Single-Cell Transcriptional Dynamics of Cell Fate Decisions during Chicken Development

Inauguraldissertation

zur

Erlangung der Würde eines Doktors der Philosophie

vorgelegt der

Philosophisch-Naturwissenschaftlichen Fakultät

der Universität Basel

von

Christian Abel Feregrino Feregrino

Basel, 2021

Genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultät

auf Antrag von

Prof. Dr. Patrick Tschopp,

Prof. Dr. Walter Salzburger, und

Prof. Dr. Joshua Payne

Basel, den 17.11.2020

Prof. Dr. Martin Spiess

Dekan

Acknowledgments

This thesis, would have not been possible without the support and contribution of many people. I would like to acknowledge them for contributing to all the scientific work presented here, the scientific work that didn't make the cut, and the non-scientific work that made all of it possible.

First, I would like to thank my PhD advisor, **Patrick Tschopp**, for giving me the opportunity to work in his research group. And also for the long, and sometimes difficult journey towards this thesis. Always keeping integrity, critical thinking, and a passion for research that I aspire to achieve.

I would also like to thank the members of my examining committee, **Prof. Dr. Walter Salzburger**, **Prof. Dr. Joshua Payne** and **Prof. Dr. Barbara Treutlein**, for their time and interest to go through my thesis and offer valuable input and discussions.

I'm thankful with all the people who directly collaborated in the production of this thesis, with data production and analysis. Especially, I thank **Fabio Sacher**, not only for providing coding and analyses assistance, but for always having interesting questions and suggestions, aside from keeping all those laughs and all those beers coming. I'm also thankful with **Emanuelle Grall** and **Chloé Moreau**, for the amazing amount and quality of work they did in the lab, and for always being an inspiration and magnificent co-workers, merci!

I thank **Prof. Dr. Henrik Kaessmann** for hosting me for a few months in the University of Heidelberg, providing an exciting and inspiring environment, and invaluable scientific discussions. I would also like to thank all the members of his research group, for all the stimulating discussions; specifically **Evgeny Leushkin**, **Francesco Lamanna** and **Florent Murat** for their input regarding my WGCNA analysis, and **Panchita**, **Marta**, **Amir** and **Celine** for being especially welcoming.

I'm very thankful with the rest of the regulatory evolution lab, past and present. **Sabrina Fischer**, whom I met a different name, but the same passion. Thank you for making all of my work possible. **Victoria Haller**, countess of the chickens, for making those early days so much fun. **Alex**, for all the questions, all the laughter. **Bianka**, always trying different things, always happy to help me, always slaying, we will have our castles, always and forever. **Maëva**, j' ne te remercierai jamais assez de m'avoir empêché de sombrer dans la folie et de m'avoir soutenu lorsque je pensais m'effondrer. Merci infiniment pour tout, et pour les cours de français. With the master students' squad: **Gabriele**, **Ana**, **Aline**, **Nicolas**, for always keeping up the spirits in the office! And thanks to **Menghan**, **Antoine** and **Navaneeth** for the short, but nice time we've spent.

I'm also thankful with the rest of the zoological institute. The participants of the interaction seminar, with valuable comments and input along my research progress. And especially those who agreed to participate in the RNA-seq club, and **Athimed El Taher**, for supporting and helping with making this idea become true, and all the good times we spent together. I'm grateful with **Jeremias**, **Nicolas Boileau**, **Peter Fields**, **Fabrizia**, and **Axel** for always providing fun scientific and non-scientific discussions, and arguments. My PhD sisters by adoption: **Maridel**, always fighting and encouraging us to fight for what's right. **Virginie**, reminding me what is it all about, the endless kindness.

I thank all my friends in Basel, who always supported me in uncountable ways, **Niko Vellnow**, **Erika**, **Selen**, **Linh**, **Nassim** and the rest of *'s Räppli Huus*. I'm especially in debt with **Telma**, who made me a much better person and scientist, in ways I never thought of. And **Roland**, who always pushed me to do more, supported all my crazy ideas, and was there for me during the last very difficult months.

None of this would have been possible without the unconditional support and affection from **my parents** and **my siblings**, from 9,594 km away.

Lastly, I would like to thank the **EMBO** for awarding me with a short term fellowship, and the **University of Heidelberg** for hosting me. And most importantly, the **University of Basel**, the **Swiss 3RCC** and the **Swiss SNF** for funding my research. And the swiss mountains, for keeping me sane.

Us Basel an mym Rhy, em Bebbi sy Stadt, Christian Feregrino

"Gentiles enim, quorum peritia in hac arte probabilis est, creant sibi basiliscos hoc modo. […] ponut duos gallos veteres duodecim aut quindecim annorum […]. Qui cum incrassate fuerint, […] convenient inter se et ponunt ova. Quibus positis eiciuntur galli, et immittuntur bufones qui ova foveant […]. Fotis autem ovis egrediuntur pulli sicut pulli gallinarum, quibus post dies septem crescent caude serpentium."

Theophilus Presbyter (Rogerus von Helmarhausen) 1125
Schedula diversarum atrium
Liber III Caput XLVIII De auro hyspanico

"The Gentiles, whose skillfulness in this art is probable, make basilisks in this manner. […] they place two old cocks of twelve or fifteen years […]. When they have become fat […], they agree together and lay eggs. As soon as the eggs are laid, the cocks are taken out and toads are put in to sit on the eggs […]. When the eggs are hatched, chicks come forth who look like young roosters, but after seven days they grow serpents' tails."

Theophilus Presbyter (Roger von Helmarhausen) 1125
List of various arts
Book III Chapter XLVIII Of Spanish gold

CONTENTS

List of figures

List of Tables

Chapter 1 - Optimizing Single-cell RNA-seq Analysis Methods for the Study of Chicken Development

List of abbreviations

scRNA-seq	Single-cell RNA sequencing
UMI	Unique molecular identifier
snRNA-seq	Single-nucleus RNA sequencing
poly(A) tail	Polyadenylated tail
poly(T) tail	Poly(thymine) tail
cDNA	Complementary DNA
UTR	Untranslated transcribed region
TSS	Transcription start site
CpG islands	DNA regions rich in CG sites
mRNA	Messenger RNA
TF	Transcription factor
CoRC	Core regulatory complex
GRN	Gene regulatory network
EvoDevo	Evolutionary developmental biology
PCA	Principal component analysis
MAD	Median absolute deviation
δS-G2M	Difference of S phase score and G2M phase score
tSNE	t-distributed stochastic neighbor embedding
CCA	Canonical correlation analysis
GO	Gene Ontology
GOE	Gene ontology terms enrichment
nn	Nearest neighbors
GTF file	Gene transfer format file
NCBI	National center for biotechnology information
BAM file	Binary alignments map file
DE	Differential expression / Differentially expressed
WGCNA	Weighted gene co-expression network analysis
HH	Hamburger-Hamilton stage
GRCg6a	Genome Reference Consortium chicken build 6a
AER	Apical ectodermal ridge
LPM	Lateral plate mesoderm
nsCT	Non-skeletal connective tissue
PFR	Phalanx forming region
CNC	Cranial neural crest
SZJ	Stylopod-zeugopod joint

GENERAL INTRODUCTION

Multicellular animals, with their different cell types, tissues, organs and outstandingly varied structures and shapes are almost miraculously build from one single cell – the fertilized zygote – with every generation. This developmental process is not only repeated for every new individual of a given species, but may also show signs of remarkable conservation amongst species of the same phylum (Duboule, 1994; Richardson, 1999). It is the slight variations during embryonic (and post-embryonic) development which ultimately result in the different forms of morphological and physiological diversity we observe. Therefore, to identify the proximate causes of morphological diversity and diversification, we need to understand the variations that development shows across different organisms. The conservation and divergence of the developmental processes has captivated the curiosity of naturalists and scientists since Aristotle (Horder, 2010). Early work from 18th and 19th century German-speaking scientists - Pander, Ernst von Baer, and Rathke - first describing chicken, amphibians and later other tetrapod embryos (Gilbert, 2000), resulted in seminal observations that keep providing a conceptual framework for an incredible amount of basic and applied research to this date.

While developmental biology is a discipline with a long history (Horder, 2010; Gilbert, 2017), and much has been learned about the molecular mechanisms active during early development and later in patterning and differentiation stages, many questions remain unanswered. We have a good understanding of developmental processes at the tissue level, yet in recent years, studies have started to focus on cellular and subcellular levels. Importantly, the advent of several seminal new technologies now allows us to revisit classical questions in developmental biology from a completely new angle. One of these latest scientific and technological advancements in molecular methods is the ability to perform high-throughput sequencing of cellular transcriptomes at single-cell resolution (Mayr *et al*, 2019; Marioni & Arendt, 2017). Describing changes at the cellular level is a powerful approach to understand a central aspect of development: cell differentiation. Cell differentiation occurs at precise temporal and spatial coordinates, defined by molecular signals. In response to these different signals, regulatory changes allow cells, which carry the same genome, to diverge into all different cell types of an organism. These regulatory changes occur independently and simultaneously within each single cell.

Studying Gene Expression Dynamics at Single-cell Resolution

Single-cell RNA sequencing (scRNA-seq) allows us to examine the dynamics of gene expression at an unprecedented resolution compared to its precursor, now termed “bulk RNA-seq”. The idea of obtaining gene expression measurements from single cells had been in the mind of scientists for several years, starting with manual separation of each cell (Tang *et al*, 2009), developing into early multiplexing methods on well plates (Islam *et al*, 2011) and highly sensitive full-length sequencing approaches (Ramsköld *et al*, 2012). Full-length sequencing of RNA of single cells isolated in well plates has been further developed (Picelli *et al*, 2014; Hagemann-Jensen *et al*, 2020) and today allows the highest detection rates of gene expression, even providing insight into expression isoforms. Nonetheless, the technological advancements in microfluidics technology, as well as increased affordability, have also opened up the possibility to make single-cell sequencing scalable. By the mid-2015 several breakthrough methods were announced (Klein *et al*, 2015; Macosko *et al*, 2015) that were capable of processing thousands of single cells at the same time. Since then, the use of single-cell RNA-sequencing has sky-rocketed in fields as diverse as cancer, evolutionary, and - of course - developmental biology.

Currently, a basic, standard scRNA-seq experiment starts with cell dissociation and results in clusters of highly similar cells and the genes that are differentially expressed in these clusters. First, the selected tissue is dissociated using a combination of mechanical an enzymatic procedures to obtain viable single cells (Nguyen *et al*, 2018). Cells are then collected in well plates using a cell sorter for full-length sequencing methods (Hagemann-Jensen *et al*, 2020), or fed into a microfluidic device which captures them into nanodroplets for massively-parallel sequencing (Zheng *et al*, 2017). Cells are lysed inside each well or nanodroplet, and their mRNA captured and sorted in molecular libraries to be later sequenced. The sequences obtained are then processed using bioinformatics tools, and mapped against a genome or transcriptome, to identify their genes or origin. The reads are counted and assigned to their cell of origin, followed by a quality control to remove non-informative cells. In the end, several bioinformatics approaches can be used to cluster the cells based on their transcriptomic similarities, and determine the genes that are differentially expressed between the clusters (Luecken & Theis, 2019).

The first crucial step of scRNA-seq is the dissociation of the tissue into a suspension of single cells. In order to obtain meaningful results, the cells need to reach the lysis step individually, alive and as undisturbed as possible (Nguyen *et al*, 2018).This fact restricts the types of tissues and experiments that can be done. Since tissues vary in their composition, a compromise is met, between dissociation efficiency and cell viability. For this reason, dissociation protocols are not universal, and a variety of commercial kits are available for the efficient dissociation of specific tissues. Most dissociation procedures today are based on both mechanical and enzymatic dissociation of the cells.

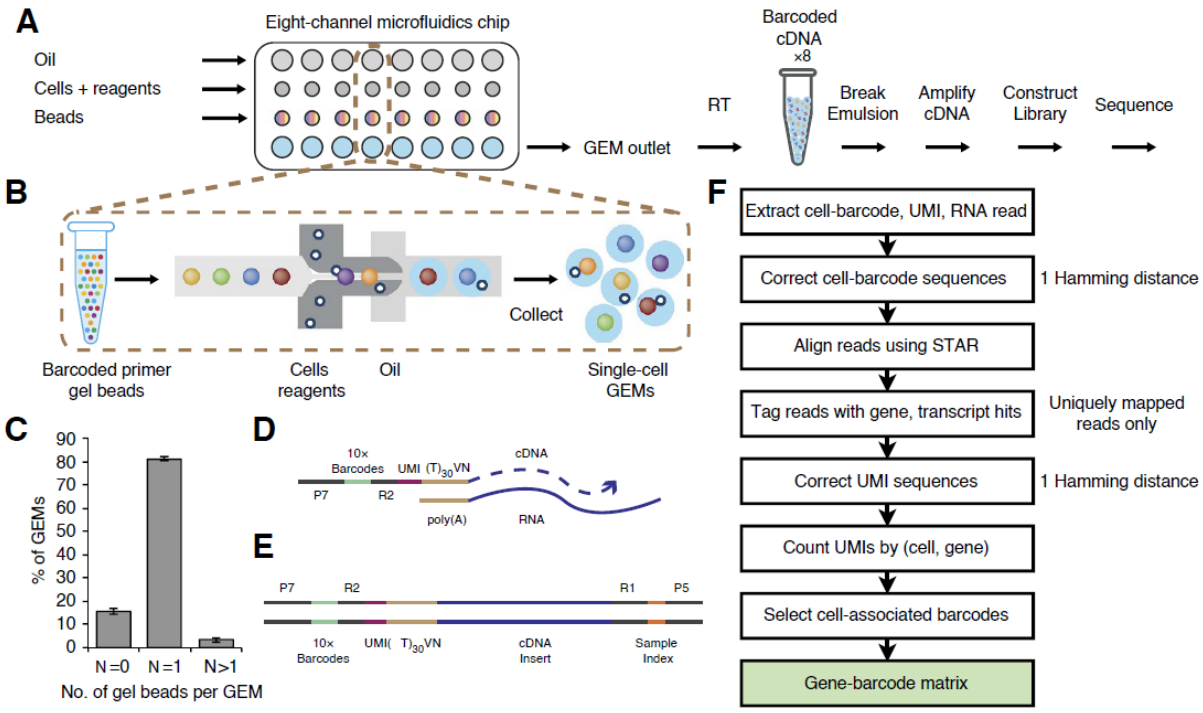


Figure 1 Single-cell RNA-seq experiment workflow. Adapted from: Massively Parallel Digital Transcriptional Profiling Of Single Cells (Zheng et al, 2017). **A** scRNA-seq workflow on Chromium 10x Genomics technology platform. Cells are combined with reagents in one channel of a microfluidic chip, and gel beads from another channel to form gel beads in emulsion (GEM) or gel bead containing droplets. RT takes place inside each GEM, after which cDNAs are pooled for amplification and library construction in bulk. **B** Gel beads loaded with primers and barcoded oligonucleotides are first mixed with cells and reagents, and subsequently mixed with oil-surfactant solution at a microfluidic junction. Single-cell GEMs are collected in the GEM outlet. **C** Percentage of gems containing 0 gel bead (n=0), 1 gel bead (n=1) and >1 gel bead (n>1). **D** gel beads contain barcoded oligonucleotides consisting of illumina adapters, 10x barcodes, UMIs and oligo dTs, which prime RT of polyadenylated RNAs. **E** finished library molecules consist of illumina adapters and sample indices, allowing pooling and sequencing of multiple libraries on a next-generation short read sequencer. **F** Bioinformatics CellRanger pipeline workflow. Gene-barcode matrix (highlighted in green) is an output of the pipeline.

Today, massively parallel single-cell RNA-seq is based on microfluidics devices and relies on microbeads coated with nucleotide strands (Figure 1 A). The nucleotide strands are barcoded with nucleotide sequences, one unique to each bead, which allows the identification of the beads from each other, and another unique to each strand which serves as a unique molecule identifier (UMI) (Figure 1 D and E). Massively parallel scRNA-seq takes advantage of the poly(A) tail (chain of multiple adenosine monophosphates) that mRNA contains. Hence, the nucleotide strands surrounding the microbeads have a sequence of repeated thymine to capture mRNA (Macosko *et al*, 2015; Klein *et al*, 2015). In a standard experiment, the single cell suspension is fed into a microfluidic device that controls the inflow of an emulsion agent, dissociated cells, enzymatic reagents and beads (Figure 1 B). The emulsion agent creates bubbles, or droplets, inside which, single cells and single beads are combined and enclosed. The efficacy to capture single cells relies on probabilities, generating an excess of droplets, many of which remain empty, only some contain one bead (Figure 1 C), and even less contain both a bead and a cell. Inside the

droplets, cells are lysed by the enzymatic reagents and their contents released, allowing the beads to interact with the mRNA transcripts.

Once the mRNA transcripts have been captured by the beads, the procedure is not very different from that of a bulk RNA-seq experiment. The emulsion is broken and the beads pooled. The transcripts are reverse-transcribed into complementary DNA (cDNA), amplified, and compiled into indexed cDNA libraries. Nowadays commercial setups like Chromium Single Cell Expression from 10x Genomics allow for fast and standardized experiments. cDNA libraries are paired-end sequenced to obtain the 3' end of the transcripts and the different barcodes and indices. Dedicated bioinformatics pipelines align the cDNA reads to the genome of the corresponding species, de-multiplex them into their cells of origin and finally counts them resulting in an expression table (Figure 1 F) (Zheng *et al*, 2017). Once reads are assigned to their cell of origin, a quality control analysis is normally performed, in order to filter out inviable cells, doublets, and in general, uninformative cells (revised in detail in Chapter 1).

An important step in the analysis of single-cell data is the grouping of cells into clusters, depending on the similarity of their transcriptomic profiles. Clusters are the base to execute different statistical tests, like differential expression analyses and marker gene inference. The cell clusters are characterized as cell populations and potentially identified as a distinct cell type, or cell state, based on the genes they express. Clustering of single cells normally relies on graph-based methods implementing community detection algorithms (Freytag *et al*, 2018; Duò *et al*, 2018). The graphs and relationship between the cells are not inferred directly from the gene expression levels (Blondel *et al*, 2008). As each cell is measured for thousands of variables (genes), the cells occupy an N - dimensional space, N referring to the number of genes measured in any given experiment. In order to make better use of the data, this is transformed using dimensionality reduction methods (Heimberg *et al*, 2016). The most popular dimensionality reductions is principal component analysis (PCA), and it is used by many algorithms as the basis to infer cell clusters (Luecken & Theis, 2019).

Several new technologies and alternatives have been developed in recent years, to overcome the limitations of single-cell RNA-seq. For example, since certain cells are not mononuclear, some are extremely tough to dissociate, and some samples impossible to obtain as live tissue, an alternative to single-cell sequencing is single-nuclei RNA sequencing (snRNA-seq) (Habib *et al*, 2017, 2016). snRNA-seq follows the same principle to capture mRNA as implemented in scRNA-seq, but it works on isolated nuclei. Nuclei extraction allows access to frozen samples and overcomes certain dissociation challenges. Nonetheless, as transcripts are obtained only from within the nucleus, and all the cytoplasmic transcripts are removed, snRNA-seq results in even more sparse data than scRNA-seq (Ding *et al*, 2020).

Another shortcoming of scRNA-seq is that positional information is lost in the cell dissociation process. While previous knowledge, stemming mostly from *in situ*

hybridization analyses, can give insight about the general position of cells expressing a particular gene, the absolute and relative position of each cell is unknown. To overcome this problem, methods to perform spatially resolved single-cell transcriptomics are also becoming widely used (Liao *et al*, 2020). These methods rely either on sequential fluorescence *in situ* hybridization (Eng *et al*, 2019) or barcoded beads arrayed on the solid surface of a slide (Rodriques *et al*, 2019). In both methods tissue is sectioned and placed in a slide, where the spatial information of transcripts is recorded. While not truly capturing spatial information at single-cell resolution, the results have been shown to be comparable with scRNA-seq.

Aside from the study of mRNA, single-cell analyses extend into other molecular territories. DNA sequencing from single cells has been successfully applied to reconstruct cell lineages by using the patterns of single-nucleotide variants (Lodato *et al*, 2015). The study of chromatin states like chromosome conformation (Nagano *et al*, 2013), chromatin accessibility (Buenrostro *et al*, 2015) and methylation (Smallwood *et al*, 2014) is also gaining popularity among molecular studies. Protein analyses, unlike the previously mentioned, allow so far the study of a few target proteins across thousands of cells (Bendall *et al*, 2011). The combination of several of these analyses in a practical and standardized single experiment, to obtain different levels of information from the same cells is an ongoing effort (Stuart & Satija, 2019), which has the potential to link genotype and phenotype at different molecular levels.

These recent advances in single-cell sequencing technologies offer great advantages for the study of complex tissues, by revealing their fine-scale cellular composition. The possibility to study certain cell types, without the need of *a priori* cellular markers or reporter transgenes has also extended the number of organisms we can analyze with such fine-grain resolution. Now, non-traditional model organisms can also be studied at the same level as classic model organisms, allowing for comparative studies spanning big evolutionary scales (Sebé-Pedrós *et al*, 2018) . The use of these technologies to construct so-called cell atlases (Regev *et al*, 2017; Cao *et al*, 2019; Farrell *et al*, 2018) has deepened our understanding of organs, systems and even whole organisms across the tree of life. And even more relevant for the work presented in this thesis, the chance to dissect patterning processes in space and time, has opened up a plethora of opportunities to perform novel studies using new and classic developmental and patterning models. Known and unknown signaling centers, proliferating cells, maturing tissue and cell fate decisions, among others, are now open to be revisited again, to describe their underlying molecular dynamics at the single-cell level.

Cell Types and Gene Regulatory Programs in Development and Evolution

In a complex concert of cell division, growth, differentiation and death, a single totipotent cell generates all the cells that build an adult organism. Therefore, almost all cells from an individual organism share a near-identical genome (Frumkin *et al*,

2005). But, although their genome is uniform throughout, adult multicellular organisms are not a mass of identical cells. Multicellular organisms subdivide their vital tasks to different cell types, from structural (like bone-forming cells) to highly dynamic (like neurons). Through a differentiation process, groups of cells with the same genome display expression regulation of different sets of genes (Arendt *et al*, 2016), which gives them characteristic physiological and morphological features. Thus, understanding the different gene regulatory programs that drive the emergence of distinct cell types – correctly patterned in both space and time – is a central goal of developmental and evolutionary biology (Marioni & Arendt, 2017).

Cell Types

Since the first observations of cells by van Leeuwenhoek and Hooke, it became clear that there were different types of them. Although, as more and more cell types were described, it became more difficult to define what a "cell type" actually is. Cell types were first classified based on their morphology and function, initially using early microscopy, later sophisticated imaging techniques, physiological measurements, molecular information or combinations thereof (Kepecs & Fishell, 2014; Zeng & Sanes, 2017). Different cells types in one organism can be quite distinct from one another, but may show remarkable similarities to certain cell types in other organisms, even among distantly related species. These similarities are sometimes due to the fact that cell types are homologous across species, meaning that they have a common origin in a shared ancestor species (Arendt, 2008; Wagner, 2014). This points out to cell types being an evolutionary unit, which poses several considerations regarding the definition and classification of cell types.

One particular perspective puts forward the idea of cell types being discrete evolutionary entities, related to each other by descent, but in completely independent evolutionary trajectories. In this framework (Arendt *et al*, 2016; Wagner, 2014), cell types are defined by having a specific set of gene regulators, termed core regulatory complex (CoRC). The CoRC, in an interplay with chromatin accessibility (Cusanovich *et al*, 2018), allows a restricted and particular access to the genome to result in the execution of a specific gene expression program. CoRCs are composed of a set of regulatory genes and their cooperative interactions, and define terminally differentiated cell types (Arendt *et al*, 2016; Hobert, 2016). If the CoRC changes, then a change in gene expression, morphology or physiology could be expected. This regulatory mechanism that maintains distinct expression profiles is evolutionary conserved, and allows for recognition of homologous cell types across species.

Seen like this, studying the specification of cell types can be considered along two discrete temporal axes, evolutionary and developmental (Marioni & Arendt, 2017). To understand the evolutionary origin of novel cell types, comparative studies detailing their underlying gene regulatory programs are necessary. By reconstructing CoRCs and other regulatory complexes across species, we can trace the history of cell types, find the evolutionary relationships they hold to each other, and arrive at a phylogenetic tree of cell types (Arendt *et al*, 2016; Sebé-Pedrós *et al*, 2017; Liang *et al*, 2015)

(Figure 2). On the developmental side, studies aim to understand how cell types are generated within an organism. Here, cells also give rise to each other, transitioning through a series of cell differentiation steps (Marioni & Arendt, 2017; Ackermann *et al*, 2005). The resulting lineages – or cell fates – that give rise to the eventually mature cell types are studied to understand how cells go from toti- or multipotent cell states to have a restricted differentiation potential, and finally, to fully mature and differentiated cells (Figure 2).

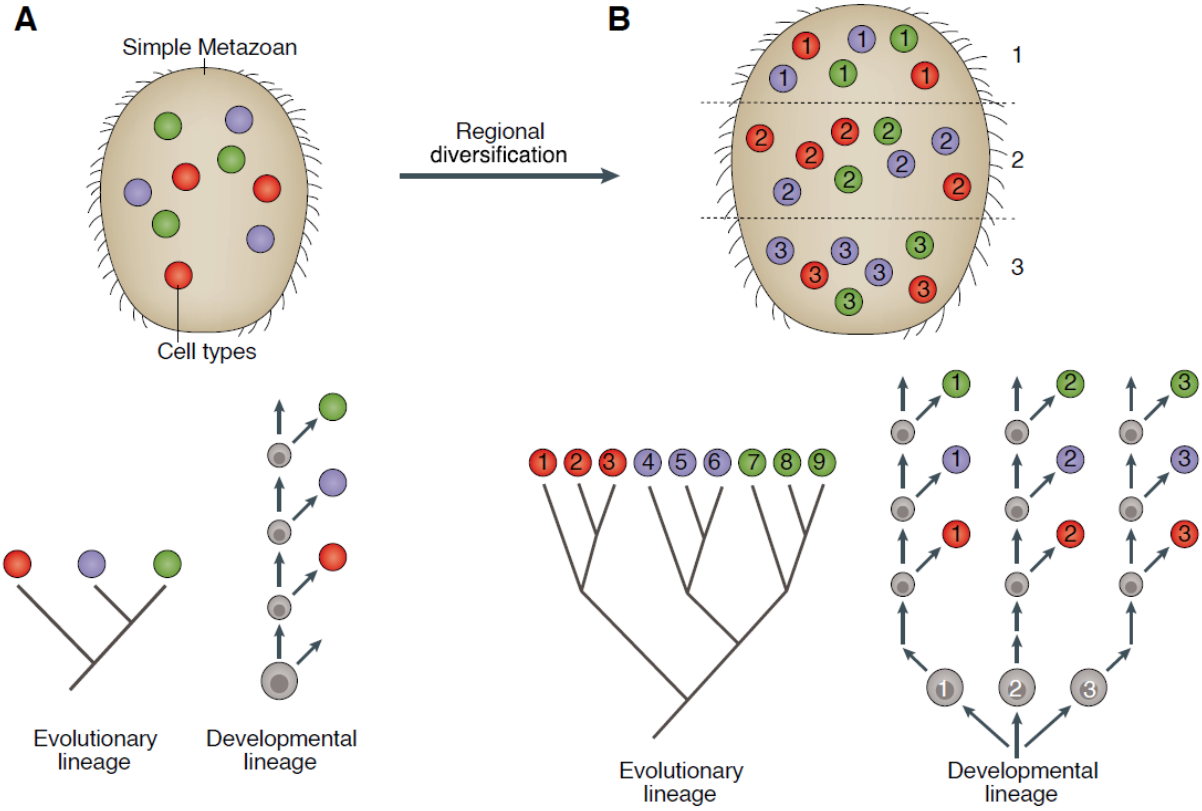


Figure 2 Interrelationship of developmental and evolutionary cell type lineages. Adapted from: The origin and evolution of cell types (Arendt et al, 2016). **A** Ancestral state. Three evolutionarily related cell types are homogenously distributed across the body in a hypothetical simple metazoan, arising from a stem cell-like developmental lineage. **B** Derived state. Cells have diversified regionally, giving rise to region-specific serial sister cell types. Within a region, cells arise from common stem cells so that developmental and evolutionary lineage differ.

The current models of development also states that the regulatory mechanism of the cells must change several times during development, as cells transition from one state into another (Waddington, 1957), these regulatory mechanisms are summarized in a so-called Gene Regulatory Network (GRN) (Thompson *et al*, 2015; Sebé-Pedrós *et al*, 2017; Davidson, 2006), which will be revised later. The regulatory changes must have order and sense in space and time, so the different structures and tissues have a normal development. Timing and a sense of space are fundamental, given that the potential of the cells becomes more and more restricted, as the regulatory and transcriptional programs become more specialized. Cells within a multicellular

organism therefore must rely on information from other cells about their relative position and next differentiation or developmental step.

Patterning and Signaling

To generate complex multicellular organisms, a cell differentiation process is necessary. Crucial to this process is coordination, cells don't differentiate at random positions, times or into fortuitous cell lineages; animals have organized and well-defined tissues, organs and body parts. Tight control of cell differentiation and maturation is achieved via molecular and physical cues, which cells carrying the corresponding receptors or internal signaling machinery can interpret (Perrimon *et al*, 2012). While physical cues (mechanics, temperature, light, etc.) are important and interesting, I will only elaborate further on the molecular signals.

Patterning is the process through which the complexity of an embryo is increased. Instead of generating homogeneous cell groups, embryos start patterning as early as in the morula stage (Barlow *et al*, 1972; Sutherland *et al*, 1990), although oocytes from some species are already asymmetric (Rossant & Tam, 2009). From there, different structures and distinct embryonic regions keep arising, until all organs are formed. All these structures produced during embryo development possess a sense of space and time within the embryo, to develop in the right position. Being three-dimensional entities, developmental axes are essential for the patterning of embryos. Axes are planes along and from which embryonic regions develop and differentiate.

The first developmental axis is defined in the oocytes before the fertilization, and following the entry of the sperm different axes are consequentially created (Rossant & Tam, 2009; Stower & Srinivas, 2018). Axes are maintained not only across the whole embryo, but also within the different developing structures. Axes are not physical entities, but created by gradients of signaling molecules, called morphogens, which can be detected and interpreted by the responding cells (reviewed in: Briscoe & Small, 2015). This provides cells and embryonic regions with boundaries, distance scalars and polarity. The morphogens are produced within the embryo itself, from so-called signaling centers (reviewed in: Albert Basson, 2012). As the embryo grows and new structures are created, new dedicated signaling centers arise in order to keep the positional information throughout.

Morphogen gradients created by the signaling centers can be found in different combinations, with primary and secondary gradients arising, generating complicated grids of relative position. These signaling molecules have different diffusion properties, depending on the tissue they are present in, adding a second layer of signaling modulation (reviewed in: Rogers & Schier, 2011; Müller *et al*, 2013). The resulting grid of differential concentration of the secreted morphogens is what instructs the cells about their position and consequent fate. Secreted signaling factors and their corresponding intracellular responding molecular machinery, or signaling pathways, are well conserved among species, and their function is also repeatedly used by forming tissues in a variety of anatomical structures. Signaling pathways involved in development can be broadly classified in the following groups groups:

Notch, EGF, WNT, cytokine JAK-STAT, bone morphogenetic protein (BMP) or transforming growth factor - β (TGF- β), hedgehog (HH), Hippo, JNK, NF-κB, retinoic acid receptor, and fibroblast growth factor (FGF) (Perrimon *et al*, 2012).

Signaling centers and the gradients defining the patterning axes constitute the first component of the signaling process. As with any signal, its purpose is futile unless it can be received and interpreted. Therefore, the second component of signaling resides within the responding cells themselves. In order to detect and receive any sort of signaling molecule, cells need to have the correspondent receptors exposed on their surface. Secreted signaling factors bind to trans-membrane proteins in a ligand - receptor fashion, and the signal is thereafter transduced into the cell. Binding of the signaling ligands to its receptor can also be modulated. Binding can be hindered by the presence of binding antagonists, or enhanced by co-receptors. The binding of the ligand to the receptor, produces molecular or structural changes that generate a cascade of biochemical reactions that ultimately cause a response from the cell (e.g. BMP signaling reviewed in: Sieber *et al*, 2009). This whole process is of course not only linear, and the response to a morphogen signal can consist of another morphogen signal, creating a feedback loop (Nguyen & Kholodenko, 2016; Bénazet *et al*, 2009). Among the different kinds of responses, changes in the gene expression are the most relevant for the work presented in this thesis.

Control of gene expression, explored below in detail, is in this sense a central process during signaling and patterning. Signaling centers must have distinct gene expression that allows the cells to secret the correct signaling factors in the precise amounts. Cells that receive the signal express the set of trans-membrane proteins that allow the recognition of specific signals, and overlook others. Signaling pathways within the cells are also composed of specifically transcribed proteins and enzymes. Different combinations of components in a certain pathway can result on attenuation (Toyoshima *et al*, 2012), amplification, integration (Schmitz *et al*, 2011) or blocking (Logue & Morrison, 2012) of the signal. As a result of the signaling pathway, sets of proteins regulate the transcriptional response to the signal received.

Understanding gene expression regulation in the context of patterning is therefore challenging and essential to understand many key developmental processes. To comprehend how patterning is realized, we should see gene regulation as cause and consequence, as complex networks of regulation that result in cell fate definitions. Studying only transcriptional data is not optimal, given all the different layers of signaling modulation and expression regulation that are not directly observed measuring gene transcription. Nonetheless, gene expression dynamics offer a glimpse into the intricate relationships between genes, and transcriptional changes are evidence of the non-observed regulatory processes.

Differential Gene Regulation in Eukaryotes

Gene regulation in eukaryotic cells – i.e. defining how much of a given gene product is being produced in a given cell, at a particular time point – is a multi-layered process with a plethora of molecular control mechanisms at different levels along its way.

These levels of regulation include: transcriptional regulation - including initiation of transcription, pausing and elongation, as well as termination -, co- and post-transcriptional modifications of the resulting mRNA transcripts, mRNA transport, regulation of ribosomic translation, and the degradation of mRNA or its resultant protein product. Given the scope of this work, I will briefly discuss only chromatin modifications and transcriptional regulation.

Eukaryotic genes consist of two main parts, the open reading frame consisting of a collection of exons and the regulatory sequences (including - in this sense - introns and UTRs). Regulatory sequences, aside from introns, can be divided in proximal and distal elements. Proximal elements are the 5' and 3' untranslated regions (UTR), one or multiple alternative core promoter elements and promoter-proximal regulatory elements. The core promoter includes the transcription starting site (TSS) enclosed in the initiator element, and TATA box sequence (5'-TATAAA-3'), upstream TFIIB recognition element, downstream promoter element, motif ten element, downstream promoter element, and downstream core element (Kadonaga, 2012). Elements surrounding the TSS and the TATA box are also sometimes referred to as proximal regulatory sequences. Distal regulatory elements, on the other hand, are not adjacent to any of the aforementioned components and can vary in number, size and distance from the main gene body (Kim & Shiekhattar, 2015; Dao *et al*, 2017).

Proximal and distal regulatory elements are not very clearly defined in terms of genomic position. Both contain binding sites for specific transcription factors and chromatin modifying enzymes that regulate transcription in *cis* (Kim & Shiekhattar, 2015). Proximal and distal regulatory elements only differ, as their name suggests, in the position they have relative to the main gene body. While the position and definition of proximal regulatory elements is unclear and debated; they are close (~200-400 bp), adjacent to, or around the TSS. On the other hand, distal regulatory elements can be several kilobases or even megabases (Lettice *et al*, 2003), away from the main gene body. The regulatory sequences found in the proximal and distal elements can be subdivided in enhancers, silencers or insulators, depending on their specific functions on transcriptional control (reviewed in: Riethoven, 2010). It has been demonstrated that changes in regulation, and specifically in *cis* regulatory elements like enhancers, can be main driver of phenotypic diversity and phenotypic novelties (Gompel *et al*, 2005; Adachi *et al*, 2016; Kvon *et al*, 2016; Thompson *et al*, 2018).

Regulatory elements within the DNA sequence by themselves provide a rather static layer of regulation, the presence or absence of the binding sites doesn't change along development or across cell types. Nonetheless their accessibility provides a dynamic layer of regulation. Since eukaryotic cells need to carry up to 150 Gbp of genetic information in their nuclei (Pellicer *et al*, 2010), their genome is tightly packed in chromatin. Compacting and protecting DNA, chromatin also makes it inaccessible and provides the setting for another level of transcriptional regulation. Chromatin modifications change DNA accessibility and can be of structural or chemical nature. Structural modifications ensure that different DNA regions are in physical contact

when necessary, for example, for distal enhancers to interact with gene promoters (reviewed in: Schoenfelder & Fraser, 2019). The most notable examples of chemical modifications are DNA cytosine methylation and the methylation or acetylation of lysine groups in the tails of the histones (reviewed in: Kouzarides, 2007).

Methylation of DNA occurs mainly, but not only, in the so-called CpG islands, which are DNA regions with unusual occurrence of the “CG” sequence (reviewed in: Schübeler, 2015). Two of the most studied histone modifications in evolutionary and developmental biology are acetylation and methylation of Lysine residues on histone 3. Different parts of the genes are associated with different modes of modifications. Regulatory histone modifications within the gene main body occurs via the tri-methylation of lysine 36 and di-methylation of lysine 79 (H3K36me3 and H3K79me2) which are responsible for gene activation. CpG islands are common in promoters, and their methylation silences the associated gene. H3K9Ac, H3K27Ac and H3K4me3 are associated with active promoters, while H3K27me3 generally represses them. Distal regulatory sequences, or enhancers, show similar modes of regulation, but here the most common activation marker is H3K4me1 (reviewed in: Kouzarides, 2007; Zhou *et al*, 2011) (Figure 3).

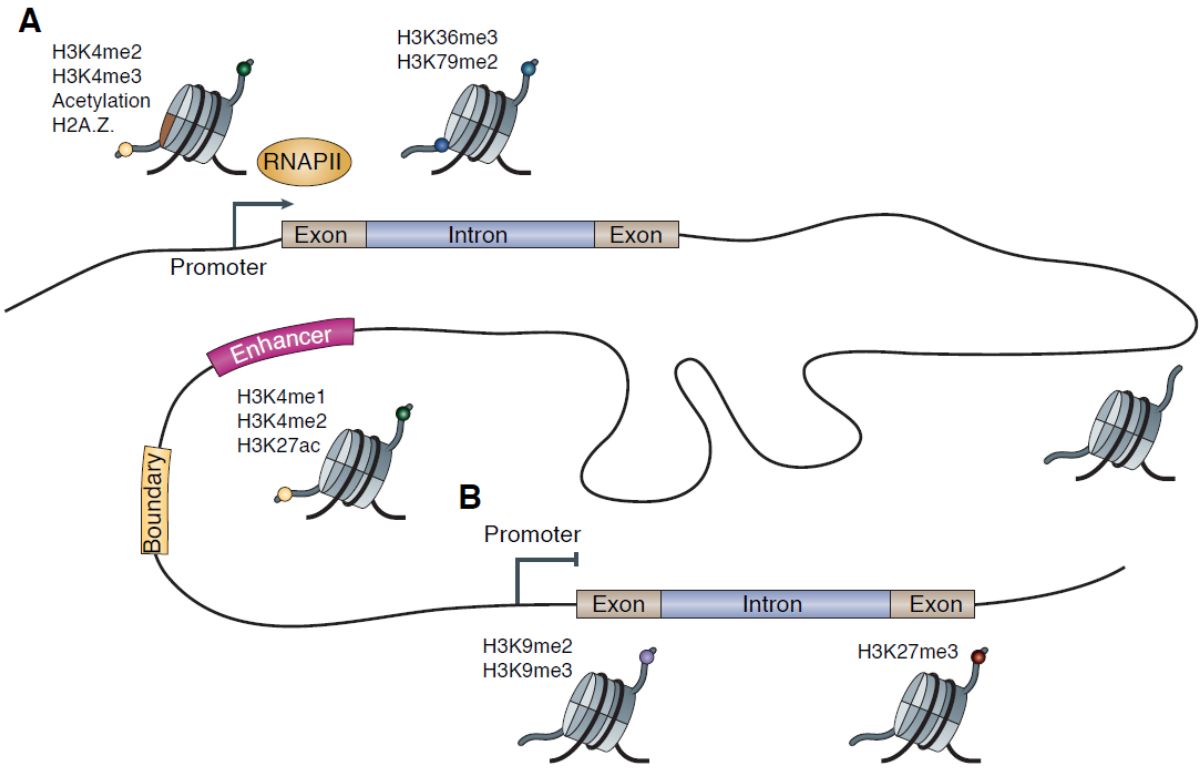


Figure 3 Histone modifications demarcate functional elements in mammalian genomes. Adapted from Charting histone modifications and the functional organization of mammalian genomes (Zhou et al, 2011). Promoters, gene bodies, an enhancer and a gene boundary element are indicated on a schematic genomic region. **A** Active promoters are commonly marked by histone H3 lysine 4 dimethylation (H3K4me2), H3K4me3, acetylation (ac), and H2A.Z. transcribed regions are enriched for H3K36me3 and H3K79me2. Enhancers are relatively enriched for H3K4me1, H3K4me2, and H3K27ac. **B** Repressed genes may be located in large domains of H3K9me2 and/or H3K9me3 or H3K27me3. RNAPII: RNA Polymerase II.

While the accessibility of regulatory sequences can be controlled via chromatin modifications, the regulatory DNA sequences don't act directly on transcription. Yet another layer of regulation comes from the proteins which bind to these regulatory sequences. These proteins, called transcription factors (TF) have binding domains which recognize and bind to DNA sequences with different degrees of specificity. Transcription factors can either cause the up or down regulation of the gene they affect (reviewed in: Spitz & Furlong, 2012). In this scenario, transcription of a gene depends on chromatin accessibility of the promoter, accessibility and proximity of one or several enhancer sequences (or other regulatory elements), and the presence of the corresponding transcription factors. This seemingly simple scenario is complicated by the fact that transcription factors, and all other proteins necessary for transcription must be first expressed themselves, and are also subject their own regulation conditions. Moreover, certain regulatory processes depend on the concentration of certain regulatory factors (Johnson *et al*, 2006; Kamath *et al*, 2008) In this sense, the relations and interactions between transcription factors and their targets is evolutionary conserved, they are evolutionary robust, yet showing potential for evolvability into other robust interactions (Payne & Wagner, 2014; Payne *et al*, 2014).

In this sense, it's possible to explore gene expression regulation beyond the scope of a single gene. And in this case we must understand gene regulation as the result of a web of interactions between different genes and other cellular components. In this intricate web of interactions we can find 1-to-1 relationships, master modulators, feedback loops and antagonisms, among others. The collection of these interactions is called a Gene Regulatory Network, and it comprises the regulatory steps that the production of a protein of group of proteins require (Davidson, 2006). GRNs can become quite complex as more and more gene interactions are discovered and variations in the interactions drive evolutionary changes (Erwin & Davidson, 2009; Peter & Davidson, 2017; Thompson *et al*, 2015).

Given the considerations described in previous sections, one could think that GRNs are an essential part of cell type identity. But the regulatory conditions of fully differentiated and developing maturing cells are substantially different. Given the intricacy of spatiotemporal coordination that is required during development, GRNs exhibit very complex hierarchy (Erwin & Davidson, 2009). On the other hand, fully differentiated cells seem to rely on a small set of transcription factors to control the majority of their cell-type specific transcriptional program (Graf & Enver, 2009). On these grounds, the term of Core Regulatory Complex, mentioned before, has been proposed to define the set of transcription factors and their cooperative interactions that define fully differentiated cell types (Arendt *et al*, 2016).

The dynamic regulation, through the modes here described, and many more not discussed, allows cells to transition from one state to the other while carrying the same genomic information. Much the variability we observe during development across species is due to changes that modify the genetic expression of signals and

their elicited responses, and not mutations in the coding sequences of genes (Wray, 2007; Rubinstein & de Souza, 2013). Exploring the changes in the regulation of gene expression is therefore a major focus of evolutionary developmental biology (Brakefield, 2011). Studying changes in chromatin accessibility, chromatin modifications and inferring differential regulation and GRNs from differential expression are the tools we possess to better understand how organisms develop from a single cell into the variation we observe in nature.

Aims of this Thesis

In this work, I intend to analyze several aspects of development at single-cell level aided by the chicken embryo as a model, and making use of high-throughput single-cell RNA-seq. The different developmental processes I will explore concern cell diversity, signaling centers, cell maturation, patterning, cell fate decisions, phenotypic convergence, and evolution of gene expression programs. As a first step, I present the work necessary to implement a novel technique like scRNA-seq on the study of chicken development. To our knowledge, we were the first to make use of this technique on chicken embryo samples, and for this reason, I describe several adaptations and optimizations of different methods, analyses and resources I performed, in order to efficiently work with chicken samples. This is an important part of this thesis, because while scRNA-seq offers the possibility to process a vast diversity of samples, bioinformatics analyses are still depending on the availability and quality of genomic references and resources.

With our bioinformatics tools and resources established, we set out to describe the cell populations which are present and active in a paradigm of developmental patterning research: the development of the tetrapod limb. We present a single-cell transcriptomic atlas of the developing chicken hind limb. There, we describe well known structures and cell types, as well as previously overlooked cell populations. We also describe the substructure within large cell populations, like the non-skeletal connective tissue. We link scRNA-seq data and spatially resolved bulk RNA-seq to distinguish the different interdigits. And lastly we infer changes in regulatory programs that result in a change of co-expression modules during the maturation of the chondrocytes.

With this atlas of the developing limb at single-cell resolution, we aimed to explore how cells know exactly where and when to make a cell-fate decision. We looked at the patterning of the digits, chains of phalanges and joints which show outstanding variation across species, to address this question. Here, during a distally-driven growth, groups of cells make the sequential decision to become either phalanx-forming chondrocytes or joint-forming chondrocytes. Genes characterizing both populations are well-known, and the molecular events leading to further maturation have been well documented in the past years. Nonetheless, the cell-fate decision process is still a mystery, as well as the sets of genes that initiate the different transcriptomic programs. Making use of our single-cell data, we aimed to reconstruct this process *in silico* using bioinformatics tools. Using progenitor cells and both kinds

of chondrocytes, we conducted an analysis on the transcriptomic dynamics. we looked for transcriptomic changes occurring prior to the cell-fate bifurcation, as well as early changes in expression in both lineages.

After analyzing the patterning process of particular skeletal structures – limb digits –, we explored the emergence of different skeletal components, which undergo different, yet convergent developmental processes. The development of the vertebrate skeleton takes place independently through three different developmental lineages. Bones and cartilage, as well as the cells that build them, are very similar throughout the body, but the axial, cranial and appendicular skeletons arise from very different progenitor populations. These skeletogenic progenitors are the axial mesoderm in the somites, the lateral plate mesoderm in the limb, and the cranial neural crest in the skull. In order to understand how similar transcriptional programs independently arise from the distinct origins and environments, we generated scRNA-seq data of the three differentiation processes. Here, we analyze the different processes independently and then aimed to do a unified analysis of the convergence process. Furthermore, reconstructing the cell maturation processes *in silico*, we try to understand the unique and shared gene expression dynamics that lead to a common transcriptomic program.

In the end, after comparing gene expression signatures, and their similarities across embryonic tissues and developmental origins, I present a test to assess the conservation of the transcriptomic signatures. This test can be implemented across tissues, samples or even across species, to explore the conservation or divergence of transcriptomic programs, the expression relationships between genes. I describe the adaptation made to gene co-expression analyses, originally developed for bulk RNA-seq data, in order to analyze scRNA-seq samples. The framework I propose produces two results. First a co-expression analysis resulting in different modules of co-expression, which reflect the transcriptomic or functional program of different cell populations. And second, a conservation test reflecting the degree of conservation of the calculated modules of co-expression in other samples, reflecting potential changes in the regulation of the module expression. We present the functionality of our approach by comparing developing tissues from chicken and mouse and exploring conservation and divergence of transcriptomic programs found therein.

References

Ackermann C, Dorresteijn A & Fischer A (2005) Clonal domains in postlarval Platynereis dumerilii (Annelida: Polychaeta). *J. Morphol.*

Adachi N, Robinson M, Goolsbee A & Shubin NH (2016) Regulatory evolution of Tbx5 and the origin of paired appendages. *Proc. Natl. Acad. Sci. U. S. A.*

Albert Basson M (2012) Signaling in cell differentiation and morphogenesis. *Cold Spring Harb. Perspect. Biol.*

Arendt D (2008) The evolution of cell types in animals: Emerging principles from molecular studies. *Nat. Rev. Genet.*

Arendt D, Musser JM, Baker CVH, Bergman A, Cepko C, Erwin DH, Pavlicev M, Schlosser G, Widder

S, Laubichler MD & Wagner GP (2016) The origin and evolution of cell types. *Nat. Rev. Genet.* **17:** 744–757

Arnold SJ & Robertson EJ (2009) Making a commitment: Cell lineage allocation and axis patterning in the early mouse embryo. *Nat. Rev. Mol. Cell Biol.*

Barlow P, Owen DA & Graham C (1972) DNA synthesis in the preimplantation mouse embryo. *J. Embryol. Exp. Morphol.*

Bénazet JD, Bischofberger M, Tiecke E, Gonçalves A, Martin JF, Zuniga A, Naef F & Zeller R (2009) A self-regulatory system of interlinked signaling feedback loops controls mouse limb patterning. *Science (80-.).*

Bendall SC, Simonds EF, Qiu P, Amir ED, Krutzik PO, Finck R, Bruggner R V, Melamed R, Trejo A, Ornatsky OI, Balderas RS, Plevritis SK, Sachs K, Pe D, Tanner SD & Nolan GP (2011) Single-Cell Mass Cytometry of Differential. *Science (80-.).*

Blondel VD, Guillaume JL, Lambiotte R & Lefebvre E (2008) Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.*

Brakefield PM (2011) Evo-devo and accounting for darwin's endless forms. *Philos. Trans. R. Soc. B Biol. Sci.*

Briscoe J & Small S (2015) Morphogen rules: Design principles of gradient-mediated embryo patterning. *Dev.*

Buenrostro JD, Wu B, Litzenburger UM, Ruff D, Gonzales ML, Snyder MP, Chang HY & Greenleaf WJ (2015) Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature*

Cao J, Spielmann M, Qiu X, Huang X, Ibrahim DM, Hill AJ, Zhang F, Mundlos S, Christiansen L, Steemers FJ, Trapnell C & Shendure J (2019) The single-cell transcriptional landscape of mammalian organogenesis. *Nature*

Cusanovich DA, Hill AJ, Aghamirzaie D, Daza RM, Pliner HA, Berletch JB, Filippova GN, Huang X, Christiansen L, DeWitt WS, Lee C, Regalado SG, Read DF, Steemers FJ, Disteche CM, Trapnell C & Shendure J (2018) A Single-Cell Atlas of In Vivo Mammalian Chromatin Accessibility. *Cell*

Dao LTM, Galindo-Albarrán AO, Castro-Mondragon JA, Andrieu-Soler C, Medina-Rivera A, Souaid C, Charbonnier G, Griffon A, Vanhille L, Stephen T, Alomairi J, Martin D, Torres M, Fernandez N, Soler E, Van Helden J, Puthier D & Spicuglia S (2017) Genome-wide characterization of mammalian promoters with distal enhancer functions. *Nat. Genet.*

Davidson E (2006) The Regulatory Genome

Ding J, Adiconis X, Simmons SK, Kowalczyk MS, Hession CC, Marjanovic ND, Hughes TK, Wadsworth MH, Burks T, Nguyen LT, Kwon JYH, Barak B, Ge W, Kedaigle AJ, Carroll S, Li S, Hacohen N, Rozenblatt-Rosen O, Shalek AK, Villani AC, et al (2020) Systematic comparison of single-cell and single-nucleus RNA-sequencing methods. *Nat. Biotechnol.*

Duboule D (1994) Temporal colinearity and the phylotypic progression: A basis for the stability of a vertebrate Bauplan and the evolution of morphologies through heterochrony. *Development*

Duò A, Robinson MD & Soneson C (2018) A systematic performance evaluation of clustering methods for single-cell RNA-seq data. *F1000Research*

Eng CHL, Lawson M, Zhu Q, Dries R, Koulena N, Takei Y, Yun J, Cronin C, Karp C, Yuan GC & Cai L (2019) Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+. *Nature*

Erwin DH & Davidson EH (2009) The evolution of hierarchical gene regulatory networks. *Nat. Rev. Genet.*

Farrell JA, Wang Y, Riesenfeld SJ, Shekhar K, Regev A & Schier AF (2018) Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis. *Science (80-.).*

Freytag S, Tian L, Lönnstedt I, Ng M & Bahlo M (2018) Comparison of clustering tools in R for medium-sized 10x Genomics single-cell RNA-sequencing data. *F1000Research*

Frumkin D, Wasserstrom A, Kaplan S, Feige U & Shapiro E (2005) Genomic variability within an organism exposes its cell lineage tree. *PLoS Comput. Biol.*

Gilbert SF (2000) Comparative Embriology. In *Developmental Biology*, Gilbert S (ed) Sunderland (MA): Sinauer Associates

Gilbert SF (2017) Developmental biology, the stem cell of biological disciplines. *PLoS Biol.*

Gompel N, Prud'Homme B, Wittkopp PJ, Kassner VA & Carroll SB (2005) Chance caught on the wing: cis-regulatory evolution and the origin of pigment patterns in Drosophila. *Nature*

Graf T & Enver T (2009) Forcing cells to change lineages. *Nature*

Habib N, Avraham-Davidi I, Basu A, Burks T, Shekhar K, Hofree M, Choudhury SR, Aguet F, Gelfand E, Ardlie K, Weitz DA, Rozenblatt-Rosen O, Zhang F & Regev A (2017) Massively parallel single-nucleus RNA-seq with DroNc-seq. *Nat. Methods*

Habib N, Li Y, Heidenreich M, Swiech L, Avraham-Davidi I, Trombetta JJ, Hession C, Zhang F & Regev A (2016) Div-Seq: Single-nucleus RNA-Seq reveals dynamics of rare adult newborn neurons. *Science (80-.).*

Hagemann-Jensen M, Ziegenhain C, Chen P, Ramsköld D, Hendriks GJ, Larsson AJM, Faridani OR & Sandberg R (2020) Single-cell RNA counting at allele and isoform resolution using Smart-seq3. *Nat. Biotechnol.*

Heimberg G, Bhatnagar R, El-Samad H & Thomson M (2016) Low Dimensionality in Gene Expression Data Enables the Accurate Extraction of Transcriptional Programs from Shallow Sequencing. *Cell Syst.*

Hobert O (2016) Terminal Selectors of Neuronal Identity. In *Current Topics in Developmental Biology*

Horder T (2010) History of Developmental Biology. In *Encyclopedia of Life Sciences*

Islam S, Kjällquist U, Moliner A, Zajac P, Fan JB, Lönnerberg P & Linnarsson S (2011) Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Res.*

Johnson KD, Kim S Il & Bresnick EH (2006) Differential sensitivities of transcription factor target genes underlie cell type-specific gene expression profiles. *Proc. Natl. Acad. Sci. U. S. A.*

Kadonaga JT (2012) Perspectives on the RNA polymerase II core promoter. *Wiley Interdiscip. Rev. Dev. Biol.*

Kamath MB, Houston IB, Janovski AJ, Zhu X, Gowrisankar S, Jegga AG & DeKoter RP (2008) Dose-dependent repression of T-cell and natural killer cell genes by PU.1 enforces myeloid and B-cell identity. *Leukemia*

Kepecs A & Fishell G (2014) Interneuron cell types are fit to function. *Nature*

Kim TK & Shiekhattar R (2015) Architectural and Functional Commonalities between Enhancers and Promoters. *Cell*

Klein AM, Mazutis L, Akartuna I, Tallapragada N, Veres A, Li V, Peshkin L, Weitz DA & Kirschner MW (2015) Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*

Kouzarides T (2007) Chromatin Modifications and Their Function. *Cell*

Kvon EZ, Kamneva OK, Melo US, Barozzi I, Osterwalder M, Mannion BJ, Tissières V, Pickle CS, Plajzer-Frick I, Lee EA, Kato M, Garvin TH, Akiyama JA, Afzal V, Lopez-Rios J, Rubin EM, Dickel DE, Pennacchio LA & Visel A (2016) Progressive Loss of Function in a Limb Enhancer during Snake Evolution. *Cell*

Lettice LA, Heaney SJH, Purdie LA, Li L, de Beer P, Oostra BA, Goode D, Elgar G, Hill RE & de Graaff E (2003) A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Hum. Mol. Genet.*

Liang C, Forrest ARR & Wagner GP (2015) The statistical geometry of transcriptome divergence in cell-type evolution and cancer. *Nat. Commun.*

Liao J, Lu X, Shao X, Zhu L & Fan X (2020) Uncovering an Organ's Molecular Architecture at Single-Cell Resolution by Spatially Resolved Transcriptomics. *Trends Biotechnol.*

Lodato MA, Woodworth MB, Lee S, Evrony GD, Mehta BK, Karger A, Lee S, Chittenden TW, D'Gama AM, Cai X, Luquette LJ, Lee E, Park PJ & Walsh CA (2015) Somatic mutation in single human neurons tracks developmental and transcriptional history. *Science (80-.).*

Logue JS & Morrison DK (2012) Complexity in the signaling network: Insights from the use of targeted inhibitors in cancer therapy. *Genes Dev.*

Luecken MD & Theis FJ (2019) Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol. Syst. Biol.*

Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, Tirosh I, Bialas AR, Kamitaki N, Martersteck EM, Trombetta JJ, Weitz DA, Sanes JR, Shalek AK, Regev A & McCarroll SA (2015) Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161:** 1202–1214

Marioni JC & Arendt D (2017) How Single-Cell Genomics Is Changing Evolutionary and Developmental Biology. *Annu. Rev. Cell Dev. Biol.* **33:** 537–553

Mayr U, Serra D & Liberali P (2019) Exploring single cells in space and time during tissue development, homeostasis and regeneration. *Dev.*

Müller P, Rogers KW, Yu SR, Brand M & Schier AF (2013) Morphogen transport. *Dev.*

Nagano T, Lubling Y, Stevens TJ, Schoenfelder S, Yaffe E, Dean W, Laue ED, Tanay A & Fraser P (2013) Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature*

Nguyen LK & Kholodenko BN (2016) Feedback regulation in cell signalling: Lessons for cancer therapeutics. *Semin. Cell Dev. Biol.*

Nguyen QH, Pervolarakis N, Nee K & Kessenbrock K (2018) Experimental considerations for single-cell RNA sequencing approaches. *Front. Cell Dev. Biol.*

Payne JL, Moore JH & Wagner A (2014) Robustness, evolvability, and the logic of genetic regulation. In *Artificial Life*

Payne JL & Wagner A (2014) The robustness and evolvability of transcription factor binding sites. *Science (80-.).*

Pellicer J, Fay MF & Leitch IJ (2010) The largest eukaryotic genome of them all? *Bot. J. Linn. Soc.*

Perrimon N, Pitsouli C & Shilo B-Z (2012) Signaling Mechanisms Controlling Cell Fate and Embryonic Patterning. *Cold Spring Harb. Perspect. Biol.* **4:** a005975–a005975

Peter IS & Davidson EH (2017) Assessing regulatory information in developmental gene regulatory networks. *Proc. Natl. Acad. Sci. U. S. A.*

Picelli S, Faridani OR, Björklund ÅK, Winberg G, Sagasser S & Sandberg R (2014) Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.*

Ramsköld D, Luo S, Wang YC, Li R, Deng Q, Faridani OR, Daniels GA, Khrebtukova I, Loring JF, Laurent LC, Schroth GP & Sandberg R (2012) Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nat. Biotechnol.*

Regev A, Teichmann S, Lander E, Amit I, Benoist C, Birney E, Bodenmiller B, Campbell P, Carninci P & Enard W (2017) Science Forum: The Human Cell Atlas. *Elife*

Richardson MK (1999) Vertebrate evolution: The developmental origins of adult variation. *BioEssays*

Riethoven J-JM (2010) Regulatory Regions in DNA: Promoters, Enhancers, Silencers, and Insulators. In *Computational Biology of Transcription Factor Binding*, Ladunga I (ed) pp 33–42. Totowa, NJ: Humana Press

Rodriques SG, Stickels RR, Goeva A, Martin CA, Murray E, Vanderburg CR, Welch J, Chen LM, Chen F & Macosko EZ (2019) Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science (80-.).*

Rogers KW & Schier AF (2011) Morphogen gradients: From generation to interpretation. *Annu. Rev. Cell Dev. Biol.*

Rossant J & Tam PPL (2009) Blastocyst lineage formation, early embryonic asymmetries and axis patterning in the mouse. *Development*

Rubinstein M & de Souza FSJ (2013) Evolution of transcriptional enhancers and animal diversity. *Philos. Trans. R. Soc. B Biol. Sci.*

Schmitz ML, Weber A, Roxlau T, Gaestel M & Kracht M (2011) Signal integration, crosstalk mechanisms and networks in the function of inflammatory cytokines. *Biochim. Biophys. Acta - Mol. Cell Res.*

Schoenfelder S & Fraser P (2019) Long-range enhancer–promoter contacts in gene expression control. *Nat. Rev. Genet.*

Schübeler D (2015) Function and information content of DNA methylation. *Nature*

Sebé-Pedrós A, Chomsky E, Pang K, Lara-Astiaso D, Gaiti F, Mukamel Z, Amit I, Hejnol A, Degnan BM & Tanay A (2018) Early metazoan cell type diversity and the evolution of multicellular gene regulation. *Nat. Ecol. Evol.*

Sebé-Pedrós A, Degnan BM & Ruiz-Trillo I (2017) The origin of Metazoa: A unicellular perspective. *Nat. Rev. Genet.*

Sieber C, Kopf J, Hiepen C & Knaus P (2009) Recent advances in BMP receptor signaling. *Cytokine Growth Factor Rev.*

Smallwood SA, Lee HJ, Angermueller C, Krueger F, Saadeh H, Peat J, Andrews SR, Stegle O, Reik W & Kelsey G (2014) Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nat. Methods*

Spitz F & Furlong EEM (2012) Transcription factors: From enhancer binding to developmental control. *Nat. Rev. Genet.*

Stower MJ & Srinivas S (2018) The Head's Tale: Anterior-Posterior Axis Formation in the Mouse Embryo. In *Current Topics in Developmental Biology*

Stuart T & Satija R (2019) Integrative single-cell analysis. *Nat. Rev. Genet.*

Sutherland AE, Speed TP & Calarco PG (1990) Inner cell allocation in the mouse morula: The role of oriented division during fourth cleavage. *Dev. Biol.*

Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, Wang X, Bodeau J, Tuch BB, Siddiqui A, Lao K & Surani MA (2009) mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods*

Thompson AC, Capellini TD, Guenther CA, Chan YF, Infante CR, Menke DB & Kingsley DM (2018) A novel enhancer near the pitx1 gene influences development and evolution of pelvic appendages in vertebrates. *Elife*

Thompson D, Regev A & Roy S (2015) Comparative Analysis of Gene Regulatory Networks: From Network Reconstruction to Evolution. *Annu. Rev. Cell Dev. Biol.*

Toyoshima Y, Kakuda H, Fujita KA, Uda S & Kuroda S (2012) Sensitivity control through attenuation of signal transfer efficiency by negative regulation of cellular signalling. *Nat. Commun.*

Waddington CH (1957) The strategy of the genes. A discussion of some aspects of theoretical biology. With an appendix by H. Kacser. *Strateg. genes A Discuss. some …*

Wagner GP (2014) Homology, genes, and evolutionary innovation

Wray GA (2007) The evolutionary significance of cis-regulatory mutations. *Nat. Rev. Genet.*

Zeng H & Sanes JR (2017) Neuronal cell-type classification: Challenges, opportunities and the path forward. *Nat. Rev. Neurosci.*

Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, Ziraldo SB, Wheeler TD, McDermott GP, Zhu J, Gregory MT, Shuga J, Montesclaros L, Underwood JG, Masquelier DA, Nishimura SY, Schnall-Levin M, Wyatt PW, Hindson CM, Bharadwaj R, et al (2017) Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8:** 1–12

Zhou VW, Goren A & Bernstein BE (2011) Charting histone modifications and the functional organization of mammalian genomes. *Nat. Rev. Genet.*

Chapter 1

OPTIMIZING SINGLE-CELL RNA-SEQ ANALYSIS METHODS FOR THE STUDY OF CHICKEN DEVELOPMENT

Christian Feregrino

Abstract

Single-cell transcriptomic data analysis is becoming more and more frequent as part of biological research. In order to expand the catalogue of organisms and tissues described at single-cell resolution, certain methodological aspects, especially the bioinformatics, must be adapted and optimized. Here, we present methodological adaptations we have made to analyze single-cell transcriptomic data originating from the chicken embryo. Using available bioinformatics toolkits like Drop-seq tools, Cell Ranger, Seurat, R, Monocle and URD in combination with scripts and calculations developed by us, we processed and analyzed several datasets stemming from different structures of the chicken embryo. Our main result is the extension of the 3' UTR tracks of the chicken genome annotation, which corrected an incompatibility with Chromium 10x Genomics scRNA-seq approaches. We also present our quality control filtering approach with thresholds calculated relative to the data itself, and a comparison of different pseudotime methods, among other details of our research. In general, we propose a framework for the analysis of scRNA-seq of chicken origin. We think our optimizations will prove helpful for other research groups planning to perform similar studies in chicken, or other organisms with poorly annotated genomes.

Introduction

The development of the chicken embryo was already studied by Aristotle (Horder, 2010) and has served as a model since the beginning of scientific embryological studies in the 18th century (Gilbert, 2000). The accessibility to the embryo itself, ease of experimental manipulation and developmental characteristics of the chicken embryo – relatively simple incubation equipment and relatively short development time –make it an optimal system to observe, describe, test and disturb different developmental processes. With such a long history and extensive body of knowledge, scientist constantly rely on the chicken embryo and make use of the newest methods and analyses to deepen our understanding of vertebrate development. One of the developmental phenomena which keeps attracting the attention of scientists, is the patterning of the tetrapod limb. For decades, the scientific community has used different methodologies to study the chicken limb development to try to understand the intricate processes that shape the complicated morphological designs we observe in nature (Petit *et al*, 2017; Zeller *et al*, 2009).

In recent years, a breakthrough concerning genomic methods was achieved, when analyses at single-cell resolution were made available on a high-throughput fashion (Macosko *et al*, 2015; Klein *et al*, 2015). First developed and optimized to be used with human and mouse tissue, single-cell RNA-seq (scRNA-seq) was quickly adapted in other model organisms. While data production on the laboratory bench is relatively easy to adjust in order to obtain sequences from different tissues and species (Nguyen *et al*, 2018), the kind and amount of data acquired presents analytical and statistical challenges as well as opportunities that are still being tackled to this day (Lähnemann *et al*, 2020). While some of these challenges are common to all scRNA-seq experiments, the use of some tissues and / or species result in very specific bench and bioinformatics problems. I will briefly discuss some aspects of the data production and data analysis of scRNA-seq experiments.

Massively parallel scRNA-seq is based on microfluidic technology, and was first developed as “home-made” laboratory bench set-ups (Macosko *et al*, 2015; Klein *et al*, 2015). Drop -seq, one of these methods, takes advantage of a microfluidic device that controls the inflow of several components: microbeads coated with DNA strands, dissociation buffer, a suspension of single cells and an emulsion agent that creates nanodroplets. Most of the reactions of the experiment occur then inside of these nanodroplets. In order to obtain faithful results that effectively represent one cell per droplet, the concentration of each component must be tightly controlled. In a standard experiment, the microbeads concentration and inflow is set so that only some droplets contain a single bead and many remain empty. In order to reduce the probability of sampling two cells at a time, the concentration and inflow of the cells is even lower, to have then a few of the droplets with one bead and one cell (Macosko *et al*, 2015).

The nucleotide-coated beads, or barcoded hydrogel particles in the case of other set-ups, used in scRNA-seq experiments have several functions. The primary function of the DNA strands coupled to the beads is to hybridize their poly(T) tail to the poly(A)

tail of the mRNA transcripts. The nucleotide strands are also barcoded, uniquely for each bead, to later identify from which bead – and therefore, from which cell – each transcript was sequenced. The nucleotide strands are additionally barcoded with a Unique Molecular Identifier (UMI) (Macosko *et al*, 2015; Klein *et al*, 2015). UMIs are unique for each of the strands in one bead, and are therefore useful to identify each single transcript that is sequenced to avoid PCR amplification bias (Islam *et al*, 2014).

A few years after the presentation of the droplet based scRNA-seq methods, the set-up was standardized and became a commercial product, distributed by 10x Genomics. While Drop-seq was the most used scRNA-seq droplet method for a couple of years, today the 10x Genomics system is the dominant agent of the commercial market. There are some technical differences between Drop-seq and the 10x system. Among the differences, in the Chromium assays we have DNA-barcoded gel beads (like in InDrop (Klein *et al*, 2015)), as well as two library preparation steps: enzymatic fragmentation and size selection (Zheng *et al*, 2017).

Regarding the data analysis of single-cell transcriptomic data, a crucial element is the quality of the genome and transcriptome annotation to which the mRNA fragments are mapped to (Zhao & Zhang, 2015). When working with well-established model organisms for genomic studies, like mice or fruit fly, this is not a major problem since the genomes have a very well curated genomic sequence and transcriptome annotation. While there have been great efforts to obtain a high quality genome and transcriptome for the chicken (Schmid *et al*, 2015), there is still a considerable difference in its annotation quality compared to mouse, fly or human genomes. Among the many challenges that arise during genome annotation, the annotation of UTRs is a particularly difficult element. While internal exons can be identified by reads that span splice junctions, the boundaries of terminal exons – containing UTRs – are only recognized as the position at which sequencing coverage decreases. Transcriptome assembly tools have different approaches to then infer a probable exon – and therefore UTR – boundary (Shenker *et al*, 2015).

Although high-throughput scRNA-seq can be performed on either end of the transcripts, 3’ sequencing is the most widely used. Therefore, and because in massively-parallel scRNA-seq transcripts are captured using their poly-A tail, and the fact that sequencing is done on small fragments (between 50 and 100 bp), the sequenced data has a strong 3’ bias. This bias means that a lot of genes are only sequenced in their 3’ UTR. Among many differences between Drop-seq and Chromium 10x Genomics assays, is the way the sequencing libraries are constructed. 10x Genomics, which has discontinued its 5’ approach, has an extra fragmentation step which is not present in Drop-seq. This makes Chromium 10x Genomics data even more 3’ biased than Drop-seq data is. Therefore, annotation of the genome, especially the 3’ UTRs, is a very important element during massively-parallel single-cell transcriptome analyses using Chromium 10x Genomics 3’ sequencing assays.

The bioinformatics pipelines developed specifically to process the raw sequencing data produced during scRNA-seq experiments aim to recover true single cell data

points. In order to disregard data produced by droplets containing no cells, dead cells or environmental RNA, these pipelines make use of a cumulative distribution of RNA reads captured by every bead. The assumption is that reads associated with a cell-containing droplet will make large contributions to a cumulative distribution, and reads associated with a droplet not containing a healthy cell will not substantially add to the cumulative sum. A threshold or “turning point” is calculated and only the “valid cells” are then processed for further downstream analyses (Macosko *et al*, 2015; Zheng *et al*, 2017).

Although UMI count tables resulting from processed raw scRNA-seq data contain only “valid cells”, further quality control filters must be used to ensure meaningful results. In this step of the analyses, putative doublets and remaining low quality or dying cells are removed from the datasets (Luecken & Theis, 2019; Amezquita *et al*, 2020). Moreover, cells dissociated from their tissue, and exposed to unfamiliar elements and conditions can present very different transcriptional responses to small variations in the disturbances they experience (Denisenko *et al*, 2020). Batch effects are therefore a big source of variation among single cell experiments. Several mathematical models and methods have been developed in the past couple of years, to successfully integrate datasets from different experiments, and even coming from different single-cell strategies, which are known to yield substantially different results (Luecken *et al*, 2020). There are other sources of variation, which have biological origin, but are nonetheless confounding factors in some analyses. It’s important to note, that biological variation is not uninteresting *per se*, but depends on what the aims of the study are. Among these factors, one can mention cell cycle, the sex of the cells or mitochondrial transcript counts in tissues with high energy turnover (Luecken & Theis, 2019).

Data originating from high throughput scRNA-seq usually comprises from 3,000 to 10,000 cells per run. For each of these cells, tens of thousands of measurements are made, representing the expression of each gene. Such an amount of data is difficult to analyze and visualize directly. Therefore, a dimensionality reduction is normally performed to capture and analyze only the main variance components. The most commonly used dimensionality reduction to this day is the principal component analysis (PCA). After calculating the dimensionality reduction, only the components which capture high levels of variance are considered for further analyses. Using graph network and community detection algorithms, clusters of transcriptionally similar cells are then detected using the dimensionality reduced data (Luecken & Theis, 2019; Amezquita *et al*, 2020). These clusters of cells are then characterized as different cell types or states, depending on the genes that they differentially expressed compared to the rest of the cells.

Cluster identification relies on published, spatially-resolved expression data. Among this kind of data, previous RNA-seq and *in situ* hybridization studies are among the most commonly used. Dedicated databases which compile gene expression profiles have proven to be highly valuable resources to annotate single-

cell datasets. From this point on, scRNA-seq studies diverge in the kind of analyses applied, depending on the aims of each study. Among the different analyses that are usually performed are: differential gene expression, transcription factor expression patterns, co-expression, and pseudotime. I will focus shortly on pseudotime analyses.

Tissues are normally not a homogeneous mass of cells, but complex collections of different cell types. Therefore, when sampling complex tissues, given the resolution of single-cell sequencing, we can observe a snapshot of the multiple cells that constitute them. Very importantly, during a dynamic process, a simultaneous snapshot of the multiple states the cells are in, can be studied. For example, cells in an organ during embryonic development are not only different from each other as different cell types, they are also present in different stages of differentiation along their maturing lineages. For dynamic systems, where cell states form a continuum of states, scRNA-seq provides the opportunity to study the gradual transcriptional changes along what is called a “pseudotime dimension” (Trapnell *et al*, 2014).

Pseudotime inference and analyses allow to dissect developmental and other dynamic processes by ordering cells along a trajectory based on expression pattern similarities (Trapnell *et al*, 2014). Changes in the expression of genes along, and at any point of the pseudotime trajectory, can be detected. Upregulation, downregulation, sporadic or cyclical expression can be observed using dedicated statistical analyses. The topology of the pseudotime trajectory can take several forms, which can also be reconstructed and examined. Different topologies reflecting biological dynamic processes include linear trajectories, bifurcations, multiple branching, and cyclical or circular trajectories. There are several trajectory inference strategies that work better depending on the underlying biological topology of the dynamic process being studied (Saelens *et al*, 2019).

Here, I present the adaptation and optimization of several methods used to analyze transcriptomic data of single cells stemming from developing chicken tissue. While scRNA-seq expression data has similar characteristics in every experiment, certain bioinformatics processes must be adapted depending on the type and design of the experiment, species, or tissue used. Moreover, given the ever-growing amount of different methods to perform certain analyses – e.g. pseudotime – benchmarking tests are desirable in order to properly compare and choose among them. In the work presented here, I show different analyses and adjustments made during the first studies of chicken development at single-cell resolution.

Results

Drop-seq bioinformatics pre-processing

For our single-cell atlas of the developing chicken hind limb (Feregrino *et al*, 2019), we used different datasets, due to availability of new technologies during the project progress. Our first dataset consisted of Hamburger-Hamilton stage 29 (HH29 ~ 6 days of development) (Hamburger & Hamilton, 1951) hind limbs pooled in 3 different Drop-seq runs performed in Boston MA at the Broad Institute back in 2015. After alignment

and counting of the reads, the Drop-seq workflow needs a manual input of the expected number of valid cells in the experiment. By sorting the cell barcodes by their total number of reads, and drawing a cumulative distribution of the reads, an inflection point can be detected in the distribution After this inflection point, the cell barcodes don't contribute meaningfully to the cumulative fraction of reads anymore (Macosko *et al*, 2015). This inflection point also represents the number of valid cells in each of the experiments. Following the documentation of Drop-seq, it was unclear to us the way this inflection point is suggested to be calculated, or whether this is about a visual inspection of a plot. Since our cumulative plots didn't show such an obvious inflection point (Figure 1 A and Supplementary figure 1) as the one depicted in the Drop-seq documentation, and to avoid any subjective assessments of the number of valid cells, we used a calculation to find the inflection point.

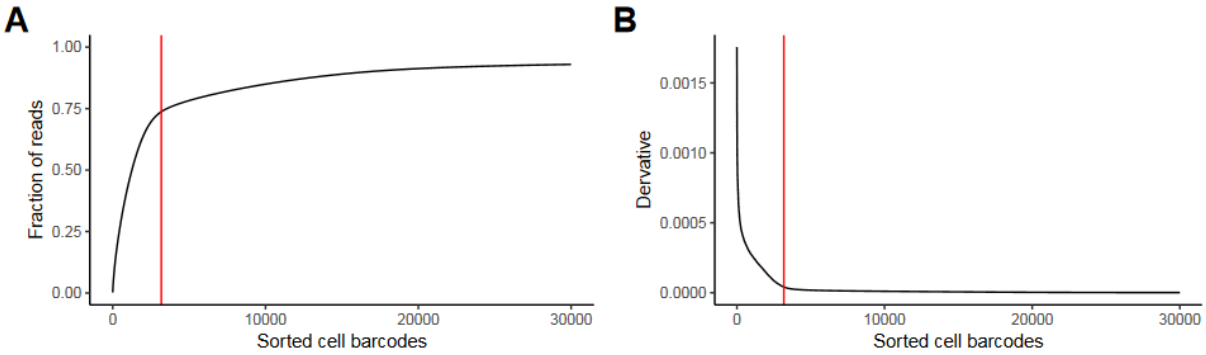


Figure 1 Drop-seq valid single cells selection. Plotted is the cumulative sum of the raw data from a representative of the 3 different HH29 hind limb runs we performed. Only the first 30,000 cell barcodes are plotted for better appreciation. Red line indicates the inflection point as calculated by the inflection package in R. **A** cumulative plot of the reads per cell barcodes. **B** First derivative of the curve in A.

For this, instead of using the cumulative curve directly, we used its first derivative. We decided to make use of the first derivative since the inflection point is sharper and more obvious there. Using the "inflection" package in R (Christopoulos, 2012), we calculated an inflection point which reflects what is represented in the Drop-seq documentation (Figure 1 B). The inflection points calculated were 3182, 3742 and 2558 for each respective run. These numbers were then used as the expected number of valid cells during the Drop-seq bioinformatics workflow.

Chicken genome annotation

For the second part of the project, in 2017, we sequenced single cells obtained from hind limbs of chicken embryos at stages HH25 and HH31 (~4.5 and ~7 days of development). For these experiments we made use of the single-cell RNA-seq approach from Chromium 10x Genomics. After our work on the single-cell atlas was finished (Feregrino *et al*, 2019), a new version of the chicken genome was released by ENSEMBL; GRCg6a (Genome Reference Consortium, 2017). In order to make use of our data for further research (see Chapter 3), we decided to re-do some of our analyses using the latest version of the chicken genome.

During our migration from genome version Gallus_gallus-5.0 (Schmid *et al*, 2015) (ENSEMBLE release 94), we accounted for 225 Gene stable IDs associated with a gene name (according to ENSEMBL BioMart and UNIPROT (Kinsella *et al*, 2011; Bateman, 2019)), which were completely absent from the genome version GRCg6a. The gene names from these IDs were also not associated to any new stable ID from GRCg6a either. We obtained the sequences of these genes, from the older genome version, and performed a batch BLAST (Zhang *et al*, 2000) against the newer version of the genome to know if the sequences of these genes were still present in it, and if they had been annotated as another gene. After comparing the sequences, and taking genomic stranding into account, we found that the sequences of 62 of these genes had no overlapping annotation in the version GRCg6a of the genome (Table 1). While many of these genes are not informative within our datasets, some of them do show different expression levels among our samples. We decided to append these genes to the GRCg6a genome annotation we used further.

Table 1 Genes present in Gallus_gallus-5.0 and appended in the GRCg6a annotation we further used.

ACTL9	KIF19	TRIM42	ISX	BPIFB6	LIF
BPIFB4	FGFBP1	APOA5	ARL13A	HIST1H111L	KIAA0408
KRT4	TAS1R3	DYRK1B	ESAM	SCRIB	SCARA3
UBL4A	POPDC3	RNF223	TMEM244	ARRB1	TERB2
NKX3-1	CYP8B1	SELENOH	AVPI1	C6orf163	CHIR-AB3
NEUROG3	SEBOX	OTUD5	gga-mir-1815	STAT6	gga-mir-1787
CLDN34	INKA1	H1F0	CTDNEP1	GABRR3	PPP1R3E
C9orf24	ADGRG3	RPS5	C2orf70	COA1	MZB1
CCDC42	C4orf45	RF00024	EFCAB3	TMEM81	OCSTAMP
CREBZF	SPATA2L	L1CAM	C2	NOXRED1	GPX2
CATIP	C2orf50				

Furthermore, during the re-analysis of our data, we realized that there was an incompatibility between our Chromium 10x Genomics single-cell RNA-seq data and the genome annotation from ENSEMBL. When inspecting the alignment of the reads to the genome using a genome browser, we observed that the alignment peaks of certain genes lies outside - or very close - to the 3' UTR annotations of GRCg6a (Supplementary figure 2). We observed this situation with genes we expected to be expressed in our samples, as is the case of SOX9 or GDF5. This problem was not observed when the Drop-seq data was used, as for the same genes the alignment peaks lie within the annotated genes. We reasoned that this is due to the fact that the 10x Genomics approach has an enzymatic digestion and size selection step that makes the sequencing libraries even more 3' biases than the ones from Drop-seq. This results in reads of certain genes not being counted, leaving us with an expression matrix that does not properly reflects the expression data of the cells.

To overcome this situation and faithfully reflect the transcriptome information we are analyzing, we decided to try to extend the 3' UTR of the ENSEMBL annotation. To do so, we followed a similar workflow as the one used to make the original chicken

genome annotation. We also used part of the data that was used to build the Galgal4 version of the annotation (Schmid *et al*, 2015). Specifically, we used bulk RNA-seq data sets from the Tabin / Ulitsky labs (see methods).

The reads of the different samples were paired-end mapped to the chicken genome version GRCg6a from ENSEMBL, individually. Then each of the mapped files were randomly subsampled to *ca.* 40 million pairs of reads. The subsampled files were then merged into a single FASTQ file and used to generate novel transcript models. According to our research goals, and to avoid further complications, we decided to focus our elongation algorithm only on protein-coding genes. We developed an R script that would extend only the 3' end of the tracks in the ENSEMBL annotation, filtered for protein-coding genes. The script goes through the following steps:

1. From the newly-generated transcript models, choose the ones which:
    - Have a minimum expression threshold of 1 fragments per kilobase of exon model per million reads mapped (FPKM).
    - Overlap only with one original gene annotation track.

2. Then, for each of the chosen novel transcript models:
    - Find the original gene and set of tracks which it overlaps.
    - If the original gene track has a shorter 3' span than the novel model:
        - Change the 3' coordinate of all tracks of type: exon, gene, 3' UTR and transcript. Only up to 5,000 bp.
        - Check if the modified track overlaps with the neighboring gene, if so, modify the extension to end 2 bp before the next gene track.

In total, we expanded 19057 annotation tracks, which are distributed in the following types (Table 2):

Table 2 Extensions of GRCg6a annotation tracks. The different track types found in the chicken annotation, the total amount per type, and how many of these tracks were modified. Also presented as percentage per track type.

Track type	Total	Modified	Percentage
Gene	16828	4485	26.65
CDS	319037	0	0
Transcript	28403	5720	20.13
Exon	332664	5720	1.71
5' UTR	29790	0	0
3' UTR	19621	3132	15.96
Start codon	27048	0	0
Stop codon	28061	0	0
Selenocysteine	5	0	0

None of the modifications resulted in a decrease in length, and most of them represented an increase of less than 1,500 bp (Figure 2 A). There is no correlation between the original length of the annotation and the length of the extensions (Figure

2 B). Based on these satisfactory indications, we tested what were the effects on the UMI counting of Chromium 10x Genomics data. We used the dedicated pipeline provided by 10x Genomics to process the raw reads of our previously mentioned HH31 sample. One time we used the original ENSEMBL chicken GRCg6a genome and annotation, filtered for protein coding genes; and another time our extended version of the annotation. We then compared the UMI counts between both annotations.

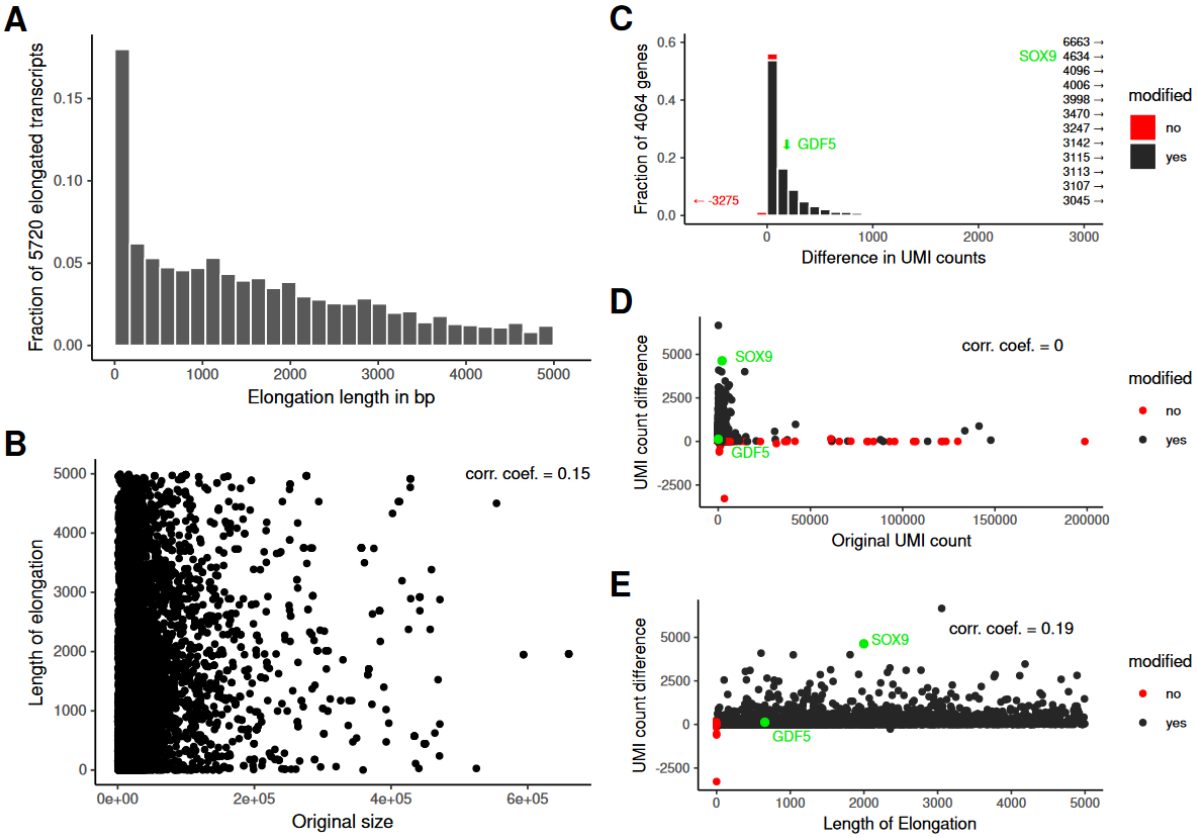


Figure 2 Chicken genome annotation improvement, transcript and gene extension statistics. **A** Frequency of the different elongation lengths. Showing only transcripts that were elongated. **B** Original size of the transcript v. the elongated size. Same genes as in A. Pearson correlation coefficient of 0.15. **C** Frequency of UMI count difference between original and elongated transcripts. Showing only genes with a UMI count difference. Modified genes in black, unmodified in red. Bins of 100 counts. Outliers labeled on each side. **D** Relation of original UMI count and UMI count difference. Showing the same genes as C. Pearson correlation coefficient = 0. **E** Relation of elongation length and UMI count difference. Showing the same genes as C. Person correlation coefficient = 0.19.

We observed that only 54 genes showed a decrease in UMI counts (Figure 2 C). One gene showed a decrease of 3,275 UMI counts, 5 other genes decreased between 100 and 600, and the rest showed decreases smaller than 36 UMI counts. Of all the genes with a decrease in UMI counts, only two were modified in length (losses of 260 and 9 UMIs). The loss of UMI counts could be explained by a cooption of the reads by a neighboring extended gene, a gene on the opposite DNA strand, or an unexpected change in the gene splicing structure, which is important for the read counting stage with STAR (Dobin *et al*, 2013). On the other hand, 4,010 genes

showed an increase in UMI counts (Figure 2 C). While most of these genes have an increase of only 1-100 UMIs, one gene shows an outstanding increase of 6,663 UMI counts. Of the genes with a gain in UMI counts, only 97 where not modified in length. The unmodified genes showed increases up to 267 UMIs. There was no linear correlation between the original UMI count and the difference in UMI counts (Figure 2 D). Likewise, we found no correlation between the elongation length and the difference in UMI counts (Figure 2 E).

We inspected the two extreme outliers of UMI count increase and decrease. First, the gene with a dramatic decrease of 3,269 counts, ENSGALG00000053871, was not modified. This gene has a short (~500 bp) cDNA sequence that overlaps on the opposite strand with a very long (~81 kb) gene, which was also not modified in any way. We couldn't find an explanation to this decrease, but deemed it negligible, since it only happened in one gene, which seems to not be very informative for our data set. On the other hand, the gene which gained 6,649 UMI counts was elongated by ~3kb, with no close-by gene on the 3' side. This makes us believe that this increase in UMI counts is not spurious. Importantly, particularly interesting genes for our research, like SOX9 and GDF5, showed a significant UMI counts increase. GDF5 had an increase of 130 UMIs, and SOX9 was the gene with the second highest increase with 4,634 new UMIs (Figure 2 C, D and E). Based on all these observations, we are confident that our extension of the 3' UTR annotations was successful.

Quality filtering and sources of variation

As explained before, the dedicated pipelines to count and assign UMI counts of scRNA-seq experiments filter cell barcodes to exclude empty beads. This, however, is usually not sufficient to filter out low quality cells, and it doesn't remove possible cell doublets. To solve this, several filters and thresholds are considered standard before the proper analysis of a sample, but also ultimately depend on the nature of the samples and research aims of each project. In our case, after using different approaches, we developed a quality filter pipeline which uses thresholds relative to the sample being processed, instead of arbitrary absolute numbers.

- Relative filter based on UMI counts:
    - Filter out cells that have more than 4 times the mean of the UMI counts, and less than 20% of the median. These represent the probable doublets and inviable, or in poor shape, cells
- Combined relative filter based on mitochondrial UMIs and total UMIs:
    - We calculate two thresholds: sample median + three times the MAD (median absolute deviation) of fraction of mitochondrial UMIs, and sample median of total UMI counts per cell. We then disregard cells that are both above the mitochondrial and below the total UMIs threshold. These represent probable dying cells.
- Combined absolute and relative filter based on the total UMI count and genes detected per cell relation:

- We calculate two thresholds: number of genes detected / UMI count relation of 0.15, and 2/3 of the maximum number of genes detected overall. We remove the cells that are under both thresholds. These cells have an unusually high amount of UMIs, representing very few genes.

As an example, we filtered the cells of the following sample: caudal portion of the lumbar neural tube HH36 (~ 10 days of development). The 10x Cell Ranger pipeline gave as a result a matrix of 5,969 cells, with UMI counts per cell ranging from 500 to 33,221, and a median of 7,145. In this case, the cell with 33,221 UMI counts is probably a doublet. During the first filtering step, 543 cells have too few UMI counts – less than 20% of the sample median (< 1,429) –, while 2 cells have too many – more than 4 times the mean (> 29,571) – (Figure 3 A).

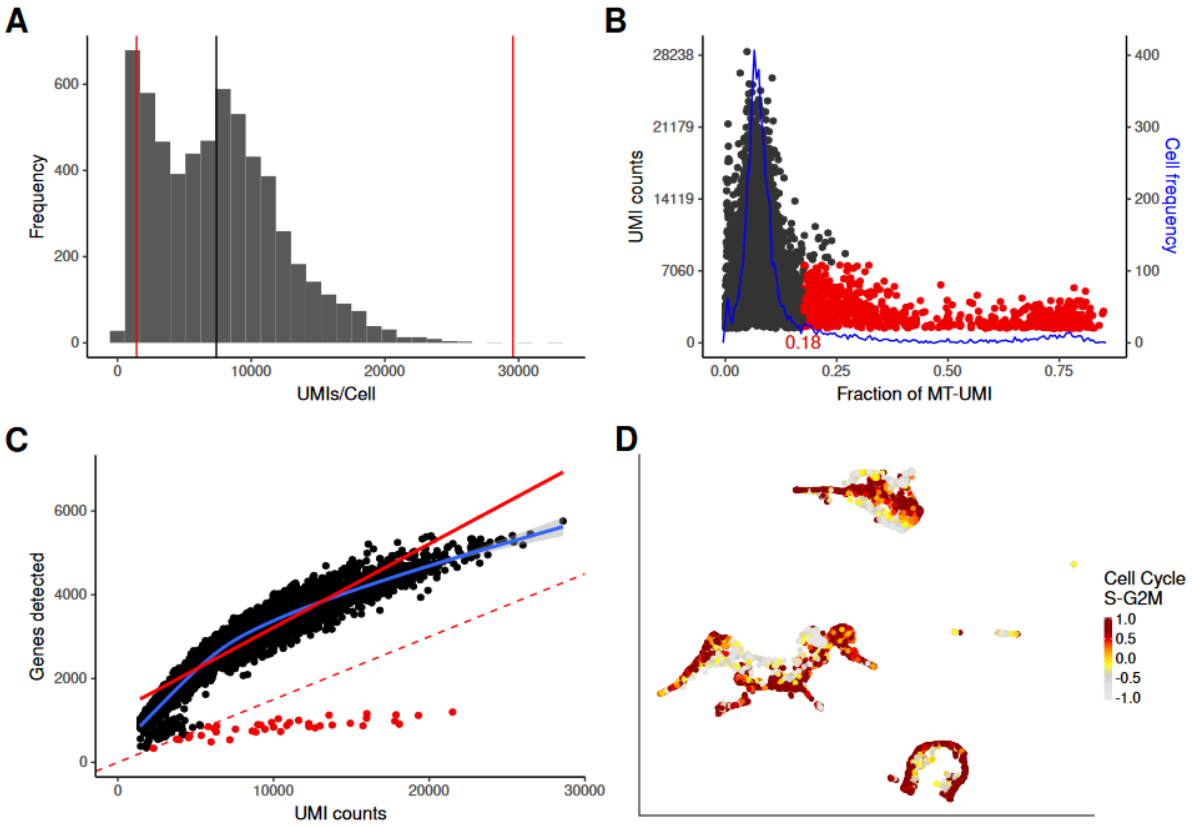


Figure 3 Quality control of scRNA-seq data and cell cycle scoring. **A** Frequency of UMI counts per cell. Black line denotes the mean of the sample, left red line threshold = 0.2*median, right red line threshold = 4*mean. **B** Relation of fraction of mitochondrial UMI v. total UMI counts per cell. Red cells do not pass the threshold 0.18 and < median UMI count. In blue the frequency of the cells with that given fraction of UMI of mitochondrial origin. **C** Relation of UMI count v. number of genes detected per cell. Dotted red line denotes the 0.15 threshold, and red dots are cells that don't pass the threshold. Red solid line shows a linear regression fit of the data, blue line shows a smoothed generalized additive model fit with confidence interval. **D** δS-G2M calculated for each cell. Plotted in an exaggerated tSNE.

During the second step, a threshold in the fraction of UMIs of mitochondrial origin – sample median plus three times the MAD – was calculated at 0.18. All cells with a mitochondrial UMIs fraction higher than the threshold and a UMI count smaller than the sample median were removed (Figure 3 B). During our last filtering step, less than

40 cells remained with too many counts in only a few genes – genes detected / UMI count < 0.15 –, they were also removed from the dataset (Figure 3 C).

When analyzing highly proliferative tissue, like is often the case with developmental samples, it is sometimes helpful to account for the variation created by the cell cycle. In order to avoid cell clusters to be divided into proliferative and non-proliferative states, we applied a correction for the variance caused by the cell cycle. This was based on the correction implemented in the original Drop-seq paper (Macosko *et al*, 2015). For this, a dataset containing pairs of genes known to covariate during the cell cycle of mouse hematopoietic cultured cells is used as a reference, and the covariance between the same pairs of genes in the test sample is then calculated. Based on this data, S, G1 and G2/M scores are calculated for every cell (Scialdone *et al*, 2015). We then calculate the difference between S and G2/M scores (δS-G2M), and use it as a continuous variable to correct for this variation. As a result, we observe that cells with different δS-G2M are evenly scattered across our data (Figure 3 D).

Dimensionality reduction and cluster identification

The most important step during the analysis of scRNA-seq data is probably the dimensionality reduction. As mentioned before, PCA has been the most used approach to reduce the dimensions of such data. Yet, when performing a PCA, the result are as many principal components as initial variables - 1. When performing a dimensionality reduction, the aim is to capture most of the data variance within a low dimensionality space. For this reason, only those principal components that represent most of the variance of the sample should be used. While several methods have been proposed to calculate the number of principal components to be taken into account, we use a rather simple approach based on the one implemented in previous works (Shekhar *et al*, 2016). Here, the squared standard deviation of each of the first 50 or 100 principal components is plotted in a frequency histogram with 500 bins. On the left-hand side of the plot, a Marchenko-Pastur distribution is observed, and many outliers to the right of it. By visual inspection of the plot, we simply count the amount of outliers that fall off of the distribution. While this is somehow subjective and not very precise regarding to the limit of the distribution, we have observed that the differences between analyses calculated on a range of +/- 3 principal components around this number are negligible.

There have been some efforts to automatize and make cell cluster annotation replicable and standardized. Nonetheless, these approaches depend on previously annotated datasets, which must contain the cell types one is trying to characterize. Given the novelty of our sampling, and the general lack of scRNA-seq datasets from chicken samples, we perform our cell annotations by manually consulting the existing bibliography. To this end we carry out cell clustering using community detection algorithms implemented in Seurat (Stuart *et al*, 2019). We then calculate the genes which are enriched in each of the clusters, relative to the rest of the cells. With the resulting gene lists, we consult the spatial expression data repositories Geisha – Chicken Embryo Gene Expression Database (Bell *et al*, 2004) and MGI – Mouse

Gene Expression Database (Smith *et al*, 2019). The spatial information, along with reported molecular function of the genes, provide us with a good understanding of what each cell cluster represents in terms of cell types or states.

To ease exploration of this kind of data – especially for members of our lab –, and inspired by several other scRNA-seq studies, I developed Shinny apps using RStudio. Here, the individual user can plot different kinds of data from each of the single-cell data sets available. For example, the expression of any gene can be plotted on different dimensionality reductions, across different samples (Figure 4 A). This information is crucial for cell cluster annotation, since it visually shows enrichment of genes across the single-cell space. To do this, the user needs to input the corresponding ENSEMBL gene code, alternatively the official gene symbol can be provided and the app will internally find the corresponding gene within the dataset. Likewise, data statistics like mitochondrial reads percentage, cell cycle score, UMI counts, clustering data or any other kind of data associated with the experiments can be easily accessed (Figure 4 B). A drop-down menu offers the user the different kind of statistics associated with the experiment. This kind of apps offer a user-friendly interface, and allow exploration of these highly complex datasets by clicking a button, and without any prior bioinformatics experience.

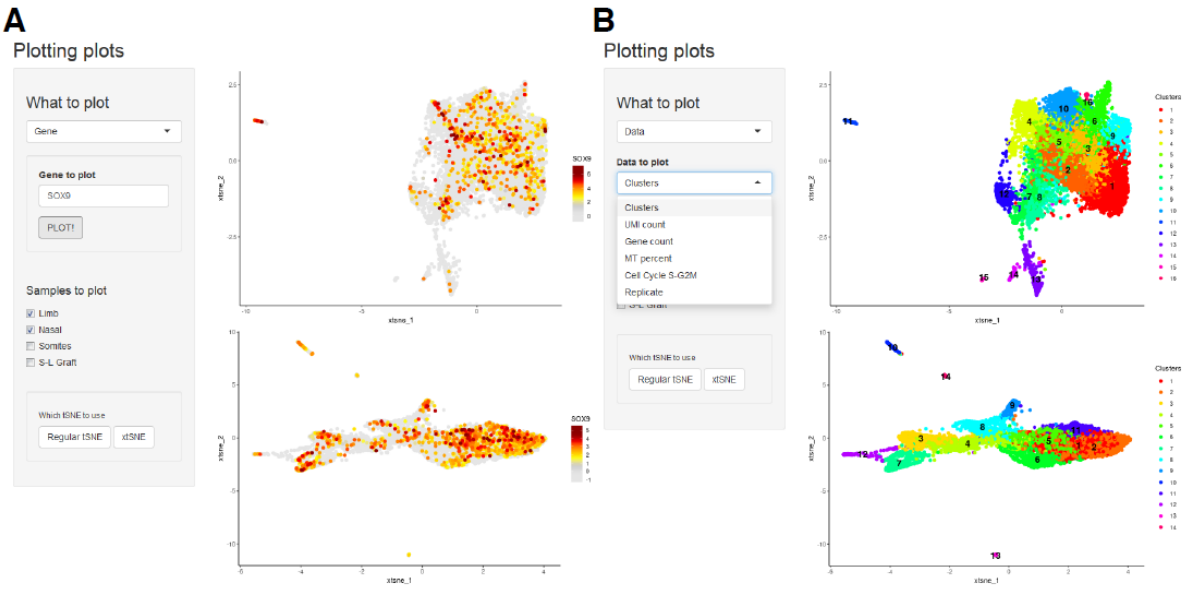


Figure 4 Shinny apps for the exploration of single-cell data. **A** Screenshot of the Shiny app interface developed. Showing the control panel on the left, in the "gene expression" mode. Users can input the gene of their choice and click on the "PLOT!" button. Other options are: the individual datasets to plot and the dimensionality reduction to use. On the right, the output of SOX9 gene expression in two of our datasets, showing expression level of each single cell and a legend as reference **B** Screenshot of the same app interface in the "data statistics" mode. Users can choose the data information to plot from the drop-down menu. Additional options are the same as in A. On the right, the output of clustering data in the same datasets as A. Showing the cluster of each single cell and a legend as reference.

Pseudotime

In order to investigate the transcriptional dynamics of developmental processes, like chondrogenesis in the chicken limb (Chapter 3), we make use of pseudotime analyses (Trapnell *et al*, 2014). Previous to achieve the results we present in chapter 3 using Slingshot (Street *et al*, 2018), we tested other pseudotime methods: Monocle v2 (Qiu *et al*, 2017) and URD (Farrell *et al*, 2018). For this, we used all of the cells we have identified as chondrocytes in our HH29 hind limb sample (see Chapter 2). After using several sets of genes (different thresholds of variability, differential expression between all clusters), differentially expressed genes between early and joint chondrocytes showed the best results, according to our expectations. We used these 63 genes to calculate the dimensionality reduction and pseudotime for both methods.

Monocle was used in an unsupervised manner and resulted in a bifurcation (1 branch dividing into 2), on which we identified a branch as the early chondrocytes trajectory and origin of the pseudotime (Figure 5 A). The other two branches coincided with late, or phalanx-forming chondrocytes, and joint-forming chondrocytes trajectories. URD was run in a semi-supervised manner, we used the cells with the lowest pseudotime values from the Monocle results as the root of the URD calculations pseudotime. The two pseudotime tips coincide with late and joint chondrocytes (Figure 5 A).

We compared the pseudotime trajectories to find out which genes had similar (and how similar) dynamics across both methods. Both resulting pseudotime models consisted of three main branches (Figure 5 A and B), hence the comparison was made branch-wise. To overcome the fact that some branches have different length, for both methods we calculated the maximum number of bins dividing the branches pseudotime, such that each bin would still contain cells. Per branch, we took the smaller of these two numbers of bins and divided both pseudotimes. The early branches were thus divided in 28 bins, the phalanx-forming in 9 bins, and the joint-forming in 16 bins. We used gene expression averaged within these bins as data points. Then we tested the correlation of each of the genes showing high averaged expression in at least one of the bins. We used the spearman's rho to quantify the correlation of the expression along the corresponding branch of the two methods. Thus resulting in a correlation coefficient for each highly expressed gene, reflecting how similar the expression dynamics between the two pseudotimes are. In a similar manner, we used 5 bins along each branch to test for cell, instead of gene, correlations.

We found that 49% of the genes tested showed a positive (rho > 0) correlation along the early branches, 62% along the phalanx-forming branches, and 72.9% along the joint branches. However, very few of these correlations were significant (FDR < 0.05). Along the early branches only the genes PLAC9 and CSRP2 showed significant correlations (Figure 5 C). Along the phalanx-forming branches, none of the correlations were statistically significant (Figure 5 D). Meanwhile, 6 genes – FRZB,

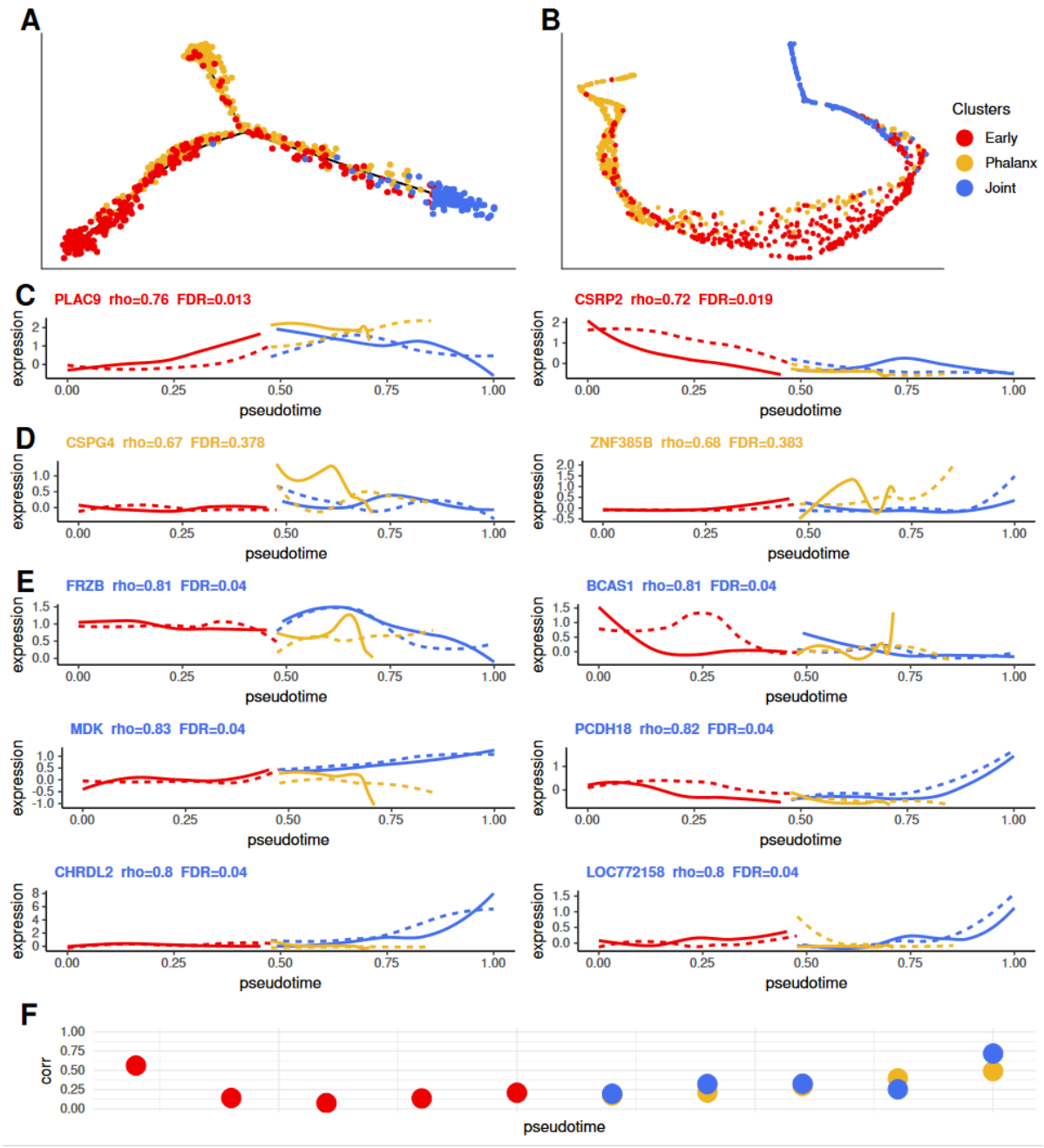


Figure 5 Pseudotime analyses comparison. Colors in B are valid throughout. **A** Monocle pseudotime cell ordering as a DDRTree. **B** URD pseudotime cell ordering as a 2D force-directed layout. Pseudotime of A and B starts in the area enriched for early chondrocytes (red). **C** Expression dynamics of significantly correlated genes along the early branches. Loess fitting on the single-cell expression data. **D** Expression of the two genes with the lowest FDR in the correlation test along the phalanx-forming branches. **E** Expression of significantly correlated genes along the joint branches. In C-E: gene name, correlation coefficient and FDR are colored by the branch in which the correlation was calculated. Branch coloration corresponds to B. Solid line: Monocle, dashed line: URD. **F** Spearman correlation of pooled cells across 10 bins of pseudotime.

BCAS1, MDK, PCDH18, CHRDL2 and LOC772158 – had significant correlation along the joint branches (Figure 5 E).

When comparing the cells, instead of the genes, across both methods, we observed a moderate (~ 0.5) correlation at the beginning of early trajectories, the correlation drops to around 0.15 - 0.2 already in the second pseudotime bin. After the branching points, along the phalanx-forming chondrocytes trajectories, correlations ascended from close to 0.2 to almost 0.5. Meanwhile, the joints trajectories showed correlation going almost to 0.75 (Figure 5 F). Taken together, the correlation analyses indicated that while the methods show similar global orderings, at the fine level they are far from identical. Moreover, it seems like the trajectory culminating in joint chondrocytes is in higher agreement between both methods.

Discussion

Here, we present several methodological applications to successfully perform single-cell RNA-seq analysis of chicken embryonic tissues, a classical model to study vertebrate development. We developed diverse approaches in order to tackle the different challenges we faced as the first research group performing scRNA-seq experiments with chicken samples. Our work spanned two different single-cell transcriptomics technologies, which presented different challenges. We show calculations which helped us to manually determine the amount of valid cell barcodes present in Drop-seq assays. On the other hand, we produced a modification of the chicken genome, which is compatible with the heavy 3' bias of Chromium 10x Genomics assays. Furthermore, we present a data-based quality filtering strategy for the processing of single-cell data. Our strategy uses thresholds which are calculated relative to the data itself, and therefore can be easily applied to a multitude of samples. Notably, I developed Shinny apps that allow in-depth data exploration by our group members with no prior bioinformatics experience. This kind of tools not only allows to share observations with other scientists, but also provides the opportunity to actively participate in the data analysis process. In this way, I have received invaluable input, for example, in the cell cluster annotation process. In addition, we also present a comparison of pseudotime methods, which exhibited a stark difference in the results they provided with the input data.

Despite the efforts to construct and annotate the chicken genome (Schmid *et al*, 2015; Genome Reference Consortium, 2017). We found that the gene annotation had an incompatibility when used to count Chromium 10x Genomics scRNA-seq reads. The pitfalls of the chicken genome and its gene annotation have been addressed in other publications. A chicken gene annotation improvement has already been published in the past (Orgeur *et al*, 2018). While their bioinformatics approach is more comprehensive and detailed than ours, it is based on the 4th version of the chicken genome and annotation Galgal4 (Hillier *et al*, 2004). Our annotation modifications, in contrast, were made using as basis the most current (6th) version of the chicken genome and annotation GRCg6a (Genome Reference Consortium, 2017). Moreover, for the previous annotation improvement, Orgeur and collaborators used two

micromass cultures as the source of their transcript data. For our approach, we opted to use data stemming from whole developing embryos, which should better reflect the transcriptomic signals of our embryonic samples.

To this date, and to our knowledge, only one study involving high-throughput scRNA-seq from chicken has been published, besides our own (Feregrino *et al*, 2019). In their study (Estermann *et al*, 2020), trying to explore gonadal sex differentiation, they also made use of 10x Genomics single-cell assays. Using the Gallus-gallus5.0, protein-coding and lncRNA from ENSEMBL v92 annotation, they also performed a gene annotation extension. In their approach, they extended 1000 bp after every gene. In another step "multiple gene annotations at a location were both counted, but genes which shared all of their aligned reads with another gene were removed". In contrast, our extension approach is based on transcriptomic data obtained from whole embryos, with considerable sequencing depth, in order to represent the gene expression from all the organs. Moreover, our approach is more directed, since the version 6 of the chicken genome contains 16,828 protein-coding genes (version 5 contains 18,346), of which only 2,498 were modified. While many of our extensions were short in length, 55% of our gene extensions were longer than 1000 bp, size which they used to elongate all genes. These 1000+ bp extensions accounted for UMI count increases with a median of 114 and up to 6663. Although we don't know if these UMI increases are due to reads mapping beyond 1000 bp of the extensions, the 3' bias of the 10x Genomics method would suggest so.

Our quality filter approach for scRNA-seq data complies with what's considered best practices (Luecken & Theis, 2019) in several aspects. We use the three covariates that are diagnostic of viable cells (Ilicic *et al*, 2016; Griffiths *et al*, 2018): number of UMI counts per cell, fraction of UMI counts with mitochondrial origin, and the number of genes detected per cell. According to the current best practices in single-cell RNA-seq analysis (Luecken & Theis, 2019), thresholds should be drawn to filter-our outlier peaks in the covariate distributions. Nonetheless, they don't propose a method to calculate said thresholds. They also recommend to have quality control thresholds determined for each sample separately. Our dynamic thresholds are therefore a good option to calculate thresholds relative to the sample in question. Nonetheless, they also argue that the covariates should not be used independently to make cutoffs, but in a combined manner, to avoid misinterpretation of the cellular signals. Here, we presented two combined cutoffs. To filter out cells rich in mitochondrial UMIs, we use a threshold based on both mitochondrial and total number of UMIs. And to remove cells with an unusual amount of UMIs we use both UMI number / gene number relation and the total count of UMIs.

While regressing out the effects of the cell cycle is considered common, it's debated whether such a data correction should be performed (Luecken & Theis, 2019). The removal of these effects has been shown to improve developmental trajectories inference (Buettner *et al*, 2015; Vento-Tormo *et al*, 2018). But the cell-cycle might also be informative for certain biological processes, and the correction of

this signal might also mistakenly correct the effects of other associated processes. Our approach corrects for the difference between the S and G2M phase scores. This corrects for differences in cell cycle phase amongst proliferating cells, but maintains the signal between cycling and non-cycling cells (Stuart & Satija, 2019). It is also important to note, that our cell cycle correction is only done in order to calculate PCAs, tSNE visualizations, and clustering algorithms. In order to perform differential expression analyses, we use data that has not been corrected for the cell cycle effects.

While a PCA step is almost universal during scRNA-seq analyses, there is no standard method to determine the number of main principal components. This is a problem that is not unique to scRNA-seq analyses, and the principles of existing methods have been previously debated (Cangelosi & Goriely, 2007). Moreover, calculations like the jackstraw permutation test have been proposed to be used in the field of single-cell genomics (Chung & Storey, 2015; Macosko *et al*, 2015). Nonetheless, we opted to use a simple, although subjective, method that also has a statistical basis. This method, described also in the context of single-cell experiments (Shekhar *et al*, 2016), is based on a distribution which follows the Marchenko-Pastur (MP) law (Marčenko & Pastur, 1967). This MP distribution predicts the upper and lower levels of a null distribution of the PC eigenvalues. Therefore, any PC with an eigenvalue lying outside of this distribution is unexpected under randomness. While this approach could be considered inexact, proponents of other methods also recognize that using a couple of components more or less than the ones computed is not a major concern (Macosko *et al*, 2015).

Lastly, we present a comparison benchmarking two single cell pseudotime analysis methods. While it was not exhaustive, it helped us to make decisions about the general way we later proceeded to do these kinds of analyses. Our comparative analyses could also be improved using pseudotime warping methods (Alpert *et al*, 2018), to better align our trajectories. The fact that we didn't have a standard data set, of which we knew the true time ordering also hindered our ability to draw further conclusions. In the meanwhile an exhaustive comparison of pseudotime analysis methods has been done (Saelens *et al*, 2019). There, they compared a large amount of methods, taking into account several characteristics, including accuracy, scalability and usability. Compared to other tree-based methods, URD ranks low among them, while Monocle ends up among the middle ranks. Slingshot (Street *et al*, 2018), which we further use in Chapter 3 of this thesis, ranks as the best tree-base method.

All together, we present a collection of adaptations and resources that allow for the analysis of scRNA-seq data coming from developing chicken embryos. Mainly, we have shown that the adjustments made to the genome annotation, although somewhat simple, show an improvement in the quantification of transcriptomic data. We believe that, collectively, our analyses workflow represents a valid improvement to analyze chicken scRNA-seq data. Some of the adaptation and adjustments we

have made can be applied to analyses for other organisms, and of course be refined and reworked to best fit the specific cell or tissue types to be analyzed.

Methods

Drop-seq bioinformatics pre-processing

To calculate the inflection point of the cumulative sum of raw Drop-seq data, we first followed the workflow of the Drop-Seq Core Computational Protocol V1.2 (Macosko *et al*, 2015). This, until the point where an expected number of cells is required to be determined. To perform these calculations, we used the read counts per cell produced by the Drop-seq pipelines. In R, we calculated cumulative sums of the reads over all cells, and used the 30,000 cell barcodes with the highest read counts to calculate a distribution curve using the function “smooth.spline” from the package stats (3.5.1., 2018). We then used the function “predict” from the package stats, to calculate the first derivative of our distribution curve. We used this derivative to calculate the inflection point using the function “ede” from the package inflection v1.3.5 (Christopoulos, 2012). The inflection calculation using this function results in several values, of which we took the index j_{F_{2}} as our inflection point.

Chicken genome annotation

To append what we deemed to be missing genes to chicken genome annotation version GRCg6a, we first compared the tables of gene stable ID and their associated gene names of both annotations, which we obtained from ENSEMBL BioMart (Kinsella *et al*, 2011). We complemented the gene names provided by BioMart with gene names found in the UNIPROT database (Bateman, 2019). When comparing these tables, we identified gene stable IDs associated with a gene name, that were absent from the GRCg6a annotation. From this subset we further identified the gene stable IDs, with an associated gene name that is also absent from the GRCg6a annotation gene names, meaning that those particular gene names were not newly associated to a different stable ID within the newer annotation. From the resulting 225 genes, we obtained a fasta file with their sequences from ENSEMBL BioMart and performed a BLAST (Zhang *et al*, 2000) against the chicken version GCA_000002315.5 (GRCg6a) in the NCBI web server (Agarwala *et al*, 2018). The resulting BLAST hit table and GRCg6a genome annotation were compared using GenomicRanges and IRanges (Lawrence *et al*, 2013) in R. We considered only those BLAST hits that didn’t overlap with any gene annotation on their corresponding strand. The annotation tracks of the resulting 62 genes were appended to the GRCg6a genome annotation GTF file. The GTF “source” field of these genes was annotated as “ENSg5BLAT” and, the “chromosome” and “start position” fields were modified according to their BLAST hit coordinates.

In order to carry out our elongation of the chicken genome annotation, we used the datasets of paired reads mentioned below (Table 3). These data sets were individually mapped to the chicken genome using a dedicated pipeline I have developed. The pipeline takes raw reads, uses Trimmomatic (Bolger *et al*, 2014) to trim low quality base pairs, and then uses STAR (Dobin *et al*, 2013) to map the reads to the given

genome. The reads and mapping information are compiled in BAM files which are then used in the following calculations. Using SAMtools (Li *et al*, 2009) and Sambamba (Tarasov *et al*, 2015), we manipulated the BAM files to first filter for reads which had both pairs successfully mapped. These filtered BAM files were subsampled down to *ca.* 40 million pairs of mapped reads and later merged. We then used Cufflinks (Trapnell *et al*, 2010) to generate transcript models using the combined BAM file. The resulting GTF file containing the newly calculated transcript models was then used as input in our 3' UTR elongating script (attached as appendix), which elongated the tracks of our chicken genome annotation. To compare the UMI counts, we processed one of our samples using CellRanger V2.0 (10x Genomics) using each of the transcriptome annotations individually.

Table 3 Different samples that were used to construct the gene and transcript models in our 3' UTR extension framework.

Sample	SRA number
HH11 whole embryo	SRX893878
HH14 whole embryo	SRX893868
HH21/22 whole embryo	SRX893872
HH25/26 whole embryo	SRX893873
HH32 whole embryo	SRX893874
HH36 whole embryo	SRX893875

Quality filtering and sources of variation

Our quality filtering workflow is also based on R. Once scRNA-seq raw data has been processed with CellRanger (10x Genomics), or the corresponding alignment-counting pipeline, the UMI count tables are obtained. With these tables, summary statistics are computed for the total UMI counts per cell barcode. The mean and median obtained are used to first remove all cells with >4*mean and <0.2*median. In a second step, the mitochondrial genes are defined, and then the proportion of mitochondrial UMI from the total UMI counts is calculated for each cell. Any cell with >3*mad of the sample-wide proportion is also removed, except if they also have more than the median of UMI counts. In a final step, the relation of UMI counts and number of genes detected is calculated, and any cell with a relation of less than 0.15 are removed, except if the cells also have more than 2/3 of the maximum number of detected genes.

Cell cycle stage scores were calculated for each cell using the R package SCRAN (Lun *et al*, 2016). For this we first obtained a list of gene pairs known to covariate with the different cell cycle stages in mouse (Scialdone *et al*, 2015). The pairs were then translated into 1:1 chicken gene orthologues from ENSEMBL release 97 using BioMart (Kinsella *et al*, 2011). The resulting lists were then filtered to keep only the pairs in which both genes had a 1-to-1 orthologue in the chicken genome. This was used as a reference and scores were calculated using the function "cyclone" from SCRAN. The effects of the cell cycle on the data variance were then removed by using the δS-G2M as a variable to regress using Seurat's v3.0 "SCTtransform" function. It is important to note that this correction is only made to calculate PCs,

tSNEs and clusters. Data used for differential expression analyses, or other tests is not corrected for the cell cycle.

Pseudotime analyses

To perform our pseudotime analyses we first selected the genes that we would use to order the cells. For this, we used the cells in clusters 15 and 17 of our HH29 fine clustering dataset (Feregrino *et al*, 2019), which correspond to early chondrocytes and joint chondrocytes respectively. We performed a differential expression test using Seurat, with the δS-G2M as a latent variable and "negative binomial" as the method. After filtering for genes that showed a significant difference (adjusted p.value < 0.05 and log2 fold change >0.5), we ended up with 63 genes. With Monocle V2.0 (Trapnell *et al*, 2014), we used all the cells in clusters 15, 17 and 3 (early, joint- and phalanx-forming chondrocytes, respectively), and the 63 genes we previously obtained to calculate a DDRTree dimensionality reduction. Based on the DDRTree, we calculated the pseudotime of each cell.

Pseudotime analyses using URD (Farrell *et al*, 2018) are done using a diffusion map instead of a DDRTree. We used the same set of cells and genes to make the calculations. For this, we also ran a clustering step using Seurat and 12 PCs, which resulted in 9 new clusters. The clustering step is necessary since URD requires to know which cells are the root of the pseudotime, and which are located at the tips. To make things comparable, we selected all the cells that had a pseudotime <0.2 in the Monocle analysis (13 cells). Pseudotime was then calculated starting from these cells, and trajectories were later inferred using the tip clusters. The cluster tips were selected based on the expression of marker genes to correspond to the Monocle tips, as well as being the ones with the largest pseudotime values.

To compare both methods we first manually closed a gap of pseudotime we found produced by URD. The gap was produced between the 13 cells with time 0 and the next cell with time 0.187. We simply added 0.187 to the pseudotime of the cells with time 0. Both pseudotimes results were then scaled to have values between 0 and 1. We compared the methods in a branch-wise manner. To make comparisons at the gene level, we first binned the cells in the branch. We calculated as many bins of equal pseudotime size as possible, which would still contain at least 3 cells per bin, and used the same number of bins to divide both pseudotimes. Exceptionally, for the phalanx-forming trajectory, our threshold was 1 cell, because otherwise we only obtained 9 bins. We then selected genes with scaled expression of at least 1 in any bin and > 0 in at least 10% of the bins. We then used SCRAN to calculate Pearson correlations, p-values and false discovery rates of the gene activities across the bins using the "correlatePairs" function. In order to compare the methods at the cell level, we used a similar approach, but this time we used only 5 bins per branch.

References

3.5.1. RDCT (2018) A Language and Environment for Statistical Computing. *R Found. Stat. Comput.* **2:** https://www.R-project.org

Agarwala R, Barrett T, Beck J, Benson DA, Bollin C, Bolton E, Bourexis D, Brister JR, Bryant SH, Canese K, Cavanaugh M, Charowhas C, Clark K, Dondoshansky I, Feolo M, Fitzpatrick L, Funk K, Geer LY, Gorelenkov V, Graeff A, et al (2018) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **46:** D8–D13

Alpert A, Moore LS, Dubovik T & Shen-Orr SS (2018) Alignment of single-cell trajectories to compare cellular expression dynamics. *Nat. Methods* **15:** 267–270

Amezquita RA, Lun ATL, Becht E, Carey VJ, Carpp LN, Geistlinger L, Marini F, Rue-Albrecht K, Risso D, Soneson C, Waldron L, Pagès H, Smith ML, Huber W, Morgan M, Gottardo R & Hicks SC (2020) Orchestrating single-cell analysis with Bioconductor. *Nat. Methods* **17:** 137–145

Bateman A (2019) UniProt: A worldwide hub of protein knowledge. *Nucleic Acids Res.* **47:** D506–D515

Bell GW, Yatskievych TA & Antin PB (2004) GEISHA, a Whole-Mount in Situ Hybridization Gene Expression Screen in Chicken Embryos. *Dev. Dyn.* **229:** 677–687

Bolger AM, Lohse M & Usadel B (2014) Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30:** 2114–2120

Buettner F, Natarajan KN, Casale FP, Proserpio V, Scialdone A, Theis FJ, Teichmann SA, Marioni JC & Stegle O (2015) Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nat. Biotechnol.* **33:** 155–160

Cangelosi R & Goriely A (2007) Component retention in principal component analysis with application to cDNA microarray data. *Biol. Direct* **2:**

Christopoulos DT (2012) Developing methods for identifying the inflection point of a convex/concave curve.

Chung NC & Storey JD (2015) Statistical significance of variables driving systematic variation in high-dimensional data. *Bioinformatics* **31:** 545–554

Denisenko E, Guo BB, Jones M, Hou R, De Kock L, Lassmann T, Poppe D, Poppe D, Clément O, Simmons RK, Simmons RK, Lister R & Forrest ARR (2020) Systematic assessment of tissue dissociation and storage biases in single-cell and single-nucleus RNA-seq workflows. *Genome Biol.* **21:**

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M & Gingeras TR (2013) STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29:** 15–21

Estermann MA, Williams S, Hirst CE, Roly ZY, Serralbo O, Adhikari D, Powell D, Major AT & Smith CA (2020) Insights into Gonadal Sex Differentiation Provided by Single-Cell Transcriptomics in the Chicken Embryo. *Cell Rep.* **31:**

Farrell JA, Wang Y, Riesenfeld SJ, Shekhar K, Regev A & Schier AF (2018) Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis. *Science (80-.).* **360:**

Feregrino C, Sacher F, Parnas O & Tschopp P (2019) A single-cell transcriptomic atlas of the developing chicken limb. *BMC Genomics* **20:**

Genome Reference Consortium (2017) Human Genome Overview. *GRCh38.12* Available at: https://www.ncbi.nlm.nih.gov/grc/human [Accessed October 16, 2020]

Gilbert SF (2000) Comparative Embriology. In *Developmental Biology*, Gilbert S (ed) Sunderland (MA): Sinauer Associates

Griffiths JA, Scialdone A & Marioni JC (2018) Using single -cell genomics to understand developmental processes and cell fate decisions. *Mol. Syst. Biol.* **14:**

Hamburger V & Hamilton HL (1951) A series of normal stages in the development of the chick embryo. *J. Morphol.* **88:** 49–92

Hillier LW, Miller W, Birney E, Warren W, Hardison RC, Ponting CP, Bork P, Burt DW, Groenen MAM, Delany ME, Dodgson JB, Chinwalla AT, Cliften PF, Clifton SW, Delehaunty KD, Fronick C, Fulton RS, Graves TA, Kremitzki C, Layman D, et al (2004) Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* **432:** 695–716

Horder T (2010) History of Developmental Biology. In *Encyclopedia of Life Sciences*

Ilicic T, Kim JK, Kolodziejczyk AA, Bagger FO, McCarthy DJ, Marioni JC & Teichmann SA (2016) Classification of low quality cells from single-cell RNA-seq data. *Genome Biol.* **17:**

Islam S, Zeisel A, Joost S, La Manno G, Zajac P, Kasper M, Lönnerberg P & Linnarsson S (2014) Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat. Methods* **11:** 163–166

Kinsella RJ, Kähäri A, Haider S, Zamora J, Proctor G, Spudich G, Almeida-King J, Staines D, Derwent P, Kerhornou A, Kersey P & Flicek P (2011) Ensembl BioMarts: A hub for data retrieval across taxonomic space. *Database* **2011:**

Klein AM, Mazutis L, Akartuna I, Tallapragada N, Veres A, Li V, Peshkin L, Weitz DA & Kirschner MW (2015) Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161:** 1187–1201

Lähnemann D, Köster J, Szczurek E, McCarthy DJ, Hicks SC, Robinson MD, Vallejos CA, Campbell KR, Beerenwinkel N, Mahfouz A, Pinello L, Skums P, Stamatakis A, Attolini CSO, Aparicio S, Baaijens J, Balvert M, Barbanson B de, Cappuccio A, Corleone G, et al (2020) Eleven grand challenges in single-cell data science. *Genome Biol.* **21:**

Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, Morgan MT & Carey VJ (2013) Software for Computing and Annotating Genomic Ranges. *PLoS Comput. Biol.* **9:**

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G & Durbin R (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25:** 2078–2079

Luecken MD, Buttner M, Chaichoompu K, Danese A, Interlandi M, Mueller MF, Strobl DC, Zappia L, Dugas M, Colome-Tatche M & Theis FJ (2020) Benchmarking atlas-level data integration in single-cell genomics. *bioRxiv*

Luecken MD & Theis FJ (2019) Current best practices in single -cell RNA -seq analysis: a tutorial. *Mol. Syst. Biol.* **15:**

Lun ATL, McCarthy DJ & Marioni JC (2016) A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. *F1000Research* **5:** 2122

Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, Tirosh I, Bialas AR, Kamitaki N, Martersteck EM, Trombetta JJ, Weitz DA, Sanes JR, Shalek AK, Regev A & McCarroll SA (2015) Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161:** 1202–1214

Marčenko VA & Pastur LA (1967) Distribution of Eigenvalues for Some Sets of Random Matrices. *Math. USSR-Sbornik* **1:** 457–483

Nguyen QH, Pervolarakis N, Nee K & Kessenbrock K (2018) Experimental considerations for single-cell RNA sequencing approaches. *Front. Cell Dev. Biol.* **6:**

Orgeur M, Martens M, Börno ST, Timmermann B, Duprez D & Stricker S (2018) A dual transcript-discovery approach to improve the delimitation of gene features from RNA-seq data in the chicken model. *Biol. Open* **7:**

Petit F, Sears KE & Ahituv N (2017) Limb development: A paradigm of gene regulation. *Nat. Rev. Genet.* **18:** 245–258

Qiu X, Mao Q, Tang Y, Wang L, Chawla R, Pliner HA & Trapnell C (2017) Reversed graph embedding resolves complex single-cell trajectories. *Nat. Methods* **14:** 979–982

Saelens W, Cannoodt R, Todorov H & Saeys Y (2019) A comparison of single-cell trajectory inference methods. *Nat. Biotechnol.* **37:** 547–554

Schmid M, Smith J, Burt DW, Aken BL, Antin PB, Archibald AL, Ashwell C, Blackshear PJ, Boschiero C, Brown CT, Burgess SC, Cheng HH, Chow W, Coble DJ, Cooksey A, Crooijmans RPMA, Damas J, Davis RVN, De Koning DJ, Delany ME, et al (2015) Third Report on Chicken Genes and Chromosomes 2015: *Cytogenet. Genome Res.* **145:** 78–179

Scialdone A, Natarajan KN, Saraiva LR, Proserpio V, Teichmann SA, Stegle O, Marioni JC & Buettner F (2015) Computational assignment of cell-cycle stage from single-cell transcriptome data. *Methods* **85:** 54–61

Shekhar K, Lapan SW, Whitney IE, Tran NM, Macosko EZ, Kowalczyk M, Adiconis X, Levin JZ, Nemesh J, Goldman M, McCarroll SA, Cepko CL, Regev A & Sanes JR (2016) Comprehensive Classification of Retinal Bipolar Neurons by Single-Cell Transcriptomics. *Cell* **166:** 1308-1323.e30

Shenker S, Miura P, Sanfilippo P & Lai EC (2015) IsoSCM: Improved and alternative 3' UTR annotation using multiple change-point inference. *Rna* **21:** 14–27

Smith CM, Hayamizu TF, Finger JH, Bello SM, McCright IJ, Xu J, Baldarelli RM, Beal JS, Campbell J, Corbani LE, Frost PJ, Lewis JR, Giannatto SC, Miers D, Shaw DR, Kadin JA, Richardson JE, Smith CL & Ringwald M (2019) The mouse Gene Expression Database (GXD): 2019 update. *Nucleic Acids Res.* **47:** D774–D779

Street K, Risso D, Fletcher RB, Das D, Ngai J, Yosef N, Purdom E & Dudoit S (2018) Slingshot: Cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics* **19:**

Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM, Hao Y, Stoeckius M, Smibert P & Satija R (2019) Comprehensive Integration of Single-Cell Data. *Cell* **177:** 1888-1902.e21

Stuart T & Satija R (2019) Integrative single-cell analysis. *Nat. Rev. Genet.* **20:** 257–272

Tarasov A, Vilella AJ, Cuppen E, Nijman IJ & Prins P (2015) Sambamba: Fast processing of NGS alignment formats. *Bioinformatics* **31:** 2032–2034

Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M, Lennon NJ, Livak KJ, Mikkelsen TS & Rinn JL (2014) The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32:** 381–386

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, Van Baren MJ, Salzberg SL, Wold BJ & Pachter L (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28:** 511–515

Vento-Tormo R, Efremova M, Botting RA, Turco MY, Vento-Tormo M, Meyer KB, Park JE, Stephenson E, Polański K, Goncalves A, Gardner L, Holmqvist S, Henriksson J, Zou A, Sharkey AM, Millar B, Innes B, Wood L, Wilbrey-Clark A, Payne RP, et al (2018) Single-cell reconstruction of the early maternal–fetal interface in humans. *Nature* **563:** 347–353

Zeller R, López-Ríos J & Zuniga A (2009) Vertebrate limb bud development: Moving towards integrative analysis of organogenesis. *Nat. Rev. Genet.* **10:** 845–858

Zhang Z, Schwartz S, Wagner L & Miller W (2000) A greedy algorithm for aligning DNA sequences. *J. Comput. Biol.* **7:** 203–214

Zhao S & Zhang B (2015) A comprehensive evaluation of ensembl, RefSeq, and UCSC annotations in

the context of RNA-seq read mapping and gene quantification. *BMC Genomics* **16:**

Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, Ziraldo SB, Wheeler TD, McDermott GP, Zhu J, Gregory MT, Shuga J, Montesclaros L, Underwood JG, Masquelier DA, Nishimura SY, Schnall-Levin M, Wyatt PW, Hindson CM, Bharadwaj R, et al (2017) Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8:** 1–12

Chapter 2

A SINGLE-CELL TRANSCRIPTOMIC ATLAS OF THE DEVELOPING CHICKEN LIMB

Christian Feregrino, Fabio Sacher, Oren Parnas and Patrick Tschopp

Abstract

Through precise implementation of distinct cell type specification programs, differentially regulated in both space and time, complex patterns emerge during organogenesis. Thanks to its easy experimental accessibility, the developing chicken limb has long served as a paradigm to study vertebrate pattern formation. Through decades' worth of research, we now have a firm grasp on the molecular mechanisms driving limb formation at the tissue-level. However, to elucidate the dynamic interplay between transcriptional cell type specification programs and pattern formation at its relevant cellular scale, we lack appropriately resolved molecular data at the genome-wide level. Here, making use of droplet-based single-cell RNA-sequencing, we catalogue the developmental emergence of distinct tissue types and their transcriptome dynamics in the distal chicken limb, the so-called autopod, at cellular resolution. Using single-cell RNA-sequencing technology, we sequenced a total of 17,628 cells coming from three key developmental stages of chicken autopod patterning. Overall, we identified 23 cell populations with distinct transcriptional profiles. Amongst them were small, albeit essential populations like the apical ectodermal ridge, demonstrating the ability to detect even rare cell types. Moreover, we uncovered the existence of molecularly distinct sub-populations within previously defined compartments of the developing limb, some of which have important signaling functions during autopod pattern formation. Finally, we inferred gene co-expression modules that coincide with distinct tissue types across developmental time, and used them to track patterning-relevant cell populations of the forming digits. We provide a comprehensive functional genomics resource to study the molecular effectors of chicken limb patterning at cellular resolution. Our single-cell transcriptomic atlas captures all major cell populations of the developing autopod, and highlights the transcriptional complexity in many of its components. Finally, integrating our data-set with other single-cell transcriptomics resources will enable researchers to assess molecular similarities in orthologous cell types across the major tetrapod clades, and provide an extensive candidate gene list to functionally test cell-type-specific drivers of limb morphological diversification.

Feregrino et al. BMC Genomics (2019) 20:401
https://doi.org/10.1186/s12864-019-5802-2

BMC Genomics

RESEARCH ARTICLE Open Access

A single-cell transcriptomic atlas of the developing chicken limb

Christian Feregrino[1], Fabio Sacher[1], Oren Parnas[2,3] and Patrick Tschopp[1*]

Abstract

Background: Through precise implementation of distinct cell type specification programs, differentially regulated in both space and time, complex patterns emerge during organogenesis. Thanks to its easy experimental accessibility, the developing chicken limb has long served as a paradigm to study vertebrate pattern formation. Through decades' worth of research, we now have a firm grasp on the molecular mechanisms driving limb formation at the tissue-level. However, to elucidate the dynamic interplay between transcriptional cell type specification programs and pattern formation at its relevant cellular scale, we lack appropriately resolved molecular data at the genome-wide level. Here, making use of droplet-based single-cell RNA-sequencing, we catalogue the developmental emergence of distinct tissue types and their transcriptome dynamics in the distal chicken limb, the so-called autopod, at cellular resolution.

Results: Using single-cell RNA-sequencing technology, we sequenced a total of 17,628 cells coming from three key developmental stages of chicken autopod patterning. Overall, we identified 23 cell populations with distinct transcriptional profiles. Amongst them were small, albeit essential populations like the apical ectodermal ridge, demonstrating the ability to detect even rare cell types. Moreover, we uncovered the existence of molecularly distinct sub-populations within previously defined compartments of the developing limb, some of which have important signaling functions during autopod pattern formation. Finally, we inferred gene co-expression modules that coincide with distinct tissue types across developmental time, and used them to track patterning-relevant cell populations of the forming digits.

Conclusions: We provide a comprehensive functional genomics resource to study the molecular effectors of chicken limb patterning at cellular resolution. Our single-cell transcriptomic atlas captures all major cell populations of the developing autopod, and highlights the transcriptional complexity in many of its components. Finally, integrating our data-set with other single-cell transcriptomics resources will enable researchers to assess molecular similarities in orthologous cell types across the major tetrapod clades, and provide an extensive candidate gene list to functionally test cell-type-specific drivers of limb morphological diversification.

Keywords: scRNA-seq, Gene expression, Cellular transcriptomics, Autopod patterning, Digits, Interdigit, Perichondrium, Phalanges

Background

Embryonic pattern formation relies on the tight coordination of numerous developmental processes, across multiple scales of complexity. From seemingly homogenous progenitor populations, different cell types get specified and arranged in intricate patterns, to give rise to functional tissues and organs. As progenitors mostly share a common genome, this phenotypic specialization relies on the precise execution of distinct gene regulatory networks, to enable cell type specification and ensuing pattern formation [1–3]. Slight deviations in these processes contribute to morphological variations within natural populations. More profound aberrations, however, can cause malformations and ultimately result in death of the embryo. To buffer such fragile balance, many cell type specification and pattering processes rely on complex feedback mechanisms, through tightly interconnected molecular loops between spatially distinct signaling centers [4–6]. Hence, integration of multiple

* Correspondence: patrick.tschopp@unibas.ch
[1]DUW Zoology, University of Basel, Vesalgasse 1, CH-4051 Basel, Switzerland
Full list of author information is available at the end of the article

signaling pathways across space and time defines a molecular coordinate grid to instruct organogenesis at the tissue level. Ultimately, however, these multifaceted signaling inputs have to be incorporated at the cellular level, via cell type-specifying gene regulatory networks, as progenitor cells undergo spatially and temporally defined cell fate decisions to contribute to proper pattern formation.

Tetrapod limb development has long served as a model to study the genetic and molecular underpinnings of vertebrate pattern formation. Due to its non-essentiality for embryo survival, many fetuses carrying mutations that affect limb development make it to full term. Accordingly, human geneticists have been able to accumulate an impressive catalogue of candidate genes for limb patterning [7–9]. Combined with the easy accessibility of the limb in chicken embryos, and molecular genetic tools in the mouse, decades of experimental work have resulted in an in-depth understanding of many of the molecular mechanisms driving limb formation at the tissue scale [5]. Moreover, given the profound morphological diversifications the basic limb structure has experienced in numerous tetrapod clades, limb development has long attracted the interests of comparative developmental biologists using 'EvoDevo' approaches [10]. This holds especially true for the most distal portion of the limb, the autopod, i.e. hands and feet. There, species-specific adaptations to distinct modes of locomotion have resulted in a diverse array of digit number formulas and individualized digit patterns [11–14].

Early in development, proliferation of a lateral plate mesoderm (LPM)-derived mesenchymal progenitor population drives overall limb bud outgrowth. Signaling crosstalk with a specialized structure of the distal overlaying ectoderm, the apical ectodermal ridge (AER), controls these dynamics. Concurrently, the major embryonic axes of the limb are defined by the coordinated action of multiple signaling centers [reviewed in 5]. As development progresses, LPM-derived progenitors start to differentiate into skeletal and other connective tissue types [15–17], while muscles cells originating from the somites migrate into the limb bud to complement formation of the musculoskeletal apparatus [18, 19]. For autopod pattern formation, digit numbers and identities are first defined by posteriorly restricted sonic hedgehog (SHH) activity, and altered by modulations therein ([10, 14, 20], reviwed in [21]). Digit elongation then relies on a specialized distal progenitor population, which supports outgrowth of individual digit bones, the phalanges [22, 23]. Digit-specific phalanx-formulas, and their stereotypic connection patterns via synovial joints, are established by signals emanating from the posterior interdigit mesenchyme [24, 25].

In this study, capitalizing on the power of droplet-based single-cell RNA-sequencing, we resolve the underlying transcriptional dynamics of autopod tissue formation and pattern emergence at single-cell resolution, across three stages of chicken hindlimb development. In total, we present transcriptomic data for 17,628 cells, allowing us to identify all major tissue types of the developing limb, as well as a substantial amount of molecular heterogeneity therein. Through weighted correlation network analysis, we define distinct gene co-expression modules that track corresponding tissue types across developmental time. Finally, we focus on the molecular make-up of cell populations involved in digit pattern formation and, hence, putative drivers of morphological diversification in the autopod.

Collectively, we present a comprehensive genomics resource that for the first time reveals the transcriptome dynamics of the developing chicken foot at the cellular level. Our study identifies a range of marker genes in co-expression modules of patterning-relevant cell populations. Thereby, we provide an extensive catalogue of candidate genes for functional follow-up studies to elucidate the molecular mechanisms of autopod pattern formation and diversification.

Results

Singe-cell sampling of the developing distal chicken limb

To follow the appearance of patterning-relevant cell populations and their associated transcriptome dynamics, we sampled three developmental stages of the embryonic chicken foot: stage Hamburger-Hamilton 25 (HH25, ~ 4.5 days of development), stage HH29 (~ 6 days of development) and stage HH31 (~ 7 days of development). This time window spans key morphogenetic events that drive species-specific patterns in the developing autopod, particularly for the skeletal apparatus and its associated tissues. Namely, stage HH25 is dominated by overall autopod outgrowth and delineation of the main embryonic axes, at HH29 digit-specific patterns differentiate, and at HH31 digit elongation is phasing out. We designed our tissue sampling strategies accordingly. At HH25, we captured the entire distal part of the growing limb (Fig. 1a), at HH29 we dissected two digits with distinct skeletal formulas, digit 3 and 4, as well as their adjacent interdigit mesenchyme (Fig. 1b), and at HH31 we focused on the tip of digit 4 with its growth-relevant progenitor population (Fig. 1c). We dissociated the micro-dissected tissue pieces using enzymatic digest combined with mechanical shearing and prepared single-cell suspensions for droplet-based high-throughput single-cell RNA-sequencing (*10X Genomics* and *Drop-Seq* [26, 27]). Using the corresponding bioinformatics pipelines, the resulting Next-Generation Sequencing libraries were mapped to the chicken genome, de-multiplexed according to their cellular barcodes and quantified to generate gene/cell UMI (unique molecular identifier) count tables. In total, we sampled over 17,000 cells and obtained single-cell transcriptomic profiles for 5982 (HH25), 6823 (HH29) and 4823 (HH31) individual

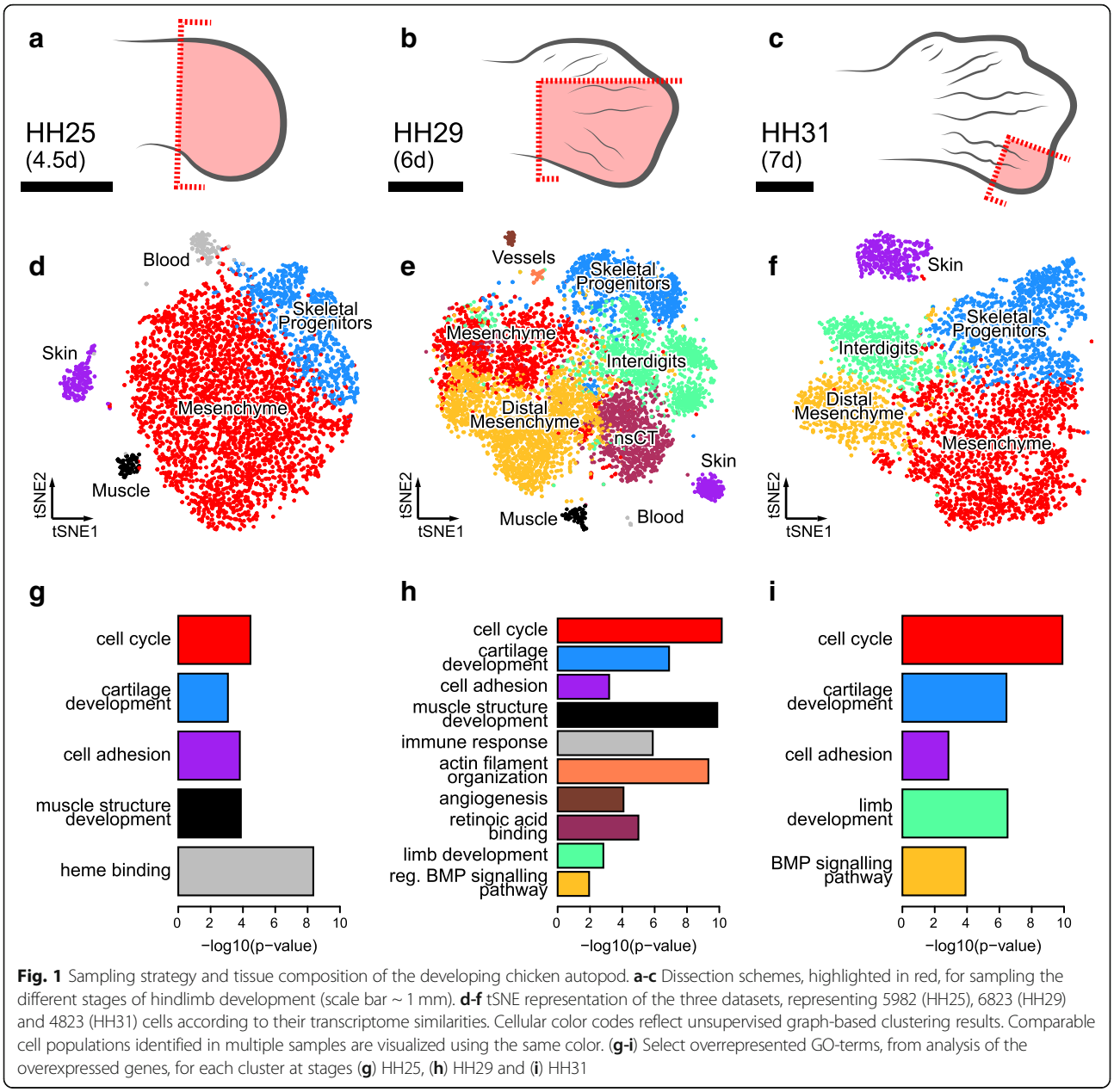


Fig. 1 Sampling strategy and tissue composition of the developing chicken autopod. **a-c** Dissection schemes, highlighted in red, for sampling the different stages of hindlimb development (scale bar ~ 1 mm). **d-f** tSNE representation of the three datasets, representing 5982 (HH25), 6823 (HH29) and 4823 (HH31) cells according to their transcriptome similarities. Cellular color codes reflect unsupervised graph-based clustering results. Comparable cell populations identified in multiple samples are visualized using the same color. (**g-i**) Select overrepresented GO-terms, from analysis of the overexpressed genes, for each cluster at stages (**g**) HH25, (**h**) HH29 and (**i**) HH31

cells, respectively (Additional file 1: Figure S1a). Quality-based exclusion of single-cell transcriptomes was implemented based on mean library size, percentage of mitochondrial reads and number of genes detected per cell. Additionally, data normalization as well as batch and cell cycle corrections were performed (for details, please refer to the *Methods* section). On average, we detected 2879 UMIs and 1081 genes per cell (Additional file 1: Figure S1b,c).

Autopod tissue composition at cellular resolution

Using unsupervised graph-based clustering, we identified 5, 10 and 5 clusters at stages HH25, HH29 and HH31, respectively. Projecting these clusters onto stage-specific tSNE (t-Distributed Stochastic Neighbor Embedding [28]), plots of our cellular transcriptomes revealed the presence of a dominant bulk of cells, with varying degrees of sub-structure, as well as distinct outlier groups (Fig. 1 d-f). Based on the expression of known marker genes and gene ontology (GO)-term enrichment analyses, we were able to attribute these broadly defined cell populations to distinct tissue types (Fig. 1g-i, Additional file 1: Figure S1a and Figure S2a-c). At stage HH25, they comprise a largely undifferentiated and proliferating mesenchymal population (red), early skeletal progenitors (blue), muscle cells invading the limb (black), as well as skin (purple) and blood cells (grey) (Fig. 1d,g). We recovered cell

populations corresponding to those same five tissue types in our HH29 sample, with the exception that the "blood cluster" was now dominated by white blood cells and not erythrocytes. Additionally, we identified cell populations matching the interdigit mesenchyme (green), non-skeletal connective tissue (nsCT, maroon), cells enriched for markers of the very distal margin of the autopod mesoderm ("distal mesenchyme", yellow), as well as endothelial (brown) and smooth muscle (orange) cells of the forming blood vessels (Fig. 1e,h). At stage HH31, we again find a largely undifferentiated mesenchymal population, the interdigit and distal margin mesenchyme, skeletal and skin cells (Fig. 1f,i). As expected according to our sampling strategy, for spatial and/or temporal context, we did not find all cell populations in every dataset. For example, while sample HH25 is biggest in relative size to the autopod, it is the earliest stage and thus predictably displayed the lowest cellular complexity. Conversely, even though development and cell type specification have advanced furthest in our HH31 sample, microdissection of only the tip of digit 4 prevented the capture of more diverse cell populations (Fig. 1c). Hence, our most complex dataset, in terms of cell number and tissue types identified, is from stage HH29. Collectively, using broad graph-based clustering and molecular profiling on our single-cell transcriptomics data, we catalogued the tissue composition of the developing autopod with cellular resolution, across three developmental stages.

Fine-scale clustering and marker gene expression across developmental time

Although all expected major tissue types were recovered in our primary analyses, smaller cell populations, some well known to be essential for limb outgrowth and patterning, remained elusive. Hence, given our sampling depth, we next examined our data for additional sub-structure. Indeed, upon closer inspection using finer-tuned clustering parameters, we did find additional sub-populations with distinct transcriptional signatures (Fig. 2a-c, Additional file 1: Figure S1a). Based on differential expression analyses, we identified marker genes for each of these sub-populations (Additional file 2, Additional file 3, Additional file 4). Certain sub-population/marker gene-combinations appeared to be conserved in all three samples, thereby allowing us to assign cellular equivalencies across developmental time (Fig. 2d-f). A subset of marker genes only showed loosely restricted expression patterns, likely a reflection of the largely undifferentiated state of the corresponding sub-population. For example, *PRRX1*, a well-established marker of the limb mesenchyme [16, 29, 30], and *PCNA*, active during DNA replication in proliferating cells [31], showed varying levels of expression beyond the proliferating mesenchyme sub-clusters. Such transcriptional ambiguities, however, seemed progressively lost, as mesenchymal progenitors committed to the different skeletal and non-skeletal lineages that define the emerging autopod patterns (Fig. 2d-f). As expected, cell sub-populations residing outside the LPM-lineage showed more pronounced transcriptome individualizations. For example, at HH25 the ectodermal 'skin' population got split into two distinct sub-clusters, one representing the bulk amount of the embryonic skin covering the autopod (sub-cluster 8), and the other corresponding to the apical ectodermal ridge (sub-cluster 7). Expression of its canonical marker *FGF8* and other highly enriched genes clearly established AER identity, demonstrating that even small cell populations can be successfully captured (Fig. 2d).

Gene co-expression modules and corresponding tissue types

To gain further insights into the regulatory programs that maintain these transcriptional signatures, and explore their potential biological significance, we tested for the occurrence of transcriptome-wide gene co-expression patterns using weighted correlation network analysis (WGCNA) [32]. This approach consists of an unsupervised clustering of genes based on their expression pattern across all cells, irrespective of the assigned cell or tissue type. In order to comprehensively screen for relevant gene co-expression modules, we conducted the analysis in our transcriptionally most complex sample at stage HH29. Starting with genes that showed high levels and variation of expression, we calculated an adjacency matrix and its topological overlap to construct a hierarchical tree. The resulting tree was cut to obtain a first set of gene co-expression modules. We then computed the first principal component of each module, to define so-called 'module eigengenes'. For each individual gene, correlation to the respective eigengenes was used to assess module membership. Genes not significantly correlated with any eigengene were discarded, after which the entire process was repeated iteratively with a reduced gene set. Eventually, we identified a total of 836 genes grouped in 16 distinct gene co-expression modules, each designated by a color (Fig. 3a). Final module sizes ranged from 15 to 215 genes (Additional file 5).

On a cell-by-cell basis, we calculated the average expression for each of the co-expression modules and visualized their distribution on our stage HH29 tSNE plot (Additional file 1: Figure S3). Compared to our initial clustering of sample HH29, we found co-expression modules specifically enriched in the following cell populations: blood cells (module Black), skin (Blue), blood vessel endothelium (Brown), nsCT (Darkgrey), distal mesenchyme (Magenta), chondrocytes (Red and Turquoise) and muscle (Yellow). Interestingly, GO-terms associated with more broadly distributed modules enabled us to attribute the sub-clustering structure of certain tissues to particular biological processes. For example, HH29 mesenchyme sub-cluster 5 showed higher activity

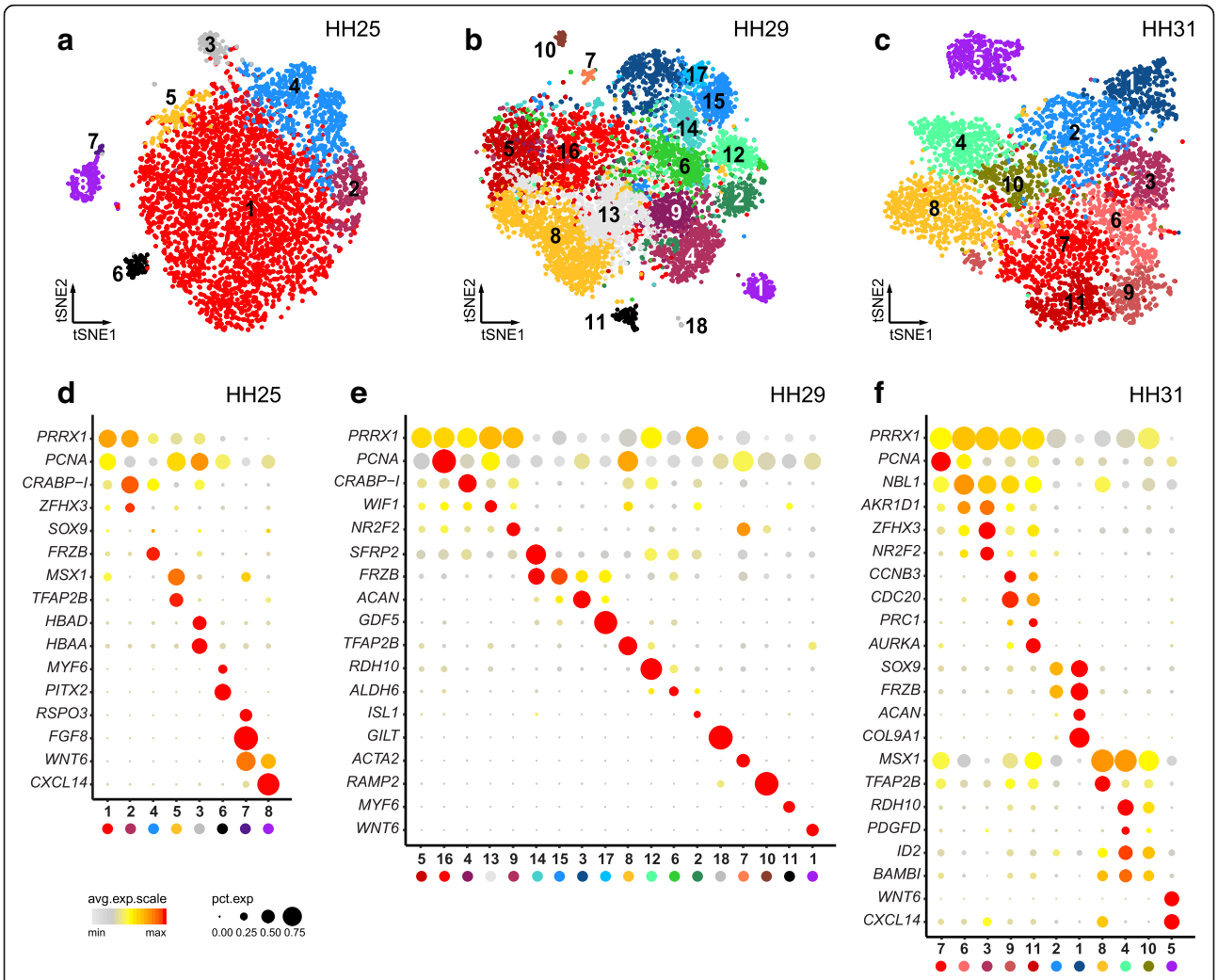


Fig. 2 Cell population sub-structure and marker gene expression (**a-c**) tSNE plots of the three datasets. Colors now represent fine-tuned unsupervised graph-based clustering, with similar colorations relating to the results of the first clustering step. Comparable cell populations identified in multiple samples are visualized using the same color. For reference, sub-cluster numbers are added. **d-f** Dot plots of sub-cluster marker gene expression. Averaged expression level (heatmap) and percentage of cells showing > 0 expression (dot size) is visualized across all samples, for all identified sub-clusters. Same color-coding for sub-clusters identification is used as in (**a-c**)

for module Green, associated with GO-terms connected to mitosis, whereas sub-cluster 16 was enriched for module Pink, linked to G2/M-transition-related genes (Additional file 1: Figure S3). Hence, we reasoned that distinct cell-cycle states underlie the subdivision of the proliferating mesenchyme cluster. Likewise, HH29 interdigit sub-clusters 2, 6 and 12 were closely matched by the activities of modules Tan, Olivegreen, Orange and Midnightblue (see below, Fig. 4a-h).

To follow the developmental dynamics of the identified modules, we calculated their averaged activities across all the three sampled time points, and visualized similarities across time and tissue types using unsupervised hierarchical clustering (Fig. 3b). Indeed, despite differences in embryonic stages and experimental platforms, we were able to confirm corresponding cell and tissue types between our samples. For example, what we refer to as the "distal mesenchyme" is a population of cells characterized by high activity of the co-expression module Magenta at all time points (Fig. 3c-f). Comparisons to published expression patterns for *TFAP2B, WNT5A*, *MSX1* and *MSX2* confirmed its distal location and, based on those genes' functions, suggested a role for this cell population in controlling distal autopod outgrowth. Using WGCNA thus enabled us to define equivalent cell populations across developmental time, and helped attribute biological functions at the sub-cluster level.

Transcriptionally and spatially distinct sub-populations in the interdigit mesenchyme

As expected by developmental stage, interdigit populations were only recovered in samples HH29 and HH31.

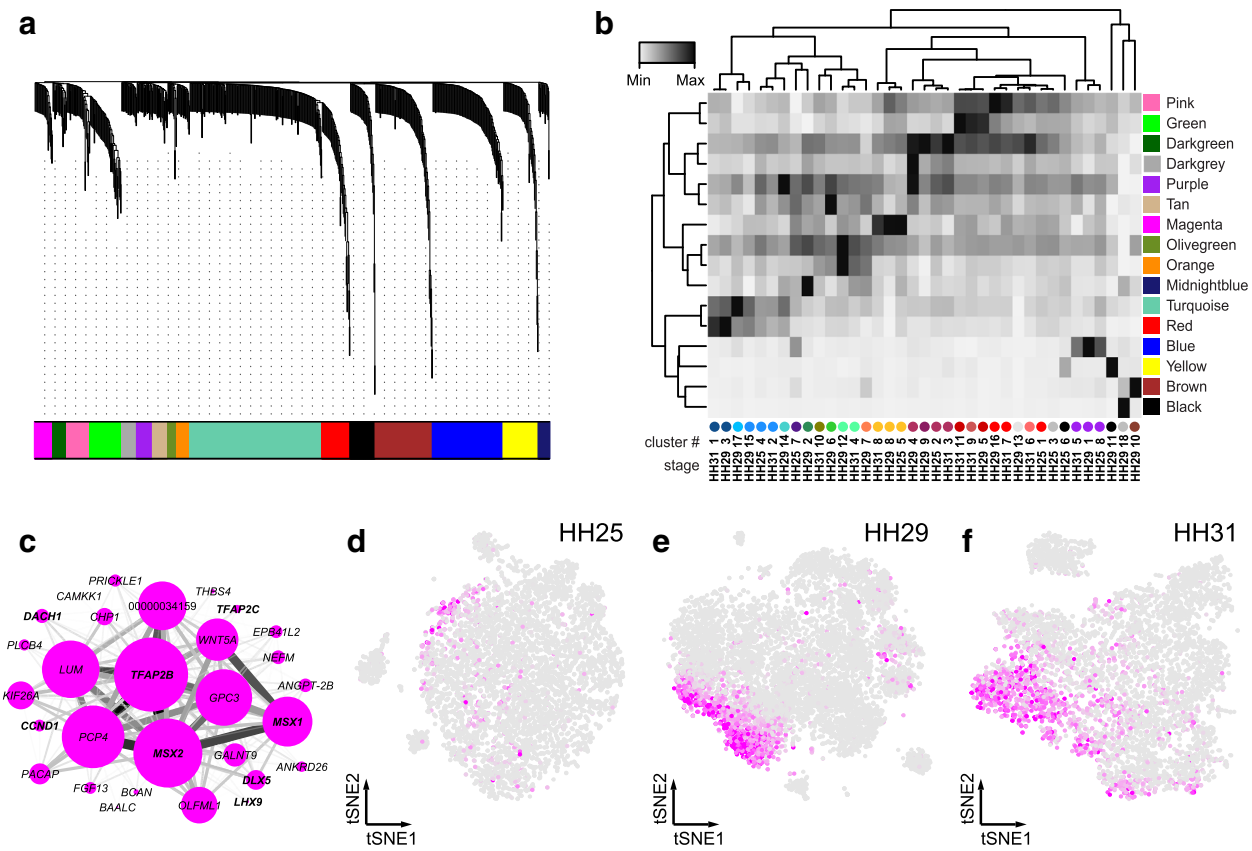


Fig. 3 Weighted correlation network analysis and gene co-expression modules**. a** WGCNA gene hierarchical clustering dendrogram and modules of co-expression. A total of 16 distinct co-expression modules are identified, visualized by colored bars at the bottom of the dendrogram (color scheme unrelated to previous cell clustering). **b** Heatmap of mean expression values per co-expression module, calculated across distinct cell sub-clusters and developmental stages. Ordering based on hierarchical clustering of averaged co-expression module activities and sub-clusters. Sub-clusters identification at bottom (number and color code) corresponds to Fig. 2a-c. **c** *Cytoscape* visualization of co-expression module Magenta. Node size is proportional to module membership of each gene, edge thickness represents correlation of pair-wise gene co-expression. **d-f** Heatmap representing the averaged cellular activity of the Magenta module, plotted on tSNE representations of the different samples. Color intensity is proportional to the mean expression of the module in each cell

In total, we identified four associated co-expression modules (Fig. 4a-d). High Orange and Olivegreen module activities were coinciding with the same interdigit sub-population (Fig. 4e, f), which was recognizable in both HH29 and HH31 samples and marked by *RDH10* expression (Fig. 2e,f). Noticeably, all genes with high membership in module Olivegreen were transcription factors (TFs), while module Orange was enriched for enzymatic activities (Fig. 4a,b). Both, however, scored high for GO-terms related to retinoic acid signaling, an important mediator of interdigit cell death [33]. Module Tan was enriched for skeletogenic and morphogenetic GO-terms, suggesting it might mediate some of the patterning information contained in the interdigit mesenchyme to the adjacently forming digits (Fig. 4c,g). Lastly, module Midnightblue showed multiple TFs and its activity was restricted to HH29 sub-cluster 2 (Fig. 4d,h).

Since relevant patterning information is contained in the interdigit, posteriorly adjacent to each forming digit, we next wondered whether some of the sub-clustering structure corresponded to spatially distinct interdigit populations along the anterior-posterior axis of the autopod. At HH29, we detected three interdigit sub-clusters (Fig. 4i). Using differential expression analyses, we defined marker genes that distinguish the three sub-clusters from each other (Fig. 4j). To assign putative spatial information to our single-cell interdigit transcriptomes, we reanalyzed a bulk RNA-seq dataset covering stages HH29 and HH31 of the developing chicken hindlimb autopod [34]. This dataset is based on dissections of individual digits, together with their posteriorly associated interdigit mesenchyme, and thus provided an opportunity to identify spatially resolved marker genes. We contrasted their transcriptomic data of digit/interdigit III against digit/interdigit IV and found a total of 54 genes to be significantly differentially expressed at both developmental time points (Fig. 4k). Comparing the digit/interdigit IV-specific subset of these genes to our

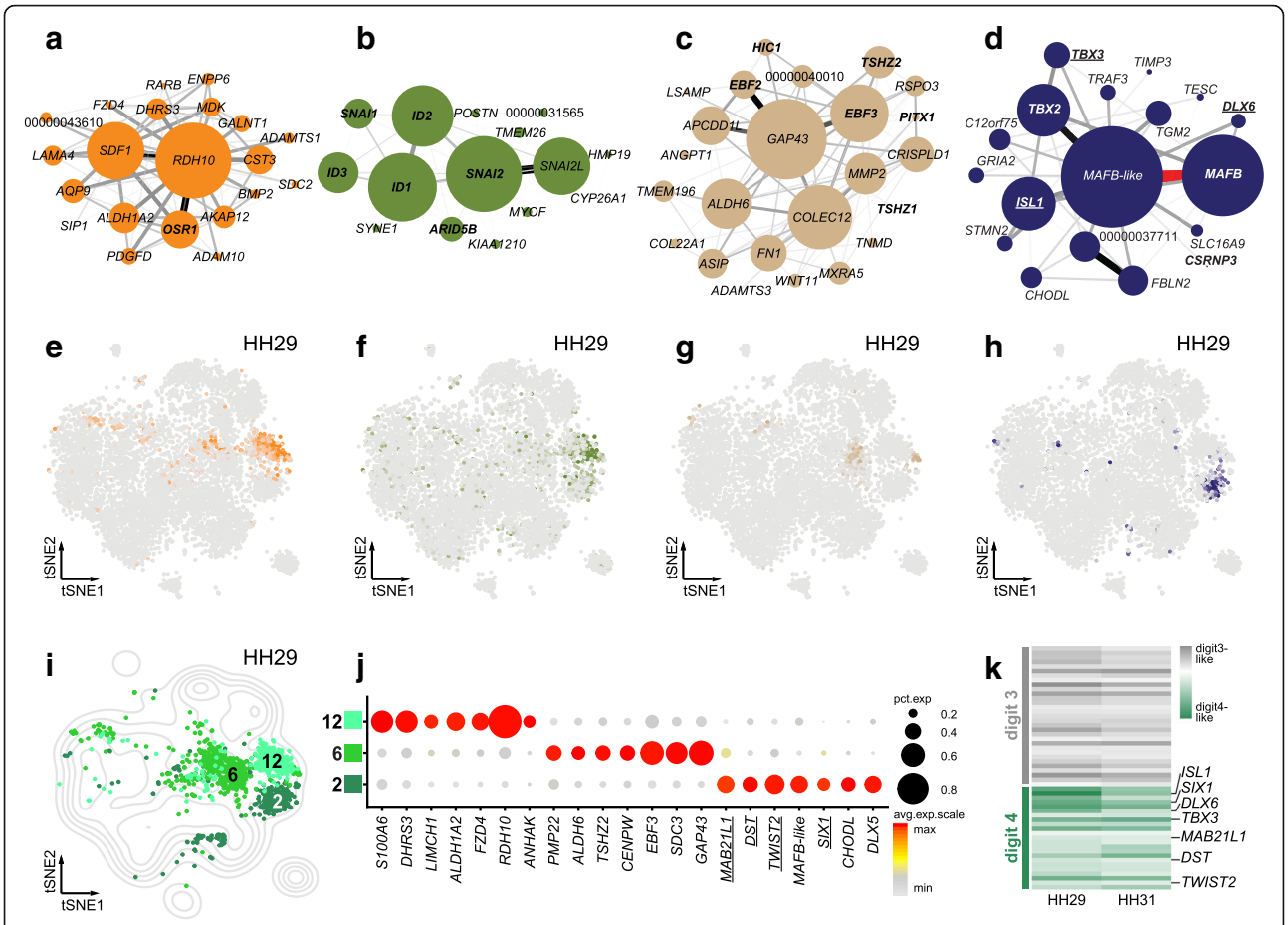


Fig. 4 Molecular and spatial heterogeneity in the interdigit mesenchyme. **a-g** Interdigit-associated co-expression modules (**a**) Orange, (**b**) Olivegreen, (**c**) Tan, and (**d**) Midnightblue. Node size represents gene module membership, edge thickness gene pair-wise correlation. Gene names in bold are classified as transcription factors, uncharacterized genes show only Ensembl numbers following the "ENSGALG" gene code. **e-h** Heatmaps of averaged activity levels of the corresponding modules, visualized on top of a tSNE plot for sample HH29. i Contour density plot of the tSNE projection for sample HH29, to delineate overall cell distribution. Partial tSNE plot on top, to visualize only cells belonging to interdigit-like sub-clusters (Color-coding and numbering according to Fig. 2b). **j** Expression dot plot of differentially expressed genes between the three interdigit sub-clusters at stage HH29. **k** Heatmap visualization of "digit3-like" and "digit4-like" gene sets at stages HH29 and HH31, based on differential expression analysis of digit-specific bulk RNA-seq data by Wang et al., 2011. Underlined gene names in **d,j** denote membership to the "digit IV-like" gene set

differential expression analysis of sub-cluster 2, and its affiliated module Midnightblue, we found an overlap of seven up-regulated genes (Fig. 4d,j, underlined). In contrast, we couldn't find any other digit/interdigit IV gene in the rest of the interdigit sub-cluster signatures or co-expression modules. We therefore concluded that HH29 sub-cluster 2 consisted of cells of the interdigit mesenchyme posterior to digit 4.

Developing digits and their associated tissues

Of the cell populations directly contributing to the making of digits, a cluster reminiscent of the non-skeletal connective tissue, the nsCT, appeared in all of the samples. In our WGCNA analyses, we identified three modules, Darkgrey, Purple, and Darkgreen, which mapped to the nsCT sub-clusters (Fig. 5a-f). The Darkgrey module was most restricted, in both time and cell numbers, and its activity pattern closely matched the HH29 sub-cluster 4 (Fig. 5d). Cellular retinoic acid binding protein I *CRABP-I*, Aquaporin *AQP1*, *DKK2* and *GLT8D2* were the genes most strongly associated with this module. Modules Purple and Darkgreen showed more widespread activities (Fig. 5e,f), and centered on *COL1A2*, *DCN*, *KCNJ2*, *SALL1*, and *AKR1D1*, *PRRX1*, *TCF12*, *ZFHX3*. By performing stage-specific differential expression analyses for our nsCT clusters (HH25–2, HH29–9/4, HH31–3; Fig. 2), we noticed a progressive maturation of nsCT signatures, with HH31–3 showing the highest degree of transcriptional differentiation (Fig. 5g). Overall, however, nsCT signatures appeared developmentally dynamic and only six genes were significantly enriched across all stages (Fig. 5g), five of which also

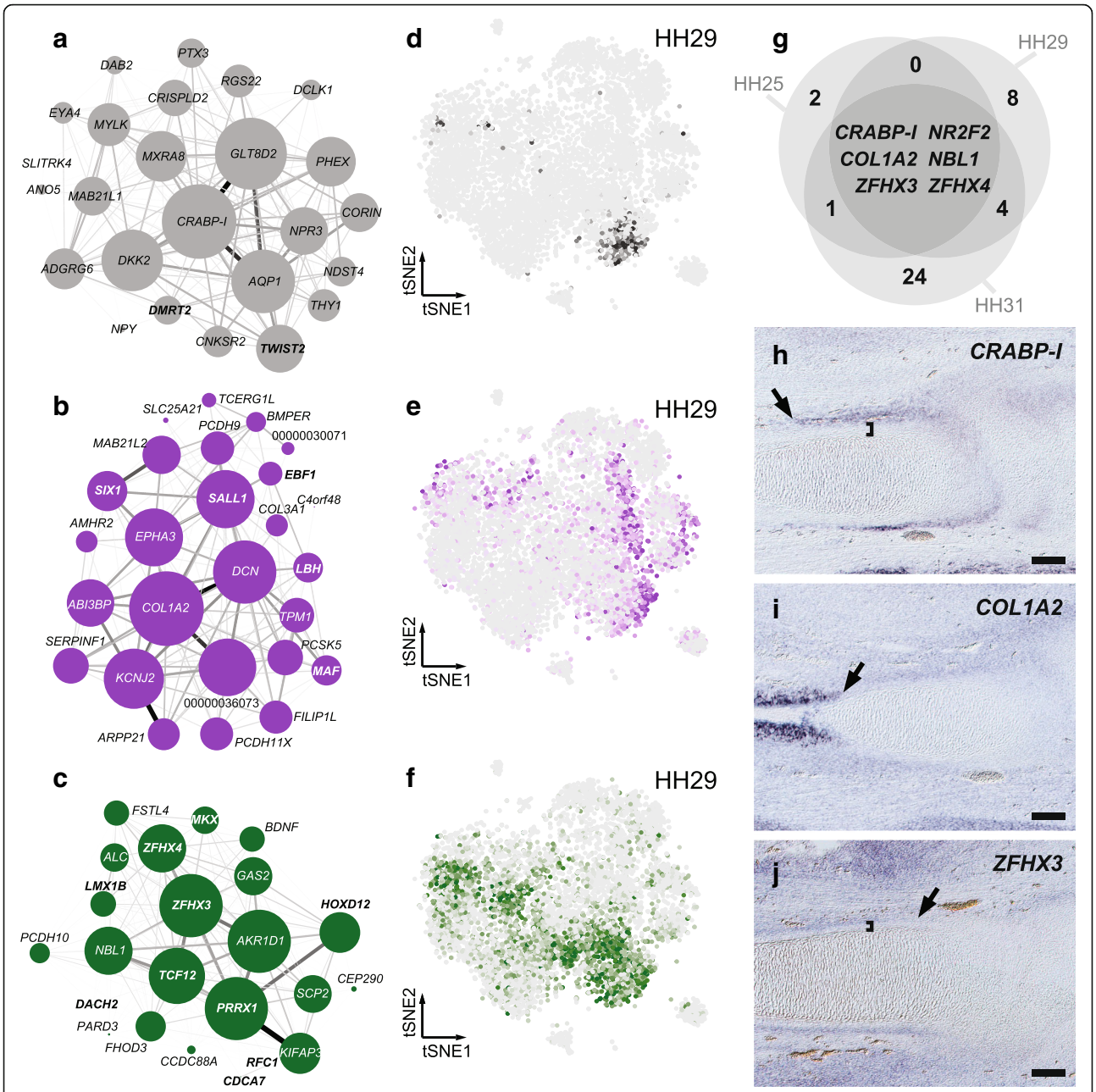


Fig. 5 Transcriptional modules in the non-skeletal connective tissue (nsCT). **a-c** Gene co-expression modules (**a**) Darkgray, (**b**) Purple and (**c**) Darkgreen enriched for peri-skeletal genes. Gene names in bold are classified as transcription factors, uncharacterized genes show only Ensembl numbers following the "ENSGALG" gene code. **d-f** Corresponding averaged module activities visualized as heatmaps on stage HH29 tSNE plots. **d** Venn diagram of shared overexpressed genes in the nsCT populations of the three samples. **h-i** Section in situ hybridization on stage HH31 chicken hindlimbs for three shared nsCT marker genes, *CRABP-I*, *COL1A2* and *ZFHX3*. Arrows denote extent of expression along the long bone axis, while brackets indicate separation from the forming skeletal element (scale bar = 100 μm)

appeared in our nsCT modules. Using in situ hybridization for the top-three of these genes, both differential expression- and module membership-wise, allowed us to attribute module activities to discrete nsCT domains along the developing skeletal elements. *CRABP-I* showed highest expression near and around the forming epiphysis, where synovial joints and ligament attachment sites develop (Fig. 5h). *COL1A2*- and *ZFHX3*-positive populations showed a graded distribution along the periskeletal tissue layer, predominantly marking the prospective periosteum and perichondrium domains, respectively (Fig. 5i,j).

Finally, we identified skeletal progenitor populations at all three time points (Fig. 6a-c). According to the developmental stages we sampled, only cartilage-producing

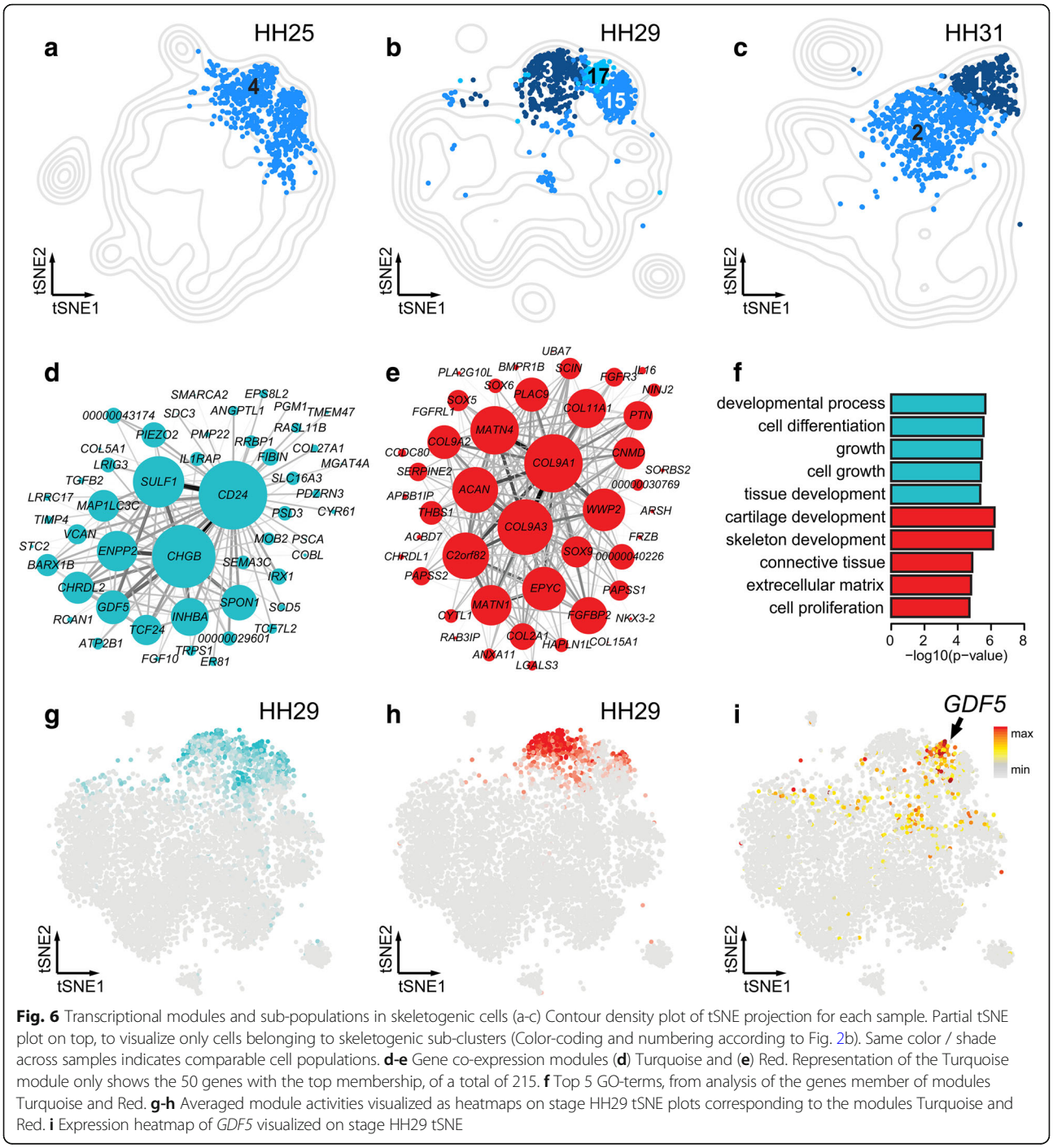


Fig. 6 Transcriptional modules and sub-populations in skeletogenic cells (a-c) Contour density plot of tSNE projection for each sample. Partial tSNE plot on top, to visualize only cells belonging to skeletogenic sub-clusters (Color-coding and numbering according to Fig. 2b). Same color / shade across samples indicates comparable cell populations. **d-e** Gene co-expression modules (**d**) Turquoise and (**e**) Red. Representation of the Turquoise module only shows the 50 genes with the top membership, of a total of 215. **f** Top 5 GO-terms, from analysis of the genes member of modules Turquoise and Red. **g-h** Averaged module activities visualized as heatmaps on stage HH29 tSNE plots corresponding to the modules Turquoise and Red. **i** Expression heatmap of *GDF5* visualized on stage HH29 tSNE

skeletal cells were recovered. In all three samples, we found a cell population resembling early chondrocytes (sub-clusters HH25–4, HH29–15 and HH31–2). At stages HH29 and HH31, a seemingly more mature chondrocyte type emerged (HH29–3, HH31–1), and an additional cartilaginous cluster was evident in the HH29 sample (HH29–17). Concomitantly, we identified two co-expression modules associated with these cell populations, Turquoise and Red (Fig. 6d,e). Turquoise is centered on *CD24*, *CHGB* and *SULF1*, whereas module Red displays a core of collagens *COL9A1* and *COL9A3*, *MATN4*, *C9H2ORF82* (also known as *SNORC* in mammals), and *ACAN*. Based on additional marker genes and GO-term enrichment analyses, we inferred the Turquoise module to be related to early chondrocyte proliferation and growth, whereas the Red module reflected chondrocyte maturation and extracellular matrix deposition (Fig. 6f). Interestingly, compared to module Turquoise, the activity of

module Red was generally more restricted and specifically excluded from sub-cluster HH29–17 (Fig. 6g,h). Upon closer inspection, we identified high expression of several known synovial joint markers genes in this population, thus identifying it as the forming interphalangeal joints (Fig. 6i, Additional file 3).

Hence, through a combination of differential gene expression and GO-term enrichment analyses, as well as gene co-expression modules, we identified spatially and/or temporally distinct sub-populations and transcriptome dynamics in the skeletal and peri-skeletal tissues of the forming digits.

Discussion

Singe-cell tissue decomposition of the developing chicken autopod

Here, using single-cell RNA-sequencing, we present a transcriptomic atlas of the developing chicken limb at cellular resolution. Focusing on the distal and morphologically diverse portion of the limb, the autopod, we sampled over 17,000 single-cell transcriptomes with an average of over 1000 genes detected in each cell. Within our atlas, we identify all major tissue types that constitute and pattern the embryonic appendage across three developmental time points. Additionally, taking advantage of our cellular and transcriptomic sampling depth, we manage to isolate even minute cell populations like the AER and assemble lists of marker genes for them. We also distinguish transcriptionally discrete sub-populations within known major tissue types, reflecting distinct spatial locations or cellular states. As such, it demonstrates the power of scRNA-seq to molecularly disentangle cell populations of the developing limb that occur in close spatial or 'lineage' proximity. Historically, such populations have proven notoriously difficult to separate and characterize transcriptionally, using either manual tissue dissection or reporter-gene based cell lineage isolation. To what extent all of our tissue sub-clusters indeed correspond to distinct lineage separations [35], or rather represent the extremes of a molecular continuum that follows the inherently stochastic nature of transcription [36, 37], remains to be addressed in future studies. Regardless, however, our results provide a toolbox of candidate genes to tackle this question in a molecularly comprehensive manner. Furthermore, our data enables a characterization of emerging embryonic cell types based on transcriptional signatures, rather than relying on the definitive morphological and/or functional features of their mature counterparts.

Cell type equivalencies across developmental and evolutionary time

Such molecular classification schemes echo recent conceptual frameworks that aim to categorize 'cell types' across developmental and evolutionary time scales, irrespective of morphology or function [2]. If, however, we consider a 'cell type' to be primarily defined by the expression of distinct regulatory programs, then detection of program activities can substantially precede our ability to distinguish morphological or functional specializations. Indeed, our sub-clustering and module analyses across developmental time reveal the appearance of certain prospective cell types long before they become morphologically distinct. For example, already at stage HH25 we recover clear gene expression signatures reminiscent of the future periskeletal nsCT, even though prominent cartilage anlagen have yet to form (Fig. 2d, Fig. 3b). As such, it suggests an early lineage priming, without necessarily implying a definite switch in cell fate or clear morphological distinctions. In agreement with this, our *ZFHX3*-containing module Darkgreen appears to be the most basic and least specific of the co-expression modules that coincide with the nsCT population. We detect its activity at all three time points, marking the prospective nsCT as well as parts of the *PRRX1*-positive mesenchymal progenitor population (Fig. 5c,f). Only later do more mature and restricted nsCT sub-divisions and their corresponding co-expression modules occur, as exemplified by the activity of module Darkgrey and some of its members known to be involved in the formation of periskeletal tissues and tendon attachment sites (Fig. 5a,d) [38, 39].

Moreover, combining such transcriptome-based 'cell type' classification schemes with comparative scRNA-seq datasets allows for a molecular assessment of homologous cell types between species, across evolutionary time scales [40, 41]. This has important implications when trying to elucidate the impact of cell type-specifying gene regulatory networks on pattern formation and diversification at its relevant cellular scale. Namely, how progenitor populations exactly perceive and process patterning-relevant cues can be modulated by species-specific alterations in the respective cell type-specifying networks. In this context, it is worth noting that we detect *RSPO3* as one of the main markers of the chicken AER (Fig. 2d, Additional file 2). R-spondins, a family of secreted ligands involved in WNT-signaling, have previously been implicated in AER maintenance and control of limb outgrowth. However, in mammals only *RSPO2*, and not *RSPO3*, seems to be implicated in AER function [42–44]. Similarly, species-specific modifications in the gene regulatory networks driving skeletal cell type maturation have been reported [45, 46]. Together with recent scRNA-seq studies in other vertebrate model organisms [30, 47, 48], our dataset now opens new avenues for a comprehensive assessment of molecular similarities and divergences in patterning-relevant cell populations of the developing limb, across all major tetrapod clades.

Digit growth and patterning at cellular resolution

Variations in digit number, size and individual digit patterns in the autopod skeletal structure reflect functional

specialization of tetrapod hands and feet. During development, condensations of mesenchymal cells first give rise to early skeletogenic progenitors, to then differentiate into distinct skeletal lineages such as chondrocytes, osteocytes or synovial joint cells [49–51]. However, unlike for skeletal elements at more proximal locations of the limb, individual phalanx condensations are sequentially added and expanded at the distal tip of each forming digit, through proliferation of an evolutionary conserved progenitor population [22, 23, 52]. Hence, identifying regulators of growth rates, as well as for the relative temporal sequence at which the different skeletal cell types are specified, becomes paramount when trying to understand digit-specific phalanx patterns [25, 53].

Early autopod outgrowth, and later digit elongation, is controlled through complex signaling interactions at the distal margin of the limb, involving the concerted action of FGFs, BMPs and WNTs (reviewed [5]). Coinciding with this distal domain, we identify a distinct sub-population of mesenchymal cell types in all of our samples, marked by elevated activity of module Magenta with *TFAP2B*, *WNT5A* and high BMP signaling (Fig. 3c-f). Certain module members have been functionally implied in regulating autopod growth and digit elongation [24, 54–56], yet others remain completely unexplored in this context.

Moreover, we identify distinct sub-populations of interdigit mesenchyme cells in our HH29 and HH31 samples, with four associated gene co-expression modules (Fig. 4a-h). Module Olivegreen contains SNAI and ID genes, known to be expressed in interdigits, and likely relates to the various BMP-driven processes in this tissue [57–62]. On the other hand, module Orange is dominated by *RDH10*, implicated in mouse interdigital apoptosis [63]. Before its apoptotic disappearance at later stages of development, interdigit mesenchyme is known to instruct the specific phalanx-formulas of its anteriorly adjacent digit [24, 25]. Moreover, we manage to spatially attribute a distinct co-expression module (Midnightblue) to interdigit 4, i.e. posterior to a digit with known regulatory individualization in tetrapods [64].

Finally, across all developmental time points we sampled, we identify skeletogenic cell populations. At those stages, the forming skeletal elements still consist exclusively of early progenitors, maturing chondrocytes, and developing synovial joints. Accordingly, we only find three distinct sub-populations, associated with two co-expression modules. Module Red shows enrichment for many canonical markers of chondrocyte maturation (Fig. 6e) [45, 51]. On the other hand, genes in module Turquoise do not, for the most part, evoke a classical chondrogenic transcriptional profile (Fig. 6d). Again, this module might rather reflect an early transcriptional priming, only this time towards the skeletogenic lineage. In agreement with this, we only detect low expression levels for the canonical early skeletogenic marker *SOX9* in HH25 sub-cluster 4 (Fig. 2d), which itself is specifically enriched for Turquoise activity. Likewise, our synovial joint-like HH29 sub-cluster 17 shows high activity for Turquoise, while excluding the more mature chondrocyte module Red (Fig. 6g-i).

Conclusion

Our single-cell transcriptomic atlas provides a comprehensive genomics resource to study chicken limb development in unprecedented detail. Thereby, it complements a classical experimental model of vertebrate pattern formation with molecular data at cellular resolution. We curate molecular catalogues to provide an in-depth description of the embryonic autopod, through the assembly of cell population-specific lists of candidate marker genes. Combined with the power of viral overexpression screens and recent CRISPR/*Cas9* genome modifications technologies, this resource will provide a roadmap for the functional elucidation of cell type specification programs in patterning-relevant populations. Moreover, by constructing cell population-specific gene co-expression modules, we provide a tool to follow tissue dynamics across developmental and evolutionary time scales. Thereby, it will enable insights into the molecular underpinnings of homologous cell types across all major tetrapod clades, and their ensuing developmental impact on pattern formation and diversification in the vertebrate autopod.

Methods

Tissue sampling

We collected tissue samples from embryonic hind limbs at different developmental stages (Fig. 1,a-c). Limbs were dissected in cold PBS, and chopped coarsely with a razorblade. Dissociation into single cells was done using 0.25% trypsin in DMEM and incubation for 15 min at 37°. Occasional mechanical shearing by careful pipetting was applied during the incubation time.

scRNA-seq library preparation

Single-cell suspensions of samples HH25 and HH31 were fed into a *10X Genomics Chromium* Single Cell System (*10X Genomics*, Pleasanton, CA, USA) aiming for a concentration of 4000 cells per microliter. Cell capture, cDNA generation, preamplification and library preparation were done using *Chromium Single Cell 3′ v2* Reagent Kit according to the manufacturer instructions. For stage HH29 the cells were processed with the *DropSeq* method according to the original protocol [26]. Once the cDNA was obtained from all the samples, the sequencing proceeded on Illumina *NextSeq 500* or *HiSeq 2000* platforms as recommended by the developers to an average depth of 400 million reads per sample.

Data processing

Using either the *Cell Ranger* software v2 (*10X Genomics*) or the *DropSeq* pipeline v1 (https://github.com/broadinstitute/Drop-seq/releases) we performed base calling, adaptor trimming, mapping to the chicken ENSEMBL genome assembly and annotation Gallus_gallus-5.0 [65], de-multiplexing of the sequences and generation of the gene / cell count matrices.

Filtering thresholds for mapped data were adapted for each sample, depending on the different library complexities. Cells with an UMI count of more than 4 times the sample mean or less than 20% of the sample median were filtered out, cells with a mitochondrial or ribosomal contribution to UMI count of more than 10% were also filtered out. Using the R package Seurat v2.3.2 [66] the UMI counts were then Log-normalized and any variation due to the library size or mitochondrial UMI counts percentage was then regressed via a variance correction using the function ScaleData.

The cell cycle stage of each cell was inferred using the R package SCRAN [67] and gene pairs that covariate with cell cycle stages in mouse [68]. The gene pairs were translated to orthologous chicken genes [69] and a cell cycle stage score was obtained cell-wise for stages S, G1 and G2/M, the difference between the G2/M and S scores (δG2M/S) was calculated to be accounted for in later steps.

Dimensionality reduction and visualization

Significant principal components were determined for each sample as those falling outside of a Marchenko-Pastur distribution [35]. A dimensionality reduction step was carried out, using the t-SNE algorithm [28] to visualize the data and clustering of the cells based on transcriptomic similarities. The cells were clustered using the Louvain method for community detection from large networks and the Jaccard similarity coefficient to compare similarity and diversity of the sets, implemented in the FindClusters function in Seurat using data which was additionally variance-corrected for δG2M/S. A first, broad clustering step was done using a resolution of 0.4 for samples HH31 and HH29 and 0.5 for HH25; a second clustering was done to find sub-clusters within the data, this time using resolutions of 1.4 and 1.1 for the corresponding samples. All clustering steps were done using a k number of 20 and the significant principal components of the sample.

Differential expression analysis

Differential expression analyses based on the negative binomial distribution were performed with Seurat, using the δG2M/S as a covariate and only genes expressed in at least 15% of any compared population (Additional files 2, Additional files 3, Additional files 4); genes expressed in at least 25% of the cells and showing differences with a log fold-change > 0.5 and an adjusted p value < 0.05 were used for GO analyses. To find expression signatures for every cell cluster, in a first step, a phylogenetic tree was obtained for the cell clusters in each sample; all directly paired clusters were tested for differential expression. Any pair of clusters with less than 15 differentially expressed genes were collapsed recursively. In a second step, specific genes for each cluster were obtained contrasting each cluster against the rest of the cells in their sample. To find genes differentially expressed genes between the interdigit clusters (Fig. 4j), we compared each of the sub-clusters against the rest of the cells in the other two clusters.

Marker genes for digit/interdigit 3 and 4 were defined using the DESeq2 R package v1.20.0 [70]. We analyzed bulk RNA data sets of digit/interdigt 3 and 4 from stage HH28/29 and HH31 of a previous study [34]. After normalization based on size factors and dispersion, we performed the differential expression analysis using a Wald test and the contrast design ~Stage+Digit to use the different stages as pseudo-replicates of the digit. We filtered for differential expression with an adjusted p-value < 0.05. For visualization, we subtracted the fold changes of early and late stages and plotted a heatmap using heatmap3 R package v1.1.1 [71] using hierarchical clustering of the genes.

Weighted co-expression analyses

A weighted correlation network analysis was done using the WGCNA R package v1.6.6 [32]. Using the function FindVariableGenes from Seurat, we calculated the genes with high variation (dispersion > 0.5) across all the cells in sample HH29, and were subsequently used in WGCNA. Adjacencies and signed topological overlaps were calculated with an inferred soft-thresholding power of 8. A hierarchical tree was constructed using the "average" method and then cut using the "tree" method at height 0.9957 and minimum module size of 15. The eigengenes of the resulting modules, as well as the membership and a Correlation Student p value of the membership of each gene to its module were calculated. All genes not significantly (p value > 0.01) correlated with any module were discarded. The process was repeated recursively, until all genes were significantly associated with a module; the only change made in every iteration was the module minimum size, set to the smallest that would yield at least the same number of modules as the first analysis.

The output of WGCNA was exported to the Cytoscape v3.7.0 software [72] where the node size was coded to represent the membership, and the edge thickness and color intensity to represent the weights of each gene-pair coexpression. For visualization purposes, the scales of thickness, color and size were made relative to the minima and maxima found in each network. Furthermore, a transparency

gradient was added to the edges, which was scaled to hide unimportant edges and avoid edge saturation, the threshold was always adjusted to make visible at least one edge per node. In only one case (module midnightblue), an edge with an outlier weight was coded to be red and thicker than any other edge, and the color/size re-scaled to the second highest weight.

Gene ontology

Gene Ontology analyses were conducted with the R package *limma* [73]. We used the list of genes in the expression signature of each computed cell cluster, and the genes members of each co-expression module as input. For each case we used all the genes detected in the corresponding sample as the contrast universe.

In situ hybridization

Probes for *CRABP-I* and *COL1A2* were described previously [38]. Primers for the *ZFHX3* probe were designed using primer3 [74]. An AA overhang and an *EcoRI* restriction site were added to each of the primers at the 5′ end. ZFHX3 (fw: [5′-AAGAATTCAGCCGTACCGGGTGCAATGAGC-3′], rev: [5′-AAGAATTCAGCGCTTCCTCTTCCCGTAGAGC-3′]). In situ hybridization was performed using standard protocols [75].

Additional files

Additional file 1: Figure S1. Sample compositions and data statistics. (a) Cellular composition of the samples and datasets, color code corresponds to Fig. 2a-c. (b) UMI count distributions across the samples. (c) Gene count distributions across the samples. **Figure S2.** Expression patterns of marker genes. Related to Fig. 1. Normalized expression patterns of selected genes to identify the different cell populations in our broad clustering, plotted on the tSNEs from sample (a) HH25, (b) HH29 and (c) HH31. **Figure S3.** Co-expression modules expression patterns. Related to Fig. 3. Average expression of each WGCNA co-expression module on the tSNE of sample HH29. (PDF 4613 kb)

Additional file 2: Genes with enriched expression per cell population in sample HH25. Genes enriched in the different cell clusters, calculated to be differentially expressed between each cell cluster and the rest of the cells in the sample. p_val: originally calculated *p* value; avg_logFC: average log fold-change relative to the rest of the cells; pct.x: percentage of cells in the focus cluster expressing the gene; pct.rest: percentage of cells in the rest of the clusters expressing the gene; p_val_ad: p value adjusted for multiple testing; cluster: cluster number in the main text and figures; gene: ENSEMBL gene identifier; name: gene symbol, or name when available; enrichment: ratio of pct.x: pct.rest. (XLSX 153 kb)

Additional file 3: Genes with enriched expression per cell population in sample HH29. Genes enriched in the different cell clusters, calculated to be differentially expressed between each cell cluster and the rest of the cells in the sample. p_val: originally calculated p value; avg_logFC: average log fold-change relative to the rest of the cells; pct.x: percentage of cells in the focus cluster expressing the gene; pct.rest: percentage of cells in the rest of the clusters expressing the gene; p_val_adj: p value adjusted for multiple testing; cluster: cluster number in the main text and figures; gene: ENSEMBL gene identifier; name: gene symbol, or name when available; enrichment: ratio of pct.x: pct.rest. (XLSX 551 kb)

Additional file 4: Genes with enriched expression per cell population in sample HH31. Genes enriched in the different cell clusters, calculated to be differentially expressed between each cell cluster and the rest of the cells in the sample. p_val: originally calculated p value; avg_logFC: average log fold-change relative to the rest of the cells; pct.x: percentage of cells in the focus cluster expressing the gene; pct.rest: percentage of cells in the rest of the clusters expressing the gene; p_val_adj: p value adjusted for multiple testing; cluster: cluster number in the main text and figures; gene: ENSEMBL gene identifier; name: gene symbol, or name when available; enrichment: ratio of pct.x: pct.rest. (XLSX 395 kb)

Additional file 5: Co-expression modules and their genes. Genes part of the different co-expression modules. nodeName: ENSMBL identifier of the genes part of the module; altName: gene symbol, or name when available; membership: membership to the module. (XLSX 51 kb)

Abbreviations

AER: Apical ectodermal ridge; EvoDevo: Evolutionary developmental biology; GO: Gene ontology; HH: Hamburger-Hamilton stages; LPM: Lateral plate mesoderm; nsCT: Non-skeletal connective tissue; scRNA-seq: Single-cell RNA sequencing; TFs: Transcription factors; tSNE: t-distributed stochastic neighbor embedding; UMIs: Unique molecular identifiers

Acknowledgements

Calculations were performed at sciCORE (http://scicore.unibas.ch/) scientific computing center at University of Basel. CF and PT wish to acknowledge Katja Eschbach and Christian Beisel for help with *10X Genomics Chromium* and sequencing. OP and PT thank Tyler Burks for help with *DropSeq* experiments. PT and OP would like to acknowledge the generous support of Cliff Tabin and Aviv Regev, in whose labs this project was initiated (with help of NIH grant HD03443 to Cliff Tabin).

Funding

Work in the Tschopp laboratory is supported by the Swiss National Science Foundation (SNSF project grant 31003A_170022), the University of Basel and the *Forschungsfonds* of the University of Basel. These funding bodies had no role in the design of the study, collection, analysis, and interpretation of data, and in writing the manuscript.

Availability of data and materials

All data generated or analyzed during this study are included in this published article and its supplementary information files. Raw sequencing data and UMI count tables have been deposited at GEO (https://www.ncbi.nlm.nih.gov/geo/) under accession number GSE130439.

Authors' contributions

PT conceived and designed the study. CF, OP and PT conducted the scRNA-seq experiments. CF conducted data analyses and in situ experiments. CF and FS conducted the bulk RNA-seq re-analysis. CF and PT drafted the manuscript. All of the authors read and approved the final manuscript.

Ethics approval and consent to participate

In accordance with Swiss national guidelines (Swiss Animal Protection Ordinance; TSchV, chapter 6, Art. 112), no formal ethics approval was required, as all experiments were carried out prior to the third trimester of incubation.

Consent for publication

Not applicable.

Competing interests

The authors declare they have no competing interests.

Publisher's Note

Author details

[1]DUW Zoology, University of Basel, Vesalgasse 1, CH-4051 Basel, Switzerland. [2]Klarman Cell Observatory, Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA. [3]Present address: The Concern Foundation Laboratories at the Lautenberg Centre for Immunology and Cancer Research, IMRIC, Hebrew University Faculty of Medicine, 91120 Jerusalem, Israel.

Received: 7 March 2019 Accepted: 14 May 2019
Published online: 22 May 2019

References

1. Stathopoulos A, Levine M. Genomic regulatory networks and animal development. Dev Cell. 2005.
2. Arendt D, Musser JM, Baker CVH, Bergman A, Cepko C, Erwin DH, Pavlicev M, Schlosser G, Widder S, Laubichler MD, Wagner GP. The origin and evolution of cell types. Nat Rev Genet. 2016;17:744–57.
3. Moris N, Pina C, Arias AM. Transition states and cell fate decisions in epigenetic landscapes. Nat Rev Genet. 2016.
4. Eldar A, Dorfman R, Weiss D, Ashe H, Shilo DZ, Barkal N. Robustness of the BMP morphogen gradient in Drosophila embryonic patterning. Nature. 2002.
5. Zeller R, Ríos JL, Zuniga A. Vertebrate limb bud development: moving towards integrative analysis of organogenesis. Nat Rev Genet. 2009;10:845–58.
6. Perrimon N, Pitsouli C, Shilo B-Z. Signaling mechanisms controlling cell fate and embryonic patterning. Cold Spring Harb Perspect Biol. 2012;4:a005975.
7. Biesecker LG. Polydactyly: how many disorders and how many genes? 2010 update. Dev Dyn. 2011.
8. Stricker S, Mundlos S. Mechanisms of digit formation: human malformation syndromes tell the story. Dev Dyn. 2011.
9. Zuniga A, Zeller R, Probst S. The molecular basis of human congenital limb malformations. Wiley Interdiscip Rev Dev Biol. 2012.
10. Petit F, Sears KE, Ahituv N. Limb development: a paradigm of gene regulation. Nat Rev Genet. 2017;18:245–58.
11. Fedak TJ, Hall BK. Perspectives on hyperphalangy: patterns and processes. J Anat. 2004.
12. Sears KE, Behringer RR, Rasweiler JJ, Niswander LA. Development of bat flight: morphologic and molecular evolution of bat wing digits. Proc Natl Acad Sci. 2006.
13. De Bakker MAG, Fowler DA, Den OK, Dondorp EM, Carmen Garrido Navas M, Horbanczuk JO, Sire JY, Szczerbińska D, Richardson MK. Digit loss in archosaur evolution and the interplay between selection and constraints. Nature. 2013.
14. Cooper KL, Sears KE, Uygur A, Maier J, Baczkowski KS, Brosnahan M, Antczak D, Skidmore JA, Tabin CJ. Patterning and post-patterning modes of evolutionary digit loss in mammals. Nature. 2014.
15. Wachtler F, Christ B, Jacob HJ. On the determination of mesodermal tissues in the avian embryonic wing bud. Anat Embryol (Berl). 1981.
16. Logan M, Martin JF, Nagy A, Lobe C, Olson EN, Tabin CJ. Expression of Cre recombinase in the developing mouse limb bud driven by a Prxl enhancer. Genesis. 2002.
17. Pearse RV, Scherz PJ, Campbell JK, Tabin CJ. A cellular lineage analysis of the chick limb bud. Dev Biol. 2007.
18. Chevallier A, Kieny M, Mauger A. Limb-somite relationship: origin of the limb musculature. J Embryol Exp Morphol. 1977.
19. Christ B, Jacob HJ, Jacob M. Experimental analysis of the origin of the wing musculature in avian embryos. Anat Embryol (Berl). 1977.
20. Riddle RD, Johnson RL, Laufer E, Tabin C. Sonic hedgehog mediates the polarizing activity of the ZPA. Cell. 1993.
21. Lopez-Rios J, Duchesne A, Speziale D, Andrey G, Peterson KA, Germann P, et al. Attenuated sensing of SHH by Ptch1 underlies evolution of bovine limbs. Nature. 2014.
22. Montero JA, Lorda-Diez CI, Gañan Y, Macias D, Hurle JM. Activin/TGFβ and BMP crosstalk determines digit chondrogenesis. Dev Biol. 2008.
23. Suzuki T, Hasso SM, Fallon JF. Unique SMAD1/5/8 activity at the phalanx-forming region determines digit identity. Proc Natl Acad Sci U S A. 2008; 105:4185–90.
24. Dahn RD, Fallon JF. Interdigital regulation of digit identity and homeotic transformation by modulated BMP signaling. Science (80-); 2000.
25. Huang BL, Trofka A, Furusawa A, Norrie JL, Rabinowitz AH, Vokes SA, Taketo MM, Zakany J, Mackem S. An interdigit signalling Centre instructs coordinate phalanx-joint formation governed by 5′Hoxd-Gli3 antagonism. Nat Commun. 2016;7:1–10.
26. Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, Tirosh I, Bialas AR, Kamitaki N, Martersteck EM, Trombetta JJ, Weitz DA, Sanes JR, Shalek AK, Regev A, McCarroll SA. Highly parallel genome-wide expression profiling of individual cells using Nanoliter droplets. Cell. 2015;161:1202–14.
27. Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, et al. Massively parallel digital transcriptional profiling of single cells. Nat Commun. 2017;8:1–12.
28. Van Der Maaten LJP, Hinton GE. Visualizing high-dimensional data using t-sne. J Mach Learn Res. 2008;9:2579–605.
29. Cserjesi P, Lilly B, Bryson L, Wang Y, Sassoon DA, Olson EN. MHox: a mesodermally restricted homeodomain protein that binds an essential site in the muscle creatine kinase enhancer. Development. 1992.
30. Gerber T, Murawala P, Knapp D, Masselink W, Schuez M, Hermann S, Gac-Santel M, Nowoshilow S, Kageyama J, Khattak S, Currie JD, Camp JG, Tanaka EM, Treutlein B. Single-cell analysis uncovers convergence of cell identities during axolotl limb regeneration. Science. 2018;(80):362.
31. Bravo R, Frank R, Blundell PA, Macdonald-Bravo H. Cyclin/PCNA is the auxiliary protein of DNA polymerase-δ. Nature. 1987.
32. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 2008;9.
33. Dupé V, Ghyselinck NB, Thomazy V, Nagy L, Davies PJA, Chambon P, Mark M. Essential roles of retinoic acid signaling in interdigital apoptosis and control of BMP-7 expression in mouse autopods. Dev Biol. 1999.
34. Wang Z, Young RL, Xue H, Wagner GP. Transcriptomic analysis of avian digits reveals conserved and derived digit identities in birds. Nature. 2011; 477:583–7.
35. Shekhar K, Lapan SW, Whitney IE, Tran NM, Macosko EZ, Kowalczyk M, Adiconis X, Levin JZ, Nemesh J, Goldman M, McCarroll SA, Cepko CL, Regev A, Sanes JR. Comprehensive classification of retinal bipolar neurons by single-cell transcriptomics. Cell. 2016;166:1308–1323.e30.
36. Nicolas D, Phillips NE, Naef F. What shapes eukaryotic transcriptional bursting? Mol BioSyst. 2017.
37. Urban EA, Johnston RJ. Buffering and Amplifying Transcriptional Noise During Cell Fate Specification. Front Genet. 2018;9 November:1–14.
38. Bandyopadhyay A, Kubilus JK, Crochiere ML, Linsenmayer TF, Tabin CJ. Identification of unique molecular subdomains in the perichondrium and periosteum and their role in regulating gene expression in the underlying chondrocytes. Dev Biol. 2008.
39. Witte F, Dokas J, Neuendorf F, Mundlos S, Stricker S. Comprehensive expression analysis of all Wnt genes and their major secreted antagonists during mouse limb development and cartilage differentiation. Gene Expr Patterns. 2009.
40. Marioni JC, Arendt D. How single-cell genomics is changing evolutionary and developmental biology. Annu Rev Cell Dev Biol. 2017;33:537–53.
41. Tschopp P, Tabin CJ. Deep homology in the age of next-generation sequencing. Philosophical Transactions of the Royal Society B: Biological Sciences. 2017.
42. Nam JS, Park E, Turcotte TJ, Palencia S, Zhan X, Lee J, Yun K, Funk WD, Yoon JK. Mouse R-spondin2 is required for apical ectodermal ridge maintenance in the hindlimb. Dev Biol. 2007.
43. Neufeld S, Rosin JM, Ambasta A, Hui K, Shaneman V, Crowder R, Vickerman L, Cobb J. A conditional allele of Rspo3 reveals redundant function of R-spondins during mouse limb development. Genesis. 2012.
44. Szenker-Ravi E, Altunoglu U, Leushacke M, Bosso-Lefèvre C, Khatoo M, Thi Tran H, et al. RSPO2 inhibition of RNF43 and ZNRF3 governs limb development independently of LGR4/5/6. Nature. 2018.
45. Gómez-Picos P, Eames BF. On the evolutionary relationship between chondrocytes and osteoblasts. Front Genet. 2015.
46. Ferguson GB, Van Handel B, Bay M, Fiziev P, Org T, Lee S, et al. Mapping molecular landmarks of human skeletal ontogeny and pluripotent stem cell-derived articular chondrocytes. Nat Commun. 2018.
47. Fabre PJ, Leleu M, Mascrez B, Lo GQ, Cobb J, Duboule D. Single-cell mRNA profiling reveals heterogeneous combinatorial expression of Hoxd genes during limb development. BMC Biol. 2018:327619.
48. Cao J, Spielmann M, Qiu X, Huang X, Ibrahim DM, Hill AJ, Zhang F, Mundlos S, Christiansen L, Steemers FJ, Trapnell C, Shendure J. The single-cell transcriptional landscape of mammalian organogenesis. Nature. 2019.
49. Hartmann C, Tabin CJ. Wnt-14 plays a pivotal role in inducing synovial joint formation in the developing appendicular skeleton. Cell. 2001.
50. Akiyama H, Kim J-E, Nakashima K, Balmes G, Iwai N, Deng JM, Zhang Z, Martin JF, Behringer RR, Nakamura T, de Crombrugghe B. Osteo-chondroprogenitor cells are derived from Sox9 expressing precursors. Proc Natl Acad Sci. 2005.
51. Kozhemyakina E, Lassar AB, Zelzer E. A pathway to bone: signaling molecules and transcription factors involved in chondrocyte development and maturation. Development. 2015.
52. Witte F, Chan D, Economides AN, Mundlos S, Stricker S. Receptor tyrosine kinase-like orphan receptor 2 (ROR2) and Indian hedgehog regulate digit outgrowth mediated by the phalanx-forming region. Proc Natl Acad Sci. 2010.

53. Hiscock TW, Tschopp P, Tabin CJ. On the formation of digits and joints during limb development. Dev Cell. 2017.
54. Sanz-Ezquerro JJ, Tickle C. Fgf signaling controls the number of phalanges and tip formation in developing digits. Curr Biol. 2003.
55. Gros J, Hu JKH, Vinegoni C, Feruglio PF, Weissleder R, Tabin CJ. WNT5A/JNK and FGF/MAPK pathways regulate the cellular events shaping the vertebrate limb bud. Curr Biol. 2010.
56. Seki R, Kitajima K, Matsubara H, Suzuki T, Saito D, Yokoyama H, Tamura K. AP-2β is a transcriptional regulator for determination of digit length in tetrapods. Dev Biol. 2015.
57. Jen Y, Manova K, Benezra R. Expression patterns of Id1, Id2, and Id3 are highly related but distinct from that of Id4 during mouse embryogenesis. Dev Dyn. 1996.
58. Ros MA, Sefton M, Nieto MA. Slug, a zinc finger gene previously implicated in the early patterning of the mesoderm and the neural crest, is also involved in chick limb development. Development. 1997.
59. Nieto MA. The snail superfamily of zinc-finger transcription factors. Nat Rev Mol Cell Biol. 2002.
60. Zuzarte-Luís V, Hurlé JM. Programmed cell death in the developing limb. Int J Dev Biol. 2002.
61. Lorda-Diez CI, Torre-Pérez N, García-Porrero JA, Hurle JM, Montero JA. Expression of Id2 in the developing limb is associated with zones of active BMP signaling and marks the regions of growth and differentiation of the developing digits. Int J Dev Biol. 2009.
62. Pignatti E, Zeller R, Zuniga A. To BMP or not to BMP during vertebrate limb bud development. Semin Cell Dev Biol. 2014.
63. Cunningham TJ, Chatzi C, Sandell LL, Trainor PA, Duester G. Rdh10 mutants deficient in limb field retinoic acid signaling exhibit normal limb patterning but display interdigital webbing. Dev Dyn. 2011.
64. Stewart TA, Liang C, Cotney J, Noonan JP, Sanger T, Wagner G. Evidence against tetrapod-wide digit identities and for a limited frame shift in bird wings. bioRxiv. 2018:224147.
65. Zerbino DR, Achuthan P, Akanni W, Amode MR, Barrell D, Bhai J, et al. Ensembl 2018. Nucleic Acids Res. 2017;2017:1–8.
66. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol. 2018.
67. Lun ATL, McCarthy DJ, Marioni JC. A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. F1000Research. 2016;5:2122.
68. Scialdone A, Natarajan KN, Saraiva LR, Proserpio V, Teichmann SA, Stegle O, Marioni JC, Buettner F. Computational assignment of cell-cycle stage from single-cell transcriptome data. Methods. 2015;85:54–61.
69. Durinck S, Spellman PT, Birney E, Huber W. Mapping identifiers for the integration of genomic datasets with the R/ Bioconductor package biomaRt. Nat Protoc. 2009;4:1184–91.
70. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014.
71. Zhao S, Guo Y, Sheng Q, Shyr Y. Heatmap3: an improved heatmap package with more powerful and convenient features. BMC Bioinformatics. 2014.
72. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003.
73. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. Limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43:e47.
74. Rozen S, Skaletsky H. In: Krawetz S, Misener S, editors. Methods and protocols: methods in molecular biology Primer3 on the WWW for general users and for biologist programmers. Totowa, NJ: Humana Press; 2000.
75. McGlinn E, Mansfield JH. Detection of gene expression in mouse embryos and tissue sections. In: Pelegri FJ, editor. Vertebrate embryogenesis: embryological, cellular, and genetic methods. Totowa, NJ: Humana Press; 2011. p. 259–92.

Chapter 3

A SINGLE-CELL PSEUDOTEMPORAL RECONSTRUCTION OF THE DIGIT PATTERNING PROCESS

Christian Feregrino

Part of:
Impact of BMP signaling on patterning cell fate decisions during digit development
by E. Grall *et al.*

Abstract

Tetrapod limbs show an outstanding variety of morphological arrangements and the part which shows the highest diversity are the autopods: hands and feet. The patterning process which results in the different numbers and designs of the limb digits has been extensively studied. We know that the phalanx-joint sequence that constitutes a digit develops and grows in a proximal to distal manner. The cells that comprise the digit originate from an undifferentiated and proliferative distal cell population called the phalanx-forming region. After leaving said progenitor population, cells diverge into phalanx or joint chondrocytes to construct the final digit pattern. Nonetheless, the mechanisms that controls this particular cell fate divergence remain unknown. Here, making use of single-cell RNA-seq data, and pseudotime analyses, we describe the gene expression dynamics that underlie the early cell fate divergence process that results in phalanx and joint cell fates. To do so, we use a subset of the cells from our previously published dataset of chicken embryo stage HH29 hind limb and Slingshot pseudotemporal analyses. Overall, we recapitulate the current models of digit and joint morphogenesis, but also infer unexpected expression dynamics. We found genes that exhibit an onset of expression previous to the upregulation of the well-established joint marker gene GDF5, as well as genes that haven't been studied yet in the context of the limb patterning process. In conclusion, we offer a resource that allows the molecular exploration of the transcriptomic dynamics along an *in silico* reconstructed early phalanx to joint interzone cell fate divergence process.

Introduction

Two pairs of limbs – fore- and hind limbs –, is the evolutionary novelty and distinctive feature that allowed tetrapods to adapt their locomotion to countless ecological niches. Swimming, walking, running, jumping or flying, are different locomotion purposes that limbs fulfil. On the other hand, limbs have also evolved even more complex functions like digging, grabbing, manipulating, running on water like some lizards do or playing a violin. Specialization of the functions a limb can perform is partially achieved through the particular morphology of the underlying limb skeleton. The skeletal structures of the limb can be anatomically divided in three parts, from proximal to distal: the stylopod, consisting of a single long and thick bone – humerus or femur –; the zeugopod, generally composed of two bones – ulna and radius or tibia and fibula –; and the autopod, composed of several thin bones, described hereafter. In general, the stylopod and zeugopod are conserved in their composition across species, but the autopod shows an outstanding diversity of shapes and combination of skeletal elements (Fedak & Hall, 2004; Sears *et al*, 2006; De Bakker *et al*, 2013; Cooper *et al*, 2014).

The skeletal pattern of autopods can be divided, as well, in three parts. From proximal to distal we find the carpals or tarsals, the metacarpals or metatarsals, and finally the phalanges. Digits are composed of a sequence of phalanges (Wagner & Chiu, 2001) and between each of the phalanges we find synovial joints, which permit flexion and extension movements. As mentioned, the composition of the autopod shows great diversity, and this is primarily found regarding the number of digits, the number of phalanges within each digit, and different lengths of each individual phalanx (Fedak & Hall, 2004; Sears *et al*, 2006; De Bakker *et al*, 2013; Cooper *et al*, 2014). It's mainly the myriad of possible combinations found in autopod design that have allowed limbs to adapt to the functions they display nowadays. The number and design of the digits is determined during the development of the limb, which has been a paradigm to study developmental patterning (Petit *et al*, 2017). Many signals, molecules and processes that determine the growth, number, position and identity of the digits have been discovered over the last few decades, advances which constitute great progress in the understanding of patterning in general (Biesecker, 2011; Stricker & Mundlos, 2011; Zuniga *et al*, 2012).

The limbs start their development with the outgrowth of the lateral plate mesoderm (LPM) from the main axis of the body. The mesenchyme of the limb is therefore characterized by the expression of PRRX1 (Logan *et al*, 2002; Cserjesi *et al*, 1992). The initial epithelial to mesenchymal transition of the LPM and subsequent limb growth is promoted by the presence of an ectodermal signaling center called the apical ectodermal ridge (AER) (Saunders, 1948). Limb growth occurs in general from within the distalmost part, where signals of the FGF gene family, particularly FGF8, are secreted by the AER and promote cell proliferation (Niswander *et al*, 1993; Mariani *et al*, 2008). In the autopod, as the mesenchymal cells leave the zone of influence of the AER, they start to differentiate into other types of cells (Tabin & Wolpert, 2007),

of which the most important for the skeletal morphology of digits are the cartilage-forming chondrocytes (later replaced by osteocytes which form bones) and the joint precursors.

While the limb grows in a proximal-distal manner, its patterning process also occurs in a posterior-anterior fashion. The main posterior-anterior patterning signal comes from mesenchymal cells in the posterior edge of the limb bud, known as the zone of polarizing activity (ZPA) which secrete SHH (Chiang *et al*, 2001; Riddle *et al*, 1993). A temporal posterior-anterior gradient of SHH establishes an early pattern of digit number and identity along this axis (Zhu *et al*, 2008), but the final digit design, in terms of number and size of the phalanges, seems to be determined later on, as we will discuss. The pattern and identity of a digit ultimately depends on the amount and periodicity of joints found in the digital rays, and not directly on its position (Wang *et al*, 2011).

Digital rays grow in a distal-driven manner, stemming from a progenitor population known as the phalanx forming region (PFR). The PFR is a domain of high p-SMAD activity, which is maintained by AER signaling (Suzuki *et al*, 2008; Montero *et al*, 2008) (Figure 1 A). There is evidence that the duration of the FGF signaling from the AER correlates with the number of phalanges produced in a given digit (Sanz-Ezquerro & Tickle, 2003). There is also evidence that the size of the already forming phalanges influences the size of the following phalanges (Kavanagh *et al*, 2013). Moreover, it's known that BMP signaling coming from the posterior interdigits instructs the PFR to induce changes in the amount and size of the phalanges, and hence periodicity of joints (Suzuki *et al*, 2008; Dahn & Fallon, 2000). But the molecular mechanisms through which posterior-anterior patterning signaling instructs a proximal-distal differentiation still remain largely unknown.

Phalanx formation starts by a condensation of mesenchymal cells and the appearance of the PFR. The PFR then maintains and extends the growing digit, by feeding new SOX9+ chondrogenic cells in it (Suzuki *et al*, 2008; Montero *et al*, 2008). SOX9+ cells later differentiate into chondrocytes expressing SOX5, SOX6, several genes of the BMP family – i.e. CHRDL1 (Nakayama *et al*, 2001; Allen *et al*, 2013) – and BMP receptors (Yoon *et al*, 2005; Zou *et al*, 1997), as well as FGF receptor proteins (Noji *et al*, 1993), among others. Proliferating chondrocytes also start to secrete several extra-cellular matrix proteins like ACAN, MATN1, COL2, COL9 and COL11, to build a cartilage bone-anlagen (Hall & Miyake, 1995; Kozhemyakina *et al*, 2015).

Many of the details that we know about joint morphogenesis, come from research done on the development of the stylopod – zeugopod joint; namely elbows and knees. Joint formation is also carried out by chondrogenic cells that are first expressing SOX9. The first histological sign of a differentiation from the rest of the chondrocytes is the appearance of an interzone. The interzone is a compact and non-vascularized mesenchymal tissue layer, which interrupts the otherwise continuous cartilaginous elements (Holder, 1977) (Figure 1 A). The cells in the interzone are rather flat and

aligned perpendicular to the main growth axis. These cells start expressing GDF5 (Koyama *et al*, 2008; Storm & Kingsley, 1996), WNT4 (Guo *et al*, 2004) and WNT9A (Hartmann & Tabin, 2001). They also down-regulate expression of SOX9 and other chondrocyte-specific genes (Schmid *et al*, 2015). The interzone later gives rise to the joint, where other cells expressing GDF5 and TGFBR2 are recruited from the surrounding tissues (Shwartz *et al*, 2016).

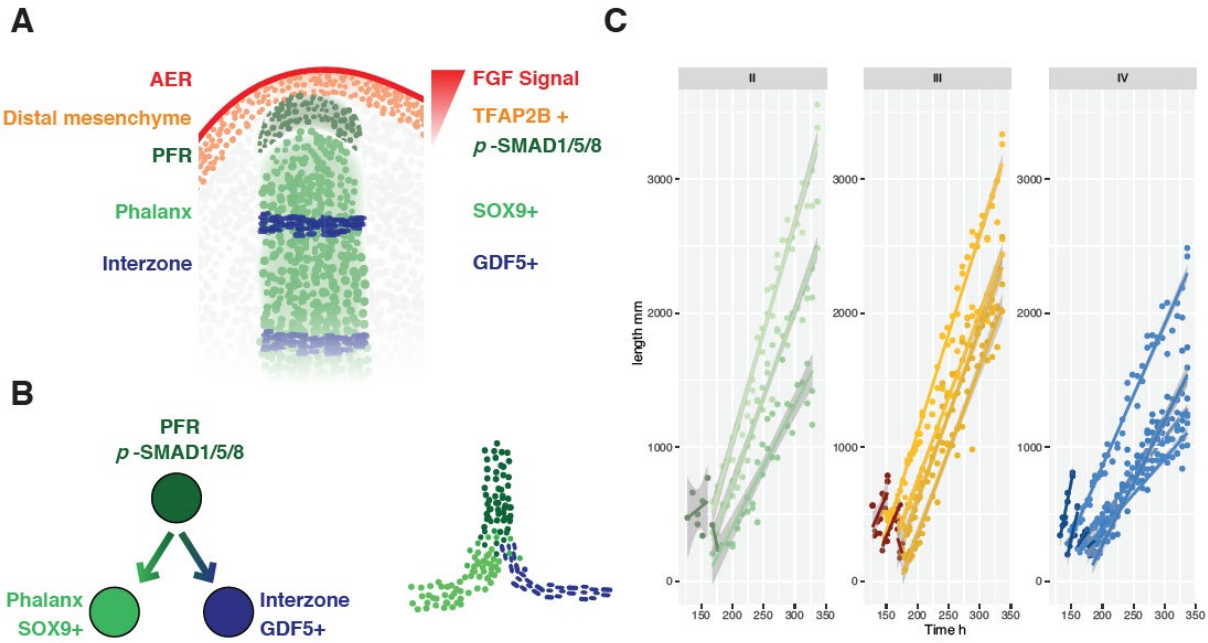


Figure 1 Digit development, patterning and growth. **A** Schematic representation of the developing digit showing the main structures involved in digit patterning, as well as their marker genes. **B** Schematic representation of the cell fate divergence that gives rise to the interzone. Left: divergence at the cellular level. Right: snapshot of the divergence at the cell population level. Colors correspond to A. **C** courtesy of E. Grall. Growth dynamics of digits II, III and IV of the chicken hind limb. Lines with low color intensity show the growth of each consecutive phalanx in the digit. Lines with high color intensity show the distance from the last interzone to the PFR. Colors independent from other panels.

Importantly, there might be differences in the general expression patterns observed between the stylopod – zeugopod joint (SZJ) and the joint development within the digits, since they also show slight developmental differences (Archer *et al*, 2003). During SZJ development, the cartilage anlagen is already present as a well-defined continuous Y-shaped structure which is then segmented by the interzone and later by the joint (Khan *et al*, 2007; Archer *et al*, 2003). In the case of the digits, the phalanx anlagen, and the chondrocytes are not as mature when an interzone first appears. Observations from our research group show that interzones in the digits, and GDF5 expression, arise very shortly after each other, and not from a continuous long bone like in the case of the SZJ (Figure 1 C) (Grall *et al.* unpublished).

Despite these insights we have about the development of phalanges and joints, the mechanisms that specify both the appearance of the interzone and the cell fate divergence remain unclear. There is evidence that active BMP signaling suppresses the formation of joints (Duprez *et al*, 1996; Tsumaki *et al*, 2002; Zhang *et al*, 2000).

On the other hand, canonical WNT signaling has the opposite effect, is essential for the maintenance of the interzone, and can even induce the formation of ectopic interzones (Guo *et al*, 2004; Hartmann & Tabin, 2001). Likewise, it has been proven that the BMP-antagonist NOG is essential for interzone formation (Brunet *et al*, 1998). Moreover, it has been found that c-Jun regulates expression of WNT genes in the early interzones (Kan & Tabin, 2013). Nevertheless, the processes upstream of this WNT regulation, GDF5 activation and interzone emergence remain poorly understood.

Given that the PFR maintains and feeds the growth of the SOX9+ digital ray, and is also the place where the periodicity of the joints is defined (Suzuki *et al*, 2008), there must be a moment during the maturation of the cells leaving the PFR, at which their fate changes and they become interzone cells (Figure 1 B). If we are able to reconstruct the molecular events that take place inside these cells, in their path from PFR cells, to proliferating phalanx chondrocyte or to joint chondrocyte, we might be able to shed light on the molecular mechanism that specifies the site of the prospective joints.

In this study, we make use of an already published single-cell RNA-seq dataset of the developing chicken autopod to explore the temporal dynamics of expression along the putative join-phalanx differentiation lineages. Given the nature of digit patterning, during certain stages of development, cells are simultaneously present at several differentiation stages. The most proximal cells have matured further, while the most distal cells, close to the PFR and AER, are just starting their differentiation process. At stage HH29, the digits of the chicken hind limb are in the patterning process, and start to show morphological differences. For this reason, we used a sample we have previously described at the transcriptome level (Feregrino *et al*, 2019), HH29 hind limb digits III and IV, as well as their adjacent interdigital space. Using the single-cell transcriptomes, and making use of pseudotemporal analyses, we propose a sequence of transcriptomic changes that take place during the cell fate divergence process into phalanx-forming chondrocytes and joint precursor cells.

Results

Hind limb scRNA-seq re-analysis

Given that high-throughput single-cell sequencing is still a relatively recent technique, advancements and new methods for analyzing this kind of data are developed constantly. As such, we reanalyzed our data set (Figure 2 A), using new and refined approaches. We use tSNE dimensionality reductions as visualizations, but calculated in a manner in which the global structure of the data is retained, so that distances in the plot are more meaningful (Kobak & Berens, 2019). Using refined unsupervised graph-clustering and community detection algorithms, we found 19 clusters of cells (Figure 2 B). Based on our previous analysis, and expression of marker genes (Figure 2 C), we identified the cell populations as different cell types and cell states. As previously, we found clusters of skin, muscle, blood vessels (endothelium) and blood (leukocytes). Furthermore, we identified two cell clusters in different cell cycle stages,

a cell population rather rich in mitochondrial transcripts, the putative mesenchymal progenitor pool expressing PPRX1, the distal mesenchyme, 3 distinct interdigital populations, an early and late non-skeletal connective tissue (nsCT) populations, a cluster of SFRP2+ cells, as well as 4 different populations expressing chondrocytes and cartilage marker genes.

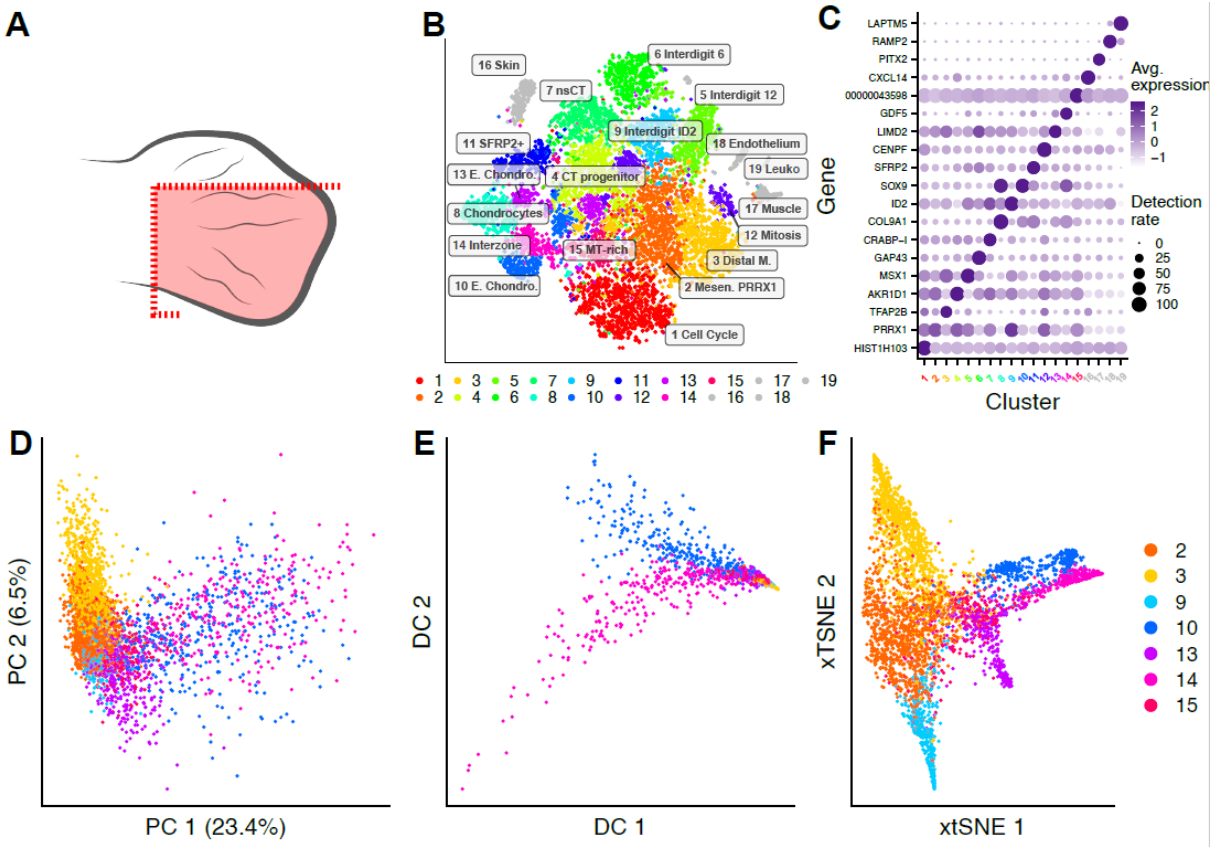


Figure 2 Re-analysis of HH29 hind limb single-cell data. **A** Dissection strategy from chicken hindlimb stage HH29. Adapted from: (Feregrino *et al*, 2019).**B** tSNE visualization with the 19 clusters we calculated for our data, annotated accordingly. **C** Dotplot showing the expression levels of different marker genes from each population. Genes with no name showing ENSEMBL code number ENSGALG-. Dot size represents the proportion of cells in that particular cluster expressing the gene, color intensity the scaled average expression in that cluster. **D** PCA of a subset of the cell populations. Clusters and coloration correspond to A and F. **E** Diffusion map of the same data as C. **F** Exaggerated tSNE of the same data as C.

We analyzed the genes that are differentially expressed in the chondrocyte-like clusters to determine which kind of chondrocytes they are. We found a population of more mature chondrocytes, showing high expression of COL9A1 and MATN1 (cluster 8). One cluster of joint-forming chondrocytes showing high GDF5 expression (cluster 14). And lastly, two clusters of early chondrocytes, one expressing high levels of SOX9 and CHRDL1 (cluster 10), and one with high levels of LIMD2 and CYTL1 expression (cluster 13).

Given the scope of this study, to understand transcriptional changes along the phalanx – joint lineage divergence, we decided to focus on a subset of clusters for our subsequent pseudotemporal analysis. We began by choosing the cluster 14 “interzone chondrocytes”, and cluster 10 “early chondrocytes” SOX9+, from which the

interzone cells theoretically diverge. While the developmental origin of all these cells is the PFR, we don't have gene expression markers to identify this population. We reasoned that if there is a smooth transition between the transcriptional state of the PFR, to that of early chondrocytes, the PFR cells would be in close proximity in the transcriptomic space. For this reason, we decided to use cluster 2 "mesenchymal progenitors" PRRX1+. Although this population probably contains progenitors and cells in the process to mature into all the other lineages (i.e. interdigits, nsCT), the chondrocyte lineage reconstruction should still be possible, although distinct, additional end-points might arise in the pseudotime reconstruction.

To not lose any valuable information, that might represent intermediate or somehow implicated steps in the divergence process, we incorporated into the analysis other close-by populations in the transcriptomic space. We used the clusters 13 "early chondrocytes" LIMD2+ and cluster 15 "rich in mitochondrial transcripts". And to confirm the additional endpoints earlier mentioned, we used cluster 9 "interdigits" and cluster 3 "distal mesenchyme". Although close-by in the transcriptomic space, we didn't use cluster 1 "cell cycle", or cluster 4 "nsCT progenitors", since their expression patterns clearly follow other trajectories (data not shown). We used principal component analysis (PCA), diffusion map, and an exaggerated tSNE to observe the relationships of the chosen cells. In the first two dimensions of the PCA (Figure 2 D) and exaggerated tSNE (Figure 2 F), we observed that indeed cluster 2 "mesenchymal progenitors" PRRX1+ seems to show a smooth transition into the transcriptional state of clusters 10 "early chondrocytes" and 14 "interzone chondrocytes". Although not having many differentially expressed genes, cluster 15 "rich in mitochondrial reads" seems to blend in the smooth transcriptional transition between these clusters. In the diffusion map, we observe a bifurcation of chondrocytes and interzone cells, but the transition from the progenitor cluster is not smooth, nor very clear (Figure 2 E). Clusters 13 "early chondrocytes" LIMD2+ and, as expected, 3 "distal mesenchyme" and "9 interdigits" show transitions that do not follow our trajectory of interest in the PCA and tSNE (Figure 2 D and F), and don't show a distinct trajectory in the diffusion map (Figure 2 E).

Based on these observations, the final data set we used in the subsequent analyses consisted of 2019 cells and 14,170 expressed genes, coming from 4 different cell populations: PRRX1+ progenitors, SOX9+ early chondrocytes, GDF5+ interzone chondrocytes and cluster 15 "rich in mitochondrial transcripts". Within these cells, we identified 397 highly variable genes, and use them to calculate new dimensionality reductions: PCA, diffusion map and exaggerated tSNE. We plotted the expression of marker genes CHRDL1 for the phalanx and GDF5 for the interzone, to make sure that divergence information is retained in the reductions (Figure 3 A). We then used the first two dimensions of each of the reductions to perform independent pseudotime analyses. Using Slingshot, we calculated pseudotime cell orderings using the progenitors' population as the origin of the pseudotime, and chondrocytes and interzones as the end tips. In each of the three pseudotime calculations, we observed two trajectories, in agreement with marker gene expression dynamics (Figure 3 B).

Pseudotime

To know which genes display a change in transcription along each of the pseudotime orderings we obtained, we performed differential expression analyses. For this, we used a zero-inflated hurdle model regression, using a loess general additive model fit of the pseudotime as the variable. In this way we tested, trajectory-wise, which genes change as a function of the pseudotime. Importantly, we only tested those genes expressed in the given trajectory.

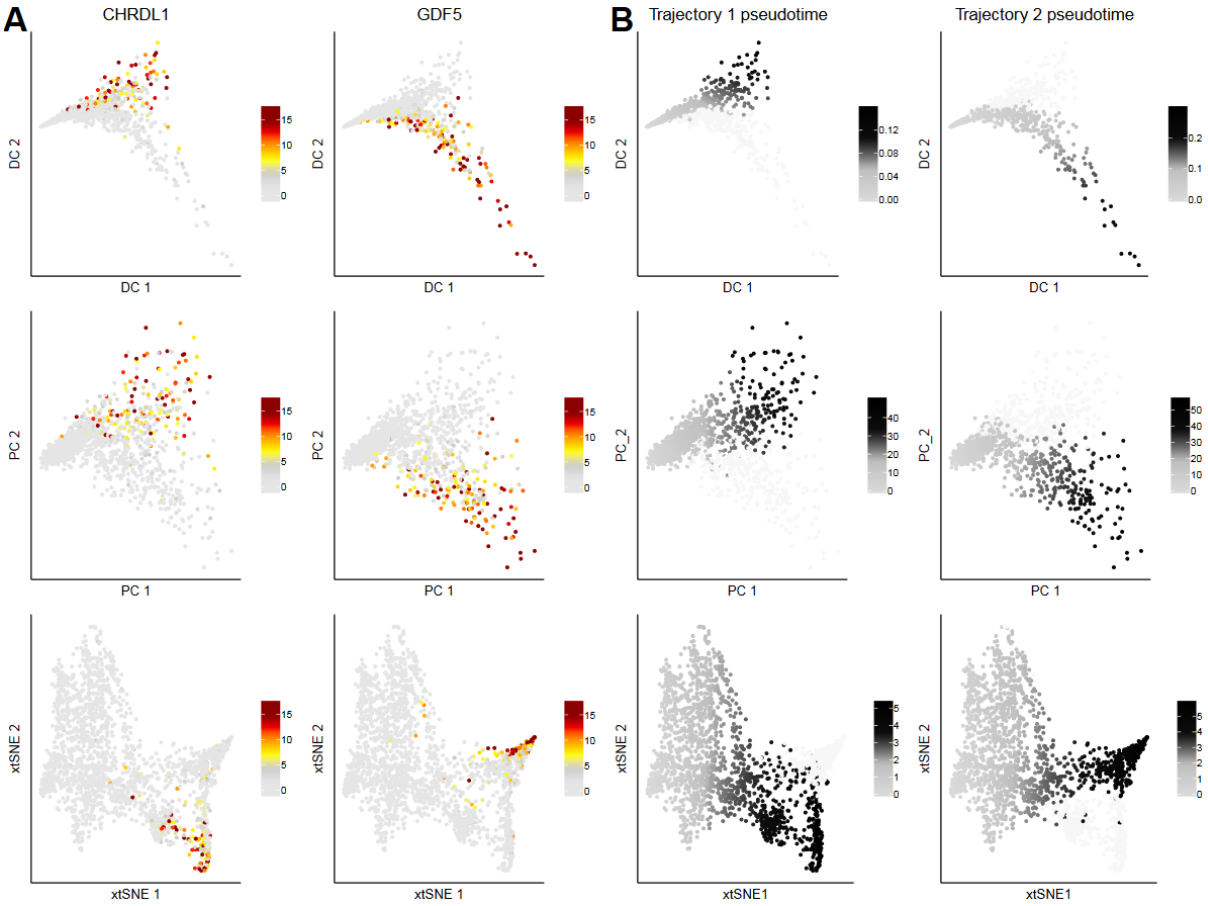


Figure 3 Pseudotime analysis of progenitors, phalanx-forming and interzone-forming chondrocytes. **A** Dimensionality reductions and gene expression of marker genes. From top to bottom: diffusion map, PCA and exaggerated tSNE. Left to right: Expression of phalanx lineage marker gene CHRDL1 and expression of interzone lineage marker gene GDF5. **B** Pseudotime of each trajectory as calculated by slingshot on the different dimensionality reductions. From top to bottom: same order as in A. From left to right: Trajectory 1 representing the phalanx lineage and trajectory 2 representing the interzone lineage.

We compared the pseudotime orderings we obtained using the PCA, diffusion map or exaggerated tSNE as basis. Since the dimensions across which the cells are distributed are different, the pseudotime trajectories are of different lengths, and cell density is different across them. Nonetheless, the identity of genes that change in expression should remain similar. For this reason, we did not compare expression dynamics nor cell identity across pseudotimes, and rather compared the identity of the genes that show differential expression along the trajectories. We took the top 100 differentially expressed genes across each trajectory in each pseudotime to compare them. We found that 82 genes are among the top 100 differentially

expressed in all pseudotimes along the phalanx trajectory, while 80 are common to all pseudotimes along the interzone trajectory. This suggested that our pseudotimes agree overall in terms of gene expression change. We chose to present the PCA-based pseudotime, as its topology shows a smoother transition from beginning to end of pseudotime. Moreover, this pseudotime shows the highest amount of shared top 100 differentially expressed genes in both trajectories. Pseudotime results from diffusion map and tSNE can be found in the Supplementary figures 3 and 4.

We first observed the expression dynamics of genes known to be involved in the specification process of the phalanx or the interzone. Most of these genes have been tested functionally and shown to be part of the process (see introduction), while CD24, ENPP2 and CHRDL2 have not been functionally tested, they show distinct expression in early and joint-forming chondrocytes (Feregrino *et al*, 2019). Of the 29 genes we considered, 8 are among the top 20 common (across all three pseudotimes) differentially expressed genes in both of the trajectories: COL9A1, SOX9, CD24, CHRDL1, PRRX1, GDF5, ENPP2 and CHRDL2. The dynamics of these genes suggested that our pseudotime reconstruction recapitulates what is known about the end points of this divergence process (Figure 4 A). By combining the 35 top DEGs from each trajectory and each pseudotime, we found 26 genes to be unique and not included in our original list of genes. We also inspected the expression dynamics of these genes. Here, we also observe genes with expression dynamics restricted to either of the trajectories (Figure 4 B). Hereafter we present our results in detail.

Departing from mesenchymal progenitors

Starting from pseudotime 0 (Figure 4 A and B).

PRRX1 is the canonical transcriptional marker for limb mesenchymal progenitors (Cserjesi *et al*, 1992; Logan *et al*, 2002), which then becomes downregulated as mesenchyme condensates to differentiate into chondrocytes and non-skeletal connective tissue (Nohno *et al*, 1993; Chesterman *et al*, 2001). Our pseudotime analysis recovers this dynamic (Figure 4 C and D). Another gene we found following the same expression pattern is BASP1, although its role in chondrogenesis is unknown. Meanwhile TUBA1A, HMGN4 and HMGB1 show a similar trend along the pseudotime, but delayed downregulation. Potential roles of TUBA1A and HMGN4 on chondrogenesis have not yet been reported. HMGB1 regulates endochondral ossification and is downregulated during early chondrogenesis (Taniguchi *et al*, 2007).

WNT5A is highly expressed from the beginning of our pseudotime reconstruction, and becomes downregulated towards the end of the interzone trajectory. WNT5A is well known to be involved in cartilage and bone patterning along the proximo-distal axis of the limb, by delaying chondrocyte maturation (Dealy *et al*, 1993; Kawakami *et al*, 1999; Church *et al*, 2002; Hartmann & Tabin, 2000; Yang *et al*, 2003; Bradley & Drissi, 2010). An opposite trend is shown by JUN and MDK, genes that become downregulated along the phalanx trajectory. The functions of these genes will be discussed later.

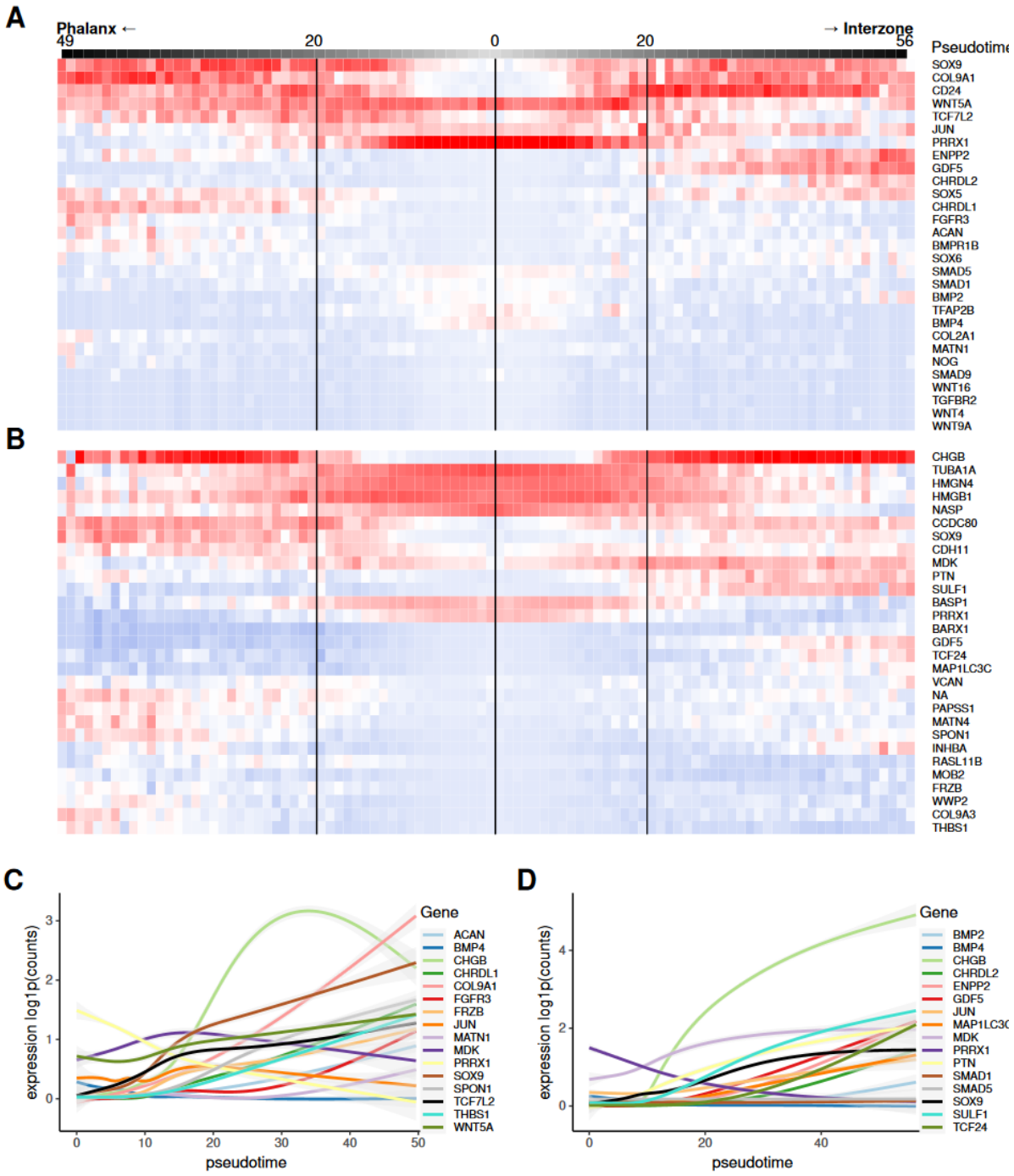


Figure 4 Expression dynamics of different genes across the phalanx-joint divergence pseudotime reconstruction as calculated on PCA. **A** Branching pseudotime heatmap of genes known to be implicated in this cell fate divergence process. Red indicates high scaled expression, blue indicates low scaled expression. The center vertical line signals the origin of the pseudotime for both branches. To the left, the phalanx trajectory, to the right, the interzone trajectory. Trajectories share cells at the beginning of the pseudotime, but after the lateral black lines, cells are not shared any longer. **B** Same representation as A. Showing expression of genes not known to be directly implicated in this process, but that were found to be differentially expressed as a function of pseudotime. PRRX1, SOX9 and GDF5 added as reference. **C** Expression dynamics of different genes along the phalanx trajectory. Area around the lines represents 0.95 confidence interval. **D** Expression dynamics of different genes along the interzone trajectory. Coloring is independent from C.

Although there are no known transcriptomic markers for the PFR, there are certain indications of its identity. The PFR is defined by its high activity of *p*-SMAD1, *p*-SMAD5 and *p*-SMAD8. For these Smad proteins to be phosphorylated as a result of BMP activity (Suzuki *et al*, 2008), they first need to be expressed. Moreover, TCF7L2, a chondrocyte maturation and proliferation regulator (Mikasa *et al*, 2011), seems to be expressed outside of the PFR in mouse limbs (Witte *et al*, 2010). However, in the chicken autopod, the expression of TCF7L2 overlaps the *p*-SMAD activity of the PFR (Grall *et al,* unpublished). In our pseudotime, we observe high expression of TCF7L2 starting before the downregulation of PRRX1, which coincides with the transient weak expression we recovered from SMAD1 and SMAD5. Therefore, we believe that our reconstruction in fact contains the PFR, from which the digit chondrocytes originate.

Early chondrogenic stage

After pseudotime 0 and before pseudotime 20 in both directions (Figure 4 A and B).

SOX9 is one of the earliest signs of chondrogenesis (Wright *et al*, 1995; Healy *et al*, 1996). It is also one of the master regulators of the chondrogenic transcriptional program. In our pseudotime reconstruction we detect a gradual upregulation of SOX9, which coincides with the gradual downregulation of PPRX1. This recapitulates the known process of early condensation and chondrogenesis. COL9A1, shows expression patterns that follow that of SOX9, with a pseudotemporal delay. COL9A1 is another well-established marker of chondrogenesis, and is directly enhanced by SOX9 (Zhang *et al*, 2003; Genzer & Bridgewater, 2007). Other genes showing a similar expression onset in our pseudotime analysis are: CD24, a surface marker of chondrocytes in juvenile stages (Lee *et al*, 2016), CHGB which together with CD24 are a central components of the early chondrogenic co-expression module (Feregrino *et al*, 2019), CCDC80, CDH11, known to be expressed during early condensation (Simonneau *et al*, 1995) and SOX5, another main controller of the emergence of chondrocytes from mesenchymal condensations (Ikeda *et al*, 2005).

Phalanx chondrocyte trajectory

From pseudotime 20 until the end of pseudotime (Figure 4 A, B and C).

SOX9 has sustained high expression until the end of this pseudotemporal trajectory. COL9A1 shows a light increase in its expression, in line with what is known about chondrocyte maturation. WNT5A also shows sustained expression, comparable to the expression found at the beginning of the pseudotime. TCF7L2 shows sustained expression in this trajectory, showing only a small decrease; as opposed to the interzone trajectory, where expression levels drop completely, reflecting the expression patterns observed *in situ* (Grall *et al* unpublished).

The expression of CHRDL1 is restricted to the phalanx pseudotime trajectory and it appears relatively early. ACAN, an integral part of the cartilage extracellular matrix (Kiani *et al*, 2002), shows weaker expression than CHRDL1, but is also expressed along this pseudotime trajectory. All of the following genes start to be highly expressed later in pseudotime, almost at the end of our phalanx trajectory: FGFR3,

which has been shown to induce chondrocyte maturation (Su *et al*, 2014), PAPSS1, MATN1, which is expressed in anlagen-forming chondrocytes (Hyde *et al*, 2007), MATN4, SPON1, a bone mass and BMP signaling regulator (Palmer *et al*, 2014), FRZB, which is a known WNT antagonist and chondrogenesis regulator (Enomoto-Iwamoto *et al*, 2002; Leimeister *et al*, 1998; Witte *et al*, 2009), COL9A3, and THBS1, which is expressed in late chondrocyte differentiation (Maumus *et al*, 2017).

CHGB, JUN and MDK show a decrease in their expression along the phalanx pseudotime trajectory. Detail on their functions are discussed hereafter.

Interzone chondrocyte trajectory

From pseudotime 20 until the end of pseudotime (Figure 4 A, B and D).

The classical marker of interzone and joint differentiation GDF5 (Storm & Kingsley, 1996; Koyama *et al*, 2008) shows expression restricted to this pseudotime trajectory, initiating well after the onset of SOX9. Certain genes that show sustained and high expression levels during the phalanx pseudotime trajectory, show downregulation by the end of the interzone trajectory, namely: SOX9, WNT5A and TCF7L2. On the other hand, the expression of CD24 slightly increases along the interzone pseudotime trajectory.

There are a couple of genes which, contrary to what is seen in the phalanx trajectory, keep being expressed through the interzone pseudotime trajectory. One of them is JUN, which is required for joint specification, and an upstream regulator of Wnt signaling (Kan & Tabin, 2013). The other gene is MDK, which has been implicated in the proliferation of articular cartilage (Deng *et al*, 2020).

PTN, has been studied in the context of joint diseases (Pufe *et al*, 2003), but not joint development. SULF1 plays a role in joint development, although its precise mechanism or function remains unknown. Both PTN and SULF1 show restricted expression along the interzone pseudotime trajectory, and they seem to be upregulated even before GDF5 shows high expression.

ENPP2, already known to be highly expressed in the joints (Bächner *et al*, 1999; Hartmann & Tabin, 2001), starts showing high expression after the onset of GDF5 in our pseudotime reconstruction. CHRDL2, a BMP inhibitor upregulated by GDF5 (Nakayama *et al*, 2004; Degenkolbe *et al*, 2015), shows expression restricted to the interzone pseudotime trajectory, and seems to be upregulated later than ENPP2. Three genes show increase in expression by the end of the interzone trajectory in our reconstruction: TCF24, MAP1LC3C and BMP2, which is known to be co-expressed together with GDF5 after the interzone has been established (Francis-West *et al*, 1999; Seemann *et al*, 2005)

Discussion

Here, using our previously published scRNA-seq data from chicken embryonic hind limb stage HH29 (Feregrino *et al*, 2019), we present a pseudotime reconstruction of the phalanx – interzone divergence process. In our pseudotime analysis, we recover

the major known transcriptional changes that occur during this differentiation process. Namely, the transitions from mesenchymal progenitor expressing PRRX1, to chondrocyte progenitors expressing SOX9, followed by the maturation process characterized by a collagen-rich expression program, or the divergence into interzone cells expressing GDF5. Along the pseudotime trajectories, we recapitulate expression dynamics of other genes known to be part of this divergence process, as well as expression dynamics of genes not previously linked with it. Importantly, we observed indications of transcriptional changes linked to the interzone preceding the upregulation of canonical marker GDF5. While we lack transcriptomic markers to undoubtedly recognize the PFR, we also show strong indications that this population is found within our analyzed cells and pseudotime reconstruction.

Importantly, in contrast with a recently published study of the synovial joint development at single-cell resolution (Bian *et al*, 2020), our study aims to analyze the processes prior to interzone formation and GDF5 expression in the digits. In their approach, they use a $Gdf5^{Cre}R26^{EYFP}$ mouse line to isolate and analyze cells from the GDF5+ lineage, which in principle would mean they exclude any molecular changes previous or leading to GDF5 expression. Another difference to our analysis, is that they exclusively analyze knee joints, which might show transcriptional differences to digital joint development. The joints within the digits show different developmental characteristics (Archer *et al*, 2003) and dynamics (Grall *et al,* unpublished). Moreover, the digits of the limb represent and evolutionary novelty of tetrapods, not present in the basal fin structures which already presented the stylopod and zeugopod elements (Coates, 1994; Wagner & Chiu, 2001; Saxena *et al*, 2017).

Our *in silico* transcriptional reconstruction of the divergence process reflects, in general, the current proposed models of interzone emergence, early joint morphogenesis and digit development (Decker *et al*, 2014; Salva & Merrill, 2017; Chijimatsu & Saito, 2019; Hiscock *et al*, 2017; Suzuki *et al*, 2008). We have a trajectory of cells which goes through the following steps: PRRX1+ cell population as the origin, then a PFR state expressing TCF7L2, SMAD1 and SMAD5, then loss of the mesenchymal state and start of a chondrogenic expression program first with SOX9 and followed by COL9A1, SOX5 and others, then, a divergence of fates occurs and cells either go through a chondrogenic maturation or through an interzone and joint emergence process, i.e. cells either become CHRDL1+, FGFR3+, ACAN+, MATN1+ or become GDF5+, ENPP2+, CHRDL2+, BMP2+.

We failed to recover the expected expression pattern of certain genes that play a central role in this divergence process. Among them NOG (Brunet *et al*, 1998), SMAD9 (Suzuki *et al*, 2008), WNT16 (Guo *et al*, 2004), WNT9A (Hartmann & Tabin, 2001), TGFBR2 (Spagnoli *et al*, 2007) and SOX6 (Ikeda *et al*, 2005). We believe the reason for this to be the sparsity and rather poor detection sensitivity of high-throughput scRNA-seq. We cannot draw any conclusions from the total absence of this gene expression data, since this does not mean the genes were not expressed.

Further experiments are needed to assert the temporal expression dynamics of these particular genes in the digital joints development.

We identified two genes with a potential expression upregulation which precedes the expression onset of GDF5, namely: SULF1 and PTN. The expression of these and other genes we found, will be analyzed in a detailed time series of spatial expression experiments. This will help us determine if the order of expression onset we observe in our pseudotime reconstruction recapitulates events *in vivo*. Furthermore, the function of these two, and several of the genes we have observed to show interesting expression dynamics will be tested by perturbing their expression *in vivo*.

Importantly, aside from the 55 genes we present with high differential expression along the pseudotime, our complete analyzed data contains information about the expression dynamics of many more genes that might be of interest to others. An in-depth exploration of the data can be performed, to analyze the pseudotemporal dynamics of other genes. In conclusion, we obtained a pseudotemporal ordering of digit single cells which reflects the early events of chondrogenic differentiation in the digits and we found several genes with potential molecular implications in the cell fate processes that generate the interzone cells. Moreover, we provide a resource that can be valuable for the discovery of new transcriptional programs, or new molecular pathways that drive cell fate decisions during the patterning of the digit.

Methods

The sampling, pre-processing and filtering of the scRNA-seq data set is detailed in our previous analysis (Feregrino *et al*, 2019). Most of our analyses were performed with the toolkit Seurat v3.1.4 (Stuart *et al*, 2019) using, otherwise stated, the default options. For the reanalysis, the cell cycle score was calculated again, using SCRAN (Lun *et al*, 2016) and a list of mouse gene pairs known to covariate with the cell cycle (Scialdone *et al*, 2015), the mouse genes were translated into 1-to-1 chicken orthologues to perform the analysis using BioMart from ENSEMBL release 97 (Kinsella *et al*, 2011; Cunningham *et al*, 2019). We used the “SCTransform” function from Seurat to scale and transform the expression data, we used the δ(S-G2M), fraction of UMI of mitochondrial origin, and the UMI count per cell, as variables to be regressed. We then performed a principal component analysis using the “RunPCA” function from Seurat, using all expressed genes except for the W-linked genes. We determined the amount of PCs to take into account as 21, by the method previously described (see Chapter 1).

In order to produce our tSNEs and exaggerated tSNEs, we followed a methodology, which retains and represents the global structure of the data (Kobak & Berens, 2019), thus giving cell and cluster distances in our visualizations more meaning. We used the FFT-accelerated Interpolation-based t-SNE (Linderman *et al*, 2019) with the following parameters: 21 first PCs, 1000 maximal iterations, learning rate of N-cells/12, perplexities 30 and ~N-cells/100 and an initialization consisting of

the two first PCs divided by their standard deviation times 0.0001. For the exaggerated tSNEs, additional parameters included a late exaggeration coefficient of 4 starting at iteration 250.

To infer cell clusters, we first identified the nearest neighbors of each cell, using the 21 first PCs, and then used the “FindClusters” function from Seurat with the classical Louvain-Jaccard algorithm, and a resolution of 1. Then, we calculated a hierarchical tree of clusters using the “BuildClusterTree” function from Seurat based on the top variable PCs. From the cluster tree, we identified the sister tips – or terminal pairs of clusters –, and performed differential expression tests on each of them. If two clusters resulted to have less than 5 genes significantly (<0.05 adjusted p.value) differentially expressed, they were merged and the process repeated with a new tree of clusters.

We performed differential expression analyses using MAST (Finak *et al*, 2015), either using the MAST package or its implementation in Seurat. We first calculated the standardized variance of each gene using the “FindVariableFeatures” function of Seurat. We then selected as highly variable genes those with a standardized variance larger than the sample median. For making comparisons across clusters we used normalized but “uncorrected” data, using the δ(S-G2M) as a latent variable. Each time we only tested the highly variable genes expressed in at least 25% of the cells of either cell population. Only genes with an adjusted p-value < 0.05 and log2 fold change > 0.5 were then taken into account as differentially expressed.

To identify our clusters, we first calculated differentially expressed genes between each of the clusters and the rest of the cells. These genes were the base for our cell cluster annotation as “marker genes” for each cell cluster. We identified expression patterns of known marker genes of cell types we were expecting to find. Then, using our lists of DE genes, we consulted spatial expression data repositories like Geisha – Chicken Embryo Gene Expression Database (Darnell *et al*, 2007) and MGI – Mouse Gene Expression Database (Smith *et al*, 2019); in some cases, we also performed a literature review. We integrate all our findings into a cell type and/or state annotation of our cell clusters.

To perform our pseudotime analysis, we first made a subset of our data as described in the main text. We calculated the highly variable genes in this subset by using the “FindVariableFeature” function from Seurat on the normalized corrected data and the “mean variable plot” method. The variability thresholds were: dispersion > 1 and average expression > 0.1. Using these highly variable genes, we calculated several dimensionality reductions: a diffusion map using the R package Destiny (Angerer *et al*, 2016), a PCA and an exaggerated tSNE using the first 20 PCs. Based on the first two dimensions from each reduction, we used Slingshot (Street *et al*, 2018) to calculate pseudotime trajectories. For each calculation we defined the start and end points as described in the main text.

We then tested for differential expression along each of the branches of our pseudotime reconstructions. For this, we used the zero inflated Hurdle model

regression implemented in the “zlm” function of MAST. Here we tested only the genes that had more than 3 UMI counts in any cell along the trajectory being tested, and a loess GAM fit of the pseudotime along the trajectory. The resulting p-values were corrected using the Bonferroni method and the number of genes tested as a variable. To detect the genes which showed consistent differential expression across all pseudotime calculations, we ordered all tested genes based on their adjusted p.values, we then chose the top 100 or 35 genes of each pseudotime calculation and found the set intersect.

References

Allen JM, McGlinn E, Hill A & Warman ML (2013) Autopodial development is selectively impaired by misexpression of chordin-like 1 in the chick limb. *Dev. Biol.* **381:** 159–169

Angerer P, Haghverdi L, Büttner M, Theis FJ, Marr C & Buettner F (2016) Destiny: Diffusion maps for large-scale single-cell data in R. *Bioinformatics* **32:** 1241–1243

Archer CW, Dowthwaite GP & Francis-West P (2003) Development of synovial joints. *Birth Defects Res. Part C - Embryo Today Rev.* **69:** 144–155

Bächner D, Ahrens M, Betat N, Schröder D & Gross G (1999) Developmental expression analysis of murine autotaxin (ATX). *Mech. Dev.* **84:** 121–125

De Bakker MAG, Fowler DA, Oude K Den, Dondorp EM, Carmen Garrido Navas M, Horbanczuk JO, Sire JY, Szczerbińska D & Richardson MK (2013) Digit loss in archosaur evolution and the interplay between selection and constraints. *Nature* **500:** 445–448

Bian Q, Cheng YH, Wilson JP, Su EY, Kim DW, Wang H, Yoo S, Blackshaw S & Cahan P (2020) A single cell transcriptional atlas of early synovial joint development. *Development*

Biesecker LG (2011) Polydactyly: How many disorders and how many genes? 2010 update. *Dev. Dyn.* **240:** 931–942

Bradley EW & Drissi MH (2010) WNT5A regulates chondrocyte differentiation through differential use of the CaN/NFAT and IKK/NF-κB pathways. *Mol. Endocrinol.* **24:** 1581–1593

Brunet LJ, McMahon JA, McMahon AP & Harland RM (1998) Noggin, cartilage morphogenesis, and joint formation in the mammalian skeleton. *Science (80-.).* **280:** 1455–1457

Chesterman ES, Gainey GD, Varn AC, Peterson RE & Kern MJ (2001) Investigation of Prx1 protein expression provides evidence for conservation of cardiac-specific posttranscriptional regulation in vertebrates. *Dev. Dyn.* **222:** 459–470

Chiang C, Litingtung Y, Harris MP, Simandl BK, Li Y, Beachy PA & Fallon JF (2001) Manifestation of the limb prepattern: Limb development in the absence of sonic hedgehog function. *Dev. Biol.* **236:** 421–435

Chijimatsu R & Saito T (2019) Mechanisms of synovial joint and articular cartilage development. *Cell. Mol. Life Sci.* **76:** 3939–3952

Church V, Nohno T, Linker C, Marcelle C & Francis-West P (2002) Wnt regulation of chondrocyte differentiation. *J. Cell Sci.* **115:** 4809–4818

Coates MI (1994) The origin of vertebrate limbs. *Development* **120:** 169–180

Cooper KL, Sears KE, Uygur A, Maier J, Baczkowski KS, Brosnahan M, Antczak D, Skidmore JA & Tabin CJ (2014) Patterning and post-patterning modes of evolutionary digit loss in mammals. *Nature* **511:** 41–45

Cserjesi P, Lilly B, Bryson L, Wang Y, Sassoon DA & Olson EN (1992) MHox: A mesodermally restricted homeodomain protein that binds an essential site in the muscle creatine kinase enhancer. *Development* **115:** 1087–1101

Cunningham F, Achuthan P, Akanni W, Allen J, Amode MR, Armean IM, Bennett R, Bhai J, Billis K, Boddu S, Cummins C, Davidson C, Dodiya KJ, Gall A, Girón CG, Gil L, Grego T, Haggerty L, Haskell E, Hourlier T, et al (2019) Ensembl 2019. *Nucleic Acids Res.* **47:** D745–D751

Dahn RD & Fallon JF (2000) Interdigital regulation of digit identity and homeotic transformation by modulated BMP signaling. *Science (80-.).* **289:** 438–441

Darnell DK, Kaur S, Stanislaw S, Davey S, Konieczka JH, Yatskievych TA & Antin PB (2007) GEISHA: An in situ hybridization gene expression resource for the chicken embryo. *Cytogenet. Genome Res.* **117:** 30–35

Dealy CN, Roth A, Ferrari D, Brown AMC & Kosher RA (1993) Wnt-5a and Wnt-7a are expressed in the developing chick limb bud in a manner suggesting roles in pattern formation along the proximodistal and dorsoventral axes. *Mech. Dev.* **43:** 175–186

Decker RS, Koyama E & Pacifici M (2014) Genesis and morphogenesis of limb synovial joints and articular cartilage. *Matrix Biol.* **39:** 5–10

Degenkolbe E, Schwarz C, Ott CE, König J, Schmidt-Bleek K, Ellinghaus A, Schmidt T, Lienau J, Plöger F, Mundlos S, Duda GN, Willie BM & Seemann P (2015) Improved bone defect healing by a superagonistic GDF5 variant derived from a patient with multiple synostoses syndrome. *Bone* **73:** 111–119

Deng Q, Yu X, Deng S, Ye H, Zhang Y, Han W, Li J & Yu Y (2020) Midkine promotes articular chondrocyte proliferation through the MK-LRP1-nucleolin signaling pathway. *Cell. Signal.* **65:**

Duprez D, Esther EJ, Richardson MK, Archer CW, Wolpert L, Brickell PM & Francis-West PH (1996) Overexpression of BMP-2 and BMP-4 alters the size and shape of developing skeletal elements in the chick limb. *Mech. Dev.* **57:** 145–157

Enomoto-Iwamoto M, Kitagaki J, Koyama E, Tamamura Y, Wu C, Kanatani N, Koike T, Okada H, Komori T, Yoneda T, Church V, Francis-West PH, Kurisu K, Nohno T, Pacifici M & Iwamoto M (2002) The Wnt antagonist Frzb-1 regulates chondrocyte maturation and long bone development during limb skeletogenesis. *Dev. Biol.* **251:** 142–156

Fedak TJ & Hall BK (2004) Perspectives on hyperphalangy: Patterns and processes. *J. Anat.* **204:** 151–163

Feregrino C, Sacher F, Parnas O & Tschopp P (2019) A single-cell transcriptomic atlas of the developing chicken limb. *BMC Genomics* **20:**

Finak G, McDavid A, Yajima M, Deng J, Gersuk V, Shalek AK, Slichter CK, Miller HW, McElrath MJ, Prlic M, Linsley PS & Gottardo R (2015) MAST: A flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.* **16:**

Francis-West PH, Parish J, Lee K & Archer CW (1999) BMP/GDF-signalling interactions during synovial joint development. *Cell Tissue Res.* **296:** 111–119

Genzer MA & Bridgewater LC (2007) A Col9a1 enhancer element activated by two interdependent SOX9 dimers. *Nucleic Acids Res.* **35:** 1178–1186

Guo X, Day TF, Jiang X, Garrett-Beal L, Topol L & Yang Y (2004) Wnt/β-catenin signaling is sufficient and necessary for synovial joint formation. *Genes Dev.* **18:** 2404–2417

Hall BK & Miyake T (1995) Divide, accumulate, differentiate: Cell condensation in skeletal development revisited. *Int. J. Dev. Biol.* **39:** 881–893

Hartmann C & Tabin CJ (2000) Dual roles of Wnt signaling during chondrogenesis in the chicken limb. *Development* **127:** 3141–3159

Hartmann C & Tabin CJ (2001) Wnt-14 plays a pivotal role in inducing synovial joint formation in the developing appendicular skeleton. *Cell* **104:** 341–351

Healy C, Uwanogho D & Sharpe PT (1996) Expression of the chicken Sox9 gene marks the onset of cartilage differentiation. In *Annals of the New York Academy of Sciences* pp 261–262.

Hiscock TW, Tschopp P & Tabin CJ (2017) On the Formation of Digits and Joints during Limb Development. *Dev. Cell* **41:** 459–465

Holder N (1977) An experimental investigation into the early development of the chick elbow joint. *J. Embryol. Exp. Morphol.* **Vol. 39:** 115–127

Hyde G, Dover S, Aszodi A, Wallis GA & Boot-Handford RP (2007) Lineage tracing using matrilin-1 gene expression reveals that articular chondrocytes exist as the joint interzone forms. *Dev. Biol.* **304:** 825–833

Ikeda T, Kawaguchi H, Kamekura S, Ogata N, Mori Y, Nakamura K, Ikegawa S & Chung U Il (2005) Distinct roles of Sox5, Sox6, and Sox9 in different stages of chondrogenic differentiation. *J. Bone Miner. Metab.* **23:** 337–340

Kan A & Tabin CJ (2013) c-Jun is required for the specification of joint cell fates. *Genes Dev.* **27:** 514–524

Kavanagh KD, Shoval O, Winslow BB, Alon U, Leary BP, Kan A & Tabin CJ (2013) Developmental bias in the evolution of phalanges. *Proc. Natl. Acad. Sci. U. S. A.* **110:** 18190–18195

Kawakami Y, Wada N, Nishimatsu S ichiro, Ishikawa T, Noji S & Nohno T (1999) Involvement of Wnt-5a in chondrogenic pattern formation in the chick limb bud. *Dev. Growth Differ.* **41:** 29–40

Khan IM, Redman SN, Williams R, Dowthwaite GP, Oldfield SF & Archer CW (2007) The Development of Synovial Joints. *Curr. Top. Dev. Biol.* **79:** 1–36

Kiani C, Chen L, Wu YJ, Yee AJ & Yang BB (2002) Structure and function of aggrecan. *Cell Res.* **12:** 19–32

Kinsella RJ, Kähäri A, Haider S, Zamora J, Proctor G, Spudich G, Almeida-King J, Staines D, Derwent P, Kerhornou A, Kersey P & Flicek P (2011) Ensembl BioMarts: A hub for data retrieval across taxonomic space. *Database* **2011:**

Kobak D & Berens P (2019) The art of using t-SNE for single-cell transcriptomics. *Nat. Commun.* **10:**

Koyama E, Shibukawa Y, Nagayama M, Sugito H, Young B, Yuasa T, Okabe T, Ochiai T, Kamiya N, Rountree RB, Kingsley DM, Iwamoto M, Enomoto-Iwamoto M & Pacifici M (2008) A distinct cohort of progenitor cells participates in synovial joint and articular cartilage formation during mouse limb skeletogenesis. *Dev. Biol.* **316:** 62–73

Kozhemyakina E, Lassar AB & Zelzer E (2015) A pathway to bone: Signaling molecules and transcription factors involved in chondrocyte development and maturation. *Dev.* **142:** 817–831

Lee J, Smeriglio P, Dragoo J, Maloney WJ & Bhutani N (2016) CD24 enrichment protects while its loss increases susceptibility of juvenile chondrocytes towards inflammation. *Arthritis Res. Ther.* **18:**

Leimeister C, Bach A & Gessler M (1998) Developmental expression patterns of mouse sFRP genes encoding members of the secreted frizzled related protein family. *Mech. Dev.* **75:** 29–42

Linderman GC, Rachh M, Hoskins JG, Steinerberger S & Kluger Y (2019) Fast interpolation-based t-SNE for improved visualization of single-cell RNA-seq data. *Nat. Methods* **16:** 243–245

Logan M, Martin JF, Nagy A, Lobe C, Olson EN & Tabin CJ (2002) Expression of Cre Recombinase in the developing mouse limb bud driven by a Prxl enhancer. *Genesis* **33:** 77–80

Lun ATL, McCarthy DJ & Marioni JC (2016) A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. *F1000Research* **5:** 2122

Mariani F V., Ahn CP & Martin GR (2008) Genetic evidence that FGFs have an instructive role in limb proximal-distal patterning. *Nature* **453:** 401–405

Maumus M, Manferdini C, Toupet K, Chuchana P, Casteilla L, Gachet M, Jorgensen C, Lisignoli G & Noël D (2017) Thrombospondin-1 partly mediates the cartilage protective effect of adipose-derived mesenchymal stem cells in osteoarthritis. *Front. Immunol.* **8:**

Mikasa M, Rokutanda S, Komori H, Ito K, Tsang YS, Date Y, Yoshida CA & Komori T (2011) Regulation of Tcf7 by Runx2 in chondrocyte maturation and proliferation. *J. Bone Miner. Metab.* **29:** 291–299

Montero JA, Lorda-Diez CI, Gañan Y, Macias D & Hurle JM (2008) Activin/TGFβ and BMP crosstalk determines digit chondrogenesis. *Dev. Biol.* **321:** 343–356

Nakayama N, Han C ya E, Scully S, Nishinakamura R, He C, Zeni L, Yamane H, Chang D, Yu D, Yokota T & Wen D (2001) A novel chordin-like protein inhibitor for bone morphogenetic proteins expressed preferentially in mesenchymal cell lineages. *Dev. Biol.* **232:** 372–387

Nakayama N, Han CYE, Cam L, Lee JI, Pretorius J, Fisher S, Rosenfeld R, Scully S, Nishinakamura R, Duryea D, Van G, Bolon B, Yokota T & Zhang K (2004) A novel chordin-like BMP inhibitor, CHL2, expressed preferentially in chondrocytes of developing cartilage and osteoarthritic joint cartilage. *Development* **131:** 229–240

Niswander L, Tickle C, Vogel A, Booth I & Martin GR (1993) FGF-4 replaces the apical ectodermal ridge and directs outgrowth and patterning of the limb. *Cell* **75:** 579–587

Nohno T, Koyama E, Myokai F, Taniguchi S, Ohuchi H, Saito T & Noji S (1993) A chicken homeobox gene related to drosophila paired is predominantly expressed in the developing limb. *Dev. Biol.* **158:** 254–264

Noji S, Koyama E, Myokai F, Nohno T, Ohuchi H, Nishikawa K & Taniguchi S (1993) Differential expression of three chick FGF receptor genes, FGFR1, FGFR2 and FGFR3, in limb and feather development. *Prog. Clin. Biol. Res.* **383 B:** 645–654

Palmer GD, Attur MG, Yang Q, Liu J, Moon P, Beier F & Abramson SB (2014) F-spondin deficient mice have a high bone mass phenotype. *PLoS One* **9:**

Petit F, Sears KE & Ahituv N (2017) Limb development: A paradigm of gene regulation. *Nat. Rev. Genet.* **18:** 245–258

Pufe T, Bartscher M, Petersen W, Tillmann B & Mentlein R (2003) Expression of pleiotrophin, an embryonic growth and differentiation factor, in rheumatoid arthritis. *Arthritis Rheum.* **48:** 660–667

Riddle RD, Johnson RL, Laufer E & Tabin C (1993) Sonic hedgehog mediates the polarizing activity of the ZPA. *Cell* **75:** 1401–1416

Salva JE & Merrill AE (2017) Signaling networks in joint development. *Dev. Dyn.* **246:** 262–274

Sanz-Ezquerro JJ & Tickle C (2003) Fgf Signaling Controls the Number of Phalanges and Tip Formation in Developing Digits. *Curr. Biol.* **13:** 1830–1836

Saunders JW (1948) The proximo-distal sequence of origin of the parts of the chick wing and the role of the ectoderm. *J. Exp. Zool.* **108:** 363–403

Saxena A, Towers M & Cooper KL (2017) The origins, scaling and loss of tetrapod digits. *Philos. Trans. R. Soc. B Biol. Sci.* **372:**

Schmid M, Smith J, Burt DW, Aken BL, Antin PB, Archibald AL, Ashwell C, Blackshear PJ, Boschiero C, Brown CT, Burgess SC, Cheng HH, Chow W, Coble DJ, Cooksey A, Crooijmans RPMA, Damas J, Davis RVN, De Koning DJ, Delany ME, et al (2015) Third Report on Chicken Genes

and Chromosomes 2015: *Cytogenet. Genome Res.* **145:** 78–179

Scialdone A, Natarajan KN, Saraiva LR, Proserpio V, Teichmann SA, Stegle O, Marioni JC & Buettner F (2015) Computational assignment of cell-cycle stage from single-cell transcriptome data. *Methods* **85:** 54–61

Sears KE, Behringer RR, Rasweiler IV JJ & Niswander LA (2006) Development of bat flight: Morphologic and molecular evolution of bat wing digits. *Proc. Natl. Acad. Sci. U. S. A.* **103:** 6581–6586

Seemann P, Schwappacher R, Kjaer KW, Krakow D, Lehmann K, Dawson K, Stricker S, Pohl J, Plöger F, Staub E, Nickel J, Sebald W, Knaus P & Mundlos S (2005) Activating and deactivating mutations in the receptor interaction site of GDF5 cause symphalangism or brachydactyly type A2. *J. Clin. Invest.* **115:** 2373–2381

Shwartz Y, Viukov S, Krief S & Zelzer E (2016) Joint Development Involves a Continuous Influx of Gdf5-Positive Cells. *Cell Rep.* **15:** 2577–2587

Simonneau L, Kitagawa M, Suzuki S & Thiery JP (1995) Cadherin 11 expression marks the mesenchymal phenotype: Towards new functions for cadherins?? *Cell Commun. Adhes.* **3:** 115–130

Smith CM, Hayamizu TF, Finger JH, Bello SM, McCright IJ, Xu J, Baldarelli RM, Beal JS, Campbell J, Corbani LE, Frost PJ, Lewis JR, Giannatto SC, Miers D, Shaw DR, Kadin JA, Richardson JE, Smith CL & Ringwald M (2019) The mouse Gene Expression Database (GXD): 2019 update. *Nucleic Acids Res.* **47:** D774–D779

Spagnoli A, O'Rear L, Chandler RL, Granero-Molto F, Mortlock DP, Gorska AE, Weis JA, Longobardi L, Chytil A, Shimer K & Moses HL (2007) TGF-β signaling is essential for joint morphogenesis. *J. Cell Biol.* **177:** 1105–1117

Storm EE & Kingsley DM (1996) Joint patterning defects caused by single and double mutations in members of the bone morphogenetic protein (BMP) family. *Development* **122:** 3969–3979

Street K, Risso D, Fletcher RB, Das D, Ngai J, Yosef N, Purdom E & Dudoit S (2018) Slingshot: Cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics* **19:**

Stricker S & Mundlos S (2011) Mechanisms of digit formation: Human malformation syndromes tell the story. *Dev. Dyn.* **240:** 990–1004

Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM, Hao Y, Stoeckius M, Smibert P & Satija R (2019) Comprehensive Integration of Single-Cell Data. *Cell* **177:** 1888-1902.e21

Su N, Jin M & Chen L (2014) Role of FGF/FGFR signaling in skeletal development and homeostasis: Learning from mouse models. *Bone Res.* **2:**

Suzuki T, Hasso SM & Fallon JF (2008) Unique SMAD1/5/8 activity at the phalanx-forming region determines digit identity. *Proc. Natl. Acad. Sci. U. S. A.* **105:** 4185–4190

Tabin C & Wolpert L (2007) Rethinking the proximodistal axis of the vertebrate limb in the molecular era. *Genes Dev.* **21:** 1433–1442

Taniguchi N, Yoshida K, Ito T, Tsuda M, Mishima Y, Furumatsu T, Ronfani L, Abeyama K, Kawahara K, Komiya S, Maruyama I, Lotz M, Bianchi ME & Asahara H (2007) Stage-Specific Secretion of HMGB1 in Cartilage Regulates Endochondral Ossification. *Mol. Cell. Biol.* **27:** 5650–5663

Tsumaki N, Nakase T, Miyaji T, Kakiuchi M, Kimura T, Ochi T & Yoshikawa H (2002) Bone morphogenetic protein signals are required for cartilage formation and differently regulate joint development during skeletogenesis. *J. Bone Miner. Res.* **17:** 898–906

Wagner GP & Chiu CH (2001) The Tetrapod Limb: A Hypothesis on Its Origin. *J. Exp. Zool.* **291:** 226–240

Wang Z, Young RL, Xue H & Wagner GP (2011) Transcriptomic analysis of avian digits reveals conserved and derived digit identities in birds. *Nature* **477:** 583–587

Witte F, Chan D, Economides AN, Mundlos S & Stricker S (2010) Receptor tyrosine kinase-like orphan receptor 2 (ROR2) and Indian hedgehog regulate digit outgrowth mediated by the phalanx-forming region. *Proc. Natl. Acad. Sci. U. S. A.* **107:** 14211–14216

Witte F, Dokas J, Neuendorf F, Mundlos S & Stricker S (2009) Comprehensive expression analysis of all Wnt genes and their major secreted antagonists during mouse limb development and cartilage differentiation. *Gene Expr. Patterns* **9:** 215–223

Wright E, Hargrave MR, Christiansen J, Cooper L, Kun J, Evans T, Gangadharan U, Greenfield A & Koopman P (1995) The Sry-related gene Sox9 is expressed during chondrogenesis in mouse embryos. *Nat. Genet.* **9:** 15–20

Yang Y, Topol L, Lee H & Wu J (2003) Wnt5a and Wnt5b exhibit distinct activities in coordinating chondrocyte proliferation and differentiation. *Development* **130:** 1003–1015

Yoon BS, Ovchinnikov DA, Yoshii I, Mishina Y, Behringer RR & Lyons KM (2005) Bmpr1a and Bmpr1b have overlapping functions and are essential for chondrogenesis in vivo. *Proc. Natl. Acad. Sci. U. S. A.* **102:** 5062–5067

Zhang P, Jimenez SA & Stokes DG (2003) Regulation of human COL9A1 gene expression: Activation of the proximal promoter region by SOX9. *J. Biol. Chem.* **278:** 117–123

Zhang Z, Yu X, Zhang Y, Geronimo B, Løvlie A, Fromm SH & Chen Y (2000) Targeted misexpression of constitutively active BMP receptor-IB causes bifurcation, duplication, and posterior transformation of digit in mouse limb. *Dev. Biol.* **220:** 154–167

Zhu J, Nakamura E, Nguyen MT, Bao X, Akiyama H & Mackem S (2008) Uncoupling Sonic Hedgehog Control of Pattern and Expansion of the Developing Limb Bud. *Dev. Cell* **14:** 624–632

Zou H, Wieser R, Massagué J & Niswander L (1997) Distinct roles of type I bone morphogenetic protein receptors in the formation and differentiation of cartilage. *Genes Dev.* **11:** 2191–2203

Zuniga A, Zeller R & Probst S (2012) The molecular basis of human congenital limb malformations. *Wiley Interdiscip. Rev. Dev. Biol.* **1:** 803–822

Chapter 4

CONVERGENT CELL FATE SPECIFICATION IN THE DEVELOPING VERTEBRATE SKELETON AT SINGLE CELL RESOLUTION

Christian Feregrino

Abstract

The development of the skeleton is a process that occurs in different places throughout the whole vertebrate embryo. The skeletal tissue, namely cartilage and bone, as well as the cells responsible for it, the chondrocytes and osteoblasts are supposed to be equivalent in function in all the different anatomical structures. Nevertheless, skeletal elements from different parts of the embryo have very different embryonic progenitor pools: the axial skeleton derives from the paraxial mesoderm, the appendicular skeleton derives from the lateral plate mesoderm, and most of the cranial skeleton derives from the cranial neural crest cells. How can cells with different developmental origins and phenotypically distinct tissue niches converge into the same functional cell type? Here, using scRNA-seq, we study the transcriptional dynamics underlying this convergence process. We sampled embryonic tissues where different parts of the skeleton arise, namely: brachial somites, the frontonasal prominence and the developing limb bud. We identify the different cell populations that are likely part of the chondrogenic process, in the three embryonic compartments, and use bioinformatics approaches to compare across our samples, in order to identify commonalities of the temporal dynamics of the transcriptional changes. Importantly, we also establish a framework for the analysis of heterotopic xenografted tissue using scRNA-seq. Collectively, in this study, we lay the basis for further exploration of the skeletogenic convergence process, and the integration with other types of genomic data.

Introduction

Vertebrates have evolved a striking diversity of morphological and physiological adaptations to thrive in most ecosystems. During their evolutionary history and transitions from water to land, then to air or back to water, countless body shape modifications have occurred. However, the basic skeletal plan of a vertebrate is preserved throughout: a vertebral column and a cranial skeleton. Gnatostomes – fishes and tetrapods – possess a full skull (cranium and jaw) and paired appendages. The paired appendages of tetrapods are called limbs, and have components which constitute evolutionary innovations not homologous to any structure in the fish fins (Wagner & Chiu, 2001). The basic modules of the tetrapod ground body plan are mainly defined by the skeleton and its three main parts: the axial, cranial and appendicular skeleton.

Skeletogenesis is thus a crucial process during vertebrate development. The skeleton begins as mesenchymal tissue, in the sites where the skeletal elements (cartilage or bone) will develop. These undifferentiated mesenchymal cells with skeletogenic potential produce an extra-cellular matrix composed of collagen-1, fibronectin and hyaluronic acid (reviewed in: Shum et al. 2003). The skeletogenic mesenchymal cells are multipotent, able to give rise to chondrocytes, osteoblasts and other skeletal cells like synovial cells, tenocytes, stromal cells, or endothelial cells. The different skeletogenic lineages are defined by the activation of specific transcriptional profiles. Among others, the chondrocyte lineage is driven by the transcription factor SOX9 (Wright *et al*, 1995; Healy *et al*, 1996; Akiyama *et al*, 2002), the osteoblast lineage is driven by IHH (Long *et al*, 2004), Wnt signaling (Gong *et al*, 2001; Rawadi *et al*, 2003), as well as the RUNX2 transcription factor (Shimoyama *et al*, 2007), and the lineages of the other skeletal cells are defined, among other mechanisms, mainly by Wnt signaling activity (Guo *et al*, 2004; Hartmann & Tabin, 2001; Zhou *et al*, 2004; Masckauchán *et al*, 2005).

After the skeletogenic mesenchymal cells have been established, skeletal elements can be formed by different processes: chondrogenesis, chondrogenesis followed by endochondral ossification, or direct intramembranous ossification. Endochondral ossification and chondrogenesis are the most widespread and better understood skeletogenic processes. During chondrogenesis, a cartilaginous skeletal element is built by chondrocytes. When endochondral ossification follows, the cartilaginous skeletal element serves as an anlagen, which will be later replaced by bone tissue. At the beginning of this process, the mesenchymal cells with skeletogenic potential stop proliferating and start to condensate, expressing several adhesion proteins (reviewed in Shum et al. 2003).

After condensation, cells undergo chondrocyte differentiation driven, among others, by the transcription factor SOX9 (Schafer *et al*, 1996), and produce an extracellular matrix rich in collagen creating a cartilage primordium. These primordia start growing longitudinally and acquire their characteristic shape. The chondrocytes in the anlagen are organized in layers of increasing maturation, ordered towards the

center of the cartilage element, starting by proliferating chondrocytes (expressing SOX9, SOX5, SOX6, MATN1 and FGFR3 among others), followed by prehypertrophic (expressing PTHRP, PPR and IHH among others) and hypertrophic chondrocytes (expressing COL10A1). After the hypertrophic stage, chondrocytes enter the terminal stage and later die. The lacunae left by the chondrocytes is then invaded by osteogenic cells that remove the cartilage matrix and lay down the bone matrix (reviewed in: Gómez-Picos & Eames 2015; Kozhemyakina et al. 2015).

Intramembranous ossification, on the other hand, only occurs in certain bones of the skull (D'Souza *et al*, 2010; Couly *et al*, 1993). This kind of ossification occurs directly, without the replacement of a cartilage anlagen. Here, skeletogenic mesenchymal cells condensate into compact nodules where the future bones will form. These cells first produce an organic matrix that later is calcified (reviewed in: Gilbert 2000). Both kinds of ossifications are nonetheless carried out at the cellular level by osteoblasts and osteocytes, expressing the transcription factor RUNX2 as a master osteogenic regulator (Shirai *et al*, 2019).

Wherever skeletal elements are formed throughout the embryo, chondrocytes and osteoblasts fulfill similar functions, and likely express similar master regulators and main components of their extracellular matrix. Nonetheless, the different parts of the skeleton have different developmental origins, as well as different evolutionary history (Wagner & Chiu, 2001; Hirasawa & Kuratani, 2015). For the majority of cell types, cell fate specification can be linearly traced back from the final cell type to their progenitor pools in a developmental kinship lineage model (Marioni & Arendt, 2017). However, in the case of skeletogenic cells, the specification process is non-linear, as there are several progenitors giving rise to the same functional cell type. The axial skeleton is derived from the somitic sclerotome, most of the cranial skeleton arises from cranial neural crest (CNC) cells, while the appendicular skeleton is derived from the lateral plate mesoderm (LPM). It's important to mention that most of the cranial bones are formed through intramembranous ossification of neural crest skeletogenic cells (Couly *et al*, 1993), and therefore can arguably be considered as a completely different cell fate specification process. Nonetheless, some neural crest-derived bones of the skull do arise trough endochondral ossification (D'Souza *et al*, 2010; Couly *et al*, 1993), making the specification process of these particular skeletal elements comparable to that of the other two lineages.

Skeletogenic cells not only arise from different progenitor lineages, but also in very different embryonic regions at different times: first along the spinal axis, then in the head, and lastly in the limbs. The properties of the surrounding tissues where the different chondrogenic processes occur, as well as the environment of patterning signals to which they are subject are different between the embryonic regions. While some patterning agents are shared among different chondrogenic processes (e.g. SHH, RA, FGF, or GLI), chondrogenic induction seems to differ among them. SHH is responsible for inducing the chondrogenic competence of the sclerotome of the somites (Zeng *et al*, 2002; Murtaugh *et al*, 1999), FGF2 (Ido & Ito, 2006; Sarkar *et al*,

2001), FGF8 (Green *et al*, 2017; John *et al*, 2011), and BMP4 (Kumar *et al*, 2012) seem to be responsible for the chondrogenic potential of the cranial neural crest, and in the limb, the combination of Wnt and Fgf8 signaling modulates the chondrogenic fate of the LPM (ten Berge *et al*, 2008; Lewandoski *et al*, 2000).

Cell fate decisions of chondrogenic cells – or any other cell –, can be understood as a progressive series of cell type commitments. These commitments, can be traced back to a progressive series of changes in their transcriptional profile and developmental potential. To produce a specific transcriptional profile, and have constrained developmental potential, a combination of a specific regulatory complexes and their corresponding access to the genome are necessary (Arendt *et al*, 2016). Somatic sclerotome, cranial neural crest and LPM cells must differ in these aspects, as they give rise to very different other cell types. Nonetheless, after chondrogenesis, their lineages all converge into a very similar transcriptional profile. Preliminary bulk RNA-seq data from mice has shown that chondrocytes from the three different lineages have very similar overall expression profiles. Intriguingly, however, the expression profiles of transcription factors alone seem to retain a lineage specific signature (Tschopp, unpublished). This suggests that different regulatory complexes activate a very similar downstream transcriptomic profile. It is thus likely that the convergence process involves the *cis*-regulation of a common set of effector genes, by different, lineage-specific complexes of transcription factors. Accordingly, this would also suggest different embryonic environment signals being transduced to the cells and activating different regulatory complexes, as well as different genome accessibility profiles allowing for the binding of these different transcription factors.

In this study, we make use of single-cell RNAseq data to study the transcriptional changes that occur during the skeletogenic process at the three different embryonic regions. While this data is not representative of the regulatory states that the genome of the cells are in, it does reflect transcriptional changes that arise due to the regulatory changes taking place. We intend to make use of pseudotemporal analyses to study the independent transcriptional dynamics that occur during the convergence process. In this way, we try to discover common and specific changes that drive the chondrogenic process in the three different developmental trajectories, in a convergent manner. To this end, we integrate data stemming from the three different embryonic regions and developmental origins, at multiple time-points along their differentiation trajectories.

Moreover, we attempt to explore the effects that cell-extrinsic versus cell-intrinsic regulatory signals have on the transcriptional profile of chondrogenic cells. The developmental potential of the somites has been extensively tested in many ways. One such test consisted in the grafting of undifferentiated somites into the developing limb, which resulted in the development of skeletogenic cartilage (Wachtler *et al*, 1982). Here, the replication of this experiment coupled to scRNA-seq allows us to study the transcriptional changes that such a modification in developmental environment produces, in order to result in a successful chondrocyte differentiation.

We analyzed the single-cell transcriptomes of chicken somites that have been xenografted into the developing limb mesenchyme of quail embryos. So-called “barnyard” experiments using avian species are not standard for scRNA-seq workflows, and this represents the first time a chicken-quail xenograft experiment is analyzed using scRNA-seq.

Results

scRNA-seq analysis of skeletogenic tissues

We sampled three specific embryonic regions, to capture the chondrogenic process of the three different skeletogenic lineages. For the somite sclerotome lineage, we sampled the brachial portion of the developing vertebral column, including all the surrounding tissues. This is, anatomically, the closest to our other two samples, and one of the first to go through chondrogenesis. For the cranial neural crest lineage, we sampled the frontonasal prominence. In this area, all the skeletal elements of the cranium are of neural crest origin. For the LPM lineage, we sampled the forelimb (Figure 1 A).

We designed a sampling strategy aimed to capture the early chondrogenic process, at the three different embryonic locations. We made use of Sox9 expression as an indicator of the early chondrogenic initiation. Performing time series *in-situ* hybridization using a SOX9 probe, we determined the earliest onset of expression in each of our focus locations (data not shown). We used this information to determine our early (right before the expression of SOX9), middle and late time points of sampling for scRNA-seq experiments. For the brachial somites, we chose stages HH12 HH15 and HH20 (~ 45, 50 and 72 hours of development). For the frontonasal neural crest cells we chose stages HH15, HH18 and HH22 (~ 2, 3 and 3.5 days of development). And for the forelimb samples we aimed for HH21, HH24 and HH27 (~ 3.5, 4.5 and 5 days of development) (Hamburger & Hamilton, 1951) (Figure 1 B). We dissected the tissue from several embryos, and pooled them per sample and time point. The tissue samples were dissociated into single cells and then processed using the Chromium 10x Genomics system.

After sequencing, mapping to our custom genome annotation and quality filtering (see Chapter 1), we recovered a comparable amount of cells from each scRNA-seq run. From the brachial region we have: 3,093 cells from stage HH12 (S12), 2,993 cells from S15, and 1,691 cells from the S20 sample. From the frontonasal prominence region we recovered: 2,702 cells from the stage HH15 (N15), 4,558 cells from N18, and 1,208 cells from the N22 sample. Meanwhile, from the forelimbs we obtained: 2,987 cells from the stage HH21 (L21), 5,293 cells from L24, and 2,070 cells from the L27 sample. The consistent low number of cells recovered at later stages might be due to the maturation of the tissue and its extracellular matrix, thereby making dissociation more complicated.

Our brachial somites region sample consisted of two big groups of cells, a mesodermal and an ectodermal one. Within the ectodermal cells we found

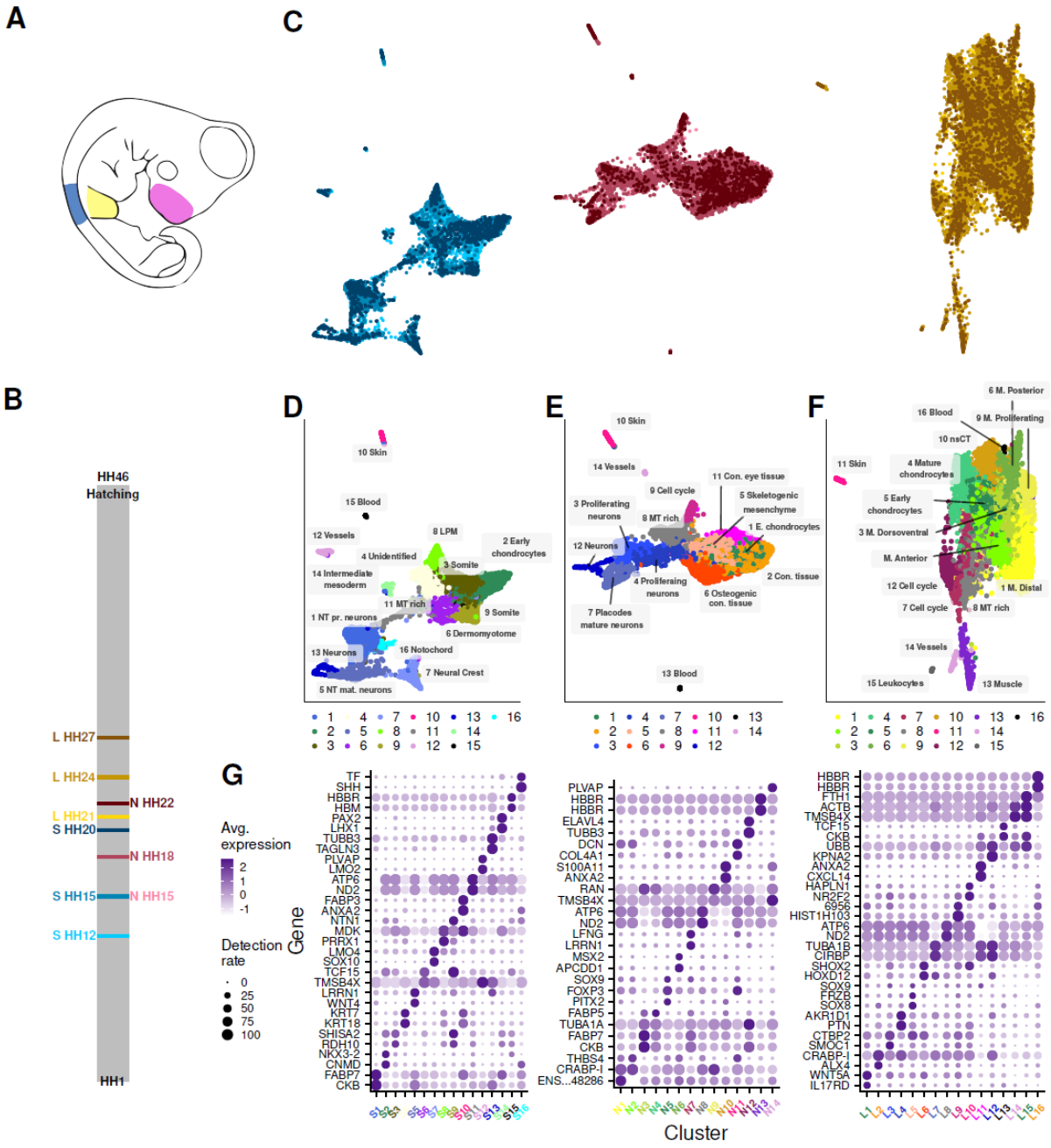


Figure 1 Samples and scRNA-seq analyses of the three skeletogenic lineages: Somite (S), Neural crest (N) and LPM (L). **A** Schematic of the dissection sampling strategy. From the brachial (blue), frontonasal (pink), and forelimb (yellow) regions. **B** Time points of our sampling per embryonic region. **C** Exaggerated tSNE and sample (time point) composition of each data set. Each panel represents an embryonic region, coloring corresponds to B. **D**, **E** and **F** Exaggerated tSNE and clustering of the different data sets. **D** Brachial / Somites samples, **E** Frontonasal / Neural Crest and **F** Limb samples. Cell types or states are annotated. The same or very similar colors across these plots represent the same or similar cell populations. **G** Expression patterns of the top two differentially expressed genes per each cluster. Panels and cluster coloring correspond to D, E and F.

proliferative neural tube cells (cluster 1, c. 1), expressing cell cycle genes and CKB, FABP7, TUBA1A, and CIRBP. We also found more mature and dorsal neural tube cells (c. 5) expressing WNT4, LRRN1 and LFNG. We found as well neurons (c. 13) expressing TAGLN3, TUBB3 and TMSB4Y, and neural crest cells (c. 7) with high expression levels of SOX10 and LMO4. Moreover, we identified notochord cells

(c.16) expressing SHH and TBXT. In the mesodermal group, we found LPM cells (c. 8) with high levels of PRRX1 expression. We also found somite dermomyotome cells (c. 6) expressing TCF15 and MEOX1. We found as well two other somitic clusters, cluster 3 expressing RDH10 and SHISA2, and cluster 9 expressing TCF15 and NTN1. Notably, we found early chondrocytes (c. 2) with high expression levels of SOX9 and CNMD (Figure 1 D and G).

Our frontonasal sample also consisted of two big groups, one with a rather neural fate and another with a rather mesenchymal fate, aside from a cluster of undifferentiated cycling cells (c. 9) expressing TMSB4X, RAN and ACTG1. In the first group we find neurons (c. 12) that are comparable to the ones found in the brachial sample, expressing high levels of TUBB3 and TMSB4Y. We also found more mature neurons probably belonging to placodes (c. 7), which are similar to the mature neural tube neurons found in the brachial sample. These cells express LRRN1 and LFNG. We found as well neural cells expressing high levels of genes associated with the cell cycle, which are similar to our proliferative neural tube cell population from the brachial sample, cluster 3 with expression of CKB and FABP7 and cluster 4 expressing TUBA1A, FABP5 and CIRBP. In the mesenchymal group, we found skeletogenic cells (c. 5) expressing PITX2 and FOXP3. We also found cells likely differentiating towards intramembranous ossification (c. 6) expressing APCDD1 and MSX2. We found as well two clusters of connective tissue cells, one expressing CRABP-I and ALX4 (c. 2), and another which probably constitutes the connective tissue part of the eye (c. 11), with high expression levels of COL4A1 and DCN. Notably, we also found a chondrogenic cluster (c. 1) with high expression levels of SOX9, as well as a gene with ENSEMBL code ENSGALG00000048286, CRABP-I and ALX4 (Figure 1 E and G).

Finally, our limb sample consisted mostly of LPM derived cells, with the exception of invading muscles coming from the somites (c. 13), expressing CKB and PITX2. We identified two clusters of undifferentiated cycling cells, cluster 7 expressing CIRBP, TUBA1B and RAN, and cluster 12 expressing KPNA2, UBB and ACTG1. We also found a cluster of mature non-skeletal connective tissue (nsCT) (c. 10) expressing NR2F2 and HAPLN1. We found as well several clusters of the limb mesenchyme, one undifferentiated mesenchymal population with cycling activity (c. 9), expressing genes HIST1H103 and MKI67, three mesenchymal populations with distinct positional information: dorso-ventral surroundings of skeletal condensations (c. 3) with high expression levels of SMOC1 and CTBP2, distal limb mesenchyme (c. 1) expressing IL17RD and WNT5A, anterior skeletogenic population (c. 2) expressing ALX4 and CRABP-I, and posterior skeletogenic population (c. 6) expressing HOXD12 and SHOX2. Notably, we also found early chondrocytes (c. 5) with expression of SOX9, SOX8 and FRZB, as well as a mature pre-hypertrophic chondrocyte cluster (c. 4) with high expression levels of PTN and AKR1D1 (Figure 1 F and G).

After less successful attempts using Seurat (Supplementary figure 5), and based on an extensive comparison of single-cell data integration methods (Luecken *et al*,

2020), we decided to integrated all of our cells using Scanorama (Hie *et al*, 2019) into a single big data set. Some of the cells that we have deemed as similar – or homologous – among the three embryonic locations indeed clustered together using this approach (Figure 2 A). For example, mature and proliferating NT and cranial neurons occupied the same spaces, where no limb cell was present. Cells we annotated as being in the cell cycle also clustered together, although the proliferative dermomyotome cells occupied the same space. Regarding the skeletogenic and mesenchymal cells, while they appeared on the same broad region of the dimensionality reduction, they were not mixed as we expected. It must be noted that the skin clusters, which appear quite distinct in all of the individual embryonic origin datasets, don't aggregate into a single isolated group of cells in the integrated data set. These cells do, however, occupy two areas around the cycling cells and proliferative dermomyotome.

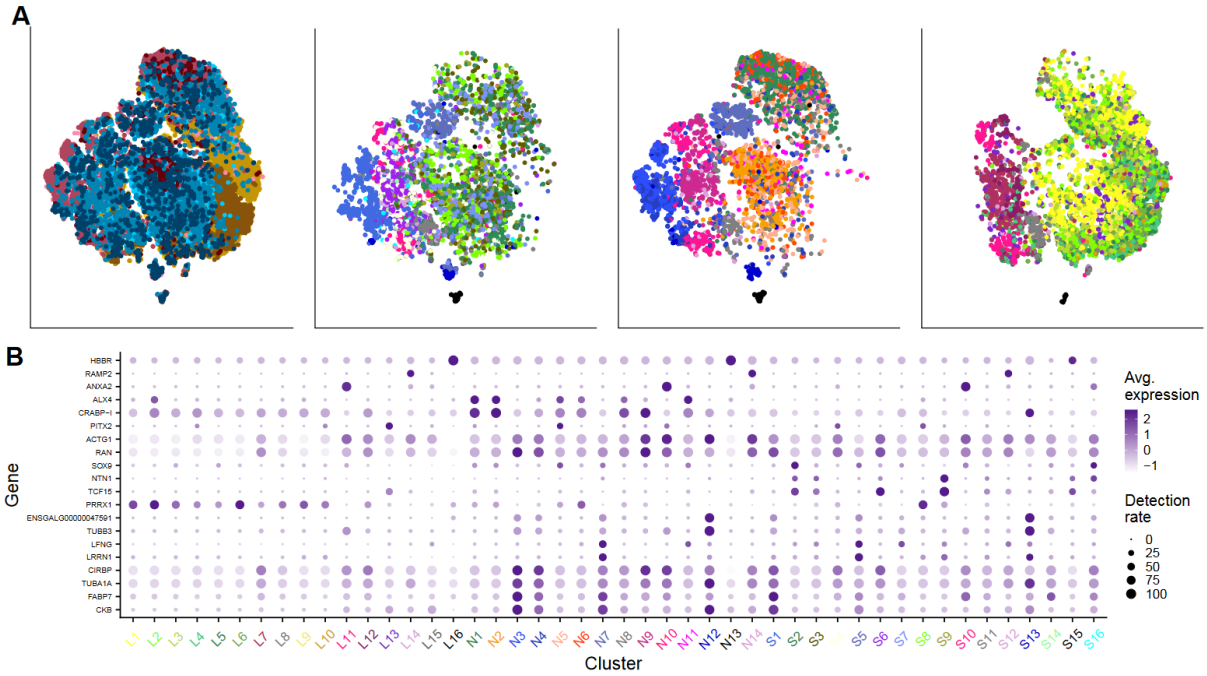


Figure 2 Data integration and cell type correspondence of cells from the three different embryonic regions. **A** Integrated datasets produced by Scanorama. Left: all samples in an integrated dataset, and then embedded using tSNE. Coloring corresponds to Figure 1 B and C. To the right: same embeddings, but showing only the cells from the brachial, frontonasal and limb samples. Order and coloring of the panels corresponds to Figure 1 D, E and F. **B** Expression pattern of selected genes, which show similar patterns across clusters in the different samples. Coloration of the cluster names correspond to Figure 1 D, E and F.

Next, we tried to integrate a subset of the data, only containing early chondrocytes, and the populations which might be part of the chondrogenic lineages. From the somites sample, we chose cluster 2 "early chondrocytes" and clusters 3 and 9, the somites (excluding cluster 8 with a more dermomyotome profile). From the frontonasal samples, we selected cluster 1 "early chondrocytes", clusters 2 "connective tissue" and cluster 5 "skeletogenic mesenchyme" (we excluded cluster 6 cells with an osteogenic profile and cluster 11 which seems to be part of the developing eye). Finally, from the forelimb samples, we chose cluster 5 "early

chondrocytes", cluster 3 "dorso-ventral mesenchyme", cluster 2 "anterior mesenchyme", cluster 1 "distal mesenchyme", as well as cluster 4 "mature pre-hypertrophic chondrocytes" (we excluded the posterior and undifferentiated mesenchyme).

We tried several methods to attempt this data integration, and we present here results from the Seurat and Scanorama approaches. We also employed two different sets of genes to perform the integration: all shared expressed genes (data not shown) and highly variable genes. We couldn't find a meaningful integration, which would show us the different cell populations as discrete clusters, or in an arrangement which would make biological sense. We then tried to run the integration with even less clusters, removing cluster 9 from the somites, cluster 2 from the neural crest and cluster 4 from the limb samples. This, in our perspective, also didn't produce any meaningful results (Figure 3). This indicates that our integration approach can still be improved in the future.

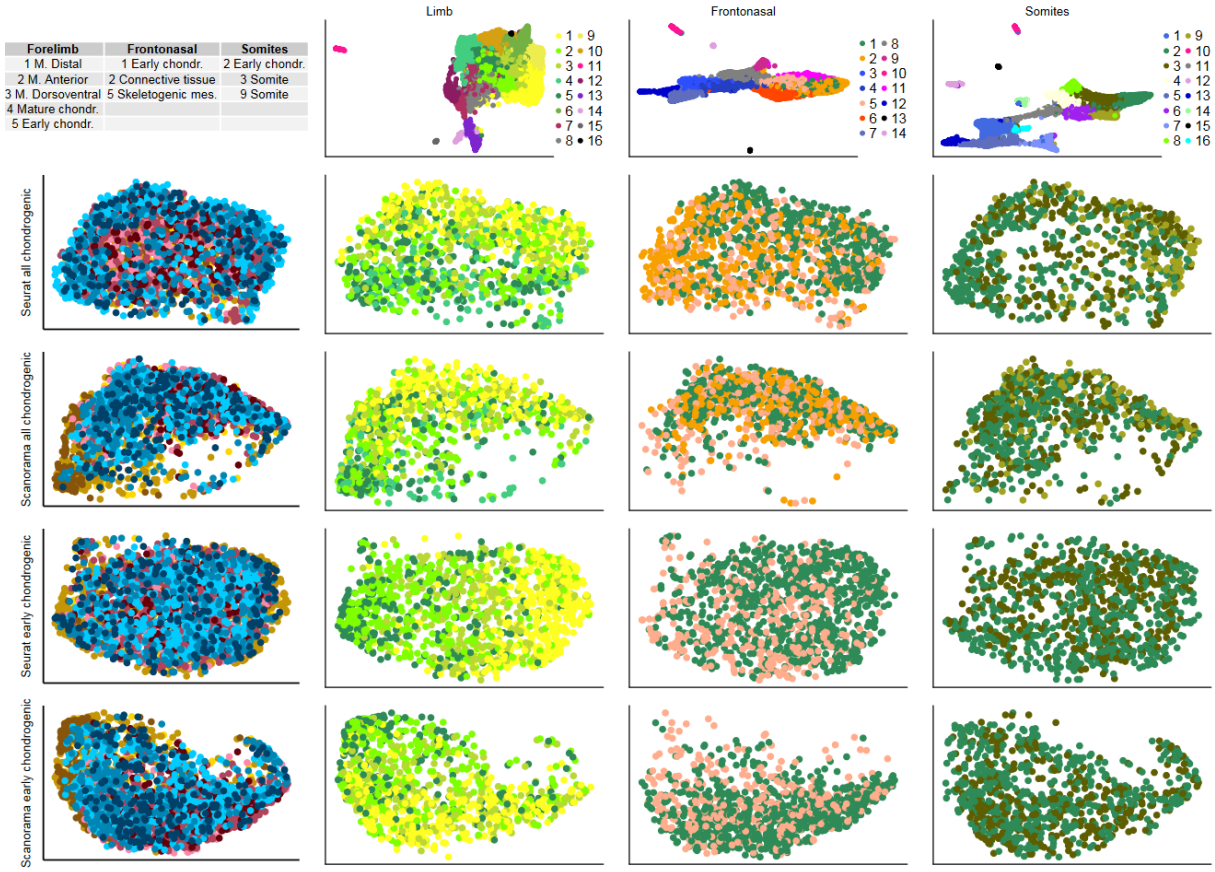


Figure 3 Data integration of the chondrogenic cells. Top left: Cell clusters we have tried to integrate into a single dataset. To the right: original tSNE corresponding to Figure 1. Rows 2 and 3: Integration using the highly variable genes and the approach labeled there. Left: Integration exaggerated tSNE embedding of the data, coloring corresponds to Figure 1 B and C. To the right: cells from each of the embryonic regions, in the integrated embedding. Colors corresponding to the panels in the top. Rows 4 and 5: Same logic as the previous rows, using only clusters L1, L2, L3, L5, N1, N5, S3 and S2.

Chicken – Quail xenograft scRNA-seq analysis.

In an effort to understand the role that cell-intrinsic and cell-extrinsic factors play during the chondrogenesis process, we decided to perform xenograft experiments. As an exploratory assay, chicken developing brachial somites were grafted into the quail developing limb. These grafting experiments were carried out by Chloé Mureau in our lab. Brachial somites were dissected out of stage HH14 chicken developing embryos, they were then soaked in a DiI labeling solution, to be able to detect them later, once grafted into the host tissue. The labeled somites were then grafted into the developing limbs of stage HH22 quail embryos. The quail embryos were then allowed to develop further for 3 more days, until approximately HH29 ~ HH30. After this time point, HH14 ungrafted somites would be in an embryo at stage HH26. The cells were then dissociated following the same procedure as with the rest of our chicken samples, and then processed for scRNA-seq using the Chromium 10x Genomics system.

We reasoned that, it should be possible to distinguish chicken and quail cells using bioinformatics tools. For this, we constructed a joint chicken and quail genome and genome annotation files. The fused genome consisted of a simple concatenation of the two entire genome files. For the genome annotation GTF file, which is required for the counting of sequenced transcripts, we followed a different approach. A simple concatenation of chicken and quail gene annotations would entail significant differences in annotation quality. Moreover, the Chromium 10x Genomics approach already proved to be partially incompatible with the current chicken gene annotation, and despite big efforts (Kawahara-Miki *et al*, 2013), the quail genome and its annotation fall far behind in quality. Therefore, we decided to create a rather brute annotation, in which all mapped reads would be counted, regardless of genomic location. We developed a bioinformatics pipeline to create a fake annotation file, in which each genomic chromosome and scaffold represents a gene, transcript and exon in each of the strand senses. This resulted in 9964 “genes” in each strand sense for the fused chicken and quail genome.

Using this fused chicken-quail genome and annotation files, we used the CellRanger pipeline (10x Genomics) to process two of our chicken samples as a first experimental control. We chose samples Limb HH27 and Frontonasal / Neural Crest HH22 – the former being of relatively low quality compared to the rest of our samples, the latter to ensure a complex combination of cell types. After the mapping and counting process, we calculated cell-by-cell the number of UMIs that aligned with the chicken or the quail part of the fused genome. We observed that practically all cells have all of their reads aligned to the chicken part of the genome, as expected for a sample of purely chicken origin (Figure 4 A). It seems that cells with a high amount of UMIs have very few UMIs that align with the quail genome, being around 0.5% of the total. We found that 1% is the expected amount of UMIs that could be wrongly mapped to the quail genome. Meanwhile, cells of low quality, can have up to 2.5% of their UMIs wrongly aligned to the quail genome (Figure 4 B).

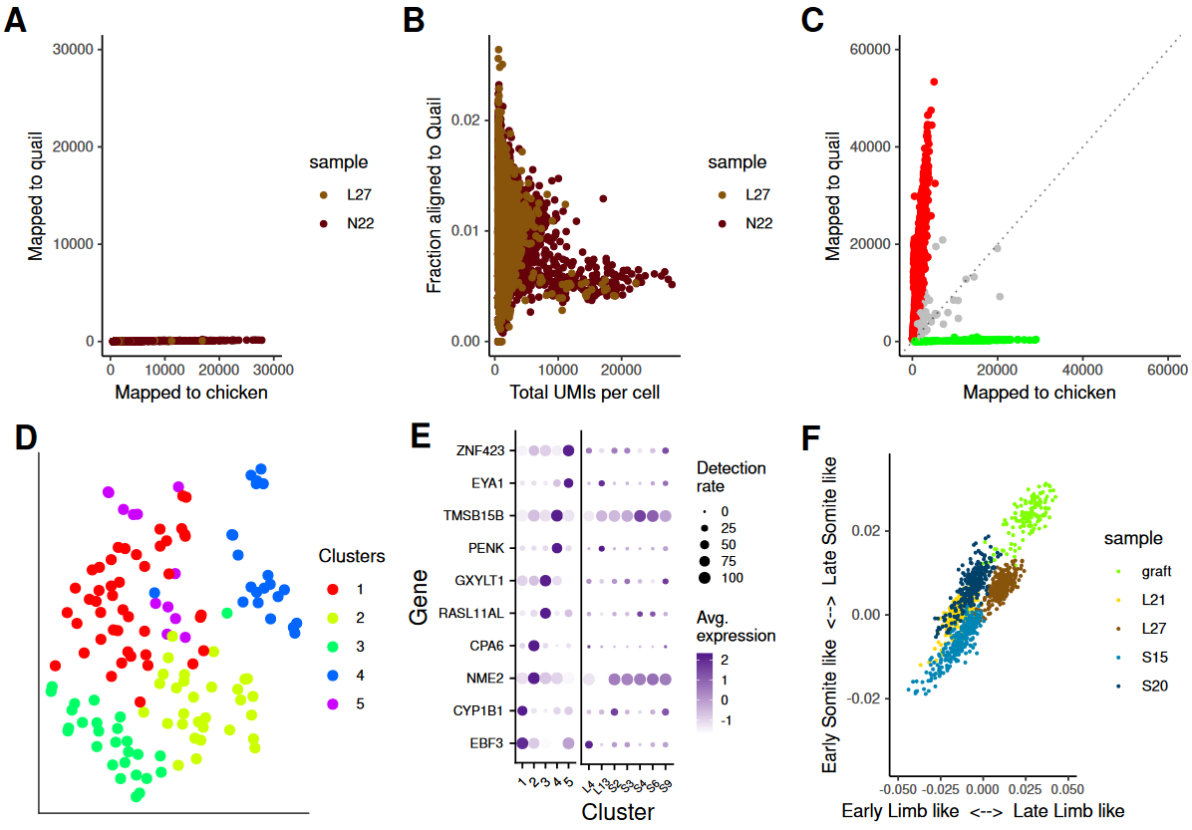


Figure 4 scRNA-seq analyses of the Somite-Limb xenograft experiment. **A** UMI count per cell aligned to the chicken or the quail portion of our concatenated genome. Here, samples from chicken origin: Limb HH27 and Frontonasal prominence HH22. **B** Relation between the total numbers of UMIs per cell and the fraction of UMIs mapped to the quail genome. Here, the same two samples as in A. **C** Number of UMIs per cell aligned to the chicken or the quail portion of our concatenated genome. Here, single cells dissociated from a chicken somite grafted into a quail limb. Diagonal line represents a proportion of 1:1. Red cells have more than 80% of their UMIs mapped to the quail portion of the genome, green cells more than 80% of their UMIs aligned to the chicken portion of the genome, grey cells have proportions in between. **D** tSNE and Louvain-Jaccard clustering of the valid chicken grafted cells. Coloring independent from any other figure or panel. **E** Expression profiles of the top 2 differentially expressed genes from D. To the right are clusters from the chicken samples: Limb cluster 4 “late chondrocytes”, Limb 5 “muscle”, Somite 2 “early chondrocytes”, Somite 3, Somite 4 “undefined”, Somite 6 “dermomyotome” and Somite 9. **F** Differential correlation to expression average of selected limb cells and selected somite cells. X axis runs from higher relative correlation to limb HH21 to higher relative correlation to limb HH27. Y axis runs from higher relative correlation to somites HH15 to higher relative correlation to somites HH20. Here, all the grafted cells, and twice as many randomly selected cells from the 4 reference samples.

Encouraged by these results, we next applied our approach to the grafted samples. As an exploratory assay, we used a relatively relaxed threshold in order to obtain as many chicken cells as possible. We divided the 4550 cells that were deemed as valid by the CellRanger pipeline in three categories based on their proportion of UMIs mapped to each species. Cells with more than 80% of their UMIs mapped to a single species were considered unambiguous, the rest were deemed ambiguous cells which probably represent doublets. Of all cells, 3777 (83%) were of quail origin, 584 (12.8%)

were ambiguous, and 189 (4.2%) of chicken origin (Figure 4 C). We then processed the grafted sample again, but using our custom annotation of the chicken genome, as with the rest of our chicken samples. We then removed all the cells which had not passed our chicken-quail alignment threshold and applied the same quality control filters as with our other data sets. In the end, we had 136 cells of high quality. Although this is a very small number of cells, we processed them using our dimensionality reduction and clustering pipelines based on Seurat.

We found 5 distinct cell clusters (Figure 4 D): cluster 1 with high expression of EBF3 and CYP1B1, cluster 2 with high expression of NME2 and CPA6, cluster 3 with high expression of RASL11AL and GXYLT1, cluster 4 with high expression of PENK and TMSB15B, and cluster 5 with high expression of EYA1 and ZNF423. PENK and EYA1, highly expressed in cluster 4 and 5, also show high specific expression in the Limb cluster 13 “muscle”, which indicates that these two grafted clusters are in a muscle differentiation trajectory. On the other hand, EBF3, enriched in cluster 1 shows high specific expression in the Limb cluster 4 “late chondrocytes”. Moreover CYP1B1, also expressed in cluster 1, shows also high expression in Somites cluster 2 “early chondrocytes”. This indicates that cluster 1 of our grafted cells appears to be on a chondrogenic differentiation trajectory, having high expression of genes that are characteristic of both embryonic regions (Figure 4 E).

We then tested whether the grafted cells are more similar to the cells in the limb, or the cells in the somites. For this, we averaged the expression of the single cells across four different samples. From the limb we independently obtained average expression of the cells from stage HH21 and HH27, the closest to the graft experiment stages. We used most of the cell clusters with a mesodermal origin: all 5 types of mesenchyme (anterior, posterior, dorso-ventral, distal and proliferative), non-skeletal connective tissue and the early and late chondrocytes. Whereas from the somites, we obtained the average expression of cells from stage HH15 and HH20, also the closest to our graft experiments. Here we only used the cell clusters from the intermediate mesoderm and somitic origin (both general somite clusters, an undetermined cluster, the early chondrocytes and the mitochondrial-rich cells). We used these 4 sets of expression averages as reference, and calculated the correlation coefficient of each of the grafted cells to them.

As a control, we randomly sampled double as many cells (272) from the reference populations, and also calculated the correlation coefficients to each of the averaged references. In the end, for each of the cells we calculated the difference between the correlations to the early and late somites expression average and between the correlations to the early and late limb expression average. We then plotted the differential correlations and found that the grafted cells don’t show the same pattern as any of the other tested sets of cells. The correlation of the grafted cells to the average expression of the late limb (HH27) and somite (HH20) samples is higher than their correlation to the early limb (HH21) and somite (HH15) sample. This difference is higher than the one found for the late limb, or late somites cells themselves. Hence,

this seems to point to a unique overall transcriptomic profile triggered within the somite cells when grafted into a limb environment (Figure 4 F).

Discussion

By applying high-throughput scRNA-seq methods on three different embryonic regions where skeletal elements develop, over three different developmental stages, we sampled more than 26,000 single cells in total. In this way, we were able to sample skeletogenic cell populations from three different developmental origins. Namely, the limb mesenchyme originating from the lateral plate mesoderm, the sclerotome mesenchyme originating from the paraxial mesoderm and the craniofacial mesenchyme originating from the neural crest ectoderm. We also recovered transcriptomic information from all the tissues that surround these skeletogenic cell populations. With our approach, we were able to sample different stages of maturation from these cell populations, meaning that we were able to capture skeletogenic cells, the cells that give rise to them, and the cells that they differentiate into. Moreover, given that the skeletogenic process is heavily driven by paracrine signaling produced by structures close to the skeleton forming sites (Lefebvre & Bhattaram, 2010), the transcriptomic profiling of all neighboring cell populations provides the setting for further exploration of cell signaling interactions underlying the skeletogenic processes.

The integration of our data into a single dataset showed that some cell types clearly show shared transcriptional profiles across our sampling. This includes obvious cell types like blood, or blood vessels cells, which should look very similar regardless of the anatomical position. Neuronal populations from the somite and frontonasal samples, as well as cycling cells also clustered together in the dimensionality reduction space. A notable exception are skin cells, which seem to be scattered around the cycling and actively proliferating cells. Another phenomenon we observed is that mesenchymal, chondrogenic and connective tissue cells all group in a big cluster, but are rather randomly scattered. In the integration tSNE, these cells don't show any particular pattern reflecting the distinct cell clusters we have detected when analyzing the embryonic locations individually. Furthermore, it appears that the embryonic origin influences this particular aspect of the integration, limb and frontonasal cells appear to aggregate away from the rest of the cells. It is known that the set of genes used to perform integration analyses, as well as the integration algorithm used directly changes the results of a data integration assay (Luecken *et al*, 2020). It is probable that the set of genes that we chose dilutes the transcriptomic signal of the skin cells, or the mesenchymal cells, and that different gene set combinations might give us better results.

Our sampling strategy ensures that the chondrocyte and skeletogenic cells that we recover indeed arise from different developmental origins. Limb skeletogenic mesenchyme arises exclusively from LPM cells (Logan *et al*, 2002; Gerber *et al*, 2018) which invade the limb field and make the appendages grow. The skeletal elements of the cranial skeleton have different origins depending on the taxonomical group. These

skeletal elements arise from the paraxial mesoderm, cranial neural crest cells, or both (Piekarski *et al*, 2014). However, in general, and in the case of the chicken, most of the mesoderm-derived cranial skeletal elements are located in the top and back of the head. All of the frontal skeletal elements in the skull emerge from migratory cranial neural crest cells. Moreover, bones of the cranial skeleton can be formed by intramembranous or endochondral ossification. And while most of the relatively big CNC-derived bones are formed without a cartilaginous intermediate step, there are many, albeit small, endochondral skeletal elements scattered across the frontal half of the skull, like the Meckel's cartilage, the ethmoid and the hyoid (Couly *et al*, 1993). Due to our dissection strategy, neural crest cells, paraxial mesoderm and lateral plate mesoderm are all present in our brachial samples. Nonetheless, we are confident that the skeletogenic cells samples there are from somitic origin. The LPM only gives rise to skeletogenic structures inside of the limb buds (Prummel *et al*, 2020); likewise, trunk neural crest cells, present at the brachial level, do not differentiate into skeletal tissue either (Nakamura & Ayer-le Lievre, 1982; McGonnell & Graham, 2002).

Once the chondrogenic process has started, cells turn on a transcriptomic program which will lead to the production of a typical chondrocyte profile. This transcriptional profile, consisting of a set of transcription factors, and a collagen-rich extracellular matrix gene set is common and can be recognized regardless of developmental or spatial origin (reviewed in: Lefebvre & Bhattaram 2010). A canonical molecular marker to recognize the onset of the chondrogenic process is its master regulator, the transcription factor SOX9 (Akiyama *et al*, 2005). Nonetheless, being a transcription factor, the detection rate of this gene in high-throughput scRNA-seq assays can be low. Conveniently, the nature of scRNA-seq allows for the identification of cell populations without the need of them expressing only one particular gene, but rather show an overall similar transcriptome. We therefore are able to detect cell populations corresponding to the early SOX9+ chondrocytes in all three of our data sets, although not all these cells have a particularly high expression of SOX9.

Given our plans to examine the pseudotemporal dynamics of gene expression that lead to the phenotypic convergence of chondrocytes, identification of the implicated cells is crucial. The first part of this task is the identification of the early chondrocytes. Thereafter, any population of cells that might derive from these early chondrocytes. But critical to our objectives is the identification of the cells that give rise to the early chondrocytes: the skeletogenic mesenchyme.

In our brachial samples, we found the population corresponding to the early chondrocytes, cluster 2 with high expression levels of SOX9. Particularly in this sample, SOX9 is also expressed as part of patterning process of the neural tube, accordingly, we also find high expression levels in cluster 1, identified as neural tube proliferating neurons. To identify the progenitor population of cluster 2 "early chondrocytes" we considered clusters 3 and 9. These two cell populations were adjacent to the early chondrocyte cluster, in an embedding which captures the global structure of the data. Moreover, with our approach, these clusters showed no clear

evidence of being differentiated cell populations, and were thus deemed as undifferentiated cells. Nonetheless, cluster 9 showed high expression of marker genes from cluster 6 “dermomyotome”, and thus we decided to use cluster 3 as the progenitor of our early chondrocytes.

In our frontonasal prominence samples, we found cluster 5 with relative high expression of SOX9 which we identified as skeletogenic mesenchyme, and cluster 1 and 2 which seem to be derived connective tissue, expressing high levels of CRABP-I. Here, during the integration analyses, we considered cluster 5 as the earliest point of the chondrogenic lineage. This decision was made based on the transcriptional signal of the two other adjacent cell populations: cluster 4 and cluster 8. Cluster 4 showed high expression of genes also expressed by proliferating neurons in the brachial sample. On the other hand, the differential expression analysis for cluster 8 resulted mainly in mitochondrial genes being enriched. This made us think of cluster 8 as uninformative cell population, and not fit to be considered as part of the chondrogenic process.

In our limb samples, cluster 5 was identified as our early chondrocytes, due to the high expression of SOX9, SOX8 and FRZB. The population deriving from this cluster in the chondrogenic process seems to be cluster 4 “mature pre-hypertrophic chondrocytes”. Nevertheless, the identification of the cluster giving rise to the early chondrocytes proved more challenging. We chose adjacent populations cluster 2 “anterior mesenchyme”, cluster 3 “dorso-ventral mesenchyme”, as well as cluster 1 “distal mesenchyme”. Cluster 2 and 3 were chosen based on their adjacency to cluster 5, while cluster 1 was chosen because the current limb development model describes the chondrogenic cells arising from the distal region of the limb.

The attempts to integrate the datasets at the level of chondrogenic cell populations, to later perform pseudotemporal analyses, were not satisfactory to our expectations. The cell clusters didn’t show any meaningful ordering or spatial relationship in the embeddings that would reflect the chondrogenic process. In the future, we will try to consider different sets of clusters or a completely different approach. For example, we have completely ignored the clusters rich in mitochondrial UMIs from the limb and frontonasal samples. The fact that the expression profiles of these clusters are not enriched with chondrogenesis-related genes, doesn’t mean that they are not part of the chondrogenic process. These clusters, aside from the mitochondrial enrichment, show indeed transcriptional profiles relatively similar to those of the imputed early chondrocytes, except for the expression of SOX9. Moreover, we have shown that similar cell clusters can be integrated as part of pseudotime reconstructions of chondrogenic processes (see Chapter 3). On the other hand, analyzing the chondrogenic process separately in each embryonic region, and then making comparisons across the results might be a better strategy. Using pseudotime reconstruction on each of our data sets, and then align and compare the pseudotime trajectory of the genes is plausible (Alpert *et al*, 2018), and has been successfully

done before (Kanton *et al*, 2019). Yet another option is the use of analyses that help to determine the fate and progenitors of cells *in silico* (Lange *et al*, 2020).

Our framework for the scRNA-seq analyses of xenografts showed satisfactory results. We were able to distinguish cells originating from two closely-related species. Chicken and quail are more closely related (ca. 40 MY divergence, calculated using TimeTree meta-analysis (Hedges *et al*, 2015)) than humans and mouse (ca. 90 MY divergence). Mouse and human cells are usually used to calibrate doublet capturing on single-cell sequencing set-ups. Our approach showed that the reads of high quality cells tend to align more faithfully to the correct genome, although the amount of reads that are aligned to the wrong genome is in any case very small. Although in our study, due to the small number of chicken cells recovered, we employed a very relaxed threshold for the amount of wrongly mapped reads per cell, we show that this threshold can be as small as 2.5% for chicken versus quail experiments. Moreover, the quality of the chicken and quail annotations forced us to use a rather crude joint annotation, which could be greatly improved using the very-well curated and maintained gene annotations from other well-stablished genomics model species. While this is not the first study that shows a single-cell transcriptome analysis of grafted tissue (Shi *et al*, 2020), to our knowledge, it's the first time that non-superficial grafted tissue is examined in this way. Furthermore, it's the first approach that attempts to sort cells *in silico* into their respective species of origin. Moreover, we set the basis for analyses of bigger data sets, which will prove helpful once the grafted tissue cell dissociation is improved.

The changes in transcriptomic profiles that drive cells with different origins to converge into the same phenotype don't only occur in a differentiated developmental context, but also have different evolutionary origins. The great challenges that are inherent to the comparison of transcriptomic data across different species disappear, as we compare expression profiles from the same species, and can even represent the same individual. In this sense, different tools and analyses designed for the study of gene expression evolution could be applied to the phenomena we present here. Altogether the work presented here represents the foundations for further analyses we plan to carry out with these and other samples. Single-cell transcriptomic data will be further analyzed to find pseudotemporal gene expression dynamics that recapitulate the convergence of the chondrogenic process. This data will be further complemented by single-cell chromatin accessibility assays, in order to perform analysis to unveil the gene regulatory processes that govern the different chondrogenic differentiation processes. We are also trying to improve our chicken – quail grafting experiments, in order to recover a higher quantity of viable chicken cells and gain statistic power to decouple the cell-intrinsic and cell-extrinsic factors, which drive chondrogenesis. Collectively, these studies constitute the foundations of a comprehensive investigation that will allow us to understand the convergent cell fate specification of the vertebrate skeleton, a defining feature of the developmental and evolutionary processes of this group of animals.

Methods

The dissociation of embryonic tissue, as well as the Chromium 10x Genomics scRNA-seq library preparation were performed as in our previous studies (Feregrino *et al*, 2019).

Filtering, dimensionality reductions, visualization, clustering

Most of our analyses were performed using the toolkit Seurat v3.1.4 (Butler *et al*, 2018) using, otherwise stated, the default options. Each of the samples (e.g. Limb HH21, Somite HH12, etc.) were first processed individually. We followed our quality control approach (see Chapter 1) to filter the cells in each of the samples. The only difference is that we used a hard threshold for the fraction of UMIs with mitochondrial origin, stablished at 0.1. After quality filtering, the cell expression matrices were imported into Seurat, where the expression was normalized using the "LogNormalize" method and a scale factor of 10,000, the data was then scaled using the "ScaleData" function. Using the scaled data, cell cycle scores were calculated for each cell using SCRAN (Lun *et al*, 2016) and a list of gene pairs known to co-variate along the cell cycle (Scialdone *et al*, 2015). The difference between the S and the G2/M phases scores δ(S-G2M) was then calculated. We used the "SCTransform" function from Seurat to normalize, scale and transform our data. We used the δ(S-G2M), fraction of UMI of mitochondrial origin, and the UMI count per cell as variables to regress.

After this, we integrated the samples according to their embryological origin. To this end, we calculated highly variably genes from each individual sample. From the variability results provided by the "SCTransform" function, we chose those genes with a variability > median + MAD. We combined the highly variable genes and chose only the genes that are expressed in all three samples. We used this list of highly variable genes as the anchors to integrate our data using Seurat and the normalization method "SCT". The integrated data was once more scaled to regress the variability produced by the fraction of UMI of mitochondrial origin, and the UMI count per cell.

We then ran a PCA dimensionality reduction using the "RunPCA" function from Seurat, using all expressed genes across the three samples, except for the W-linked genes. We determined the amount of PCs to take into account as 19 for the limb and frontonasal samples, and 21 for the somites samples. This was done by the method previously described (see Chapter 1). For visualization we produced tSNEs and exaggerated tSNEs, following a method that retains global structure of the data (Kobak & Berens, 2019), with parameters previously discussed (see Chapter 3). The cell cluster calculation also is identical to the ones we have previously used (see Chapter 3)

Differential expression analyses were done using the implementation of MAST (Finak *et al*, 2015) in Seurat. We first calculated the standardized variance of each gene in the entire dataset by using the "FindVariableFeatures" function of Seurat, and selected as variable genes those with a standardized variance larger than the sample media. For making cluster to cluster comparisons we used scaled but "untransformed"

data, we used the δ(S-G2M) as a latent variable. We tested only those highly-variable genes expressed in at least 25% of the cells of either cell population. Only genes with an adjusted p-value < 0.05 and log2 fold change >0.5 are then taken into account as differentially expressed genes.

To annotate our clusters, we used our previous strategy with differentially expressed genes, expression patterns of known marker genes, spatial expression data repositories Geisha (Darnell *et al*, 2007) and MGI (Smith *et al*, 2019), as well as a literature review (see Chapter 3).

Data integration

The integration presented in Figure 2 was done using Scanorama (Hie *et al*, 2019). For this, first we use Seurat to normalize and "transform" the data for each of the individual samples, as in the embryonic region integration described above. We then used all of the genes in the count tables and used the "correct" function of Scanorama to obtain a dimensionality reduction. We don't use Scanorama to transform our data, but to obtain a joint hyper plane (JHP), which then we use instead of a PCA reduction for the calculation of tSNEs and exaggerated tSNEs. For the integrations on Figure 3, we followed a different strategy. We first subset our individual datasets based on the annotated clusters, as described in the main text. The data was then normalized and transformed individually. Then variable features were selected using the "SelectIntegrationFeatures" function from Seurat, requesting 3000 features. Using these genes we then used the integration method "SCT" from Seurat using 20 CCA dimensions and 100 k nearest neighbors. Using the same genes, we used the function "correct" from Scanorama to obtain a JHP reduction. Exaggerated tSNEs were then calculated using the relevant PCA or JHP dimensions.

Bioinformatics for xenograft experiment

For the single-cell analysis of the grafted tissue we used a combination of tools, using bash, python and R. In a first step, the chicken GRCg6a and quail Coturnix_japonica_2.0 genomes we obtained from ENSEMBL release 97 (Cunningham *et al*, 2019) and then concatenated. After, using awk in bash we generate a file listing each scaffold of the concatenated gene, and its full length. We then developed a python script that uses this file as input and generates a custom gtf file. The custom gtf file has all the columns and information that the chicken genome annotation gtf file has: chromosome / scaffold, source, feature type, start, end, score, strand, frame and record attributes. For each of the scaffolds, we created 6 records for the gtf file, a gene, exon and transcript for each strand sense. The chromosome name is obtained from the input file, the start is always 1, and the end is the length of the scaffold. We don't assign scores or frames, but add important attributes like a gene id with the name of the scaffold, and the biotype as "protein coding". The gtf file is then processed using the "mkref" function from CellRanger (10x Genomics).

In the next step, we used the custom gtf file and the concatenated genome file to map and count scRNA-seq reads using CellRanger. After this is completed, we load the UMI count matrices into R and calculate the proportions of UMIs mapped either

to the chicken or the quail portion of the genome. For the grafted cells analyses, we then filtered out all the cells for which the quail UMI counts divided by the chicken UMI counts resulted in more than 0.75. The grafted cells were processed further in the same way as the rest of our chicken samples.

To perform our differential correlation analyses we used averaged expression data. We first subset our limb and somite data into four different data sets in the way described in the main text. For each of the four data subsets we calculated the genes that are expressed in at least 3 cells, and compiled them in a unified gene set. From this combined expressed genes, we selected only those which are also expressed in at least 3 of the grafted cells. This final subset of genes was then used afterwards for the rest of the analysis. For each of the data subsets, we used the function "AverageExpression" from Seurat to obtain the average of the normalized "untransformed" expression per cell cluster. The cluster averages were then further averaged into a single number per gene for each sample subset. We randomly sampled 272 cells from each of the data subsets, meaning that together with the grafted cells data set we had in total 5 test single-cell datasets and 4 average reference datasets. We calculated the spearman correlation of each cell from the 5 test datasets to each of the 4 averaged reference data sets. Then, cell-wise, we subtracted the correlation coefficient to the averaged expression of the early limb from the correlation coefficient to the averaged expression of the late limb. This subtraction was repeated for the somite reference datasets correlations. We then plotted the differential correlations against each other.

References

Akiyama H, Chaboissier MC, Martin JF, Schedl A & De Crombrugghe B (2002) The transcription factor Sox9 has essential roles in successive steps of the chondrocyte differentiation pathway and is required for expression of Sox5 and Sox6. *Genes Dev.* **16:** 2813–2828

Akiyama H, Kim JE, Nakashima K, Balmes G, Iwai N, Deng JM, Zhang Z, Martin JF, Behringer RR, Nakamura T & De Crombrugghe B (2005) Osteo-chondroprogenitor cells are derived from Sox9 expressing precursors. *Proc. Natl. Acad. Sci. U. S. A.* **102:** 14665–14670

Alpert A, Moore LS, Dubovik T & Shen-Orr SS (2018) Alignment of single-cell trajectories to compare cellular expression dynamics. *Nat. Methods* **15:** 267–270

Arendt D, Musser JM, Baker CVH, Bergman A, Cepko C, Erwin DH, Pavlicev M, Schlosser G, Widder S, Laubichler MD & Wagner GP (2016) The origin and evolution of cell types. *Nat. Rev. Genet.* **17:** 744–757

ten Berge D, Brugmann SA, Helms JA & Nusse R (2008) Wnt and FGF signals interact to coordinate growth with cell fate specification during limb development. *Development* **135:** 3247–3257

Butler A, Hoffman P, Smibert P, Papalexi E & Satija R (2018) Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36:** 411–420

Couly GF, Coltey PM & Le Douarin NM (1993) The triple origin of skull in higher vertebrates: A study in quail-chick chimeras. *Development* **117:** 409–429

Cunningham F, Achuthan P, Akanni W, Allen J, Amode MR, Armean IM, Bennett R, Bhai J, Billis K, Boddu S, Cummins C, Davidson C, Dodiya KJ, Gall A, Girón CG, Gil L, Grego T, Haggerty L, Haskell E, Hourlier T, et al (2019) Ensembl 2019. *Nucleic Acids Res.* **47:** D745–D751

D'Souza RN, Ruest L-B, Hinton RJ & Svoboda KKH (2010) Development of the Craniofacial Complex. In *Bone and Development* pp 153–181.

Darnell DK, Kaur S, Stanislaw S, Davey S, Konieczka JH, Yatskievych TA & Antin PB (2007) GEISHA: An in situ hybridization gene expression resource for the chicken embryo. *Cytogenet. Genome Res.* **117:** 30–35

Feregrino C, Sacher F, Parnas O & Tschopp P (2019) A single-cell transcriptomic atlas of the developing chicken limb. *BMC Genomics* **20:**

Finak G, McDavid A, Yajima M, Deng J, Gersuk V, Shalek AK, Slichter CK, Miller HW, McElrath MJ, Prlic M, Linsley PS & Gottardo R (2015) MAST: A flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.* **16:**

Gerber T, Murawala P, Knapp D, Masselink W, Schuez M, Hermann S, Gac-Santel M, Nowoshilow S, Kageyama J, Khattak S, Currie JD, Camp JG, Tanaka EM & Treutlein B (2018) Single-cell analysis uncovers convergence of cell identities during axolotl limb regeneration. *Science (80-.).* **362:**

Gï¿½mez-Picos P & Eames BF (2015) On the evolutionary relationship between chondrocytes and osteoblasts. *Front. Genet.* **6:**

Gilbert SF. Developmental Biology. (2000) Osteogenesis: The Development of Bones. In *Development Biology: 6th edition* p http://www.ncbi.nlm.nih.gov/books/NBK10056/. Sunderland (MA): Sinauer Associates

Gong Y, Slee RB, Fukai N, Rawadi G, Roman-Roman S, Reginato AM, Wang H, Cundy T, Glorieux FH, Lev D, Zacharin M, Oexle K, Marcelino J, Suwairi W, Heeger S, Sabatakos G, Apte S, Adkins WN, Allgrove J, Arslan-Kirchner M, et al (2001) LDL receptor-related protein 5 (LRP5) affects bone accrual and eye development. *Cell* **107:** 513–523

Green RM, Fish JL, Young NM, Smith FJ, Roberts B, Dolan K, Choi I, Leach CL, Gordon P, Cheverud JM, Roseman CC, Williams TJ, Marcucio RS & Hallgrímsson B (2017) Developmental nonlinearity drives phenotypic robustness. *Nat. Commun.* **8:**

Guo X, Day TF, Jiang X, Garrett-Beal L, Topol L & Yang Y (2004) Wnt/β-catenin signaling is sufficient and necessary for synovial joint formation. *Genes Dev.* **18:** 2404–2417

Hamburger V & Hamilton HL (1951) A series of normal stages in the development of the chick embryo. *J. Morphol.* **88:** 49–92

Hartmann C & Tabin CJ (2001) Wnt-14 plays a pivotal role in inducing synovial joint formation in the developing appendicular skeleton. *Cell* **104:** 341–351

Healy C, Uwanogho D & Sharpe PT (1996) Expression of the chicken Sox9 gene marks the onset of cartilage differentiation. In *Annals of the New York Academy of Sciences* pp 261–262.

Hedges SB, Marin J, Suleski M, Paymer M & Kumar S (2015) Tree of life reveals clock-like speciation and diversification. *Mol. Biol. Evol.* **32:** 835–845

Hie B, Bryson B & Berger B (2019) Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. *Nat. Biotechnol.* **37:** 685–691

Hirasawa T & Kuratani S (2015) Evolution of the vertebrate skeleton: morphology, embryology, and development. *Zool. Lett.* **1:**

Ido A & Ito K (2006) Expression of chondrogenic potential of mouse trunk neural crest cells by FGF2 treatment. *Dev. Dyn.* **235:** 361–367

John N, Cinelli P, Wegner M & Sommer L (2011) Transforming growth factor β-mediated sox10 suppression controls mesenchymal progenitor generation in neural crest stem cells. *Stem Cells*

29: 689–699

Kanton S, Boyle MJ, He Z, Santel M, Weigert A, Sanchís-Calleja F, Guijarro P, Sidow L, Fleck JS, Han D, Qian Z, Heide M, Huttner WB, Khaitovich P, Pääbo S, Treutlein B & Camp JG (2019) Organoid single-cell genomic atlas uncovers human-specific features of brain development. *Nature* **574:** 418–422

Kawahara-Miki R, Sano S, Nunome M, Shimmura T, Kuwayama T, Takahashi S, Kawashima T, Matsuda Y, Yoshimura T & Kono T (2013) Next-generation sequencing reveals genomic features in the Japanese quail. *Genomics* **101:** 345–353

Kobak D & Berens P (2019) The art of using t-SNE for single-cell transcriptomics. *Nat. Commun.* **10:**

Kozhemyakina E, Lassar AB & Zelzer E (2015) A pathway to bone: Signaling molecules and transcription factors involved in chondrocyte development and maturation. *Dev.* **142:** 817–831

Kumar M, Ray P & Chapman SC (2012) Fibroblast growth factor and bone morphogenetic protein signaling are required for specifying prechondrogenic identity in neural crest-derived mesenchyme and initiating the chondrogenic program. *Dev. Dyn.* **241:** 1091–1103

Lange M, Klein M, Restrepo Lopez JL, Theis FJ & Pe'er D (2020) CellRank.

Lefebvre V & Bhattaram P (2010) Vertebrate skeletogenesis

Lewandoski M, Sun X & Martin GR (2000) Fgf8 signalling from the AER is essential for normal limb development. *Nat. Genet.* **26:** 460–463

Logan M, Martin JF, Nagy A, Lobe C, Olson EN & Tabin CJ (2002) Expression of Cre Recombinase in the developing mouse limb bud driven by a Prxl enhancer. *Genesis* **33:** 77–80

Long F, Chung U Il, Ohba S, McMahon J, Kronenberg HM & McMahon AP (2004) Ihh signaling is directly required for the osteoblast lineage in the endochondral skeleton. *Development* **131:** 1309–1318

Luecken MD, Buttner M, Chaichoompu K, Danese A, Interlandi M, Mueller MF, Strobl DC, Zappia L, Dugas M, Colome-Tatche M & Theis FJ (2020) Benchmarking atlas-level data integration in single-cell genomics. *bioRxiv*

Lun ATL, McCarthy DJ & Marioni JC (2016) A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. *F1000Research* **5:**

Marioni JC & Arendt D (2017) How single-cell genomics is changing evolutionary and developmental biology. *Annu. Rev. Cell Dev. Biol.* **33:** 537–553

Masckauchán TNH, Shawber CJ, Funahashi Y, Li CM & Kitajewski J (2005) Wnt/β-catenin signaling induces proliferation, survival and interleukin-8 in human endothelial cells. *Angiogenesis* **8:** 43–51

McGonnell IM & Graham A (2002) Trunk neural crest has skeletogenic potential. *Curr. Biol.* **12:** 767–771

Murtaugh LC, Chyung JH & Lassar AB (1999) Sonic hedgehog promotes somitic chondrogenesis by altering the cellular response to BMP signaling. *Genes Dev.* **13:** 225–237

Nakamura H & Ayer-le Lievre CS (1982) Mesectodermal capabilities of the trunk neural crest of birds. *J. Embryol. Exp. Morphol.* **70:** 1–18

Piekarski N, Gross JB & Hanken J (2014) Evolutionary innovation and conservation in the embryonic derivation of the vertebrate skull. *Nat. Commun.* **5:**

Prummel KD, Nieuwenhuize S & Mosimann C (2020) The lateral plate mesoderm. *Dev.* **147:**

Rawadi G, Vayssière B, Dunn F, Baron R & Roman-Roman S (2003) BMP-2 Controls Alkaline

Phosphatase Expression and Osteoblast Mineralization by a Wnt Autocrine Loop. *J. Bone Miner. Res.* **18:** 1842–1853

Sarkar S, Petiot A, Copp A, Ferretti P & Thorogood P (2001) FGF2 promotes skeletogenic differentiation of cranial neural crest cells. *Development* **128:** 2143–2152

Schafer AJ, Foster JW, Kwok C, Weller P, Guioli S & Goodfellow PN (1996) Campomelic dysplasia with XY sex reversal: Diverse phenotypes resulting from mutations in a single gene. In *Annals of the New York Academy of Sciences* pp 137–149.

Scialdone A, Natarajan KN, Saraiva LR, Proserpio V, Teichmann SA, Stegle O, Marioni JC & Buettner F (2015) Computational assignment of cell-cycle stage from single-cell transcriptome data. *Methods* **85:** 54–61

Shi Y, Sun L, Wang M, Liu J, Zhong S, Li R, Li P, Guo L, Fang A, Chen R, Ge WP, Wu Q & Wang X (2020) Vascularized human cortical organoids (vOrganoids) model cortical development in vivo. *PLoS Biol.* **18:**

Shimoyama A, Wada M, Ikeda F, Hata K, Matsubara T, Nifuji A, Noda M, Amano K, Yamaguchi A, Nishimura R & Yoneda T (2007) Ihh/Gli2 signaling promotes osteoblast differentiation by regulating Runx2 expression and function. *Mol. Biol. Cell* **18:** 2411–2418

Shirai Y, Kawabe K, Tosa I, Tsukamoto S, Yamada D & Takarada T (2019) Runx2 function in cells of neural crest origin during intramembranous ossification. *Biochem. Biophys. Res. Commun.* **509:** 1028–1033

Shum L, Coleman CM, Hatakeyama Y & Tuan RS (2003) Morphogenesis and dysmorphogenesis of the appendicular skeleton. *Birth Defects Res. Part C - Embryo Today Rev.* **69:** 102–122

Smith CM, Hayamizu TF, Finger JH, Bello SM, McCright IJ, Xu J, Baldarelli RM, Beal JS, Campbell J, Corbani LE, Frost PJ, Lewis JR, Giannatto SC, Miers D, Shaw DR, Kadin JA, Richardson JE, Smith CL & Ringwald M (2019) The mouse Gene Expression Database (GXD): 2019 update. *Nucleic Acids Res.* **47:** D774–D779

Wachtler F, Christ B & Jacob HJ (1982) Grafting experiments on determination and migratory behaviour of presomitic, somitic and somatopleural cells in avian embryos. *Anat. Embryol. (Berl).* **164:** 369–378

Wagner GP & Chiu CH (2001) The Tetrapod Limb: A Hypothesis on Its Origin. *J. Exp. Zool.* **291:** 226–240

Wright E, Hargrave MR, Christiansen J, Cooper L, Kun J, Evans T, Gangadharan U, Greenfield A & Koopman P (1995) The Sry-related gene Sox9 is expressed during chondrogenesis in mouse embryos. *Nat. Genet.* **9:** 15–20

Zeng L, Kempf H, Murtaugh LC, Sato ME & Lassar AB (2002) Shh establishes an Nkx3.2/Sox9 autoregulatory loop that is maintained by BMP signals to induce somitic chondrogenesis. *Genes Dev.* **16:** 1990–2005

Zhou S, Eid K & Glowacki J (2004) Cooperation between TGF-β and Wnt pathways during chondrocyte and adipocyte differentiation of human marrow stromal cells. *J. Bone Miner. Res.* **19:** 463–470

Chapter 5

CROSS-SPECIES COMPARISON OF CELL TYPE SPECIFIC GENE CO-EXPRESSION MODULES

Christian Feregrino

Abstract

The advent of single-cell RNA-seq has provided the opportunity to generate gene expression data with cellular resolution. Moreover, the number of species from which this data is being produced is constantly increasing. Therefore, the development of computational methods and approaches to compare single-cell transcriptomic data across species is highly relevant today. Here, we present an approach, to circumvent some of the challenges and difficulties of gene expression data comparisons across species. By adapting Weighted Gene Co-expression Network Analysis (WGCNA), an approach developed for bulk RNA-seq analyses, we propose to compare single cell gene co-expression modules instead of comparing gene expression data directly. The workflow we present here results, first, in a complete WGCNA analysis with the detection of co-expression modules, their expression along the single-cell dataset and their functional enrichments. The second and main result of our approach is the statistical scoring of overall module conservation across species boundaries, as well as the decomposition of this conservation into two main components: conservation of module density and module connectivity. Our workflow is highly customizable and adaptable to different experimental designs. It can be used to make cross-species comparisons, but also for comparisons across developmental time or experimental conditions.

Introduction

Homologous cell types and tissues in different species, even in very distantly related ones, are more similar to each other than to other cells within the same organism. For example, hepatocytes from a zebrafish are more similar to the hepatocytes of chicken, than to neuronal cells of the same zebrafish. The similarity between cells of the same cell type is recognizable in terms of cell structure, function and transcriptomic programs (Brawand *et al*, 2011). Although all cells from an organism share [almost] exactly the same genome (Frumkin *et al*, 2005), they can display very distinct phenotypes. Cells are different from each other because they have particular epigenetic states and developmental potentials, which are acquired through developmental differentiation that progressively restricts their full access to the genome (Marioni & Arendt, 2017). In this sense, cell types can be molecularly discriminated from one another based on specific interactions of transcription factor sets, called “core regulatory complexes”, which activate particular genomic features and drive the expression of different transcriptomic programs (Arendt *et al*, 2016). These characteristics can be used not only to delineate cell types within an organism, but also to identify homologous cell types, or cell type families, shared across species. Thus, comparative studies of the transcriptome – reflective of a cell’s underlying regulatory state – can be used to find “signatures” or “core transcriptional programs”, indicative of cell types or tissues (Shapiro *et al*, 2013; Trapnell *et al*, 2014; Cardoso-Moreira *et al*, 2019a). Such studies have the potential to yield insights into novel gene roles, evolutionary and developmental history, as well as medically relevant cellular functions.

Finding shared transcriptomic programs and their underlying regulatory input is, nonetheless, a challenging process. A very elaborate and expensive experimental design would be needed to perform classical differential expression analyses between different tissues across different species. Moreover, classical differential expression analyses are challenging when performed across different experimental and sequencing platforms; and this incompatibility only increases when dealing with multiple species (Zhou *et al*, 2019). Different methodologies have been developed to successfully analyze bulk RNA-seq data across experimental methods and species (Brawand *et al*, 2011; Cardoso-Moreira *et al*, 2019a; Lin *et al*, 2014; Kouzarides, 2007; Sudmant *et al*, 2015). However, with the advent of single-cell RNA-seq and the inherent statistical challenges that the resulting data poses (Shafer, 2019; Bacher & Kendziorski, 2016; Stuart & Satija, 2019), different approaches have to be developed, or adapted, to make cross-species comparisons of thousands of transcriptomes at a time. Cross species comparisons of scRNA-seq cell populations can be performed on separately analyzed and annotated datasets, or in datasets that are integrated in a single analysis and annotation. While the former type of analysis retains intra-dataset heterogeneity, batch effects can negatively affect comparisons. Integrative analyses have increased statistical power to detect variation, but the data corrections necessary might hide species-specific characteristics (Shafer, 2019).

The development of approaches to perform scRNA-seq analyses integrating different datasets is an active field of research, and a plethora of methods with different strengths and weaknesses have been presented in the last couple of years (Luecken *et al*, 2020). Moreover, many of these approaches can be implemented to find common cell populations in multi-species studies. However, the attempts to compare independently analyzed an annotated scRNA-seq cell populations across species have been less systematic. Among the methods presently employed we find correlation based on a so-called "gene-specificity index" (Tosches *et al*, 2018), ward linkage and correlation of shared variable genes expression (Zilionis *et al*, 2019), and neighbor voting algorithms based on correlation networks (Han *et al*, 2020; Crow *et al*, 2018). Moreover, pseudotime warping methods are useful to compare transcriptomic dynamics across species, during differentiation processes (Alpert *et al*, 2018; Kanton *et al*, 2019).

To further expand such cross-species comparisons of cell populations, here we propose the use of Weighted Gene Correlation Network Analysis (WGCNA) (Langfelder & Horvath, 2008), a bioinformatics tool developed for the analysis of bulk RNA-seq and microarray assays. Our research group has successfully employed WGCNA to analyze scRNA-seq datasets, and find signatures of cell-type-specific expression patterns across different samples of single-cell RNA-seq (Feregrino *et al*, 2019). The work presented here, however, is not the first time that WGCNA has been used in the setting of cross-species comparisons. At the bulk RNA-seq level, it has been applied in plants, to find expression programs conserved between maize and rice (Ficklin & Feltus, 2011). In animals, a WGCNA analysis was performed on human data and based on the gene sets they found, they analyzed the expression differences in mammalian model species and chicken (Cardoso-Moreira *et al*, 2019b). Notably, WGCNA has also been used for the cross-species comparison of telencephalon scRNA-seq (Tosches *et al*, 2018), their comparison consisted in testing the gene overlap between co-expression modules.

WGCNA is based on the calculation of adjacency networks of co-expression, where each node represents a gene, and each edge the adjacency of two genes based on their co-expression coefficient. An important step in a WGCNA analysis is the detection of co-expression modules, or clusters of densely interconnected genes according to their topological overlap within the network. Topological overlap of two genes refer to how similar are their connections with the rest of the genes (Langfelder & Horvath, 2008). When working with scRNA-seq data, these modules of co-expression are calculated with no prior information about the clustering of the cells, which makes this process unsupervised. After modules have been detected, several comparisons can be made. Modules can be compared in their identity (the genes that compose them), density (average connection strength among all nodes), or connectivity (the pattern of the connectivity strengths within the module) (Langfelder *et al*, 2011).

Here, we present the adaptation and use of WGCNA and its related module conservation tests, to make cross-species comparisons of scRNA-seq datasets. With this approach, we are able to assess the conservation of gene co-expression modules that represent putative core transcriptomic programs of different cell populations. It's important to note that using this methods we don't directly compare gene expression, but rather measurements of co-expression (Langfelder *et al*, 2011). By doing so, we circumvent some of the main challenges of scRNA-seq data integration, like correction of technical batch effects or overfitting of the data. The methodology we use starts with independently analyzed and annotated datasets, and follows this logic:

1. Construction of pseudocells, cluster-wise for each data set.
2. Calculation of co-expression modules in one of the samples, as a reference.
3. Calculation of weighted gene co-expression coefficients and construction of co-expression networks in the test samples.
4. Comparison of density and connectivity of the reference co-expression modules across all samples, or species, to assess module conservation.

This approach also allows to use and integrate the vast and ever-growing collection of single-cell transcriptomic data that the scientific community produces. To show the functionality of our methodology, we make use of a published data set and compare it to our own produced data. Specifically, we compare a published dataset of the developing neural tube of the mouse embryo (Delile *et al*, 2019), and our own dataset obtained from the developing neural tube of the chicken. We demonstrate, overall, that the methodology is applicable for single-cell RNA-seq data, and that it produces satisfactory results comparing across samples of different species, with data stemming from completely independent experiments.

Results

While developing and testing our approach, we considered diverse module comparison methods. A comparison approach mentioned before (Tosches *et al*, 2018), consists in detecting modules of co-expression in each sample, and then test the identity of each module using a hypergeometric test. However, we have found that the detection of modules is not always reproducible. Therefore, we opted to compare the network information of defined modules, rather than detecting modules and compare their identity. Testing for the conservation of defined modules was introduced by the developing team of WGCNA (Langfelder *et al*, 2011). In this approach, modules are detected from within the co-expression network of one sample, and then the conservation of these modules is tested in the co-expression networks of other samples. The tests take advantage of the gene co-expression information contained within the network, and not just the identity of the genes composing the modules. The conservation index resulting from a conservation test is calculated using a combination of statistical tests that compare several network metrics, like connectivity and density. We then set out to test the capabilities of our chosen approach.

We used our modified scRNA-seq WGCNA approach to compare the transcriptomes of three neural tube datasets, one as a reference and two as test samples. The reference data set is from a mouse spinal cord study, where the authors sequenced the transcriptomes of single cells from the cervical and thoracic sections at mouse embryonic stages E9.5, E10.5, E11.5, E12.5 and E13.5 (Delile *et al*, 2019). From these datasets, we only used the neuronal cells from the cervical section at stage E13.5 (Figure 1 A). We took these cells, since they are the closest ones in space and time to our brachial sample. The two test datasets were produced in our laboratory, and stem from the embryonic chicken neural tube at stage HH36 (day 10 of development) (Hamburger & Hamilton, 1951). Tissue was dissected from the spinal cord at the brachial and lumbar sections, where the motor neurons innervate the limbs (Figure 1 D). Each location was sampled with two technical replicates of single-cell 10x Genomics Chromium assays.

As a preliminary step to our WGCNA analysis, we processed all datasets using a standardized workflow (see Chapter 1), in order to have distinct cell clusters and dimensionality reduction visualizations. The mouse data was processed only for visualization purposes, as the cell clustering had already been defined. We normalized the count data and corrected for effects of the percentage of mitochondrial UMIs and cell library size. We ran a PCA and used the first 18 principal components to calculate an exaggerated tSNE and then plotted the cell annotation provided by the authors on the resulting embeddings. We observed that our exaggerated tSNE in general recapitulated their findings, by separating the neurons and progenitors (Figure 1 B), but also reproducing the neurogenesis process they describe (Figure 1 B insert). We also observed the separation of the different neuron classes they defined, as well as clustering of inhibitory GABAergic neurons (d4, V1 and V2) and excitatory neurons (Figure 1 C). The chicken data was analyzed and clustered following our full single-cell transcriptome workflow (see Chapter 1). After verifying the expression of genes used to mark the mouse cell populations – like SOX2, TUBB3, MEOX1 and SOX10 (Delile *et al*, 2019) –, we confirmed that our chicken samples contained comparable cell populations, *i. e.* neuronal progenitors and mature neurons. Our samples also contained oligodendrocytes and other smaller populations. We did not further annotate this data, as this was not in the scope of our WGCNA test. From the chicken samples, nonetheless, we found 30 cell clusters for the brachial sample (Figure 1 E) and 28 clusters for the lumbar sample (Figure 1 F).

As a first step in our comparative WGCNA workflow, we generated pseudocells from each of the datasets, in order to reduce noise and have more robust data for the subsequent calculations. Pseudocell construction was based on previous strategies (Kanton *et al*, 2019). First, we calculate the 10 nearest neighbors of each cell in the PCA space. We then randomly pick a fifth of the cells from each previously calculated cluster and use them as seeds to which their nearest neighbors are aggregated. The average expression is then calculated from each of the aggregates and used for further analyses.

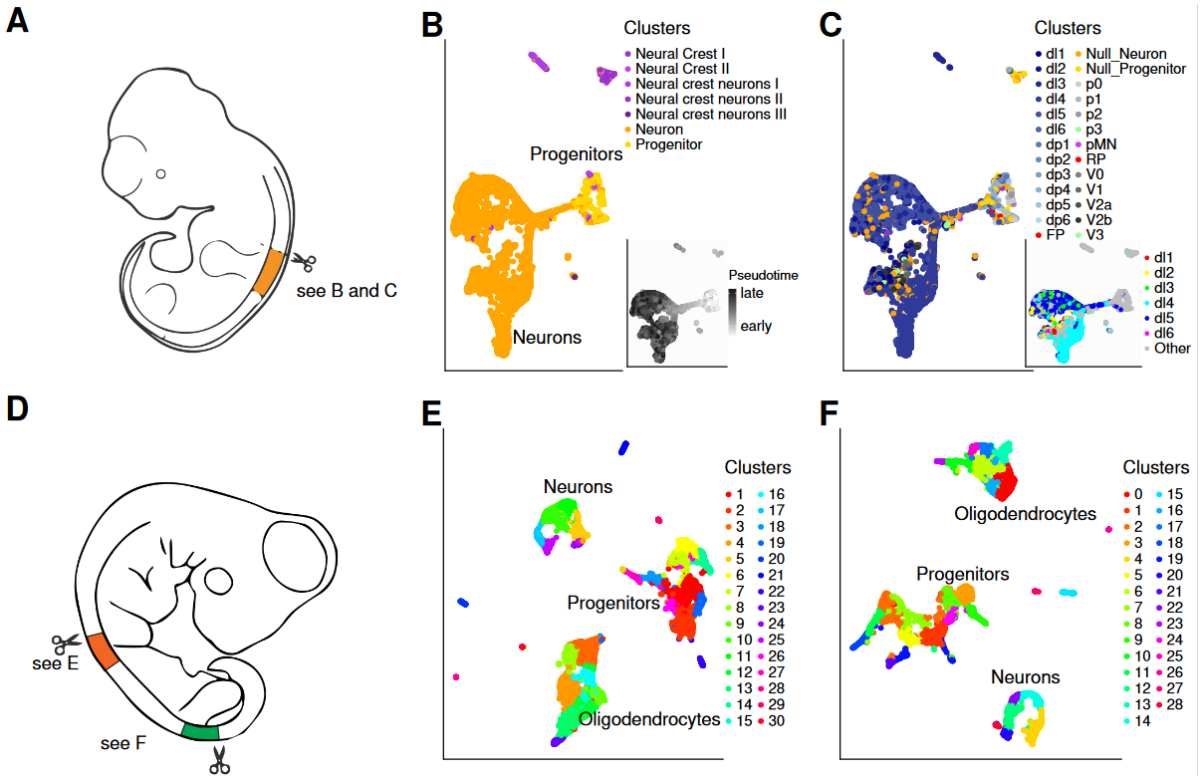


Figure 1 Mouse and chicken neural tube single-cell data. **A** Dissection strategy of the mouse neural tube data. Adapted from: Delile *et al*. 2019. **B** Neuronal cells from stage 13.5 of the mouse neural tube data. Exaggerated tSNE embedding showing the main clusters defined by the authors, colored similarly to the original publication Fig. 1 D. Insert: Showing the pseudotime as defined in the original publication. **C** Same embeddings as B, excluding the neural crest cells. Showing the clustering corresponding to the neural tube sections, colored following a similar color scheme to the original publication Fig. 1 C and D. Insert: Showing a different color scheme of the neural tube mature neurons, for better appreciation. **D** Dissection strategy for the sampling of the chicken neural tube data. **E** Exaggerated tSNE and Louvain-Jaccard clustering of the brachial chicken data. Color scheme independent. **F** Exaggerated tSNE and cell clustering of the lumbar chicken data. Color scheme independent.

The next step of our workflow is the WGCNA analysis *per se*, and is performed in by an in-home developed R script. The script takes input data contained in a Seurat objet (Stuart *et al*, 2019). In case pseudocells are being used, the original single cell data should also be provided. The script also uses several variables as options: for example, if only a subset of the data is to be analyzed, or which species the data belongs to. Then, the variable genes are calculated, in case these are not already provided by the user. The set of genes used for downstream analysis is critical, since it directly affects the modules that potentially can be detected (Langfelder & Horvath, 2008, 2014).In the next stage a soft-thresholding power is chosen, a process inherent to WGCNA which reduces the noise and makes the network resemble a scale-free topology. In a network with scale-free topology, its underlying structure and characteristics are independent of changes in the network size ("the probability $p(k)$ of having a vertex of degree k is of the form $p(k) \propto k^{-\gamma}$, where γ is referred to as the scaling parameter" (Payne & Eppstein, 2009)). This stage is also a control stage: if a scale-free topology index is not reached, the set of genes or cells used could be reconsidered. The next stage constitutes the main WGCNA analysis, where the topological overlap and distance matrices are calculated, and module detection is

performed. In short: genes are assigned to discrete modules of co-expression, according to their topological distances, and then, the membership of the genes to their modules is tested. Genes without significant membership are discarded and the process is repeated until all genes pass the membership test. The last stage of our script performs data processing. The mean expression of the modules is calculated in the single-cell space and plotted in the chosen dimensionality reduction. Moreover, a graph representation of the module is generated, and GO term enrichment analysis is performed. The results are presented in a report (Supplement 2).

We first performed this WGCNA co-expression module detection in the mouse data. Using the 2052 top variable genes as input, we obtained 1065 genes distributed in 12 different modules of co-expression (Figure 2 A). The modules varied in size from 34 to 211 genes, and showed distinct Gene Ontology term enrichments (GOE). The expression pattern of each module was also distinct, we calculated the expression of a module as the average expression of all the genes which are part of it. The different modules are named after different colors according to WGCNA (Figure 2 B).

Modules yellow (GOE: cell cycle, 92 genes), green (GOE: morphogenesis, nervous system development, 81 genes) and magenta (GOE: epithelium and neuro-differentiation, 56 genes) have higher expression in the neuronal progenitors, reflective of the early stage in the neuronal maturation process. The expression of modules red (GOE: translation, 75 genes), tan (GOE: development, transcription 34 genes) and greenyellow (GOE: patterning, 37 genes) seem to be ordered following the maturation process. Module blue (GOE: neurodevelopment, 202 genes) is expressed in mature neurons and module black (GOE: neurodevelopment, 59) shows high expression at the very end of the process. Module pink (GOE: trans-synapsis, 57) has expression in the mature inhibitory neurons, and not in in excitatory neurons (mainly clusters dl3 and dl5 (Delile *et al*, 2019)). The rest of the modules mark isolated cell clusters, *e. g.* brown (GOE: immune response, 112 genes) with high expression in a few cells, purple (GOE: neurodevelopment, 49 genes) and turquoise (GOE: proliferation, 211 genes) in the neural crest cells (Figure 2 B).

Once the WGCNA results of the reference data are obtained, the next step in our workflow are the comparative tests across samples. This is also carried out by a dedicated R script we have developed. The input of this script is a list of 1-to-1 orthologous genes, the expression data of the test datasets and the results from the reference WGCNA analysis. The conservation tests depend on gene expression variance, for which the genes need to be expressed in all samples. Therefore, the script, in a first stage, tests for presence of the module genes in the 1-to-1 orthologous gene list, and for their expression across all datasets. The comparison is then done using only these genes in all samples. The second stage is the so-called “preservation test”, done using WGCNA tools. The rest of this script is for data processing and plotting. As a result of this analysis, we have measurements of overall module quality in the reference relative to the test samples, overall module preservation, as well as preservation of density and connectivity of each module.

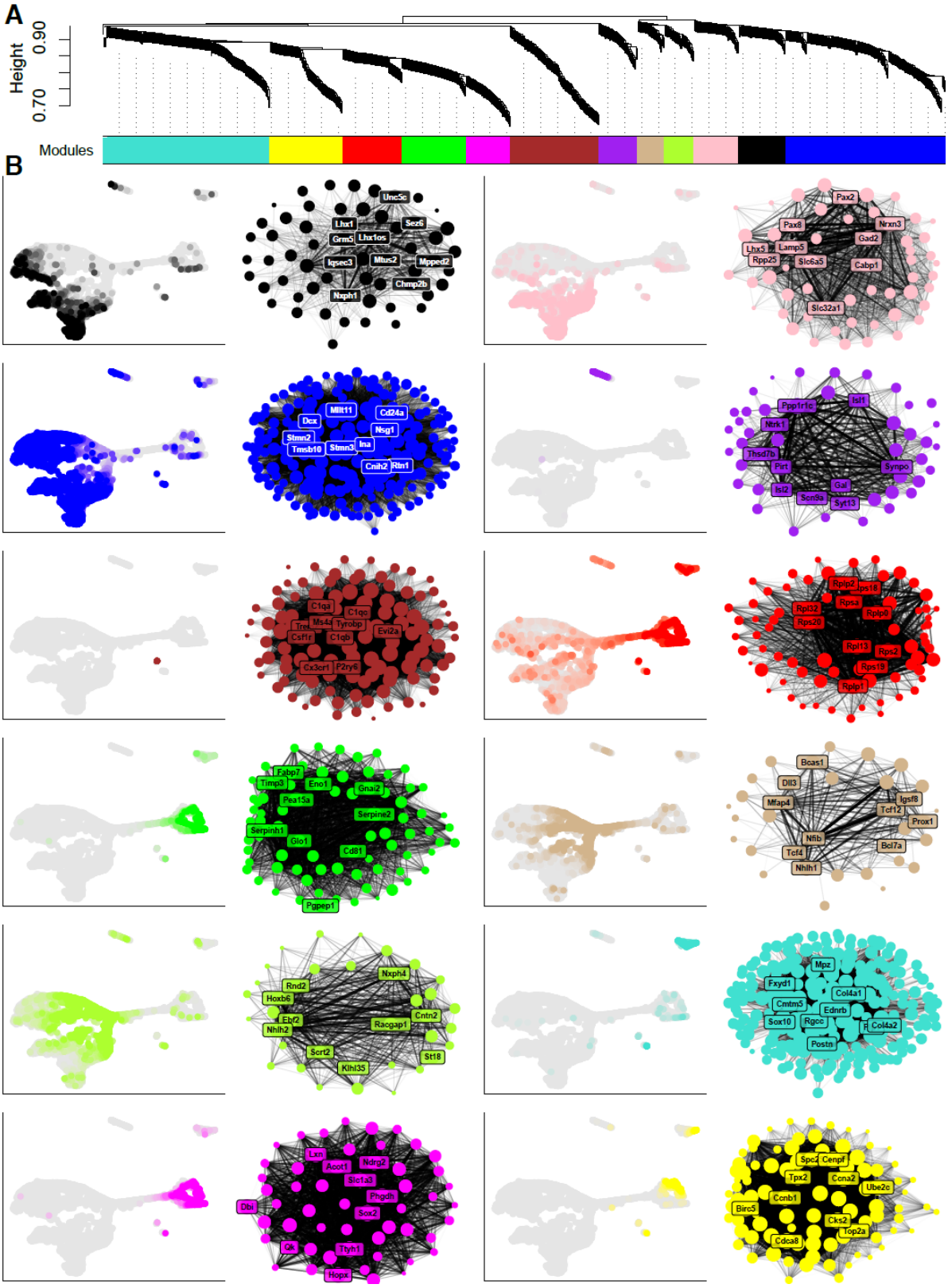


Figure 2 scRNA-seq WGCNA analysis of mouse E13.5 cervical neural tube. **A** Hierarchical clustering genes based on topological overlap, modules underneath. **B** Modules of co-expression shown as paired panels. tSNEs on the left show mean expression of the module. Color intensity corresponds to expression levels, values scaled per panel. Force-directed networks on the right are module representations. Nodes represent genes, size corresponds to membership of the gene, scaled per panel. Edges represent co-expression as topological distance, thickness and intensity represent co-expression relationship. Scaled per panel.

We applied this conservation analysis with the mouse WCGNA analysis results as a reference and our chicken samples as test sets. Our first observation was that the fraction of genes present in the 1-to-1 orthologues list and expressed in the test samples was different from module to module. Genes removed due to non-expression amount to only 6 across all modules. Modules black, blue, green, greenyellow, magenta and pink, all part of the neuronal maturation processes, had around 80% of their genes as 1-to-1 orthologous present in the chicken genome. On the other hand, modules purple, red, tan, turquoise and yellow contained around 70% of 1-to-1 orthologous genes. Notably, module brown is composed of only 50% 1-to1 orthologous genes (Figure 3 A), which can be explained by the immune nature of this co-expression module, and immune responses and the implicated genes are known to be fast-evolving (Lazzaro & Clark, 2013).

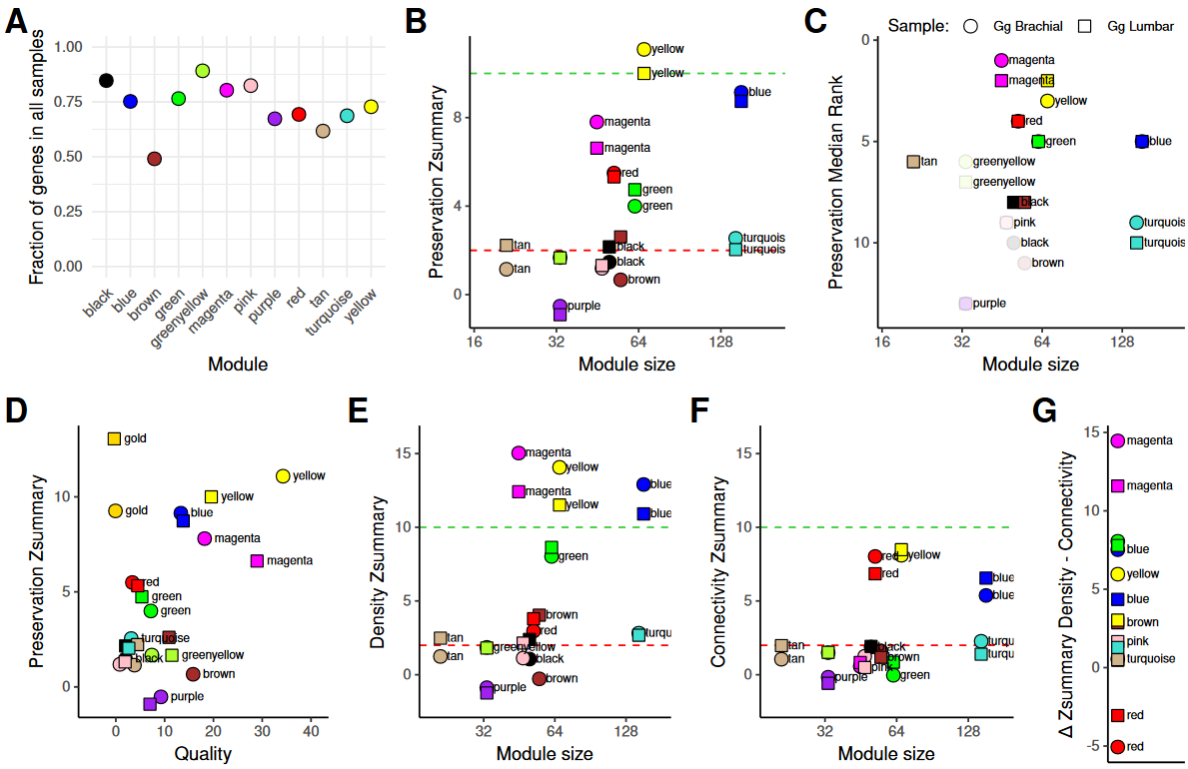


Figure 3 Conservation of mouse neural tube co-expression modules in the neural tube chicken samples. **A** Fraction of module genes present as 1-to-1 orthologues and expressed in both test samples. Color corresponding to the module name. **B** Preservation Zsummary scores of each module. The shapes correspond to the test sample, signaled in C and valid for the rest of the panels. **C** Median Rank of the preservation statistics across the modules. **D** Relation of preservation score and the quality of each module relative to the test samples. **E** Scores of density preservation across the co-expression modules. **F** Scores of connectivity preservation across the co-expression modules. **G** Difference of Density and Connectivity preservation scores. Only those genes with a score > 2 in any of the two metrics are plotted.

The conservation test of WGCNA showed that our modules have indeed different levels of conservation across our test samples. We mainly used two summary statistics of preservation, the Zsummary and the Median Rank. The Zsummary statistic is the combination of several density and connectivity tests, into a single index, which can be then formulated as a threshold. According to the developers of

WGCNA, thresholds of 2 and 10 is advised, where >10 means strong evidence of module preservation and <2 means no evidence of preservation (Langfelder *et al*, 2011). We observed that module yellow (GOE: cell cycle), highly expressed in neuronal progenitors, shows strong evidence for preservation in the brachial and lumbar samples of the chicken neural tube. On the other hand, modules tan (GOE: development, transcription – highly expressed in early maturation), greenyellow (GOE: patterning – high expression in early maturation), purple (GOE: neurodevelopment – expressed in the neural crest), pink (GOE: trans-synapsis – expressed in inhibitory GABAergic and glycinergic interneurons), black (GOE: neurodevelopment – expressed in late maturation), brown (GOE: immune response – expression in a few, putative immune cells), and turquoise (GOE: proliferation –high expression in neural crest and progenitors), all show little or no evidence of preservation in our chicken samples (Figure 3 B).

While the Zsummary index can be compared within modules, across test samples, it shouldn't be compared across modules, since it generally shows correlation with the module size (Langfelder *et al*, 2011). For preservation comparison across modules, a median ranking of the preservation statistics is provided, in which the first rankings mean higher compared preservation. Here, we observed that module magenta (GOE: epithelium and neuro-differentiation), expressed in the progenitors, and yellow (GOE: cell cycle), also highly expressed in neuronal progenitors, are the most conserved of the modules in both chicken samples. They are followed by modules red (GOE: translation), expressed in progenitors and the early maturation process, blue (GOE: neurodevelopment), expressed in mature neurons, and green (GOE: morphogenesis, nervous system development), expressed in progenitors. The modules with little or no evidence of preservation are also the worst ranked (Figure 3 C). The conservation scores depend on many factors, and one that can be evaluated during this analysis is the quality of the module in the reference data. This quality score also reflects several statistical tests of how well-defined the module is in the reference set, relative to the test datasets. Our analysis shows that in our case the quality of the modules is correlated with their Zsummary preservation score (Figure 3 D). Here, a module consisting of a random sample of genes from all the modules, named gold module, is added for reference. In our case, as expected, it shows the worst quality.

The last result from our analysis decomposes the preservation scores into two main components: the density and connectivity. Here, we observed that overall, the modules have a higher conservation of density (Figure 3 E) than of connectivity (Figure 3 F). As previously mentioned, density refers to the average of all interactions between the genes and connectivity refers to the patterns of the interactions and their strength. Notably, we found three modules showing big differences between the two scores (Figure 3 G). Module magenta (GOE: epithelium and neuro-differentiation), expressed in the progenitors, shows the biggest discrepancy, with strong evidence of density preservation, and no evidence of connectivity preservation (Figure 3 G). Module green (GOE: morphogenesis, nervous system development) expressed in

progenitors shows the same difference pattern, albeit not as strong. On the other hand, module red (GOE: translation) expressed in progenitors and the early maturation process shows the opposite trend, weak evidence of density preservation, and stronger evidence of connectivity preservation (Figure 3 G).

As examples of the different comparative insights our approach produces, we present two modules in detail. Module purple (GOE: neurodevelopment), expressed in the neural crest, shows no evidence of overall preservation, and no evidence of preservation of density or connectivity. An inspection of this module in all samples shows that indeed density and connectivity are visually very different between mouse and chicken, but similar within chicken samples. Moreover, the expression patterns is also not conserved (Figure 4 A). On the other hand, module red (GOE: translation), expressed in progenitors and the early maturation process, is among the most conserved modules, and shows different density and connectivity conservation scores. Module representations across the samples show that connection patterns between the genes are visually very similar across mouse and chicken samples. The expression patters is also similar, as they show higher expression in the neuronal progenitor clusters as in the mature neurons (Figure 4 B).

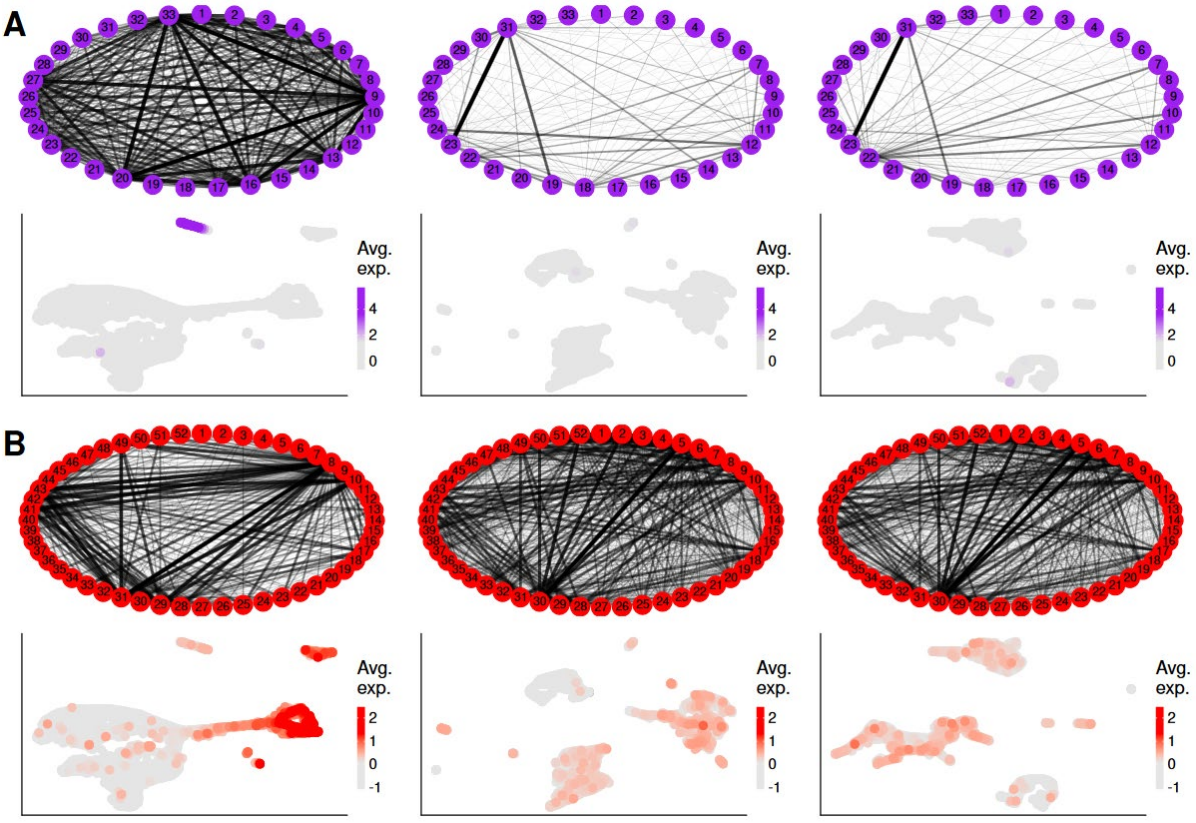


Figure 4 Comparison of module connectivity and their average expression. First column corresponds to mouse data, second column to chicken brachial data, and the third column to chicken lumbar data. **A** Module purple (GOE: neurodevelopment) expressed in the neural crest. **B** Module red (GOE: translation) expressed in progenitors and the early maturation process. Each node in the module representations corresponds to one gene, number and position corresponds to 1-to-1 orthologues. Size of the nodes is not scaled. The edges represent the co-expression levels between two genes. Thickness and color intensity are scaled for each panel.

In total, our workflow consists of 3 main steps, for which we have developed dedicated scripts. In summary, the first step produces pseudocells in a cluster-aware fashion from previously analyzed data. The second step performs the iterative WGCNA analysis in one sample, to detect modules of gene co-expression. These results are then used as the reference for the comparative analysis. Then, in the final step our script carries out the actual conservation test using the WGCNA analysis as a reference and pseudocells of the test samples. All scripts are designed using the R notebook format, to produce html reports complete with tables and figures that can be easily shared. The scripts can easily be modified and adapted for different experimental setups.

Discussion

Here, we present an integrative analysis to perform a comparative gene co-expression analysis. We showed its functionality by testing it with neural tube samples coming from mouse and chicken embryos. The chicken samples were produced as part of a project aiming to study the motor neurons of the brachial and lumbar levels at developmental stage HH36. Given the lack of a precisely corresponding sample of mouse origin, we decided to use the most similar available sample. The first high-throughput single-cell study of the neural tube development in mouse spanned from stages e9.5 to e13.5 (Delile *et al*, 2019), which should roughly correspond to chicken developmental stages HH10 to HH28 (Hill, 2020). We naturally chose to use the latest of these stages, to make our comparison. This means that we are comparing chicken samples 4.5 days – or 8 morphological stages – older than what the actual embryonic correspondence would suggest (Hill, 2020), and so, accordingly, our results should be interpreted with caution.

Our initial processing of the mouse E13.5 neural tube data resulted in a dimensionality reduction visualization which meaningfully represents the different cell populations, as well as the neuronal maturation process, as defined and described in the original publication. The initial processing of the chicken samples resulted in 30 and 28 different clusters, which we didn't further characterize, since this is not within the scope of the comparison we intended to make here. Since we are not computing modules of co-expression based on the chicken data, the cell composition of the test data is not critical. Indeed, if the co-expression modules are truly conserved across samples, we expect them to be present in the test datasets regardless of the presence or absence of unrelated cell populations. We acknowledge, nonetheless, that the correlation of the genes, and therefore the rest of the calculations depend on the cell composition of the sample. We therefore advice for an informed selection of the cell populations that will constitute both the reference and test datasets, in order to obtain more meaningful results in the future.

The WGNA analysis of the mouse neural tube data revealed 12 different modules of gene co-expression. Although some modules show high expression in the same cell populations, or part of the neurogenesis process, they are defined as distinct co-expression modules, reflecting the different process that co-occur in the same cells

and developmental state. We found the maturation process to be reflected by the succession of several co-expression modules. For example, in the progenitor populations and early neurogenesis process we find 4 distinct co-expression modules with different functional enrichments. These 4 co-expression modules are also similar to each other, as they appear next to each other on the hierarchical clustering of the genes. Module green and magenta, with GO terms enrichment like nervous system development are the closest of these modules. They also show very similar expression patterns, when their average expression is plotted in our dimensionality reduction. Module red shows extended expression into the maturation process, with GO term enrichment of translation and metabolic processes, pointing to a change in expression pattern and cellular functions. The next stage of maturation is characterized by the expression of modules greenyellow and tan, with different functional signals reflected in their GO term enrichments. Greenyellow shows evidence of patterning processes, while tan is enriched for terms related to transcription. By the end of the neurogenesis process, three modules show high expression: pink, blue and black. Blue and black are enriched for GO terms like neuron development, maturation and differentiation, including axon projection. Meanwhile, the pink module shows enrichment for synaptic signaling.

While some of the co-expression modules are closely related, and show similar expression patterns, the conservation scores of these modules are not always alike. All four modules present in the early neurogenesis show high conservation scores, and are the best ranked in terms of conservation. Moreover, magenta and yellow have the highest ranks. The conservation of yellow can be somewhat expected, as a co-expression module representing the cell cycle process should be conserved even among distantly related phyla. Module magenta is composed of canonical marker genes of neural tube progenitors, like SOX9 (Stolt *et al*, 2003), SOX2 (Graham *et al*, 2003) and PAX6 (Ericson *et al*, 1997), and shows high expression in the populations we have also characterized as progenitors in the chicken samples. This shows the ability to recognize a transcriptional signature program that characterizes the progenitor populations across species. The only module with high expression in the late neurogenesis process and high conservation score is blue. This module is one of the biggest ones, and seems to show substructure in its hierarchical clustering. Although conservation of modules can be correlated with their size (Langfelder *et al*, 2011), the conservation of module blue is not completely explained by its size. This is evidenced by module turquoise, which is even bigger and shows very little evidence of preservation. Collectively, this suggests that the expression signatures at the beginning and end of the neurogenesis process are conserved between the mouse and chicken samples we used, while the middle and specific stages show weak or no evidence of conservation. However, this might also simply reflect the heterochronic differences in our samples, compared to the reference data.

The separate analysis of conservation of density and connectivity of the modules also turned out to be highly informative. We observe how the conservation of a module is driven by either of the two characteristics. This is very well illustrated in Figure 4

where the connectivity of module red shows remarkable conservation across all samples.

This is not the first time that WGCNA is applied to high throughput single-cell data (Wu *et al*, 2017; Niu *et al*, 2020; Korrapati *et al*, 2019), nor the first time that pseudocells are used to perform WGCNA calculations (Tosches *et al*, 2018). These studies, however, did not employ an iterative approach to the WGCNA calculations. Nonetheless, this work is also not the first time that an iterative approach is proposed for the use of WGCNA, either bulk or single-cell data (Greenfest-Allen *et al*, 2017; Feregrino *et al*, 2019; Kee *et al*, 2017). These iterative approaches are very similar, making use of statistical measurements produced with the WGCNA tools themselves, to measure significance of gene membership to their assigned modules. Genes with significant membership values are then grouped and used as input data in the next iteration. Moreover WGCNA has been used for comparisons of scRNA-seq of different studies (Tosches *et al*, 2018). However, here we use a more robust comparison approach, using a combination of statistical tests, which not only compare the identity of modules, but test for the conservation of co-expression network metrics.

It's sometimes assumed that WGCNA reveals regulatory networks, or functional relationships between genes (Niu *et al*, 2020; Korrapati *et al*, 2019). However, the input data and calculations made simply reflect modules of co-expression (Langfelder & Horvath, 2008) and not regulatory interactions *per se*. While the co-expression of transcription factors and target genes might reflect their interaction, relying on gene expression alone to infer this process is likely to result in a high proportion of false positives. Other approaches, like SCENIC (Aibar *et al*, 2017), which make use of *cis*-regulatory sequence information to make binding motif enrichment analyses and link this to the co-expression data seems more appropriate for these purposes. SCENIC is specifically designed to infer gene regulatory networks from single-cell data, and can even be used to make cross species comparisons. Nonetheless, the use of SCENIC is restricted to three model species (human, mouse and fly), due to the requirement of an extensive and high-quality binding motif dataset.

In summary, we present a comparative approach of co-expression module detection for single-cell data. We have described the basic logic of our workflow and demonstrated that it provides meaningful insights when comparing completely independent data sets, even from different species like mouse and chicken. Our proposed approach helped us recover signatures of expression programs that might define the same cell type across species. While we cannot infer regulatory relationships among these genes based on co-expression alone, conservation of density or connectivity might reflect the underlying evolutionary changes in regulation of the co-expression modules. In conclusion, we provide a tool to better understand and explore transcriptional signatures across species at single cell resolution, as well as provide clues of regulatory changes that might be driving the underlying cell type evolutionary process.

Methods

The dissociation of the chicken embryonic tissue, as well as the Chromium 10x Genomics scRNA-seq library preparation were performed as in our previous studies (Feregrino *et al*, 2019). The ground single-cell analysis, of filtering, dimensionality reductions and clustering was done following our previously stablished workflow (see Chapter 1).

Pseudocells

The production of pseudocells is performed using different functions from Seurat v3.1.4 (Stuart *et al*, 2019) using, otherwise stated, the default options. In a first step, the 10 nearest neighbors (nn) of each cell are calculated using the function "FindNeighbors", based on the PCA space and only the first relevant components (see Chapter 1) (determined for every dataset, same as for example number of PCs used for tSNE calculations). From each of the previously calculated cell clusters, 20% of the cells were chosen randomly as seed cells, to which the nn cells were then aggregated. Since this is a random sampling, not all cells in the dataset end up as the nn of the seeds cells. In order to maximize the amount of cells aggregated into pseudocells, we performed a sampling of 50 sets of randomly chosen seed cells, we counted the total amount of cells that are their nn and chose the set with the largest count. We created a table of all the cells, and their assigned seed cell (none at the beginning). Then using the table of the seed cells and their 10 nn cells, we performed and iterative pseudocell assignment. In a first step, the one seed cell with the lowest amount of remaining nn was chosen, to avoid greedy seed cells. Then, from the pool of nn cells one was chosen at random. The corresponding pseudocell is recorded and the nearest neighbor cell removed from the pool of cells to choose from, since cells are typically nearest neighbors to several cells. Finally, we creates a subset of the original Seurat object, containing only the cells assigned to a pseudocell, and using the function "AverageExpression" we constructed the pseudocells using the scaled data.

Iterative WGCNA

Our WGCNA approach uses mainly functions and tools from the WGCNA package v1.6.6 (Langfelder & Horvath, 2008) combined with some from the Seurat toolkit. We first created a subset of the pseudocells expression matrix using variable genes calculated using Seurat on the single-cell data. Using the function "pickSoftThreshold", we asses a soft thresholding power to use in the calculations of the adjacency matrix. The iterative WGCNA analysis is then started. In brief, a topological overlap matrix is produced from the subset of expression data using the function "TOMsimilarityFromExpr", the previously calculated soft thresholding power and the bidweight midcorrelation. A hierarchical clustering tree is computed using the topological overlap distances and the calculation of the cut height to construct modules. A series of cut heights are set around the automatically generated cut height, in steps of 0.0001 until +- 0.0005, and the size of the detected modules for each cut height is recorded. The cutting height is then chosen as the one that

produces the smallest or no gray module (unassigned genes), and the same amount of modules as the previous iteration (20, in the first run). Once a height is chosen, modules are detected and the module membership of each to its parent module is calculated using the function “geneModuleMembership”. Genes that are not assigned, or without a significant module membership are removed, and the remaining genes used to start the process again.

Once all genes have a significant membership to their assign module, we calculate the eigengenes and average expression of each module in the single-cell data and plot them. We create network visualizations for each module using the R packages network (Butts, 2008) and GGally (Schloerke *et al*, 2020). In order to maintain the visualization of each module meaningful, the node sizes representing the module membership of the gene, and the edge thickness and intensity representing the topological overlap are scaled module-wise. GO term enrichment analyses are performed using Limma (Ritchie *et al*, 2015) and our previous approach (Feregrino *et al*, 2019).

Comparative WGCNA

For the comparative analysis, we used pseudocell data, of the reference and the test samples. Using a 1-to1 orthologous genes list obtained from ENSEMBL BioMart (Kinsella *et al*, 2011), we subset the modules to contain only those genes. We further analyze the test data to assess genes that are not expressed in those samples, and also remove it from the modules. The conservation test is done using the function “modulePreservation” with the filtered module assignments, using the bidweight midcorrelation, a maximal gold modules size of 300, and 20 permutations.

References

Aibar S, González-Blas CB, Moerman T, Huynh-Thu VA, Imrichova H, Hulselmans G, Rambow F, Marine JC, Geurts P, Aerts J, Van Den Oord J, Atak ZK, Wouters J & Aerts S (2017) SCENIC: Single-cell regulatory network inference and clustering. *Nat. Methods* **14:** 1083–1086

Alpert A, Moore LS, Dubovik T & Shen-Orr SS (2018) Alignment of single-cell trajectories to compare cellular expression dynamics. *Nat. Methods* **15:** 267–270

Arendt D, Musser JM, Baker CVH, Bergman A, Cepko C, Erwin DH, Pavlicev M, Schlosser G, Widder S, Laubichler MD & Wagner GP (2016) The origin and evolution of cell types. *Nat. Rev. Genet.* **17:** 744–757

Bacher R & Kendziorski C (2016) Design and computational analysis of single-cell RNA-sequencing experiments. *Genome Biol.* **17:**

Brawand D, Soumillon M, Necsulea A, Julien P, Csárdi G, Harrigan P, Weier M, Liechti A, Aximu-Petri A, Kircher M, Albert FW, Zeller U, Khaitovich P, Grützner F, Bergmann S, Nielsen R, Pääbo S & Kaessmann H (2011) The evolution of gene expression levels in mammalian organs. *Nature* **478:** 343–348

Butts CT (2008) network: A package for managing relational data in R. *J. Stat. Softw.* **24:**

Cardoso-Moreira M, Halbert J, Valloton D, Velten B, Chen C, Shao Y, Liechti A, Ascenção K, Rummel C, Ovchinnikova S, Mazin P V., Xenarios I, Harshman K, Mort M, Cooper DN, Sandi C, Soares MJ, Ferreira PG, Afonso S, Carneiro M, et al (2019a) Gene expression across mammalian organ development. *Nature* **571:** 505–509

Cardoso-Moreira M, Velten B, Mort M, Cooper D, Huber W & Kaessmann H (2019b) Developmental gene expression differences between humans and mammalian models. *bioRxiv***:** 747782

Crow M, Paul A, Ballouz S, Huang ZJ & Gillis J (2018) Characterizing the replicability of cell types defined by single cell RNA-sequencing data using MetaNeighbor. *Nat. Commun.* **9:**

Delile J, Rayon T, Melchionda M, Edwards A, Briscoe J & Sagner A (2019) Single cell transcriptomics reveals spatial and temporal dynamics of gene expression in the developing mouse spinal cord. *Dev.* **146:**

Ericson J, Rashbass P, Schedl A, Brenner-Morton S, Kawakami A, Van Heyningen V, Jessell TM & Briscoe J (1997) Pax6 controls progenitor cell identity and neuronal fate in response to graded Shh signaling. *Cell* **90:** 169–180

Feregrino C, Sacher F, Parnas O & Tschopp P (2019) A single-cell transcriptomic atlas of the developing chicken limb. *BMC Genomics* **20:**

Ficklin SP & Feltus FA (2011) Gene coexpression network alignment and conservation of gene modules between two grass species: Maize and rice. *Plant Physiol.* **156:** 1244–1256

Frumkin D, Wasserstrom A, Kaplan S, Feige U & Shapiro E (2005) Genomic variability within an organism exposes its cell lineage tree. *PLoS Comput. Biol.* **1:** 0382–0394

Graham V, Khudyakov J, Ellis P & Pevny L (2003) SOX2 functions to maintain neural progenitor identity. *Neuron* **39:** 749–765

Greenfest-Allen E, Cartailler J-P, Magnuson M & Stoeckert C (2017) iterativeWGCNA: iterative refinement to improve module detection from WGCNA co-expression networks. *bioRxiv***:** 234062

Hamburger V & Hamilton HL (1951) A series of normal stages in the development of the chick embryo. *J. Morphol.* **88:** 49–92

Han X, Zhou Z, Fei L, Sun H, Wang R, Chen Y, Chen H, Wang J, Tang H, Ge W, Zhou Y, Ye F, Jiang M, Wu J, Xiao Y, Jia X, Zhang T, Ma X, Zhang Q, Bai X, et al (2020) Construction of a human cell landscape at single-cell level. *Nature* **581:** 303–309

Hill MA (2020) Embryology Carnegie Stage Comparison. *UNSW Embryol.* Available at: https://embryology.med.unsw.edu.au/embryology/index.php/Carnegie_Stage_Comparison [Accessed October 2, 2020]

Kanton S, Boyle MJ, He Z, Santel M, Weigert A, Sanchís-Calleja F, Guijarro P, Sidow L, Fleck JS, Han D, Qian Z, Heide M, Huttner WB, Khaitovich P, Pääbo S, Treutlein B & Camp JG (2019) Organoid single-cell genomic atlas uncovers human-specific features of brain development. *Nature* **574:** 418–422

Kee N, Volakakis N, Kirkeby A, Dahl L, Storvall H, Nolbrant S, Lahti L, Björklund ÅK, Gillberg L, Joodmardi E, Sandberg R, Parmar M & Perlmann T (2017) Single-Cell Analysis Reveals a Close Relationship between Differentiating Dopamine and Subthalamic Nucleus Neuronal Lineages. *Cell Stem Cell* **20:** 29–40

Kinsella RJ, Kähäri A, Haider S, Zamora J, Proctor G, Spudich G, Almeida-King J, Staines D, Derwent P, Kerhornou A, Kersey P & Flicek P (2011) Ensembl BioMarts: A hub for data retrieval across taxonomic space. *Database* **2011:**

Korrapati S, Taukulis I, Olszewski R, Pyle M, Gu S, Singh R, Griffiths C, Martin D, Boger E, Morell RJ & Hoa M (2019) Single Cell and Single Nucleus RNA-Seq Reveal Cellular Heterogeneity and Homeostatic Regulatory Networks in Adult Mouse Stria Vascularis. *Front. Mol. Neurosci.* **12:**

Kouzarides T (2007) Chromatin Modifications and Their Function. *Cell* **128:** 693–705

Langfelder P & Horvath S (2008) WGCNA: An R package for weighted correlation network analysis. *BMC Bioinformatics* **9:**

Langfelder P & Horvath S (2014) Tutorial for the WGCNA package for R : 1. Data input and cleaning. *Tutorials WGCNA Packag.* Available at: https://horvath.genetics.ucla.edu/html/CoexpressionNetwork/Rpackages/WGCNA/Tutorials/FemaleLiver-01-dataInput.pdf [Accessed October 9, 2019]

Langfelder P, Luo R, Oldham MC & Horvath S (2011) Is my network module preserved and reproducible? *PLoS Comput. Biol.* **7:** e1001057

Lazzaro BP & Clark AG (2013) Rapid evolution of innate immune response genes. In *Rapidly Evolving Genes and Genetic Systems* pp 203–210.

Lin S, Lin Y, Nery JR, Urich MA, Breschi A, Davis CA, Dobin A, Zaleski C, Beer MA, Chapman WC, Gingeras TR, Ecker JR & Snyder MP (2014) Comparison of the transcriptional landscapes between human and mouse tissues. *Proc. Natl. Acad. Sci. U. S. A.* **111:** 17224–17229

Luecken MD, Buttner M, Chaichoompu K, Danese A, Interlandi M, Mueller MF, Strobl DC, Zappia L, Dugas M, Colome-Tatche M & Theis FJ (2020) Benchmarking atlas-level data integration in single-cell genomics. *bioRxiv*

Marioni JC & Arendt D (2017) How single-cell genomics is changing evolutionary and developmental biology. *Annu. Rev. Cell Dev. Biol.* **33:** 537–553

Niu J, Huang Y, Liu X, Zhang Z, Tang J, Wang B, Lu Y, Cai J & Jian J (2020) Single-cell RNA-seq reveals different subsets of non-specific cytotoxic cells in teleost. *Genomics* **112:** 5170–5179

Payne JL & Eppstein MJ (2009) Evolutionary dynamics on scale-free interaction networks. *IEEE Trans. Evol. Comput.* **13:** 895–912

Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W & Smyth GK (2015) Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43:** e47

Schloerke B, Briatte F, bigbeardesktop, Crowley J, justsomeone1001, Cook D, Ibanez E, Ross, Ogden K, Sievert C, Joseph, Spiller T, Gilligan J, Wallace E, elbamos, Beck MW, Toomet O, Richter J, Thoen E, Jones O, et al (2020) ggobi/ggally: v1.5.0.

Shafer MER (2019) Cross-Species Analysis of Single-Cell Transcriptomic Data. *Front. Cell Dev. Biol.* **7:**

Shapiro E, Biezuner T & Linnarsson S (2013) Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nat. Rev. Genet.* **14:** 618–630

Stolt CC, Lommes P, Sock E, Chaboissier MC, Schedl A & Wegner M (2003) The Sox9 transcription factor determines glial fate choice in the developing spinal cord. *Genes Dev.* **17:** 1677–1689

Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM, Hao Y, Stoeckius M, Smibert P & Satija R (2019) Comprehensive Integration of Single-Cell Data. *Cell* **177:** 1888-1902.e21

Stuart T & Satija R (2019) Integrative single-cell analysis. *Nat. Rev. Genet.* **20:** 257–272

Sudmant PH, Alexis MS & Burge CB (2015) Meta-analysis of RNA-seq expression data across species, tissues and studies. *Genome Biol.* **16:**

Tosches MA, Yamawaki TM, Naumann RK, Jacobi AA, Tushev G & Laurent G (2018) Evolution of pallium, hippocampus, and cortical cell types revealed by single-cell transcriptomics in reptiles. *Science (80-.).* **360:** 881–888

Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M, Lennon NJ, Livak KJ, Mikkelsen TS & Rinn JL (2014) The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32:** 381–386

Wu YE, Pan L, Zuo Y, Li X & Hong W (2017) Detecting Activated Cell Populations Using Single-Cell RNA-Seq. *Neuron* **96:** 313-329.e6

Zhou Y, Zhu J, Tong T, Wang J, Lin B & Zhang J (2019) A statistical normalization method and differential expression analysis for RNA-seq data between different species. *BMC Bioinformatics* **20:**

Zilionis R, Engblom C, Pfirschke C, Savova V, Zemmour D, Saatcioglu HD, Krishnan I, Maroni G, Meyerovitz C V., Kerwin CM, Choi S, Richards WG, De Rienzo A, Tenen DG, Bueno R, Levantini E, Pittet MJ & Klein AM (2019) Single-Cell Transcriptomics of Human and Mouse Lung Cancers Reveals Conserved Myeloid Populations across Individuals and Species. *Immunity* **50:** 1317-1334.e10

DISCUSSION AND OUTLOOK

Ever since the first exploratory observations of developmental processes, the chicken embryo has served as a prime model of study (Horder, 2010). It's easy to understand its scientific value: although not as closely related to humans as mice are, for example, chicken embryos are easy to obtain, maintain, access and manipulate. Although not a genetics model species, the study of the chicken development has produced vast amounts of observations and scientific breakthroughs, particularly in the field of experimental embryology (Davey & Tickle, 2007; Abramyan & Richman, 2018). The study of developmental biology, and thus of the chicken embryo, is now revisited as new technologies become available (Gilbert, 2017; Marioni & Arendt, 2017). With the advent of single-cell RNA-seq, developmental biologists are now trying to study chicken embryogenesis at the molecular level, with cellular.

In this thesis, I present a set of methods optimizations developed to study the development of the chicken embryo at single-cell resolution. I present also the study of cell fate decisions and patterning events at different scales, and from different perspectives. We produced a transcriptomic atlas of the different cell populations implicated in the patterning of the limb, made an in silico reconstruction of the cell-fate decisions occurring during digit formation, explored the cell fate convergence driving skeletogenesis in different embryonic regions with different origins, and finally stablished a workflow for cross-species comparisons of co-expression signatures along developmental trajectories.

Optimization

As the first research group to make use of single-cell RNA-seq techniques to study the development of the chicken, we presented a series of method optimizations necessary to make use of this novel technique. We made use of two different high-throughput scRNA-seq methods, namely Drop-seq and 10x Genomics Chromium Single Cell Gene Expression.

Owing to a difference in library preparation steps, the Chromium 10x Genomics assays result in a heavy 3' sequencing bias. This bias, in combination with the current state of the chicken genome annotation, leads to underrepresentation of the expression of certain genes, some of which have developmental relevance, i.e. SOX9. We performed the optimization of the chicken genome annotation, to be compatible with Chromium 10x Genomics assays. We showed that our approach increases the UMI counts of more than 4,000 genes, which would otherwise be underrepresented in the final expression matrix. We are confident that our modification of the genome annotation, more complex than other not data-driven modifications (Estermann *et al*, 2020), will be of use for future research projects involving the sequencing of single-cell transcriptomics from chicken tissue. Additionally, the general logic of our approach may serve as the basis for similar modifications, in other species.

We presented a single-cell RNA-seq quality filtering framework consisting of relative filters based on total UMI count per cell, to remove probable doublets and cells of low quality; a combined relative filter based on percentage of mitochondrial UMIs and total UMI counts, to remove cell barcodes associated with dying cells; and a threshold to remove cells with high UMI count coming representing only a few genes in total. Our filtering framework, besides being in line with the best practices of single-cell RNA-seq analyses (Luecken & Theis, 2019), has proven very useful during our research. Our framework will continue to be useful during future studies from our research group, and will serve as the basis for future and updated filtering techniques.

We also performed a comparison of the pseudotime ordering methods Monocle (Trapnell *et al*, 2014) and URD (Farrell *et al*, 2018). We observed that both approaches can produce an overall similar pseudotime topology, showing the same branching pattern. However, we found very small correlations between the gene trajectories along the branches. Moreover, we observed that the two resulting pseudotimes differ the most in the middle of the total trajectory, which is the most interesting part of a cell-fate bifurcating process. Nonetheless, after the publication of an extensive comparison of pseudotime analyses (Saelens *et al*, 2019), we decided to use Scanorama (Hie *et al*, 2019), another, well-reviewed method for our analyses. Furthermore, we acknowledge that our comparison is far from systematic: it doesn't asses any quality from the two methods we used and it's only restricted to its performance with our data.

Single cell atlas of the developing chicken limb

Using our optimized analytical tools, we set out to describe the cellular composition of one of the best studied models of tissue patterning, the developing tetrapod limb (Zeller *et al*, 2009; Petit *et al*, 2017). We characterized the transcriptomic profiles of the cell populations present at three key stages of chicken limb development: growth and embryonic axes establishment at stage HH25, differential digit patterning at stage HH29, and final digit elongation at HH31. Notably, we characterized the transcriptome of minute cell populations like the AER, and other difficult to sample populations, like perichondrium and non-skeletal connective tissue.

During our analyses we also calculated co-expression modules which reflected the transcriptional signatures of populations known to be essential to the patterning process. We found four different co-expression modules marking three distinct interdigits cell populations. Of these, we were able to identify a cell population and one corresponding co-expression module which represent the fourth interdigit, thus conciliating single-cell expression data with spatial gene expression information (Wang *et al*, 2011). We also presented the expression signature of the distal mesenchyme, a distinct cell population which is likely implicated in the processing and integration of different signals coming from the AER (Zeller *et al*, 2009). Finally, we present two co-expression modules, which reflect the chondrogenic process of the developing digits. We found a module consistent with a chondrogenic signature (Gï¿½mez-Picos & Eames, 2015; Kozhemyakina *et al*, 2015) highly expressed in differentiated chondrocytes. On the other hand, a co-expression module composed of genes that have not yet been directly implicated in the chondrogenic process is highly expressed in early chondrocytes, thus suggesting a transcriptional priming process. This study represented the first single-cell transcriptomic description of the developing limb, and has served since as a reference for other studies concerning limb patterning.

Pseudotemporal reconstruction of digit patterning

After describing the cellular composition of the developing limb, we set out to describe the dynamic patterning process shaping the phalanx – joint design of the digits at single-cell resolution. For this, we reconstructed the digit morphogenesis process in silico making use of single-cell pseudotime analysis tools. We obtained pseudotemporal reconstructions reflecting the bifurcation of the phalanx versus joint cell fate decision. We describe the in silico, reconstructed process and found that it reflects the in vivo transcriptional dynamics of several key genes. We confirmed this along the pre-bifurcation, phalanx and joint trajectories. We show that our pseudotime reconstruction suggests that at least two genes, namely PTN and SULF1, are upregulated before GDF5, the earliest known interzone molecular marker (Koyama et al, 2008; Storm & Kingsley, 1996). This observations will serve as the basis for further in vivo experiments, to validate their expression dynamics and functionality in the digit patterning process. Furthermore, our analyses could be reinforced by

employing different and novel methods designed to analyze pseudotime and cell fate processes (La Manno et al, 2018; Bergen et al, 2020; Montero et al, 2008).

Convergent skeleton cell-fate specification

We further analyze the cell specification process of the skeleton in other parts of the developing embryo. We analyzed single cell data from different skeletogenic cell populations, stemming from different embryonic regions, and, importantly, different developmental origins. We characterized and identified the different cell populations present in our samples and found cells resembling chondrogenic mesenchyme populations, as well as early chondrocytes. Our attempts to integrate the chondrogenic populations in a single dataset have not been entirely successful, and we will probably follow a different strategy to compare the chondrogenic processes. For this, as with our digit patterning project, we could benefit from the use of alternative techniques to calculate pseudotime (La Manno *et al*, 2018; Bergen *et al*, 2020) and infer cell fate progression (Lange *et al*, 2020). This study has also served as the basis of further assays to understand the regulatory dynamics underlying the convergence process. Specifically, our research group has produced single-cell ATAC-seq data from equivalent locations and samples, which is currently being analyzed using our transcriptomic data as a basis.

Moreover, we performed single-cell analyses of xenografted tissue in order to understand the effects of cell-intrinsic and cell-extrinsic factors during the skeletal developmental process. We developed a framework which successfully identifies cells of the donor species from the dissociated xenograft. Although still working with a limited number of cells, we are on the trail to improve the dissociation of xenografted tissue, in order to recover a higher number of viable grafted cells. Nonetheless, we have already stablished the basis for a large scale analysis of xenografted tissue. This also represents, to our knowledge, the first time that non-superficial xenografted tissue is analyzed using high-throughput single-cell transcriptomic methods.

Cross-species comparison of co-expression modules

After describing cell-fate specification processes across the embryo, we implemented an approach to compare these processes across species, to understand development in its evolutionary context. We made use of WGCNA (Langfelder & Horvath, 2008) and its different tools to perform comparative analyses of co-expression. We presented an iterative implementation of WGCNA adapted to work with single-cell transcriptomic data, similar to existing approaches which aim to refine the co-expression results (Greenfest-Allen *et al*, 2017; Kee *et al*, 2017). Furthermore, we use the conservation tests developed by the same research team that created WGCNA (Langfelder *et al*, 2011), and present their functionality with single-cell data stemming from completely independent experiments and different species. Here, we observed the conservation and divergence of gene expression signatures present in homologous cell types of mouse and chicken.

Conclusion

Overall in this thesis, I present an integrative framework of different analyses that complement each other for the study of developmental processes at different scales with single-cell resolution, in the chicken embryo. I present several novelties regarding the study of the chicken development which will complement long-established and powerful embryology, molecular biology and even genetics methods. Moreover, our implementations can be adapted for the study of other model species aside from human, mouse and fruit-fly. During this work, I described transcriptional profiles of important and iconic signaling centers, analyzed a computational reconstruction showing cells transitioning from one state to another as they mature, explored the developmental convergence of cells from different developmental and evolutionary origins into the same functional cell type, and even recovered transcriptional signatures conserved in cells separated by millions of years of evolutionary divergence. I expect these contributions, both of observations and analyses, to further expand the potential of chicken embryology in the field of developmental biology.

References

Abramyan J & Richman JM (2018) Craniofacial development: Discoveries made in the chicken embryo. *Int. J. Dev. Biol.* **62:** 93–103

Bergen V, Lange M, Peidli S, Wolf FA & Theis FJ (2020) Generalizing RNA velocity to transient cell states through dynamical modeling. *Nat. Biotechnol.*

Davey MG & Tickle C (2007) The chicken as a model for embryonic development. *Cytogenet. Genome Res.* **117:** 231–239

Estermann MA, Williams S, Hirst CE, Roly ZY, Serralbo O, Adhikari D, Powell D, Major AT & Smith CA (2020) Insights into Gonadal Sex Differentiation Provided by Single-Cell Transcriptomics in the Chicken Embryo. *Cell Rep.* **31:**

Farrell JA, Wang Y, Riesenfeld SJ, Shekhar K, Regev A & Schier AF (2018) Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis. *Science (80-.).* **360:**

Gï¿½mez-Picos P & Eames BF (2015) On the evolutionary relationship between chondrocytes and osteoblasts. *Front. Genet.* **6:**

Gilbert SF (2017) Developmental biology, the stem cell of biological disciplines. *PLoS Biol.* **15:** e2003691

Greenfest-Allen E, Cartailler J-P, Magnuson M & Stoeckert C (2017) iterativeWGCNA: iterative refinement to improve module detection from WGCNA co-expression networks. *bioRxiv***:** 234062

Hie B, Bryson B & Berger B (2019) Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. *Nat. Biotechnol.* **37:** 685–691

Horder T (2010) History of Developmental Biology. In *Encyclopedia of Life Sciences*

Kee N, Volakakis N, Kirkeby A, Dahl L, Storvall H, Nolbrant S, Lahti L, Björklund ÅK, Gillberg L, Joodmardi E, Sandberg R, Parmar M & Perlmann T (2017) Single-Cell Analysis Reveals a Close Relationship between Differentiating Dopamine and Subthalamic Nucleus Neuronal Lineages. *Cell Stem Cell* **20:** 29–40

Koyama E, Shibukawa Y, Nagayama M, Sugito H, Young B, Yuasa T, Okabe T, Ochiai T, Kamiya N, Rountree RB, Kingsley DM, Iwamoto M, Enomoto-Iwamoto M & Pacifici M (2008) A distinct cohort of progenitor cells participates in synovial joint and articular cartilage formation during mouse limb skeletogenesis. *Dev. Biol.* **316:** 62–73

Kozhemyakina E, Lassar AB & Zelzer E (2015) A pathway to bone: Signaling molecules and transcription factors involved in chondrocyte development and maturation. *Dev.* **142:** 817–831

Lange M, Klein M, Restrepo Lopez JL, Theis FJ & Pe'er D (2020) CellRank.

Langfelder P & Horvath S (2008) WGCNA: An R package for weighted correlation network analysis. *BMC Bioinformatics* **9:**

Langfelder P, Luo R, Oldham MC & Horvath S (2011) Is my network module preserved and reproducible? *PLoS Comput. Biol.* **7:** e1001057

Luecken MD & Theis FJ (2019) Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol. Syst. Biol.* **15:**

La Manno G, Soldatov R, Zeisel A, Braun E, Hochgerner H, Petukhov V, Lidschreiber K, Kastriti ME, Lönnerberg P, Furlan A, Fan J, Borm LE, Liu Z, van Bruggen D, Guo J, He X, Barker R, Sundström E, Castelo-Branco G, Cramer P, et al (2018) RNA velocity of single cells. *Nature* **560:** 494–498

Marioni JC & Arendt D (2017) How single-cell genomics is changing evolutionary and developmental

biology. *Annu. Rev. Cell Dev. Biol.* **33:** 537–553

Montero JA, Lorda-Diez CI, Gañan Y, Macias D & Hurle JM (2008) Activin/TGFβ and BMP crosstalk determines digit chondrogenesis. *Dev. Biol.* **321:** 343–356

Petit F, Sears KE & Ahituv N (2017) Limb development: A paradigm of gene regulation. *Nat. Rev. Genet.* **18:** 245–258

Saelens W, Cannoodt R, Todorov H & Saeys Y (2019) A comparison of single-cell trajectory inference methods. *Nat. Biotechnol.* **37:** 547–554

Storm EE & Kingsley DM (1996) Joint patterning defects caused by single and double mutations in members of the bone morphogenetic protein (BMP) family. *Development* **122:** 3969–3979

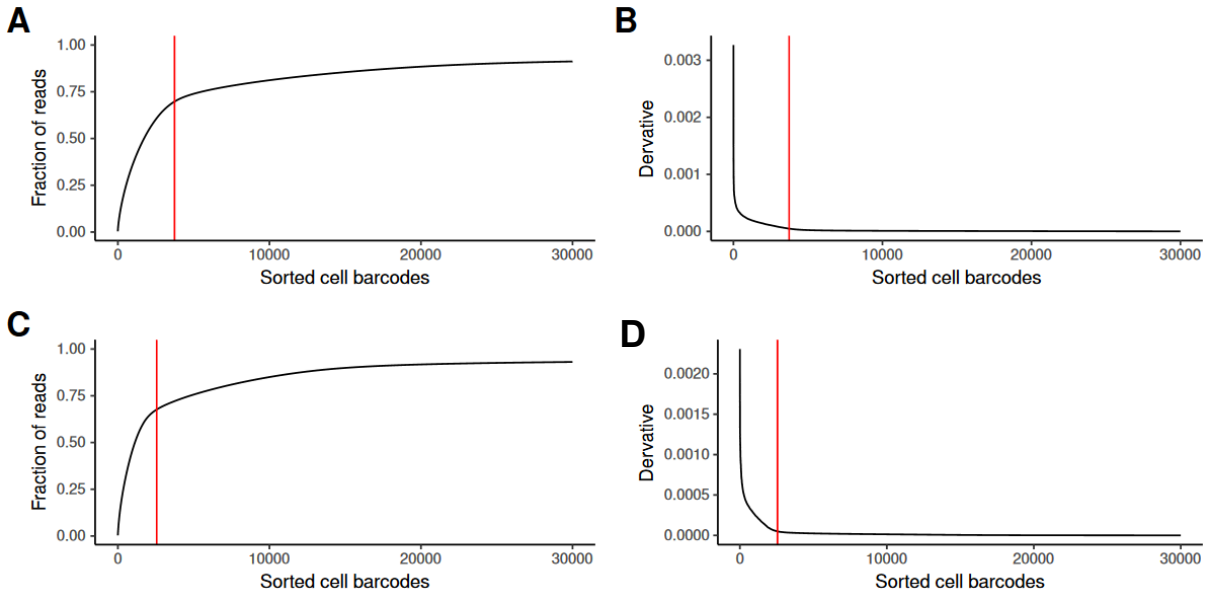
Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M, Lennon NJ, Livak KJ, Mikkelsen TS & Rinn JL (2014) The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32:** 381–386

Wang Z, Young RL, Xue H & Wagner GP (2011) Transcriptomic analysis of avian digits reveals conserved and derived digit identities in birds. *Nature* **477:** 583–587

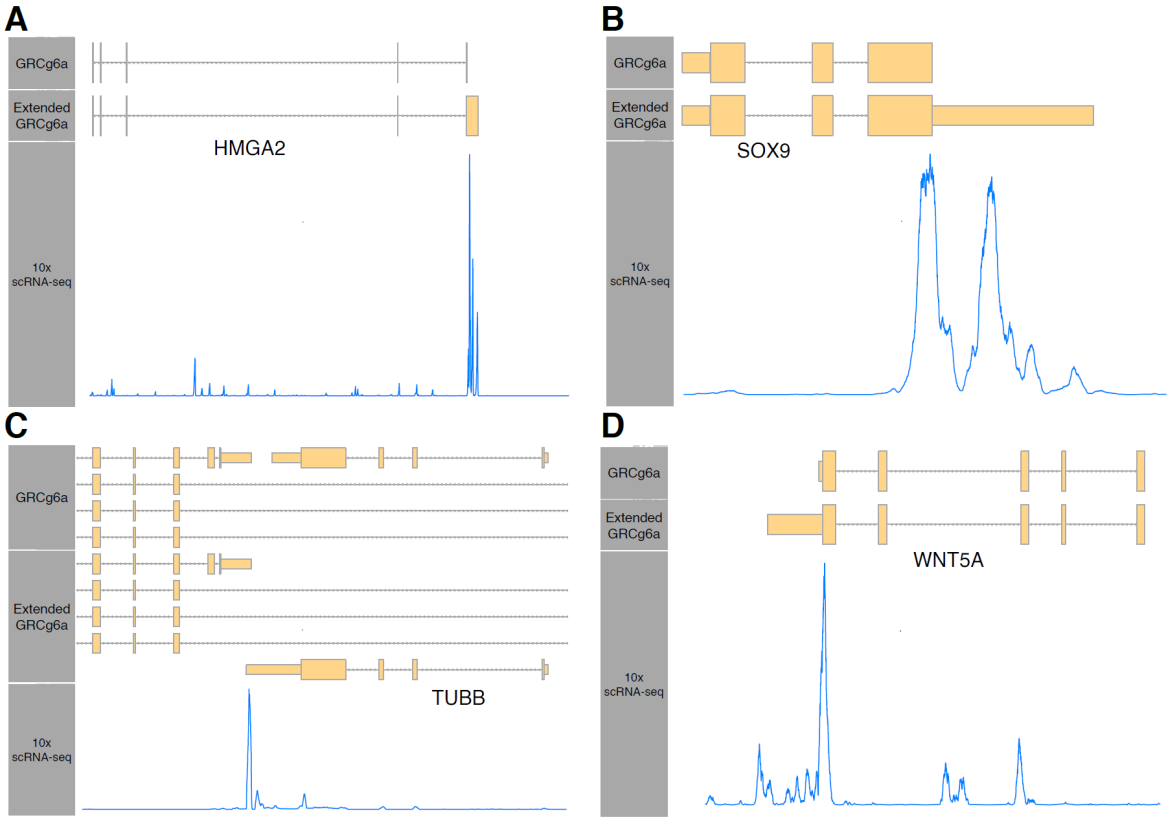
Zeller R, López-Ríos J & Zuniga A (2009) Vertebrate limb bud development: Moving towards integrative analysis of organogenesis. *Nat. Rev. Genet.* **10:** 845–858

SUPPLEMENT 1

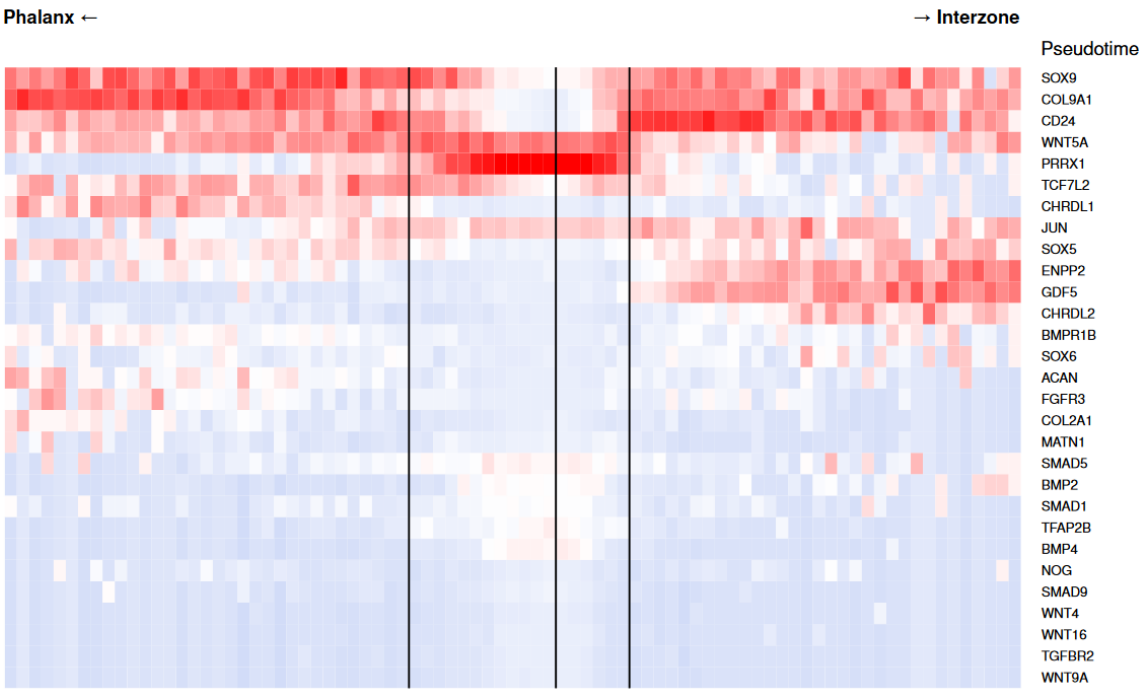
Supplementary figures 1, 2 and 3



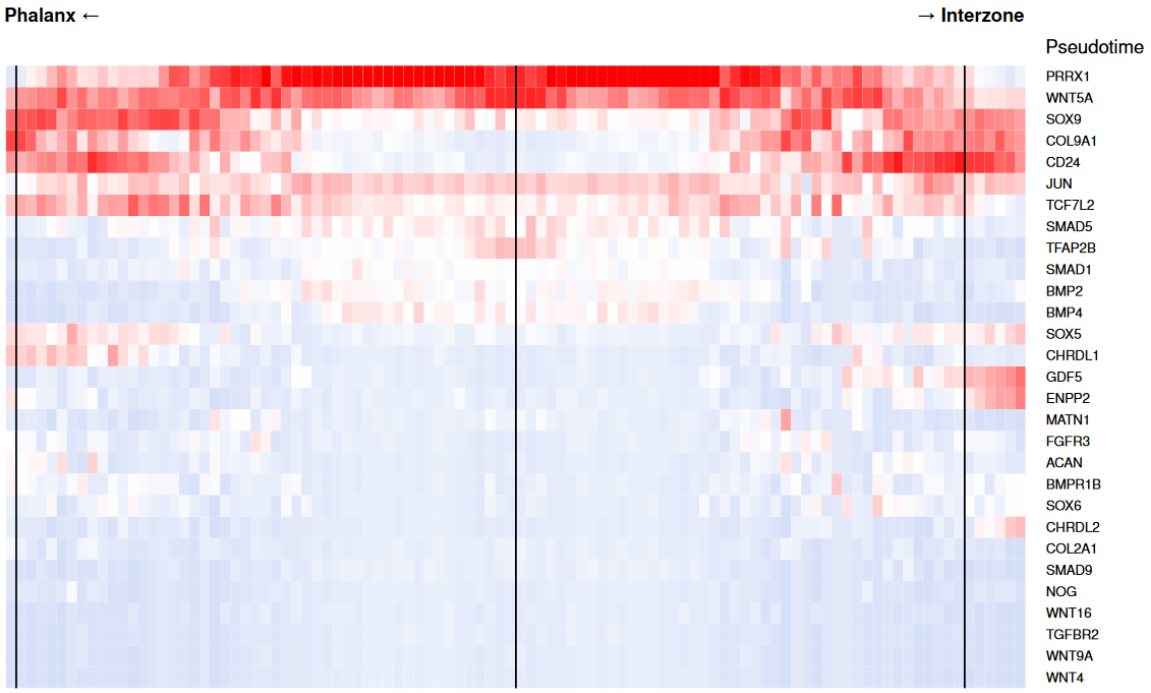
Supplementary Figure 1 Threshold calculation for the other two Drp-seq samples. Same representation as in Chapter 1 Figure 1.
A and **C** cumulative plot of the reads per cell barcodes. **B** and **D** First derivative of the curve in A and C.



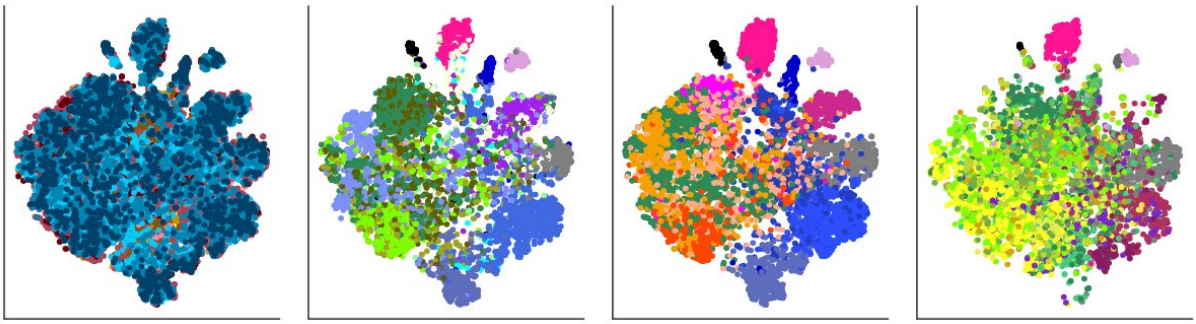
Supplementary Figure 2 Alignment of reads to the original annotation of chicken GRCg6a and our transcript extension. Depicted ate the genes with the highest UMI increases. **A** HMGA2, forward strand. **B** SOX9, forward strand. **C** TUBB, reverse strand. **D** WNT5A, reverse strand.



Supplementary Figure 3 Expression dynamics of different genes across the phalanx-joint divergence pseudotime reconstruction as calculated on Diffusion Map.
Corresponds to Chapter 3 Figure 4



Supplementary Figure 4 Expression dynamics of different genes across the phalanx-joint divergence pseudotime reconstruction as calculated on an exaggerated tSNE.
Corresponds to Chapter 3 Figure 4



Supplementary Figure 5 Data integration and cell type correspondence of cells from the three different embryonic regions using Seurat. Colors and panels correspond to Chapter 4 Figure 2

SUPPLEMENT 2

Output example of our WGCNA analysis

WGCNA workflow, Briscoe 2019

Christian Feregrino

Date: 14.05.20

We use WGCNA to run an iterative analysis on a data set.

Pre-analysis

We need to first set up our working environment.

Code

Get our single cells and the pseudocells to be able to run the analysis

We need the data in either seurat or data-frame format. This must contain at least 20 samples (no problems with sc data), according to the documentation of the package itself (https://horvath.genetics.ucla.edu/html/CoexpressionNetwork/Rpackages/WGCNA/faq.html).

Code

```
the condition has length > 1 and only the first element will be used
```

Code

```
[1] "We have 2052 genes in the variable genes object"
```

Code

```
[1] 2052 1626
```

Code

Now we need to calculate the soft threshold power. First it calculates the similarity and then transforms this similarity to a weighted network. The scale-free topology is calculated for each of the powers.

We choose the smallest power for which the scale-free topology fit index reaches 0.90. If none of the powers reaches 0.90, we take the one with the maximum, as long as we have a number above 0.75. If none of them reaches at least 0.75 we need to check our dataset.

Code

```
bicor: zero MAD in variable 'x'. Pearson correlation was used for individual columns with zero (or missing) MAD.bi
cor: zero MAD in variable 'y'. Pearson correlation was used for individual columns with zero (or missing) MAD.
```

Power <dbl>	SFT.R.sq <dbl>	slope <dbl>	truncated.R.sq <dbl>	mean.k. <dbl>	median.k. <dbl>	max.k. <dbl>
1	0.182	20.80	0.819	1.03e+03	1.03e+03	1080.00
2	0.179	-10.00	0.847	5.26e+02	5.20e+02	596.00
3	0.603	-11.30	0.771	2.71e+02	2.65e+02	345.00
4	0.953	-9.65	0.979	1.41e+02	1.36e+02	210.00
5	0.957	-7.27	0.973	7.50e+01	7.04e+01	135.00
6	0.953	-5.60	0.966	4.06e+01	3.67e+01	90.40
7	0.951	-4.45	0.974	2.25e+01	1.94e+01	63.10
8	0.939	-3.63	0.968	1.28e+01	1.04e+01	45.60
9	0.940	-2.97	0.977	7.60e+00	5.57e+00	33.90
10	0.925	-2.50	0.954	4.68e+00	3.03e+00	25.90

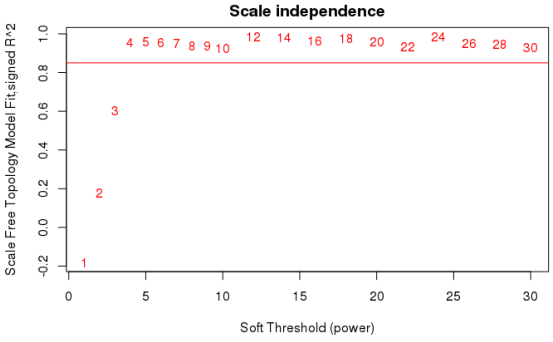
1-10 of 20 rows Previous 1 2 Next

Code

Code

```
 [1]  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20
```

Code



Code

```
[1] "Or power is 4"
```

WGCNA analysis

Now, the following piece of code will run WGCNA iteratively, to end up with out final modules. The iterations follow these steps:

- Tree cutting for module calculation
- Calculate an adjacency matrx from the data, then turn it into topological overlap and then into a distance matrix
- Calculate the tree based on the topological overlap distance
- Calculate the automatic height to cut out the 0.05 quantile
- Generate a matrix where we calculate the amount of modules, size of de modules and wheter if we have a grey module based on:
  - Different minimum module sizes, arbitrarely set to 7:30
  - Different cut-heiughts going 0.0005 up and down from the automatic height in steps of 0.0001
- Check if any combination of the parameters will get rid of the grey module
  - If we only get grey modules, subset the matrix for the height at which the grey module is the smallest
  - If we have a combination without grey module AND we have at least the same number of modules as in the beginning, we subset for whichever those heights are
- Take wichever min module sizes gives us at least the same amount of clusters as in the beginning, if none, then the highest
- Chose the max of the reminding min module sizes.
- Calculate the actual modules
- Calculate the eigengenes
- Calculate the module membership per gene
- Calculate the p.value of the membership to a given module
- Get rid of the genes in the grey module
- Delete the genes that are not significantely associated with their module
- Save the remaining geneset
- Print how many genes were deleted due to significance
- Unless 0 genes were deleted in the last step, update the expression matrices and beginn again.

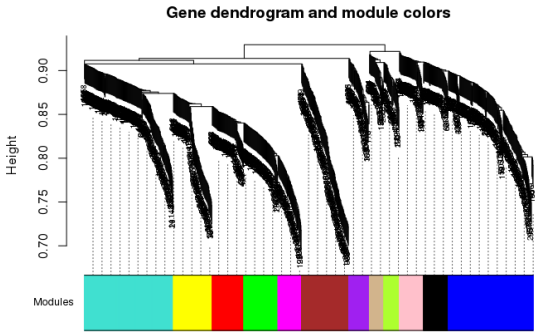
ONLY THE FIRST TWO ITERATIONS ARE DIFFERENT.

FIRST:

- During the tree cutting
- Set the minimum module size to arbitrary 15
- Subset for the heights that get rid of at least 50% of the genes with that min module size
- Choose the height that gives us the most clusters

SECOND:

- Set the resulting number of modules as the ground number of modules

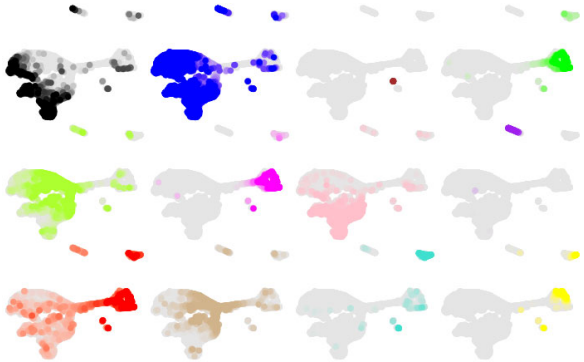


Code

Modules of co-expression

For single cells We can see what are the expression levels of our co-expression modules. If we provided a seurat object. We look, in this case, at a tSNE

Code



Create a list object with all the essential data from the analyses

Code

To create the visualizations of the actual co-expression networks, we use this code. It also creates files that can be read into cytoscape

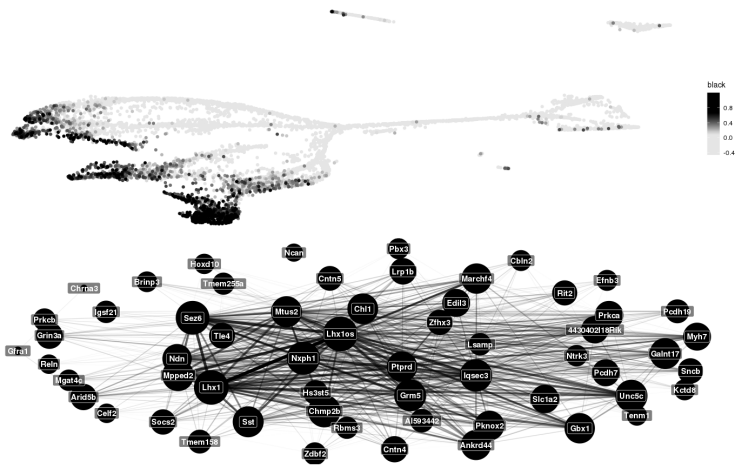
Code

To check the results module by module, we report some GO terms and individual plots, with module sizes and gene names. First the GO terms analyses

Code

And here the report of each of the modules

Code



NA

	module	color	size
1	black		59

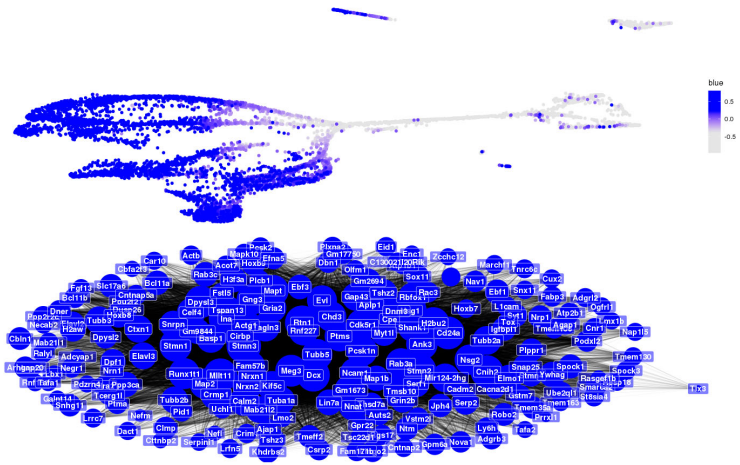
NA

.	.	.	.	.	.	.	.	.	.
Sst	Lhx1os	Zfhx3	Lhx1	Chmp2b	Cbln2	Hoxd10	Mpped2	Grm5	Al593442
Brinp3	Pbx3	Rbms3	Mgat4c	Unc5c	Nxph1	Sez6	Reln	Chl1	Gbx1
Cntn5	Celf2	Socs2	Lsamp	Lrp1b	Ntrk3	Rit2	Ptprd	Grin3a	Efnb3
Ndn	4430402I18Rik	Pcdh7	Slc1a2	Mtus2	Hs3st5	Myh7	Tmem255a	Igsf21	Zdbf2
Ankrd44	Iqsec3	Edil3	Pcdh19	Prkcb	Prkca	Tmem158	Tle4	Gfra1	Arid5b
Kctd8	Chrna3	Sncb	Cntn4	Tenm1	Pknox2	Marchf4	Galnt17	Ncan	

NA

.	.
nervous system development	neuron development
neuron differentiation	synapse organization
generation of neurons	cell development
locomotory behavior	neuron projection development
neurogenesis	adult behavior

NA



NA

	module	color	size
2	blue		202

NA

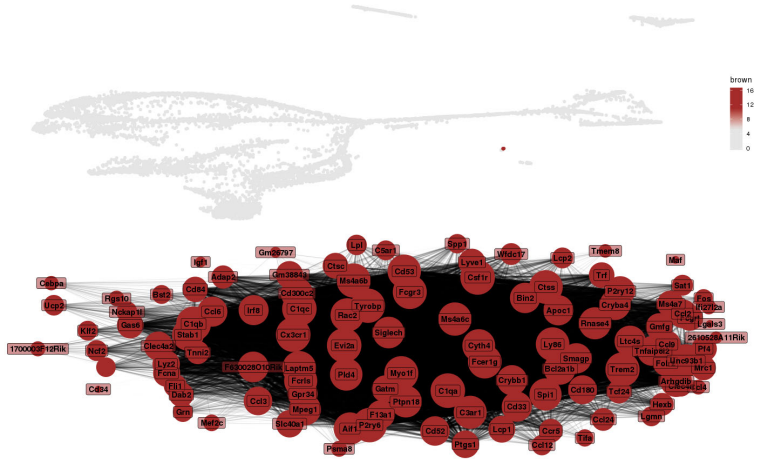
.	.	.	.	.	.	.	.	.	.
Snhg11	Nefl	Actb	Meg3	Nefm	Runx1t1	Gm2694	Gap43	Ebf1	Igfbpl1
Map1b	Tshz2	Stmn2	Basp1	Tmsb10	Tafa1	Tuba1a	Hoxb9	Pid1	Tubb3
Tlx3	Sox11	Mapt	Ncam1	Hoxb8	Ina	Ntm	Mab21l2	Pcsk1n	Nrp1
Ly6h	Stmn1	Bcl11b	Nrn1	Nrxn1	Stmn3	Cd24a	Cbln1	Stmn4	Celf4
Bcl11a	Ebf3	Rtn1	Gm17750	Dpysl2	Gng3	H2bu2	Spock3	Car10	Prrxl1
Efna5	Csrp2	Fgf13	Snrpn	Atp1b1	Tmem163	Actg1	Vstm2l	C130021I20Rik	Cnih2
Mllt11	Khdrbs2	Lrfn5	Dcx	Elavl2	Crmp1	Adcyap1	Elmo1	Thsd7a	Pdzrn4
Nnat	Map2	Auts2	Lingo2	Ptma	Olfm1	Tubb2b	Cacna2d1	Adgrl2	Nap1l5
Tubb2a	Nova1	Dpysl3	Lmx1b	Tmeff2	Uchl1	Cntnap2	Tubb5	Atp2b1	Robo2
Enc1	Ajap1	Tshz3	Ralyl	Cntnap5a	Pou2f2	Podxl2	Mab21l1	Tagln3	Lbx1
Calm2	Slc17a6	NA	Ogfrl1	Plcb1	Arhgap20	Ank3	Cdk5r1	Rnf227	Mir124-2hg
Shank1	Chd3	Rbfox1	Lin7a	Elavl3	H3f3a	Cnr1	Dusp26	Marchf1	Tafa2
Nsg1	Gria2	Plxna2	Lmo2	Rac3	Tmem108	Tcerg1l	Crim1	Fstl5	Dner
Myt1l	Necab2	Tspan13	Cbfa2t3	Grin2b	Resp18	Tsc22d1	Rnf152	Syt1	Nsg2
Glra2	Rasgef1b	Cirbp	Ctxn1	Hoxb7	Dnm3	Aplp1	Ppp3ca	St8sia4	Gm1673
Dact1	Arpp21	Negr1	Cttnbp2	Ptms	Acot7	Zcchc12	Snap25	Cux2	Evl
Nrxn2	Snx11	H2aw	Spock1	Pcsk2	Tox	Serpini1	Fam171b	Tnrc6c	Plppr1
Rab3c	Clmp	Nav1	Cadm2	Rab3a	Mapk10	Fabp3	Serf1	Eid1	Jph4
Gpm6a	Tmem130	Ppp2r2c	Dpf1	Smarca2	Agap1	Dbn1	Serp2	Gpr22	Galnt14
Kif5c	Fam57b	Gstm7	Rgs17	Ube2ql1	Tmem35a	L1cam	Lrrc7	Gm9844	Ywhag
Adgrb3	Cpe								

NA

.	.

.	.
nervous system development	neurogenesis
neuron projection development	neuron projection morphogenesis
neuron development	cell part morphogenesis
neuron differentiation	plasma membrane bounded cell projection morphogenesis
generation of neurons	cell projection morphogenesis

NA



NA

	module	color	size
	3	brown	112

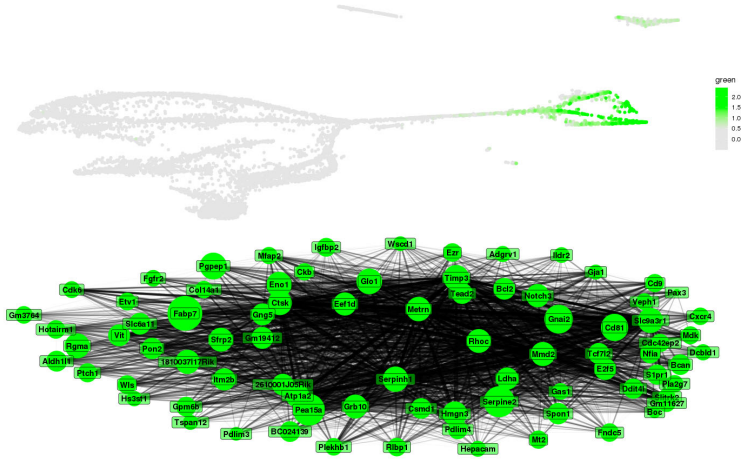
NA

.	.	.	.	.	.	.	.	.	.
Tyrobp	C1qb	C1qc	C1qa	Fcer1g	Cx3cr1	Csf1r	Pf4	Lyve1	F13a1
Dab2	Maf	Stab1	Ly86	Aif1	Ctss	Mpeg1	P2ry12	Mrc1	Fcrls
Ptpn18	Spp1	Fcgr3	Ccl24	Igf1	Ms4a6c	Lcp2	Trem2	Crybb1	Ccl9
Gmfg	Rac2	Slc40a1	Laptm5	Lyz2	Arhgdib	Pld4	Tcf24	Lpl	Wfdc17
Apoc1	Ccl6	Mef2c	Rnase4	Ccl4	Cd300c2	Fcgrt	Spi1	Clec4a2	Ccl3
Gpr34	Gatm	Ccl2	Hexb	Rgs10	Adap2	Cd53	P2ry6	C5ar1	Siglech
Cebpa	Cd34	Ucp2	Ifi27l2a	Cd52	Ncf2	Clec4n	Tnni2	Trf	Bin2
Unc93b1	Ccl12	Bcl2a1b	Ms4a7	Tmem8	Folr2	NA	Lcp1	Ms4a6b	Irf8
Cryba4	Gas6	Smagp	Gm38843	Tnfaip8l2	Cd84	Cyth4	Lgmn	Fcna	Ltc4s
C3ar1	Fli1	Grn	1700003F12Rik	Tifa	Ptgs1	Cd33	Klf2	Gm26797	Ctsc
Bst2	Psma8	Fos	Sat1	Cd180	F630028O10Rik	Evi2a	Nckap1l	Lgals3	2610528A11Rik
Myo1f	Ccr5								

NA

.	.
immune system process	myeloid leukocyte migration
defense response	neutrophil chemotaxis
response to external stimulus	cell chemotaxis
immune response	positive regulation of response to external stimulus
leukocyte chemotaxis	inflammatory response

NA



NA

	module	color	size
	4	green	81

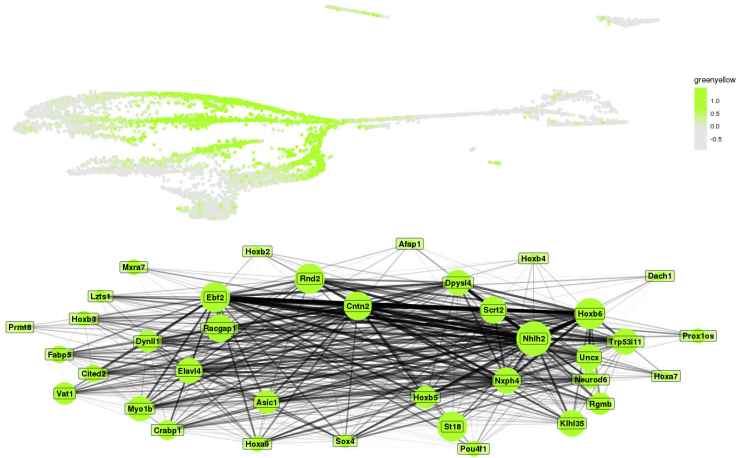
NA

.	.	.	.	.	.	.	.	.	.
Fabp7	Sfrp2	Ckb	Nfia	Mt2	Pea15a	Spon1	Pgpep1	Serpine2	Cxcr4
Pax3	Slc6a11	Timp3	Mdk	Metrn	Gm3764	Fndc5	Bcan	Col14a1	Eno1
Serpinh1	Gnai2	Igfbp2	Mmd2	Ldha	Slitrk2	Atp1a2	Gng5	Etv1	Cdk6
Cd81	Gas1	Mfap2	Grb10	Gja1	Pdlim4	Hotairm1	Pla2g7	Ezr	Tead2
Ctsk	Aldh1l1	Cd9	Itm2b	Ddit4l	Rgma	Wls	Bcl2	Pdlim3	Rlbp1
Wscd1	Slc9a3r1	Hs3st1	Gpm6b	Glo1	Hmgn3	Rhoc	BC024139	Veph1	Csmd1
Boc	1810037I17Rik	Eef1d	Hepacam	Vit	2610001J05Rik	Gm19412	S1pr1	Plekhb1	Tspan12
Dcbld1	Tcf7l2	Cdc42ep2	Pon2	Gm11627	Ptch1	Adgrv1	Notch3	E2f5	Fgfr2
Ildr2									

NA

.	.
anatomical structure morphogenesis	nervous system development
animal organ development	animal organ morphogenesis
tissue development	anatomical structure development
growth	cellular component morphogenesis
developmental growth	system development

NA



NA

	module	color	size
	5	greenyellow	37

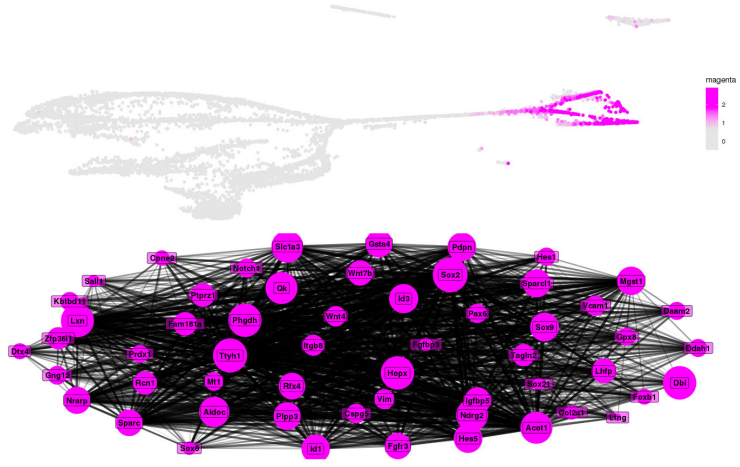
NA

.	.	.	.	.	.	.	.	.	.
Crabp1	Pou4f1	Neurod6	Nhlh2	Elavl4	Cntn2	Hoxb6	Ebf2	Sox4	Uncx
Rnd2	Hoxa9	Racgap1	St18	Cited2	Hoxa7	Nxph4	Dpysl4	Klhl35	Lzts1
Trp53i11	Hoxb5	Fabp5	Asic1	Scrt2	Dach1	Prox1os	Dynll1	Hoxb2	Myo1b
Hoxb4	Rgmb	Mxra7	Hoxb3	Vat1	Afap1	Prmt8			

NA

.	.
embryonic skeletal system development	regulation of biosynthetic process
skeletal system morphogenesis	pattern specification process
embryonic skeletal system morphogenesis	anterior/posterior pattern specification
skeletal system development	regulation of RNA metabolic process
regulation of cellular biosynthetic process	positive regulation of transcription by RNA polymerase II

NA



NA

	module	color	size
	6	magenta	56

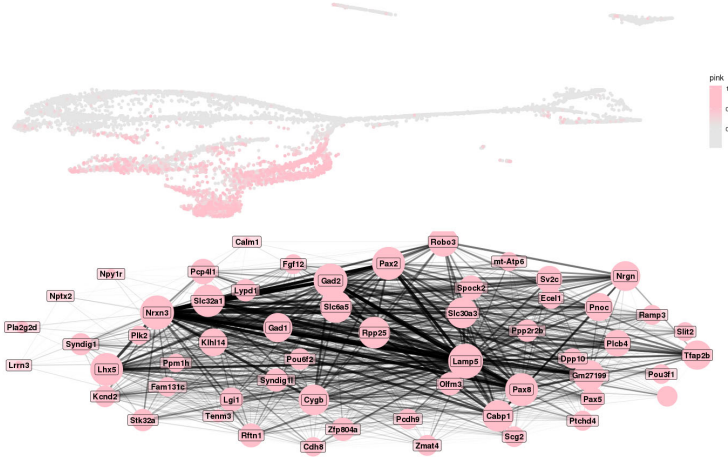
NA

.	.	.	.	.	.	.	.	.	.
Fgfbp3	Dbi	Hopx	Vim	Igfbp5	Hes5	Id3	Aldoc	Sparcl1	Slc1a3
Lxn	Sparc	Pax6	Sox2	Id1	Qk	Mt1	Ttyh1	Phgdh	Sox9
Ptprz1	Tagln2	Acot1	Wnt4	Prdx1	Ndrg2	Nrarp	Mgst1	Wnt7b	Pdpn
Zfp36l1	Rfx4	Hes1	Fgfr3	Fam181a	Foxb1	Plpp3	Gsta4	Lhfp	Notch1
Rcn1	Ddah1	Sox21	Cspg5	Kbtbd11	Gng12	Col2a1	Gpx8	Lfng	Vcam1
Sox6	Sall1	Itgb8	Cpne2	Dtx4	Daam2				

NA

.	.
epithelium development	tube development
central nervous system development	negative regulation of multicellular organismal process
negative regulation of developmental process	epithelial cell differentiation
negative regulation of cell differentiation	glial cell differentiation
tissue development	negative regulation of nervous system development

NA



NA

	module	color	size
	7	pink	57

NA

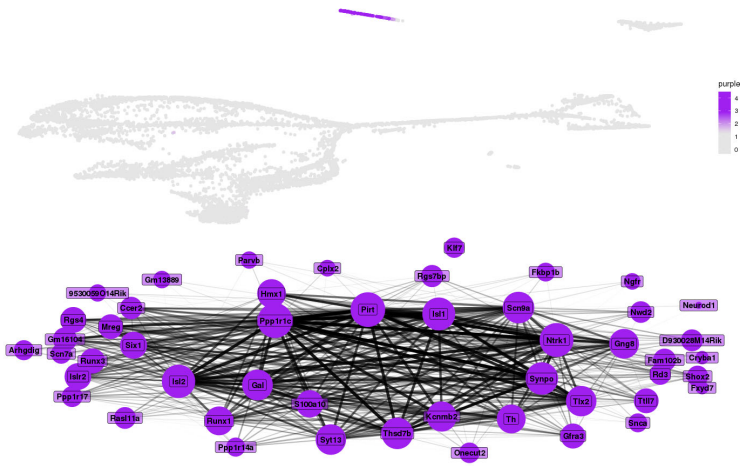
.	.	.	.	.	.	.	.	.	.
Nrgn	Tfap2b	Slit2	Lamp5	Robo3	Pou3f1	Pnoc	Pax2	Pou6f2	Nrxn3
Ramp3	Lypd1	Lhx5	Cabp1	Gad2	Rpp25	Pax8	Slc6a5	Ppp2r2b	Slc32a1
Ppm1h	Klhl14	Ptchd4	Cygb	Pcp4l1	Cdh8	Gm27199	Plcb4	Pax5	Tenm3
Spock2	Syndig1	Zfp804a	Pcdh9	Slc30a3	Rftn1	Ecel1	Lrrn3	Npy1r	Calm1
Nptx2	NA	Lgi1	Gad1	Sv2c	Plk2	Stk32a	Olfm3	Syndig1l	Scg2
Dpp10	Fgf12	Fam131c	Zmat4	Pla2g2d	mt-Atp6	Kcnd2			

NA

.	.
anterograde trans-synaptic signaling	cell-cell signaling
chemical synaptic transmission	collecting duct development
trans-synaptic signaling	apoptotic process involved in metanephric collecting duct development

.	.
synaptic signaling	apoptotic process involved in metanephric nephron tubule development
distal tubule development	kidney field specification

NA



NA

	module	color	size
	8	purple	49

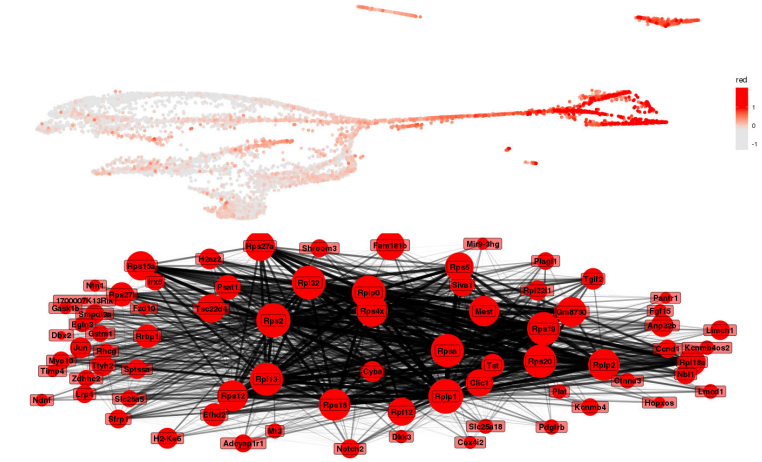
NA

.	.	.	.	.	.	.	.	.	.
Gal	Isl1	Ntrk1	Ppp1r14a	Shox2	Neurod1	D930028M14Rik	Rgs4	Tlx2	S100a10
Ppp1r1c	Scn9a	Ccer2	Fxyd7	Snca	Ppp1r17	Islr2	Kcnmb2	Fkbp1b	Ngfr
Gm16104	Onecut2	Pirt	Six1	Runx1	Thsd7b	Synpo	Runx3	Klf7	Fam102b
Isl2	Gfra3	Hmx1	Syt13	Ttll7	9530059O14Rik	Scn7a	Parvb	Mreg	Arhgdig
Rd3	Nwd2	Rasl11a	Cplx2	Rgs7bp	Gng8	Gm13889	Cryba1	Th	

NA

.	.
behavioral response to pain	regulation of cell development
nervous system process	axon development
neuron development	nervous system development
system process	neuron differentiation
cell development	cell morphogenesis involved in differentiation

NA



NA

	module	color	size
	9	red	75

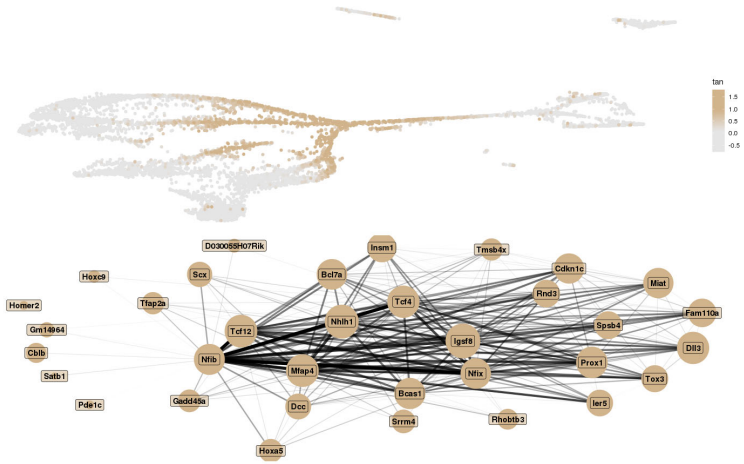
NA

.	.	.	.	.	.	.	.	.	.
Mest	Sptssa	Notch2	Fam181b	Mt3	Gstm1	Ccnd1	Ntn1	Fgf15	Dbx2
H2az2	Psat1	Pantr1	Gask1b	Lrp4	Slc25a5	Rplp0	Nbl1	Ttyh2	Rps12
Rplp1	Anp32b	Sfrp1	Zdhhc2	Rps19	Kcnmb4	Rpl22l1	Plagl1	Timp4	Irx5
Rps4x	Mir9-3hg	Ctnna3	Tsc22d4	Fzd10	Siva1	Shroom3	Clic1	Jun	Gm8730
Rpl32	Limch1	Cyba	Rps18	Rps2	Rpl12	Efhd2	Rhcg	Smpdl3a	Pdgfrb
Slc25a18	Ndnf	Cox4i2	Rps20	Adcyap1r1	Rpl13	Plat	Rps27a	Lmcd1	Tst
1700007K13Rik	H2-Ke6	Hopxos	Myo10	Rpsa	Rpl18a	Rps5	Egln3	Kcnmb4os2	Rps27l
Rps15a	Rrbp1	Dkk3	Rplp2	Tgif2					

NA

.	.
translation	ribosome assembly
peptide biosynthetic process	ribosomal small subunit assembly
amide biosynthetic process	organonitrogen compound biosynthetic process
cellular amide metabolic process	cellular biosynthetic process
peptide metabolic process	biosynthetic process

NA



NA

module	color	size
10	tan	34

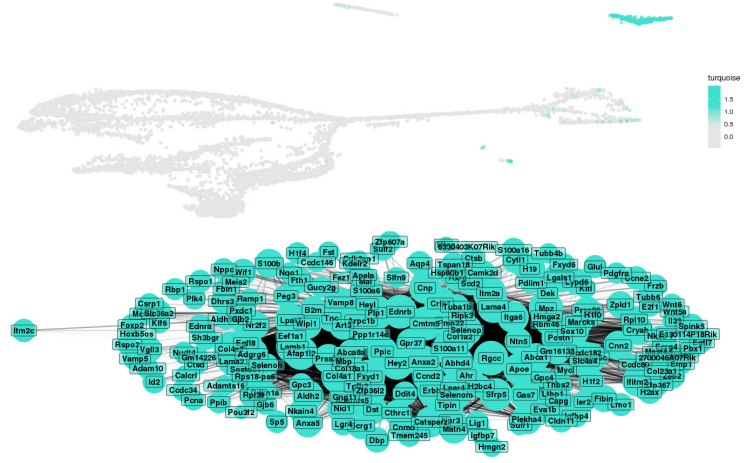
NA

.	.	.	.	.	.	.	.	.	.
Cdkn1c	Mfap4	Tmsb4x	Nfib	Tcf4	Igsf8	Nfix	Scx	Dll3	Nhlh1
Tcf12	Hoxc9	Tfap2a	Bcas1	Insm1	Rnd3	Miat	Prox1	Ier5	Spsb4
Tox3	Gm14964	Pde1c	Bcl7a	Fam110a	Dcc	Cblb	Homer2	Gadd45a	Rhobtb3
D030055H07Rik	Hoxa5	Satb1	Srrm4						

NA

.	.
regulation of RNA metabolic process	positive regulation of nucleic acid-templated transcription
regulation of transcription, DNA-templated	positive regulation of RNA biosynthetic process
regulation of nucleic acid-templated transcription	anatomical structure development
regulation of RNA biosynthetic process	cellular developmental process
positive regulation of transcription, DNA-templated	transcription, DNA-templated

NA



NA

module	color	size
11	turquoise	211

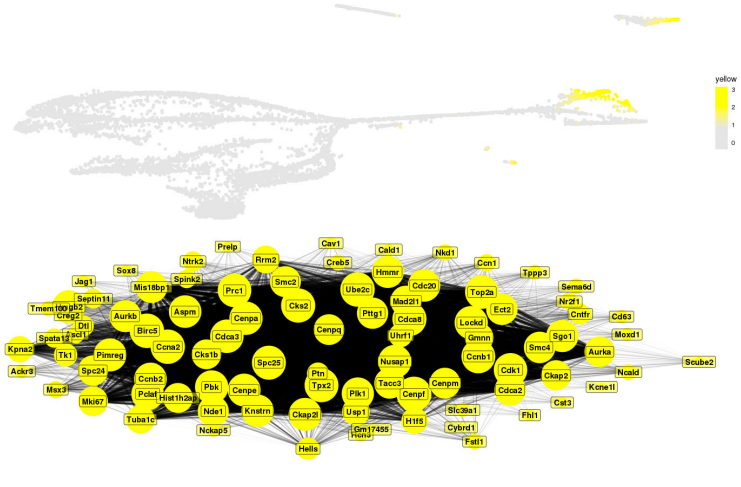
NA

.	.	.	.	.	.	.	.	.	.
Rgcc	Apoe	H19	Ednrb	Selenop	Col1a2	Rbp1	Id2	Plp1	Lgals1
Prss56	Sostdc1	Arpc1b	Col4a1	Foxp2	Meis2	Cryab	Frzb	Sfrp5	Gm14226
Cldn11	Mpz	H2ax	Rspo1	Ctsd	Sulf1	Ccnd2	Egfl8	Nr2f2	Sox10
S100a11	Igfbp7	Col4a2	Hmga2	2700046A07Rik	Nppc	Tuba1b	Slc36a2	Adamts5	Itm2a
Ctsb	Gpr37	Crip1	Postn	Ifitm2	Hoxb5os	Gng11	Rspo3	Nid1	Apela
Cmtm5	Gpc3	Vamp5	Nkx6-1	Anxa2	Zeb2	Sulf2	Ntn5	H1f0	Pdgfra
E130114P18Rik	Fxyd1	Mbp	Hey2	Zpld1	Ppp1r14c	Anxa5	Csrp1	B2m	Lpar4
Aldh1a2	Calcrl	Il33	Hmgn2	Spink5	Wnt6	Fst	6330403K07Rik	Ahr	Sema3c
Mfng	Dhrs3	Igfbp4	S100a6	Egfl7	Tnc	Col18a1	Rpl39	Fez1	Cytl1
Erbb3	Selenoh	Glul	Lamb1	Adgrg6	Tubb6	Pou3f2	S100a16	Cdk2ap1	Eva1b
Camk2d	Adamts16	Slc4a4	Fth1	Aldh2	Fxyd6	Nudt4	Rps18-ps6	Vgll3	Cthrc1
Ripk3	Pbx1	Dek	Sp5	Prss23	Catsperz	Scd2	Zfp36l2	Ddit4	Lmo1
Pmp22	Slfn9	Fibin	E2f1	H1f2	Capg	Abhd4	Rxrg	Ramp1	Nkain4
Tubb4b	Lgr4	Ier2	H2bc4	Ppic	Plk4	Aqp4	Abca8a	Cnn2	Abca1
Art3	Ccdc34	Dst	Tmem245	Lama4	Scrg1	Peg3	Itga6	Afap1l2	Zfp367
Ccdc182	Eef1a1	Gjb2	Kitl	Cnmd	Emp1	Lypd6	Npr3	Gucy2g	Tgfb2
Gjb6	Sh3bgr	Cnp	Wnt5a	Pcna	Adam10	Ccdc80	Mal	Ltbp1	Wipi1
Lama2	Lig1	Ecrg4	Col23a1	Selenom	Heyl	Plekha4	Ccdc146	Fbln1	Thbs2
S100b	Ccne2	Gm16133	Ppib	Dbp	Marcks	Tipin	Pxdc1	Mast4	Matn4
Gpc4	Wif1	Gas7	Zfp607a	Rpl10	Kdelr2	Mycl	Lpar1	Itm2c	Mcm2
Nqo1	Ednra	Vamp8	Pdlim1	Cdkn1a	Tspan18	Nedd1	Rbm46	Hsp90b1	Klf6
H1f4									

NA

.	.
system development	tissue development
anatomical structure development	animal organ development
multicellular organism development	cell proliferation
multicellular organismal process	regulation of cell proliferation
developmental process	regulation of multicellular organismal process

NA



NA

module	color	size
12	yellow	92

NA

.	.	.	.	.	.	.	.	.	.
Ptn	Hmgb2	Hist1h2ap	Ube2c	Pclaf	Cenpf	Top2a	Cks2	Cenpa	Msx3
Ascl1	Ccnb1	Cdc20	Birc5	Spc25	Pttg1	Prc1	Nusap1	Cdk1	Cdca8
Jag1	Rrm2	Ccna2	Pbk	Lockd	Ccnb2	Cald1	Cst3	Pimreg	Tpx2
Cav1	Smc4	Nkd1	Cdca3	Knstrn	Moxd1	Mis18bp1	Cks1b	Nr2f1	Cenpe
Slc39a1	Smc2	Gmnn	Fstl1	Ntrk2	Cd63	Tppp3	Mki67	Aspm	Tmem100
Spata13	Kcne1l	Ackr3	Cenpq	Ckap2l	Plk1	Cenpm	Ect2	Tuba1c	Uhrf1
Cntfr	Nde1	Aurkb	Ncald	Sgo1	H1f5	Ccn1	Cybrd1	Spc24	Dtl
Septin11	Creg2	Kpna2	Mad2l1	Tacc3	Hmmr	Ckap2	Usp1	Sema6d	Cdca2
Hells	Nckap5	Tk1	Aurka	Spink2	Prelp	Fhl1	Scube2	Gm17455	Sox8
Creb5	Rcn3								

NA

.	.
cell division	chromosome segregation
cell cycle	nuclear division
mitotic cell cycle	mitotic nuclear division
mitotic cell cycle process	nuclear chromosome segregation
cell cycle process	organelle fission

NA