Characterising cell fate decision in space and time via transcriptomic data analysis

Kumulative Dissertation
der Fakultät für Biologie
der Ludwig-Maximilians-Universität München

vorgelegt von
Mayra L. Ruiz Tejada Segura

München, den 09.11.2022

Diese Dissertation wurde angefertigt

unter der Leitung von

Prof. Dr. Maria-Elena Torres-Padilla und Dr. Antonio Scialdone

am Institut für Epigenetik und Stammzellen

des Helmholtz Zentrum Münchens

Erstgutachter: Prof. Dr. Maria-Elena Torres-Padilla

Zweitgutachter: Prof. Dr. Wolfgang Enard

Tag der Einreichung: 09.11.2022

Tag der mündlichen Prüfung: 18.07.2023

Eidesstattliche Erklärung

Ich versichere hiermit an Eides statt, dass die vorgelegte Dissertation von mir selbständig und ohne unerlaubte Hilfe angefertigt ist.

München, den 9.11.2022 Mayra L. Ruiz Tejada Segura

(Unterschrift)

Erklärung

Hiermit erkläre ich, *

☒ dass die Dissertation nicht ganz oder in wesentlichen Teilen einer anderen Prüfungskommission vorgelegt worden ist.

☒ dass ich mich anderweitig einer Doktorprüfung ohne Erfolg **nicht** unterzogen habe.

☐ dass ich mich mit Erfolg der Doktorprüfung im Hauptfach ....

und in den Nebenfächern ....

bei der Fakultät für .... der ....

(Hochschule/Universität)

unterzogen habe.

☐ dass ich ohne Erfolg versucht habe, eine Dissertation einzureichen oder mich der Doktorprüfung zu unterziehen.

München, den 9.11.2022 Mayra L. Ruiz Tejada Segura

(Unterschrift)

*) Nichtzutreffendes streichen

Table of Contents

Abbreviations -- **5**

List of Publications --- **6**

List of unpublished manuscripts -- **6**

Declarations of contribution as a co-author --- **7**

Summary -- **9**

Aims --- **10**

Introduction --- **12**

I. Adding spatial dimensions to transcriptional profiles ------------------------ **14**

I.I Placing transcriptomes in the 3D space: The biological relevance of cell types' spatial positions -- 14

I.II Spatial Transcriptomics -- 14

I.III Studying cell identity changes in the olfactory mucosa ---------------- 15

I.IV The chromatographic hypothesis in olfaction ---------------------------- 17

I.V Spatial mapping of Olfr gene expression: a brief summary of previous studies -- 18

I.VI Creating the Olfactory Mucosa 3D transcriptomic atlas ---------------- 19

I.VII An unbiased and quantitative definition of olfactory zones ----------- 20

I.VIII A machine learning algorithm to map the spatial expression of the entire repertoire of *Olfr* genes -- 21

I.IX Deconvolving cell types through single-cell RNA sequencing and TOMO-seq data integration --- 22

II. Studying cell identity transitions as a function of time with single-cell RNA-seq time course experiments -- **24**

II.I Cellular plasticity in the early mouse embryo ------------------------------ 24

II.II Two cell-like cells: an *in vitro* model of cellular totipotency ------------ 24

Results -- **27**

Chapter I. A 3D transcriptomics atlas of the mouse nose sheds light on the anatomical logic of smell -- **28**

Chapter II. Retinoic acid signaling is critical during the totipotency window in early mammalian development --- **61**

Discussion -- **96**

I. A 3D transcriptomics atlas of the mouse nose sheds light on the anatomical logic of smell --- **96**

I.I Interrogating our 3D gene expression atlas of the olfactory mucosa. -- 96

I.II Refining our transcriptional atlas through machine learning ------------ 97

I.III Revisiting the Chromatographic Hypothesis ------------------------------ 97

I.IV scRNA-seq data enabled us to get cell type abundance spatial distributions --- 98

I.V Perspectives -- 98

II. Retinoic acid signaling is critical during the totipotency window in early mammalian development: Insights from single cell transcriptomic profiling 100

II.I Low doses of retinoic acid on embryonic stem cell cultures affect cell state transitions of mouse embryonic stem cells ------------------------------------ 100

II.II Other cell identities induced by a low-dose retinoic acid treatment --- 100

Closing remarks -- **102**

References --- **103**

Acknowledgements -- **110**

Annex I. Curriculum Vitae --- **113**

Annex II. Copyright statements -- **119**

List of abbreviations

Abbreviation	**Definiition**
OSN	Olfactory Sensory Neuron
OM	Olfactory Mucosa
ISH	*In Situ* Hybridization
ISS	*In Situ* Sequencing
OB	Olfactory Bulb
OC	Olfactory Cortex
M/T cells	Mitral/Tufted cells
Olfr	Olfactory Receptor
PCA	Principal Component Analysis
tSNE	t - distributed Stochastic Neighbour Embedding
DPT	Diffusion Pseudo-Time
scRNA-seq	Single-cell RNA sequencing
DWLS	Dampened Weighted Least Squares
NNLS	Non Negative Least Squares
ESCs	Embryonic Stem Cells
2CLCs	2-cell-like-cells
mESCs	mouse Embryonic Stem Cells
RA	Retinoic Acid
CIARA	Cluster-Independent Algorithm for the identification of markers of RAre cell types
smFISH	Single-molecule Fluorescence in situ Hybridization

List of publications

Ruiz Tejada Segura, M. L.*, Abou Moussa, E.*, Garabello, E., Nakahara, T. S., Makhlouf, M., Mathew, L. S., ... & Saraiva, L. R. (2022). A 3D transcriptomics atlas of the mouse olfactory mucosa sheds light into the anatomical logic of smell. *Cell Reports,* 38(12), 110547.

Iturbide, A., **Ruiz Tejada Segura, M. L.***, Noll, C.*, Schorpp, K.*, Rothenaigner, I., Ruiz-Morales, E. R., ... & Torres-Padilla, M. E. (2021). Retinoic acid signalling is critical during the totipotency window in early mammalian development. *Nature Structural & Molecular Biology*, 28(6), 521-532.

Yin, W., Cerda-Hernández, N., Castillo-Morales, A., **Ruiz Tejada Segura, M. L.**, Monzón-Sandoval, J., Moreno-Castilla, P., ... & Gutiérrez, H. (2020). Transcriptional, Behavioural and Biochemical Profiling in the 3xTg-AD Mouse Model Reveals a Specific Signature of Amyloid Deposition and Functional Decline in Alzheimer's Disease. *Frontiers in neuroscience*, 1322.

Huang, S. S., Makhlouf, M., AbouMoussa, E. H., **Ruiz Tejada Segura, M. L.**, Mathew, L. S., Wang, K., ... & Saraiva, L. R. (2020). Differential regulation of the immune system in a brain-liver-fats organ network during short-term fasting*. Molecular metabolism*, 40, 101038.

List of unpublished manuscripts

Gabriele Lubatti, Marco Stock, Ane Iturbide, **Mayra L. Ruiz Tejada Segura**, Richard Tyser, Fabian J. Theis, Shankar Srinivas, Maria-Elena Torres-Padilla, Antonio Scialdone (2022). CIARA: a cluster-independent algorithm for the identification of markers of rare cell types from single-cell RNA seq data. bioRxiv 2022.08.01.501965; doi: https://doi.org/10.1101/2022.08.01.501965

Jitesh Neupane, Gabriele Lubatti, Mayra Luisa Ruiz Tejada Segura, Joao Pedro Alves Lopes, Sabine Dietmann, Antonio Scialdone, M Azim Surani (Unpublished manuscript). Human embryonic organoids reveal origin of primordial germ cells and neuromesodermal progenitors.

* These authors contributed equally to this work

Statement of contribution

I hereby state that my contribution to the publication:

> **Ruiz Tejada Segura, M. L.***, Abou Moussa, E.*, Garabello, E., Nakahara, T. S., Makhlouf, M., Mathew, L. S., ... & Saraiva, L. R. (2022). A 3D transcriptomics atlas of the mouse olfactory mucosa sheds light into the anatomical logic of smell. *Cell Reports,* 38(12), 110547.

consisted in writing an initial version of the manuscript and doing all the data analysis presented here and related figures; except for the clustering of *Olfrs* 3D expression patterns via Topic modeling, which was done by Elisa Garabello, whom I co-supervised during her internship with the Scialdone lab.

Mayra L. Ruiz Tejada Segura
München, August 16th, 2022

Confirmation of contribution

We hereby confirm that the statement of contribution reproduced above is both truthful and accurate and represents a substantial enough contribution to warrant a first-co-authorship.

Prof. Dr. Maria Elena Torres Padilla

Eman Abou Moussa

Dr. Antonio Scialdone

Statement of contribution

I hereby state that my contribution to the publication:

> Iturbide, A., **Ruiz Tejada Segura, M. L.**, Noll, C., Schorpp, K., Rothenaigner, I., Ruiz-Morales, E. R., ... & Torres-Padilla, M. E. (2021). Retinoic acid signalling is critical during the totipotency window in early mammalian development. *Nature Structural & Molecular Biology*, 28(6), 521-532.

consisted in performing the single cell RNA-seq data analysis related to the timeline of the 2-cell-stage-like phenotype induction through retinoic acid and producing the related figures.

Mayra L. Ruiz Tejada Segura
München, August 16th, 2022

Confirmation of contribution

I hereby confirm that the statement of contribution reproduced above is both truthful and accurate.

Dr. Antonio Scialdone Prof. Dr. Maria Elena Torres Padilla

Summary

Cell identities can be described in terms of cell types and states, whose specialised functions are reflected in their transcriptome. The broad spectrum of cell identities that mammals possess varies often gradually across space and time, and some of them are restricted to specific developmental stages and spatial locations in an organism or tissue. Recent advances in molecular biology, like single-cell transcriptome sequencing, provide gene expression profiles of individual cells; therefore allowing a precise transcriptomic characterization of cell types. In this way, single-cell transcriptomics has helped define cell type identity in several biological systems, yet the standard protocols do not provide info about space or time.
Mapping cell identity changes in space and time is fundamental to understanding the functions and mechanisms of specification of cell types
For instance, spatial gradients of gene expression enable organisms to carry out complex biological processes by orchestrating different functions at specific locations in tissues.
In my dissertation, I describe how the analysis of spatial and time-course single-cell transcriptomic data revealed insights into cell fate decisions across space and time in two different contexts: activation of different olfactory receptor genes across the olfactory mucosa and cell state transitions in mouse embryonic stem cell cultures.

Aims

Chapter 1. A 3D transcriptomics atlas of the mouse nose sheds light on the anatomical logic of smell

- Build a tridimensional gene expression atlas of the whole mouse olfactory mucosa.

- Find spatially localised genes and get insight into the spatial aspect of biological processes happening in the olfactory mucosa.

- Define olfactory receptor genes' expression zones in an unbiased and unsupervised way.

- Look into the anatomical logic of the sense of smell through associations between *Olfr* genes' spatial expression patterns and properties of the chemical compounds they detect.

Chapter 2. Retinoic acid signalling is critical during the totipotency window in early mammalian development

- Identify and characterise the cell state changes that mouse embryonic stem cells can undergo in response to treatment with low doses of retinoic acid

- Describe the transcriptional changes driving cells' transition between totipotent-like and pluripotent states.

- Identify new pathways regulating the 2-cell (totipotent-like) state programme.

Introduction

Historically, biologists have tried to deconstruct complex biological systems by breaking them down to their basic unit, cells, and then classifying cells into different cell types according to their phenotype and function [1]. Mammals contain a multitude of distinct cell types, each of which may be composed of multiple cell states. This combination of cell type and state together defines a cell identity (Figure 1). Each of these cell identities has specialised functions [2], which are reflected in the set of genes they express [3]. Therefore new advances in molecular biology, like transcriptome sequencing, have given us the chance to describe cell identities through the expression profiles of thousands of genes [1]. This broad spectrum of cell identities varies gradually across tissues and also along time as an organism develops (Figure 1), which raises fundamental biological questions: How does this cell diversity arise? How do the different types of cells distribute and interact in a tissue, and ultimately, an organism? Although much has been learned, these fundamental questions still captivate us today [4], and they represent the main motivation behind lots of transcriptomic research. Gradual cell identity changes in time are accompanied by transcriptional changes (Figure 1) [5], whereas gene expression spatial gradients can drive tissue formation [6].

This gives us a hint that transcriptomic profiles could be used as a proxy of cell identity, through which cell identity changes in space and time could be analysed.

The following chapters of this dissertation describe how we analyzed the establishment of cell diversity and the spatial distribution of cell (sub)types across a tissue in specific systems, using single cell and spatial gene expression profiling approaches. The first chapter describes our research on the mouse olfactory system, by estimating the spatial distribution of different Olfactory Sensory Neuron (OSN) subtypes in the olfactory mucosa (OM) and the role it plays in smell. The second chapter concerns the investigation of changes in cell identity in a population of mouse embryonic stem cells treated with low doses of retinoic acid.

In the next paragraphs, I'll introduce the two biological systems I studied and the main technologies and computational tools I used.

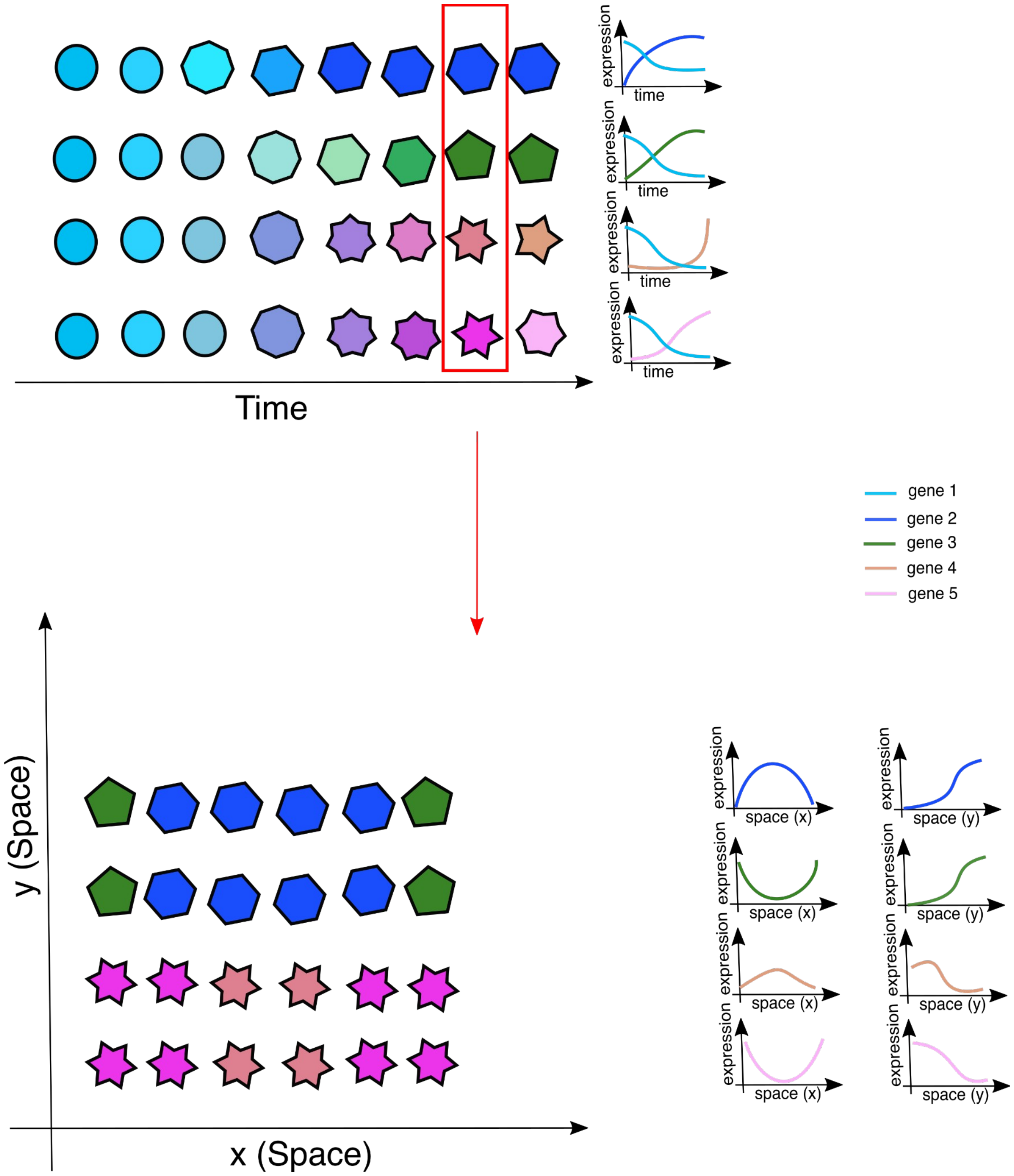


Figure 1. Cell identity changes across space and time. Shapes indicate cell types and colours cell states. The combination of cell type and state together defines a cell identity. Cell identities change gradually along time and across space. Cell identity changes are often accompanied by changes in gene expression.

I. Adding spatial dimensions to transcriptional profiles

I.I Placing transcriptomes in the 3D space: The biological relevance of cell types' spatial positions

Spatial positions of cells allow us to start exploring how the identity of a cell is affected by the types or states of surrounding cells. Cells in different tissue microenvironments express specific sets of genes [7]. As mentioned above, this can be seen during development, with the formation of gene expression gradients along the main embryonic body axes. In the *Drosophila* embryo, for example, 'coordinate' genes determine different embryonic regions along the anteroposterior axis**.** Briefly, a gene product is localised in a specific region of a freshly laid egg. This works as a spatial signal that results in the asymmetrical distribution of transcription factors, which are organized in concentration gradients. These gradients then define the spatial limits of expression of zygotic target genes [6], directing the activation of the correct developmental gene programs needed for the construction of specific organs. In such a way, cell fate decisions are based on spatial relationships between cells. Therefore, cell spatial relationships are key to understanding the properties of individual cells within multicellular organisms [7]. Moreover, many diseases are characterised by abnormal cell type composition of tissues, with some cell types in higher or lower proportions than usual, or some misplaced cell types [8], highlighting the medical importance of cell spatial locations.

Spatial variability of expression levels can also be observed within single organs, even among cells of the same type. An example of this lies in the Olfactory Mucosa. In this tissue, there are ~1100 different subtypes of Olfactory Sensory Neurons, whose transcriptional profile and identity depend on their spatial location [9], according to mechanisms and functions that are still unclear. While transcriptional profiling could help address these open questions, with the standard protocols the spatial information gets lost [10]. This triggered the development of new techniques that preserve spatial information while providing unbiased profiling of the entire transcriptome. These techniques go under the name of spatial transcriptomics, and we describe some of them in the next paragraphs.

I.II Spatial Transcriptomics

The spatial transcriptomics techniques can be divided into subcategories as follows: i) Technologies based on microdissected gene expression profiling, consisting of isolating regions of interest and then performing RNA extraction and sequencing on them. One of the main advantages of these techniques is that they allow whole transcriptome profiling across large pieces of tissues. On the other hand, many of them can not reach single-cell resolution [11] or they miss other features in order to reach it, such as spatial resolution within the region of interest [12] or number of cells characterised [13,14]. ii) In Situ Hybridization (ISH) technologies, where a labelled probe must be hybridised to a complementary target of interest [15]. These methods allow the visualisation of RNA molecules directly in their original environment. However, targets must be known a priori and the computational demand of image processing increases with the number of targets. This puts a limit on the field of view and the number of targets visualised. iii) In Situ Sequencing (ISS) technologies, where RNA sequencing is performed directly on the RNA content of a cell while it remains in its tissue context. This can achieve subcellular resolution in some cases [16], but these approaches are also targeted, and the number of targets that can be analysed is limited by technical constraints as well; iv) in situ capturing technologies, capturing transcripts in situ and then performing sequencing ex situ. These would ideally allow unbiased whole transcriptome analysis; however, RNA capture efficiency gets compromised as resolution increases, detecting often under 10% of the genes [15].

In general, spatial transcriptomics techniques present a trade-off between number of captured genes and spatial resolution. Some techniques based on microdissected gene expression return transcriptome-wide profiles, a big advantage against *in situ* hybridization and some *in situ* sequencing technologies such as seqFISH [17], MERFISH [18] and STARmap [19], which, on the other side, produce higher resolution data, at subcellular levels in some cases [18]. Microdissected gene expression techniques with higher spatial resolution, like Spatial Transcriptomics (1-10 cells) [20] or Slide-seq (1-3 cells) [21], do not need specific gene targets, but they have low RNA capture efficiency, which keeps them from detecting many genes in most cases [20,21]. Low gene capture efficiency from these techniques creates the need for having scRNA-seq data to confirm the identities of the targeted cells as cell type marker genes are often not captured. Interestingly, scRNA-seq data can also be used to deconvolve the signal coming from different cell types when single-cell resolution can not be reached, as in the case of TOMO-seq [15]. For example, TOMO-seq allows the estimation of 3D gene expression profiles of whole tissues and although its spatial resolution is not as high as other techniques', it allows localised whole transcriptome profiling across large tissue pieces and even across some whole tissues [11,15].

I.III Studying cell identity changes in the olfactory mucosa

As mentioned above, the sense of smell relies on a specific spatial distribution of OSN sub-types in the olfactory mucosa. Animals' ability to distinguish and interpret chemical signals in their surroundings through the olfactory system is essential for their survival. Essential activities like finding food, mates or avoiding infection depend on the sense of smell being able to recognize millions of compounds in the environment. This discriminative process starts when, after a sniff, air enters the nasal cavity and reaches the OM, where odours activate Olfactory Receptors located in the cilia of OSNs. Each neuron expresses one *Olfr* gene randomly chosen from an extensive repertoire of about 1100 genes in mice, each of which encodes an Olfactory Receptor protein able to bind to a specific group of ligands. Thus, an OSN can detect a subset of odours depending on the Olfactory Receptor it expresses. Then, as a ligand binds an olfactory receptor in an OSN, a signalling cascade is activated. This makes the neuron fire, transporting information about ligand binding to the Olfactory Bulb through its axon, which is directed to a specific glomerulus according to its active Olfactory Receptor. Glomeruli associated with different Olfactory Receptors are consistently localised in specific areas of the Olfactory Bulb. Thus, each ligand will induce unique spatial patterns of glomerular activation, resulting in the release of neurotransmitters at specific locations in the Olfactory Bulb (Figure 2). Then localised glomerular activation triggers specific behavioural responses [22].

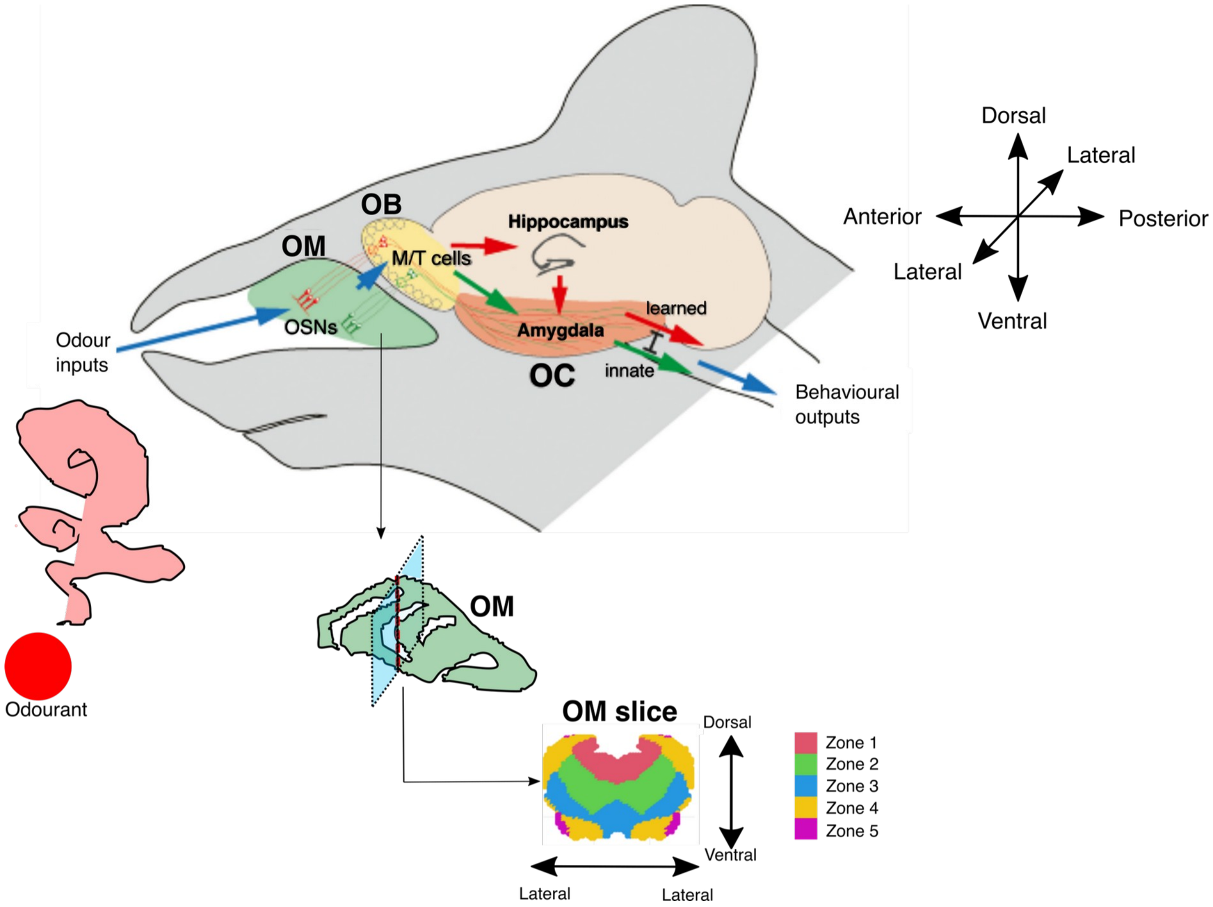


Figure 2. Odour detection (Adapted from [23]). Different odours are detected by OSNs in different zones of the olfactory mucosa, which activates a signalling cascade. As a result, neurons fire, transporting information about odourant binding to specific glomeruli in the Olfactory Bulb according to the active Olfactory Receptors. In turn, glomerular activation causes specific behavioural responses. (OM=Olfactory Mucosa, OSNs=Olfactory Sensory Neurons, OB=Olfactory Bulb, OC=Olfactory Cortex, M/T cells=Mitral/Tufted cells)

The first step in the generation of the specificity of these odour-triggered signals is the choice of the Olfactory Receptor (*Olfr*) gene to express by each Olfactory Sensory Neuron (OSN), which defines the compounds they will be able to detect, as well as a unique transcriptional programme [[22]]. The choice of the *Olfr* gene to activate is random, and the spatial location of the OSN determines the probability of activation of each *Olfr* gene. Thus, in this system, cell identity depends on spatial location. The regulation of the activation of *Olfr* genes is not completely understood. It has been proposed that epigenetic mechanisms regulate *Olfrs*' activation probabilities in the different regions of the OM, resulting into groups of *Olfrs* with similar activation probabilities across OM regions [[24]].

Olfactory receptors were identified as a group of hundreds of proteins that belong to a superfamily of receptors that transduce signals via interactions with G proteins. They are characterised by shared sequence motifs that are not present in the rest of the superfamily; however, this subfamily is still highly diverse, consistent with their hypothesised ability to bind structurally diverse ligands [25]. *Olfrs* are organised in clusters in the genome and it has been observed that *Olfrs* in the same cluster, which are not more than 300kb away from each other, tend to be expressed in similar areas of the Olfactory Mucosa [26]. *Olfrs'* expression is restricted to the OM, specifically to Olfactory Sensory Neurons located in this tissue [25,27].

Since Olfactory Receptors' family was identified [27], the question of how OSNs achieve through them the high discrimination level of odourants observed in mammals was raised. As observations suggested that *Olfrs* were a large family, it was likely that just a small subset of OSNs expressed each *Olfr*. Moreover, it was already known that some genes, mainly *Olfrs*, had particular spatial expression patterns in the OM, which opened the next questions: How are the neuronal subsets expressing each *Olfr* defined? Could they have specific locations such that the olfactory system uses physical space within the OM to encode sensory information? This would imply the presence of a "topographic map" of odours in the OM. Then the olfactory system could employ these maps to discriminate among the numerous odorants. However, no complete map of gene expression in the OM existed, i.e., the spatial location in the OM of only a subset of OSN subtypes was known. This constitutes a strong motivation to apply spatial transcriptomics to explore OSN identity changes across the OM.

I.IV The chromatographic hypothesis in olfaction

The link between OSNs' spatial location and function -the anatomical logic of smell- is still a longstanding question. Could it be that the locations where different odourants are detected contribute to the high discriminatory power that characterizes mammalian olfaction?

In an attempt to answer this question, electrophysiological observations of the detection of 15 compounds in different regions of the OM were done [28]. The authors of this work confirmed that odorants caused different neuronal activity intensities at the medial and lateral regions of the OM; and that the activity intensity ratios between regions varied between odorants [28]. They also noticed that the elapsed time between the onset of OSN activity in the two regions varied depending on the odorant used. Therefore, they suggested that these observations could be explained by the rate at which different molecules migrate across the OM. Being this the case, the OM could separate different molecules by their ability to move across it. Thus we could think that this separation follows the same principle as chromatography, which also takes advantage of differential molecular migration caused by differential affinity of molecules to the medium through which they pass [28]. This hypothesis is known as "the chromatographic hypothesis" and remains to be tested as it is based on observations of a small group of odorants.

Olfr genes were not known when the hypothesis was formulated. Thus, this odour perception model has been kept independent from *Olfrs'* spatial expression patterns. Later, it has been speculated that the spatial distribution of OSNs sub-types and their associated *Olfrs* might reflect the spatial patterns of migration of the odourants they detect, which could improve odour discrimination [29].

In the first part of this dissertation, I provide a reconstruction of 3D gene expression patterns for thousands of genes, including most of *Olfrs*. As one application of this map, I use it to quantitatively test for the first time the chromatographic hypothesis of olfaction.

I.V Spatial mapping of Olfr gene expression: a brief summary of previous studies

Olfr genes' expression across the OM started being assessed via *in situ* hybridization experiments, which demonstrated that they follow topographically distinct expression patterns in this tissue, shaped as concentric rings which are bilaterally symmetrical in the 2 nasal cavities [25]. As the thought that the diversity of *Olfrs* expression patterns could be related to the diversity of detected ligands became popular, attempts to profile and classify *Olfrs'* spatial patterns started.

In order to represent the diversity of *Olfr* patterns in the OM and classify them, researchers divided the OM into discrete zones where different *Olfrs* are expressed. Four was the initial number of zones identified, as the expression area of each of 31 *Olfrs* profiled back then covered about one-quarter of the total surface area of the OM [25]. Then, the finding of overlaps between *Olfrs* spatial expression patterns triggered the idea of using a continuous index to classify patterns of *Olfr* expression. Thus, indexes were assigned to *Olfrs* according to fraction of OSNs expressing them in each of the previously defined four contiguous zones where they can be found [9]. For example, an *Olfr* gene with an index of 1.5 signifies that roughly one-half of the OSNs expressing it is estimated to be in zone 1 and the other half in zone 2 [9]. While the definition of indexes allowed a more precise classification of the *Olfr* expression patterns, they were affected by at least two major limitations: first, the initial definition of the zones and their number was chosen in a rather arbitrary way; second, the analysis was based on the spatial expression profiles of less than 10% of the existing *Olfrs*.

More recently, to profile the spatial expression of more *Olfrs*, Tan and Xie performed RNA-seq on a single sample of OM sectioned in 12 pieces along the dorso-medial axis [26]. Despite including information about thousands of genes, this dataset generated a low-resolution spatial map of gene expression along a single axis, assuming that the 3D expression patterns of *Olfrs* are entirely determined by their patterns along the dorso-medial axis. Therefore this data would also not allow an unsupervised definition of zones. Nevertheless, this experiment allowed assigning a spatial index to 1033 *Olfrs* based on the similarity between their patterns and the patterns of *Olfrs* indexed by Miyamichi et al. [9]. However, *Olfrs* described by Miyamichi et. al follow all ring shaped patterns. Hence, genes whose patterns resulted inconsistent with those profiled by Miyamichi et al., like *Olfr459* [9], could not get an index.

Another attempt to better characterise *Olfr* genes' expression patterns was done by Mombaerts et. al. in 2019 [30]. They used three colour fluorescence in situ hybridization, semi-automated image segmentation and 3D reconstruction to map 68 *Olfrs* in 3D and OM-wide. They qualitatively discerned 9 overlapping zones of expression where these *Olfrs* were expressed. However, they were still looking at a very small subset of *Olfr* genes. Thus the main limitation until now in order to get to the description of the whole spectrum of *Olfr* expression and, thus, of OSN identities has been the small number of characterized genes and/or the limited spatial resolution achieved [9,26,30].

One of the main goals of my work described in the first chapter of this thesis was to profile the spatial expression pattern of all genes within the mouse OM in 3D by using spatial transcriptomics.

I.VI Creating the Olfactory Mucosa 3D transcriptomic atlas

For our project, we chose a spatial transcriptomic technique that would allow us to profile as many genes as possible in the 3D space, even if not at the single-cell level. Thus, our collaborators from the Saraiva group at Sidra Medicine (Doha, Qatar) carried out the OM spatial transcriptomic profiling via TOMO-seq [11], a technique which is based on performing RNA-seq on cryosections of a tissue of interest cut along the main body axes (Figure 3). The transcriptomic profiling of these cryosections gives information on the expression patterns of genes along single spatial axes. These uni-dimensional patterns can then be computationally combined to yield a 3D transcriptional profile of the tissue, by using algorithms such as the Iterative Proportional Fitting [15].

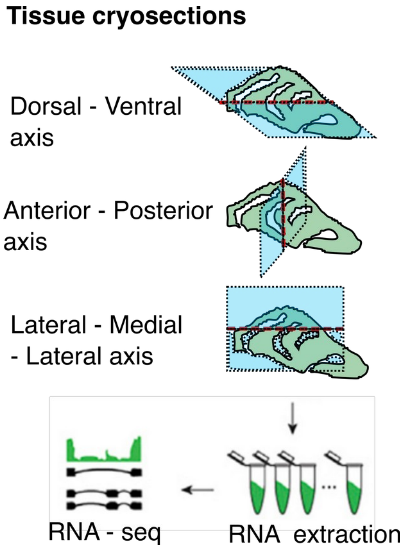


Figure 3. Schematic illustration of the TOMO-seq protocol applied to the olfactory mucosa. Tissues were cryosectioned along the three main body axes; RNA from each slice of tissue was extracted and sequenced.

This transcriptome-wide approach reaches an RNA capture efficiency comparable to bulk RNA-seq, giving us the possibility to notably increase the number of spatially characterised

Olfrs [31]. Furthermore, a 3D characterization of gene expression in the OM would serve as a starting point to describe biological processes taking place in the OM, like toxin detection, from a spatial perspective; and starting from these data, a more unbiased, quantitative definition of *Olfrs* zones can be achieved (see section below).
Overall, our work resulted in the first 3D transcriptome atlas of the Olfactory Mucosa, as described in the first chapter of this dissertation. And to facilitate the access of this resource to the scientific community, I built a web app where the data can be downloaded and explored in 3D, available from the website http://atlas3dnose.helmholtz-muenchen.de:3838/atlas3Dnose **.**

I.VII An unbiased and quantitative definition of olfactory zones

With TOMO-seq, we could detect and robustly profile the spatial patterns of ~50% (689) of *Olfrs*. Our next goal was to use them for an unbiased definition of Olfrs' zones of expression, grouping *Olfr* genes according to spatial expression pattern similarity. As mentioned before, *Olfrs'* expression patterns overlap in space, following a continuous distribution across the OM; such continuous distribution is an aspect that should be taken into account when defining and describing the zones.

Frequently, clustering algorithms are applied to gene expression data to find genes whose expression values change in a similar way across different samples, but most of these algorithms [32–34] divide data into discrete clusters, with each element assigned to a single cluster.
With our data, we decided to use Topic Modelling, an approach originally designed for text mining. In the original context, this method receives as input a set of documents with the aim to describe them based on the topics they contain. In order to achieve this, the contained topics must be inferred. Then, the distributions of word frequencies representing the different topics, as well as the distributions of topic proportions describing each document are obtained as output [35]. This is particularly convenient for us to represent the fact that *Olfr* genes can be expressed in more than one zone. So with the goal to describe *Olfrs* spatial expression trends in terms of proportions of expression in different zones, we used the reconstructed 3D expression patterns of *Olfrs* as inputs for a Topic Modelling algorithm. As output, we obtained an unbiased definition of the zones (defined as probability distributions over the spatial coordinates) and a decomposition of *Olfr* expression patterns in terms of the zones (through the “degrees of belonging”, representing a quantification of how much a given expression pattern fits in each zone).

Based on the behavior of the likelihood as a function of the number of topics, we estimated the existence of at least five *Olfr* expression zones. Then, each *Olfr* spatial pattern could be identified by five numbers representing its degrees of belonging to the five zones. The observation of continuous overlap between contiguous *Olfr* genes' expression areas brought the representation of the spatial expression pattern of the 82 profiled *Olfrs* through a continuous index [9]. As we extended the fraction of spatially characterized *Olfrs* repertoire, we observed that *Olfrs* expression patterns could be described by only a subset of all possible combinations of zones. In particular, we showed that each *Olfrs* expression pattern could be represented by a single number. To show this, we performed dimensionality reduction on our *Olfrs* spatial expression patterns represented by the degrees of belonging to the five zones we inferred using diffusion maps [36].

Diffusion maps have recently gained popularity among dimensionality reduction techniques due to their accuracy at modelling continuous trajectories in biological processes such as the gradual transcriptional changes cells undergo during differentiation [36]. This technique models the state transitions involved in biological processes via diffusion dynamics. In this model, elements can randomly diffuse from their position, which represents a state, through an isotropic Gaussian wave function. So the transition probability from state x to state y is defined by the interference of the two wave functions Y(x) and Y(y). In this way, states can be ordered according to transition probabilities and visualised in a manifold defined by eigenvectors of the transition probability matrix. Given its ability to order states, this dimensionality reduction method has the advantage of preserving the nonlinear structure of data as a continuum [36]. This makes it optimal for representing nonlinear continuous processes in comparison with other dimensionality reduction techniques, like principal component analysis (PCA) or t - distributed Stochastic Neighbour Embedding (tSNE), which are based on linear or clustering methods [37,38].

When applied to *Olfrs* spatial expression trends, the diffusion map showed a continuous distribution of *Olfrs* along a uni-dimensional curve, which hinted to the possibility of simplifying their representation to a single coordinate tracking the position of each *Olfr* along this curve. We defined this coordinate, which we named “3D index”, via a method called diffusion pseudo-time (DPT) [39]. DPT is a random-walk-based distance that is computed based on Euclidean distances in the 'diffusion map space'. So, this technique orders *Olfrs* according to the similarities of their degrees of belonging using diffusion-like random walks [39]. While the 3D index correlated well with the previously defined expression index [9], its definition is fully quantitative and allows the description of the spatial patterns of a much larger set of *Olfrs*.

I.VIII A machine learning algorithm to map the spatial expression of the entire repertoire of *Olfr* genes

As mentioned above, our TOMO-seq experiment allowed us to characterise the 3D spatial expression pattern of 689 *Olfrs*, which corresponds to approximately half of the mouse *Olfr* repertoire. The rest of the *Olfr* genes were either too lowly expressed or not detected at all in our data. Given that previous studies have demonstrated that machine learning models can estimate gene expression values using genomic sequences and features (eg, k-mer frequencies, GC content…), we decided to try such approaches to get some insight about the expression of the *Olfr* genes we could not detect. In previous studies, machine learning models predicting gene expression values were trained on features such as the separation between the coding and the regulatory sequence of a gene, codon frequency and other important features for gene expression regulation [40,41]. For *Olfr* genes, in addition to their genomic features (e.g., position of genomic cluster, gene length, etc), we also know the positions of many loci acting as enhancers, which are called Greek islands [42,43]. So, we decided to verify whether the genomic features of *Olfrs* and those of their known enhancers could predict *Olfrs* zonal expression via machine learning algorithms.

Machine learning models can be trained to predict gene expression based on diverse features. Deep learning architecturesdecode information directly from genomic sequences, which shallow models can not do. However they lack interpretability and addressing questions about the importance of specific features takes a full model training process without the tested feature(s). [41]. On the other hand, when predictions are made using a few non-sequence features, standard machine learning models perform well and the importance of specific features is retrievable [41].
In our case we had around ten non-sequence features to predict on, so we decided to try a standard machine learning approach, which would also allow us to ask questions about the importance of the features in the predictions. Random forest models keep a feature in the model if its absence significantly affects the prediction, meaning that feature importance is intrinsically calculated by the model. This was a big advantage for us, given that it allowed us to compare the importance of different genomic features for spatial expression patterns.
By combining our TOMO-seq data with machine learning methods, using the 689 *Olfrs* we profiled as a training dataset, we were able to spatially characterise nearly all *Olfr* genes (N= 1378, ~98%) in the mouse.

I.IX Deconvolving cell types through single-cell RNA sequencing and TOMO-seq data integration

TOMO-seq does not achieve single-cell resolution. Thus, conclusions about the location of different cell types in the OM could not be made via this dataset. Therefore we thought of combining our TOMO-seq data with previously published single-cell RNA-seq data to estimate a 3D map of cell type composition of the OM.

Single-cell RNA sequencing (scRNA-seq) provides gene expression profiles of individual cells and allows a precise transcriptomic characterization of cell types [1], even those that are rare [44]. On the other hand, as tissues must be dissociated to collect single cells for scRNA-seq, this kind of data lacks information about spatial relationships among cells [10].

The idea to integrate data from different sequencing methods through common features has become popular as these combined datasets have revealed novel insights that would not be found using one single method [45]. In this case, in order to find how different cell types were distributed across the OM, we needed to get reference transcriptomic profiles for different cell types from scRNA-seq data. Then we would use them to deconvolve the transcriptomic signal from every slice of tissue obtained with TOMO-seq to estimate the cell-type composition.

Several Machine learning methods have been recently published to deconvolve bulk transcriptomic profiles using the transcriptomic profiles of single cell types derived, eg, from scRNA-seq data. One of the first steps for cell type deconvolution is the choice of the subset of features (i.e., genes) to use in the algorithm. This can be done in supervised and unsupervised ways. Supervised approaches rely on predefined signature matrices containing expression data of known marker genes of well-defined cell types. scRNA-seq data contains data from many more genes, which has opened the possibility of having better descriptions of specific cell types and even discovering new ones [7]. So in order to make the most of scRNA-seq data to define cell type-specific signatures, we decided to use an unsupervised approach.

Different unsupervised feature selection techniques have been used for this task. The most common ones select genes according to a single criterion, such as highest expression, coefficient of variation or q values from t-tests comparing different cell types [46,47]. Cell type deconvolution using cell type transcriptomic profiles obtained using such gene selection techniques has shown consistency with previous knowledge on cell types location [48]; however, the addition of known marker genes to those cell type profiles has improved the deconvolution results [49], pointing to the possibility of improving feature selection strategies. Moreover, the gene selection methods mentioned before would sometimes select cell type unspecific genes [50].

Due to these issues, we decided to adopt a gene selection strategy called AutogeneS that minimises the correlation among cell type profiles and maximises the euclidean distance among them [50]. This unsupervised approach has been tried on different datasets where the ground truth is known, performing almost as well as supervised approaches based on known marker genes [50].

Once the feature selection is done, the deconvolution algorithm can be run on the data. Typically, deconvolution algorithms assume that gene expression signals from a cell type in a mixed sample are proportional to the fraction of this cell type in the mixture [50,51]. This, in turn, means that linear models could describe mixed transcriptional profiles in terms of fractions of gene expression signals coming from different cell types' transcriptional profiles. Different types of models have been used for this purpose, like Dampened Weighted Least Squares (DWLS), Non Negative Least Squares (NNLS), and diverse penalised linear regression models like Support Vector Machines [50,52]. These approaches have shown consistent performance when cell type profiles are based on the same list of genes [48].

In the first chapter of this dissertation, we integrated scRNA-seq data from [53] with our TOMO-seq data and estimated the spatial distribution of different cell types across the OM through cell type deconvolution analysis. Specifically, we built cell type profiles using scRNA-seq data and used them to generate a model that describes gene expression from TOMO-seq data in terms of proportions of RNA coming from different cell types.

II. Studying cell identity transitions as a function of time with single-cell RNA-seq time course experiments

II.I Cellular plasticity in the early embryo

Single-cell transcriptomics has been extensively applied in developmental biology. In particular, adding the time dimension to single-cell transcriptomic data has allowed the study of cell state trajectories that generate the different cell types [54,55] and the loss of cellular plasticity involved in this process [55].

In mammals, only cells in the earliest embryonic pre-implantation stages can autonomously form a whole organism and therefore give rise to any cell type. This property is called totipotency and it is lost as development progresses and cellular plasticity is gradually reduced until cells reach a differentiated state. In mice, the totipotent window of embryos is limited to the zygote and the 2-cell stage embryo (Figure 4). After this window, cells commit to two different fates: the embryonic cell lineage (inner cell mass), marked by the presence of OCT4, SOX2 and NANOG, and the GATA4/6+ or CDX2+ extraembryonic cell lineages [56]. Cells from the inner cell mass can give origin to any cell in the embryonic lineage, but not in the extraembryonic lineages, thus, these cells are not totipotent anymore but pluripotent [57].

Embryonic Stem Cells (ESCs), derived from the inner cell mass, can stay pluripotent in culture; so they have great potential in regenerative medicine. Moreover, pluripotent stem cells have been successfully induced by manipulating the transcriptional and epigenetic networks of various differentiated cell types [58,59]. The possibility of pluripotency induction and maintenance in culture has facilitated the study of cell state trajectories from the pluripotent state to differentiated states. However, totipotency and the factors that confer the ability to give rise to cells in both embryonic and extraembryonic lineages remained poorly understood.

II.II Two cell-like cells: an *in vitro* model of cellular totipotency

In ESC cultures, the presence of rare groups of cells resembling the blastomeres of 2-cell stage embryos has been observed [56]. These cells, referred to as ‘2-cell-like-cells’ (2CLCs), share many features with the two cell stage embryos such as the expression of genes like *Zscan4,* and retrotransposons from the MERVL family. Furthermore, they lack pluripotency proteins OCT4, SOX2, and NANOG, and have the capacity to contribute to extraembryonic tissues [56]. 2CLCs are considered totipotent-like cells, and a powerful model to study totipotency-related molecular features. Therefore, identifying conditions that can induce and maintain 2CLCs in culture can facilitate their isolation and transcriptional profiling. This, in turn, can help us uncover key factors involved in the onset of totipotency and the underlying gene regulatory networks [60].

In the second chapter of this thesis, I show how we found that a treatment with low doses of retinoic acid can promote the transition of mouse ESC (mESC) into 2CLCs. In particular, I analysed scRNA-seq datasets of mESC at 0 hours, 2 hours, 12 hours and 48 hours after retinoic acid treatment, which led to the characterization of the transcriptional dynamics that accompanies the transition of mESC into 2CLC or precursor cells.

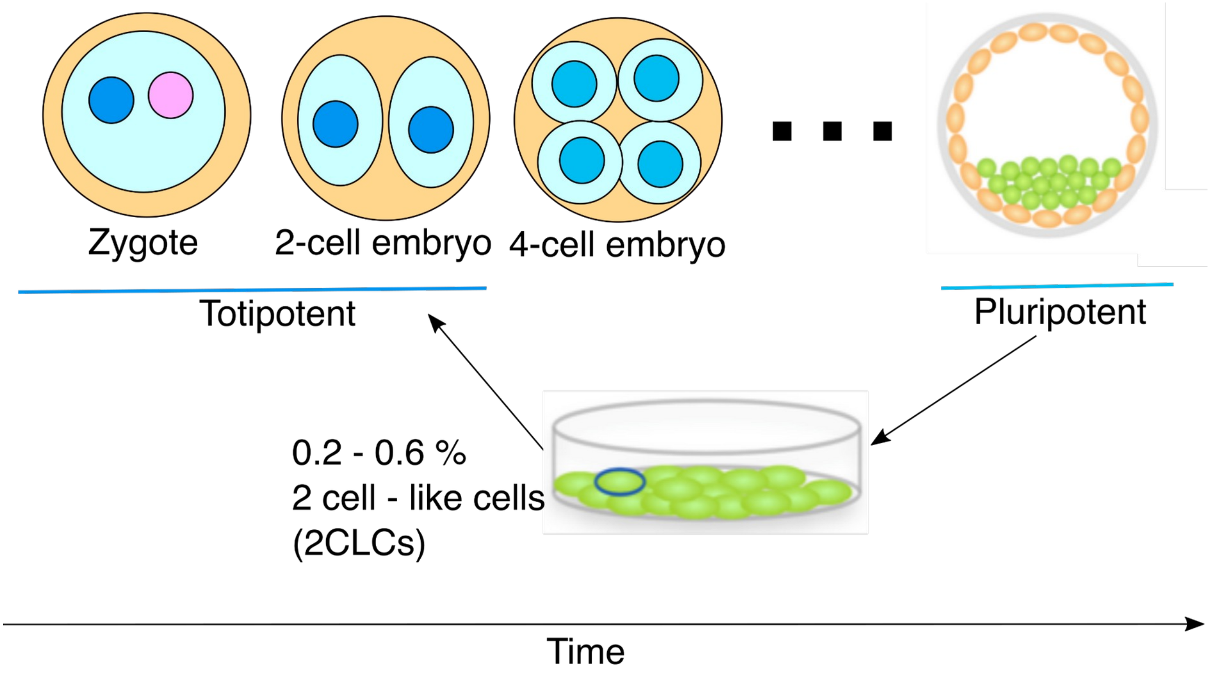


Figure 4. Schematics of totipotency-pluripotency transition in mouse embryos. 2-cell-like-cells arise spontaneously in mouse pluripotent stem cell cultures. (adapted from https://www.helmholtz-munich.de/ies/news-and-events/news/news/article/48544/index.html)

Overall, the two chapters presented in this cumulative dissertation highlight the importance of integrating space and time in gene expression data analysis, which allows investigating how cell identities are distributed across tissues and how they can dynamically vary following a perturbation. Transcriptional features associated with specific cell types have turned essential for cell type description. And looking at transcriptional changes related to cell identity dynamics in time and space could be crucial for addressing further unsolved questions, like the formation and distribution of different cell types in biological contexts as diverse as embryonic development and olfaction.

Results

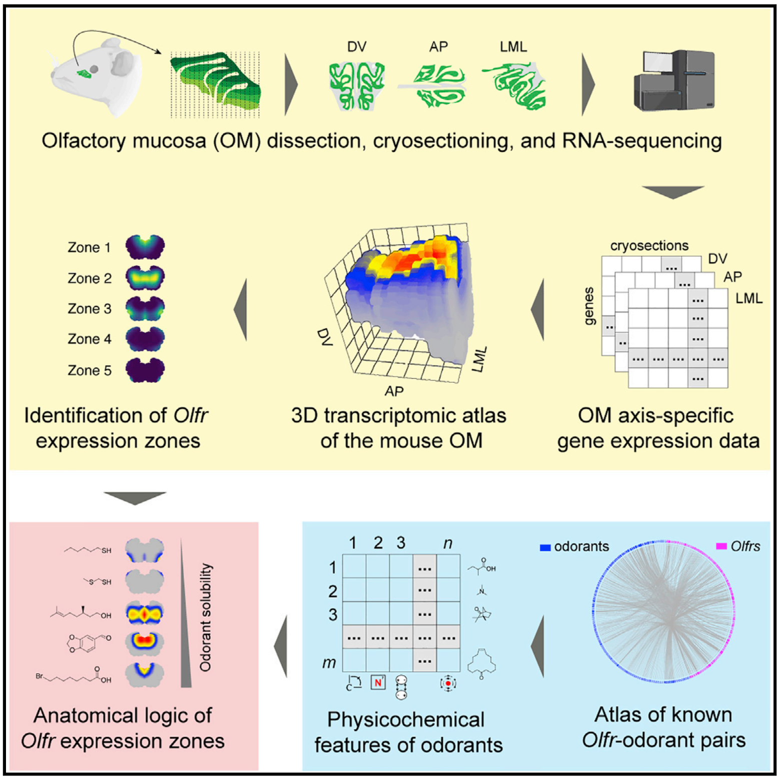
Chapter I. A 3D transcriptomics atlas of the mouse nose sheds light on the anatomical logic of smell

Cell Reports

Article

A 3D transcriptomics atlas of the mouse nose sheds light on the anatomical logic of smell

Graphical abstract


DV
AP
LML
Olfactory mucosa (OM) dissection, cryosectioning, and RNA-sequencing
Zone 1
Zone 2
Zone 3
Zone 4
Zone 5
Identification of *Olfr* expression zones
DV
AP
LML
3D transcriptomic atlas of the mouse OM
cryosections
DV
AP
LML
genes
OM axis-specific gene expression data
Odorant solubility
Anatomical logic of *Olfr* expression zones
1 2 3 n
1
2
3
m
Physicochemical features of odorants
odorants
Olfrs
Atlas of known *Olfr*-odorant pairs

Highlights

- We generate a browsable 3D transcriptomic atlas of the mouse olfactory mucosa (OM)
- We identify potential functional hotspots in the mouse OM
- Odorant receptor genes (*Olfrs*) are continuously distributed over at least five zones
- Spatial locations of *Olfrs* correlate with the mucus solubility of their ligands

Authors

Mayra L. Ruiz Tejada Segura,
Eman Abou Moussa, Elisa Garabello, ...,
Bettina Malnic, Antonio Scialdone,
Luis R. Saraiva

Correspondence

antonio.scialdone@helmholtz-muenchen.de (A.S.),
saraivalmr@gmail.com (L.R.S.)

In brief

Ruiz Tejada Segura et al. employ a spatial transcriptomics approach to create a 3D map of gene expression of the mouse nose and combine it with single-cell RNA-seq, machine learning, and chemoinformatics to resolve its molecular architecture and shed light into the anatomical logic of smell.

Ruiz Tejada Segura et al., 2022, Cell Reports *38*, 110547
March 22, 2022 © 2022 The Authors.
https://doi.org/10.1016/j.celrep.2022.110547

CellPress

Cell Reports

CellPress
OPEN ACCESS

Article

A 3D transcriptomics atlas of the mouse nose sheds light on the anatomical logic of smell

Mayra L. Ruiz Tejada Segura,[1,2,3,12] Eman Abou Moussa,[4,12] Elisa Garabello,[1,5,6] Thiago S. Nakahara,[7] Melanie Makhlouf,[4] Lisa S. Mathew,[4] Li Wang,[4] Filippo Valle,[5] Susie S.Y. Huang,[4] Joel D. Mainland,[8,9] Michele Caselle,[5] Matteo Osella,[5] Stephan Lorenz,[4,10] Johannes Reisert,[8] Darren W. Logan,[10] Bettina Malnic,[7] Antonio Scialdone,[1,2,3,13,*] and Luis R. Saraiva[4,8,11,13,14,*]

[1]Institute of Epigenetics and Stem Cells, Helmholtz Zentrum München, Feodor-Lynen-Strasse 21, 81377 München, Germany
[2]Institute of Functional Epigenetics, Helmholtz Zentrum München, Ingolstädter Landstraße 1, 85764 Neuherberg, Germany
[3]Institute of Computational Biology, Helmholtz Zentrum München, Ingolstädter Landstraße 1, 85764 Neuherberg, Germany
[4]Sidra Medicine, P.O. Box 26999, Doha, Qatar
[5]Physics Department, University of Turin and INFN, Via P. Giuria 1, 10125 Turin, Italy
[6]Department of Civil and Environmental Engineering, Cornell University, Ithaca, NY 14853, USA
[7]Department of Biochemistry, University of São Paulo, São Paulo, Brazil
[8]Monell Chemical Senses Center, 3500 Market Street, Philadelphia, PA 19104, USA
[9]Department of Neuroscience, University of Pennsylvania, Philadelphia, PA 19104, USA
[10]Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, UK
[11]College of Health and Life Sciences, Hamad Bin Khalifa University, P.O. Box 34110, Doha, Qatar
[12]These authors contributed equally
[13]These authors contributed equally
[14]Lead contact
*Correspondence: antonio.scialdone@helmholtz-muenchen.de (A.S.), saraivalmr@gmail.com (L.R.S.)
https://doi.org/10.1016/j.celrep.2022.110547

SUMMARY

The sense of smell helps us navigate the environment, but its molecular architecture and underlying logic remain understudied. The spatial location of odorant receptor genes (*Olfrs*) in the nose is thought to be independent of the structural diversity of the odorants they detect. Using spatial transcriptomics, we create a genome-wide 3D atlas of the mouse olfactory mucosa (OM). Topographic maps of genes differentially expressed in space reveal that both *Olfrs* and non-*Olfrs* are distributed in a continuous and overlapping fashion over at least five broad zones in the OM. The spatial locations of *Olfrs* correlate with the mucus solubility of the odorants they recognize, providing direct evidence for the chromatographic theory of olfaction. This resource resolves the molecular architecture of the mouse OM and will inform future studies on mechanisms underlying *Olfr* gene choice, axonal pathfinding, patterning of the nervous system, and basic logic for the peripheral representation of smell.

INTRODUCTION

The functional logic underlying the topographic organization of primary receptor neurons and their receptive fields is well known for all sensory systems but olfaction (Kandel et al., 2013). The mammalian nose is constantly flooded with odorant cocktails. Powered by a sniff, air enters the nasal cavity until it reaches the olfactory mucosa (OM). There, myriad odorants activate odorant receptors (Olfrs) present in the cilia of olfactory sensory neurons (OSNs), triggering a cascade of events that culminate in the brain and result in odor perception (Buck and Axel, 1991; Kandel et al., 2013). Most mouse mature OSNs express a single allele of one out of ~1,100 Olfr genes (*Olfrs*) (Chess et al., 1994; Hanchate et al., 2015; Malnic et al., 1999; Saraiva et al., 2015b). Olfrs employ a combinatorial strategy to detect odorants, which maximizes their detection capacity (Malnic et al., 1999; Nara et al., 2011). OSNs expressing the same *Olfr* share similar odorant response profiles (Malnic et al., 1999; Nara et al., 2011) and drive their axons to the same glomeruli in the olfactory bulb (Mombaerts et al., 1996; Ressler et al., 1994; Vassar et al., 1994). Thus, *Olfrs* define functional units in the olfactory system and function as genetic markers to discriminate between different mature OSN subtypes (Ibarra-Soria et al., 2017; Saraiva et al., 2015b).

Another remarkable feature of the OSN subtypes is their spatial distribution in the OM. Early studies postulated that OSNs expressing different *Olfrs* are spatially segregated into four broad areas within the OM, called "zones," and which define hemicylindrical rings with different radii (Ressler et al., 1993; Vassar et al., 1993). Subsequent studies identified *Olfrs* expressed across multiple zones, making clear that a division in four discrete zones might not accurately reflect the system, and a continuous numerical index representing the pattern of expression of each *Olfr* along the zones was implemented

This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

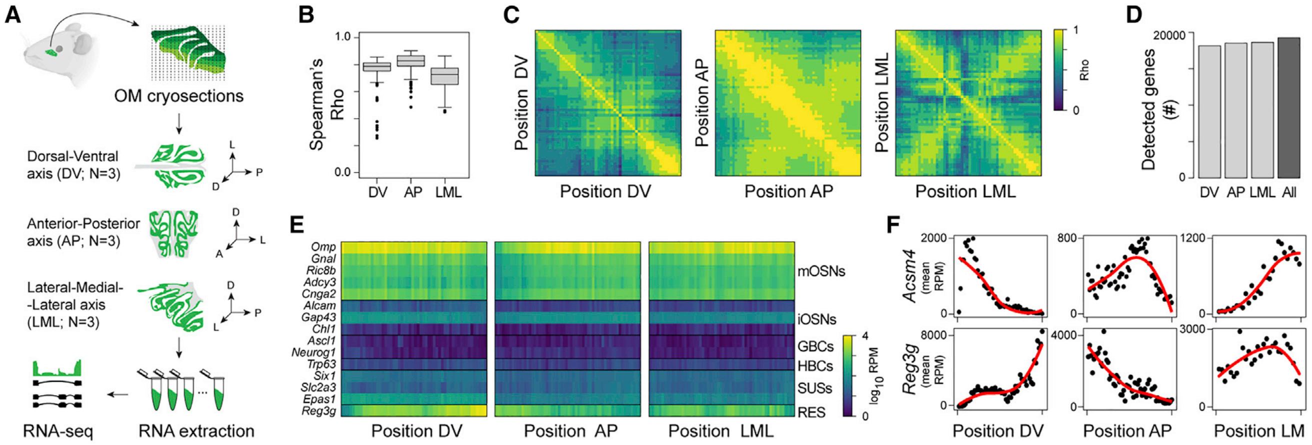


Figure 1. Application of TOMO-seq to mouse OM

(A) Experimental design. TOMO-seq was performed on nine tissue samples, from which three were sliced along the dorsal-ventral axis (DV), three along the anterior-posterior axis (AP), and three along the lateral-medial-lateral axis (LML).

(B) Boxplots showing the distributions of Spearman's correlation coefficients (rho) between replicates in each axis.

(C) Heatmaps showing Spearman's correlation between gene expression patterns at different positions along the three axes.

(D) Number of detected genes along each axis separately or across the whole dataset. Genes were considered as detected when they had at least one normalized count in at least 10% of the samples from one axis.

(E) Heatmaps of $\log_{10}$ normalized expression (after combining the three replicates per axis) of OM canonical markers along the three axes (GBCs, globose basal cells; HBCs, horizontal basal cells; iOSNs, immature olfactory sensory neurons; mOSNs, mature olfactory sensory neurons; RESs, respiratory epithelium cells; RPM, reads per million; SUSs, sustentacular cells).

(F) Normalized expression of canonical OM spatial marker genes along the three axes. Red line shows fits with local polynomial models.

(Miyamichi et al., 2005; Strotmann et al., 1992). More recently, a study reconstructed *Olfr* expression patterns in three dimensions (3D) and qualitatively classified the expression areas of 68 *Olfrs* in nine overlapping zones (Zapiec and Mombaerts, 2020). However, all these studies sampled a fraction (~10%) of the total intact olfactory receptor gene repertoire and, most importantly, lack a quantitative and unbiased definition of zones or indices. We do not currently understand the full complexity of the OM and lack an unbiased and quantitative definition of zones. In effect, the exact number of zones, their anatomical boundaries, molecular identity, and functional relevance are yet to be determined.

One hypothesis is that the topographic distribution of *Olfr* and OSN subtypes evolved because it plays a key role in the process of *Olfr* choice in mature OSNs and/or in OSN axon guidance (Bashkirova et al., 2020; Coppola et al., 2013). An alternative hypothesis is that the spatial organization of *Olfr*/OSN subtypes is tuned to maximize the detection and discrimination of odorants in the peripheral olfactory system (Ressler et al., 1993). Interestingly, the receptive fields of mouse OSNs vary with their spatial location (Ma and Shepherd, 2000), which in some cases correlates with the patterns of odorant sorption in the mouse OM—this association was proposed as the "chromatographic hypothesis" decades before the discovery of the *Olfrs* (Mozell, 1966) and later rebranded as the "sorption hypothesis" in olfaction (Schoenfeld and Cleland, 2006; Scott et al., 2014). While some studies lend support to these hypotheses (reviewed in Secundo et al. 2014), others question their validity (Abaffy and Defazio, 2011; Coppola et al., 2019). Thus, the logic underlying the representation of smell in the peripheral olfactory system still remains unknown, and it is subject of great controversy (Kurian et al., 2021; Secundo et al., 2014).

Spatial transcriptomics, which combines spatial information with high-throughput gene expression profiling, expanded our knowledge of complex tissues, organs, or even entire organisms (Achim et al., 2015; Asp et al., 2019, 2020; Junker et al., 2014; Peng et al., 2016). In this study, we employed a spatial transcriptomics approach to create a 3D map of gene expression of the mouse nose, and we combined it with single-cell RNA sequencing (RNA-seq), machine learning, and chemoinformatics to resolve its molecular architecture and shed light onto the anatomical logic of smell.

RESULTS

A high-resolution spatial transcriptomic map of the mouse olfactory mucosa

We adapted the RNA-seq tomography (Tomo-seq) method (Junker et al., 2014) to create a spatially resolved genome-wide transcriptional atlas of the mouse nose. We obtained cryosections (35 μm) collected along the dorsal-ventral (DV), anterior-posterior (AP), and lateral-medial-lateral (LML) axes (n = 3 per axis) of the OM (Figure 1A) and performed RNA-seq on individual cryosections (see STAR Methods). After quality control (Figures S1A–S1D; Table S1; STAR Methods), we computationally refined the alignment of the cryosection along each axis, and we observed a high correlation between biological replicates (Figure 1B). Hence, we combined the three replicates into a single series of spatial data, including 54, 60, and 56 positions along the DV, AP, and LML axis, respectively (Figure 1C; STAR Methods). On average, we detected >18,000 genes per axis, representing a total of 19,249 genes for all axes combined (Figure 1D). Molecular markers for all canonical cell types known to populate the mouse OM were detected in all axes (Figure 1E) and expressed at the expected levels (Saraiva et al., 2015b).

Next, we verified the presence of a spatial signal with the Moran's I (Schmal et al., 2017; Figure S1E), whose value is significantly higher than 0 for the data along all axes ($p < 2 \times 10^{-16}$ for all axes), indicating that nearby sections have more similar patterns of gene expression than expected by chance. Given the left/right symmetry along the LML axis (Figure 1C), the data were centered and averaged on the two sides—henceforth, the LML axis will be presented and referred to as the lateral-medial (LM) axis (see STAR Methods). We could reproduce the expression patterns for known OM spatial markers, including the dorsomedial markers *Acsm4* and *Nqo1* (Gussing and Bohm, 2004; Oka et al., 2003) and the ventrolateral markers *Ncam2* and *Reg3g* (Alenius and Bohm, 1997; Yu et al., 2005; Figures 1F and S1F).

Together, these results show that RNA tomography is a sensitive and reliable method to examine gene expression patterns in the mouse OM.

Spatial differential gene expression analysis identifies cell-type-specific expression patterns and functional hotspots in the OM

In the last 3 decades, multiple genes with spatially segregated expression patterns across the OM have been identified. Most of these genes are expressed in mature OSNs and encode chemosensory receptors, transcription factors, adhesion molecules, and many molecules involved in the downstream signaling cascade of receptor activation (Cho et al., 2007; Cloutier et al., 2002; Fulle et al., 1995; Greer et al., 2016; Gussing and Bohm, 2004; Juilfs et al., 1997; Liberles and Buck, 2006; Miyamichi et al., 2005; Norlin et al., 2001; Oka et al., 2003; Pacifico et al., 2012; Saraiva et al., 2015b; Tietjen et al., 2003, 2005; Vassar et al., 1993; Wang et al., 2004; Yoshihara et al., 1997; Yu et al., 2005; Zapiec and Mombaerts, 2020). A smaller number of zonally expressed genes (e.g., metabolizing enzymes, chemokines, and transcription factors) were found to be expressed in sustentacular cells, globose basal cells, olfactory ensheathing cells, Bowman's gland cells, and respiratory epithelial cells (Cloutier et al., 2002; Duggan et al., 2008; Heron et al., 2013; Juilfs et al., 1997; Miyawaki et al., 1996; Norlin et al., 2001; Peluso et al., 2012; Whitby-Logan et al., 2004; Yu et al., 2005). Despite this progress, our knowledge on what genes display true zonal expression patterns and what cell types they are primarily expressed in is still very limited.

To identify axis-specific differentially expressed genes (DEGs) (hereafter referred to as spatial DEGs), we first filtered out lowly expressed genes, then binarized the expression levels at each position according to whether they were higher or lower than their median expression, and applied the Ljung-Box test to the autocorrelation function calculated on the binarized expression values (Figure S2A; STAR Methods). After correcting for multiple testing, we obtained a total of 12,303 spatial DEGs for the three axes combined (false discovery rate [FDR] < 0.01; Figure 2A)—the AP axis showed the highest number of spatial DEGs (10,855), followed by the DV axis (3,658) and the LM (1,318).

Next, we added cell-type resolution to the spatial axes by combining our data with a single-cell RNA-seq (scRNA-seq) dataset from 13 cell types present in the mouse OM (Fletcher et al., 2017). We cataloged spatial DEGs based on their expression in mature OSNs (mOSNs) versus the 12 other cell types (non-mOSNs; Figures 2B and 2C; Table S2). This led to the identification of 456 spatial DEGs expressed exclusively in non-mOSNs, which are associated with gene ontology (GO) terms, such as transcription factors, norepinephrine metabolism, toxin metabolism, bone development, regulation of cell migration, T cell activation, and others (Table S2). Some genes are expressed across many cell types, but others are specific to one cell type (Figure 2C; Table S2). As expected, some of these genes are cell-specific markers with known spatial expression patterns, such as the sustentacular cells and Bowman's glands markers *Cyp2g1* and *Gstm2* (Yu et al., 2005), the neural progenitor cell markers *Eya2* and *Hes6* (Tietjen et al., 2003), and the basal lamina and olfactory ensheathing cell markers *Aldh1a7* and *Aldh3a1* (Norlin et al., 2001; Table S2). We also identified spatial DEGs along a single axis or multiple axes and specific to one or few cell types (Figures S2B and S2C). For example, the ribosomal protein *Rps21* plays a key role in ribosome biogenesis, cell growth, and death (Wang et al., 2020) and is primarily expressed in horizontal basal cells (HBCs), consistent with their role in the maintenance and regeneration of the OM (Leung et al., 2007). Another example is the extracellular proteinase inhibitor *Wfdc18*, which induces the immune system and apoptosis (Jung et al., 2004) and is expressed in microvillous cells type 1 (MVC1s), consistent with their role in immune responses to viral infection (Baxter et al., 2020). Two more examples are the fibroblast growth factor *Fgf20* in immature sustentacular cells (iSCs) and the adapter protein *Dab2* in mature sustentacular cells (mSCs) (Figures S2B and S2C). *Fgf20* is expressed in several cell types, regulates the horizontal growth of the olfactory turbinates, and is preferentially expressed in the lateral OM (Yang et al., 2018), consistent with our data. *Dab2* regulates mechanisms of tissue formation, modulates immune responses, and participates in the absorption of proteins (Finkielstein and Capelluto, 2016; Park et al., 2019), consistent with the known maintenance and support roles of mSCs in the OM (Brann et al., 2020).

A GO enrichment analysis on the axis-specific DEGs for non-mOSNs genes revealed a very wide variety of biological processes and molecular functions. Some of the notable terms identified were water and fluid transport (e.g., *Ctfr* and *Aqp3*), transcription factors (e.g., *Hes1* and *Dlx5*), oxidation-reduction processes (e.g., *Scd2* and *Cyp2f2*), microtubule cytoskeleton organization involved in mitosis (e.g., *Stil* and *Aurkb*), cell cycle (e.g., *Mcm3* and *Mcm4*), cell division (e.g., *Kif11* and *Cdca3*), negative regulation of apoptosis (e.g., *Dab2* and *Scg2*), sensory perception of chemical stimulus (e.g., *Olfr870* and *Gnas*), and cellular processes (e.g., *Mal* and *Pthlh*), among many others (Table S2).

The identification of thousands of spatial DEGs prompted us to examine their distribution patterns along each axis and the putative functions associated with such spatial clusters of gene expression. We started by using uniform manifold approximation and projection (UMAP) (Becht et al., 2018) and hierarchical clustering to visualize and cluster all spatial DEGs along the three axes. This analysis uncovered nine patterns of expression in the DV and AP axes each and five patterns in the LM axis (Figures 2D and 2E). These patterns include variations of four major

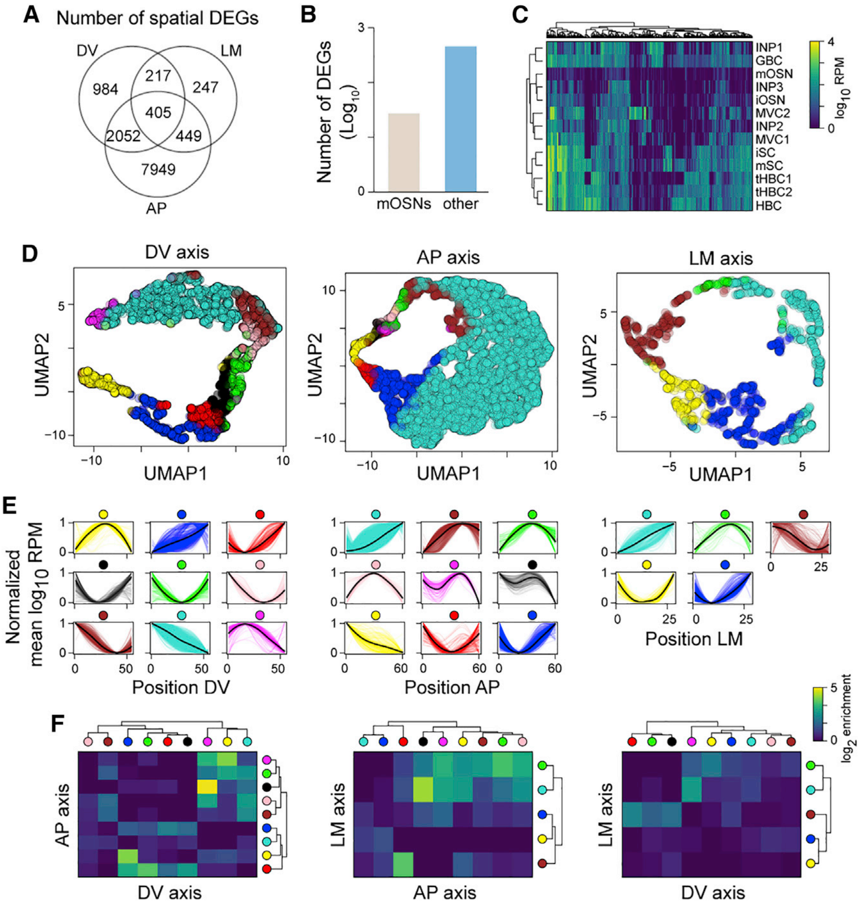


Figure 2. Genes with non-random spatial patterns across different cell types in the OM

(A) Venn diagram showing the numbers of spatial differentially expressed genes (DEGs) along each axis.

(B) Bar plot showing the $\log_{10}$ number of spatial DEGs that are mOSN specific (“mOSNs”) or that are detected only in cell types other than mOSNs (“other”).

(C) Heatmap of $\log_{10}$ mean expression per cell type of genes that are not expressed in mOSNs but only in other OM cell types (INPs, immediate neuronal precursors; iSCs, immature sustentacular cells; mSCs, mature sustentacular cells; MVCs, microvillous cells; mSCs, mature sustentacular cells).

(D) UMAP plots of spatial DEGs along the three axes (n = 3 per axis). Each gene is colored according to the cluster it belongs to.

(E) Normalized average expression patterns of spatial DEGs clusters along the three axes.

(F) Heatmap showing the $\log_2$ enrichment over the random case for the intersection between lists of genes belonging to different clusters (indicated by colored circles) across pairs of axes.

shapes: monotonically increasing (/), monotonically decreasing (\), U-shape (∪), and inverted U-shape (∩) (Figure 2E). The latter two patterns present clear maximum and minimum at different positions along the axis—for example, the brown, green, pink, magenta, and black AP clusters show a similar inverted U-shape pattern, but their maximum moves along the axis (Figure 2E). As expected, the dorsomedial markers *Acsm4* and *Nqo1* belong to the turquoise clusters in both the DV and LM axes, while the ventrolateral marker *Reg3g* belongs to the blue cluster from the DV axis (Figure 1F; Table S3), mimicking their respective expression pattern in the mouse OM.

The total number of genes per cluster had a median value of 236 but varied greatly between clusters—ranging from 57 in the green LM cluster to 8,551 in the turquoise AP cluster (Figure 2D; Table S3). GO enrichment analyses on the spatial DEGs yielded enriched terms for 14 of the 23 spatial clusters (Table S3). For example, the turquoise AP cluster displaying a monotonically increasing pattern (Figure 2E) yielded GO terms associated with the molecular machinery of mOSNs—such as axonal transportation, RNA processing, ribosomal regulation, and regulation of histone deacetylation (Table S3). Interestingly, the brown DV cluster, which displays a monotonically decreasing expression pattern (Figure 2E), had similar GO term enrichment (Table S3). In agreement with these results, we found that most known OSN activity-dependent markers (Wang et al., 2017) are spatial DEGs belonging to the AP turquoise and brown clusters, which contain genes with expression peaks in the posterior region (Figure S2E; Table S3). We also observed a similar trend in the DV axis, with many of these markers being more highly expressed in the dorsal region (Figure S2E).

The results above show that OSN activity is enriched in the dorsoposterior region of the OM, which could be due to an enrichment of OSNs in that region. To test this hypothesis, we estimated the abundance and spatial variability of OSNs and five additional major cell types (HBCs, globose basal cells [GBCs], SCs, MVCs, and immediate neuronal precursors [INPs]) in each section through a cell deconvolution analysis (see STAR Methods). We observed statistically significant changes in the abundance of OSNs, which is predicted to be higher in the dorsoposterior region of the OM, as previously suggested (Nickell et al., 2012; Vedin et al., 2009). Conversely, other cell types like HBCs are predicted to have an opposite pattern, as they tend to be more abundant in the anteroventral region (Figure S2F; Table S2).

Next, we extended our GO analysis to the remaining spatial clusters and found additional terms enriched or shared between several clusters among the three axes. For example, GO terms enriched in the dorsomedial region (turquoise DV, pink AP, and LM green clusters) include detoxification of several metabolites and multiple metabolic and catabolic processes, suggesting that this region is involved in the OM detoxification (Table S3). Another example is the enrichment in terms related to the immune system—such as defense response and humoral immune response—in the anteromedial section along the AP axis (yellow, black, and magenta AP and turquoise LM clusters), which strongly hints at a role of this area in defending OM from pathogenic invaders (Table S3). The anteroventral and posteroventral regions (blue DV and blue AP clusters) are enriched in terms related to the cellular and anatomical organization (e.g., extracellular matrix organization and regulation of cell communication) and bone and cartilage development (e.g., ossification and biomineral tissue development), suggesting these locations are hotspots for the development and regulation of the OM structure. Finally, the ventral portion of the DV (red DV cluster) is associated with terms related to cilia movement and function (e.g., regulation of cilium movement and microtubule-based movement), consistent with both the location and functions of the respiratory epithelium (Yu et al., 2005).

Next, we further explored the relationships between the genes populating each cluster. We found that ventral genes (blue DV cluster) tend to reach a peak in expression in the anterior area of the OM (yellow AP cluster) more often than expected by chance (Figure 2F). We also observed that medial genes (turquoise LM cluster) are more highly expressed in the dorsal (magenta DV cluster) and anterior regions (black, yellow, and magenta AP cluster), while genes peaking in the lateral region (brown LM cluster) tend to be ventral (red DV cluster; Figure 2F). These conclusions hold, even when we exclude *Olfrs* from the analysis (Figure S2D).

These associations between the clusters of spatial DEGs along different axes suggest that the presence of complex 3D expression patterns in OM is not restricted to either *Olfrs* or OSNs. Moreover, our results show that our experimental approach can uncover spatially restricted functional hotspots within the OM.

A 3D transcriptomic atlas of the mouse OM

Since the OR discovery 3 decades ago (Buck and Axel, 1991), *in situ* hybridization (ISH) has been the method of choice to study spatial gene expression patterns across the OM. This method is technically challenging and inherently a very low-throughput experimental approach.

As we showed above, our Tomo-seq data enable a systematic and quantitative estimation of gene expression levels along the three axes of the OM. Here, we take this analysis one step further and generate a fully browsable tridimensional (3D) gene expression atlas of the mouse OM. First, we reconstructed the 3D shape of OM based on publicly available images of OM sections (STAR Methods). We then fed the shape information combined with the gene expression data along the three axes into the iterative proportional fitting (IPF) algorithm (Fienberg, 1970; Junker et al., 2014; Figure 3A). The 3D atlas of the OM faithfully reproduced the known 3D pattern of the dorsomedial marker *Acsm4* (Oka et al., 2003; Figure 3B). To further validate our 3D gene expression atlas of the OM, we compared the 3D reconstructed patterns with conventional ISH patterns for five spatial DEGs identified in this study. The first gene validated was *Cytl1*, which we confirmed to be expressed along the septum throughout the OM (Figures 3C and 3D), consistent with the role *Cytl1* plays in osteogenesis, chondrogenesis, and bone and cartilage homeostasis (Shin et al., 2019; Zhu et al., 2019). The four additional genes (*Olfr309*, *Olfr618*, *Olfr727*, and *Moxd2*) validated via ISH are presented elsewhere in this manuscript (Figures 4, 5, and S4).

To make this 3D gene expression atlas of the mouse OM available to the scientific community, we created a web portal (available at http://atlas3dnose.helmholtz-muenchen.de:3838/atlas3Dnose) providing access to the spatial transcriptomic data described here in a browsable and user-friendly format. This portal contains search functionalities allowing the users to perform pattern search by gene, which returns (1) the normalized counts along each of the three axes, (2) the predicted expression pattern in 3D with a zoom function, (3) visualization of the expression patterns in virtual cryosections along the OM by selecting any possible pairwise intersection between two given axes (i.e., DVxAP, DVxLM, and APxLM), (4) the degrees of belonging for each "zone" (see results section below), and (5) single-cell expression data across 14 different cell types.

In sum, here, we generated and made publicly available a highly detailed and fully browsable 3D gene expression atlas of the mouse OM, which allows the exploration of the expression patterns for ~20,000 genes.

Topographical expression patterns of *Olfrs*

In our combined dataset, we detected a total of 959 *Olfrs* (Figure 4A), of which we confidently reconstructed the spatial expression patterns for 689 differentially expressed in space (FDR < 0.01; Figure 4B)—a number six times larger than the combined 112 *Olfrs* characterized by previous ISH studies (Miyamichi et al., 2005; Ressler et al., 1993; Vassar et al., 1994; Zapiec and Mombaerts, 2020). To define *Olfr* expression in 3D space in a rigorous, unbiased, and quantitative way, we ran a latent Dirichlet allocation (LDA) algorithm (STAR Methods; Liu et al., 2016) on the 689 spatially differentially expressed *Olfrs*. LDA is a generative statistical model that can infer the topics of a collection of documents based on the variability and frequency of specific words. In the context of this study, if the spatial expression data of *Olfrs* are considered equivalent to "documents," the inferred topics correspond to "zones" (STAR Methods). We ran LDA for different numbers of zones, and the trend of the log likelihood function suggested that the minimal number of topics required to represent the diversity of patterns is five (Figure S3A; STAR Methods). Next, we visualized the spatial distribution of these five zones in our 3D OM model, with colors representing the probability that a given spatial position belongs to each zone. These five zones extend from the dorsomedial-posterior to the lateroventral-anterior region (Figure 4C), consistent with the previously described zones (Miyamichi et al., 2005; Ressler et al., 1993; Vassar et al., 1993).

The majority of *Olfrs* with known spatial patterns are restricted to a single zone, but a small number of *Olfrs* are expressed across multiple zones in a continuous or non-continuous fashion (Miyamichi et al., 2005; Strotmann et al., 1992; Zapiec and Mombaerts, 2020). Under this logic, each *Olfr* has a different probability of belonging to the five topics and zones we identified. To test this assumption, we used the same mathematical framework as above to compute the probabilities that the expression pattern of each *Olfr* belongs to a given zone, i.e., the "degree of belonging" (DOB) (Table S4). The DOBs represent a decomposition of the expression patterns in terms of the five zones (Figure 4C) and quantitatively describe the changes in patterns of genes with overlapping areas of expression (e.g., see Figure S3B). The width

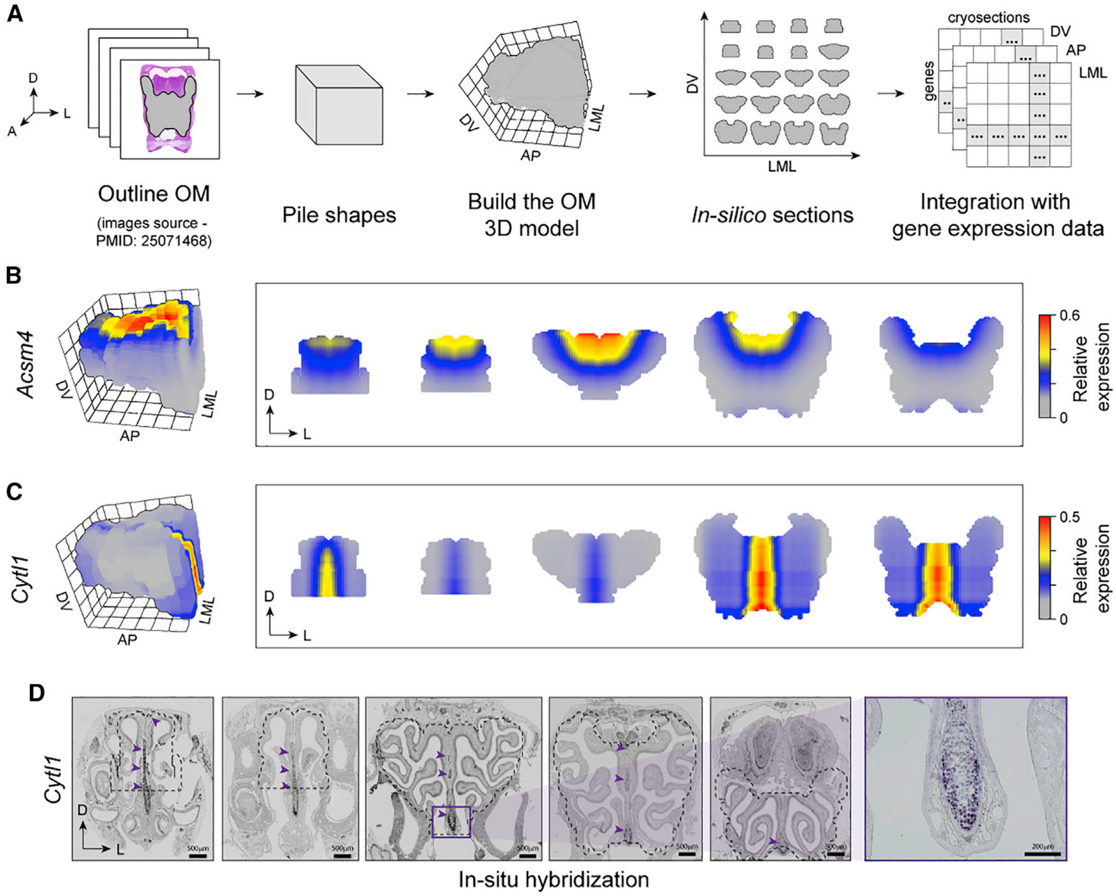


Figure 3. The 3D reconstruction of the OM

(A) Schematic of 3D shape reconstruction strategy. Images of 2D slices along the AP axis of the OM were piled together to build an *in silico* 3D model of OM, which can also be used to visualize *in silico* sections. This 3D model, together with the gene expression data along each axis, was the input of the iterative proportional fitting algorithm, which allowed us to estimate a 3D expression pattern for any gene.

(B and C) Reconstruction of the 3D expression patterns of the *Acsm4* (B) and *Cytl1* (C) in the OM, visualized in 3D and in OM coronal sections taken along the AP axis.

(D) ISH experiment validating *Cytl1* spatial expression pattern reconstructed in (C); note that *Cytl1* is expressed in the septal region all along the OM. Purple arrowheads indicate the location of labeled cells. The dotted outline marks the borders of the OM dissected and used in the RNA-seq experiments and for the construction of the 3D model.

of the distribution of DOBs across the five zones, which can be measured with entropy, can distinguish genes whose patterns mostly fit in a single zone from those spanning multiple zones (Figure S3C; STAR Methods).

To visualize the global distribution of the 689 *Olfrs*, we applied the diffusion map algorithm (Haghverdi et al., 2015) to their DOBs. This showed that the genes are approximately distributed along a continuous line spanning the five zones and without clear borders between zones (Figure 4D), consistent with previous studies (Miyamichi et al., 2005; Strotmann et al., 1992; Zapiec and Mombaerts, 2020). With the diffusion pseudo-time algorithm (Haghverdi et al., 2016), we calculated an index (hereafter referred to as "3D index") that tracks the position of each *Olfr* gene along the 1D curve in the diffusion map and represents its expression pattern (Figure 4E).

While our approach yielded an index for the 689 spatially differentially expressed *Olfr* genes used to build the diffusion map, there were 697 *Olfrs* that could not be analyzed, either because they were too lowly expressed or not detected at all in our dataset (Figure 4A). Since the spatial expression patterns for some *Olfrs* are partly associated with their chromosomal and genomic coordinates (Sullivan et al., 1996; Tan and Xie, 2018; Zhang et al., 2004), we hypothesized that we could use a machine-learning algorithm to predict the 3D indices for the 697 *Olfrs* missing from our dataset. Thus, we trained a random forest algorithm on the 3D indices of the spatially differentially expressed *Olfrs* in our dataset using nine genomic features as predictors, such as the chromosomal position, number of *Olfrs* in cluster, and distance to nearest known enhancer (Figure 4F; STAR Methods). The algorithm performance was confirmed by over 100 cross-validation iterations, which revealed a low root-mean-square error ($\lesssim 10\%$) on the mean 3D index (Figure S4A; STAR Methods). The five most important predictors were features associated with

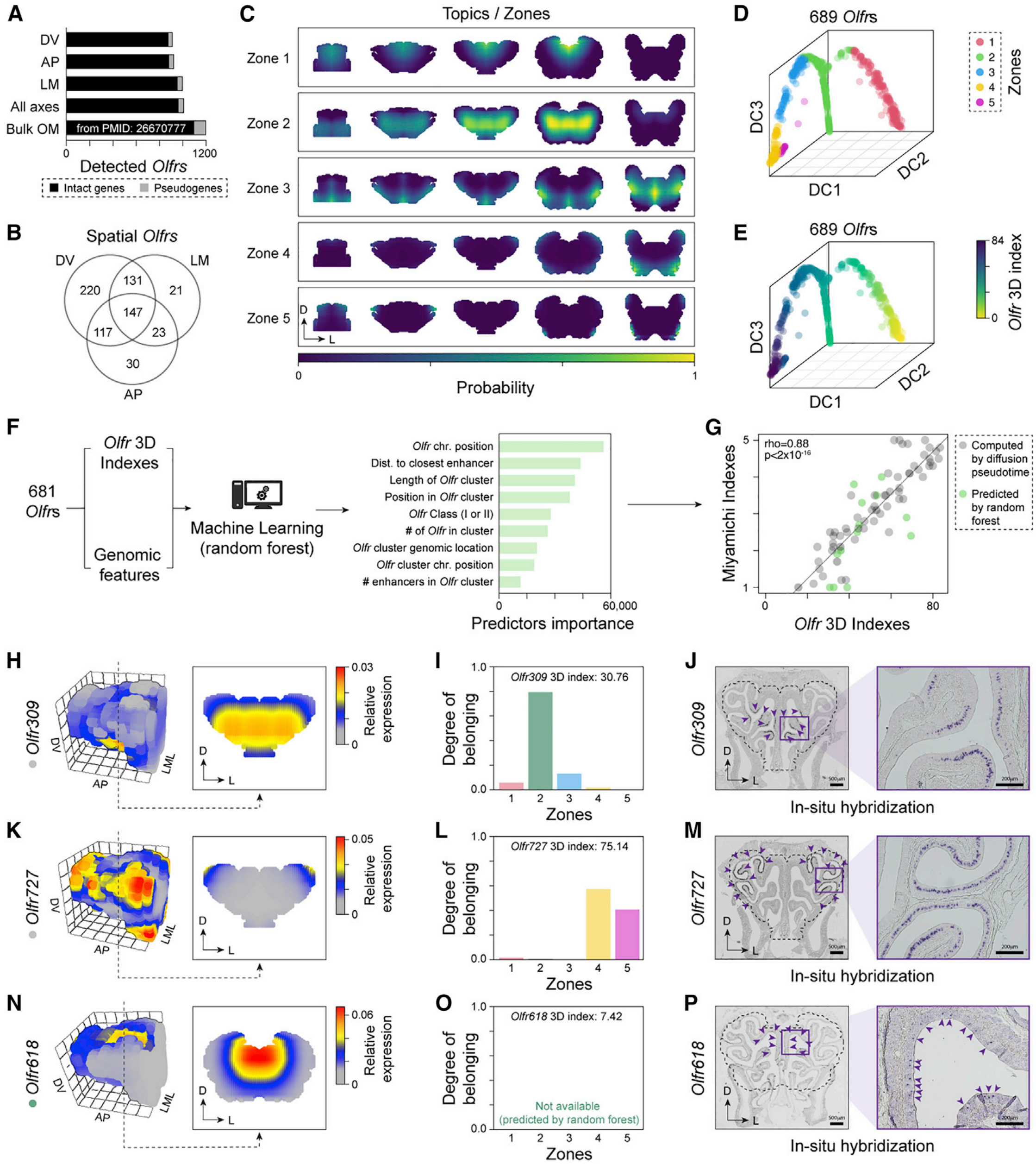


Figure 4. Zonal organization of *Olfrs* in the OM

(A) Number of *Olfrs* detected in our data and in an OM bulk RNA-seq data (Saraiva et al., 2015b).

(B) Venn diagram of spatially differentially expressed *Olfrs* per axis (n = 3 per axis).

(C) Visualization of the five zones across the OM (coronal sections) estimated with a latent Dirichlet allocation algorithm. The colors indicate the probability (scaled by its maximum value) that a position belongs to a given zone.

(D) Diffusion map of *Olfrs*. Genes are colored based on the zone they fit in the most. DC, diffusion component.

(E) Same as (D), with *Olfrs* colored by their 3D index.

(F) We fitted a random forest algorithm to the 3D indices of 681 spatial *Olfrs* using nine genomic features as predictors. After training, the random forest was used to predict the 3D indices of 697 *Olfrs* that have too low levels in our data.

(G) 3D indices versus the indices of 80 *Olfrs* estimated in Miyamichi et al. (2005) from ISH data. Black circles indicate *Olfrs* detected in our dataset; green circles are *Olfrs* whose indices were predicted with random forest. The correlation coefficients computed separately on these two sets of *Olfrs* are, respectively, rho = 0.92 ($p < 2 \times 10^{-16}$) and rho = 0.69 (p = 0.009).

(H–P) Predicted expression patterns (H, K, and N), degrees of belonging (I, L, and O), and ISH (J, M, and P) for *Olfr309*, *Olfr727*, and *Olfr618*, respectively. Purple arrowheads indicate the location of labeled cells. The dotted outline marks the borders of the OM dissected and used in the RNA-seq experiments and for the construction of the 3D model.

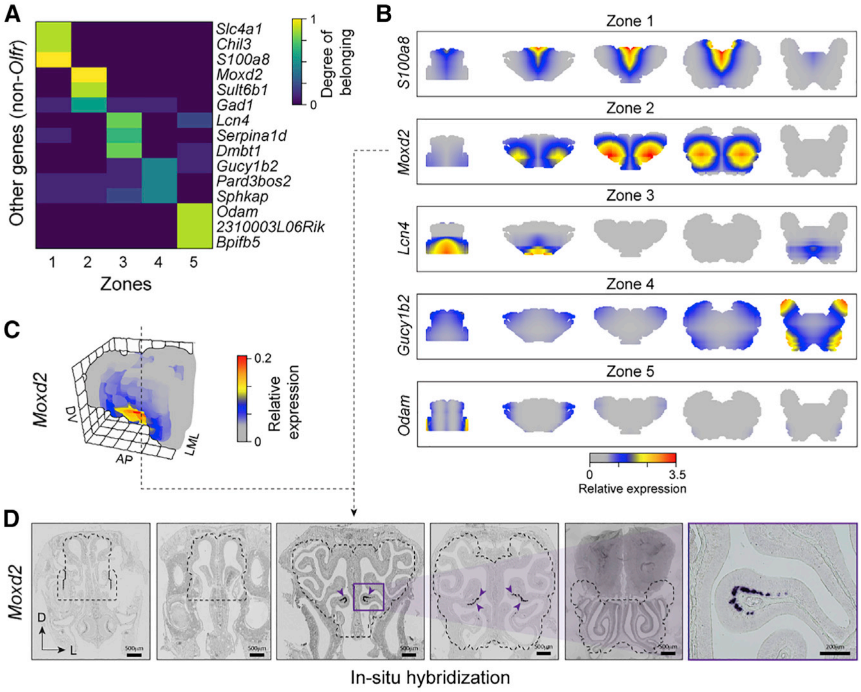


Figure 5. Zonal organization of non-*Olfr* genes in the OM

(A) Heatmap of degrees of belonging of most zone-specific non-*Olfr* genes.

(B) 3D gene expression pattern (coronal sections) of most topic-specific non-*Olfrs* for each topic along the AP axis.

(C) Reconstruction of the 3D expression pattern of the gene *Moxd2* in the OM.

(D) ISH experiment validating *Moxd2* spatial expression pattern reconstructed in (B) and (C). Purple arrowheads indicate the location of ISH-labeled cells. The dotted outline marks the borders of the OM dissected and used in the RNA-seq experiments and for the construction of the 3D model.

chromosomal position, distance to the closest *Olfr* enhancer (Monahan et al., 2017), length of the *Olfr* cluster, position in the *Olfr* cluster, and phylogenetic *Olfr* class (Figure 4F). Using this machine-learning algorithm, we predicted the 3D indices for the 697 *Olfrs* missing reliable expression estimates in our dataset (Table S4).

Overall, through multiple unsupervised and supervised computational methods, we have quantitatively defined five spatial expression domains in the OM (called zones) and have provided accurate 3D spatial indices for 1,386 *Olfrs*, which represents ~98% of the annotated *Olfrs*.

Importantly, we found strong correlations between the “Miyamichi indices” inferred using ISH in Miyamichi et al. (2005) and our 3D indices (rho = 0.88; $p < 2 \times 10^{-16}$; Figure 4G). This correlation remains significant when we separately analyze the 3D indices computed by diffusion pseudo-time (rho = 0.92; $p < 2 \times 10^{-16}$) or predicted by random forest (rho = 0.69; p = 0.009). In addition, our indices also correlated with the “Zolfr indices” (Zapiec and Mombaerts, 2020; rho = 0.88; $p < 2 \times 10^{-16}$; Figure S4B), and with the “Tan indices” (Tan and Xie, 2018; rho = 0.89; $p < 2 \times 10^{-16}$; Figure S4C), inferred by ISH and RNA-seq, respectively.

To confirm our predictions, we performed ISH for three *Olfrs* that have not been characterized before—two detected in our dataset and for which the 3D index was calculated via diffusion pseudo-time (DPT) (*Olfr309* and *Olfr727*) and one not detected in our dataset and for which the 3D index was predicted with the random forest algorithm (*Olfr 618*). Notably, all three *Olfrs* were expressed primarily within the zones they were predicted to be expressed in: zone 2 for *Olfr309* (3D index = 30.76), zones 4 and 5 for *Olfr727* (3D index = 75.14), and zone 1 for *Olfr618* (3D index = 7.42; Figures 4H–4P and S4D–S4F; Table S4).

Topographical expression patterns for non-Olfr genes

A recent study performed RNA-seq in 12 randomly dissected OM pieces along the DV axis and identified ~700 non-*Olfr* genes with putatively spatial patterns (Tan and Xie, 2018), including many genes with zonal expression patterns identified previously (Duggan et al., 2008; Gussing and Bohm, 2004; Liberles and Buck, 2006; Ling et al., 2004; Norlin et al., 2001; Oka et al., 2003; Tietjen et al., 2003; Whitby-Logan et al., 2004; Yoshihara et al., 1997). By identifying 11,538 non-*Olfr* spatial DEGs (Figures 2 and 3; Table S5), we increased the list of non-*Olfr* genes with spatial zonation in the OM by 16-fold.

Using the mathematical framework based on topic modeling described above, we decomposed the expression patterns of non-*Olfr* genes onto the five zones we identified. This allowed us to identify genes showing zone specificity by calculating the entropy of the DOBs distributions. Interestingly, we found 28 genes highly specific for each of the five zones (i.e., with entropy <1; STAR Methods; Figure 5A; Table S5). For example, *S100a8* (zone 1) codes for a calcium-binding protein involved in calcium signaling and inflammation (Yoshikawa et al., 2018), *Moxd2* (zone 2) is a mono-oxygenase dopamine hydroxylase-like protein possibly involved in olfaction (Kim et al., 2014), *Lcn4* (zone 3) is a lipocalin involved in transporting odorants and pheromones (Charkoftaki et al., 2019; Miyawaki et al., 1994), *Gucy1b2* (zone 4) is a soluble guanylyl cyclase oxygen and nitric oxide (Bleymehl et al., 2016; Koglin et al., 2001), and *Odam* (zone 5) is a secretory calcium-binding phosphoprotein involved in cellular differentiation and matrix protein production and with antimicrobial functions of the junctional epithelium (Lee et al., 2012; Springer et al., 2019; Figure 5B). The high zone specificity of the expression pattern of these genes gives clues into possible biological processes taking place in the zones. Indeed, *Gucy1b2* is a known genetic marker for a small OSN subpopulation localized in cul-de-sac regions in the lateral OM, consistent with our reconstruction (Figure 5B), and it

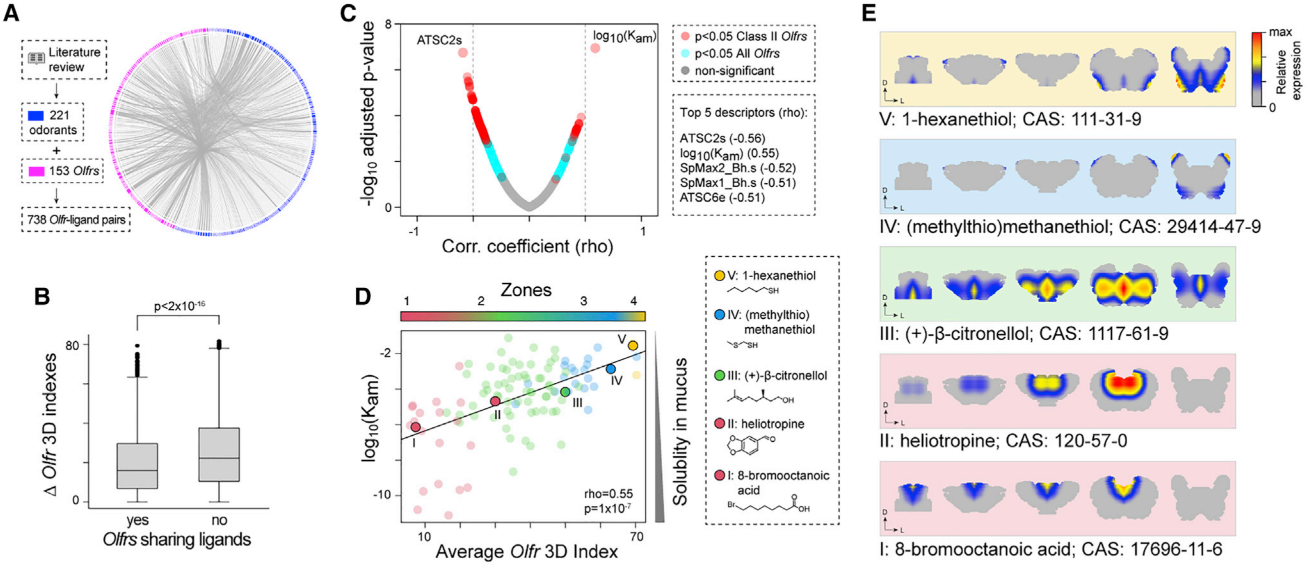


Figure 6. Physiological role of the zones

(A) Circular network illustrating the pairs of Olfrs and ligands that we found in the literature.

(B) Boxplots showing the distributions of the absolute value of 3D index differences between pairs of Olfrs sharing at least one ligand versus pairs of Olfrs without cognate ligands in common. The difference between the two distributions is statistically significant ($p < 2 \times 10^{-16}$; Wilcoxon rank-sum test).

(C) Scatterplot showing the Spearman correlation coefficients between the ligands' mean 3D indices and molecular descriptors and the corresponding $-\log_{10}$(adjusted p value). Turquoise circles indicate the descriptors having a significant correlation only when both class I and II *Olfrs* are considered; red circles mark the descriptors with a significant correlation also when class I *Olfr*s are removed.

(D) Scatterplot illustrating the correlation between air/mucus partition coefficients of the odorants and the average 3D indices of their cognate Olfrs. Only odorants for which we know at least two cognate Olfrs (110) were used here. Odorants are colored according to the zone they belong to (defined as the zone with the highest average degree of belonging computed over all cognate receptors). The five odorants highlighted in the plot by larger circles are indicated on the right-hand side, along with their molecular structure and common name.

(E) Average expression pattern of the cognate Olfrs recognizing each of the five odorants highlighted in (D), including their respective CAS numbers.

regulates the sensing of environmental oxygen levels through the nose (Omura and Mombaerts, 2015; Saraiva et al., 2015b). In addition, our ISH experiments revealed that *Moxd2* is expressed in a small ventrolateral patch of the OM (Figure 5D), validating its predicted 3D spatial pattern (Figures 5B and 5C) and highlighting a potential highly localized role of this protein in neurotransmitter metabolism (Goh et al., 2016) in the mouse OM.

A recent study showed that the transcription factors Nfia, Nfib, and Nfix regulate the zonal expression of *Olfrs* (Bashkirova et al., 2020). To get some insights into the signaling pathways involved in this process, we mined our dataset for genes encoding ligands and receptors (Efremova et al., 2020) correlated with the expression patterns of the Nfis (STAR Methods). This analysis returned 476 genes involved in biological processes associated with the regulation of neurogenesis, regulation of cell development, anatomical structure development, cellular component organization or biogenesis, and regulation of neuron differentiation (Table S5). As expected, some of these genes have known functions in the OM, such as segregating different cell lineages for *Notch1-3* (Carson et al., 2006), genes associated with the development of the nervous system (e.g., *Erbb2* and *Lrp2*; Britsch et al., 1998; Spuch et al., 2012), and many others associated with the semaphorin-plexin, ephrin-Eph, and Slit-Robo signaling complexes—which regulate OSN axon guidance and spatial patterning of the OM (Cloutier et al., 2002; Cutforth et al., 2003; Huber et al., 2003; Kania and Klein, 2016). Excitingly, the majority of these 476 genes still have unknown functions in the OM, thus highlighting the potential of our approach to discover genes and pathways involved in the regulation of zonal expression in the OM.

The anatomical logic of smell

For most sensory systems, the functional logic underlying the topographic organization of primary receptor neurons and their receptive fields is well known (Kandel et al., 2013). In contrast, the anatomic logic of smell still remains unknown, and it is subject of great controversy and debate (Kurian et al., 2021; Secundo et al., 2014).

To explore the underlying logic linked to the zonal distribution of Olfrs, we investigated possible biases between their expression patterns and the physicochemical properties of their cognate ligands. First, we compiled a list of known 738 Olfr-ligand pairs, representing 153 Olfrs and 221 odorants (Figure 6A; Table S6). Interestingly, Olfr pairs sharing at least one common ligand have more similar expression patterns (i.e., more similar 3D indices) than Olfrs detecting different sets of odorants (Wilcoxon rank-sum test; $p < 2 \times 10^{-16}$; Figure 6B). This observation is consistent with the hypothesis that the Olfr zonal distribution depends, at least partially, on the properties of the odorants they bind to.

Next, we considered a set of 1,210 physicochemical descriptors, including the molecular weight, the number of atoms, aromaticity index, lipophilicity, and the air/mucus partition coefficient (K_{am}), which quantifies the mucus solubility of each ligand (Rygg et al., 2017; Scott et al., 2014; STAR Methods). We then computed the Spearman's correlation of each of these descriptors of the ligands with the average 3D indices of the Olfrs detecting them. We found a statistically significant correlation for 744 descriptors (FDR < 0.05; Figure 6C; Table S6). The top five highest correlations were with the air/mucus partition coefficient K_{am} (rho = 0.55; adjusted p = 1×10^{-7}), ATSC2S (rho = −0.56; adjusted p = 2×10^{-7}), SPmax2_Bh.s (rho = −0.52; adjusted p = 2×10^{-6}), SPmax1_Bh.s (rho = −0.51; adjusted p = 3×10^{-6}), and ATSC6e (rho = −0.51; adjusted p = 3×10^{-6}; Figure 6C; Table S6). Interestingly, ATSC2S, SPmax1_Bh.s, and SPmax2_Bh.s are also related to solubility (Consonni and Todeschini, 2008; Devillers and Domine, 1997; Hollas, 2003). Notably, the association between K_{am} and the mean 3D indices does not depend on the number of zones defined with LDA (STAR Methods). Furthermore, it remains robust to changes in the set of ligands and/or Olfrs used for the analysis, namely, when we excluded Olfrs for which the 3D indices were predicted with the random forest model (rho = 0.48; p = 2×10^{-6}; Figure S5B) or when only 3D indices from class II *Olfrs* were included in the analysis (rho = 0.5; p = 1×10^{-7}; Figure S5C).

In particular, the positive correlation of the 3D indices with K_{am} (Figure 6D) indicates that the most soluble odorants (lower) preferentially activate Olfrs located in the most antero-dorsomedial region (zone 1) of the OM, while the least soluble odorants (higher K_{am}) activate Olfrs in the postero-ventrolateral region (zones 4 to 5). In other words, gradients of odorants sorption (as defined by their K_{am}) correlate with the gradients of Olfr expression from zone 1 to zone 5, consistent with the chromatographic/sorption hypothesis in olfaction (Mozell, 1966; Scott et al., 2014). This is exemplified by the plots in Figure 6E, illustrating the predicted average expression levels across OM sections of the Olfrs binding to five odorants with different values of K_{am}. These results show a direct association between *Olfr* spatial patterns and the calculated sorption patterns of their cognate ligands in the OM, providing a potential explanation for the physiological function of the zones in the OM.

DISCUSSION

Past studies yielded inconclusive and sometimes contradictory views on the basic logic underlying the peripheral representation of smell, partly because the topographic distribution of OSN subtypes and their receptive fields still remained vastly uncharted, data on Olfr-ligand pairs were scarce, and there were pitfalls associated with electro-olfactogram recordings used to study spatial patterns of odor recognition in the nose (Kurian et al., 2021; Scott and Scott-Johnson, 2002; Secundo et al., 2015). Here, we combined RNA-seq and computational approaches that utilize unsupervised and supervised machine learning methods to discover and quantitatively characterize spatial expression patterns in the OM. We created a 3D transcriptional map of the mouse OM, which allowed us to spatially characterize 17,628 genes, including ~98% of the annotated *Olfrs*. We identified and validated by ISH several spatial marker genes, and a clustering analysis pinpointed the OM locations where specific functions related to, e.g., the immune response might be carried out. We also mathematically defined *Olfr* expression zones in the OM with an unsupervised machine-learning method based on topic modeling. We estimated that the OM includes at least five zones, which can be used to decompose the expression patterns of all genes. However, our analysis showed that there is a continuous distribution of *Olfrs* patterns in the OM. Thus, while a discrete number of zones might be convenient to provide a first classification of *Olfrs*, these might obscure the complexity of the OM spatial patterns. To account for this, we adopted a mathematical framework that can rigorously define zones while capturing finer structures in the data, via the degrees of belonging and the 3D index, which are more suitable to describe *Olfrs* patterns crossing multiple zones.

The global transcriptomic landscape of the vertebrate OM is similar between individuals and broadly conserved among different vertebrate species, ranging from zebrafish to human (Bear et al., 2016; Saraiva et al., 2015a, 2019). Similarly, the spatial segregation of *Olfrs* into partially overlapping rings of expression, centered around the midline structure of the OM, is also conserved among vertebrates (Freitag et al., 1995; Horowitz et al., 2014; Marchand et al., 2004; Miyamichi et al., 2005; Octura et al., 2018; Ressler et al., 1993; Strotmann et al., 1992; Vassar et al., 1993; Weth et al., 1996). While the number of *Olfr* zones in zebrafish, frog, and salamander still remain unknown (Freitag et al., 1995; Marchand et al., 2004; Weth et al., 1996), ISH studies suggested that the total number of *Olfr* expression zones can vary between mammals—ranging from two in macaque (Horowitz et al., 2014) to four in rat (Vassar et al., 1993) and goat (Octura et al., 2018), and between four and nine in mouse (Miyamichi et al., 2005; Ressler et al., 1993; Zapiec and Mombaerts, 2020). While the exact number of *Olfr* expression zones in OM still remains under debate, our results are consistent with both another recent RNA-seq study (Tan and Xie, 2018) and the largest *Olfr* ISH study in the mouse OM (Miyamichi et al., 2005), thus supporting the existence of at least five overlapping *Olfr* expression zones in the mouse nose.

Taking into account how conserved the molecular organization of the OM is in vertebrates, the 2-fold reduction in the number of *Olfr* expression zones in macaque compared with rodents and goat (an ungulate) is puzzling. While we cannot exclude the presence of confounding factors (e.g., limited *Olfr* sampling and/or inconsistent definitions of "zones"), it is interesting that the 2-fold reduction in number of zones is associated with a 2-fold reduction in the number of annotated intact *Olfrs* in macaque (and other higher primates, including human) compared with other rodents and ungulates (Horowitz et al., 2014; Niimura et al., 2014; Saraiva et al., 2019). Since the accelerated loss of *Olfr* genes during primate evolution has been linked to the acquisition of trichromatic acute vision and dietary changes (Niimura et al., 2018), it is plausible that these evolutionary pressures also helped shape the spatial distribution of *Olfrs* in macaques and other primates, including human.

The quantitative framework we built for this dataset will facilitate the interrogation of gene expression patterns via an online tool we provide and help answer important questions on the physiology of the nose. Our approach could be easily applied to spatial transcriptomic data collected from other tissues to perform comparisons across tissues from different species or the same tissue across multiple developmental stages. Moreover, the results from this study serve as a template to start answering other important questions about olfaction, such as whether *Olfr* spatial expression maps can encode maps of odor perception. Because the general molecular mechanisms of olfaction, zonal organization of *Olfrs*, conservation of ligands among Olfr orthologs, and components of olfactory perception are conserved in mammals (Adipietro et al., 2012; Bear et al., 2016; Freitag et al., 1995; Horowitz et al., 2014; Kurian et al., 2021; Manoel et al., 2021; Octura et al., 2018; Saraiva et al., 2019; Weth et al., 1996), the association we uncovered here between Olfr zones and the solubility of odorants they detect can likely be extrapolated to other mammals, including humans. Finally, the functional logic underlying the mammalian topographic organization of primary receptor neurons and their receptive fields in smell is now starting to be exposed.

Limitations of the study

This study enabled us to answer fundamental and long-standing questions about the rationale behind the spatial organization of the peripheral olfactory system. Specifically, we provide evidence to the hypothesis that the spatial zones increase the discriminatory power of the olfactory system by distributing Olfrs in the areas of the OM more likely to be reached by their cognate ligands, based on their solubility in mucus. A caveat of this approach is that the Olfr-ligand list we compiled from the literature includes odorant libraries of different size and composition and tested using different experimental approaches. Moreover, highly abundant Olfrs have a higher probability of being deorphanized than lowly abundant Olfrs, and ecologically relevant odorants are more likely to activate Olfrs when compared with other odorants (Dunkel et al., 2014; Saraiva et al., 2019; Trimmer and Mainland, 2017). Despite having compiled and performed our analysis on the largest set of Olfr-ligand pairs assembled to date and carrying out multiple robustness checks, we cannot rule out that ascertainment bias might contribute to the associations we found between the Olfr spatial location and the properties of their respective ligands. Future studies investigating the activation profiles for all mouse Olfrs and/or mapping the *in vivo* activation patterns of mouse Olfrs in the olfactory mucosa will be key to stress test the conclusions of our study.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - ○ Lead contact
  - ○ Materials availability
  - ○ Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - ○ Animals
- METHOD DETAILS
  - ○ Dissection of the olfactory mucosa, cryosections, and RNA-sequencing
  - ○ RNA-seq data mapping and gene counting
  - ○ Quality control
  - ○ Data normalization
  - ○ Identification of differentially expressed genes and gene clustering
  - ○ Combining Tomo-seq with single-cell RNA-seq data
  - ○ Gene ontology (GO) enrichment analysis
  - ○ Cell type deconvolution analysis
  - ○ Identification of ligands and receptors associated with the NfiA, NfiB or NfiX transcription factors
  - ○ 3D spatial reconstruction
  - ○ Definition of zones by topic modelling
  - ○ Definition of *Olfr* 3D indexes via diffusion pseudo-time
  - ○ Prediction of zone index for undetected *Olfrs* with Random Forest
  - ○ Odorant information and Olfr-ligand pairs
  - ○ Correlation analysis of physico-chemical descriptors with 3D index
  - ○ *In-situ* hybridization
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.celrep.2022.110547.

ACKNOWLEDGMENTS

We would like to thank Ximena Ibarra-Soria for help with the initial analysis and Maria-Elena Torres-Padilla, Bob Datta, Giorgia Greco, and the members of the Saraiva and Scialdone Labs for the constructive feedback, insightful discussions, and helpful comments. This work was supported by the Helmholtz Association (A.S.), the Deutsche Forschungsgemeinschaft (SC280/2–1 to A.S.), the National Institutes of Health (R01DC016647 to J.R.), and Sidra Medicine (SDR400040 to L.R.S.)—a member of Qatar Foundation.

AUTHOR CONTRIBUTIONS

M.L.R.T.S. analyzed data and wrote the initial version of the manuscript. E.A.M. performed the RNA-seq experiments and analyzed data. T.S.N., L.S.M., L.W., and S.L. performed experiments. M.M., S.S.Y.H., J.D.M., F.V., M.C., and M.O. analyzed data. E.G., J.R., D.W.L., and B.M. analyzed data and helped write the manuscript. A.S. and L.R.S. conceived and supervised the project, analyzed data, and wrote the final version of the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: September 13, 2021
Revised: January 26, 2022
Accepted: March 1, 2022
Published: March 22, 2022

REFERENCES

Abaffy, T., and Defazio, A.R. (2011). The location of olfactory receptors within olfactory epithelium is independent of odorant volatility and solubility. BMC Res. Notes *4*, 137.

Abaffy, T., Matsunami, H., and Luetje, C.W. (2006). Functional analysis of a mammalian odorant receptor subfamily. J. Neurochem. *97*, 1506–1518.

Achim, K., Pettit, J.-B., Saraiva, L.R., Gavriouchkina, D., Larsson, T., Arendt, D., and Marioni, J.C. (2015). High-throughput spatial mapping of single-cell RNA-seq data to tissue of origin. Nat. Biotechnol. *33*, 503–509.

Adipietro, K.A., Mainland, J.D., and Matsunami, H. (2012). Functional evolution of mammalian odorant receptors. PLoS Genet. *8*, e1002821.

Aldinucci, M., Bagnasco, S., Lusso, S., Pasteris, P., Rabellino, S., and Vallero, S. (2017). OCCAM: a flexible, multi-purpose and extendable HPC cluster. J. Phys. Conf. Ser. *898*, 082039.

Alenius, M., and Bohm, S. (1997). Identification of a novel neural cell adhesion molecule-related gene with a potential role in selective axonal projection. J. Biol. Chem. *272*, 26083–26086.

Aliee, H., and Theis, F.J. (2021). AutoGeneS: automatic gene selection using multi-objective optimization for RNA-seq deconvolution. Cell Syst. *12*, 706–715.e4.

Anders, S., Pyl, P.T., and Huber, W. (2014). HTSeq - a Python framework to work with high-throughput sequencing data. Bioinformatics *31*, 166–169.

Angerer, P., Haghverdi, L., Buttner, M., Theis, F.J., Marr, C., and Buettner, F. (2016). destiny: diffusion maps for large-scale single-cell data in R. Bioinformatics *32*, 1241–1243.

Araneda, R.C., Peterlin, Z., Zhang, X., Chesler, A., and Firestein, S. (2004). A pharmacological profile of the aldehyde receptor repertoire in rat olfactory epithelium. J. Physiol. *555*, 743–756.

Asp, M., Giacomello, S., Larsson, L., Wu, C., Furth, D., Qian, X., Wardell, E., Custodio, J., Reimegard, J., Salmen, F., et al. (2019). A spatiotemporal organ-wide gene expression and cell atlas of the developing human heart. Cell *179*, 1647–1660.e19.

Asp, M., Bergenstrahle, J., and Lundeberg, J. (2020). Spatially resolved transcriptomes-next generation tools for tissue exploration. Bioessays *42*, e1900221.

Barrios, A.W., Nunez, G., Sanchez Quinteiro, P., and Salazar, I. (2014). Anatomy, histochemistry, and immunohistochemistry of the olfactory subsystems in mice. Front. Neuroanat. *8*, 63.

Bashkirova, E., Monahan, K., Campbell, C.E., Osinski, J.M., Tan, L., Schieren, I., Barnea, G., Xie, X.S., Gronostajski, R.M., and Lomvardas, S. (2020). Homeotic regulation of olfactory receptor choice via NFI-dependent heterochromatic silencing and genomic compartmentalization. Preprint at bioRxiv. https://doi.org/10.1101/2020.08.30.274035.

Baxter, B.D., Larson, E.D., Merle, L., Feinstein, P., Polese, A.G., Bubak, A.N., Niemeyer, C.S., Hassell, J., Shepherd, D., Ramakrishnan, V.R., et al. (2020). Transcriptional profiling reveals potential involvement of microvillous TRPM5-expressing cells in viral infection of the olfactory epithelium. Preprint at bioRxiv. https://doi.org/10.1101/2020.05.14.096016.

Bear, D.M., Lassance, J.-M., Hoekstra, H.E., and Datta, S.R. (2016). The evolving neural and genetic architecture of vertebrate olfaction. Curr. Biol. *26*, R1039–R1049.

Becht, E., McInnes, L., Healy, J., Dutertre, C.A., Kwok, I.W.H., Ng, L.G., Ginhoux, F., and Newell, E.W. (2018). Dimensionality reduction for visualizing single-cell data using UMAP. Nat. Biotechnol. *37*, 38–44.

Blei, D.M., Ng, A.Y., and Jordan, M.I. (2003). Latent dirichlet allocation. J. Mach. Learn. Res. *3*, 993–1022.

Bleymehl, K., Perez-Gomez, A., Omura, M., Moreno-Perez, A., Macias, D., Bai, Z., Johnson, R.S., Leinders-Zufall, T., Zufall, F., and Mombaerts, P. (2016). A sensor for low environmental oxygen in the mouse main olfactory epithelium. Neuron *92*, 1196–1203.

Bozza, T., Feinstein, P., Zheng, C., and Mombaerts, P. (2002). Odorant receptor expression defines functional units in the mouse olfactory system. J. Neurosci. *22*, 3033–3043.

Brann, D.H., Tsukahara, T., Weinreb, C., Lipovsek, M., Van den Berge, K., Gong, B., Chance, R., Macaulay, I.C., Chou, H.J., Fletcher, R.B., et al. (2020). Non-neuronal expression of SARS-CoV-2 entry genes in the olfactory system suggests mechanisms underlying COVID-19-associated anosmia. Sci. Adv. *6*, eabc5801.

Britsch, S., Li, L., Kirchhoff, S., Theuring, F., Brinkmann, V., Birchmeier, C., and Riethmacher, D. (1998). The ErbB2 and ErbB3 receptors and their ligand, neuregulin-1, are essential for development of the sympathetic nervous system. Genes Dev. *12*, 1825–1836.

Buck, L., and Axel, R. (1991). A novel multigene family may encode odorant receptors: a molecular basis for odor recognition. Cell *65*, 175–187.

Carson, C., Murdoch, B., and Roskams, A.J. (2006). Notch 2 and Notch 1/3 segregate to neuronal and glial lineages of the developing olfactory epithelium. Dev. Dyn. *235*, 1678–1688.

Charkoftaki, G., Wang, Y., McAndrews, M., Bruford, E.A., Thompson, D.C., Vasiliou, V., and Nebert, D.W. (2019). Update on the human and mouse lipocalin (LCN) gene family, including evidence the mouse Mup cluster is result of an “evolutionary bloom”. Hum. Genomics *13*, 11.

Chess, A., Simon, I., Cedar, H., and Axel, R. (1994). Allelic inactivation regulates olfactory receptor gene expression. Cell *78*, 823–834.

Cho, J.H., Lepine, M., Andrews, W., Parnavelas, J., and Cloutier, J.F. (2007). Requirement for Slit-1 and Robo-2 in zonal segregation of olfactory sensory neuron axons in the main olfactory bulb. J. Neurosci. *27*, 9094–9104.

Cloutier, J.F., Giger, R.J., Koentges, G., Dulac, C., Kolodkin, A.L., and Ginty, D.D. (2002). Neuropilin-2 mediates axonal fasciculation, zonal segregation, but not axonal convergence, of primary accessory olfactory neurons. Neuron *33*, 877–892.

Consonni, and Todeschini, R. (2008). New spectral indices for molecule description. MATCH Commun. Math. Comput. Sci. *60*, 3–14.

Coppola, D.M., Fitzwater, E., Rygg, A.D., and Craven, B.A. (2019). Tests of the chromatographic theory of olfaction with highly soluble odors: a combined electro-olfactogram and computational fluid dynamics study in the mouse. Biol. Open *8*, bio047217.

Coppola, D.M., Waggener, C.T., Radwani, S.M., and Brooks, D.A. (2013). An electroolfactogram study of odor response patterns from the mouse olfactory epithelium with reference to receptor zones and odor sorptiveness. J. Neurophysiol. *109*, 2179–2191.

Cutforth, T., Moring, L., Mendelsohn, M., Nemes, A., Shah, N.M., Kim, M.M., Frisen, J., and Axel, R. (2003). Axonal ephrin-As and odorant receptors: coordinate determination of the olfactory sensory map. Cell *114*, 311–322.

Devillers, J., and Domine, D. (1997). Comparison of reliability of log P values calculated from a group contribution approach and from the autocorrelation method. SAR QSAR Environ. Res. *7*, 195–232.

Dey, K.K., Hsiao, C.J., and Stephens, M. (2017). Visualizing the structure of RNA-seq expression data using grade of membership models. PLoS Genet. *13*, e1006599.

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

Duggan, C.D., DeMaria, S., Baudhuin, A., Stafford, D., and Ngai, J. (2008). Foxg1 is required for development of the vertebrate olfactory system. J. Neurosci. *28*, 5229–5239.

Dunkel, A., Steinhaus, M., Kotthoff, M., Nowak, B., Krautwurst, D., Schieberle, P., and Hofmann, T. (2014). Nature’s chemical signatures in human olfaction: a foodborne perspective for future biotechnology. Angew. Chem. Int. Ed. Engl. *53*, 7124–7143.

Efremova, M., Vento-Tormo, M., Teichmann, S.A., and Vento-Tormo, R. (2020). CellPhoneDB: inferring cell-cell communication from combined expression of multi-subunit ligand-receptor complexes. Nat. Protoc. *15*, 1484–1506.

Fienberg, S.E. (1970). An iterative procedure for estimation in contingency tables. Ann. Math. Stat. *41*, 907–917.

Finkielstein, C.V., and Capelluto, D.G. (2016). Disabled-2: a modular scaffold protein with multifaceted functions in signaling. Bioessays *38*, S45–S55.

Fletcher, R.B., Das, D., Gadye, L., Street, K.N., Baudhuin, A., Wagner, A., Cole, M.B., Flores, Q., Choi, Y.G., Yosef, N., et al. (2017). Deconstructing olfactory stem cell trajectories at single-cell resolution. Cell Stem Cell *20*, 817–830.e8.

Floriano, W.B., Vaidehi, N., Goddard, W.A., 3rd, Singer, M.S., and Shepherd, G.M. (2000). Molecular mechanisms underlying differential odor responses of a mouse olfactory receptor. Proc. Natl. Acad. Sci. U S A *97*, 10712–10716.

Freitag, J., Krieger, J., Strotmann, J., and Breer, H. (1995). Two classes of olfactory receptors in Xenopus laevis. Neuron *15*, 1383–1392.

Fulle, H.J., Vassar, R., Foster, D.C., Yang, R.B., Axel, R., and Garbers, D.L. (1995). A receptor guanylyl cyclase expressed specifically in olfactory sensory neurons. Proc. Natl. Acad. Sci. U S A *92*, 3571–3575.

Gaillard, I., Rouquier, S., Pin, J.P., Mollard, P., Richard, S., Barnabe, C., Demaille, J., and Giorgi, D. (2002). A single olfactory receptor specifically binds a set of odorant molecules. Eur. J. Neurosci. *15*, 409–418.

Godfrey, P.A., Malnic, B., and Buck, L.B. (2004). The mouse olfactory receptor gene family. Proc. Natl. Acad. Sci. U S A *101*, 2156–2161.

Goh, C.J., Choi, D., Park, D.B., Kim, H., and Hahn, Y. (2016). MOXD2, a gene possibly associated with olfaction, is frequently inactivated in birds. PLoS One *11*, e0152431.

Greer, P.L., Bear, D.M., Lassance, J.M., Bloom, M.L., Tsukahara, T., Pashkovski, S.L., Masuda, F.K., Nowlan, A.C., Kirchner, R., Hoekstra, H.E., et al. (2016). A Family of non-GPCR chemosensors defines an alternative logic for mammalian olfaction. Cell *165*, 1734–1748.

Grosmaitre, X., Vassalli, A., Mombaerts, P., Shepherd, G.M., and Ma, M. (2006). Odorant responses of olfactory sensory neurons expressing the odorant receptor MOR23: a patch clamp analysis in gene-targeted mice. Proc. Natl. Acad. Sci. U S A *103*, 1970–1975.

Grosmaitre, X., Fuss, S.H., Lee, A.C., Adipietro, K.A., Matsunami, H., Mombaerts, P., and Ma, M. (2009). SR1, a mouse odorant receptor with an unusually broad response profile. J. Neurosci. *29*, 14545–14552.

Gussing, F., and Bohm, S. (2004). NQO1 activity in the main and the accessory olfactory systems correlates with the zonal topography of projection maps. Eur. J. Neurosci. *19*, 2511–2518.

Haghverdi, L., Buettner, F., and Theis, F.J. (2015). Diffusion maps for high-dimensional single-cell analysis of differentiation data. Bioinformatics *31*, 2989–2998.

Haghverdi, L., Buttner, M., Wolf, F.A., Buettner, F., and Theis, F.J. (2016). Diffusion pseudotime robustly reconstructs lineage branching. Nat. Methods. *13*, 845–848.

Hanchate, N.K., Kondoh, K., Lu, Z., Kuang, D., Ye, X., Qiu, X., Pachter, L., Trapnell, C., and Buck, L.B. (2015). Single-cell transcriptomics reveals receptor transformations during olfactory neurogenesis. Science *350*, 1251–1255.

Heron, P.M., Stromberg, A.J., Breheny, P., and McClintock, T.S. (2013). Molecular events in the cell types of the olfactory epithelium during adult neurogenesis. Mol. Brain. *6*, 49.

Hollas, B. (2003). An analysis of the autocorrelation descriptor for molecules. J. Math. Chem. *33*, 91–101.

Horowitz, L.F., Saraiva, L.R., Kuang, D., Yoon, K.H., and Buck, L.B. (2014). Olfactory receptor patterning in a higher primate. J. Neurosci. *34*, 12241–12252.

Huber, A.B., Kolodkin, A.L., Ginty, D.D., and Cloutier, J.F. (2003). Signaling at the growth cone: ligand-receptor complexes and the control of axon growth and guidance. Annu. Rev. Neurosci. *26*, 509–563.

Ibarra-Soria, X., Nakahara, T.S., Lilue, J., Jiang, Y., Trimmer, C., Souza, M.A., Netto, P.H., Ikegami, K., Murphy, N.R., Kusma, M., et al. (2017). Variation in olfactory neuron repertoires is genetically controlled and environmentally modulated. Elife *6*, e21476.

Jiang, Y., Gong, N.N., Hu, X.S., Ni, M.J., Pasi, R., and Matsunami, H. (2015). Molecular profiling of activated olfactory neurons identifies odorant receptors for odors *in vivo*. Nat. Neurosci. *18*, 1446–1454.

Jones, E.M., Jajoo, R., Cancilla, D., Lubock, N.B., Wang, J., Satyadi, M., Cheung, R., de March, C., Bloom, J.S., Matsunami, H., et al. (2019). A scalable, multiplexed assay for decoding GPCR-ligand interactions with RNA sequencing. Cell Syst. *8*, 254–260.e6.

Juilfs, D.M., Fulle, H.J., Zhao, A.Z., Houslay, M.D., Garbers, D.L., and Beavo, J.A. (1997). A subset of olfactory neurons that selectively express cGMP-stimulated phosphodiesterase (PDE2) and guanylyl cyclase-D define a unique olfactory signal transduction pathway. Proc. Natl. Acad. Sci. U S A *94*, 3388–3395.

Jung, D.J., Bong, J.J., and Baik, M. (2004). Extracellular proteinase inhibitor-accelerated apoptosis is associated with B cell activating factor in mammary epithelial cells. Exp. Cell Res. *292*, 115–122.

Junker, J.P., Noel, E.S., Guryev, V., Peterson, K.A., Shah, G., Huisken, J., McMahon, A.P., Berezikov, E., Bakkers, J., and van Oudenaarden, A. (2014). Genome-wide RNA tomography in the zebrafish embryo. Cell *159*, 662–675.

Kajiya, K., Inaki, K., Tanaka, M., Haga, T., Kataoka, H., and Touhara, K. (2001). Molecular bases of odor discrimination: reconstitution of olfactory receptors that recognize overlapping sets of odorants. J. Neurosci. *21*, 6018–6025.

Kandel, E., Schwartz, J., Jessell, T., Siegelbaum, S., and Hudspeth, A. (2013). In Principles of Neural Science, Fifth edition (McGraw-Hill).

Kania, A., and Klein, R. (2016). Mechanisms of ephrin-Eph signalling in development, physiology and disease. Nat. Rev. Mol. Cell. Biol. *17*, 240–256.

Kim, D.S., Wang, Y., Oh, H.J., Lee, K., and Hahn, Y. (2014). Frequent loss and alteration of the MOXD2 gene in catarrhines and whales: a possible connection with the evolution of olfaction. PLoS One *9*, e104085.

Kinsella, R.J., Kahari, A., Haider, S., Zamora, J., Proctor, G., Spudich, G., Almeida-King, J., Staines, D., Derwent, P., Kerhornou, A., et al. (2011). Ensembl BioMarts: a hub for data retrieval across taxonomic space. Database *2011*, bar030.

Koglin, M., Vehse, K., Budaeus, L., Scholz, H., and Behrends, S. (2001). Nitric oxide activates the beta 2 subunit of soluble guanylyl cyclase in the absence of a second subunit. J. Biol. Chem. *276*, 30737–30743.

Kurian, S.M., Naressi, R.G., Manoel, D., Barwich, A.S., Malnic, B., and Saraiva, L.R. (2021). Odor coding in the mammalian olfactory epithelium. Cell Tissue Res. *383*, 445–456.

Lee, H.K., Park, S.J., Oh, H.J., Kim, J.W., Bae, H.S., and Park, J.C. (2012). Expression pattern, subcellular localization, and functional implications of ODAM in ameloblasts, odontoblasts, osteoblasts, and various cancer cells. Gene Expr. Patterns. *12*, 102–108.

Leung, C.T., Coulombe, P.A., and Reed, R.R. (2007). Contribution of olfactory neural stem cells to tissue maintenance and regeneration. Nat. Neurosci. *10*, 720–726.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The sequence alignment/map format and SAMtools. Bioinformatics *25*, 2078–2079.

Liaw, A., and Wiener, M. (2002). Classification and regression by randomForest. R News *2*, 18–22.

Liberles, S.D., and Buck, L.B. (2006). A second class of chemosensory receptors in the olfactory epithelium. Nature *442*, 645–650.

Ling, G., Gu, J., Genter, M.B., Zhuo, X., and Ding, X. (2004). Regulation of cytochrome P450 gene expression in the olfactory mucosa. Chem. Biol. Interact. *147*, 247–258.

Liu, L., Tang, L., Dong, W., Yao, S., and Zhou, W. (2016). An overview of topic modeling and its current applications in bioinformatics. Springerplus *5*, 1608.

Liu, X., Su, X., Wang, F., Huang, Z., Wang, Q., Li, Z., Zhang, R., Wu, L., Pan, Y., Chen, Y., et al. (2011). ODORactor: a web server for deciphering olfactory coding. Bioinformatics *27*, 2302–2303.

Loader, C. (2007). R Package "locfit": Local Regression, Likelihood and Density Estimation (Version 1).

Lun, A.T., McCarthy, D.J., and Marioni, J.C. (2016). A step-by-step workflow for low-level analysis of single-cell RNA-seq data with bioconductor. F1000Res. *5*, 2122.

Ma, M., and Shepherd, G.M. (2000). Functional mosaic organization of mouse olfactory receptor neurons. Proc. Natl. Acad. Sci. U S A *97*, 12869–12874.

Malnic, B., Godfrey, P.A., and Buck, L.B. (2004). The human olfactory receptor gene family. Proc. Natl. Acad. Sci. U S A *101*, 2584–2589.

Malnic, B., Hirono, J., Sato, T., and Buck, L.B. (1999). Combinatorial receptor codes for odors. Cell *96*, 713–723.

Manoel, D., Makhlouf, M., Arayata, C.J., Sathappan, A., Da'as, S., Abdelrahman, D., Selvaraj, S., Hasnah, R., Mainland, J.D., Gerkin, R.C., et al. (2021). Deconstructing the mouse olfactory percept through an ethological atlas. Curr. Biol. *31*, 2809–2818.e3.

Marchand, J.E., Yang, X., Chikaraishi, D., Krieger, J., Breer, H., and Kauer, J.S. (2004). Olfactory receptor gene expression in tiger salamander olfactory epithelium. J. Comp. Neurol. *474*, 453–467.

McInnes, L., Healy, J., and Melville, J. (2018). UMAP: uniform manifold approximation and projection for dimension reduction. Preprint at arXiv. https://doi.org/10.48550/arXiv.1802.03426.

Miyamichi, K., Serizawa, S., Kimura, H.M., and Sakano, H. (2005). Continuous and overlapping expression domains of odorant receptor genes in the olfactory epithelium determine the dorsal/ventral positioning of glomeruli in the olfactory bulb. J. Neurosci. *25*, 3586–3592.

Miyawaki, A., Homma, H., Tamura, H., Matsui, M., and Mikoshiba, K. (1996). Zonal distribution of sulfotransferase for phenol in olfactory sustentacular cells. EMBO J. *15*, 2050–2055.

Miyawaki, A., Matsushita, F., Ryo, Y., and Mikoshiba, K. (1994). Possible pheromone-carrier function of two lipocali proteins in the vomeronasal organ. EMBO J. *13*, 5835–5842.

Mombaerts, P., Wang, F., Dulac, C., Chao, S.K., Nemes, A., Mendelsohn, M., Edmondson, J., and Axel, R. (1996). Visualizing an olfactory sensory map. Cell *87*, 675–686.

Monahan, K., Schieren, I., Cheung, J., Mumbey-Wafula, A., Monuki, E.S., and Lomvardas, S. (2017). Cooperative interactions enable singular olfactory receptor expression in mouse olfactory neurons. Elife *6*, e28620.

Mozell, M.M. (1966). The spatiotemporal analysis of odorants at the level of the olfactory receptor sheet. J. Gen. Physiol. *50*, 25–41.

Nara, K., Saraiva, L.R., Ye, X., and Buck, L.B. (2011). A large-scale analysis of odor coding in the olfactory epithelium. J. Neurosci. *31*, 9179–9191.

Nguyen, M.Q., Zhou, Z., Marks, C.A., Ryba, N.J., and Belluscio, L. (2007). Prominent roles for odorant receptor coding sequences in allelic exclusion. Cell *131*, 1009–1017.

Nickell, M.D., Breheny, P., Stromberg, A.J., and McClintock, T.S. (2012). Genomics of mature and immature olfactory sensory neurons. J. Comp. Neurol. *520*, 2608–2629.

Niimura, Y., Matsui, A., and Touhara, K. (2014). Extreme expansion of the olfactory receptor gene repertoire in African elephants and evolutionary dynamics of orthologous gene groups in 13 placental mammals. Genome Res. *24*, 1485–1496.

Niimura, Y., Matsui, A., and Touhara, K. (2018). Acceleration of olfactory receptor gene loss in primate evolution: possible link to anatomical change in sensory systems and dietary transition. Mol. Biol. Evol. *35*, 1437–1450.

Norlin, E.M., Alenius, M., Gussing, F., Hagglund, M., Vedin, V., and Bohm, S. (2001). Evidence for gradients of gene expression correlating with zonal topography of the olfactory sensory map. Mol. Cell. Neurosci. *18*, 283–295.

Octura, J.E.R., Maeda, K.I., and Wakabayashi, Y. (2018). Structure and zonal expression of olfactory receptors in the olfactory epithelium of the goat, Capra hircus. J. Vet. Med. Sci. *80*, 913–920.

Oka, Y., Katada, S., Omura, M., Suwa, M., Yoshihara, Y., and Touhara, K. (2006). Odorant receptor map in the mouse olfactory bulb: *in vivo* sensitivity and specificity of receptor-defined glomeruli. Neuron *52*, 857–869.

Oka, Y., Kobayakawa, K., Nishizumi, H., Miyamichi, K., Hirose, S., Tsuboi, A., and Sakano, H. (2003). O-MACS, a novel member of the medium-chain acyl-CoA synthetase family, specifically expressed in the olfactory epithelium in a zone-specific manner. Eur. J. Biochem. *270*, 1995–2004.

Oka, Y., Omura, M., Kataoka, H., and Touhara, K. (2004). Olfactory receptor antagonism between odorants. EMBO J. *23*, 120–126.

Oka, Y., Takai, Y., and Touhara, K. (2009). Nasal airflow rate affects the sensitivity and pattern of glomerular odorant responses in the mouse olfactory bulb. J. Neurosci. *29*, 12070–12078.

Omura, M., and Mombaerts, P. (2015). Trpc2-expressing sensory neurons in the mouse main olfactory epithelium of type B express the soluble guanylate cyclase Gucy1b2. Mol. Cell. Neurosci. *65*, 114–124.

Pacifico, R., Dewan, A., Cawley, D., Guo, C., and Bozza, T. (2012). An olfactory subsystem that mediates high-sensitivity detection of volatile amines. Cell Rep. *2*, 76–88.

Paradis, E., Claude, J., and Strimmer, K. (2004). APE: analyses of phylogenetics and evolution in R language. Bioinformatics *20*, 289–290.

Park, J., Levic, D.S., Sumigray, K.D., Bagwell, J., Eroglu, O., Block, C.L., Eroglu, C., Barry, R., Lickwar, C.R., Rawls, J.F., et al. (2019). Lysosome-rich enterocytes mediate protein absorption in the vertebrate gut. Dev. Cell *51*, 7–20.e6.

Peluso, C.E., Jang, W., Drager, U.C., and Schwob, J.E. (2012). Differential expression of components of the retinoic acid signaling pathway in the adult mouse olfactory epithelium. J. Comp. Neurol. *520*, 3707–3726.

Peng, G., Suo, S., Chen, J., Chen, W., Liu, C., Yu, F., Wang, R., Chen, S., Sun, N., Cui, G., et al. (2016). Spatial transcriptome for the molecular annotation of lineage fates and cell identity in mid-gastrula mouse embryo. Dev. Cell *36*, 681–697.

Pfister, P., Smith, B.C., Evans, B.J., Brann, J.H., Trimmer, C., Sheikh, M., Arroyave, R., Reddy, G., Jeong, H.Y., Raps, D.A., et al. (2020). Odorant receptor inhibition is fundamental to odor encoding. Curr. Biol. *30*, 2574–2587.e6.

Rather, T.A., Kumar, S., and Khan, J.A. (2020). Multi-scale habitat modelling and predicting change in the distribution of tiger and leopard using random forest algorithm. Sci. Rep. *10*, 11473.

Rehurek, R., and Sojka, P. (2010). Software framework for topic modelling with large corpora. In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks (ELRA), pp. 45–50.

Repicky, S.E., and Luetje, C.W. (2009). Molecular receptive range variation among mouse odorant receptors for aliphatic carboxylic acids. J. Neurochem. *109*, 193–202.

Ressler, K.J., Sullivan, S.L., and Buck, L.B. (1993). A zonal organization of odorant receptor gene expression in the olfactory epithelium. Cell *73*, 597–609.

Ressler, K.J., Sullivan, S.L., and Buck, L.B. (1994). Information coding in the olfactory system: evidence for a stereotyped and highly organized epitope map in the olfactory bulb. Cell *79*, 1245–1255.

Rygg, A.D., Van Valkenburgh, B., and Craven, B.A. (2017). The influence of sniffing on airflow and odorant deposition in the canine nasal cavity. Chem. Senses *42*, 683–698.

Saito, H., Kubota, M., Roberts, R.W., Chi, Q., and Matsunami, H. (2004). RTP family members induce functional expression of mammalian odorant receptors. Cell *119*, 679–691.

Saito, H., Nishizumi, H., Suzuki, S., Matsumoto, H., Ieki, N., Abe, T., Kiyonari, H., Morita, M., Yokota, H., Hirayama, N., et al. (2017). Immobility responses are induced by photoactivation of single glomerular species responsive to fox odour TMT. Nat. Commun. *8*, 16011.

Saraiva, L.R., Ahuja, G., Ivandic, I., Syed, A.S., Marioni, J.C., Korsching, S.I., and Logan, D.W. (2015a). Molecular and neuronal homology between the olfactory systems of zebrafish and mouse. Sci. Rep. *5*, 11487.

Saraiva, L.R., Ibarra-Soria, X., Khan, M., Omura, M., Scialdone, A., Mombaerts, P., Marioni, J.C., and Logan, D.W. (2015b). Hierarchical deconstruction of mouse olfactory sensory neurons: from whole mucosa to single-cell RNA-seq. Sci. Rep. *5*, 18178.

Saraiva, L.R., Riveros-McKay, F., Mezzavilla, M., Abou-Moussa, E.H., Arayata, C.J., Makhlouf, M., Trimmer, C., Ibarra-Soria, X., Khan, M., Van Gerven, L., et al. (2019). A transcriptomic atlas of mammalian olfactory mucosae reveals an evolutionary influence on food odor detection in humans. Sci. Adv. *5*, eaax0396.

Schmal, C., Myung, J., Herzel, H., and Bordyugov, G. (2017). Moran's I quantifies spatio-temporal pattern formation in neural imaging data. Bioinformatics *33*, 3072–3079.

Schoenfeld, T.A., and Cleland, T.A. (2006). Anatomical contributions to odorant sampling and representation in rodents: zoning in on sniffing behavior. Chem. Senses *31*, 131–144.

Scott, J.W., and Scott-Johnson, P.E. (2002). The electroolfactogram: a review of its history and uses. Microsc. Res. Tech. *58*, 152–160.

Scott, J.W., Sherrill, L., Jiang, J., and Zhao, K. (2014). Tuning to odor solubility and sorption pattern in olfactory epithelial responses. J. Neurosci. *34*, 2025–2036.

Secundo, L., Snitz, K., and Sobel, N. (2014). The perceptual logic of smell. Curr. Opin. Neurobiol. *25*, 107–115.

Secundo, L., Snitz, K., Weissler, K., Pinchover, L., Shoenfeld, Y., Loewenthal, R., Agmon-Levin, N., Frumin, I., Bar-Zvi, D., Shushan, S., et al. (2015). Individual olfactory perception reveals meaningful nonolfactory genetic information. Proc. Natl. Acad. Sci. U S A *112*, 8750–8755.

Shin, Y., Won, Y., Yang, J.I., and Chun, J.S. (2019). CYTL1 regulates bone homeostasis in mice by modulating osteogenesis of mesenchymal stem cells and osteoclastogenesis of bone marrow-derived macrophages. Cell Death Dis. *10*, 47.

Shirasu, M., Yoshikawa, K., Takai, Y., Nakashima, A., Takeuchi, H., Sakano, H., and Touhara, K. (2014). Olfactory receptor and neural pathway responsible for highly selective sensing of musk odors. Neuron *81*, 165–178.

Shirokova, E., Schmiedeberg, K., Bedner, P., Niessen, H., Willecke, K., Raguse, J.D., Meyerhof, W., and Krautwurst, D. (2005). Identification of specific ligands for orphan olfactory receptors. G protein-dependent agonism and antagonism of odorants. J. Biol. Chem. *280*, 11807–11815.

Simes, R.J. (1986). An improved Bonferroni procedure for multiple tests of significance. Biometrika *73*, 751–754.

Springer, M.S., Emerling, C.A., Gatesy, J., Randall, J., Collin, M.A., Hecker, N., Hiller, M., and Delsuc, F. (2019). Odontogenic ameloblast-associated (ODAM) is inactivated in toothless/enamelless placental mammals and toothed whales. BMC Evol. Biol. *19*, 31.

Spuch, C., Ortolano, S., and Navarro, C. (2012). LRP-1 and LRP-2 receptors function in the membrane neuron. Trafficking mechanisms and proteolytic processing in Alzheimer's disease. Front. Physiol. *3*, 269.

Strotmann, J., Wanner, I., Krieger, J., Raming, K., and Breer, H. (1992). Expression of odorant receptors in spatially restricted subsets of chemosensory neurones. Neuroreport *3*, 1053–1056.

Sullivan, S.L., Adamson, M.C., Ressler, K.J., Kozak, C.A., and Buck, L.B. (1996). The chromosomal distribution of mouse odorant receptor genes. Proc. Natl. Acad. Sci. U S A *93*, 884–888.

Taddy, M. (2012). On estimation and selection for topic models. In Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics, D.L. Neil and G. Mark, eds., pp. 1184–1193, Proceedings of Machine Learning Research: PMLR.

Tan, L., and Xie, X.S. (2018). A near-complete spatial map of olfactory receptors in the mouse main olfactory epithelium. Chem. Senses *43*, 427–432.

Tietjen, I., Rihel, J., and Dulac, C.G. (2005). Single-cell transcriptional profiles and spatial patterning of the mammalian olfactory epithelium. Int. J. Dev. Biol. *49*, 201–207.

Tietjen, I., Rihel, J.M., Cao, Y., Koentges, G., Zakhary, L., and Dulac, C. (2003). Single-cell transcriptional analysis of neuronal progenitors. Neuron *38*, 161–175.

Trimmer, C., and Mainland, J.D. (2017). Simplifying the odor landscape. Chem. Senses *42*, 177–179.

Vassar, R., Chao, S.K., Sitcheran, R., Nunez, J.M., Vosshall, L.B., and Axel, R. (1994). Topographic organization of sensory projections to the olfactory bulb. Cell *79*, 981–991.

Vassar, R., Ngai, J., and Axel, R. (1993). Spatial segregation of odorant receptor expression in the mammalian olfactory epithelium. Cell *74*, 309–318.

Vedin, V., Molander, M., Bohm, S., and Berghard, A. (2009). Regional differences in olfactory epithelial homeostasis in the adult mouse. J. Comp. Neurol. *513*, 375–384.

von der Weid, B., Rossier, D., Lindup, M., Tuberosa, J., Widmer, A., Col, J.D., Kan, C., Carleton, A., and Rodriguez, I. (2015). Large-scale transcriptional profiling of chemosensory neurons identifies receptor-ligand pairs *in vivo*. Nat. Neurosci. *18*, 1455–1463.

Wang, Q., Titlow, W.B., McClintock, D.A., Stromberg, A.J., and McClintock, T.S. (2017). Activity-Dependent gene expression in the mammalian olfactory epithelium. Chem. Senses *42*, 611–624.

Wang, S.S., Lewcock, J.W., Feinstein, P., Mombaerts, P., and Reed, R.R. (2004). Genetic disruptions of O/E2 and O/E3 genes reveal involvement in olfactory receptor neuron projection. Development *131*, 1377–1388.

Wang, T., Wang, Z.Y., Zeng, L.Y., Gao, Y.Z., Yan, Y.X., and Zhang, Q. (2020). Down-regulation of ribosomal protein rps21 inhibits invasive behavior of osteosarcoma cells through the inactivation of MAPK pathway. Cancer Manag. Res. *12*, 4949–4955.

Weth, F., Nadler, W., and Korsching, S. (1996). Nested expression domains for odorant receptors in zebrafish olfactory epithelium. Proc. Natl. Acad. Sci. U S A *93*, 13321–13326.

Whitby-Logan, G.K., Weech, M., and Walters, E. (2004). Zonal expression and activity of glutathione S-transferase enzymes in the mouse olfactory mucosa. Brain Res. *995*, 151–157.

Wolf, F.A., Angerer, P., and Theis, F.J. (2018). SCANPY: large-scale single-cell gene expression data analysis. Genome Biol. *19*, 15.

Yang, L.M., Huh, S.H., and Ornitz, D.M. (2018). FGF20-expressing, wnt-responsive olfactory epithelial progenitors regulate underlying turbinate growth to optimize surface area. Dev. Cell *46*, 564–580.e5.

Yoshihara, Y., Kawasaki, M., Tamada, A., Fujita, H., Hayashi, H., Kagamiyama, H., and Mori, K. (1997). OCAM: a new member of the neural cell adhesion molecule family related to zone-to-zone projection of olfactory and vomeronasal axons. J. Neurosci. *17*, 5830–5842.

Yoshikawa, K., Nakagawa, H., Mori, N., Watanabe, H., and Touhara, K. (2013). An unsaturated aliphatic alcohol as a natural ligand for a mouse odorant receptor. Nat. Chem. Biol. *9*, 160–162.

Yoshikawa, K., and Touhara, K. (2009). Myr-Ric-8A enhances G(alpha15)-mediated Ca2+ response of vertebrate olfactory receptors. Chem. Senses *34*, 15–23.

Yoshikawa, K., Wang, H., Jaen, C., Haneoka, M., Saito, N., Nakamura, J., Adappa, N.D., Cohen, N.A., and Dalton, P. (2018). The human olfactory cleft mucus proteome and its age-related changes. Sci. Rep. *8*, 17170.

Yu, T.T., McIntyre, J.C., Bose, S.C., Hardin, D., Owen, M.C., and McClintock, T.S. (2005). Differentially expressed transcripts from phenotypically identified olfactory sensory neurons. J. Comp. Neurol. *483*, 251–262.

Yu, Y., de March, C.A., Ni, M.J., Adipietro, K.A., Golebiowski, J., Matsunami, H., and Ma, M. (2015). Responsiveness of G protein-coupled odorant receptors is partially attributed to the activation mechanism. Proc. Natl. Acad. Sci. U S A *112*, 14966–14971.

Zapiec, B., and Mombaerts, P. (2020). The zonal organization of odorant receptor gene choice in the main olfactory epithelium of the mouse. Cell Rep. *30*, 4220–4234.e5.

Zhang, X., Rogers, M., Tian, H., Zhang, X., Zou, D.J., Liu, J., Ma, M., Shepherd, G.M., and Firestein, S.J. (2004). High-throughput microarray detection of olfactory receptor gene expression in the mouse. Proc. Natl. Acad. Sci. U S A *101*, 14168–14173.

Zhu, S., Kuek, V., Bennett, S., Xu, H., Rosen, V., and Xu, J. (2019). Protein Cytl1: its role in chondrogenesis, cartilage homeostasis, and disease. Cell Mol. Life Sci. *76*, 3515–3523.

Zhuang, H., and Matsunami, H. (2007). Synergism of accessory factors in functional expression of mammalian odorant receptors. J. Biol. Chem. *282*, 15284–15293.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Anti-Digoxigenin-AP, Fab fragments	Merck (Roche)	Cat# 11093274910, RRID:AB_514497
Biological samples		
Olfactory mucosae from C57Bl/6J mice (adult males)	The Jackson Laboratory	Stock # 00664
Chemicals, peptides, and recombinant proteins		
30% Hydrogen Proxyde Solution	Merck (Sigma-Aldrich)	Cat. # H1009
Triethanolamine	Merck (Sigma-Aldrich)	Cat. # T58300
Acetic anhydride	Merck (Sigma-Aldrich)	Cat. # 320102
Deoinized formamide	Merck (Sigma-Aldrich)	Cat. # F9037
Yeast tRNA	Merck (Roche)	Cat. # 10109495001
Denhardt's solution (50×)	Merck (Sigma-Aldrich)	Cat. # D9905
Dextran sulfate solution (50%)	Merck (Chemicon)	Cat. # S4030
20× SSC	Merck (Calbiochem)	Cat. # 8310-OP
Tween-20	Merck (Sigma-Aldrich)	Cat. # 822184
TSA Blocking Reagent	Perkin-Elmer	Cat. # FP1020
NBT/BCIP Stock Solution	Merck (Roche)	Cat. # 11681451001
Critical commercial assays		
SMART-Seq v4 Ultra Low Input RNA Kit for Sequencing	Clontech (Takara Bio)	Cat. # 634892
Bioanalyzer DNA High-Sensitivity kit	Agilent Technologies	Cat. # 5067-4626
Nextera XT DNA Library Preparation Kit (96 samples)	Illumina	Cat. # FC-131-1096
Nextera XT Index Kit v2 Set A (96 indexes, 384 samples)	Illumina	Cat. # FC-131-2001
pGEM®-T Easy Vector Systems	Promega	Cat. # A1360
DIG RNA Labeling Kit (SP6/T7)	Merck (Roche)	Cat. # 11175025910
ProbeQuant G-50 Micro Columns	Cytiva Biosciences	Cat. # 28903408
Deposited data		
TOMO-seq Olfactory Mucosa dataset	This study	https://www.ebi.ac.uk/arrayexpress/E-MTAB-10211
Single cell RNA-seq data from the Olfactory Mucosa	Fletcher et al., 2017	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE95601
Dragon database of molecular descriptors	Talete S.R.L.	http://www.talete.mi.it
CellphoneDB ligands and receptors database	Efremova et al., 2020	https://github.com/ventolab/CellphoneDB
Experimental models: Organisms/strains		
Adult male C57Bl/6J mice	The Jackson Laboratory	Stock # 00664
Oligonucleotides		
See "method details" section for oligonucleotides	This study	N/A
Software and algorithms		
samtools version 0.1.19-44428cd	Li et al., 2009	http://samtools.sourceforge.net/
htseq-count version 0.11.2	Anders et al., 2014	https://github.com/htseq/htseq/
R 4.1.2	The R Foundation	https://www.r-project.org/

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Python 3.9.6	Python Software Foundation	https://www.python.org/
Scripts for TOMO-seq data analysis	This study	https://doi.org/10.5281/zenodo.6036047 https://zenodo.org/badge/DOI/10.5281/zenodo.6045897.svg

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and data should be directed to and will be fulfilled by the Lead Contact Luis R. Saraiva (saraivalmr@gmail.com).

Materials availability

This study did not generate new unique reagents.

Data and code availability

RNA-seq raw data have been deposited and are publicly available as of the date of publication at ArrayExpress: E-MTAB-10211. All original code and scripts for the 3D nose atlas shiny app has been deposited at Github and can be found at the Github Repository: https://doi.org/10.5281/zenodo.6036047https://zenodo.org/badge/DOI/10.5281/zenodo.6045897.svg. The 3D nose atlas processed data can be browsed and visualized here: http://atlas3dnose.helmholtz-muenchen.de:3838/atlas3Dnose.

Any additional information required to reanalyze the data reported in this paper is available from the lead contacts upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Animals

The animals used in this study were adult male C57Bl/6J mice (aged 8–14 weeks, The Jackson Laboratory, Stock # 00664) maintained in group-housed conditions on a 12:12 h light:dark schedule (lights on at 0700 hours). Each mouse was randomly assigned for cryosectioning along one of the three cartesian axes.

The use and care of animals used in this study was approved by the Internal Animal Care and Use Committee (IACUC) of Monell Chemical Senses Center, by the IACUC of the University of São Paulo, and by the Wellcome Trust Sanger Institute Animal Welfare and Ethics Review Board in accordance with UK Home Office regulations, the UK Animals (Scientific Procedures) Act of 1986.

METHOD DETAILS

Dissection of the olfactory mucosa, cryosections, and RNA-sequencing

The olfactory mucosa (OM) of 9 mice was carefully dissected, and all the surrounding tissue (including glands and bone) removed – this was necessary to ensure that the transcripts present in the surrounding tissue do not contaminate the RNA isolated from the OM. The OMs were then embedded in OCT (Tissue Tek), immediately frozen in dry-ice and kept at −80°C. Each OM was then cryosectioned along each of the 3 cartesian axes: dorsal-ventral (DV, n = 3), anterior-posterior (AP, n = 3), or lateral-medial-lateral (LML, N = 3). Every second cryosections (35 μm thick) was collected into 1.5 mL eppendorf tubes containing 350 μL RLT Plus Buffer (Qiagen) supplemented with 1% 2-mercaptoethanol, immediately frozen in dry-ice and kept at −80°C until extraction. RNA was extracted using the RNeasy Plus Micro Kit (Qiagen), together with a genomic DNA eliminator column and a 30-minute incubation with DNAse I (Qiagen). Reverse transcription and cDNA pre-amplification were performed using the SMART-Seq v4 Ultra Low Input RNA Kit for Sequencing (Clontech/Takara). cDNA was harvested and quantified with the Bioanalyzer DNA High-Sensitivity kit (Agilent Technologies). Libraries were prepared using the Nextera XT DNA Sample Preparation Kit and the Nextera Index Kit (Illumina). Multiplexed libraries were pooled and paired-end 150-bp sequencing was performed on the Illumina HiSeq 4000 platform at Sidra Medicine, except for one library (DV-I) for which 125-bp paired-end sequencing was performed on the Illumina HiSeq 2500 platform at the Wellcome Sanger Institute. The raw data are available through ArrayExpress under accession number E-MTAB-10211.

RNA-seq data mapping and gene counting

Reads were aligned to the mm10 mouse genome (release 99). The sequences of the genes "*Xntrpc*" and "*Capn5*" were removed from the genome files as in Saraiva et al. (2015b). The alignment was performed with the software STAR version 2.7.3a (Dobin et al., 2013). Genome indexes were generated using STAR –runMode genomeGenerate with default parameters. Then, alignment of reads was performed with the following options: –runThreadN 48 –outSAMunmapped Within –outFilterMultimapNmax 1000 –outFilterMismatchNmax 4 –outFilterMatchNmin 100 –alignIntronMax 50000 –alignMatesGapMax 50500 –outSAMstrandField intronMotif –outFilterType BySJout. The resulting SAM files were converted to bam format and sorted using samtools (version 0.1.19-44428cd) (Li et al., 2009). The multi mapping reads were eliminated using the same software (samtools view -q 255). Finally, the reads for each gene were counted using htseq-count (version 0.11.2) with the options -m intersection-nonempty -s no -i gene_name -r pos (Anders et al., 2014).

Quality control

We excluded all the samples that fulfilled any of these criteria: they had less than 50% mapped reads, less than 4,000 detected genes, more than 20% mitochondrial reads, less than 10,000 total number of reads, or did not express any of the 3 canonical OSN markers *Omp*, *Cnga* and *Gnal*. This resulted in ~51 good-quality sections along the DV axis (~84% out of the collected sections), ~76 (~91% of total) along the AP axis and ~59 (~93% of total) along the LML axis, as averaged across the three replicates per axis.

Data normalization

Gene expression counts were normalized by reads-per-million (RPM), then genes detected in only one replicate and genes that were detected in less than 10% of all samples along one axis were eliminated. To check the similarity between replicates, we calculated Spearman correlations between the transcriptional profiles of sections along each axis (using the top 1000 Highly Variable Genes per axis). Close positions had the most similar transcriptional profiles (Figure 1C). Afterward, the 3 replicates for each axis were aligned as follows: the top 3,000 highly variable genes (HVGs) from each replicate were identified using the method implemented in the scran library in R (Lun et al., 2016) and the intersection of these 3 groups was used in the next steps. For the replicates' alignment, we took as reference the replicate with the smallest number of slices. We used a sliding window approach that identified the range of consecutive positions on each replicate along which the average value of the Spearman's correlation coefficient computed with the reference replicate over the HVG was maximum (mean Spearman's Rho = 0.80, $p < 0.05$). To mitigate batch effects, the level of every gene was scaled in such a way that their average value in each replicate was equal to the average calculated across all replicates. After this scaling transformation, the data was then averaged between replicates. Once the 3 biological replicates were combined, we had 54 sections along the DV axis, 60 along the AP and 56 along the LML. Along the LML axis a symmetric pattern of expression is expected around the central position, where the septal bone is located. To confirm this in our data, first we identified the central position by analyzing the expression pattern of neuronal markers like *Cnga2*, *Omp* and *Gnal*, whose expression is lowest in the area around the septal bone. Indeed, all three marker genes reach a minimum at the same position along the LML axis (slice 28), which we considered to be the center. The expression patterns of ~90% of genes on either side of the central position show a positive correlation, and ~70% reach statistical significance (Spearman's correlation computed on the highly variable genes having more than 50 normalized counts in at least 3 slices), further supporting the hypothesis of the bilateral symmetry. Hence, after replicates were averaged, LML axis was made symmetric averaging positions 1:28 and 56:29. Moreover, *Olfrs* were normalized by the geometric mean of neuronal markers *Omp*, *Gnal* and *Cnga2*, as done previously (Ibarra-Soria et al., 2017).

To verify the presence of a spatial signal, we calculated the Moran's I and the associated p-values for the top 100 Highly Variable genes along each axis using the "Moran.I" function from the "ape" library in R with default parameters (Paradis et al., 2004). The p-values of the genes along each axis were combined with the Simes' method (Simes, 1986) using the function combinePValues from the scran R library (Figure S1E).

Identification of differentially expressed genes and gene clustering

Before testing for differential expression along a given axis, we filtered out genes whose expression levels had low variability. To this aim, for each gene we estimated their highest and lowest expression by taking the average of its three highest and three lowest values respectively. Then, we considered for downstream analyses only the genes that meet either of these two criteria: the highest expression value is greater than or equal to 5 normalized counts and the fold-change between the highest and lowest value is greater than 2; or the difference between the highest and the lowest value is greater than or equal to 4 normalized counts. The expression levels of the genes were binarized according to whether their value was higher or lower than their median expression along the axis. Finally, we used the "ts" function in R to transform the binarized expression values into time series objects, and we applied on them the Ljung-Box test (Box.test function in R with lag = (axis length)-10) which identifies genes with statistically significant autocorrelations, i.e., with non-random expression patterns along an axis. The resulting p-values were adjusted using the FDR method and genes with an FDR <0.01 were considered as differentially expressed. For the next steps, the $\log_{10}$ normalized expression of differentially expressed genes along each axis was fitted with a local regression using the locfit function in the R library locfit (Loader, 2007). Smoothing was defined in the local polynomial model term of the locfit model using the function "lp" from the same library with the following parameters: nn = 1 (Nearest neighbor component of the smoothing parameter) and deg = 2 (degree of polynomial). The fitted expression values of these genes along each axis were normalized between 0 and 1. Clustering was performed separately for each axis on the fitted and normalized patterns of the differentially expressed genes. We used the R function "hclust" to perform hierarchical clustering on the gene expression patterns, with a Spearman's correlation-based distance (defined as $\sqrt{0.5\,(1-\rho)}$) and the "average" aggregation method. The number of clusters were defined with the cutreeDynamic function from the dynamicTreeCut R library, with the parameters minClusterSize = 50, method = "hybrid" and deepSplit = 0. To visualize the data in two dimensions, we applied the UMAP dimensionality reduction algorithm (umap function in the R library umap with default options; see Figure 2D) (Becht et al., 2018; McInnes et al., 2018). To analyze the relationship between the expression patterns of genes along different axes, we computed the intersections of the gene clusters between any pair of axes. The expected number of elements in each intersection under the assumption of independent sets is given by:

$$|A \cap B|_{exp} = \frac{|A||B|}{|A \cup B|}$$

where A and B indicate the sets of genes in two clusters identified along two different axes and $|\cdot|$ indicates the cardinality of a set (i.e., the number of its elements). The ratio between the observed and the expected number of elements in the intersection $|A \cap B|_{obs} \ / \ |A \cap B|_{exp}$ quantifies the enrichment/depletion of genes having a given pair of patterns across two axes with respect to the random case. The $\log_2$ values of $(1 + |A \cap B|_{obs} \ / \ |A \cap B|_{exp})$ are shown in Figure 2F.

Combining Tomo-seq with single-cell RNA-seq data

The TPM (transcripts per million)-normalized single cell RNA-seq (scRNA-seq) data collected from mouse olfactory epithelium available from Fletcher et al. (2017) was used to identify cell-type specific genes. To this aim, we computed the average expression level for each cell type in the scRNA-seq dataset for all the differentially expressed genes that we identified in our TOMO-seq data. The genes with an average expression above 100 TPM in mOSNs and below 10 TPM in all other cell types were considered mOSN-specific. Conversely, genes with an average expression above 100 TPM in any of the non-mOSN cell types and below 10 in mOSNs were considered to be specific for non-mOSN cells.

Gene ontology (GO) enrichment analysis

GO Enrichment analyses were performed using the GOrilla online tool (http://cbl-gorilla.cs.technion.ac.il) with the option "Two unranked lists of genes (target and background lists)". For each axis, we used as background list the list of the genes we tested.

Cell type deconvolution analysis

To perform cell type deconvolution analysis, we used a previously published single-cell RNA-seq (scRNA-seq) data from the mouse OM (Fletcher et al., 2017). First, the cells included in unclassified clusters were removed and the data was rescaled using the function "pp.log1p" from the scanpy library (Wolf et al., 2018). Then, we obtained 2000 highly variable genes using the function "pp.highly_variable_genes" (scanpy library). In the following analysis, we merged clusters of similar cell populations and considered the following 6 cell types: 1-HBC = HBC1+HBC2+HBC3; 2-INP = INP1+INP2+INP3; 3-GBC = GBC, 4-SC = mSC + iSC, 5-OSN = iOSN + mOSN, 6-MVC = MVC1+MVC2.

This scRNA-seq data was used as input for the AutoGeneS algorithm (Aliee and Theis, 2021). The cell type assignment as well as the list of highly variable genes were passed as input to the function "ag.init" from AutogeneS, and then we estimated the optimal subset of genes to perform cell type deconvolution with the function "ag.optimize" (with parameters: "ngen" = 5000, "nfeatures" = 400 and "mode" = "fixed"). Finally, we deconvolved the Tomo-seq data along the three axes with the function "ag.deconvolve" using Nu Support Vector regression models ("model" = "nusvr"). The results were normalized such that the sums of cell type proportions per slice is equal to 1 (Figure S2F). To identify the cell types with non-random spatial distribution along the axes, we applied the Ljung-Box test as explained above (section "identification of differentially expressed genes and gene clustering"); the p values are reported in Table S2.

Identification of ligands and receptors associated with the NfiA, NfiB or NfiX transcription factors

The genes in the CellphoneDB ligands and receptor database (Efremova et al., 2020) that were among our spatially differentially expressed genes were selected and Spearman correlation tests between their 1D expression patterns and the 1D patterns for the Nfi transcription factors were performed. Correlation coefficients from the three axes were averaged and FDRs from the 3 axes were combined with the Simes' method (Simes, 1986) using the function combinePValues from the scran R library. Combined FDR values <0.01 were taken as significant.

3D spatial reconstruction

The olfactory mucosa shape was obtained from publicly available images of the mouse nasal cavity along the posterior to the anterior axis published in Barrios et al. (2014). The area of the slices corresponding to the OM was manually selected and images of their silhouettes were made. Those images were then transformed into binary matrices having 1's in the area occupied by the OM and 0's in the remaining regions. The binary matrices were rescaled to match the spatial resolution in our dataset, which is composed of 54 voxels along the DV axis, 56 along the LML axis and 60 along the AP axis. Finally, matrices were piled in a 3D array in R to obtain an in-silico representation of the 3D shape of the OM, which, in total, was composed of 77,410 voxels. To perform the 3D reconstruction of the expression pattern for a given gene, first we normalized its expression levels by the volume of the slice at each corresponding position along the three axes, which was estimated using our 3D in silico representation of the OM. Then, we rescaled the data in such a way that the sum of the expression levels along each axis was equal to the average expression computed across the whole dataset. This rescaled dataset together with the binary matrix representing the 3D OM shape was used as input of the Iterative Proportional Fitting algorithm, which produced an estimation of the expression level of a gene in each voxel (Junker et al., 2014). Iterations stopped when the differences between the true and the reconstructed 1D values summed across the three axes was smaller than 1.

Definition of zones by topic modelling

In order to identify zones, we fitted a Latent Dirichlet Allocation (LDA) (Blei et al., 2003) algorithm to the 3D gene expression patterns (in $\log_{10}$ scale) of the differentially expressed *Olfrs* (689 *Olfrs* x 77,410 voxels).

The LDA algorithm was originally employed for document classification: based on the words included in each document, LDA can identify "topics", in which the documents can then be classified. Using this linguistic analogy, in our application of LDA, we considered the genes as “documents”, and the spatial locations as “words”, with the matrix of gene expression levels being the analogous of the "bag-of-word" matrix (Liu et al., 2016). In this representation, the zones are the equivalent of “topics”, and they are automatically identified by LDA. We used the LDA implementation included in the R package “Countclust” (Dey et al., 2017), developed based on the “maptpx” library (Taddy, 2012), which performs a maximum a posteriori estimation to for model fitting. LDA was run for all possible numbers of topics $K \in [2,9]$. The following parameters were chosen: convergence tolerance = 0.1; max time optimization step = 180 seconds; n_init = 3. For each number of topics k, three independent runs were performed with different starting points, in order to avoid biases due to the choice of the initial condition. We estimated the number of topics by computing the log likelihood for each value of $K \in [2,9]$. As seen in Figure S3A, while the log-likelihood is a monotonically increasing function of the number of topic (as expected), for a number of topics around ~5 it shows a “knee” and starts to increase more slowly. This suggests that ~5 is the minimal number of topics needed to describe the complexity of the data without overfitting. Hence, we fix a number of topics equal to 5; however, we also verified that all our conclusions remain substantially unaffected if a different number of topics is chosen.

After running LDA with K = 5, we retrieved the model output, which consists of two probability distributions: the first is P(position| k) with $k \in [1,5]$, which is the conditional probability distribution defining the topic k; the second probability distribution is P(k | gene), namely the probability distribution that quantifies the degrees of belonging of a given gene to the topics $k \in [1,5]$. With these probability distributions, we can identify the spatial positions that form each topic and how the different topics can be combined to generate the spatial expression pattern of each gene.

Being a generative model, once trained, LDA can also decompose into topics the spatial expression patterns of genes that were not used during the training procedure. We exploited this feature of LDA to estimate the degrees of belonging of non-olfactory receptor genes. To this aim, we utilized an algorithm based on the python gensim library Lda.Model.inference function (Rehurek and Sojka, 2010), using as input the estimated probability distribution P(position | k) with $k \in [1,5]$. The model fitting was performed using the Open Computing Cluster for Advanced data Manipulation (OCCAM), the High-Performance Computer designed and managed in collaboration between the University of Torino and the Torino division of the Istituto Nazionale di Fisica Nucleare (Aldinucci et al., 2017).

Definition of *Olfr* 3D indexes via diffusion pseudo-time

As explained in the section above, we can describe the spatial expression pattern of each gene through a set of five numbers, which represent the degrees of belonging to the five topics identified by LDA. We applied a diffusion map (Haghverdi et al., 2015) to the degrees of belonging of the *Olfrs* to visualize them in two dimensions by using the “DiffusionMap” function from the “destiny” R package (Angerer et al., 2016) (with distance = “rankcor” and default parameters). In this two-dimensional map, the *Olfrs* are approximately distributed along a curve that joins the most dorsal/medial genes (those in zones 1–2) with those that are more ventral/lateral (zones 3–5). To track the position of the genes along this curve, we computed a diffusion pseudo-time (DPT) coordinate (Haghverdi et al., 2016) with the “DPT” function from the “destiny” R package (taking as starting point the gene with the smallest first diffusion component among the genes suggested by the function find_tips from the same package). In order to make the indexes go from Dorsal to Ventral, as in previous studies (Miyamichi et al., 2005), we reversed the order of the DPT coordinates by substracting the maximum coordinate from all coordinates and multiplying them by (−1). By doing this, we obtained for each *Olfr* an index, which we called 3D index, representing its spatial expression pattern in the 3D space: more dorsal/medial genes (zones 1–2) have smaller 3D indexes than *Olfrs* expressed in the ventral/lateral regions (zones 3–5).

Prediction of zone index for undetected *Olfrs* with Random Forest

We fitted a Random Forest model to the 3D indexes of 681 of the 689 *Olfrs* we characterized with our dataset (i.e., those that are located in genomic clusters). The following nine features of each *Olfr* were used as predictors: genomic position (i.e., gene starting position divided by chromosome length); genomic cluster; genomic cluster length; number of *Olfrs* in the genomic cluster; number of enhancers in the genomic cluster; cluster position (i.e., starting position of the cluster divided by the chromosome length); distance to the closest enhancer; gene position within the cluster (i.e., the distance of the gene starting position from the end of the cluster divided by the cluster length); and phylogenetic class. These features were computed using the mm10 mouse genome in Biomart (Kinsella et al., 2011), while the list of enhancers and the genomic clusters assigned to each *Olfr* were taken from Monahan et al. (2017). The Random Forest model was fitted with the function “randomForest” (R library “randomForest” (Liaw and Wiener, 2002), with option “na.action = na.omit”). Afterward, we performed a cross-validation test with the function “rf.crossValidation” from the “rfUtilities” package (Rather et al., 2020) with default parameters. Over 100 cross-validation iterations, the root mean square error (RMSE) was $\lesssim$10% of the mean 3D index. The feature importance was computed with the “importance” function from the randomForest library with default parameters. Finally, the Random Forest model trained on the 681 *Olfrs* was used to predict the 3D indexes of 697 *Olfrs* that were too lowly expressed or were undetected in our dataset. Overall, we were able to compute or predict with Random Forest a 3D index for all the *Olfrs* annotated in the mouse genome, except for 28 of them that do not have any genomic cluster assigned. To quantify the consistency between our *Olfr* 3D indexes and indexes defined previously, we calculated the Spearman’s correlation coefficients between our indexes and those defined in three previous studies (Miyamichi et al., 2005; Tan and Xie, 2018; Zapiec and Mombaerts, 2020) (see Figures 4G, S4B, and S4C).

Odorant information and Olfr-ligand pairs

All odorant structures and associated CAS numbers were retrieved from either Sigma-Aldrich (www.sigmaaldrich.com) or PubChem (https://pubchem.ncbi.nlm.nih.gov). A comprehensive catalog of the cognate mouse Olfr-ligand pairs was collected (last update: March 2021) by combining data from the ODORactor database (Liu et al., 2011) and additional literature sources (Abaffy et al., 2006; Araneda et al., 2004; Bozza et al., 2002; Dunkel et al., 2014; Floriano et al., 2000; Gaillard et al., 2002; Godfrey et al., 2004; Grosmaitre et al., 2006, 2009; Jiang et al., 2015; Jones et al., 2019; Kajiya et al., 2001; Malnic et al., 1999, 2004; Nara et al., 2011; Nguyen et al., 2007; Oka et al., 2004, 2006, 2009; Pfister et al., 2020; Repicky and Luetje, 2009; Saito et al., 2004, 2017; Saraiva et al., 2019; Shirasu et al., 2014; Shirokova et al., 2005; von der Weid et al., 2015; Yoshikawa et al., 2013; Yoshikawa and Touhara, 2009; Yu et al., 2015; Zhuang and Matsunami, 2007).

This catalog includes 738 Olfr-ligand interactions for a total of 153 Olfrs and 221 odorants. These 153 *Olfrs* include 100 spatial *Olfrs* in our dataset and for which we have 3D indexes, and 49 additional *Olfrs* with predicted 3D indexes (see above). Next, we checked whether *Olfrs* pairs sharing at least one cognate ligand have more similar spatial expression patterns than pairs not sharing ligands. To do this, we computed the absolute values of the differences between the 3D indexes (Δ) of 1706 pairs of ORs sharing at least one odorant and 9,922 pairs of ORs that are known to bind to different odorants (Figure 6B). The two sets of Δ values were significantly different (Mann-Whitney U test, p value $< 2 \times 10^{-16}$). This test remained significant when excluding Olfrs for which 3D indexes were estimated by the Random Forests model (p value $< 2 \times 10^{-16}$), and also when excluding class I Olfrs (p value $< 2 \times 10^{-16}$).

Correlation analysis of physico-chemical descriptors with 3D index

Physicochemical descriptors for ligands were obtained from the Dragon 6.0 software (http://www.talete.mi.it/). After removing the descriptors showing 0 variance, a table of 1911 descriptors for 205 ligands was obtained. In addition to these, we estimated the air/mucus partition coefficients (K_{am}) of the odorants as done previously (Rygg et al., 2017; Scott et al., 2014). Briefly, we calculated the air/water partition coefficients (K_{aw}) for each odorant from the Henry's Law constants obtained using the HENRYWIN model in the US EPA Estimation Program Interface (EPI) Suite (version 4.11; www.epa.gov/oppt/exposure/pubs/episuite.htm). Then, we computed the air/mucus partition coefficients (K_{am}) according to the formula:

$$Log(K_{am}) = 0.524 \cdot Log(K_{aw}) \cdot Log(K_{ow})$$

where K_{ow} indicates the octanol/water partition coefficient, which were obtained using the KOWWIN model in the EPI Suite.

To increase the robustness of our correlation analysis, we removed the descriptors with 20 or more identical values across our set of ligands, and we initially considered only the ligands having 2 or more known cognate receptors; these filters gave us 1,210 descriptors (including K_{am}) for 101 ligands.

We performed Spearman's correlation tests between the physicochemical descriptors and mean 3D index of the cognate *Olfrs*, and we considered as statistically significant those correlations with FDR <0.05 (see Table S6). The descriptors with the largest correlation coefficients were K_{am} (rho = 0.55, p = 1×10^{-7}) and ATSC2s (Centred Broto-Moreau Autocorrelation of lag 2 weighted by I-state, rho = -0.56, p = 2×10^{-7}). We obtained statistically significant correlations between K_{am} and the mean 3D indexes also when excluding *Olfrs* with 3D indexes predicted by Random Forest (Rho = 0.48, p value = 2×10^{-6}, based on 87 ligands; Figure S5B) or excluding class I *Olfrs* (Rho = 0.5, p value = 1×10^{-7}, based on 101 ligands; Figure S5C).

In-situ hybridization

In-situ hybridization was basically performed as previously described (Ibarra-Soria et al., 2017). Adult 12-week-old male C57BL/6J mice anesthetized, and then perfused with 4% paraformaldehyde. The snouts containing the OM were dissected out, decalcified in RNase-free 0.45M EDTA solution (in 1× PBS) for two weeks – the bone and tissue encapsulating the OM are necessary to preserve the OM tissue integrity during the ISH. Next, the decalcified snouts were cryoprotected in RNase-free 30% sucrose solution (1× PBS), dried, embedded in OCT embedding medium, and frozen at −80°C. Sequential 16 μm sections were prepared with a cryostat and the sections were hybridized to digoxigenin-labeled cRNA probes prepared from the different genes using the following oligonucleotides: *Cytl* (5′-AAAGACACTACCTCTGTTGCTGCTG-3′ and 5′-GTAAGCAGAGACCAGAAAGAAGAGTG-3′), *Moxd2* (5′-TGTACCTTTCTCCCACTCCCTATTGTC-3′ and 5′-CCCATGCAACTGGAGATGTTAATTCTG-3′), *Olfr309* (5′-TACAATGGCCTATGACCGCTATGTG-3′ and 5′-TCCTGACTGCATCTCTTTGTTCCTG-3′), *Olfr727* (5′-CGCTATGTTGCAATATGCAAGCCTC-3′ and 5′-GCTTTGACATTGCTGCTTTCACCTC-3′), and *Olfr618* (5′-CATGAACCAATGTACCTTTTCCTCTC-3′ and 5′-AAACCTGTCTTGAATTTGCTTTGTC-3′). The PCR products were cloned into pGEM-T Easy vector and the probes were obtained by *in vitro* transcription of the plasmids, using SP6 or T7 RNA Polymerases (Roche) and DIG RNA Labeling mix (Roche).

QUANTIFICATION AND STATISTICAL ANALYSIS

Information on gene expression thresholds for spatial differential expression analysis is described in the method details section. The presence of a spatial signal along the 3 axes was verified via the Moran's I statistic (see relevant section above). The presence of spatial non-random patterns was tested using the Ljung-Box test and the resulting p values were adjusted using the FDR method (see relevant section above). Consistency between different datasets and replicates, as well as association between independent data were tested using Spearman correlation tests. Mann-Whitney U tests were employed to test the statistical significance of differences between two distributions. Finally, a cross validation test was used to quantify the accuracy of our Random Forests model through the root mean square error (RMSE). Statistical tests were performed using R (version 4.1.2). Statistical details are reported in the Main text, Figures and Figure legends, the STAR methods section and supplementary tables. N represents the number of biological replicates (animals) we analyzed. Boxplots are centered at the median of the distribution, the bottom and top of the box represent the 1st and 3rd quartiles respectively, and the whiskers extend for an additional 1.5 times the interquartile range.

Cell Reports, Volume *38*

Supplemental information

A 3D transcriptomics atlas of the mouse nose sheds light on the anatomical logic of smell

Mayra L. Ruiz Tejada Segura, Eman Abou Moussa, Elisa Garabello, Thiago S. Nakahara, Melanie Makhlouf, Lisa S. Mathew, Li Wang, Filippo Valle, Susie S.Y. Huang, Joel D. Mainland, Michele Caselle, Matteo Osella, Stephan Lorenz, Johannes Reisert, Darren W. Logan, Bettina Malnic, Antonio Scialdone, and Luis R. Saraiva

SUPPLEMENTAL FIGURES

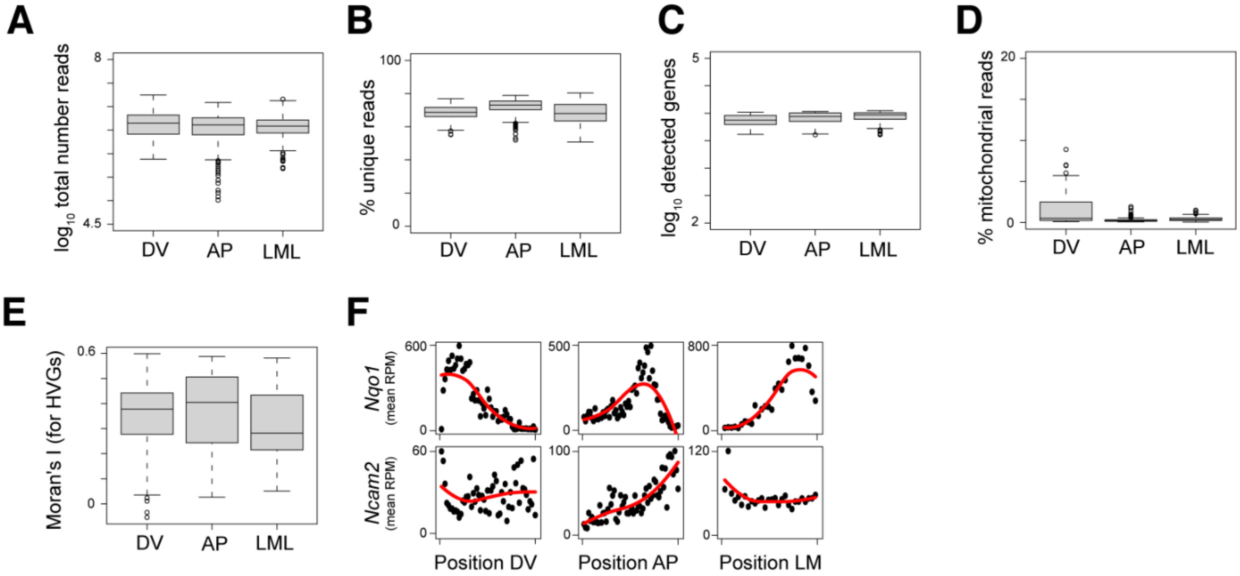


Figure S1. TOMO-seq data QC, Related to Figure 1 and Table S1. (A) Boxplots showing the distributions of the $\log_{10}$ total number of reads per sample in each axis (DV = dorsal-ventral; AP = anterior – posterior; LML = lateral-mid-lateral). (B) Boxplots of percentage of uniquely mapped reads per sample per axis. (C) Boxplots of distributions of $\log_{10}$ detected genes per sample per axis. (D) Boxplots of percentage of mitochondrial reads per sample per axis. (E) Boxplots showing the distribution of the Moran's I statistics calculated for the top 100 Highly Variable Genes per axis. P-values are computed for each gene and then combined with the Simes' method. The combined p-values are $< 2.2\text{x}10^{-16}$ for all axes. (F) Normalized expression of canonical OM spatial marker genes along the three axes. Red line showing fits with local polynomial models.

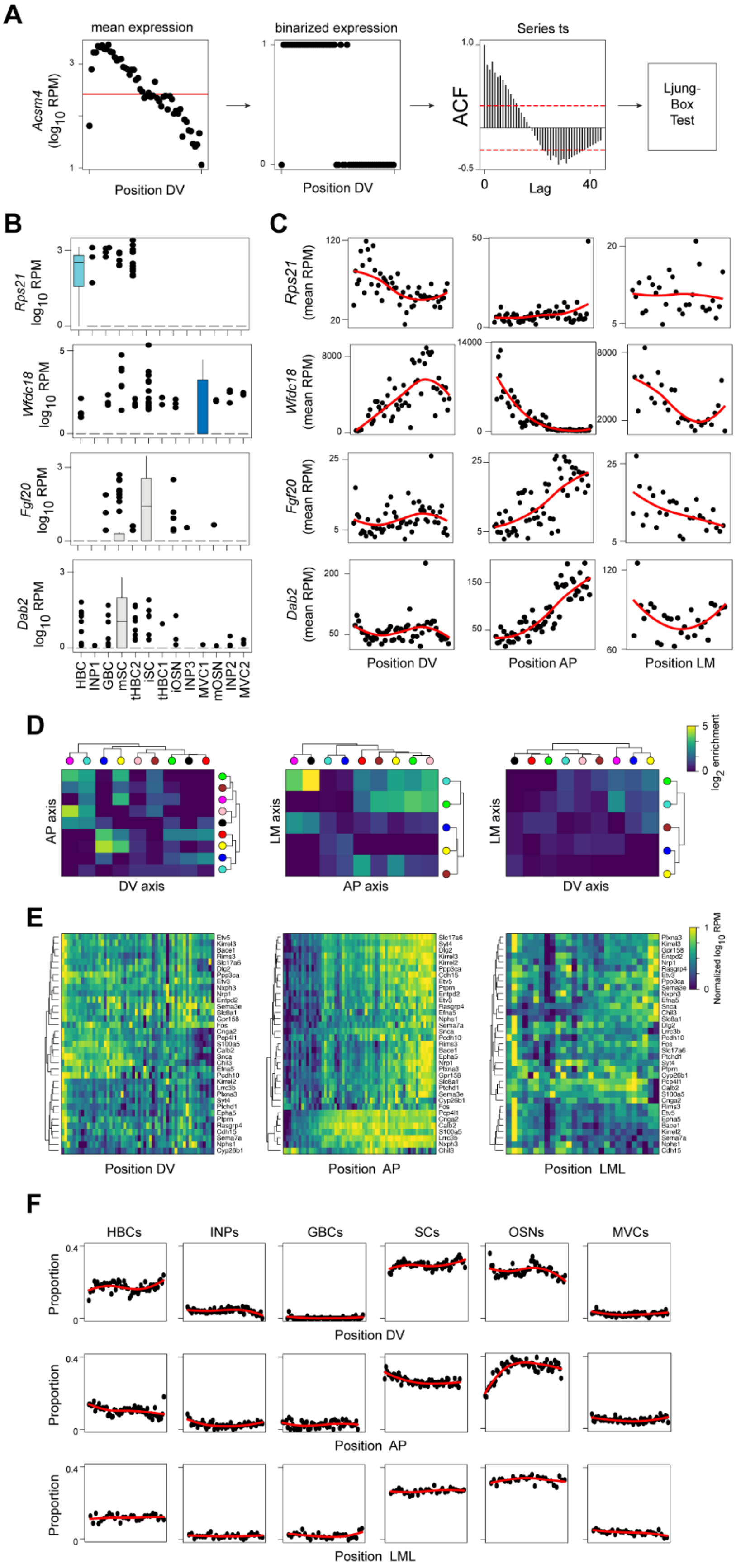

A
mean expression
Acsm4 (log10 RPM)
Position DV
binarized expression
Position DV
Series ts
ACF
Lag
Ljung-Box Test
B
Rps21 log10 RPM
Wfdc18 log10 RPM
Fgf20 log10 RPM
Dab2 log10 RPM
HBC
INP1
GBC
mSC
tHBC2
iSC
tHBC1
iOSN
INP3
MVC1
mOSN
INP2
MVC2
C
Rps21 (mean RPM)
Wfdc18 (mean RPM)
Fgf20 (mean RPM)
Dab2 (mean RPM)
Position DV
Position AP
Position LM
D
AP axis
DV axis
LM axis
AP axis
LM axis
DV axis
log2 enrichment
E
Etv5
Kirrel3
Bace1
Rims3
Slc17a6
Dlg2
Ppp3ca
Etv3
Nxph3
Nrp1
Entpd2
Sema3e
Slc8a1
Gpr158
Fos
Cnga2
Pcp4l1
S100a5
Calb2
Snca
Chil3
Efna5
Pcdh10
Kirrel2
Lrrc3b
Plxna3
Syt4
Ptchd1
Epha5
Ptprn
Rasgrp4
Cdh15
Sema7a
Nphs1
Cyp26b1
Position DV
Slc17a6
Syt4
Dlg2
Kirrel3
Kirrel2
Ppp3ca
Cdh15
Etv5
Ptprn
Entpd2
Etv3
Rasgrp4
Efna5
Nphs1
Sema7a
Snca
Pcdh10
Rims3
Bace1
Epha5
Nrp1
Plxna3
Gpr158
Slc8a1
Ptchd1
Sema3e
Cyp26b1
Fos
Pcp4l1
Cnga2
Calb2
S100a5
Lrrc3b
Nxph3
Chil3
Position AP
Plxna3
Kirrel3
Gpr158
Entpd2
Nrp1
Rasgrp4
Etv3
Ppp3ca
Sema3e
Nxph3
Efna5
Snca
Chil3
Slc8a1
Dlg2
Lrrc3b
Pcdh10
Fos
Slc17a6
Ptchd1
Syt4
Ptprn
Cyp26b1
Pcp4l1
Calb2
S100a5
Cnga2
Rims3
Etv5
Epha5
Bace1
Kirrel2
Sema7a
Nphs1
Cdh15
Position LML
Normalized log10 RPM
F
HBCs
INPs
GBCs
SCs
OSNs
MVCs
Proportion
Position DV
Position AP
Position LML

Figure S2. Spatial differential expression analysis, **Related to Figure 2 and Table S2.** (A) Schematics of strategy to find spatially differentially expressed genes; as an example, data for *Acsm4* along the dorsal-ventral (DV) axis is shown: Gene expression was binarized according to whether the expression per slice was higher or lower than the median expression (red horizontal line). Then, we computed the autocorrelation function for different values of the lags, and we applied the Ljung-Box test to verify whether the autocorrelation values are significantly higher than zero. (B) Box plots of example genes' expression ($\log_{10}$ reads-per-million, RPMs) distributions in different cell types. None of these genes is expressed in mOSNs (INP = Immediate Neuronal Precursors; GBC = Globose Basal Cells; mOSNs = mature Olfactory sensory neurons; iOSNs = immature Olfactory Sensory Neurons; MVC = Microvillous Cells; iSC = Immature Sustentacular Cells; mSC = Mature Sustentacular Cells; HBCs = Horizontal Basal Cells). (C) Spatial gene expression trends along each axis of the example genes shown in panel B. (D) Heatmap showing the $\log_2$ enrichment for the intersection between different gene clusters (indicated by colored circles) across pairs of axes, after excluding *Olfr* genes. (E) Heatmaps showing normalized mean expression of the neuronal activity marker genes listed in Table S2 from (Wang et al., 2017) along the three axes. (F) We used cell type deconvolution analysis to estimate the cell type composition per section along the three axes. The red line marks the fit with local polynomial models.

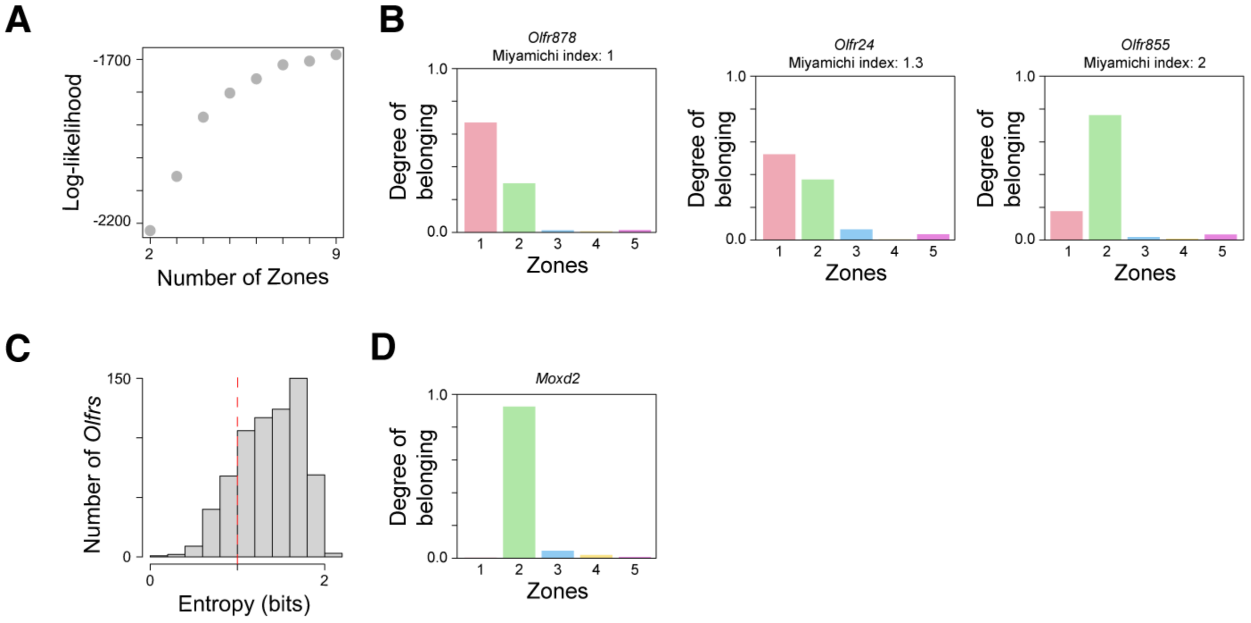


Figure S3. *Olfr* genes 3D zones, Related to Figure 3. (A) Log-likelihood values for fits with LDA models as a function of the number of zones. (B) Bar plot showing the degrees of belonging of *Olfr* genes with overlapping spatial patterns (Miyamichi indexes of 1, 1.3 and 2 respectively). (C) Distribution of entropy values of our 689 spatially differentially expressed *Olfrs*. The *Olfrs* with entropy values less than 1 bit (vertical red line) can be considered to fit mostly in one zone. (D) Bar plot showing the degrees of belonging of *Moxd2*.

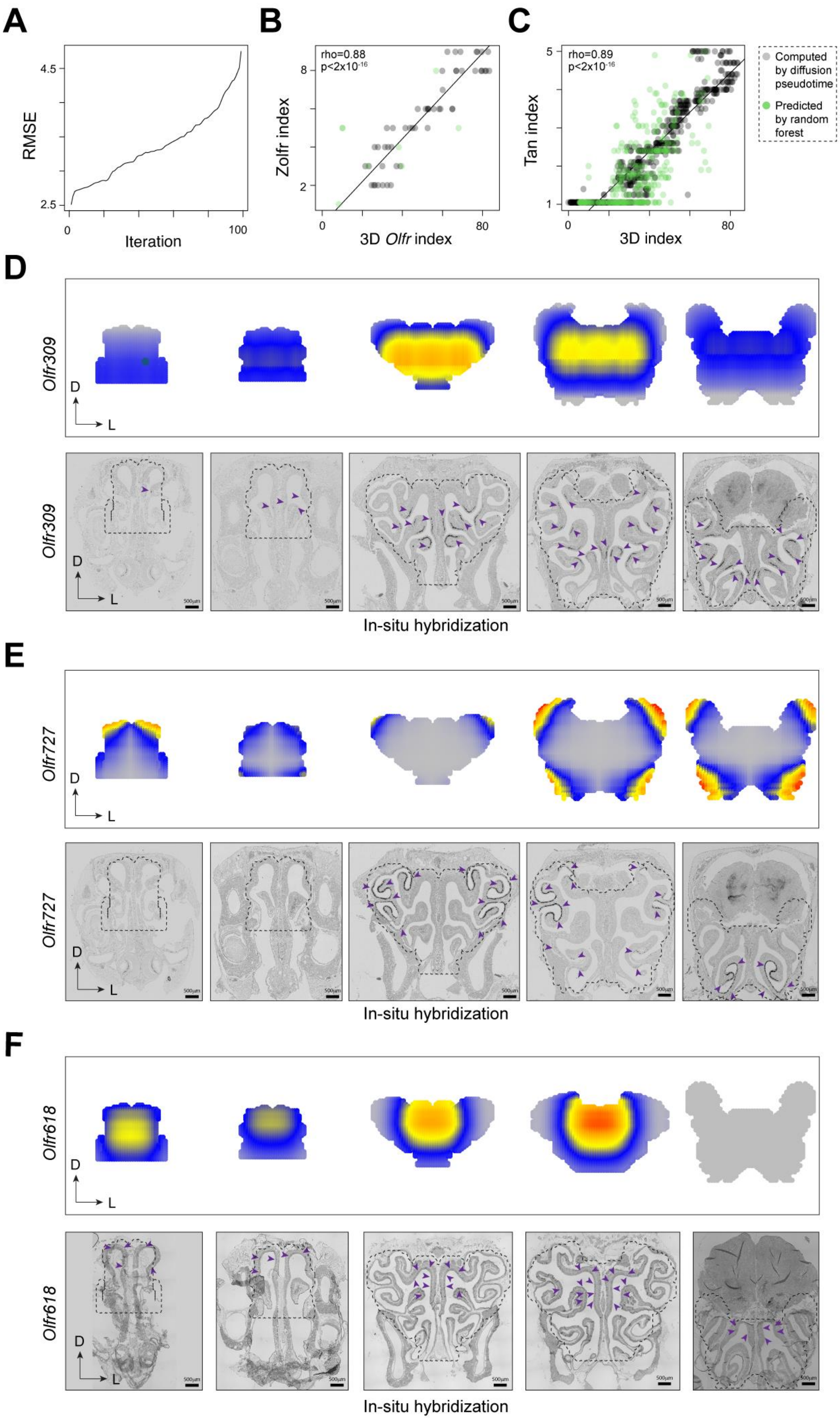

A
RMSE
4.5
2.5
Iteration
0
100
B
rho=0.88
p<2x10-16
Zolfr index
8
2
3D Olfr index
0
80
C
rho=0.89
p<2x10-16
Tan index
5
1
3D index
0
80
Computed by diffusion pseudotime
Predicted by random forest
D
Olfr309
D
L
Olfr309
500μm
In-situ hybridization
E
Olfr727
D
L
Olfr727
500μm
In-situ hybridization
F
Olfr618
D
L
Olfr618
500μm
In-situ hybridization

Figure S4. *Olfr* 3D index prediction, Related to Figure 4 and Tables S3 and S4. (A) Root mean square error (RMSE) per iteration of the cross-validation test for the Random Forest model used to predict 3D indexes. (B) Scatter plot illustrating the comparison of our 3D indexes versus the "Zolfr indexes" defined by (Zapiec and Mombaerts, 2020) from ISH data. For this comparison, these zones were numbered from 1 to 9 from the most dorsal to the most ventral. Black circles indicate *Olfrs* detected in our dataset; green circles are *Olfrs* for which indexes were predicted with Random Forest. The correlation coefficients computed separately on these two sets of *Olfrs* are respectively rho=0.92, p-value$<2x10^{-16}$ and rho=0.44, p-value>0.05. (C) Scatter plot showing the correlation of our 3D indexes with the "Tan Indexes" estimated by (Tan and Xie, 2018), who performed RNA-seq on 12 samples at different positions along the dorsal-ventral axis of the OM and estimated indexes using as reference the ~80 *Olfrs* analyzed in (Miyamichi et al., 2005) via ISH. Black circles indicate *Olfrs* detected in our dataset; green circles are *Olfrs* for which indexes were predicted with Random Forest. The correlation coefficients computed separately on these two sets of *Olfrs* are respectively rho=0.95, p-value$<2x10^{-16}$, and rho=0.68, p-value $< 2x10^{-16}$.(D-F) In-situ hybridization experiment validating the predicted 3D spatial expression patterns for *Olfr309* (D), *Olfr727* (E), and *Olfr618* (F). Note that *Olfr618* is expressed in Zone 1, consistent with its predicted spatial expression pattern and calculated 3D index of 7.42 (Figure 4 N, O). Purple arrowheads indicate the location of ISH labeled cells. The dotted outline indicates the borders of the OM dissected and used in the RNA-seq experiments and for the construction of the 3D model.

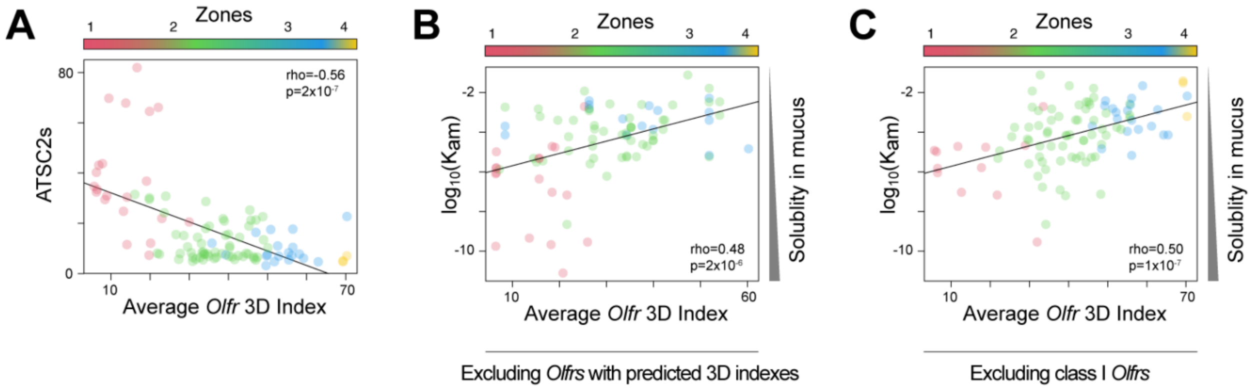


Figure S5. Physiological role of the zones, Related to Figure 6 and Table S6. Scatter plot illustrating the correlation between ATSC2s of the odorants and the average 3D indexes of their cognate Olfrs. Only odorants for which we know at least two cognate Olfrs (110) were used here. Odorants are colored according to the zone they belong to (defined as the zone with the highest average degree of belonging computed over all cognate receptors). (B) Scatter plot illustrating the correlation between air/mucus partition coefficients of the odorants and the average 3D indexes of their cognate Olfrs. Only odorants which are detected by Olfrs present in our TOMO-seq dataset (87) were used here. (C) Scatter plot illustrating the correlation between air/mucus partition coefficients of the odorants and the average 3D indexes of their cognate Olfrs. Only odorants which are detected by Class II Olfrs (101) were used here.

Chapter II. Retinoic acid signaling is critical during the totipotency window in early mammalian development: Insights from single cell transcriptomic profiling

nature structural & molecular biology

RESOURCE

https://doi.org/10.1038/s41594-021-00590-w

Check for updates

OPEN

Retinoic acid signaling is critical during the totipotency window in early mammalian development

Ane Iturbide[1], Mayra L. Ruiz Tejada Segura[1,2,3,7], Camille Noll[1,7], Kenji Schorpp[4,7], Ina Rothenaigner[4], Elias R. Ruiz-Morales[1,5], Gabriele Lubatti[1,2,3], Ahmed Agami[1], Kamyar Hadian[4], Antonio Scialdone[1,2,3] and Maria-Elena Torres-Padilla[1,6] ✉

Totipotent cells hold enormous potential for regenerative medicine. Thus, the development of cellular models recapitulating totipotent-like features is of paramount importance. Cells resembling the totipotent cells of early embryos arise spontaneously in mouse embryonic stem (ES) cell cultures. Such '2-cell-like-cells' (2CLCs) recapitulate 2-cell-stage features and display expanded cell potential. Here, we used 2CLCs to perform a small-molecule screen to identify new pathways regulating the 2-cell-stage program. We identified retinoids as robust inducers of 2CLCs and the retinoic acid (RA)-signaling pathway as a key component of the regulatory circuitry of totipotent cells in embryos. Using single-cell RNA-seq, we reveal the transcriptional dynamics of 2CLC reprogramming and show that ES cells undergo distinct cellular trajectories in response to RA. Importantly, endogenous RA activity in early embryos is essential for zygotic genome activation and developmental progression. Overall, our data shed light on the gene regulatory networks controlling cellular plasticity and the totipotency program.

Totipotency is the ability of a cell to give rise to a full organism[1,2] and encompasses the broadest cellular plasticity in the mammalian body. Totipotency is a transient feature of the cells in the early embryo, which in mice is limited to the zygote and 2-cell embryo, because only the blastomeres of these stages can autonomously generate a full organism[3–5]. As development progresses, totipotency is lost and cellular plasticity is gradually reduced. Three days after fertilization, the blastocyst forms and pluripotent cells emerge within the inner cell mass (ICM)[2]. In contrast to totipotent cells, pluripotent cells can no longer contribute to the extra-embryonic derivatives of the trophectoderm[6].

Pluripotent embryonic stem (ES) cells derive from the ICM. The establishment of ES cell lines over 30 years ago[7] has enabled their use as model system to study pluripotency. Depending on the culture conditions, ES cell cultures can be highly heterogeneous, in which distinct cell populations with diverse developmental potentials coexist. Among these, cells resembling the blastomeres of 2-cell stage embryos, referred to as '2-cell-like-cells' (2CLCs), arise spontaneously, constituting less than 1% of the cells[8]. 2CLCs share several features with 2-cell stage embryos, including a '2 C' transcriptional program, characterized by genes expressed upon zygotic genome activation (ZGA), which occurs in late 2-cell embryos[8–10]. This includes the transcription factor ZSCAN4[11] and retrotransposons from the MERVL family[12]. In addition, 2CLCs recapitulate other features of 2-cell embryos including their chromatin accessibility landscape[9], greater global histone mobility[13] and the capacity to contribute to extra-embryonic tissues[8].

Although not strictly totipotent, 2CLCs are considered totipotent-like cells and are therefore a powerful cellular model to study molecular features related to totipotency. 2CLCs emerge most often from naive ES cells, but downregulate protein levels of pluripotency factors[10]. Upon exit from pluripotency, 2CLCs arise from an intermediate cellular population characterized by the expression of ZSCAN4. The number of ZSCAN4$^+$ cells fluctuates in cell cultures, and can increase following changes in metabolites in the medium or the addition of signaling molecules such as retinoic acid (RA)[14,15]. Much effort has been made towards understanding the mechanisms regulating the transcriptional program in 2CLCs and in 2-cell stage embryos[8–10,16–21]. However, it is still unclear how 2CLCs arise, and the factors that activate the 2-cell program and regulate ZGA in vivo remain elusive. Thus, identifying conditions that can robustly induce and stably maintain 2CLCs in culture can shed light into their regulatory networks and potentially uncover key factors activating the earliest developmental program in mammals.

Results

Low concentrations of RA induce 2CLCs. To identify the molecular pathways underlying 2CLC identity, we performed a large-scale, small-molecule screen using an ES cell line with a stable integration of the '*2C::tbGFP*' reporter, driving turbo GFP expression under MERVL long-terminal repeat (LTR; Supplementary Fig. 1a), used to identify 2CLCs[8–10,16,17]. We set up a pilot screen with 1,280 FDA-approved compounds using the percentage of tbGFP-expressing cells as primary readout. As a positive control for 2CLC induction we used acetate[14]. Our pilot set-up performed robustly across experiments (Supplementary Fig. 1b–d). We then screened 30,000 compounds from a diversity library and obtained 393 hits (Supplementary Fig. 1b), which we further assayed in

---

[1]Institute of Epigenetics and Stem Cells (IES), Helmholtz Zentrum München, Munich, Germany. [2]Institute of Functional Epigenetics (IFE), Helmholtz Zentrum München, Neuherberg, Germany. [3]Institute of Computational Biology (ICB), Helmholtz Zentrum München, Neuherberg, Germany. [4]Assay Development & Screening Platform, Institute of Molecular Toxicology & Pharmacology (TOXI), Helmholtz Zentrum München, Neuherberg, Germany. [5]Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge, UK. [6]Faculty of Biology, Ludwig-Maximilians Universität, Munich, Germany. [7]These authors contributed equally: Mayra L. Ruiz Tejada Segura, Camille Noll, Kenji Schorpp. ✉e-mail: torres-padilla@helmholtz-muenchen.de

triplicates and under two concentrations, incorporating ZSCAN4 expression as additional readout. This resulted in 16 confirmed hits, which we tested in a tertiary screen using a concentration gradient and a viability test. In general, higher concentrations of these 16 hits led to reduced cell numbers (Supplementary Fig. 1e), suggesting dose-dependent toxicity. The tertiary screen identified three retinoids as major hits for their ability to increase the number of 2CLCs: RA, isotretinoin and acitretin (Supplementary Fig. 2a,b). Because RA is the only natural retinoid among them, we focused primarily on RA for further studies. We validated the screening using fluorescence-activated cell sorting (FACS), which confirmed that RA induces 2CLCs, with an effect size of ~10-fold (Supplementary Fig. 2c).

Next, we characterized the conditions that allow robust reprogramming to 2CLCs by RA. We also aimed to reduce the DMSO concentration because DMSO hampers 2CLC emergence (Supplementary Fig. 2c). Because, in our screen, we observed 2CLC induction at the lowest RA doses, we probed these RA concentrations with reduced DMSO concentrations and different treatment lengths (Fig. 1a). Remarkably, we identified conditions under which RA induced a more than 50-fold increase of 2CLCs (up to 30% of the culture; Fig. 1b). Although we observed an increase in 2CLC induction with higher RA concentration and length of treatment, just 30 min of RA treatment at the lowest concentration (0.16 μM) robustly increased (approximately fourfold) 2CLCs (Fig. 1b). We obtained similar results, albeit with slightly lower induction rates, for the other retinoid, acitretin (Supplementary Fig. 3a).

RA has been used for decades to induce ES cell differentiation[22], which appears at odds with its ability to induce 2CLCs. However, RA induces differentiation at higher doses (1–10 μM) than those we report here to induce 2CLCs, and when added for longer time periods. Indeed, increasing the RA concentration (up to 10 μM) did not lead to a higher proportion of 2CLCs (Fig. 1c). Instead, we observed maximal 2CLC induction at 0.53 μM RA, and higher concentrations gradually decreased this effect (Fig. 1c). Thus, RA mediates 2CLC reprogramming most efficiently at lower concentrations. 2CLCs induced with RA express 2CLC markers such as ZSCAN4 (Fig. 1d). The simultaneous addition of RA or acitretin with acetate—also known to induce 2CLCs[14]—resulted in a synergistic effect, leading to a conversion of more than 40% of the ES population into 2CLCs (Fig. 1e and Supplementary Fig. 3b). We next addressed whether RA plays a role in the transition from ZSCAN4$^+$ cells to 2CLCs. We used a double ‘2C’ and *Zscan4* reporter cell line[10], sorted *Zscan4*$^+$/*2C::tbGFP*$^-$ cells, and treated them with RA. RA treatment increased the number of 2CLCs arising from ZSCAN4$^+$ cells (Fig. 1f), and induction of 2CLCs from ZSCAN4$^+$ cells was blocked by an antagonist of RA signaling (Fig. 1f). These data indicate that RA promotes the transition to the 2CLC state from the intermediary ZSCAN4$^+$ cell population. Thus, we conclude that low doses of RA robustly induce 2CLC reprogramming.

The RA pathway is active in spontaneously emerging 2CLCs. We next explored whether RA signaling is responsible for the spontaneous emergence of 2CLCs. Analysis of 2CLC RNA-seq datasets[16] revealed an increase in the expression of some of the genes encoding proteins mediating the conversion of retinol to RA, such as RDH10 and ALDH1A2 and ALDH1A3[23]. The nuclear receptors RAR (retinoic acid receptor) and RXR (retinoid X receptor) also showed increased expression in 2CLCs (Fig. 2a). This suggests that the RA pathway might be active in 2CLCs, and possibly also in totipotent cells in vivo.

To investigate the mechanism whereby RA induces 2CLCs, we disrupted the RA signaling and degradation pathways. First, we disrupted cellular RA metabolism by perturbing RA degradation through the downregulation of CRABP1, which mediates RA clearance (Fig. 2b)[24]. siRNA for *Crabp1* increased 2CLC induction in response to RA (Fig. 2c and Supplementary Fig. 4a) and led to a strong upregulation of *Zscan4* and endogenous *Mervl* transcripts (Fig. 2d). Importantly, *Crabp1* downregulation also increased the 2CLC population in control conditions (Fig. 2c), indicating that the RA pathway might be involved in triggering spontaneous reprogramming of 2CLCs. Second, we addressed whether 2CLC induction relies on nuclear RA function. We performed siRNA against the RA importers CRABP2 and FABP5, which bind RA and translocate into the nucleus to facilitate RA binding to RAR or PPAR, respectively, enabling transcriptional activation of RA-response genes[24] (Fig. 2b). Downregulation of *Crabp2* or *Fabp5* did not prevent 2CLC induction and resulted instead in a small, reproducible increase in RA-mediated 2CLC reprogramming (Fig. 2e). We observed similar results, albeit not significant, without RA addition (Fig. 2e). The slight increase in 2CLC was accompanied by an increase in *Zscan4* and *Mervl* expression (Fig. 2f). Because altering the levels of the nuclear RA importers affects 2CLC number, these results suggest that the RA pool in the nucleus plays a role in 2CLC induction.

The transcription factor RARγ mediates 2CLC reprogramming. We next addressed whether 2CLCs depend on downstream transcriptional activity of RA. Following RA import into the nucleus, RA binds to RARs and RXRs[25]. In the canonical pathway, these receptors form heterodimers upon ligand binding and activate transcription of targets containing retinoic acid response elements (RAREs). RXRs can also form non-canonical heterodimers with other nuclear receptors[26]. Thus, we tested whether specific transcription factors are necessary for RA-induced 2CLC reprogramming. We first asked whether 2CLC induction by RA and acitretin is affected by a general RAR antagonist, AGN193109[27,28]. AGN193109 clearly blocked 2CLC induction by RA and acitretin (Fig. 2g,h), indicating that 2CLC reprogramming upon retinoid stimulation depends on RAR activity. Interestingly, AGN193109 also reduced the effect of acetate on 2CLCs (Fig. 2g,h), suggesting that 2CLC induction by acetate is mediated partly through RAR activity. Importantly, addition of AGN193109 led to a significant reduction of the endogenous 2CLCs in control conditions, leading to a practically undetectable 2CLC population (Fig. 2g,h). Consistently, AGN193109 abolished the effect of *Crabp1*, *Crabp2* and *Fabp5* siRNA on 2CLC induction in control conditions and upon RA stimulation (Supplementary Fig. 4b). These results indicate that RAR activity mediates endogenous and RA-induced 2CLC reprogramming, pointing towards a key role for the RA pathway and its receptors in the core 2CLC network.

We next investigated whether RA activity signals through RAR homodimers or RAR/RXR heterodimers by treating ES cells with RXR antagonists in combination with RA. In contrast to the RAR antagonist (AGN193109), neither of the RXR antagonists tested affected 2CLC induction (Fig. 2i), suggesting that a non-canonical RAR dimer mediates RA activity during 2CLC induction. Because AGN193109 inhibits all RAR subtypes (α, β and γ), we next determined which RAR subtype is necessary for 2CLC induction. Inhibiting RARα and RARβ decreased RA-mediated 2CLC induction slightly, but did not abolish it (Fig. 2j). However, blocking RARγ with LY2955303 had the strongest effect in inhibiting 2CLC emergence, with an almost complete disappearance of detectable 2CLCs in control conditions, and a dramatic reduction upon RA stimulation (Fig. 2j,k and Supplementary Fig. 4c). Accordingly, RARγ participates in 2CLC induction by RA and in the spontaneous emergence of 2CLCs.

To test whether RA can activate transcription in 2CLCs, we used a RARE reporter, whereby a minimal promoter (cytomegalovirus, CMV) and an upstream RARE[29] drive GFP expression (Fig. 2l), which we transfected into a *2C::tdTomato* ES cell line[16]. RARE reporter activity increased upon RA addition compared to the control plasmid containing the minimal promoter alone. In addition, the 2CLC population (tdTOMATO$^+$) contains GFP$^+$ cells (~25% of the cells;

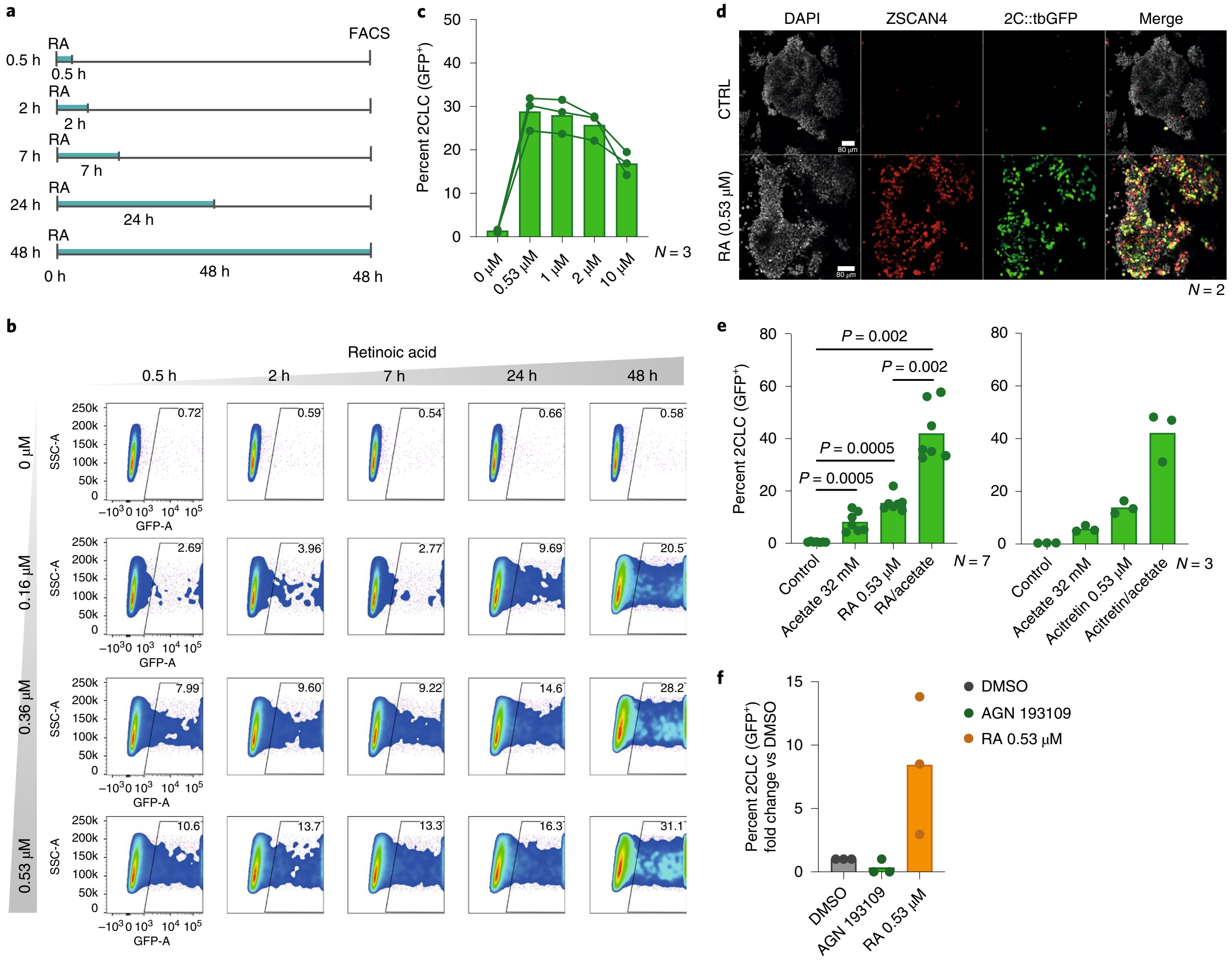


Fig. 1 | Low concentrations of RA robustly induce 2CLCs. a, Experimental design. Embryonic stem (ES) cells were treated with a range of RA concentrations for different time periods. 2CLC induction was measured by FACS, 48 h after treatment. **b**, Representative scatter plot for the experiment in **a**, showing *2C::tbGFP* fluorescence measurements of individual cells as assayed by FACS. **c**, Effect of high RA concentrations on 2CLCs induction. The percentage of 2CLCs (GFP$^+$) quantified by FACS 48 h after treatment is shown (bars show the mean of the indicated number of replicates). Each line and connecting dots correspond to measurements of one replicate. **d**, Immunofluorescence using antibodies for the indicated proteins. The merge images show 4′,6-diamidino-2-phenylindole (DAPI; gray), ZSCAN4 (red) and tbGFP (green) expression. Scale bars, 80 μm. **e**, Effect of treatment with retinoids in combination with acetate on 2CLC induction. The percentage of 2CLCs (GFP$^+$) was quantified by FACS, 48 h after treatment. The mean of the indicated replicates (represented by individual dots) is shown. *P* values were calculated by two-sided Mann–Whitney test. **f**, Induction of 2CLCs from ZSCAN4$^+$ cells upon RA treatment. The percentage of 2CLCs (GFP$^+$/mCherry$^+$) was quantified by FACS, 24 h after sorting ZSCAN4$^+$ (GFP$^-$/mCherry$^+$) cells.

Fig. 2l). Altogether, this indicates that endogenous 2CLCs possess RARE activity and that the fraction of 2CLCs showing this activity increases upon RA stimulation. To investigate this further, we asked whether genes expressed in 2CLCs contain RARE motifs by examining 2CLC-regulatory regions from assay for transposase-accessible chromatin sequencing (ATAC-seq) datasets[30]. The RARE motif was significantly enriched in 2CLCs compared to a random distribution, which appeared both in the 'gained' and 'lost' peaks compared to ES cells (Fig. 2m). The RARE motif in 2CLC-specific peaks was also significantly enriched compared to ATAC-seq peaks shared between 2CLCs and ES cells ($P = 1.14 \times 10^{-95}$). We obtained similar results in ES cell-specific peaks ($P = 1.05 \times 10^{-132}$). Thus, enrichment of the RARE motif in accessible regions in 2CLCs correlates with the RARE activity observed in 2CLCs and suggests that RA activity functions through the binding of RARE elements in ES cells to induce 2CLC reprogramming.

RA induces 2CLC reprogramming without inducing differentiation. 2CLCs arise preferentially from naive ES cells[10]. Because RA promotes ES cell differentiation[22], we next addressed whether the ability of RA to reprogram 2CLCs depends on culture conditions. We tested conditions that promote (1) naive, ground-state pluripotency (+LIF (leukemia inhibitory factor) and +2i), (2) primed pluripotency (+LIF without 2i) or (3) exit of pluripotency towards differentiation (withdrawal of LIF and 2i). We treated ES cells with RA for one to five days and quantified 2CLCs (Fig. 3a). For the three conditions analyzed, 2CLC induction was highest 48 or 72 h following RA addition, beyond which timepoint the 2CLC population gradually decreased (Fig. 3a). Although the addition of 2i decreased the number of RA-induced 2CLCs, LIF removal also led to a decrease in the percentage of 2CLCs (Fig. 3a). Of the three conditions, the highest reprogramming efficiency by RA was observed when LIF was maintained, but 2i was removed (Fig. 3a). These data

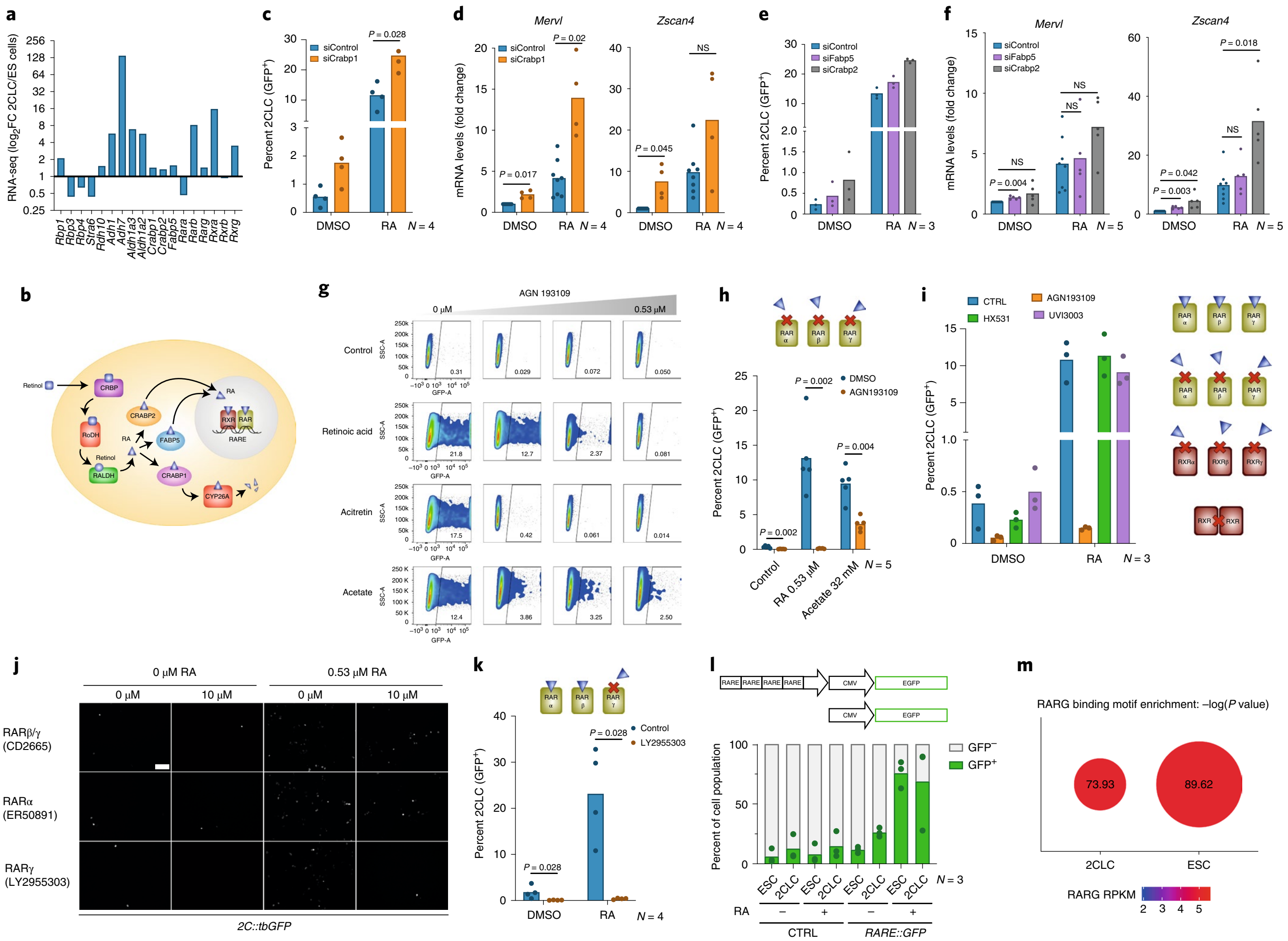


Fig. 2 | RARγ is required for 2CLC emergence. a, Expression levels ($\log_2$FC) (FC, fold change) of selected RA-pathway-related genes in 2CLCs and ES cells (ESCs) based on RNA-seq data ($N=2$, from ref. [16]). **b**, Schematic of the RA pathway. **c**, Induction of 2CLCs upon siRNA for *Crabp1* and RA treatment. The percentage of 2CLCs was quantified by FACS. The mean ± s.d. of the indicated number of replicates is shown. *P* values were calculated by two-sided Mann–Whitney test. **d**, Quantitative polymerase chain reaction (qPCR) analysis upon transfection of siRNA for *Crabp1* and RA treatment. Mean ± s.d. values of the indicated number of replicates are shown. *P* values were calculated by two-sided Student's *t*-test. NS, not significant. **e**, Induction of 2CLCs upon transfection of siRNA for *Fabp5* and *Crabp2* and RA treatment. The percentage of 2CLCs was quantified by FACS. The mean ± s.d. of the indicated number of replicates is shown. **f**, qPCR analysis after transfection of siRNA for *Fabp5* and *Crabp2* and RA treatment. Mean ± s.d. values of the indicated number of replicates are shown. *P* values were calculated by two-sided Student's *t*-test. **g**, Representative scatter plots from data in 3 h showing *2C::tbGFP* fluorescence measurements of individual cells as assayed by FACS. **h**, Induction of 2CLCs upon treatment with AGN193109. The percentage of 2CLCs was quantified by FACS, 48 h after treatment. Mean values of the indicated replicates are shown. *P* values were calculated by two-sided Mann–Whitney test. **i**, Induction of 2CLCs upon treatment with RAR and RXR antagonists. The percentage of 2CLCs was quantified by FACS, 48 h after treatment. Mean ± s.d. values of the indicated replicates are shown. **j**, Representative fluorescence images of ES cell colonies harboring the *2C::tbGFP* reporter, 48 h after treatment with the indicated antagonists and RA. Scale bar, 100 μm. **k**, Induction of 2CLCs upon treatment with LY2955303. The percentage of 2CLCs was quantified by FACS, 48 h after treatment. The mean of the indicated replicates is shown. *P* values were calculated by two-sided Mann–Whitney test. **l**, Percentage of 2CLCs displaying RARE activity. The percentage of 2CLCs (tdTOMATO+) and ES cells (tdTOMATO−) with RARE activity (GFP+) was quantified by FACS, 48 h after *RARE::EGFP* reporter transfection and 24 h after RA treatment. The mean of the indicated replicates is shown. **m**, RARγ binding motif enrichment in open chromatin regions, using 2CLC and ES cell specific peaks. Dot size: $-\log_{10}(P\text{ value})$.

suggest that a constant pool of pluripotent cells is required for 2CLC reprogramming upon RA addition and that, upon longer treatment, ES cells start to differentiate and are no longer able to transition towards the 2CLC state. Next, we determined the time it takes for ES cells to reprogram into 2CLCs in response to RA by adding RA to the medium for only 2 h and analyzing the percentage of 2CLCs at several timepoints thereafter (Fig. 3b). We first detected 2CLC induction 18 h after treatment and maximal induction 48 h after RA removal, suggesting that short exposure to RA induces reprogramming a few hours after the pulse. Overall, a short RA treatment is sufficient to robustly induce 2CLCs and RA may be important early during the reprogramming process.

The above results indicate that low RA concentrations robustly induce 2CLC reprogramming under a defined temporal window. To better understand how RA induces 2CLCs, we performed single cell (sc) RNA-seq at 0, 2, 12 and 48 h of RA treatment (Fig. 3c). We also analyzed cells cultured under identical RA conditions, but in the absence of LIF, as a reference for cells undergoing differentiation[31] (Fig. 3c). We sequenced 14,742 cells across timepoints, of which 11,432 passed stringent quality criteria (Supplementary

Fig. 5a,b). Clustering all data points cultured with RA and LIF revealed six clusters, visualized using uniform manifold approximation and projection (UMAP; Fig. 3d). These clusters (A–F) corresponded roughly to (A) cells with high expression levels of pluripotency factors (*Rex1/Zfp42*, *Sox2*, *Nanog*); (B) cells with a more intermediate expression level of pluripotency factors, presumably exiting pluripotency; (C) a cluster of 'RA-responsive' cells exclusively present in the 48 h RA treatment, which express low levels of 2CLC markers such as *Zscan4a*,*c*,*d*,*e* and *Gm47924*; (D) and (E) cells expressing 2CLC markers, such as *Zscan4a*,*c*,*d*,*e*, *Gm47924* and *Tcstv1*; (F) cells expressing early differentiation markers (*Gata6*, *Sox17*, *Sox7*) (Fig. 3e–h and Supplementary Fig. 5c). The transcriptional differences between the clusters extended beyond the known 2CLC and pluripotency markers (Supplementary Fig. 5c and Supplementary Table 1).

We analyzed each timepoint individually based on the six clusters identified, which comprise all cellular heterogeneity across timepoints. To assess whether any cluster represents the 2CLC population, we plotted *2C::tbGFP* and *Zscan4* expression over the UMAP (Fig. 3g). Both *tbGFP* and *Zscan4* were expressed highest in clusters D and E in all timepoints, indicating that unbiased clustering identifies 2CLCs based on transcriptional data (Fig. 3e). In agreement with our observations above, the number of 2CLCs (GFP$^+$ cells) was maximal in the 48 h RA-treated timepoint, reaching up to 60% of the population (Fig. 3g,h and Supplementary Fig. 5d). Accordingly, *Zscan4*$^+$ cells represented almost 80% of the cells captured at this timepoint (Supplementary Fig. 5e).

Differential gene expression (DE) analysis between clusters revealed the '2C' signature in clusters D and E (Fig. 3h, Supplementary Fig. 5c and Supplementary Tables 2–7), which contained genes expressed in 2-cell embryos, including *Zscan4*, *Tcstv1 and Gm20767*. The gene signature specific to cluster D overlapped significantly with that of cluster E (Fig. 3f; Fisher's exact test $P<2.2\times10^{-16}$). This indicates that endogenous 2CLCs (cluster E, already detected in early timepoints), overall, share the transcriptional profile of RA-induced 2CLCs (cluster D, upon induction at 48 h), including expression of *Dux* (Supplementary Fig. 5f). We also identified new 2CLC markers (Supplementary Tables 2–7), such as *Tmem72*, a transmembrane protein of unknown function (Supplementary Fig. 6a,b). The RA-responsive cluster (cluster C) emerging at 48 h displayed a partial '2C' signature too (Supplementary Fig. 6c). This includes expression of *2C::tbGFP* and *Zscan4a,c,d,e*, albeit at low levels, as well as *Tcstv1* and *Gm47924* (Fig. 3e and Supplementary Fig. 5c).

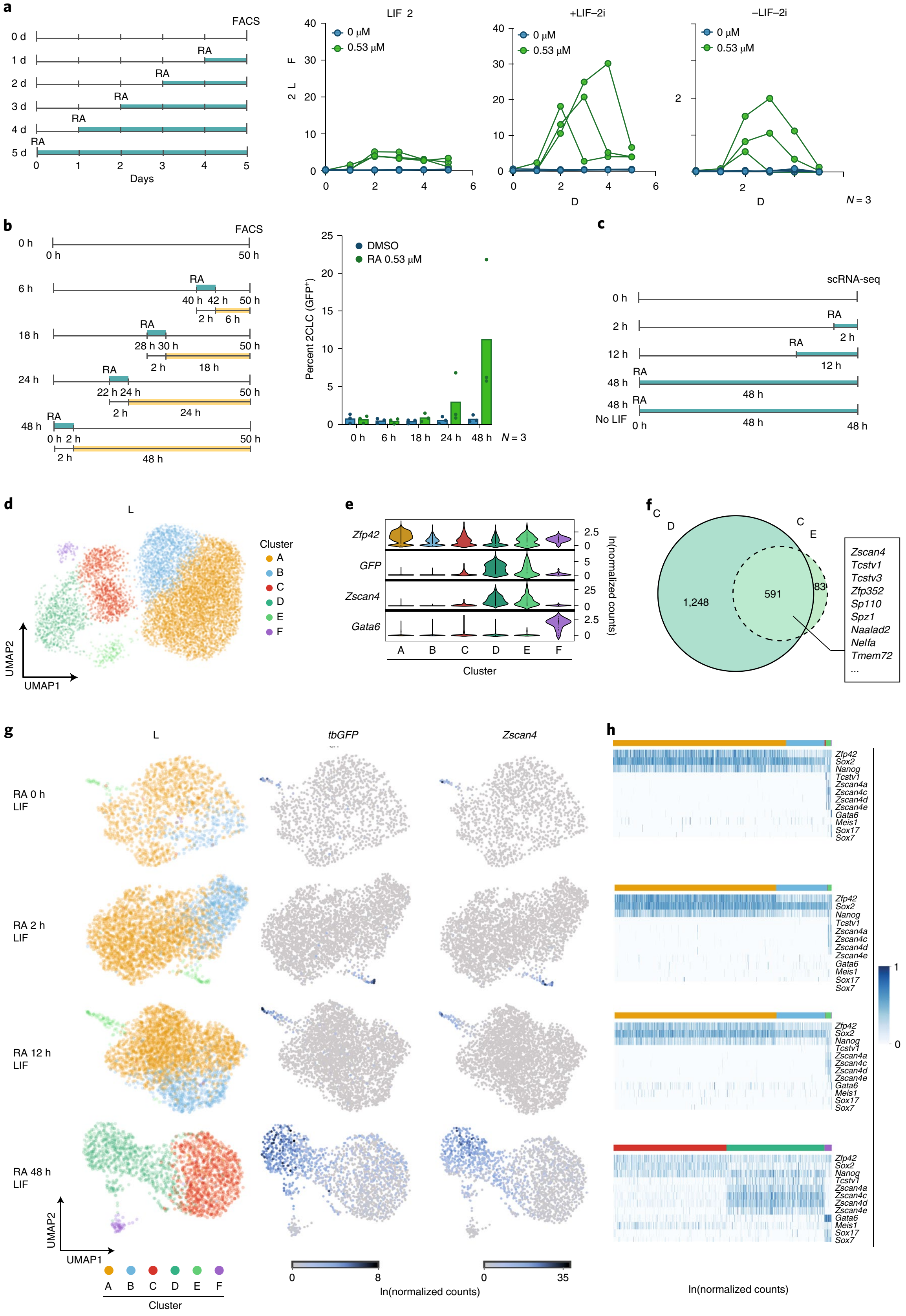
In addition to the 2CLC clusters, the two clusters comprising pluripotent ES cells exhibiting high and medium levels of *Rex1* and *Nanog* (clusters A and B) were consistently present across early timepoints (0, 2 and 12 h) and represented the majority of the cells at these timepoints (Fig. 3g). Specifically, at time 0 h, the two largest clusters expressed pluripotency markers, while the 2CLC cluster exhibited lower expression of pluripotency genes (Fig. 3h), as expected[8,10]. With longer timepoints with RA exposure, pluripotency markers expression decreased and, by 48 h, the number of 2CLCs increased drastically and a cluster of cells expressing differentiation markers emerged (cluster F; Fig. 3g,h). Importantly, the 2CLCs and the differentiating cluster do not share expression patterns and are clearly distinguishable from each other (Fig. 3g,h). This was further demonstrated when comparing scRNA-seq profiles of cells grown for 48 h with RA with LIF and without LIF (Fig. 4a). LIF removal resulted in a larger population of cells undergoing differentiation, visible as a cluster of cells expressing markers like *Gata6* (Fig. 4a,b). In line with our results above, LIF removal resulted in fewer 2CLCs compared to cells grown in LIF, upon RA stimulation (Fig. 4b). Importantly, the 2CLC cell population (*tbGFP*$^+$ and *Zscan4*$^+$) did not overlap with the population of differentiating precursor cells (*Gata6*$^+$) under these conditions either (Fig. 4a). We note that another feature that distinguishes 2CLCs (clusters D and E) from differentiating cells (cluster F) is the expression of some RA-signaling components, such as *Rxra*, which display higher expression levels in 2CLCs (see below and Fig. 5a). Thus, cells differentiating upon RA addition constitute a distinct population from 2CLCs, and ES cells can respond differently to RA stimulation, thereby generating different populations and potential cell trajectories.

To address whether RA elicits different cellular trajectories we performed RNA velocity analysis[32]. We first asked whether the scRNA-seq transcriptional dynamics faithfully recapitulates the origin of the 2CLCs that emerge from ES cells[8,10]. RNA velocity on all early timepoints (0, 2 and 12 h of RA treatment) revealed indeed a directional flow emerging from ES cells (Fig. 4c). In addition, we observed arrows denoting flow between clusters A and B, suggestive of fate transitions between naive (*Nanog/Rex1*-high) and more primed (*Nanog/Rex1*-low) ES cells, as expected[33,34]. We asked if trajectories for 2CLCs versus differentiation in response to RA can be distinguished based on transcriptional dynamics. We applied RNA velocity to our later timepoint, which revealed a strong separation between the path of differentiating precursors (purple, cluster F) and that of 2CLCs (green, cluster D) (Fig. 4d). Thus, 2CLCs undertake a clearly distinct trajectory to that of early differentiating precursors.

Next, we explored potential reasons why cells may undertake these two different trajectories. We used Slingshot to map the trajectory depicting the transition towards 2CLCs (cluster D) and the trajectory towards differentiation (cluster F) across the late timepoint. We then asked whether genes are differentially expressed along each trajectory. Different genes become activated during each transition, displaying either a sharp or a more gradual increase in gene expression (Fig. 4e,f). Among these, *Gsk3b* is downregulated in the 2CLC trajectory, suggesting potential differences in Wnt

Fig. 3 | 2CLC induction by RA is time-regulated and captured by scRNA-seq. a, Left: experimental design. ES cells containing the *2C::tbGFP* reporter were treated for a range of time periods with RA under the indicated culture conditions. 2CLC (GFP$^+$) induction was measured for all samples at the same end point by FACS. Right: percentage of 2CLCs (GFP$^+$) determined by FACS. Each line with connected dots corresponds to the measurement of one replicate. **b**, Left: experimental design. ES cells containing the *2C::tbGFP* reporter were treated with RA for 2 h, and the emergence of 2CLCs was measured at different timepoints after treatment. Right: percentage of 2CLCs (GFP$^+$) quantified by FACS. The mean of the indicated replicates (represented by individual dots) is shown. **c**, Experimental design for scRNA-seq. ES cells containing *2C::tbGFP* reporter were treated with RA for different time periods. **d**, UMAP plot from scRNA-seq comprising all cells grown with serum/LIF and treated with RA for 0 h, 2 h, 12 h or 48 h. Cells are colored based on the clusters identified by the Leiden algorithm. **e**, Violin plots showing the expression levels of selected marker genes (rows) in each cluster (columns): *Zfp42/Rex1*, marker of naive ES cells (corresponding to cluster A); *Zscan4* (computed as the sum of expression counts of genes in the *Zscan4* family) and *tbGFP* (MERVL) marking 2CLCs (clusters D and E); *Gata6* for differentiating cells (cluster F). **f**, Venn diagram comparing upregulated genes in cluster D and cluster E. **g**, UMAP plots depicting scRNA-seq data from cells grown in LIF and RA for different periods of time (rows) and colored by cluster (left column), by expression level of *GFP* (MERVL) (central column) and by expression level of *Zscan4* (calculated as the sum of the levels of genes from the *Zscan4* family; right column). **h**, Heatmaps displaying the expression levels of selected marker genes in cells at different times after RA treatment as in **g** (0 h, 2 h, 12 h, 48 h). *Zfp42/Rex1* is a marker of naive ES cells; *Sox2* and *Nanog* mark ES cells; *Tcstv1*, *Zscan4a*, *Zscan4c*, *Zscan4d* and *Zscan4e* are upregulated in 2CLCs; *Gata6*, *Sox17* and *Sox7* display higher expression levels in differentiating cells.

signaling underlying the differential response to RA (Fig. 4e and Supplementary Table 8). DE analysis of genes displaying opposite expression changes across the two trajectories identified 104 genes upregulated in the trajectory towards 2CLCs and downregulated towards differentiation (Fig. 4g). Furthermore, 750 genes were downregulated in the trajectory towards 2CLCs but upregulated towards differentiation. Altogether, 854 genes displayed transcriptional changes in response to RA across both trajectories.



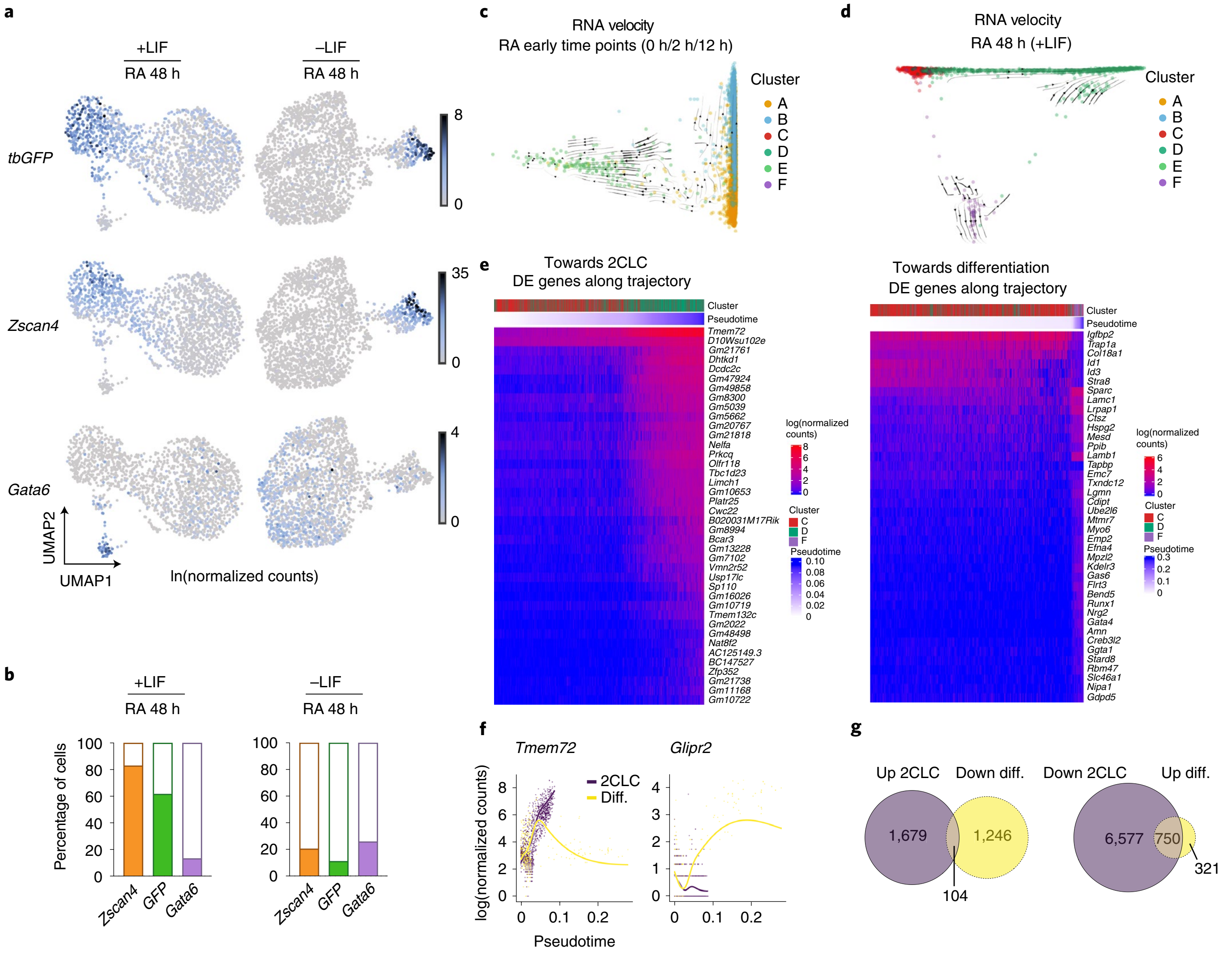


Fig. 4 | RA-reprogrammed 2CLCs differ from differentiating cells. **a**, UMAP plots of cells treated with RA for 48 h with LIF (left column) or without LIF (right column). Rows from top to bottom are colored by expression of *tbGFP* (MERVL), *Zscan4* (marking 2CLCs) and *Gata6* (marking differentiating cells). **b**, Percentages of cells where the indicated marker gene is detected (counts > 0). The left barplots refer to cells grown with LIF and the right barplots to cells grown without LIF; in both cases, cells were treated with RA for 48 h. **c**, Diffusion map with RNA velocity overlaid for cells grown in LIF and treated with RA for 0 h, 2 h and 12 h. The RNA velocity vectors indicate that cells from the ES cell clusters (A and B) are transitioning into the 2CLC cluster (E). **d**, Diffusion map with RNA velocity overlaid for cells grown in LIF and treated with RA for 48 h. Here, 2CLCs (clusters C and D) and differentiating cells (cluster F) lie on different transcriptional trajectories. **e**, Heatmaps displaying the expression of DE genes along the trajectories towards 2CLCs and towards cell differentiation based on the 48 h scRNA-seq timepoint. The cell clusters (as in Fig. 3e) and pseudotime values are indicated. **f**, Expression levels of *Tmem72* and *Glipr2* genes plotted according to the pseudotime along the cellular trajectories towards differentiation (yellow line) or 2CLCs (purple line). **g**, Venn diagram of DE genes within each of the two trajectories.

Gene list enrichment analysis revealed that GATA2 target genes (*P* value = 0.01089) were enriched in upregulated genes towards 2CLCs, in line with the known role of GATA2 in 2CLC induction[21]. By contrast, genes upregulated towards the differentiation trajectory were enriched in MAX targets ($P = 4.952 \times 10^{-24}$). Indeed, *Max* expression is downregulated exclusively across the 2CLC trajectory (Supplementary Table 8), suggesting a potential role for MAX in the distinctive response of ES cells to RA. Although the role of each of these pathways needs to be investigated, these data provide a basis for understanding the different responses elicited upon RA stimulation in ES cells.

Early embryos display endogenous RA activity. The above results indicate that RA is a primary gatekeeper of 2CLC reprogramming. Accordingly, our scRNA-seq data reveal that components of the RA signaling pathway are expressed in 2CLCs (Fig. 5a). Whether such a signaling response is a 'cell culture' feature of 2CLCs or part of the regulatory network of totipotent cells in 2-cell embryos is unclear. Indeed, while RA plays a key role in cell differentiation at later developmental stages[22,35], its receptors are expressed earlier[36]. We thus addressed whether the RA pathway is active in pre-implantation embryos. RNA-seq analysis revealed expression of proteins responsible for metabolizing retinol, RA transporters and the RA nuclear receptors prior to the blastocyst stage (Fig. 5b). RARγ displayed the highest expression levels at the late 2-cell stage (Fig. 5b), suggesting that RA may regulate gene expression in 2-cell embryos through RARγ. To test this, we asked if regulatory elements in 2-cell stage embryos contain RARE motifs. We interrogated ATAC-seq datasets[37] and found that the RARγ motif is enriched in accessible regions in early stages compared to the ICM (Fig. 5c). The enrichment in

RARE motifs was observed in 2-cell and 8-cell stage embryos, suggesting that RA activity may be important during several stages of pre-implantation embryogenesis.

Next, we addressed whether the embryos display RA activity. First, we examined the localization of the nuclear RA importers, which translocate to the nucleus to mediate RA signaling[24]. Because CRABP2 is the RA donor for RARs and FABP5 for RXRs, we focused on CRABP2 and found that its mRNA is maternally deposited (Fig. 5b). Immunostaining revealed nuclear localization of CRABP2 from the 2-cell stage onwards, but cytoplasmic in zygotes (Fig. 5d). This change in localization suggests that RA signaling may be activated at the 2-cell stage. Second, we addressed whether embryos display RA-dependent RARE transcriptional activity by microinjecting the RARE-GFP reporter in a late 2-cell stage blastomere (Fig. 5e). We monitored embryos 42–44 h later to allow for detectable GFP fluorescence. We detected RARE activity in the large majority of microinjected embryos, based on GFP fluorescence (Fig. 5f,g). This activity was RARE-dependent, because GFP was undetectable in most embryos injected with the reporter lacking RARE (Fig. 5f,g). Note that the fact that we did not see GFP expression in all embryos is expected in this type of experiment due to potential mosaicism upon plasmid injection[38]. The number of embryos expressing GFP was similar in controls (DMSO) and with RA (Fig. 5g), indicating that early embryos have endogenous RA activity. Thus, the pre-implantation embryo displays endogenous RA activity and has the machinery to regulate RARE-driven transcription.

Inhibiting RA activity compromises cleavage development. Finally, we investigated a potential role of RA signaling during the totipotency transition in embryos. To address whether RA signaling is important for pre-implantation development, we inhibited RAR signaling using a RARγ antagonist. We cultured zygotes with LY2955303 or the vehicle (DMSO). Control embryos formed blastocysts after three days (88%, $n=51$). By contrast, inhibiting RARγ prevented developmental progression, with most embryos arrested at the 2-cell or 4-cell stage (78%, $n=59$) (Fig. 6a,b). To investigate the potential involvement of other RA receptors, we treated embryos with three other antagonists against RXR homo- and heterodimers (HX531), RARα (ER50891) or both RARβ and RARγ (CD2665), but the latter with much lower affinity than LY2955303 (CD2665 Ki for RARγ is 100 times higher than LY2955303). None of these antagonists affected blastocyst formation, suggesting that only specific and robust chemical inhibition of RARγ affects developmental progression (Fig. 6c). To test this further we used siRNA against RARγ in zygotes, which led to a reduction of RARγ mRNA levels to ~8% of the controls (Fig. 6d). Knockdown of RARγ resulted in compromised developmental progression, with only ~60% of the embryos reaching the blastocyst stage (Fig. 6e). The milder phenotype observed with siRNA—as opposed to the RARγ antagonist—may be due to either incomplete protein knockdown and maternal deposition of RARγ, potential compensatory effects of other RA receptors upon RNAi, or LY2955303 potentially targeting other receptors. Unfortunately, our attempts to perform a RARγ western blot after siRNA were unsuccessful due to the low amount of material. Thus, although the RARγ antagonist treatment results in a much stronger phenotype, our siRNA results support a role for RARγ in regulating early developmental progression. However, we cannot formally exclude the possibility that other RA receptors may also be involved in RA signaling in early embryos.

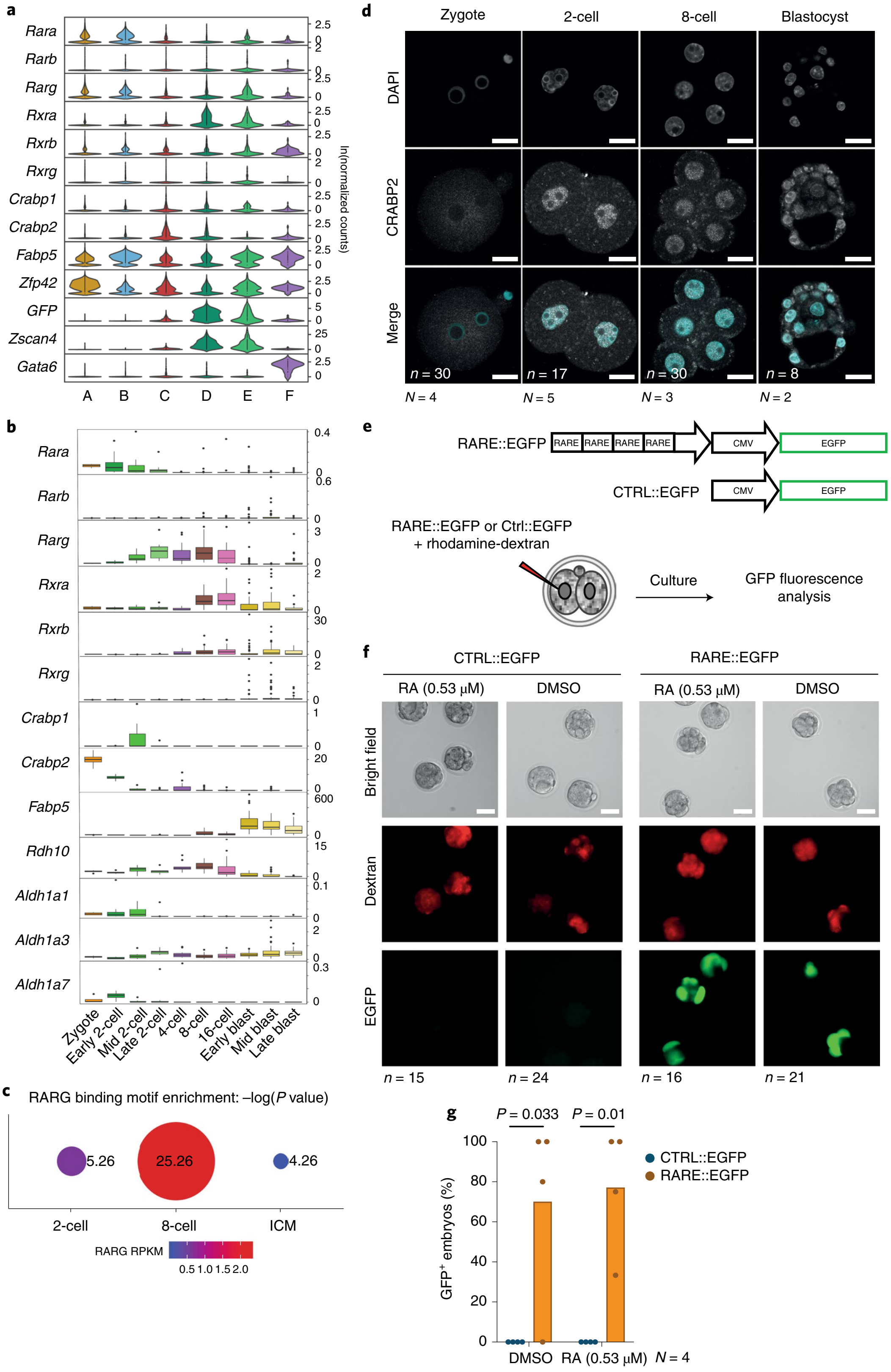
Blocking ZGA with a general RNA PolII inhibitor results in most embryos arresting at the 2-cell stage[39], similarly to the phenotype observed upon LY2955303 treatment. Thus, we next addressed if inhibiting RARγ affects ZGA by analyzing MERVL expression—a key ZGA marker—in embryos treated with LY2955303. qPCR revealed a striking reduction in MERVL transcripts in 2-cell embryos upon RARγ inhibition (Fig. 6f). These data suggest that RAR activity is necessary to ensure correct development prior to the 4-cell stage, presumably through regulation of ZGA. To address this, we performed RNA-seq[40] in late 2-cell embryos upon LY2955303 treatment (Supplementary Fig. 7a,b). DE analysis revealed no significant differences between DMSO (vehicle) and potassium simplex optimized medium (KSOM) (control) embryos, so we performed all subsequent analyses against the DMSO group. Embryos grown with LY2955303 displayed a transcriptional program that differed from controls (Supplementary Fig. 7b). LY2955303 treatment led to significant changes in gene expression, with 1,780 upregulated and 2,339 downregulated genes ($\log_2\mathrm{FC} > 1$ and $\log_2\mathrm{FC} < -1$, respectively; $P_{\mathrm{adj}} < 0.05$) (Fig. 6g and Supplementary Table 9). The majority of upregulated genes are normally highly expressed in zygotes and early 2-cell embryos (Fig. 6h), suggesting that LY2955303-treated embryos fail to progress into the transcriptional program of late 2-cell embryos. By contrast, most downregulated genes are highly expressed at the late 2-cell stage, which demarcates ZGA (Fig. 6h). Thus, chemical inhibition of RA signaling results in a failure to fully activate ZGA. Indeed, major ZGA genes were under-represented in the upregulated genes ($P = 2.2 \times 10^{-16}$, Fisher test) and over-represented in the downregulated genes ($P = 2.723 \times 10^{-11}$, Fisher test). Repetitive element expression was also affected by LY2955303, including downregulation of MERVL elements (MT2B2, MT2C_Mm and several MaLR) (Supplementary Table 9). Overall, our data suggest that RA signaling can control the '2-cell' transcriptional program both in vitro, in cell culture, as well as in vivo, in mouse embryos.

Discussion

Using a high-throughput, large-scale chemical screening, our work identifies a new regulatory pathway of 2CLC reprogramming and early mouse development. Consistent with our findings in 2CLCs, we identified a previously unappreciated activity of RA signaling at the earliest stages of embryogenesis. Thus, this work also helps to validate the use of 2CLCs as a model system for understanding the biology of the early embryo, enabling the discovery of a crucial signaling pathway at this stage of development.

Fig. 5 | The RA pathway is active in totipotent cells of the mouse embryo. a, Violin plots showing the distribution of expression of RA receptors per cluster. The lower four genes are markers for naive ES cells (*Zfp42*; cluster A); 2CLCs (*Zscan4* and *tbGFP*; clusters C, D and E); and differentiating cells (*Gata6*; cluster F). **b**, Box plots depicting the expression level of the indicated RA-pathway-related genes in pre-implantation embryos at zygote ($n=4$), early 2-cell ($n=8$), mid 2-cell ($n=12$), late 2-cell ($n=10$), 4-cell ($n=14$), 8-cell ($n=28$), 16-cell ($n=50$), early blastocyst ($n=43$), mid blastocyst ($n=60$) and late blastocyst ($n=30$) stages. The boxes denote the 25th and 75th percentiles (bottom and top of box) and median values (horizontal band inside box). The whiskers indicate the values observed within up to 1.5 times the interquartile range above and below the box. **c**, RARG motif enrichment in the open chromatin regions of the ±10 kb TSS by indicated developmental stage. Dot size, $-\log_{10}(P\ \text{value})$. **d**, Immunostaining of CRABP2 at the indicated developmental stages. Images are single confocal sections of single embryos. *n*, number of embryos analyzed. *N*, number of experimental replicates. Scale bars, 20 μm. **e**, Experimental design for the data in Fig. 6f,g. A RARE::EGFP reporter or a control plasmid lacking the *RARE* motifs was injected in one random blastomere of 2-cell-stage embryos. **f**, Representative fluorescence images of embryos with the *RARE::EGFP* reporter 44 h after microinjection of the reporter with or without RA treatment, showing embryos between late 8-cell and cavitating morula. **g**, Percentage of embryos expressing GFP from the control (CTRL) or RARE reporter. Median values of the indicated replicates (represented by individual dots) are shown. *P* values were calculated by one-sided Mann–Whitney test.

Although several factors preventing the progression to a 2CLC state are known, much less is known about positive regulators promoting 2CLCs other than DUX[9,17,41], DPPA2/4 (refs. [18,19,42]) and miR-344 (ref. [21]). Our data identify the RA signaling pathway as a core component of 2CLC identity and key regulator of 2CLC emergence. Previous work has shown that RA can increase the number of *Zscan4*$^{+}$ cells in ES cell cultures[15,43], which constitute around 5% of the ES cell population and are an intermediate cellular state between ES and 2CLCs[10]. In contrast to 2CLCs, RAR activity is not necessary for the emergence of the ZSCAN4^{+} population, although



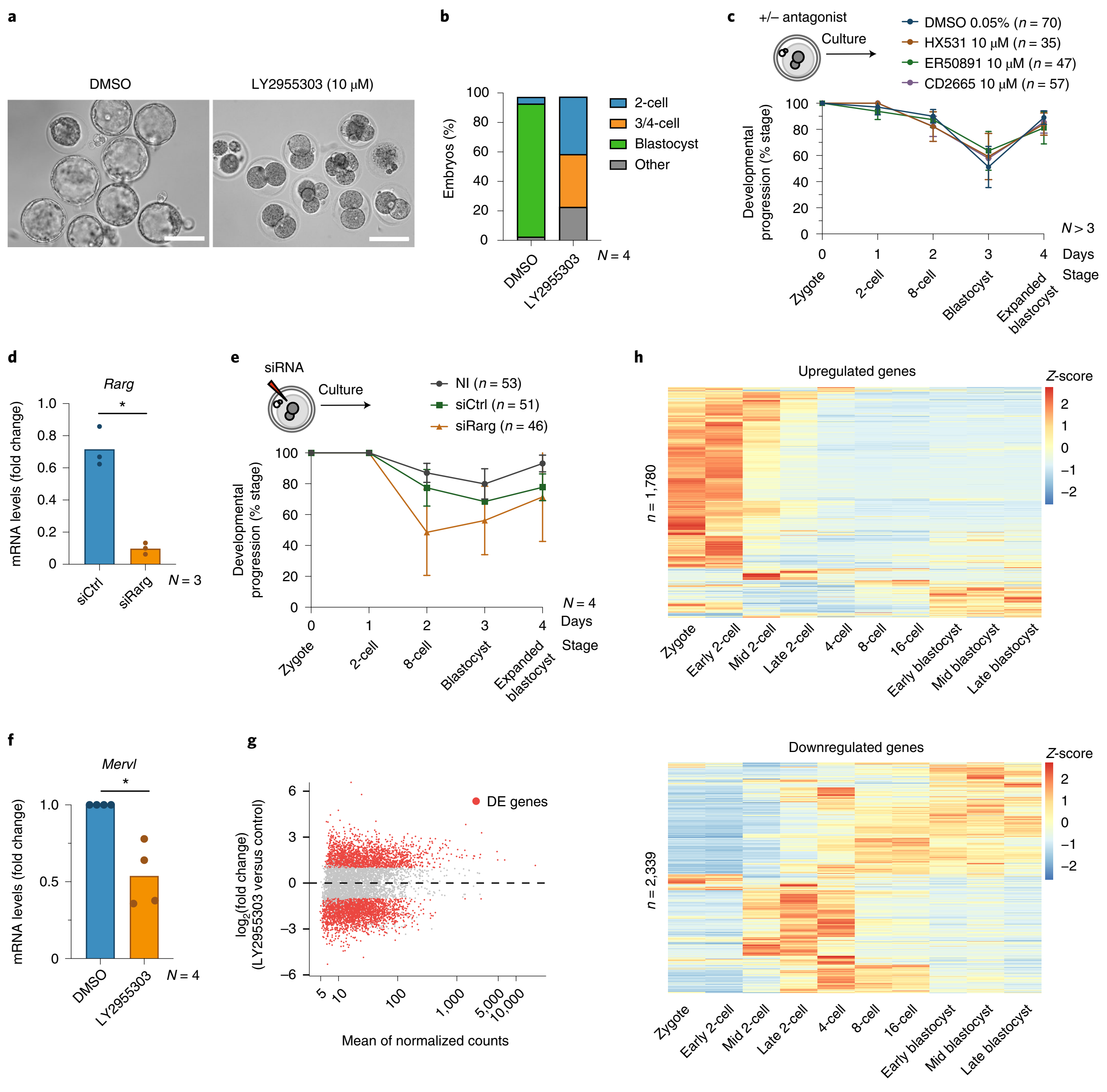


Fig. 6 | Perturbing RA signaling in the early mouse embryo affects developmental progression. **a**, Phase-contrast images of representative embryos treated with the RARγ antagonist LY2955303 or control DMSO. $N = 4$. Scale bars, 100 μm. **b**, Developmental progression (in percentage) of control (DMSO, $n = 51$) or embryos treated with the RARγ antagonist LY2955303 ($n = 59$ embryos). N, number of experimental replicates. **c**, Developmental progression of control (DMSO, $n = 70$) or embryos treated with the indicated antagonists against RXR (HX531, $n = 35$), RARα (ER50891, $n = 47$) and both RARβ and RARγ (CD2665, $n = 57$). Data are presented as mean values, and error bars represent s.d. N, number of independent replicates. **d**, qPCR analysis of *Rarg* in 2-cell stage embryos after siRNA for *Rarg* in zygotes. N, number of experimental replicates. P value calculated by two-sided Student's t-test. **e**, Developmental progression of zygotes non-injected ($n = 53$) or microinjected with scramble siRNA (control; $n = 51$) or with siRNA against *Rarg* ($n = 46$). Data are presented as mean values, and error bars represent s.d. N, number of experimental replicates. **f**, qPCR analysis of *Mervl* transcripts after LY2955303 treatment. N, number of experimental replicates. P value calculated by two-sided Student's t-test. **g**, MA plot showing differentially expressed genes in control (DMSO) 2-cell stage embryos versus LY2955303-treated embryos. Differential gene expression analysis was performed using DESeq2 (P values obtained by two-sided Wald test and corrected for multiple testing using the Benjamini and Hochberg method). Red color indicates $\log_2 FC > 1$ or < -1; $P_{adj} < 0.05$. **h**, Heatmaps depicting the endogenous expression patterns of the up- and downregulated genes between embryos treated with LY2955303 versus control embryos at the late 2-cell stage. Z-score values are shown. RNA-seq datasets are from ref. [52] (Methods).

their numbers decrease when treated with a RAR inhibitor[15]. Together with previous work, our data support a model whereby RA induces both the ZSCAN4$^+$ cells[43] as well as the transition from the ZSCAN4$^+$ state towards the 2CLC state. The identification of additional hits from our screening together with our findings on RA will enable the investigation of culture conditions to stably maintain

2CLCs. Our scRNA-seq dataset indicates that ES cells can undertake several paths in response to RA signaling and that 2CLCs are a clearly distinguishable, non-overlapping cell population, compared to early differentiating precursors. The fact that we did not detect additional cell populations between ES cells and 2CLCs in our scRNA-seq and velocity analyses may suggest that reprogramming towards the 2CLC state involves fast cellular transitions.

Whether the ability of ES cells to adopt distinct fates in response to RA signaling depends on the ability of RAR to target different genomic regions deserves further investigation. A possible mechanism whereby different doses of RA may cause different cellular responses could be the existence of different types of RA-responsive genes, for example, target genes with low versus high affinity for RARs binding, or with a different spacer length between the DR motifs. In such a scenario, a different output regarding gene expression results from different levels of transcription factor occupancy. This phenomenon has been documented for other nuclear receptors[44–46], but has not been explored for RAR/RXR. Although pan-RAR antibodies have been used in the past[47], the lack of antibodies specific for each RAR transcription factor has precluded this type of analysis. Notwithstanding, our observations that RAR motifs are significantly enriched in regulatory regions of 2CLCs and embryos at the 2- and 8-cell stages anticipates direct gene regulation by RA. Binding motifs for some transcription factors important for mouse development, such as *Nr5a2* and *Rarg*, do not show an enrichment in regulatory regions at the same stages in human pre-implantation embryos[48]. This suggests potential species-specific regulation, so a potential response to RA signaling of human induced pluripotent stem cells or ES cells will be exciting to investigate.

Identifying RA as a robust inducer of bona fide 2CLC reprogramming has allowed us to discover a new role for RA signaling in promoting the 2-cell stage program in vivo. In line with cell culture observations, chemical inhibition of RARγ results in developmental arrest, most probably due to a failure to fully trigger ZGA. Double compound mutants for RARα/RARγ are embryonic lethal at E7.5, and RARγ/RARβ double-deficient animals survive until birth[49,50]. In addition, although it is unclear whether RARγ$^{-/-}$ females display reduced fertility, they can give rise to offspring[51]. Thus, although these studies did not reveal a pre-implantation phenotype when knocked out zygotically, their function during early development may have been obscured due to maternal inheritance and redundant activities. Indeed, the intricate functional redundancy of RAR and RXR, together with the compensatory effects by their different isoforms, renders their individual analysis complex[35].

Altogether, our work sheds light into the regulatory networks underlying the reprogramming to a totipotent-like state in culture and suggests a previously unappreciated role for RA signaling at the earliest stages of mammalian embryogenesis.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41594-021-00590-w.

Received: 5 June 2020; Accepted: 7 April 2021;
Published online: 27 May 2021

References

1. Ishiuchi, T. & Torres-Padilla, M.-E. Towards an understanding of the regulatory mechanisms of totipotency. *Curr. Opin. Genet. Dev.* **23**, 512–518 (2013).
2. Wu, G. & Schöler, H. R. Lineage segregation in the totipotent embryo. *Curr. Top. Dev. Biol.* **117**, 301–317 (2016).
3. Tarkowski, A. K. Experiments on the development of isolated blastomeres of mouse eggs. *Nature* **184**, 1286–1287 (1959).
4. Togashi, M. Production of monozygotic twins by splitting of 2-cell stage embryos in mice. *Jpn J. Anim. Reprod.* **33**, 51–57 (1987).
5. Sotomaru, Y., Kato, Y. & Tsunoda, Y. Production of monozygotic twins after freezing and thawing of bisected mouse embryos. *Cryobiology* **37**, 139–145 (1998).
6. Rossant, J. & Tam, P. P. L. Blastocyst lineage formation, early embryonic asymmetries and axis patterning in the mouse. *Development* **136**, 701–713 (2009).
7. Shahbazi, M. N. & Zernicka-Goetz, M. Deconstructing and reconstructing the mouse and human early embryo. *Nat. Cell Biol.* **20**, 878–887 (2018).
8. Macfarlan, T. S. et al. Embryonic stem cell potency fluctuates with endogenous retrovirus activity. *Nature* **487**, 57–63 (2012).
9. Hendrickson, P. G. et al. Conserved roles of mouse DUX and human DUX4 in activating cleavage-stage genes and MERVL/HERVL retrotransposons. *Nat. Genet.* **49**, 925–934 (2017).
10. Rodriguez-Terrones, D. et al. A molecular roadmap for the emergence of early-embryonic-like cells in culture. *Nat. Genet.* **50**, 106–119 (2018).
11. Cerulo, L. et al. Identification of a novel gene signature of ES cells self-renewal fluctuation through system-wide analysis. *PloS ONE* **9**, e83235 (2014).
12. Peaston, A. E. et al. Retrotransposons regulate host genes in mouse oocytes and preimplantation embryos. *Dev. Cell* **7**, 597–606 (2004).
13. Bošković, A. et al. Higher chromatin mobility supports totipotency and precedes pluripotency in vivo. *Genes Dev.* **28**, 1042–1047 (2014).
14. Rodriguez-Terrones, D. et al. A distinct metabolic state arises during the emergence of 2-cell-like cells. *EMBO Rep.* **21**, e48354 (2020).
15. Tagliaferri, D. et al. Retinoic acid induces embryonic stem cells (ESCs) transition to 2 cell-like state through a coordinated expression of *Dux* and *Duxbl1*. *Front. Cell Dev. Biol.* **7**, 385 (2019).
16. Ishiuchi, T. et al. Early embryonic-like cells are induced by downregulating replication-dependent chromatin assembly. *Nat. Struct. Mol. Biol.* **22**, 662–671 (2015).
17. De Iaco, A. et al. DUX-family transcription factors regulate zygotic genome activation in placental mammals. *Nat. Genet.* **49**, 941–945 (2017).
18. De Iaco, A., Coudray, A., Duc, J. & Trono, D. DPPA2 and DPPA4 are necessary to establish a 2C-like state in mouse embryonic stem cells. *EMBO Rep.* **20**, e47382 (2019).
19. Eckersley-Maslin, M. et al. Dppa2 and Dppa4 directly regulate the Dux-driven zygotic transcriptional program. *Genes Dev.* **33**, 194–208 (2019).
20. Choi, Y. J. et al. Deficiency of microRNA *miR-34a* expands cell fate potential in pluripotent stem cells. *Science* **355**, eaag1927 (2017).
21. Yang, F. et al. DUX-miR-344-ZMYM2-mediated activation of MERVL LTRs induces a totipotent 2C-like state. *Cell Stem Cell* **26**, 234–250 (2020).
22. Rhinn, M. & Dollé, P. Retinoic acid signalling during development. *Development* **139**, 843–858 (2012).
23. Cunningham, T. J. & Duester, G. Mechanisms of retinoic acid signalling and its roles in organ and limb development. *Nat. Rev. Mol. Cell Biol.* **16**, 110–123 (2015).
24. Napoli, J. L. in *The Biochemistry of Retinoid Signaling II*: *The Physiology of Vitamin A—Uptake*, *Transport*, *Metabolism and Signaling* (eds Asson-Batres, M. A. & Rochette-Egly, C.) 21–76 (Springer, 2016).
25. Benbrook, D. M., Chambon, P., Rochette-Egly, C. & Asson-Batres, M. A. in *The Biochemistry of Retinoic Acid Receptors I*: *Structure*, *Activation and Function at the Molecular Level* (eds Asson-Batres, M. A. & Rochette-Egly, C.) 1–20 (Springer, 2014).
26. Lee, S. & Privalsky, M. L. Heterodimers of retinoic acid receptors and thyroid hormone receptors display unique combinatorial regulatory properties. *Mol. Endocrinol.* **19**, 863–878 (2005).
27. Agarwal, C., Chandraratna, R. A., Johnson, A. T., Rorke, E. A. & Eckert, R. L. AGN193109 is a highly effective antagonist of retinoid action in human ectocervical epithelial cells. *J. Biol. Chem.* **271**, 12209–12212 (1996).
28. Germain, P. et al. Differential action on coregulator interaction defines inverse retinoid agonists and neutral antagonists. *Chem. Biol.* **16**, 479–489 (2009).
29. Monaghan, J. R. & Maden, M. Visualization of retinoic acid signaling in transgenic axolotls during limb development and regeneration. *Dev. Biol.* **368**, 63–75 (2012).
30. Eckersley-Maslin, M. A. et al. MERVL/Zscan4 network activation results in transient genome-wide DNA demethylation of mESCs. *Cell Rep.* **17**, 179–192 (2016).
31. Fraichard, A. et al. In vitro differentiation of embryonic stem cells into glial cells and functional neurons. *J. Cell Sci.* **108**, 3181–3188 (1995).
32. La Manno, G. et al. RNA velocity of single cells. *Nature* **560**, 494–498 (2018).
33. Kalmar, T. et al. Regulated fluctuations in Nanog expression mediate cell fate decisions in embryonic stem cells. *PLoS Biol.* **7**, e1000149 (2009).
34. Osorno, R. & Chambers, I. Transcription factor heterogeneity and epiblast pluripotency. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **366**, 2230–2237 (2011).

35. Mark, M., Ghyselinck, N. B. & Chambon, P. Function of retinoic acid receptors during embryonic development. *Nucl. Recept. Signal.* **7**, e002 (2009).
36. Ulven, S. M. et al. Identification of endogenous retinoids, enzymes, binding proteins and receptors during early postimplantation development in mouse: important role of retinal dehydrogenase type 2 in synthesis of all-*trans*-retinoic acid. *Dev. Biol.* **220**, 379–391 (2000).
37. Wu, J. et al. The landscape of accessible chromatin in mammalian preimplantation embryos. *Nature* **534**, 652–657 (2016).
38. Iqbal, K. et al. Cytoplasmic injection of circular plasmids allows targeted expression in mammalian embryos. *BioTechniques* **47**, 959–968 (2009).
39. Warner, C. M. & Versteegh, L. R. In vivo and in vitro effect of α-amanitin on preimplantation mouse embryo RNA polymerase. *Nature* **248**, 678–680 (1974).
40. Picelli, S. et al. Smart-Seq2 for sensitive full-length transcriptome profiling in single cells. *Nat. Methods* **10**, 1096–1098 (2013).
41. Whiddon, J. L., Langford, A. T., Wong, C.-J., Zhong, J. W. & Tapscott, S. J. Conservation and innovation in the DUX4-family gene network. *Nat. Genet.* **49**, 935–940 (2017).
42. Yan, Y.-L. et al. DPPA2/4 and SUMO E3 ligase PIAS4 opposingly regulate zygotic transcriptional program. *PLoS Biol.* **17**, e3000324 (2019).
43. Tagliaferri, D. et al. Retinoic acid specifically enhances embryonic stem cell metastate marked by Zscan4. *PloS ONE* **11**, e0147683 (2016).
44. Penvose, A., Keenan, J. L., Bray, D., Ramlall, V. & Siggers, T. Comprehensive study of nuclear receptor DNA binding provides a revised framework for understanding receptor specificity. *Nat. Commun.* **10**, 2514 (2019).
45. Watson, L. C. et al. The glucocorticoid receptor dimer interface allosterically transmits sequence-specific DNA signals. *Nat. Struct. Mol. Biol.* **20**, 876–883 (2013).
46. Giguère, V. Orphan nuclear receptors: from gene to function. *Endocr. Rev.* **20**, 689–725 (1999).
47. Chatagnon, A. et al. RAR/RXR binding dynamics distinguish pluripotency from differentiation associated *cis*-regulatory elements. *Nucleic Acids Res.* **43**, 4833–4854 (2015).
48. Wu, J. et al. Chromatin analysis in human early development reveals epigenetic transition during ZGA. *Nature* **557**, 256–260 (2018).
49. Lohnes, D. et al. Function of the retinoic acid receptors (RARs) during development (I). Craniofacial and skeletal abnormalities in RAR double mutants. *Development* **120**, 2723–2748 (1994).
50. Mendelsohn, C. et al. Function of the retinoic acid receptors (RARs) during development (II). Multiple abnormalities at various stages of organogenesis in RAR double mutants. *Development* **120**, 2749–2771 (1994).
51. Lohnes, D. et al. Function of retinoic acid receptor γ in the mouse. *Cell* **73**, 643–658 (1993).
52. Deng, Q., Ramsköld, D., Reinius, B. & Sandberg, R. Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science* **343**, 193–196 (2014).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2021, corrected publication 2022

Methods

Cell culture. Cells were grown in medium containing DMEM-GlutaMAX-I, 15% FBS, 0.1 mM 2-beta-mercaptoethanol, non-essential amino acids, penicillin and streptomycin and 2× LIF over gelatin-coated plates. Medium was supplemented with 2i (3 µM CHIR99021 and 1 µM PD0324901, Miltenyi Biotec) for maintenance and expansion. The 2i was removed 24 h before starting experiments.

Flow cytometry. Before cytometry, cells were washed with PBS, trypsinized with trypsin-EDTA 0.1% and resuspended in 0.5% BSA PBS solution at 4 °C. Cells were kept on ice until sorting, performed using a BD BioSciences FACS Aria III. Analysis was done with FlowJo software (the gating strategy is shown in Supplementary Fig. 7c). For the RA effect on GFP$^-$ cells experiment, the GFP$^-$ gate was defined based on the fluorescence of wild-type (WT) ES cells and 2CLCs were removed before RA treatment. For scRNA-seq, treatments started at different timepoints so that all experimental conditions were collected at the same time. Samples were sorted to enrich the population in living single cells and library preparation was conducted immediately.

Real-time polymerase chain reaction. Total RNA was extracted using phenol–chloroform extraction using TRIzol reagent (Invitrogen). Reverse transcription was performed with a First Strand cDNA synthesis kit (Roche) following the manufacturer's instructions with random hexamers. Real-time PCR was performed with GoTaq qPCR Master Mix (Promega) on a LightCycler 96 Real-time PCR system (Roche). The relative expression level of each gene was normalized to *Rps28* and *Actb*. The primers used are listed in Supplementary Table 10. Data were plotted with GraphPad Prism.

siRNA transfection. One day before transfection, 2i inhibitors were removed. siRNA transfection was performed using Lipofectamine RNAi MAX (Life Technologies). A total of 75,000 cells were transfected per condition and well in 24-well gelatin-coated plates, with a final siRNA concentration of 30 nM. Silenced Negative Control No.1 (Life Technologies) was used. The siRNAs are listed in Supplementary Table 9. The effect of siRNA silencing was examined three days after transfection and two days after RA treatment (qPCR primers are listed in Supplementary Table 11).

Immunofluorescence. The *2C::turboGFP* cell line was cultured on gelatin-coated coverslips. At 48 h after RA treatment, cells were washed with PBS, fixed with 4% PFA for 10 min at room temperature and, after four washes with PBS, permeabilized with 0.3% Triton X-100 for 10 min at room temperature. After washing with PBS, primary antibodies were incubated overnight at 4 °C, followed by another three washes in PBS. The antibodies used were mouse turboGFP (TA140041, Origene) and rabbit Zscan4 (AB4340, EMD Millipore). Secondary antibodies were incubated for 1 h at room temperature. Mounting was done in Vectashield mounting medium (Vector Labs). Images were acquired using a Leica SP8 confocal microscope.

Reporter cell lines. The *2C::tdTomato* and *2C::turboGFP/Zscan4::mCherry* lines have been previously described[10,16]. To generate *2C::turboGFP* reporter, ES cells were transfected with a plasmid containing a destabilized NLS-tagged turboGFP cassette under the regulation of *Mervl* LTR using Lipofecramine 2000. A single clone was selected from successfully transfected cells and has been fully characterized elsewhere (Nakatani et al., manuscript in preparation).

Small-molecule screening. Plate and liquid handling was performed using an HTS platform system composed of a Sciclone G3 liquid handler from PerkinElmer with a Mitsubishi robotic arm (Mitsubishi Electric, RV-3S11), a MultiFlo dispenser (Biotek Instruments) as well as a Cytomat incubator (Thermo Fisher Scientific). Cell seeding and assays were performed in black 384-well CellCarrier plates (PerkinElmer, 6007558). The plates were coated with gelatin 0.1% for 20 min at 37 °C to facilitate better cell adherence. Cells were seeded in 384-well microplates with 10,000 cells per well. Image acquisition and image-based quantification was done using the Operetta/Harmony high-throughput imaging platform (PerkinElmer). Z' factors were calculated according to the formula $Z' = 1 - (3(\theta_p + \theta_n)/(\mu_p - \mu_n))$, where p is the positive control, n is the negative control, θ is the standard deviation and μ is the mean.

Screening assay. *2C::turboGFP* ES cells were washed with 1× PBS, trypsinized and resuspended to a density of 90,909 cells ml^{-1} in cell culture medium. The cell suspension (10,000 cells per well; 110 µl per well) was dispensed into assay 384-well plates and incubated at 37 °C in 5% CO_2. The same day, cells were treated either with compound (1 mM stock solution) dissolved in 100% dimethyl sulfoxide (DMSO) or DMSO alone, then 0.7 µl of compounds/DMSO were transferred to 110 µl cell culture medium per well to keep the final DMSO volume concentration below 0.7%. The positive control (10,000 cells per 110 µl) with 32 mM acetate and 0.7% DMSO was seeded separately after compound transfer in columns 23 and 24 of the 384-well assay plates. The cells were then incubated (37 °C, 5% CO_2) for 48 h before fixation and antibody staining. Cells were permeabilized with PBS-Triton 0.3% for 5 min at room temperature (RT). After washing with PBS and blocking with PBS-BSA 1% for 1 h, primary anti-tbGFP antibody (TA140041) was added overnight at 4 °C. After washes with PBS, cells were incubated with Alexa488 anti-mouse secondary antibody, for 1 h at RT. After washes with PBS, cells were incubated with PBS-Hoechst 33342 (1 µg ml^{-1}) for 15 min at RT. Cells were again washed with PBS. Finally, plates were recorded using the automated Operetta microscope using the ×20 NA objective for high-resolution images (PerkinElmer). For quantification, six images of each condition were recorded. This resulted in a cell number of ~100 cells of each condition in control wells with DMSO.

Image analysis. Multiparametric image analysis was performed using Columbus high-content imaging and analysis software version 2.8.0 (PerkinElmer Life Sciences). Hoechst signal was used to detect cell nuclei using method C with the following parameters: common threshold (parameter determining the lower level of pixel intensity for the whole image that may belong to nuclei), 0.30; area (to tune the merging and splitting of nuclei during nuclei detection), >30 µm^2; split factor (parameter influencing the decision of the computer of whether a large object is split into two or more smaller objects or not), 10; individual threshold (parameter determining the intensity threshold for each object individually), 0.2; contrast (parameter setting a lower threshold to the contrast of detected nuclei), 0.1. Next, the area of nuclei and the Hoechst intensity were determined and the nuclei were filtered by these properties (nucleus area >20 µm^2 and <400 µm^2; intensity > 100). For this subpopulation called 'Nuclei selected' the median intensity of the GFP signal was calculated and used to select the green cell population (intensity > 600). The percentage of the green cells was calculated. In addition, the whole image area was defined and the mean GFP signal was calculated to exclude wells with green fluorescent compounds (intensity < 400).

Embryo collection and immunostaining. Experiments were carried out according to valid legislation and in compliance with the local government (Government of Upper Bavaria). Mice were bred in a 12-h light cycle. Housing conditions were according to ETS 123 guidelines: 20–24 °C and 45–65% humidity. Embryos were collected for immunostaining as described in ref. [53] from CD1 ~6-week-old females that were crossed with CD1 males upon natural matings. Embryos were fixed immediately after collection. The zona pellucida was removed with acid Tyrode's solution (Sigma), and embryos were washed three times in PBS and fixed[54]. After permeabilization, embryos were washed three times in PBS-T (0.1% Tween in PBS), free aldehydes were removed by short incubation in NH_4Cl (2.6 mg ml^{-1}) and the embryos were washed twice in PBS-T. The embryos were blocked and incubated with anti-CRABP2 antibody, then washed three times in PBS-T, blocked and incubated with the corresponding secondary antibodies (A488-conjugated goat anti rabbit immunoglobulin-G). After washes in PBS-T and PBS, embryos were mounted in Vectashield with DAPI (Vector Laboratories) and imaged under a Leica SP8 inverted confocal microscope using a ×63 oil objective across 0.5-µm stacks. Blastocysts were mounted in three dimensions and imaged across a 1-µm stack.

Microinjection and embryo manipulation. For the RARE::GFP reporter plasmid experiments, 2-cell-stage embryos were collected from 5–8-week-old F1 (CBAxC57BL/6J) females mated with F1 males 42–44 h post hCG injection. Ovulation was induced by injecting 10 IU pregnant mare serum gonadotropin (PMSG) (IDT Biologika) and human chorionic gonadotrophin (hCG) (MSD Animal Health) 48 h later. A single, random blastomere was microinjected with 1–2 pl of 20 ng µl^{-1} of the RARE plasmid or the plasmid without the RARE sequences. Dextran rhodamine (1 mg ml^{-1}) was added as the microinjection control. Embryos were cultured in KSOM and monitored regularly. For RNAi, zygotes were collected from 5–8-week-old F1 (CBAxC57BL/6J) females mated with F1 males at 17–19 h post hCG injection and microinjected with 1–2 pl of 25 µM siRarg pool (Horizon Discovery M-04974-01-005) or siControl[10]. GFP mRNA (100 ng) was added as positive control for microinjection. Embryos were cultured in KSOM and monitored regularly. At 20 h post injection, some embryos were washed in PBS and frozen for qPCR. For the experiments with antagonists, zygotes were collected at 18 h post hCG injection and randomly allocated to the experimental groups, then cultured in the presence of 10 µM LY2955303, HX531, ER50891 or CD2665 (Tocris 3912, 2823 and 3800, respectively) in 0.05% DMSO or DMSO 0.05% in KSOM and scored daily for developmental progression. The data were plotted with GraphPad Prism.

Embryo real-time qPCR. Total RNA was obtained from 20–25 2-cell embryos using the Arcutus PicoPure RNA isolation kit (Applied Biosystems 12204-01). Reverse transcription was performed with Superscript IV reverse transcriptase (Invitrogen 18090010) following the manufacturer's instructions, with random hexamers. Real-time PCR was performed with Roche SYBR Green I Master Mix (04707516001) on a LightCycler 96 real-time PCR system (Roche). The relative expression level of each gene was normalized to *Gapdh* and *Actb*.

Single embryo RNA-seq. Zygotes were collected at 18 h post hCG injection and cultured in the presence of 10 µM LY2955303 in 0.05% DMSO, 0.05% DMSO in KSOM or KSOM alone. Embryos were cultured until the late 2-cell stage (48 h post hCG), washed in PBS at 37 °C and flash-frozen in lysis buffer according to the Smart-Seq2 protocol. Libraries were verified using a 2100 Bioanalyzer

(Agilent). Samples were paired-end sequenced at PE250 on an Illumina NovaSeq 6000 platform.

Single-cell RNA-seq. Cells were collected after RA treatment and sorted for live single cells by FACS. Cell were then counted and tested for viability with an automated cell counter. Five thousand cells of the sample were then input into the 10X protocol. Gel bead-in-emulsion (GEM) generation, reverse transcription, cDNA amplification and library construction steps were performed according to the manufacturer's instructions (Chromium Single Cell 3′ v3, 10X Genomics). Samples were run on an Illumina NovaSeq 6000 platform.

Gene counting. Unique molecular identifier (UMI) counts were obtained using the kallisto (version 0.46.0) bustools (version 0.39.3) pipeline[55]. First, mouse transcriptome and genome (release 98) fasta and gtf files were downloaded from the Ensembl website, and 10X barcodes list version 3 was downloaded from the bustools website. We built an index file with the 'kallisto index' function with default parameters. Then, pseudoalignment was done using the 'kallisto bus' function with default parameters and the barcodes for 10X version 3. The BUS files were corrected for barcode errors with 'bustools correct' (default parameters), and a gene count matrix was obtained with 'bustools count' (default parameters). To estimate the *tbGFP* read counts, we used the *tbGFP* sequence available from GenBank (ID ASW25889.1) and followed the same procedure.

Quality control and normalization. To remove barcodes corresponding to empty droplets, we used the 'emptyDrops' function from the R library 'DropletUtils' version 1.6.1 (ref. [56]). For this, a lower threshold of 1,000 UMI counts per barcode was considered. Afterwards, quality control was performed using Python library 'scanpy' version 1.4.2 (ref. [57]). Cells were filtered by fraction of mitochondrial reads and number of detected genes. Cells having more than 10% counts mapped to mitochondrial genes or fewer than 1,000 detected genes were removed (Supplementary Fig. 4). Then data from *tbGFP* expression were integrated and count tables from each timepoint were normalized separately using the R library 'scran' (version 1.14.0)[58] as follows. First, the function 'quickCluster' was run, then size factors were calculated based on this clustering using the function 'computeSumFactors' with default parameters. Finally, the data were normalized using the computed size factors.

Batch correction and regressing out of confounding effects. We performed batch correction on the data with LIF with the mutual nearest neighbors (MNN) method[59] (function 'mnn_correct' from the 'mnnpy' library; https://github.com/chriscainx/mnnpy), using as input the log-transformed normalized counts of the genes that were in the list of top 3,000 highly variable genes (HVGs) at every timepoint, as done in ref. [59] (highly variable genes were identified with the function 'highly_variable_genes' in the scanpy library with the following parameters: min_disp=0.3, inplace=False, n_top_genes=3000). Afterwards, only genes with more than two counts in at least two cells were kept for further analysis and the data were scaled using the function 'pp.scale' from scanpy. On this batch-corrected data, the number of detected genes was regressed out using the scanpy function 'regress_out'.

Data visualization, clustering and diffusion maps. We used UMAP[60] for data visualization ('umap' function in scanpy, with options n_components=2, min_dist=1). Leiden clustering was performed on the top 3,000 HVGs calculated across the whole dataset (with $k = 15$ and resolution $= 0.4$) using a correlation distance in the 'pp.neighbors' function from scanpy. To identify marker genes for a given cluster, first we found differentially expressed genes between that cluster and any other cluster (Wilcoxon's rank sum test, false discovery rate (FDR) < 0.1, $\log_2\mathrm{FC} > 1$), then genes were ranked according to their mean FDRs computed across all pairwise comparisons. To validate the differentiation state of the clusters suggested by the markers, the expression of some previously known relevant genes (*Rex1*, *Sox2*, *Nanog*, *Tcstv1*, *Zscan4a*, *Zscan4c*, *Zscan4d*, *Zscan4e*, *Gata6*, *Meis1*, *Sox17* and *Sox7*) was plotted on UMAP. Cells were aligned along a pseudotime trajectory using a diffusion map[61], which was computed with the 'diffmap' function from the scanpy package on the first 20 principal components. We performed all differential gene expression analyses with Wilcoxon's rank sum test, with an FDR threshold of 0.1 and $\log_2$FC threshold of 1.

RNA velocity. To estimate RNA velocities[62], we obtained loom files as described in the following. Fastq files were aligned using STAR (version 2.7.3a)[63]. Genome indices were generated using STAR --runMode genomeGenerate with default parameters. Then, alignment of reads was performed with the following options: --runThreadN 8 --outSAMunmapped Within. The resulting SAM files were converted to bam format and sorted using samtools[64] (version 0.1.19-44428cd). Uniquely aligned reads from cells that passed the quality control were selected and distributed in separate bam files. We ran velocyto (version 0.17.17)[62] with the option run-smartseq2 on bam files from cells corresponding to each timepoint to generate one loom file of spliced and unspliced counts per timepoint. On these loom files, we ran 'scvelo'[65] to perform RNA velocity analysis. This was done separately for the early timepoints (0 h, 2 h and 12 h) and the 48 h + LIF dataset. Second-order moments (steady-state levels) were calculated with the function 'pp.moments'. These values were used for computing velocities using the function 'tl.velocity' with the following options: mode='stochastic', min_r2=0.001. RNA velocity was plotted on a diffusion map colored by cluster with the function 'pl.velocity_embedding_stream' from scvelo.

Cellular trajectory analysis. The trajectories analysis was performed in R (version 4.0.2) using the R package slingshot[66] (version 1.6.1) on the 48 h dataset with the main clusters. As input for slingshot, we used the original main clusters (2, 3 and 5) and the diffusion map (function DiffusionMap from the R library destiny[67] computed on the top 3,000 HVGs identified with the function FindVariableFeatures (with selection.method='vst') from the R library Seurat. Data were normalized using the function NormalizeData (with parameter normalization.method equal to 'LogNormalize') from the R library Seurat[68] (version 3.2.0). DE analysis was done with the R package tradeSeq[69] (version 1.2.1). For detecting the DE genes along the two trajectories we used the function startVsEndTest. Identification of the genes that are most different between the two trajectories was performed with the function patternTest with parameters l2fc equal to $\log_2(1.5)$ and nPoints equal to 50.

Single-embryo RNA-seq analysis. Data quality was assessed with FastQC (version 0.11.7). Reads were processed with Trimmomatic (version .0.39) to remove Nextera adaptors and over-represented sequences. Reads were subsequently mapped to the mouse genome M25 (GRCm38.p6) and quantified using kallisto (version 0.44.0). Reads were imported into R (version 4.0.2) by the tximport package and the Scater and Single Cell Experiment packages were used to perform quality control tests by comparing library size, number of expressed genes and proportion of mitochondrial genes, for which the applied thresholds were 30,000 reads as the minimum for library size, 5,000 genes as minimum for the number of expressed genes and 20% as the maximum for the proportion of mitochondrial genes. Accordingly, one of the LY2955303 samples was removed as an 'outlier', because it did not pass the QC threshold (Supplementary Fig. 7a). Embryos with an average number of counts of $\geq$10 were kept for subsequent analysis. The average number of counts was calculated using the calculateAverage function from the scater package, where size-adjusted average count is defined by dividing each count by the size factor and taking the average across embryos. Principal component analysis was used to analyze the three groups of embryos (KSOM, DMSO or LY2955303) using log-transformed and library size-normalized counts using the top 3,650 high variable genes, which were calculated using modelGeneVar() and getTopHVGs() functions from the scran package. Differential gene expression analysis was performed using DESeq2 (version 1.28.1) with the threshold of an adjusted P value < 0.05 to select DE genes. Upregulated and downregulated DE genes from LY2955303 versus DMSO embryos with $\log_2$FC of >1 and <-1, respectively, were selected to show how they were expressed in WT embryos, based on RPKM values of published data[52]. RPKM values of the genes with non-zero counts were transformed to Z-scores to produce the relevant heatmaps. For repetitive elements analysis, trimmed reads were mapped to the primary assembly of the mouse genome M25 (GRCm38.p6) using STAR (version 2.7.6a) with the following parameters: --readFilesCommand zcat --outFilterType BySJout --outFilterMultimapNmax 100 --winAnchorMultimapNmax 200 --alignSJoverhangMin 8 --alignSJDBoverhangMin 1 --outFilterMismatchNmax 999 --alignIntronMin 20 --alignIntronMax 0 --alignMatesGapMax 0 --outSAMprimaryFlag AllBestScore --outMultimapperOrder Random --outSAMstrandField intronMotif --runRNGseed 13 --outSAMtype BAM Unsorted --quantMode GeneCounts --twopassMode Basic. Mapped reads to genes and TEs were counted using TEtranscripts (v.2.1.4), where the used GTF file for TE annotations was mm10_rmsk_TE.gtf. Finally, DE analysis was performed as described above using the count table generated from TEtranscripts. The list of 'major' ZGA genes has already been published[70].

Assay for transposase-accessible chromatin sequencing analysis and transcription factor binding site enrichment analysis. ATAC-seq data from 2CLC and ES cells[30] (GSE75751) was downloaded, reads were trimmed using trimmomatic (version 0.38) with parameters 3:30:8:1:true LEADING:10 TRAILING:10 SLIDINGWINDOW:5:10 MINLEN:30. The output was aligned to the mm10 (vM21 GRCm38.p6) mouse genome from GENCODE, using bowtie2 with the parameters --dovetail --no-discordant --no-mixed -X 1500. BAM files were cleaned keeping the uniquely mapped reads using the samtools functions fixmate, sort and view -q 14. Peaks were called using macs2 v2.1.2.20181002 --bdg -q 0.01 -SPMR --keep-dup all --call-summits. The ATAC-seq data from mouse embryos[37] (GSE66390) were preprocessed and aligned as above. Peak-calling was also done with macs2, with parameters --bdg -q 0.01 --nomodel --nolambda --keep-dup, all as reported by the authors of that study. The transcription factor binding site enrichment analysis was done using the software Analysis of Motif Enrichment (AME) from the MEME suite v5.0.5, using Fisher's exact test to assess the relative enrichment and --kmer 1. The binding motif matrices used for the scanning were downloaded from JASPAR. 2CLC and ES cell RNA-seq (GSE75751) reads were trimmed in the same way as just described. The output reads were pseudoaligned with kallisto v0.44.0, using the mm10 (vM21 GRCm38.p6) mouse transcriptome available in GENCODE. Counts were normalized as RPKM. The

RNA-seq data from mouse embryos were from GSE66390 and were processed following the same pipeline as for 2CLCs and ES cells RNA-seq.

Statistical analyses. Statistical tests were performed keeping in mind the data distribution and the number of data points available. For all the qPCR analyses, because each replicate represents the mean expression level of the particular gene for thousands of cells, the data follow a normal distribution according to the central limit theorem. We thus applied the *t*-test (unpaired) for all statistically relevant comparisons. Across the manuscript, data on the percentage of 2CLCs in control conditions were gathered ($n=99$) and a Shapiro–Wilk test was used to test if they were normally distributed. The test returned a significant P value, discarding a normal distribution. Therefore, a non-parametric test was used (Mann–Whitney, unpaired) to compare the 2CLC percentage between conditions whenever $N \geq 4$. Additional details on sample sizes, in addition to the statistical tests conducted, are presented in the corresponding figure legends.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this Article.

Data availability

scRNA-seq data generated in this study are available under ArrayExpress accession no. E-MTAB-8869 and single-embryo RNA-seq data under accession no. E-MTAB-9940. All other data supporting the findings of this study are available from the corresponding author on reasonable request.

Code availability

All scRNA-seq data were analyzed with standard programs and packages, as detailed in the Methods. Code is available on request.

References

53. Hogan, B., Beddington, R. & Costantini, F. (eds) *Manipulating the Mouse Embryo*: *A Laboratory Manual* (Cold Spring Harbor Laboratory Press, 1994).
54. Torres-Padilla, M. E. & Zernicka-Goetz, M. Role of TIF1α as a modulator of embryonic transcription in the mouse zygote. *J. Cell Biol.* **174**, 329–338 (2006).
55. Melsted, P. et al. Modular, efficient and constant-memory single-cell RNA-seq preprocessing. *Nat. Biotechnol.* https://doi.org/10.1038/s41587-021-00870-2 (2021).
56. Lun, A. T. L. et al. EmptyDrops: distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data. *Genome Biol.* **20**, 63 (2019).
57. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).
58. Lun, A. T. L., McCarthy, D. J. & Marioni, J. C. A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. *F1000Res.* **5**, 2122 (2016).
59. Haghverdi, L., Lun, A. T. L., Morgan, M. D. & Marioni, J. C. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat. Biotechnol.* **36**, 421–427 (2018).
60. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for dimension reduction. Preprint at https://arxiv.org/pdf/1802.03426.pdf (2018).
61. Haghverdi, L., Buettner, F. & Theis, F. J. Diffusion maps for high-dimensional single-cell analysis of differentiation data. *Bioinformatics* **31**, 2989–2998 (2015).
62. La Manno, G. et al. RNA velocity of single cells. *Nature* **560**, 494–498 (2018).
63. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
64. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
65. Bergen, V., Lange, M., Peidli, S., Wolf, F. A. & Theis, F. J. Generalizing RNA velocity to transient cell states through dynamical modeling. *Nat. Biotechnol.* **38**, 1408–1414 (2020).
66. Street, K. et al. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics* **19**, 477 (2018).
67. Angerer, P. et al. destiny: diffusion maps for large-scale single-cell data in R. *Bioinformatics* **32**, 1241–1243 (2016).
68. Stuart, T. et al. Comprehensive integration of single-cell data. *Cell* **177**, 1888–1902 (2019).
69. Van den Berge, K. et al. Trajectory-based differential expression analysis for single-cell sequencing data. *Nat. Commun.* **11**, 1201 (2020).
70. Park, S.-J. et al. Inferring the choreography of parental genomes during fertilization from ultralarge-scale whole-transcriptome analysis. *Genes Dev.* **27**, 2736–2748 (2013).

Acknowledgements

We thank D. Pich and W. Hammerschmidt for advice and access to FACS, L. Altamirano-Pacheco for advice regarding bioinformatic analyses and A. Burton and S. Hamperl for critical reading of the manuscript. Work in the Torres-Padilla laboratory is funded by the Helmholtz Association, HMGU Small Molecule projects (Developmental projects), the German Research Council (CRC 1064) and H2020 Marie-Curie Actions ITN EpiSystem and ChromDesign. A.I. is a recipient of a long-term EMBO fellowship (ALTF 383-2016).

Author contributions

A.I. and M.-E.T.-P. conceived the project. A.I., C.N. and K.S. performed and designed experiments. M.L.R.T.S., I.R., E.R.R.-M., G.L. and A.A. performed computational analysis with the supervision of K.H., A.S. and M.-E.T.-P. M.-E.T.-P wrote the manuscript with input from all authors.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41594-021-00590-w.

Correspondence and requests for materials should be addressed to M.-E.T.-P.

Peer review information *Nature Structural & Molecular Biology* thanks Bin Gu and Duanqing Pei for their contribution to the peer review of this work. Beth Moorefield was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Reprints and permissions information is available at www.nature.com/reprints.

nature research

Corresponding author(s): Maria-Elena Torres-Padilla

Last updated by author(s): Apr 2, 2021

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's d, Pearson's r), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	For small molecule screening, multiparametric image analysis was performed using Columbus high-content imaging and analysis software (version 2.8.0). For cell and embryo immunofluoresce image collections, Fiji (version 1.0) was used. FACS data was collected using Diva software from BD.
Data analysis	GraphPad Prism (version 8) and RStudio (version 1.1.383) were used for data analysis. Adobe Creative Suite was used for Figure preparation: Illustrator (CS6 version 16.0.0). The R programming language (versions R-3.6.3 and R-4.0.2) (https://www.R-project.org/) was widely used within the study for statistical analysis and data plotting, all custom code is avilable on request. For FACS experiments, data was analyzed using FlowJo (version 10). For single cell RNAseq analysis, UMI counts were obtained using the kallisto (version 0.46.0) – Bustools (version 0.39.3) pipeline and the barcodes for 10x version 3. For quality control and normalization, R libraries DropletUtils (version 1.6.1) and scran (version 1.14.0) and Python library scanpy (version 1.4.256) were used. Data visualization was done using Leiden algorithm for clustering and plotting using UMAP with Python library scanpy (version 1.4.256). For RNA velocity, alignment was done with STAR (version 2.7.3a) and analysis with velocyto (version 0.17.17) and scvelo (version 0.1.24). For single embryo RNAseq analysis, data quality was checked using FastQC (version 0.11.7), reads were processed with Trimmomatic (version 0.39) and quantified using kallisto (0.44.0). Reads were imported into R (version 4.0.2) by tximport package (version 1.16.1) and then Scater (version 1.16.2) and Single Cell Experiment (version 1.10.1) packages were used to perform quality control tests. Differential gene expression analysis was performed using DESeq2 (version 1.28.1). For repetitive elements analysis, trimmed reads were mapped using STAR (version 2.7.6a), mapped reads to genes and TEs were counted using TEtranscripts (version 2.1.4). For ATAC-seq analysis, reads were trimmed using trimmomatic (version 0.38) and aligned with bowtie2. Peaks were called using macs2 (version 2.1.2.20181002). The transcription factor binding site enrichment analysis was done using MEME suite (version 5.0.5).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:
- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All scRNAseq data are available at the ArrayExpress accession E-MTAB-8869.
Single cell embryo RNAseq data are available at the ArrayExpress accession E-MTAB-9940.

Previously published datasets re-analysed here are available under accession codes GSE75751 and GSE66390 (ATAC-seq) ; E-MTAB-2684 and GSE66390 (RNA-seq).

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample size was chosen in order to ensure that the data was consistent and reproducible. See Figures and Figure legends for each experiment.
Data exclusions	Data in single cell and embryo RNAseq that did not pass quality control were excluded. The criteria for exclusion in quality control was pre-established as follows. For single cell RNA-seq, to remove barcodes corresponding to empty droplets, a lower threshold of 1000 UMI counts per barcode was considered. Afterwards, cells having more than 10% counts mapped to mitochondrial genes or less than 1,000 detected genes were removed. For single embryo RNA-seq, applied quality control thresholds were 30,000 reads as minimum for library size, 5000 genes as minimum for number of expressed genes and 20% as maximum for proportion of mitochondrial genes.
Replication	All data was replicated at least twice and the total replicate number is indicated in the respective panel. All attempts at replication were successful as reported in the manuscript with the exception of Figure 5g. For this experiment, four independent experiments were performed (as indicated in the panel). In one of the replicates, injection of RARE::GFP construct in DMSO condition did not present GFP+ embryos. This is represented in the Figure.
Randomization	2C blastomere to be injected was selected randomly and embryos were allocated at random to experimental groups as stated in the Methods.
Blinding	No experiment presented a subjective data collection that would require blinding. Experimentors were not blinded during experimental group allocation, embryos were divided randomly between groups.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study	
☐	☒	Antibodies
☐	☒	Eukaryotic cell lines
☒	☐	Palaeontology
☐	☒	Animals and other organisms
☒	☐	Human research participants
☒	☐	Clinical data

Methods

n/a	Involved in the study	
☒	☐	ChIP-seq
☐	☒	Flow cytometry
☒	☐	MRI-based neuroimaging

Antibodies

Antibodies used	Antibodies used were as follows: (dilutions): anti-turboGFP (TA140041, Origene), ZSCAN4 (AB4340, EMD Millipore)(1:1000),

	CRABP2 (TA349827, Origene)(1:300). Secondary antibodies used were: A-11029, A32732, A32731. Dilutions: 1:1000 for cells, 1:500 for embryos.
Validation	Anti-turboGFP antibody was validated by FACS using ES WT cell line (Supplementary Figure 7c). Anti-ZSCAN4 antibody was validated using a Zscan4c::tdTomato reporter cell line in Rodriguez-Terrones, D., Nat Genet 50, 106–119 (2018). Anti-CRABP2 was validated by the manufacturer (https://www.origene.com/catalog/antibodies/primary-antibodies/ta349827/crabp2-rabbit-polyclonal-antibody).

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	The 2C::tdTomato and 2C::turboGFP/Zscan4::mCherry cell lines were previously described (Ishiuchi, T. et al., Nat. Struct. Mol. Biol. 2015; Rodriguez-Terrones, D. et al., Nat. Genet. 2018). To generate 2C::turboGFP reporter cell line, ES cells were transfected with a plasmid containing a destabilized NLS-tagged turboGFP cassette under the regulation of Mervl LTR using Lipofecramine 2000. A single clone was selected from successfully transfected cells and has been fully characterized elsewhere (Nakatani et al., submitted).
Authentication	The 2C::tdTomato and 2C::turboGFP/Zscan4::mCherry cell lines were characterized in Ishiuchi, T. et al., Nat. Struct. Mol. Biol. 2015; Rodriguez-Terrones, D. et al., Nat. Genet. 2018). 2C::turboGFP reporter cell line has been also characterized (Nakatani et al., submitted).
Mycoplasma contamination	All cell lines tested negative for mycoplasma contamination.
Commonly misidentified lines (See ICLAC register)	No commercially misidentified cell lines were used.

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	Preimplantation mouse embryos were collected from 5-7 week old F1 (C57BL/6J x CBA/H) superovulated females crossed with F1 males (3-6 months old). Superovulation was induced by intraperitoneal injection of pregnant mare serum gonadotropin (PMSG, Intervet, 5 IU) and human chorionic gonadotropin (hCG, Intervet, 7.5 IU) 46-48 hours later.
Wild animals	This study did not use wild animals.
Field-collected samples	This study did not involve field-collected samples.
Ethics oversight	All experiments were approved by and performed under the compliance of the Government of Upper Bavaria.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Flow Cytometry

Plots

Confirm that:

☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☒ All plots are contour plots with outliers or pseudocolor plots.

☒ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	Mouse ES cells were washed with PBS, trypsinized and resuspendedin 3% BSA PBS.
Instrument	FACS Aria IIIu
Software	FlowJo v10
Cell population abundance	Whenever cell numbers were not an issue, fluorescence was verified after sorting and was usually 95 - 100%. Downstream experiments always confirmed a very high degree of sorting purity.
Gating strategy	Stringent gatings were always used, leaving a significant gap in between negative/positive populations.

☒ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.

nature portfolio

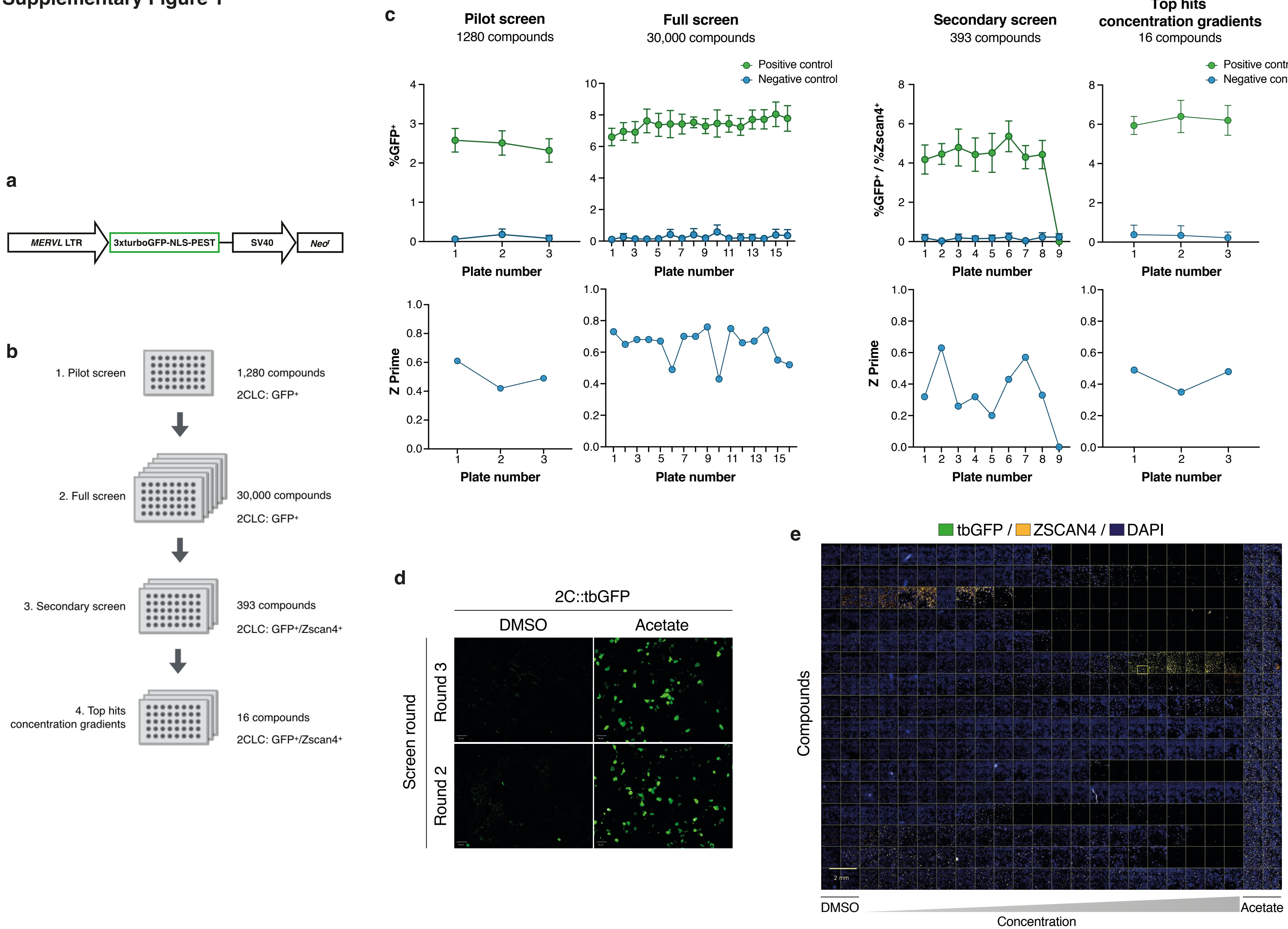
https://doi.org/10.1038/s41594-021-00590-w

Supplementary information

Retinoic acid signaling is critical during the totipotency window in early mammalian development

In the format provided by the authors and unedited

Supplementary Figure 1



Supplementary Figure 1. Small molecule screen identifies retinoids as inducers of 2CLCs

a. Schematic representation of the 2C::reporter.

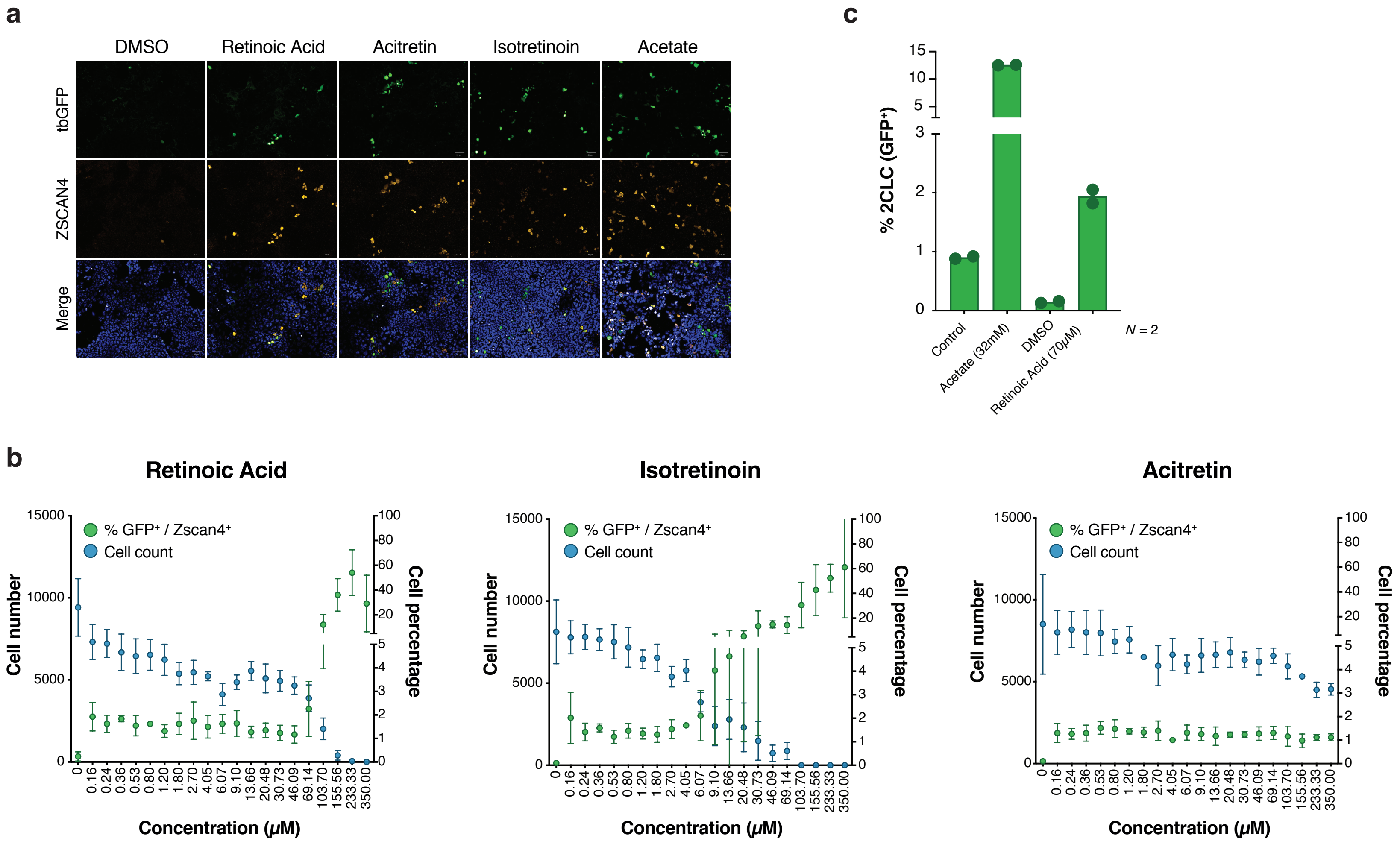
b. Design of the small molecule screen and the validation assays.

c. (Top) Quantification of GFP$^+$ and GFP$^+$/Zscan4$^+$ cells in control conditions (DMSO as negative control and Acetate as positive control) for plates across the pilot screen, one typical screening round with 30.000 diversity library compounds, secondary screen and final screen with top hits. Mean values ± s.d. from all control wells of a plate are shown. (Bottom) Z Prime values were calculated for each screening plate based on GFP$^+$ and GFP$^+$/Zscan4$^+$ quantification in control conditions (DMSO and Acetate).

d. Representative images of positive and negative controls in two different rounds of the full screen after immunofluorescence using a tbGFP antibody. Each plate of the screen counted with 32 wells for negative and positive controls.

e. Representative image of the final round of the screen (comprising 16 top hit compounds) after immunofluorescence with the indicated antibodies.

Supplementary Figure 2



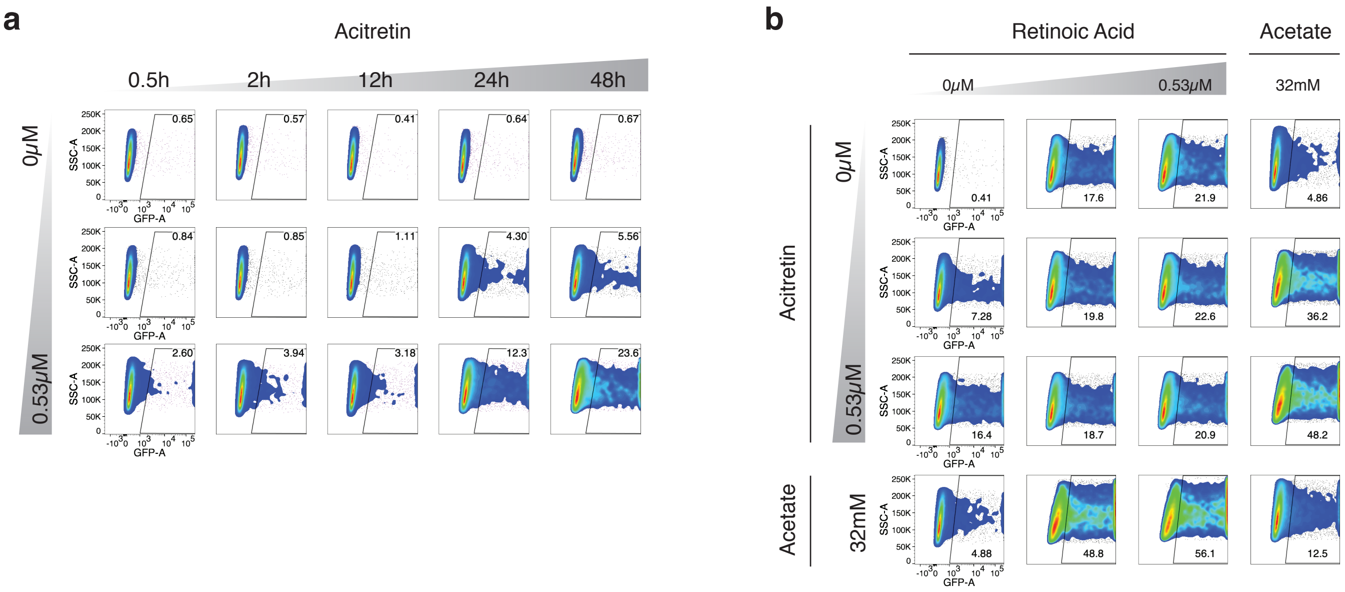
Supplementary Figure 2. Quality control of the small molecule screen identifying retinoids as 2CLC inducers

a. Representative images of ES cells with the retinoids identified as hits and the negative (DMSO) and positive (Acetate) controls from the last round of the screen (n=3 plate replicates). Scale bar, 50µm.

b. Quantification of 2CLCs, identified as double positive for GFP and ZSCAN4 ($GFP^+/Zscan4^+$), induced upon treatment with retinoids in a range of concentrations from the last round of the screen. Mean values ± s.d. from triplicate wells are shown. Total cell number is represented in blue.

c. Induction of 2CLCs (GFP^+) upon RA treatment measured by FACS. Control (no treatment), DMSO (RA vehicle) and Acetate (positive control) are shown. Mean of 2 replicates are shown. Each dot corresponds to the measurement of one replicate.

Supplementary Figure 3

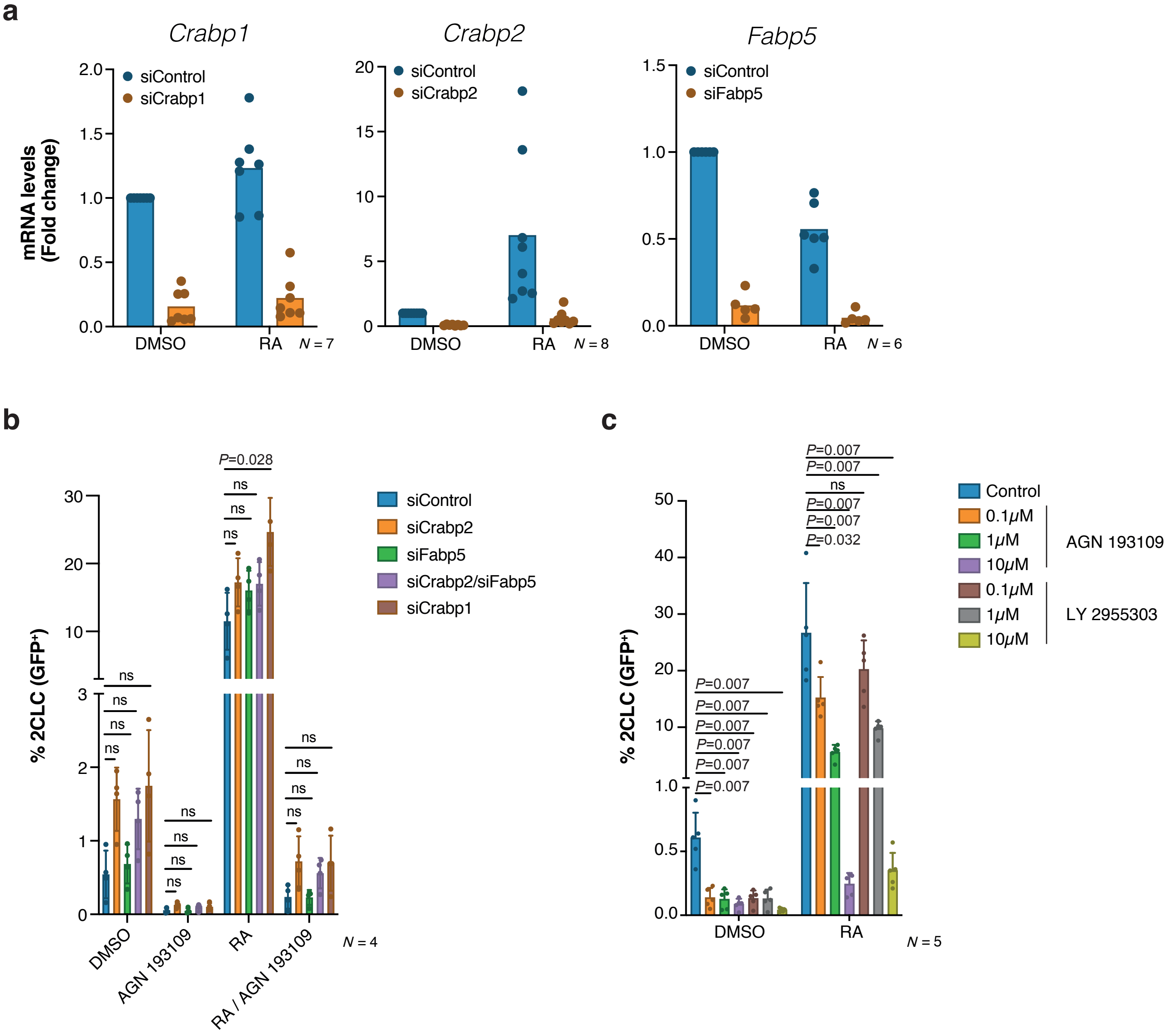


Supplementary Figure 3. Analysis of 2CLC induction by different retinoids and synergistic effects with Acetate

a. Representative scatter plots for experiment in Fig. 2a with Acitretin showing *2C::tbGFP* fluorescent measurements of individual cells as assayed by FACS.

b. Representative scatter plots for experiment shown in Fig. 2e showing *2C::tbGFP* fluorescent measurements of individual cells as assayed by FACS.

Supplementary Figure 4



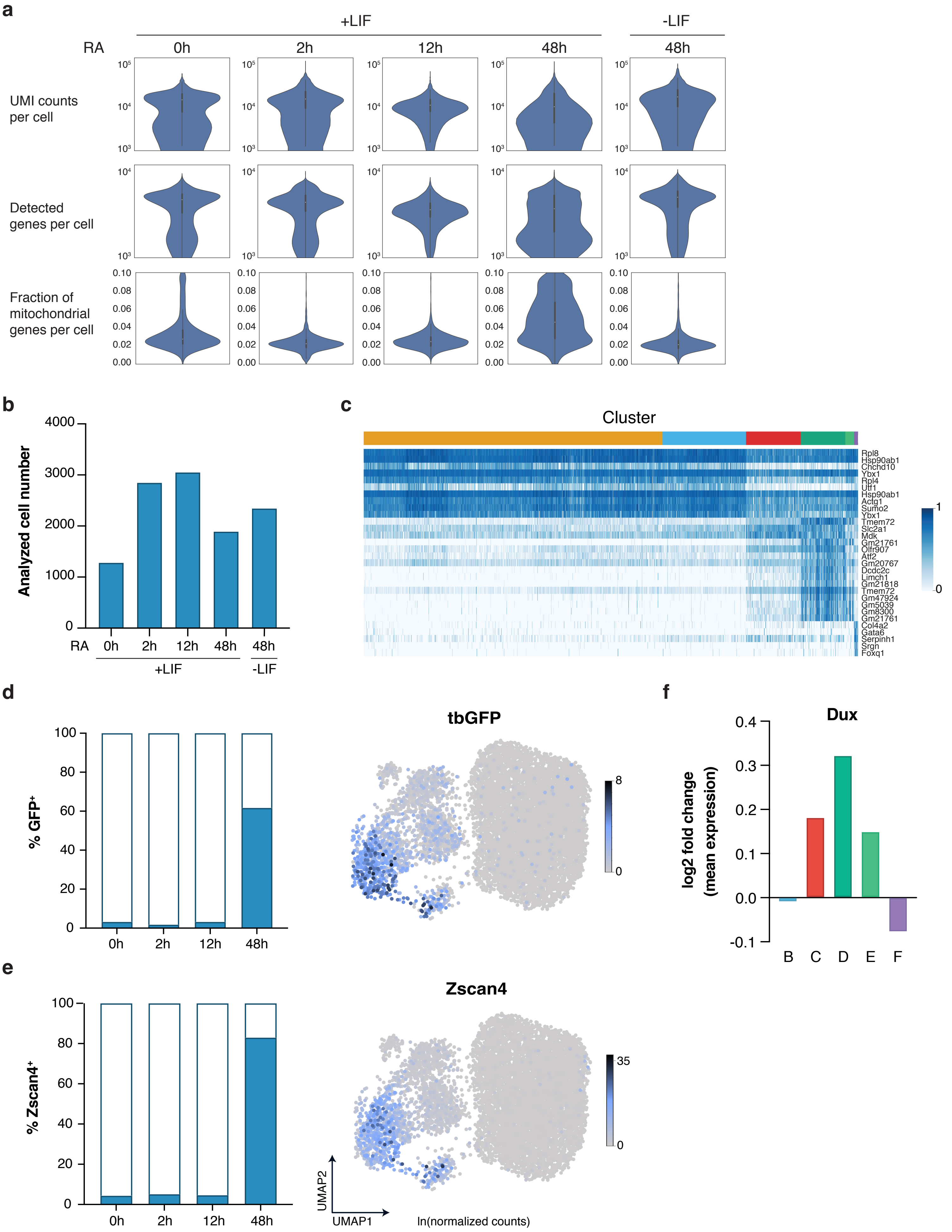
Supplementary Figure 4. Effect of perturbing RA pathway on 2CLC reprogramming.

a. RT-qPCR analysis of the indicated transcripts after siRNA for Crabp1, Crabp2 and Fabp5. Mean of the indicated replicates (represented by individual dots) is shown.

b. Induction of 2CLCs upon siRNA for Crabp1, Crabp2 and Fabp5 and RA and/or RAR antagonist AGN 193109 treatment. Scramble siRNA was used as control. Percentage of 2CLCs (GFP$^+$) quantified by FACS 48 hours after treatment. Mean ± s.d. of the indicated number of replicates is shown. *P*-value by paired two-sided Student's *t* test.

c. Induction of 2CLCs upon treatment with the RARγ antagonist LY2955303 or the RAR antagonist AGN193109. Percentage of 2CLCs (GFP$^+$) quantified by FACS 48 hours after treatment. Mean ± s.d. of the indicated number of replicates is shown. *P*-value by two-sided Mann-Whitney test.

Supplementary Figure 5



Supplementary Figure 5. Parameters of scRNA-seq analysis and quality controls.

a. The distribution of UMI counts, number of detected genes, and fraction of reads mapped to mitochondrial genes (rows from top to bottom) are shown with violin plots for cells in each condition (columns) after quality filtering.

b. Number of cells in each condition that passed quality control and were used for downstream analyses.

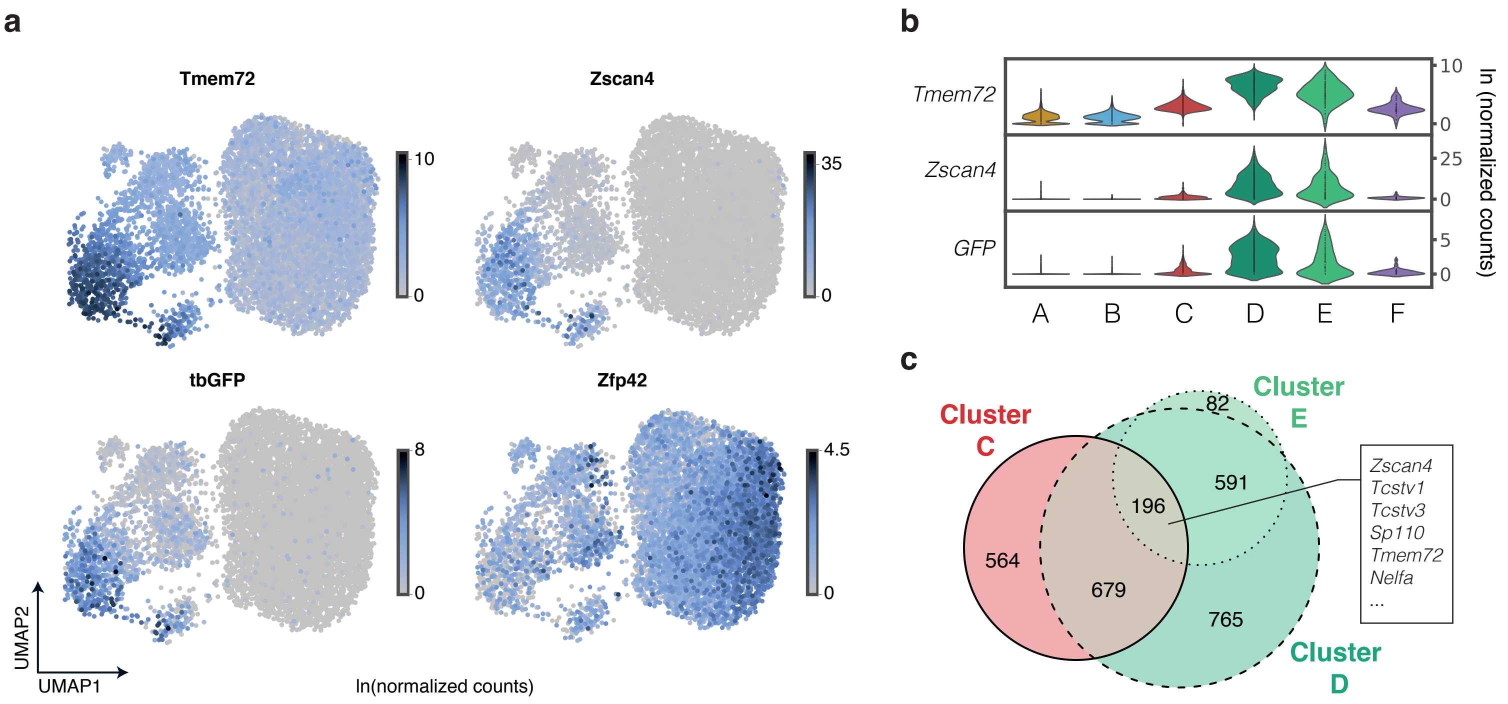
c. Heatmap illustrating the expression levels (normalized by the maximum) of the top 5 marker genes for each of the six clusters, indicated by the color bar at the top.

d. Percentage of cells where GFP is detected (left) and corresponding UMAP with cells colored by GFP expression (right) for cells grown in LIF and treated with RA during different times.

e. Percentage of cells where Zscan4 is detected (left) and corresponding UMAP with cells colored by Zscan4 expression (right) for cells grown in LIF and treated with RA during different times.

f. Bar plot showing the log2 fold change expression levels of Dux in different clusters. Fold change was calculated by dividing the mean of the ln(normalized counts) in each cluster by the mean of the ln(normalized counts) in cluster A.

Supplementary Figure 6



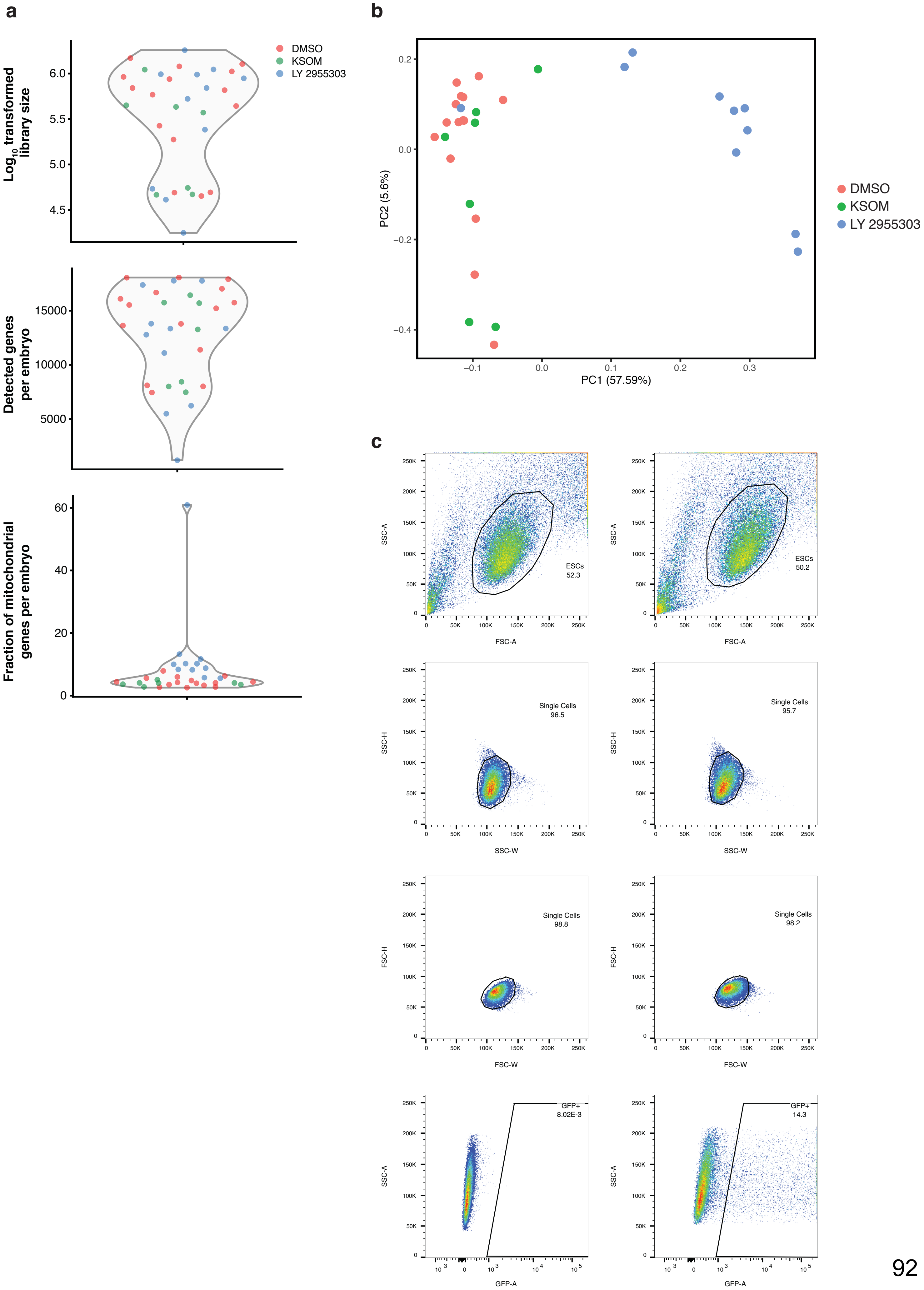
Supplementary Figure 6. Example of identification of new 2CLC markers

a. UMAP of all cells grown in LIF and Retinoic Acid, colored by expression levels of Tmem72 (top left), Zscan4 (top right); GFP (bottom left) and Zfp42/Rex1 (bottom right).

b. Violin plots showing the distribution of expression of 2CLCs marker genes Tmem72, Zscan4 and GFP (rows from top to bottom) per cluster (columns).

c. Venn diagram comparing up-regulated genes (compared to cluster A) in Cluster C, Cluster D and Cluster E.

Supplementary Figure 7



Supplementary Figure 7. Quality control of singe embryo RNA-seq analysis

a. Log10 transformed library size, number of detected genes per embryo, and fraction of reads mapped to mitochondrial genes per embryo are shown. Colour code corresponds to each experimental condition and each dot represents one embryo.

b. PCA of the single-embryo top HVG expression dataset. Each point corresponds to a single embryo, which is colored according to the experimental group it corresponds to.

c. Example of gating strategy used to quantify GFP$^+$ cells in FACS experiments. Left and right column corresponding to ES WT and 2C::tbGFP cells respectively.

Discussion

In the work described in this dissertation, we showed the importance of the space and time context of transcriptomic data for its analysis and further biological interpretation. Therefore, my work also highlights the urge for transcriptomic profiling methods that allow having spatial and time information at high resolution, as well as for the development of computational pipelines for the analysis of the resulting data.
Placing transcriptomes along time and spatial axes allowed us to address fundamental biological questions in very different contexts, such as the olfactory system and embryonic stem cells.
In the next paragraphs, I will briefly summarize our main results and discuss our contributions to understanding the relevance of OSN subtype location for the sense of smell (Chapter I), and the onset of a totipotent-like cell state in embryonic stem cell cultures as a consequence of treatment with low doses of retinoic acid (Chapter II).

I. A 3D transcriptomics atlas of the mouse nose sheds light on the anatomical logic of smell

I.I Interrogating our 3D gene expression atlas of the olfactory mucosa.

The existence of stereotypic spatial gene expression patterns in the olfactory mucosa (OM) has been known for some time. However, limited information about spatial gene expression patterns is available, and their function is unclear as well. The most familiar example of such patterns in this tissue is provided by olfactory receptor genes (*Olfrs*), of which less than 10% had been characterised. As a consequence, previous analyses about the logic of the peripheral representation of smell often remained inconclusive. In the first chapter of this dissertation, we used spatial transcriptomics (with the TOMO-seq protocol) to create the most complete transcriptional characterisation of the olfactory mucosa to date. We reconstructed the spatial expression pattern of > 17000 genes in the OM across the 3D space. Then, by combining our spatial transcriptomics computational analysis pipeline with machine learning, we went beyond the technical limitations of the experimental protocol and provided a spatial reconstruction of ~98% of *Olfr* genes' expression patterns**.** We built a user-friendly web app for our OM transcriptional atlas and made it publicly available at [http://atlas3dnose.helmholtz-muenchen.de:3838/atlas3Dnose]**,** which will facilitate the study of diverse biological processes occurring in the OM along with their transcriptional and spatial context.

Our 3D transcriptional atlas allowed us to make a new unbiased definition of *Olfrs*' spatial expression zones through a mathematically rigorous procedure that was able to capture two key features: the definition of a discrete number of zones; and the expression of *Olfr* genes in multiple zones. To represent the continuous distribution of *Olfr* across the zones, we also defined a continuous "3D index" for each *Olfr* that can describe their spatial distribution from the Dorsal/Medial/Anterior to Ventral/Lateral/Posterior region of the OM. For the *Olfr* genes previously characterised, our predictions on their spatial expression patterns and the values of their 3D indexes matched those found with orthogonal experimental techniques (e.g., in situ hybridization).

Apart from odourant detection, other biological processes take place in the OM, like toxin detection [61] and neurotransmitter metabolism [62] . However, in most cases, we do not know whether they take place at specific locations in the OM, and the identity of the genes involved. Our topic modelling approach allowed us to easily identify the genes showing the most localised expression patterns: first, we projected all genes onto the topics defined from *Olfr* genes; then, we computed the entropy associated with their expression patterns, by which we found highly localised genes. Such genes are involved in several biological processes, like neurotransmitter metabolism, which had been previously observed in the OM [62,63]. Thus, these analyses pointed out the possibility that some processes known to happen in the OM are spatially localised. Furthermore, we set a starting point for studies focused on these processes, indicating the spatial location of some related genes, which might suggest their specific roles and the pathways they are involved in.

I.II Tuning our atlas through machine learning

As ~600 *Olfrs* were not detected in our TOMO-seq data, we employed a machine learning approach to estimate their 3D index. More specifically, we trained a random forest model on 681 detected *Olfr* genes to predict the 3D index from their genomic features (e.g., distance to closest enhancer element, genomic cluster).

By doing so, we obtained the spatial expression of ~98% of the mouse olfactory receptor genes, showing the potential of genomic features as predictors of spatial gene expression trends. Furthermore, this analysis allowed the identification of the genomic features that can best predict the expression patterns of *Olfrs* (i.e., genomic position and distance to closest enhancer), which could help reveal the mechanism of *Olfr* random choice.

I.III Revisiting the Chromatographic Hypothesis

Our 3D indexes, as a simpler representation of the almost complete repertoire of *Olfrs'* spatial expression patterns, made it possible to look for associations between *Olfrs'* spatial trends and features of the ligands that *Olfrs* detect. One of the most statistically significant associations was with the air / mucus partition coefficient, defined by the ratio between the odorant concentration in the air phase and the concentration in mucus at the air–mucus interface [64]. This coefficient can be a way of quantifying the affinity of ligands to mucus relative to air. Overall, this result suggests that olfactory receptors are distributed through the OM according to their ligands' affinity to mucus, which affects the ability of those ligands to diffuse in the OM.

In this way, we provided quantitative evidence for the chromatographic hypothesis, i.e., the hypothesis that the *Olfrs* spatial distribution mirrors the patterns of diffusion of the associated ligands in the OM, and could optimize the ability to detect and distinguish different odourants.

While we created and used the largest database of *Olfr* - ligand pairs (including 153 *Olfr* genes and 221 odourants), the collection of additional data will be important to confirm our conclusions and explore further associations between *Olfr* positions in the OM and the physico-chemical properties of their ligands.

I.IV scRNA-seq data enabled us to get cell type abundance spatial distributions.

Different cell types might be involved in localised biological processes in the OM. However, the spatial transcriptomic technique we used (TOMO-seq), while it allows the unbiased profiling of the transcriptome of the entire OM, it does not achieve single-cell resolution, which is needed to distinguish the various cell types.

To circumvent this limitation, we integrated our TOMO-seq data with previously published scRNA-seq datasets to computationally estimate changes in abundances of different cell types across the OM [53]. Neuronal abundance trends were consistent with previous observations, showing the potential of scRNA-seq and spatial (TOMO-seq) data integration. In particular, we observed an increase in the relative number of neurons towards the posterior side of the OM. On the other hand, sustentacular cells and horizontal basal cells showed opposite patterns, tending to be more abundant in the anterior region. Our estimated cell type abundance trends could be potentially useful for cell communication analysis in the OM and for the study of localised biological processes in this tissue.

I.V Perspectives

Our study sets a basis for future research on the sense of smell and other biological processes taking place in the OM.

There is several follow-up computational work that could be done on our data. For example, as the importance of coding and regulatory DNA sequences in gene expression estimation is pointed out in some studies [40,41], the inclusion of DNA sequences might improve the prediction of spatial patterns of undetected *Olfrs*. Algorithms based on Convolutional Neural Networks can, for instance, include both DNA sequences and genomic features as predictors, as it was shown in [41].

Another interesting future direction is the extension of the analysis of *Olfr*-ligand pairs. The list we compiled from the literature includes odourant libraries of different sizes and compositions and tested using different experimental approaches. Moreover, highly abundant *Olfrs* have a higher probability of being deorphanised, and ecologically relevant odorants are more likely to activate *Olfrs* when compared with other odorants [65–67]. These elements might have determined a bias in our analysis towards highly abundant *Olfrs* detecting ecologically relevant ligands. So, it will be important in the future to repeat the analysis on the association between odourants' physicochemical descriptors and *Olfr* spatial patterns using a larger number of *Olfr* - ligand pairs. As a possibility to extend the available knowledge on *Olfr* - ligand pairs, machine learning approaches based on chemical features of odourants have been used to predict interactions between odourants and olfactory receptors and shown to be effective [68]. *Olfr* - ligand pairs resulting from such predictions could be used in the near future to revisit associations between *Olfrs* spatial patterns and ligands' properties.

Extending the list of *Olfr* - ligand pairs would make it possible to carry out further association studies on olfaction. An example of this would be investigating the relationship between the behaviours elicited by odours and the spatial distribution of the associated receptors. Some studies about genetic variation in olfactory receptors have shown links between specific odourant receptors' activity and behavioural responses to certain odours; it was observed that some single nucleotide polymorphisms altered *Olfr* genes' activity, causing different behavioural responses to odours [69]. Data about *Olfr* - odour correspondence together with odour-specific behavioural responses and *Olfr* genes' expression data might allow us to get this sort of insight as well, without the need for genetic variants. Having this information would allow us to test directly for associations between quantitative behavioural data and *Olfr* genes' expression.

Finally, it might be possible to implement a computational method for the distinction between different sources of gene expression spatial variation. Indeed, transcriptional differences across a tissue can occur for two reasons: Either because the transcriptome of cells of the same type varies depending on cells' location in the tissue; or because cell type composition changes across space. Thus, a method to distinguish these two kinds of transcriptional trends would be a good addition to current spatial transcriptomics data analysis pipelines. Our spatial cell type deconvolution analysis assumes that most of the genes in each cell type do not change their average expression levels as a function of position; however, a priori, we can not exclude this possibility.

II. Retinoic acid signaling is critical during the totipotency window in early mammalian development: Insights from a time-course single-cell transcriptomic profiling

II.I Low doses of retinoic acid on embryonic stem cell cultures affect cell state transitions of mouse embryonic stem cells

Retinoic acid (RA) is a known differentiation inducer [70]; however, we found that a low dose treatment of this agent can induce 2CLCs in mESC. To understand how this cell identity change occurs, we obtained and analysed a time-course single-cell RNA - seq dataset from mESCs at different times following RA treatment.
Unsupervised clustering with the Leiden algorithm of transcriptional profiles across all time points revealed six clusters: two comprising pluripotent ES cells; a cluster of ‘RA-responsive’ cells exclusively present in the 48 h RA treatment, which express low levels of 2CLC markers; two clusters that corresponded to 2CLCs, as indicated by the expression of both MERVL and *Zscan4* ; and one cluster of cells expressing early differentiation markers, which could be mostly found at the last time point only [57].

One of the 2CLCs clusters was composed mostly by cells which were profiled early after treatment (endogenous 2CLCs); whereas the other one was mostly populated by cells sequenced 48 hours after treatment (induced 2CLCs). These two clusters showed a highly significant overlap of marker genes, which pointed out new 2CLC markers and indicates that RA-induced 2CLCs overall share the transcriptional profile of endogenous 2CLCs. Future work will be needed to identify culture conditions that stably maintain 2CLCs in culture.

II.II Other cell identities induced by a low-dose retinoic acid treatment

The RA treatment also leads to the differentiation of a small group of cells. Importantly, the 2CLCs and the differentiating cluster do not share expression patterns and are clearly distinguishable from each other. To get insights into the different pathways that induce a 2CLC or differentiation, we identified and analysed the transcriptional trajectory joining mESCs and these two alternative cell identities using RNA velocity analysis [71] and the trajectory analysis pipeline implemented in [72] using the Slingshot and Tradeseq R libraries [72,73]. We found that a feature that distinguishes 2CLCs from differentiating cells is the expression of some RA-signalling components, such as Rxra. This hinted at possible different responses to RA, each triggering a different cell state trajectory. A model where different responses to RA trigger alternative trajectories also matched the RNA velocity analysis results, which depicted two different cell state trajectories: one towards 2CLCs and one towards differentiating cells. And consistent with this, further analysis revealed that different genes become activated during each transition.

Our work helped characterise RA-induced 2CLCs as a model to study the early embryonic development, identifying the retinoic acid signalling pathway as a key component of the 2CLC identity and regulator of 2CLCs reprogramming. Furthermore, our data provide a basis for understanding ES cells' different responses upon RA stimulation. The ability of ES cells to activate different sets of genes in response to RA needs further investigation to identify the specific molecular pathways involved and their roles in the onset of distinct fates. A start towards the inference of gene regulatory networks involved in the onset of cell types could be done based on the SCENIC R pipeline, which identifies sets of genes that are coexpressed with TFs in each cell and then infers stable cell states based on co-expression module activity [74].

It is also worth noticing that the Leiden algorithm was able to identify early 2CLCs as a separate cluster only once we put the data from all time points together. Then the fraction of cells of this type was big enough to be identified. Identifying and characterising rare cells has a great relevance in many biological contexts. Here for example, it could reveal the time point when different cell types start to emerge. Although scRNA-seq gives us the possibility of identifying novel rare cell types, these cells sometimes share markers with other more abundant cell types, which makes it hard for standard clustering methods to distinguish them. Many algorithms have been specifically designed for identifying rare cells, which perform well when these cells have strong markers [75,76]. However their efficiency decreases when the targeted cell population is very small (<1%) and has just a few unique markers. In order to assess this problem, we developed an algorithm called CIARA (Cluster-Independent Algorithm for the identification of markers of RAre cell types), which identifies genes that are highly expressed in small groups of cells with similar transcriptomic signatures as potential marker genes for rare cell types [77].

As mentioned before, we showed that a group of reprogrammed mESCs into 2CLCs and a group of differentiating cells were detectable after a 48 hours treatment of low doses of RA. However it is unclear how long the RA treatment should be to cause effects on cell fate decisions [77]. Hence, a new scRNA-seq dataset of mESCs after a 24 hours treatment was generated and analysed with CIARA for cell type composition. Apart from a group of mESCs and a group of 2CLCs, the marker genes selected by CIARA helped find a small group of cells characterised by the expression of differentiation markers such as *Gata4* and *Gata6*. This indicates that these cell types are present at this time point, but the change in their proportions in the population comes after 24 hours. Thus here we also show the potential of CIARA for the detection of small changes in cell type composition.

Closing remarks

Overall, during my PhD I characterised cell identity changes in space and time using transcriptome data from different sequencing techniques. My work highlights the importance of combining information across different modalities and scales (e.g., spatial transcriptomic patterns across whole tissue, single-cell transcriptomes, genomic positions of olfactory receptors and their enhancers) and the potential of interrogating them via machine learning methods to get further biological insights.

Here, I also pointed out some limitations, potential ways to address them, and the possibility of extending some of the analyses done in this dissertation. The data acquired in the chapters of this thesis has been made freely accessible to the community at http://atlas3dnose.helmholtz-muenchen.de:3838/atlas3Dnose and the Array Express database [[78]].

References

1. Morris, S. A. The evolving concept of cell identity in the single cell era. *Development* **146**, dev169748 (2019).
2. Kimmel, J. C. *et al.* Murine single-cell RNA-seq reveals cell-identity- and tissue-specific trajectories of aging. *Genome Res.* **29**, 2088–2103 (2019).
3. Elmentaite, R., Domínguez Conde, C., Yang, L. & Teichmann, S. A. Single-cell atlases: shared and tissue-specific cell types across human organs. *Nat. Rev. Genet.* **23**, 395–410 (2022).
4. Trapnell, C. Defining cell types and states with single-cell genomics. *Genome Res.* **25**, 1491–1498 (2015).
5. Savulescu, A. F., Jacobs, C., Negishi, Y., Davignon, L. & Mhlanga, M. M. Pinpointing Cell Identity in Time and Space. *Front. Mol. Biosci.* **7**, 209 (2020).
6. Nusslein-Volhard, C. Determination of the embryonic axes of Drosophila. 12.
7. Adil, A., Kumar, V., Jan, A. T. & Asger, M. Single-Cell Transcriptomics: Current Methods and Challenges in Data Acquisition and Analysis. *Front. Neurosci.* **15**, 591122 (2021).
8. Murry, C. E. & Keller, G. Differentiation of Embryonic Stem Cells to Clinically Relevant Populations: Lessons from Embryonic Development. *Cell* **132**, 661–680 (2008).
9. Miyamichi, K. Continuous and Overlapping Expression Domains of Odorant Receptor Genes in the Olfactory Epithelium Determine the Dorsal/Ventral Positioning of Glomeruli in the Olfactory Bulb. *J. Neurosci.* **25**, 3586–3592 (2005).
10. Rao, A., Barkley, D., França, G. S. & Yanai, I. Exploring tissue architecture using spatial transcriptomics. *Nature* **596**, 211–220 (2021).
11. Junker, J. P. *et al.* Genome-wide RNA Tomography in the Zebrafish Embryo. *Cell* **159**, 662–675 (2014).
12. Medaglia, C. *et al.* Spatial reconstruction of immune niches by combining photoactivatable reporters and scRNA-seq. *Science* **358**, 1622–1626 (2017).

13. Haimovich, G. & Gerst, J. Single-molecule Fluorescence in situ Hybridization (smFISH) for RNA Detection in Adherent Animal Cells. *BIO-Protoc.* **8**, (2018).
14. Nichterwitz, S. *et al.* Laser capture microscopy coupled with Smart-seq2 for precise spatial transcriptomic profiling. *Nat. Commun.* **7**, 12139 (2016).
15. Asp, M., Bergenstråhle, J. & Lundeberg, J. Spatially Resolved Transcriptomes—Next Generation Tools for Tissue Exploration. *BioEssays* **42**, 1900221 (2020).
16. Alon, S. *et al.* Expansion sequencing: Spatially precise in situ transcriptomics in intact biological systems. *Science* **371**, eaax2656 (2021).
17. Eng, C.-H. L. *et al.* Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+. *Nature* **568**, 235–239 (2019).
18. Moffitt, J. R. *et al.* Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region. *Science* **362**, eaau5324 (2018).
19. Wang, X. *et al.* Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science* **361**, eaat5691 (2018).
20. Ståhl, P. L. *et al.* Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science* **353**, 78–82 (2016).
21. Marshall, J. L. *et al.* High-resolution Slide-seqV2 spatial transcriptomics enables discovery of disease-specific cell neighborhoods and pathways. *iScience* **25**, 104097 (2022).
22. Wang, I.-H. *et al.* Spatial transcriptomic reconstruction of the mouse olfactory glomerular map suggests principles of odor processing. *Nat. Neurosci.* **25**, 484–492 (2022).
23. Sakano, H. Developmental regulation of olfactory circuit formation in mice. *Dev. Growth Differ.* **62**, 199–213 (2020).
24. Bashkirova, E. *et al. Homeotic Regulation of Olfactory Receptor Choice via NFI-dependent Heterochromatic Silencing and Genomic Compartmentalization*. http://biorxiv.org/lookup/doi/10.1101/2020.08.30.274035 (2020) doi:10.1101/2020.08.30.274035.

25. Ressler, K. J., Sullivan, S. L. & Buck, L. B. A zonal organization of odorant receptor gene expression in the olfactory epithelium. *Cell* **73**, 597–609 (1993).
26. Tan, L. & Xie, X. S. A Near-Complete Spatial Map of Olfactory Receptors in the Mouse Main Olfactory Epithelium. *Chem. Senses* (2018) doi:10.1093/chemse/bjy030.
27. Buck, L. & Axel, R. A Novel Multigene Family May Encode Odorant Receptors: A Molecular Basis for Odor Recognition. 13.
28. Mozell, M. M. & Jagodowicz, M. Chromatographic Separation of Odorants by the Nose: Retention Times Measured across in vivo Olfactory Mucosa. *Science* **181**, 1247–1249 (1973).
29. Schoenfeld, T. A. & Cleland, T. A. Anatomical Contributions to Odorant Sampling and Representation in Rodents: Zoning in on Sniffing Behavior. *Chem. Senses* **31**, 131–144 (2006).
30. Zapiec, B. & Mombaerts, P. The Zonal Organization of Odorant Receptor Gene Choice in the Main Olfactory Epithelium of the Mouse. *Cell Rep.* **30**, 4220-4234.e5 (2020).
31. Saraiva, L. R. *et al.* Hierarchical deconstruction of mouse olfactory sensory neurons: from whole mucosa to single-cell RNA-seq. *Sci. Rep.* **5**, 18178 (2015).
32. Kalehbasti, P. R., Ushijima-Mwesigwa, H., Mandal, A. & Ghosh, I. Ising-Based Louvain Method: Clustering Large Graphs with Specialized Hardware. in vol. 12695 350–361 (2021).
33. Traag, V. A., Waltman, L. & van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. *Sci. Rep.* **9**, 5233 (2019).
34. Gil-Garcia, R. J., Badia-Contelles, J. M. & Pons-Porrata, A. A General Framework for Agglomerative Hierarchical Clustering Algorithms. in *18th International Conference on Pattern Recognition (ICPR'06)* 569–572 (IEEE, 2006). doi:10.1109/ICPR.2006.69.
35. Blei, D. M. Latent Dirichlet Allocation. 30.
36. Haghverdi, L., Buettner, F. & Theis, F. J. Diffusion maps for high-dimensional single-cell analysis of differentiation data. *Bioinformatics* **31**, 2989–2998 (2015).

37. Wold, S., Esbensen, K. & Geladi, P. Principal Component Analysis. 16.

38. Laurens van der Maaten & Geoffrey Hinton. Visualizing Data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579−2605 (2008).

39. Haghverdi, L., Büttner, M., Wolf, F. A., Buettner, F. & Theis, F. J. Diffusion pseudotime robustly reconstructs lineage branching. *Nat. Methods* **13**, 845–848 (2016).

40. Avsec, Ž. *et al.* Effective gene expression prediction from sequence by integrating long-range interactions. *Nat. Methods* **18**, 1196–1203 (2021).

41. Zrimec, J. *et al.* Deep learning suggests that gene expression is encoded in all parts of a co-evolving interacting gene regulatory structure. *Nat. Commun.* **11**, 6141 (2020).

42. Smedley, D. *et al.* BioMart – biological queries made easy. *BMC Genomics* **10**, 22 (2009).

43. Monahan, K. *et al.* Cooperative interactions enable singular olfactory receptor expression in mouse olfactory neurons. *eLife* **6**, e28620 (2017).

44. Hwang, B., Lee, J. H. & Bang, D. Single-cell RNA sequencing technologies and bioinformatics pipelines. *Exp. Mol. Med.* **50**, 1–14 (2018).

45. Perkel, J. M. Single-cell analysis enters the multiomics age. *Nature* **595**, 614–616 (2021).

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

47. Crow, M., Paul, A., Ballouz, S., Huang, Z. J. & Gillis, J. Characterizing the replicability of cell types defined by single cell RNA-sequencing data using MetaNeighbor. *Nat. Commun.* **9**, 884 (2018).

48. Avila Cobos, F., Alquicira-Hernandez, J., Powell, J. E., Mestdagh, P. & De Preter, K. Benchmarking of cell type deconvolution pipelines for transcriptomics data. *Nat. Commun.* **11**, 5650 (2020).

49. Elosua-Bayes, M., Nieto, P., Mereu, E., Gut, I. & Heyn, H. SPOTlight: seeded NMF

regression to deconvolute spatial transcriptomics spots with single-cell transcriptomes. *Nucleic Acids Res.* **49**, e50–e50 (2021).

50. Aliee, H. & Theis, F. J. AutoGeneS: Automatic gene selection using multi-objective optimization for RNA-seq deconvolution. *Cell Syst.* **12**, 706-715.e4 (2021).
51. Zaitsev, K., Bambouskova, M., Swain, A. & Artyomov, M. N. Complete deconvolution of cellular mixtures based on linearity of transcriptional signatures. *Nat. Commun.* **10**, 2209 (2019).
52. Newman, A. M. *et al.* Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nat. Biotechnol.* **37**, 773–782 (2019).
53. Fletcher, R. B. *et al.* Deconstructing Olfactory Stem Cell Trajectories at Single Cell Resolution. *Cell Stem Cell* **20**, 817-830.e8 (2017).
54. Han, X. *et al.* Mapping human pluripotent stem cell differentiation pathways using high throughput single-cell RNA-sequencing. *Genome Biol.* **19**, 47 (2018).
55. Gulati, G. S. *et al.* Single-cell transcriptional diversity is a hallmark of developmental potential. 8 (2020).
56. Macfarlan, T. S. *et al.* Embryonic stem cell potency fluctuates with endogenous retrovirus activity. *Nature* **487**, 57–63 (2012).
57. Iturbide, A. *et al.* Retinoic acid signaling is critical during the totipotency window in early mammalian development. *Nat. Struct. Mol. Biol.* **28**, 521–532 (2021).
58. Takahashi, K. & Yamanaka, S. Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors. *Cell* **126**, 663–676 (2006).
59. Takahashi, K. *et al.* Induction of Pluripotent Stem Cells from Adult Human Fibroblasts by Defined Factors. *Cell* **131**, 861–872 (2007).
60. Rodriguez-Terrones, D. *et al.* A molecular roadmap for the emergence of early-embryonic-like cells in culture. *Nat. Genet.* **50**, 106–119 (2018).
61. Bryche, B., Baly, C. & Meunier, N. Modulation of olfactory signal detection in the olfactory epithelium: focus on the internal and external environment, and the emerging

role of the immune system. *Cell Tissue Res.* **384**, 589–605 (2021).

62. Goh, C. J., Choi, D., Park, D.-B., Kim, H. & Hahn, Y. MOXD2, a Gene Possibly Associated with Olfaction, Is Frequently Inactivated in Birds. *PLOS ONE* **11**, e0152431 (2016).
63. Mania-Farnell, B. L., Farbman, A. I. & Bruch, R. C. Bromocriptine, a dopamine d2 receptor agonist, inhibits adenylyl cyclase activity in rat olfactory epithelium. *Neuroscience* **57**, 173–180 (1993).
64. Scott, J. W., Sherrill, L., Jiang, J. & Zhao, K. Tuning to Odor Solubility and Sorption Pattern in Olfactory Epithelial Responses. *J. Neurosci.* **34**, 2025–2036 (2014).
65. Dunkel, A. *et al.* Nature's Chemical Signatures in Human Olfaction: A Foodborne Perspective for Future Biotechnology. *Angew. Chem. Int. Ed.* **53**, 7124–7143 (2014).
66. Trimmer, C. & Mainland, J. D. Simplifying the Odor Landscape. *Chem. Senses* **42**, 177–179 (2017).
67. Ruiz Tejada Segura, M. L. *et al.* A 3D transcriptomics atlas of the mouse nose sheds light on the anatomical logic of smell. *Cell Rep.* **38**, 110547 (2022).
68. Kowalewski, J. & Ray, A. Predicting Human Olfactory Perception from Activities of Odorant Receptors. *iScience* **23**, 101361 (2020).
69. Trimmer, C. *et al.* Genetic variation across the human olfactory receptor repertoire alters odor perception. *Proc. Natl. Acad. Sci.* **116**, 9475–9480 (2019).
70. Rohwedel, J., Guan, K. & Wobus, A. M. Induction of Cellular Differentiation by Retinoic Acid in vitro. *Cells Tissues Organs* **165**, 190–202 (1999).
71. La Manno, G. *et al.* RNA velocity of single cells. *Nature* **560**, 494–498 (2018).
72. Van den Berge, K. *et al.* Trajectory-based differential expression analysis for single-cell sequencing data. *Nat. Commun.* **11**, 1201 (2020).
73. Street, K. *et al.* Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics* **19**, 477 (2018).
74. Aibar, S. *et al.* SCENIC: single-cell regulatory network inference and clustering. *Nat.*

Methods **14**, 1083–1086 (2017).

75. Wegmann, R. *et al.* CellSIUS provides sensitive and specific detection of rare cell populations from complex single-cell RNA-seq data. *Genome Biol.* **20**, 142 (2019).
76. Jiang, L., Chen, H., Pinello, L. & Yuan, G.-C. GiniClust: detecting rare cell types from single-cell gene expression data with Gini index. *Genome Biol.* **17**, 144 (2016).
77. Lubatti, G. *et al. CIARA: a cluster-independent algorithm for the identification of markers of rare cell types from single-cell RNA seq data*. http://biorxiv.org/lookup/doi/10.1101/2022.08.01.501965 (2022) doi:10.1101/2022.08.01.501965.
78. Parkinson, H. *et al.* ArrayExpress--a public database of microarray experiments and gene expression profiles. *Nucleic Acids Res.* **35**, D747–D750 (2007).

Acknowledgements

I would like to dedicate this thesis to all people who took part in the many stories happening in my four years of PhD. Those stories, the awesome ones, the absurd ones, the worst ones, the surreal ones, the ones we thought could never ever happen, the ones people will remember and tell again and again, those will stay in this world at least for a few generations…

And well, let’s start from the beginning, from my family, who were with me from the beginning of this adventure. I will always be grateful for 2 sayings “El no ya lo tienes”, which I have taken sometimes a little bit too far. Never regretted it though. And “Lo importante no es no caerse, sino levantarse”, which helped me keep going for my goals, getting back with improved articles, applications, etc after each rejection. Thanks Mom and Dad for encouraging me to pursue what I want no matter how crazy it can sound. Also for making the Atlantic ocean seem small as we talked. I’ve always felt close to you and hope it will always be that way. I want to also thank my brother Rafa, my grandparents Geo and Rruizte, my aunts Gaby, Claudia and Itziar, uncles Edgar and Jaime and cousins Pris and Sam, for spontaneously reaching out for a chat and making the time difference between Mexico and Germany fade away.

Then it is time to thank my school friends, my second family. Thanks Liz and Hazel, Mafer and Marisol, for sticking with me since we were little kids, for helping me get better at explaining things even since then. Kassia, Diana and Frangy as well, my best highschool friends, for those escapes to the mall and cinema, for making taekwondo training even more fun, for showing me the streets of my own city. I might have always been considered the smart girl (that nerdy weirdo), but actually I’ve learned a lot from you about life outside of school and tests, outside of academia. I’m so glad and thankful for having friends outside of academia, thanks for letting me see the world through your eyes and sharing your way through life with me. I see you being successful in your own areas and I feel so proud of having grown together.

Now those who met me on my way to independence and freedom, my university friends, Fabian, Pavel, Turing, Diego, Elias, Vianka, thanks guys for the cool nights, the trips, the international encounters. Cheers for those experiences that made our bonds tighter. Your names come often in lots of stories I tell again and again. Thanks for the stories. And thanks for the openness, easiness, and derived conversations. You’ve contributed a lot to my current characteristic open-mindedness, changing some of my points of view, and I really appreciate it. Now

we're spread around the world doing great science and it is always awesome to hear from you. Great people behind great science.

And talking about great people behind great science, I would like to of course thank friends I met during PhD, Gabriele, Adam, Elmir, Carsten, Marco, Maria, Tamas, Allwyn, Marlies, Hiromi, to mention a few names. Thanks for the good times, sporty activities, calory recoveries, italian lessons, study nights, career talks, advice and cool collaborations. I always feel happy to see familiar names together in a nice paper. And I'd like to add a special mention to the people I met at AchemS 2022, the event that brought my enthusiasm for science back after the heavy lockdowns. The jump into the jacuzzi in the middle of the night summarizes my feeling towards this conference pretty well. Talking about great topics sometimes with a beer, or in the pool, getting to know who was behind the zoom screen and behind the latest cool ideas in the olfaction and taste fields.

And finally, I would like to thank the great minds guiding my work during my PhD: Antonio, Luis, Maria Elena and Wolfgang. I was very lucky to get to work with you in very cool projects and to get to experience your way through science. Thanks for the good advice, corrections, the discussions and talks leading to new ideas. That is something I find awesome: great minds developing new ideas together. It was fascinating to see how the ideas behind projects originate and then to participate in their consolidation. A huge thanks to Antonio for this chance. It's been a pleasure to work together.

Annex I. Curriculum Vitae

Curriculum Vitae — Mayra Luisa Ruiz Tejada Segura

Date of Birth: 12.08.1995
Place of Birth: Mexico City
Nationality: Mexican

Education

Sep. 2018 - present	**Ludwig Maximilians Universität München (LMU),** **Helmholtz Zentrum München** Ph.D. Biology (Expected: Summer 2022)
Aug. 2013 - May 2017	**Universidad Nacional Autónoma de México (UNAM)** BSc. Genomic Sciences (Graduation with Honours)

Research Experience

Sep. 2018 - present

Ph.D. Researcher
Helmholtz Zentrum München,
Munich, Germany

Project title: ***A 3D transcriptomics atlas of the mouse nose sheds light on the anatomical logic of smell***
-Design of a bioinformatic pipeline for identifying unidimensional non-random trends in spatial transcriptomics data in R.
-3D gene expression model reconstruction and 3D spatial transcriptomics atlas generation, public visualisation via R Shiny app.
-3D gene expression pattern indexing through clustering and dimensionality reduction (diffusion map, diffusion pseudotime, topic modeling).
-Undetected genes' index inference via machine learning models trained on genomic features.
-Cheminformatics analysis to link gene expression spatial trends and properties of compounds detected by the encoded proteins.

	Project title: ***Retinoic acid signalling is critical during the totipotency window in early mammalian development*** -Annotation and quality control of 10x single cell RNA sequencing data from a totipotent-like phenotype induction time course experiment. -Single cell RNA-seq data clustering and cluster characterization through marker genes in Python. -Cell type trajectory analysis and visualisation using RNA velocity and diffusion maps in Python.
Sep. - Dec. 2017	***Graduate Researcher*** Helmholtz Zentrum München Munich, Germany Funding: Boehringer Ingelheim Fonds Travel Grant -Genome annotation and quality control statistics of RNA Illumina sequencing data from cryosectioned mice olfactory mucosa (TOMO-seq). -Identification of one-dimensional spatial gene expression patterns in the main olfactory mucosa of mice
Oct. 2016 - Aug. 2017	***Undergraduate Researcher*** University of Lincoln and University of Bath Lincoln and Bath, UK Funding: Korner Award (University of Sussex), Beca Movilidad UNAM (Fundación UNAM) -Bachelor's thesis Topic: **Differential gene expression and co-expression analysis comparing transcriptomic data from a mouse model of Alzheimer’s disease and wildtype mice.**
Sep. 2015 - Oct. 2016	***Undergraduate Researcher*** Universidad Nacional Autónoma de México (UNAM) Morelos, México -Use of molecular biology techniques like DNA purification and gel electrophoresis, for the construction of fluorescent reporter plasmids -DNA transformation for the analysis of protein expression in models of neuronal development
June - Aug. 2015	***Summer Undergraduate Researcher*** University of Bath Bath, UK -Genome annotation and quality control statistics of RNA Illumina sequencing data derived from a mouse model of Alzheimer’s disease.

Publications

2022	**Ruiz Tejada Segura, M. L.**, Abou Moussa, E., Garabello, E., Nakahara, T. S., Makhlouf, M., Mathew, L. S., ... & Saraiva, L. R. (2022). A 3D transcriptomics atlas of the mouse olfactory mucosa sheds light into the anatomical logic of smell. *Cell Reports,* 38(12), 110547.
2021 K.,	Iturbide, A., **Ruiz Tejada Segura, M. L.**, Noll, C., Schorpp, Rothenaigner, I., Ruiz-Morales, E. R., ... & Torres-Padilla, M. E. (2021). Retinoic acid signalling is critical during the totipotency window in early mammalian development. *Nature Structural & Molecular Biology*, 28(6), 521-532.
2020	Yin, W., Cerda-Hernández, N., Castillo-Morales, A., **Ruiz Tejada Segura, M. L.**, Monzón-Sandoval, J., Moreno-Castilla, P., ... & Gutiérrez, H. (2020). Transcriptional, Behavioural and Biochemical Profiling in the 3xTg-AD Mouse Model Reveals a Specific Signature of Amyloid Deposition and Functional Decline in Alzheimer's Disease. *Frontiers in neuroscience*, 1322.
	Huang, S. S., Makhlouf, M., AbouMoussa, E. H., **Ruiz Tejada Segura, M. L.**, Mathew, L. S., Wang, K., ... & Saraiva, L. R. (2020). Differential regulation of the immune system in a brain-liver-fats organ network during short-term fasting. *Molecular metabolism*, 40, 101038.

Manuscripts in review

2022	Neupane, J., Lubatti, G., **Ruiz Tejada Segura, M. L.**, Alves Lopes, J. P., Dietmann, S., Scialdone, A., Surani, M. A. (Unpublished manuscript). Human embryonic organoids reveal origin of primordial germ cells and neuromesodermal progenitors
	Lubatti, G., Stock, M., Iturbide, A., **Ruiz Tejada Segura, M. L.**, Tyser, R., Theis, F., Srinivas, S., Torres-Padilla, M. E., Scialdone, A. (2022). CIARA: a cluster-independent algorithm for the identification of markers of rare cell types from single-cell RNA seq data. bioRxiv 2022.08.01.501965; doi: https://doi.org/10.1101/2022.08.01.501965.

Scholarships

Sep. - Dec. 2017	**Boehringer Ingelheim Fonds Travel Grant** Boehringer Ingelheim, Ingelheim am Rhein, Germany
Oct. 2016 - Aug. 2017	**Korner Award** University of Sussex Sussex, UK **Beca Movilidad UNAM (Fundación UNAM)** Universidad Nacional Autónoma de México (UNAM) Ciudad de México, México

Skills

Programming languages:	R, Python, Bash, Perl
Data analysis:	Bulk RNA-seq, single cell RNA-seq, spatial transcriptomics
Tools:	R Shiny, Git
Lab:	DNA purification, transformation, gel electrophoresis
Languages:	Spanish (native language) English (C1) German (Basic)

Selected talks

2020	**Deconstructing the spatial organisation of mouse olfactory epithelium by spatial transcriptomics** Wellcome Genome Campus Scientific Conferences: “Single Cell Biology 2020”. November 9th-12th, 2020, virtual due to COVID-19
2019	**Deconstructing the spatial organisation of mouse olfactory epithelium by spatial transcriptomics** The Institute of Physics: “Quantitative Methods in Gene Regulation V”. December 9th-10th, 2019, London, UK

Posters

2019	**Deconstructing the spatial organisation of mouse olfactory epithelium by spatial transcriptomics** Keystone Symposia: “L1 Single Cell Biology”. January 13th - 17th, 2019, Breckenridge, Colorado, USA

2022	**A 3D transcriptomics Atlas of the Mouse Nose sheds light into the Anatomical Logic of Smell** AChemS 2022. April 27th-30th, 2022, Bonita Springs, FL, USA

Leadership & Teaching experience

Sep. 2018 - present

Ph.D. Researcher
Helmholtz Zentrum München, LMU
Munich, Germany

- Supervised internship MSc (1) and BSc students (1)
- Assisted in teaching Single Cell RNA-seq analysis course from LMU
- Prepared course material including practice problems and presentations

Sep. - Dec. 2015

Seminar organiser
Universidad Nacional Autónoma de México (UNAM)
Morelos, México
Organisation of cycle of conferences about Epigenetics for the Undergraduate Program on Genomic Sciences at UNAM.

References

Dr. Antonio Scialdone
PhD supervisor
E-mail: antonio.scialdone@helmholtz-muenchen.de

Dr. Luis Saraiva
Main PhD project collaborator
E-Mail: lsaraiva@sidra.org

Dr. Araxi Urrutia
Bachelor's thesis supervisor
E-Mail: A.Urrutia@bath.ac.uk

Munich, 16.09.2022

Annex II. Copyright statements

RightsLink

Home | Help | Email Support | Sign In | Create Account

A 3D transcriptomics atlas of the mouse nose sheds light on the anatomical logic of smell

Author:
Mayra L. Ruiz Tejada Segura,Eman Abou Moussa,Elisa Garabello,Thiago S. Nakahara,Melanie Makhlouf,Lisa S. Mathew,Li Wang,Filippo Valle,Susie S.Y. Huang,Joel D. Mainland,Michele Caselle,Matteo Osella,Stephan Lorenz,Johannes Reisert,Darren W. Logan et al.

Publication: CELL REPORTS

Publisher: Elsevier

Date: 22 March 2022

© 2022 The Authors.

Welcome to RightsLink

Elsevier has partnered with Copyright Clearance Center's RightsLink service to offer a variety of options for reusing this content.

Note: This article is available under the Creative Commons CC-BY-NC-ND license and permits non-commercial use of the work as published, without adaptation or alteration provided the work is fully attributed.

For commercial reuse, permission must be requested below.

I would like to... ? make a selection

To request permission for a type of use not listed, please contact Elsevier Global Rights Department.

Are you the author of this Elsevier journal article?

© 2022 Copyright - All Rights Reserved | Copyright Clearance Center, Inc. | Privacy statement | Data Security and Privacy | For California Residents | Terms and ConditionsComments? We would like to hear from you. E-mail us at customercare@copyright.com

RightsLink

? Help

Email Support

SPRINGER NATURE

Retinoic acid signaling is critical during the totipotency window in early mammalian development

Author: Ane Iturbide et al
Publication: Nature Structural & Molecular Biology
Publisher: Springer Nature
Date: May 27, 2021

Copyright © 2021, The Author(s)

Creative Commons

This is an open access article distributed under the terms of the Creative Commons CC BY license, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

You are not required to obtain permission to reuse this article.
To request permission for a type of use not listed, please contact Springer Nature

© 2022 Copyright - All Rights Reserved | Copyright Clearance Center, Inc. | Privacy statement | Data Security and Privacy | For California Residents | Terms and ConditionsComments? We would like to hear from you. E-mail us at customercare@copyright.com