **Mauro**, thank you for being such a great listener, dinner host and for all your help with my post-PhD orientation. **Christoph**, mein unerwarteter Pflanzenfreund, der Tag wurde immer besser wenn ich dich in rosa Socken und gleichfarbigen Pulli gesehen habe. **Nico**, when we both managed to stick our heads through the clouds, I really appreciated your honest feedback and suggestions (only few ideas I wish I would have never heard ;) ). Thank you, **Reinier**, for your competent and calming FACS-sorting experience. I always trusted my experiments more when the cells were approved by your eyes. **Josi**, thank you for teaching me everything I know about handling zebrafish. I would have been completely lost in my first years without you. It was a great to have you as bench neighbour and be able to chat with you during long pipetting days. During my first project we worked intensely on scartrace and I thank **Anna A.**, **Josi**, **Chloé** and **Alexander** for all their enthusiasm.

I also want to thank the newer group-members who greatly changed up the lab. **Vincent**, I am glad another person joined who understands *the office* jokes. **Helena**, **Jeroen**, **Freddy**, **Mike** and **Vivek**, it's great that you are keeping the noisy office spirit alive. **Joe**, arigato for proof-reading my thesis and for teaching me so many British synonyms. **Marius**, my master student, thank you for being a dedicated and entertaining student. I don't know how you manage to make me laugh about the darkest stories. **Jake**, thank you for re-viving my R skills and for appreciating great food. I survived many afternoons on your snap-frozen

mushrooms or salt'n'vinegar chips. **Marijn**, thank you, for sharing your sailing enthusiasm with us and for all your awesome scientific feedback. **Peter**, danke, dass du mich mit dem Chromatin-Fieber angesteckt hast. Ich habe das neue Thema und Lernerfahrung sehr genossen auch wenn die Projekte nicht so ganz gelaufen sind wie geplant.

**Li ting**, how great that you joined the group and I got to know you outside the lab. Thanks for the great dinners, baking sessions and your curiosity for European culture.

**Corina**, how awesome that we get to finish more or less at the same time. It was fun to share this ride with you and thank you so much for all our hang-outs and your big smile.

**Dylan**, you're a wise man and things always went smooth when I listened to your adivce. Thank you both, **Steph** and **Dylan**, for the yummy dinners and fun game evenings. You're both such inspiring people and I always enjoy our talks, serious and non-serious ones. I hope our families will keep in touch.

**Marloes**, you're one of a kind and that's great. I'll definitely miss our office hang-outs (luckily, **Jelmer** doesn't mind too much to hang out with us as well, thank you!). It was always great that we could communicate without words and just got the same impression from a situation. I deeply admire your strong sense of fairness and your determination. I wish you all the best with your further PhD journey which you will surely master perfectly. I'm confident about it.

**Ännchen**, danke für den atemberaubenden Wanderurlaub und dein grosses Verständnis. Danke, dass du immer ein offenes Ohr hast und es hat mir immer geholfen deine andere Perspektive zu verschieden Dingen zu hören. Ich bin stolz, dass Du tapfer und positiv deinen Weg gehst trotz all den Schwierigkeiten.

Meine Studienzeit in der Schweiz war unvergesslich und ich bin besonders froh, dass ein Paar Freundschaften der Distanz und Zeit getrotzt haben. **Marie-Louise**, danke füer das tolle Neujahr in der Valser Landschaft und für all deine lieben Päckchen und Karten aus der ganzen Welt. Deine positive Energie ist immer super ansteckend.
**Isa**, danke, für die Besuche hier im hohen Norden von Dir und Joshua. Es ist immer schön was mit euch zu unternehmen. Es war wirklich super, dass Du mir beim Umzug so sehr geholfen hast und wir ein Paar Tage die Niederlande auf dem

Rad, zu Fuss oder im Kanu zusammen erkunden konnten. Das hat den Beginn hier total versüsst.
**Irina**, wir haben's beinah beide geschafft! Trotz all dem Stress und der Ferne hat es immer gut getan Dir von all den Ereignissen hier zu erzählen und von Dir zu hören. Wir haben es trotz unserer Abneigung gegen Chats geschafft uns zu sehen und in Kontakt zu bleiben. Danke für all die Besuche von Dir und auch von Marco.

Most importantly, I want to thank my ever growing family.

**Alex** şi **Irina**, ce frumos că aţi venit şi voi aicea în Holanda. Mulţumesc mult pentru corecţile care au îmbunătaţit mult rezumatul.

**Yolande**, **Erik**, **Linde** & **Xander** en **Haye** & **Louise** , bedankt dat jullie mij zo hartelijke opgenomen hebben in jullie familie en voor all jullie steun door de PhD. Het is doch altijd gezellig bij jullie met diepe discussie en mooie wandelingen.

**Raimund**, danke, dass du dich auch so für Lego begeistern kannst und dass du so liebevoll auf meine Mama acht gibst.

Vreau să mulţumesc lui **Carmen**, **Elena** şi **Constantin**. Vă multumesc pentru toate vizitele, sprijin şi momente caragioase de care îmi amintesc. Mă bucur mereu să vă văd şi să ştiu că vă merge bine.

**Cătălin**, **Bunica** si **Cristi**. Ce m-aş face fără voi ? Mă bucur că aveţi aşa de mult success cu firma voastră, chiar dacă munciţi aşa de mult că nu prea putem să ne vedem. Sunt mandră de voi şi vă mulţumesc pentru tot sprijin care mi l-aţi dat.

**Mama** şi **Tata**. Eu nu a-ş fi aşa cum sunt fără voi. Vă multumesc din inimă că aveti aşa de multă încredere in tot ce fac si mă sprijiniţi in aventurile mele chiar daca de multe ori nu înţelegeţi motivile mele. Ultimii ani au fost cam greii cu plecarea din Eleveţia, mutatul în Holanda şi cu toate surprizele care mau găsit aicea. Uite că totuşi sa incheiat cu bine si vă mulţumesc mult pentru sprijinul vostru.

**Buys**. Every day I feel like the luckiest person who can spent so much time with you. Thank you for all the creative projects we tackle and for sparking my joy for this small place in the world with seemingly infinite bike paths which are so smoothly paved like the so often enjoyed peanut butter. I simply have the best time with you and I am grateful for having found my partner and best friend in one.

# List of publications

Whole-organism clone tracing using single-cell sequencing
A. Alemany*, M. Florescu*, C. Baron*, J. Peterson-Maduro* and A. van Oudenaarden, *Nature (2018)*

Neuromesodermal Progenitors are Conserved Source of Spinal Cord with Divergent Growth Dynamics
A. Attardi*, T. Fulton*, M. Florescu et al., *Development (2018)*

Exon/Intron-split analysis quantifies both transcriptional and post-transcriptional contributions to differential expression in RNA-seq
D. Gaidatzis*, L. Burger*, M. Florescu and M. Stadler, *Nature Biotech (2015)*

**Preprint or in preparation**

Deconvolving multiplexed histone modifications in single cells
J. Yeung*, M. Florescu*, P. Zeller, B.A. de Barbanson and A. van Oudenaarden, *in preparation (2020)*

Massively parallel whole-organism lineage tracing using CRISPR / Cas9 induced genetic scars
J.P. Junker, B. Spanjaard, J. Peterson-Maduro, A. Alemany, B. Hu, M. Florescu, A. van Oudenaarden, *bioRxiv (2017)*

* equal contribution

# Curriculum Vitae

Maria Florescu was born on the 31st of october 1989 in Buzău, Romania. After finishing high-school in Freiburg in Germany she moved to Basel, Switzerland, to study Biology at the Biozentrum, University of Basel. She specialized in molecular biology during her studies. Already during her Bachelors she found a strong interest in computational biology and joined the group of Simon Berneche to work on simulations of membrane proteins. For her Master degree in Molecular Biology she then continued to work in the computational biology field, but decided to explore the sequencing field at the Friedrich-Miescher Institute in Michael Stadler's group. Here she first explored single-cell sequencing data towards the end of her internship.
In August 2015 Maria started her PhD at the Hubrecht Institute in the group of Prof. Dr. A. van Oudenaarden with a strong interest in being able to measure multiple modalities from the same single cell and embryogenesis. The results of this research are described in this thesis.