Technische Universität München
TUM School of Computation, Information and Technology

Characterizing cell states in early embryonic development with single-cell RNA sequencing and mathematical modelling.

Gabriele Lubatti

Vollständiger Abdruck der von der TUM School of Compuation, Information and Technology der Technischen Universität München zur Erlangung eines Doktors der Naturwissenschaften (Dr. rer. nat.) genehmigten Dissertation.

Vorsitz: Prof. Dr. Johannes Mueller

Prüfende der Dissertation:

1. Prof. Dr. Fabian J. Theis
2. Prof. Dr. Maria Colomé-Tatché
3. Prof. Dr. Bertie Göttgens

Die Dissertation wurde am 20.06.2023 bei der Technischen Universität München eingereicht und durch die TUM School of Computation, Information and Technology am 17.11.2023 angenommen

Characterizing cell states in early embryonic development with single-cell RNA sequencing and mathematical modelling

Gabriele Lubatti

June 2023

Acknowledgments

First of all, I would like to thank my advisors, Antonio Scialdone and Fabian Theis, for their supervision and support throughout my PhD.
Special thanks also to the members of Scialdone's lab, especially Mayra Ruiz, Marco Stock, Matteo Zambon, Elmir Mahammadov, Jonathan Fiorentino, Farid Ahadli, Melina Riepl, Carla Ares, Aleksei Taradin, Karoline Holler and Tobias Krauß for productive discussions, collaborations and good times together.
It was a pleasure to work with all the collaborators I had in my projects, in particular Ana Lima, Tristan Rodriguez, Marion Genet, Maria Elena Torres Padilla, Jitesh Neupane, Azim Surani.
A big thank also goes to all the colleagues and friends with whom I shared my time here in Munich, especially Yung-Li Chen, Maria Solovey, Federico Pecori, Adam Burton, Tamas Schauer, Tomas Zikmund, Luis Altamirano, Mrinmoy Pal, Frank Ziemann, Yicong Chen and Gabriele Managoli.
Finally I want to thank my family, especially my parents. They taught me the value of putting effort and working hard to get results.

Abstract

Every cell in an adult arises from a single zygote through a sequence of cell divisions and fate decisions, in which a cell makes a transition from one type or state to another. Cell states can be characterized by specific transcriptomic signatures and can lead to different functions and abilities for the cells to deal with internal and external stimuli. While such changes in cell states are crucial for the proper functioning of organisms, they can also contribute to diseases if they become dysregulated. Understanding the molecular mechanisms regulating cell state transitions is fundamental in many fields, ranging from developmental biology to cancer biology.
The advent of single-cell sequencing techniques has offered an unprecedented opportunity to explore cell state changes in great detail. However, the huge amount of data produced is complex to analyze and interpret. For example, sequencing techniques capture only a snapshot of the single-cell molecular features at a given time point, but computational modelling is needed to reconstruct the dynamics, unlike with non-destructive but lower throughput technologies such as microscopy-based approaches. Thus, to exploit single-cell sequencing data, the development of dedicated machine learning methods is needed alongside the use of mathematical models that can help interpret the results.
Developmental biology is one of the fields that have benefited the most from single-cell sequencing technologies, which have been extensively used to study cell state transitions during embryonic development. As the embryo develops, changes in cell states can mark variations in the differentiation potential of stem cells, affecting the cell types they can give rise to. This is the case of the transition between the totipotency to the pluripotency state, which is crucial for the development of an organism but is governed by molecular drivers that we still don't fully understand. During development, there are changes in cell states that can also modify the way in which cells interact with their neighbors. For example, the pluripotent epiblast cells in mouse can exist in a "winner" and a "loser" state, where the cells in the latter state are eliminated by the "winners" via cell competition. While this process might ensure that only the fittest cells will contribute to the embryo, we don't know what the molecular features that distinguish "loser" from "winner" cells are, and how the transition between these states occurs.

During my research, I focused on the totipotency/pluripotency and the "winner"/"loser" cell state transitions, which are key during embryonic development, aiming to find their molecular drivers and their function with a combination of single-cell RNA-seq data analysis and mathematical modeling.
More specifically, by examining the transcriptomes of in vivo as well as in vitro models of totipotency and pluripotency in several mammalian species, I aimed to gain a better understanding of the mechanisms that control the totipotency-pluripotency transition, and whether they are conserved among mammalian species. In particular, it has been shown that transitions from a pluripotent to a totipotent-like state rarely occur in in vitro cultures of embryonic stem cells in mouse and human. To study if such rare transitions also occur in other species, I developed a new machine-learning algorithm to identify in an unsupervised manner rare cell types in single-cell sequencing datasets, called CIARA. With this method, I have found previously uncharacterized rare cell types in human and mouse embryos, as well as in in vitro cell cultures.

I also used single-cell RNA-seq to characterize the "winner" and "loser" states of epiblast cells in mouse embryos. Using an algorithm I developed called MitoHEAR, I discovered that the transition into a lower fitness state (the "loser" state) is likely caused by mitochondrial mutations that impair mitochondrial performance. Additionally, using mathematical modelling, I showed that cellular competition might contribute to the regulation of embryo size. This study provided evidence that cellular competition in mouse embryos is a purifying selection ensuring the presence of epiblast cells with an optimal mitochondrial pool in preparation for the following steps in development.
Overall, the work that I have done during the thesis has led to the characterisation of the molecular mechanisms underlying fundamental cell state transitions occurring during embryonic development. Moreover, the computational methods I developed are very general and will be useful in studying cell state transitions in many other biological contexts.

Zusammenfassung

Jede Zelle in einem erwachsenen Organismus entsteht aus einer einzigen Zygote durch eine Abfolge von Zellteilungen und Schicksalsentscheidungen, bei denen eine Zelle von einem Typ oder Zustand zu einem anderen übergeht. Zellzustände können durch spezifische transkriptomische Signaturen charakterisiert werden und zu unterschiedlichen Funktionen und Fähigkeiten der Zellen führen, um mit internen und externen Reizen umzugehen. Während solche Veränderungen in den Zellzuständen für das ordnungsgemäße Funktionieren von Organismen entscheidend sind, können sie auch zu Krankheiten beitragen, wenn sie aus dem Gleichgewicht geraten. Das Verständnis der molekularen Mechanismen, die die Übergänge der Zellzustände regulieren, ist in vielen Bereichen von grundlegender Bedeutung, angefangen von der Entwicklungsbiologie bis zur Krebsbiologie. Das Aufkommen von Einzelzell-Sequenzierungstechniken hat eine beispiellose Möglichkeit geschaffen, Zellzustandsänderungen im Detail zu erforschen. Die große Menge an erzeugten Daten ist jedoch komplex in der Analyse und Interpretation. Zum Beispiel erfassen Sequenzierungstechniken nur einen Momentaufnahme der einzelzellulären molekularen Merkmale zu einem bestimmten Zeitpunkt, aber zur Rekonstruktion der Dynamik sind rechnergestützte Modelle erforderlich, im Gegensatz zu zerstörungsfreien, jedoch niedrigeren Durchsatztechnologien wie mikroskopiebasierten Ansätzen.
Um also die Daten der Einzelzell-Sequenzierung zu nutzen, ist die Entwicklung spezialisierter maschineller Lernmethoden erforderlich, die zusammen mit mathematischen Modellen eingesetzt werden können, um die Ergebnisse zu interpretieren. Die Entwicklungsbiologie ist eine der Disziplinen, die am meisten von Einzelzell-Sequenzierungstechnologien profitiert haben, die ausgiebig genutzt wurden, um Zellzustandsübergänge während der embryonalen Entwicklung zu untersuchen. Mit fortschreitender Entwicklung des Embryos können Veränderungen der Zellzustände Variationen im Differenzierungspotenzial von Stammzellen kennzeichnen und die Zelltypen beeinflussen, aus denen sie hervorgehen können. Dies ist zum Beispiel der Fall bei dem Übergang von der Totipotenz zur Pluripotenz, der für die Entwicklung eines Organismus entscheidend ist, aber von molekularen Treibern gesteuert wird, die wir noch nicht vollständig verstehen.
Während der Entwicklung gibt es Veränderungen der Zellzustände, die auch die Art und Weise beeinflussen können, wie Zellen mit ihren Nachbarn interagieren. Zum Beispiel können pluripotente Epiblastzellen bei Mäusen in einem "Gewinner"- und einem "Verlierer"-Zustand existieren, wobei die Zellen in letzterem Zustand durch die "Gewinner" durch zellulären Wettbewerb eliminiert werden. Während dieser Prozess sicherstellt, dass nur die am besten geeigneten Zellen zum Embryo beitragen, wissen wir nicht, welche molekularen Merkmale "Verlierer"- von "Gewinner"-Zellen unterscheiden und wie der Übergang zwischen diesen Zuständen erfolgt.
In meiner Forschung habe ich mich auf diese beiden Zellzustandsübergänge konzentriert, die während der embryonalen Entwicklung entscheidend sind, um ihre molekularen Treiber und ihre Funktion mithilfe einer Kombination aus Einzelzell-RNA-Sequenzierungsdatenanalyse und mathematischer Modellierung zu finden.
Genauer gesagt habe ich durch die Untersuchung der Transkriptome von in vivo- und in vitro-Modellen der Totipotenz und Pluripotenz in mehreren Säugetier-

arten versucht, ein besseres Verständnis der Mechanismen zu gewinnen, die den Übergang von der Totipotenz zur Pluripotenz kontrollieren. Insbesondere wurde gezeigt, dass Übergänge von einem pluripotenten zu einem totipotenten-ähnlichen Zustand selten in in vitro-Kulturen von embryonalen Stammzellen bei Mäusen und Menschen auftreten. Um zu untersuchen, ob solche seltenen Übergänge auch bei anderen Arten vorkommen, habe ich einen neuen maschinellen Lernalgorithmus entwickelt, um in einer nicht überwachten Weise seltene Zelltypen in Einzelzell-Sequenzierungsdatensätzen zu identifizieren, genannt CIARA. Mit dieser Methode habe ich zuvor uncharakterisierte seltene Zelltypen in menschlichen und mausartigen Embryos sowie in in vitro-Zellkulturen gefunden.
Ich habe auch Einzelzell-RNA-Sequenzierung verwendet, um die "Gewinner"- und "Verlierer"-Zustände von Epiblastzellen in Mäuseembryos zu charakterisieren. Mit einem von mir entwickelten Algorithmus namens MitoHEAR habe ich entdeckt, dass der Übergang in einen Zustand geringerer Fitness (der "Verlierer"-Zustand) wahrscheinlich durch mitochondriale Mutationen verursacht wird, die die mitochondriale Leistung beeinträchtigen. Darüber hinaus habe ich mit mathematischer Modellierung gezeigt, dass zellulärer Wettbewerb zur Regulation der Embryogröße beitragen könnte. Diese Studie lieferte Hinweise darauf, dass zellulärer Wettbewerb in Mäuseembryos eine reinigende Selektion ist, die das Vorhandensein von Epiblastzellen mit einem optimalen mitochondrialen Pool zur Vorbereitung auf die folgenden Entwicklungsschritte sicherstellt.
Insgesamt hat die Arbeit, die ich während der Dissertation durchgeführt habe, zur Charakterisierung der molekularen Mechanismen beigetragen, die grundlegende Zellzustandsübergänge während der embryonalen Entwicklung steuern. Darüber hinaus sind die von mir entwickelten rechnergestützten Methoden sehr allgemein und werden in vielen anderen biologischen Zusammenhängen nützlich sein.

List of contributed publications

This cumulative thesis is based on the following publications. The "*" symbol marks co-first authorship.

Core publications as main author

- Ana Lima*, **Gabriele Lubatti***, Jorg Burgstaller, Di Hu, Alistair Green, Aida Di Gregorio, Tamzin Zawadzki, Barbara Pernaute, Elmir Mahammadov, Salvador Perez Montero, Marian Dore, Juan Miguel Sanchez, Sarah Bowling, Margarida Sancho, Mohammed Karimi, David Carling, Nick Jones, Shankar Srinivas, Antonio Scialdone, Tristan A Rodriguez. Cell competition acts as a purifying selection to eliminate cells with mitochondrial defects during early mouse development. Nat Metab 3, 1091–1108 (2021). `https://doi.org/10.1038/s42255-021-00422-7`.

- **Gabriele Lubatti**, Elmir Mahammadov, Antonio Scialdone. (2022). MitoHEAR: an R package for the estimation and downstream statistical analysis of the mitochondrial DNA heteroplasmy calculated from single-cell datasets. Journal of Open Source Software, 7(74), 4265. `https://doi.org/10.21105/joss.04265`.

- **Gabriele Lubatti**, Marco Stock, Ane Iturbide, Mayra L. Ruiz Tejada Segura, Melina Riepl, Richard C. V. Tyser, Anna Danese, Maria Colomé-Tatché, Fabian J. Theis, Shankar Srinivas, Maria-Elena Torres-Padilla, Antonio Scialdone; CIARA: a cluster-independent algorithm for identifying markers of rare cell types from single-cell sequencing data. Development 1 June 2023; 150 (11): dev201264. doi: `hhttps://doi.org/10.1242/dev.201264`.

Further publications as co-author

- Ane Iturbide, Mayra L Ruiz Tejada Segura, Camille Noll, Kenji Schorpp, Ina Rothenaigner, Elias R Ruiz-Morales, **Gabriele Lubatti**, Ahmed Agami, Kamyar Hadian, Antonio Scialdone, Maria-Elena Torres-Padilla. Retinoic acid signaling is critical during the totipotency window in early mammalian development. Nat Struct Mol Biol 28, 521–532 (2021). `https://doi.org/10.1038/s41594-021-00590-w`.

Further manuscripts as main author under review or currently in preparation, for which a preprint is not available

- Jitesh Neupane*, **Gabriele Lubatti***, Mayra Luisa Ruiz Tejada Segura, Sabine Dietmann, Antonio Scialdone, Azim Surani. Human embryo organoids as a model for peri-and post-gastrulation embryo development. Currently under review in Nature.

- **Gabriele Lubatti***, Marion Genet*, Maria Elena Torres Padilla, Antonio Scialdone. Single cell dissection of pluripotency states across mammals. Currently in preparation.

- **Gabriele Lubatti**, Tobias Krauß, Antonio Scialdone. Cell competition as a mechanism to control embryo size at the onset of gastrulation. Currently in preparation.

Detailed descriptions of my specific contributions for the three publications ("Cell competition acts as a purifying selection to eliminate cells with mitochondrial defects during early mouse development", "MitoHEAR: an R package for the estimation and downstream statistical analysis of the mitochondrial DNA heteroplasmy calculated from single-cell datasets" and "CIARA: a cluster-independent algorithm for the identification of markers of rare cell types from single-cell sequencing data" for which I am main author can be found in **chapter 3**, in the sections **3.1**, **3.2** and **3.3**, respectively.

List of software packages developed

A detailed description of the methods that I developed during my PhD is provided in **chapter 2** and **chapter 3**.

- **MitoHEAR** (`https://CRAN.R-project.org/package=MitoHEAR`)
- **CIARA** (`https://CRAN.R-project.org/package=CIARA`)
- **SCOPRO** (`https://CRAN.R-project.org/package=SCOPRO`)
- **WOTPLY** (`https://CRAN.R-project.org/package=WOTPLY`)

Overview of the thesis

The thesis consists of four chapters. The **first** (Introduction) is a summary of the biological background of my PhD projects. The **second** (Materials and Methods) includes a description of the algorithms and computational workflows adopted during my analyses of single-cell sequencing datasets as well as an introduction to the original methods that I developed.
The **third** chapter (Summary of contributed articles) is a summary of the core publications as main author with also a short overview of the projects with manuscripts currently under review or in preparation.
Detailed author contributions are given at the beginning of each section.
Finally, the **fourth** chapter (Discussion) provides an outlook for the projects I worked on in my PhD.

Contents

Chapter 1

Introduction

The chapter includes an introduction to the concepts of cell states and cell potency. The state of a cell encompasses many different functions. One of these, which is particularly important during development, concerns the ability of cells to differentiate into different types. Cell potency refers to the varying ability of cells to differentiate into specialized cell types. Cells with the greatest potency can generate more cell types or states than those with lower potency.
In this chapter, I will introduce the biological background of my PhD projects. I will focus on changes in cell states and potency in both in vivo models (developing embryos) and in vitro models (stem cells and embryo organoids). A more detailed description of the projects will be provided in the **third chapter** (Summary of contributed articles) and the **fifth chapter** (Appendices) of the thesis.

1.1 Cell states and cell potency

Every cell in an adult arises from a single zygote through a sequence of cell divisions and fate decisions, in which a cell transitions from one type or state to another. Although several definitions are possible, in this thesis, I refer to "cell state" as the functional features of the cell (i.e., what cells can do) [1].
The state of a cell encompasses many different functions. One of these, which is particularly important during development, concerns the ability of cells to differentiate into different types. Cell potency refers to the varying ability of cells to differentiate into specialized cell types [2]. Cells with the greatest potency can generate more cell types or states than those with lower potency.
A developing embryo is a prominent example of a system where cells change continuously and rapidly their states and potency. In particular, during embryo development, the potency of the cells decreases over time. These changes are driven by complex gene regulatory interactions [3]. Recently developed single-cell methods are allowing the characterization of cell-to-cell heterogeneity and the tracking of differentiation pathways [4].
Understanding how cell potency works can open up unexplored scenarios with important applications in regenerative medicine, including the ability to manipulate cell identity and create any desired cell type.
Cells change their state in a very precise manner. A crucial question is how the

ability of changing state coexists with the robustness of cell fate decision (i.e. the process in which a particular cell develops into a final cell type). Indeed, it has been shown that the embryo is very robust to changes in the number of cells or the construction of chimeric embryos with cells of varying fitness levels. However, the molecular mechanisms by which cells change state, the timing and location of these changes, and how cells maintain their identity and robustness against external perturbations remain largely unknown. A comprehensive and in-depth understanding of cell potency is still lacking.

1.2 Cells states in embryo development and stem cells

Cells from the earliest stages of development can generate both embryonic and extraembryonic tissues and are called totipotent cells. In mouse embryos, cells from the zygote and 2-cell stage are totipotent [5]. During the late 2-cell stage, a network of genes (in particular, the *Zscan4* genes family) is activated. These genes are highly specific to this stage and not expressed in later stages [6] (**Figure 1.1**).
After the late 2-cell stage, the cells of the mouse embryo become committed to two lineages: either the embryonic lineage (represented by the inner cell mass that will give rise to the epiblast) or the extra-embryonic lineages (primitive endoderm and trophoblast cells that form a large part of the placenta). Cells from the inner cell mass (ICM) are pluripotent because they can contribute only to embryonic but not extra-embryonic lineages [6]. In mice, the transition from totipotency to pluripotency occurs between the 4-cell stage and the morula stage. Pluripotent cells from inner cell mass give rise to hypoblast and epiblast. Hypoblast originates the yolk sac, while the epiblast, at the gastrula stage, differentiates into the three primary germ layers that will constitute the embryo proper.All these transitions between different pluripotent states occur during early stages of embryonic development, which are difficult to study in vivo for both technical and ethical reasons.
However, there are also in vitro models of the different potency states. For example, embryonic stem cells (ESCs) and induced pluripotent stem cells (iPSCs) are two models of pluripotency that are available for multiple species ([7], [8]). Embryonic stem cells (ESCs) can be derived from ICM cells and maintain pluripotency in culture. Differentiated cells can be reprogrammed into pluripotent cells through forced expression of pluripotency-associated master transcription factors to get induced pluripotent stem cells (iPSC).
The geometry of tissues can have important consequences on how cells interact with each other, which can ultimately affect cell state transitions. While standard stem cell models do not generate 3D structures, more advanced in vitro models of embryos have been developed in recent years that can form a 3D structure and reproduce aspects of in vivo embryonic development more faithfully[9]. How well these embryo models reproduce in vivo systems and how they differ from real embryos are important and still largely unanswered questions. In the next sections, I focus on the open questions regarding the characterization of heterogeneity in embryo development and stem cells that I faced during the PhD.

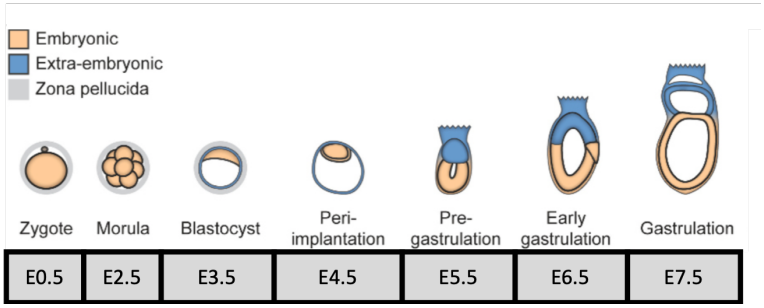


Figure 1.1: **Mouse embryo development from zygote stage to gastrulation**. Scheme of mouse embryo development from early totipotent stages until gastrulation. E0.5 (and similar for all the other stages) stands for embryonic day 0.5. Image adapted from [10]

1.3 Cell states in pluripotent cells at the beginning of gastrulation

In mouse embryos, in the 24 hours time window before gastrulation (between embryonic stages 5.5 and 6.5) around 35% of epiblast cells are eliminated through apoptosis [11]. This has led to the hypothesis about the existence of two states of epiblast cells, one with higher level of fitness and the other with a lower level of fitness that is eliminated [11]. The process whereby cells with higher quality eliminate cells with lower quality is known as cell competition [12]. Cell competition is a fitness-sensing mechanism that eliminates cells that, although viable, are less fit than their neighbours (**Figure 1.2**).
There are several in vitro models of cell competition available. For example, it has been shown that ESCs that display defective bone morphogenetic protein (BMP) signaling or defective autophagy or are tetraploid are eliminated in an apoptosis-dependent manner. The elimination takes place only when they are co-cultured with wild types cells but not when they are cultured alone [13] (**Figure 1.3**).
In the mouse embryo at the onset of gastrulation, the eliminated cells show features such as a low level of *Myc* and *mTor*, and a high level of p53, which have been associated with cell competition observed in vitro [11, 14]
However, very little is known about the characteristics of the cells that are eliminated in the embryo in a physiological context.
Our research, in collaboration with Prof. Dr. Tristan Rodriguez and Dr. Ana Lima, aims to characterize the transcriptional features of the loser cell state and how the transition from winner to loser cells occurs. Additionally, using mathematical modeling, we have begun to explore whether cell competition plays additional beneficial roles for the embryo, aside from eliminating cells with lower fitness (see **chapter 3** (sections 3.1.1, 3.2.3) and **chapter 5** (Appendix A) of the thesis for more information).

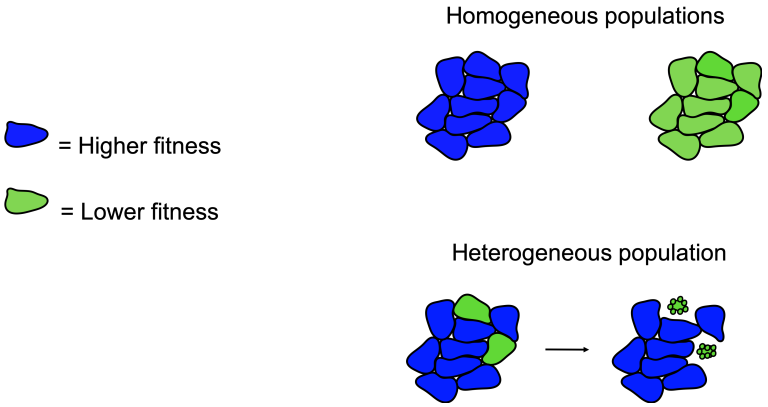


Figure 1.2: **Cell competition**. Schematic representation of cellular competition. Cells with lower fitness (represented in green) are eliminated when they are cultured together with higher-fitness cells (represented in blue) in an heterogeneous population. Instead, cells with lower fitness are able to survive and proliferate if they are cultured alone (homogeneous population).

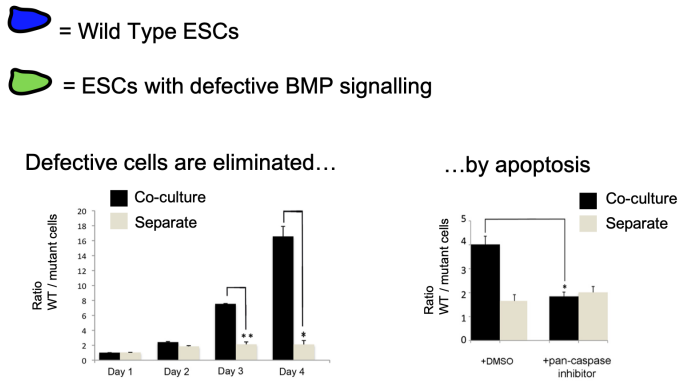


Figure 1.3: **Example of an vitro model of cell competition.** Wild type ESCs (WT) are the winners of the competition, while ESCs with defective BMP signaling (mutant cells) are the losers.
When WT and mutant cells are cultured separately, they both survive and proliferate over time (from day 1 to day 4). On the contrary, when they are co-cultured together, the mutant cells are eliminated by the WT.
This elimination is dependent on apoptosis because if a pan-caspase inhibitor is used to block apoptosis, the ratio of WT to mutant cells in separate and co-culture conditions remains the same. However, under normal conditions (DMSO), the ratio between WT and mutant cells is higher when they are co-cultured.Image adapted from [13]

1.4 Characterizing the transition between pluripotency and totipotency in stem cells

Studies in mouse ESCs and iPSCs have revealed that a rare subpopulation of ESCs (less than 0.5%) expresses much lower levels of *Oct4*, *Nanog* and *Sox2* than the majority of ESCs while expressing a group of genes that are only detected in 2-cell mouse embryos (*Zscan4* genes) [15].
Based on these transcriptional features, they are named 2-cell embryo-like cells (2CLC). The 2CLC have the ability to contribute to both embryonic and extraembryonic tissues and so they are totipotent-like cells [15] (**Figure 1.4**).
Recently, it has been shown the existence of totipotent-like cells in human embryonic stem cells (hESCs) [16]. For humans, the totipotent-like cells are transcriptionally similar to the cells from the 8-cell stage, which is a totipotent stage. The transition from pluripotent cells to totipotent like cells is still not well characterized.
Our project, in collaboration with Prof. Dr Maria Elena Torres Padilla and Ane Iturbide, aims at finding pathways that can regulate the 2CLC program. We identified retinoic acid (RA)-signaling pathway as robust inducers of 2CLC.
Using single-cell RNA-seq, we revealed the transcriptional dynamics of 2CLC reprogramming and showed that ES cells undergo distinct cellular trajectories in response to RA.
The comprehension of the mechanism that regulates the transition from pluripotency to totipotency in stem cells is crucial to obtain an in vitro model of totipotency (**Figure 1.5**). Having an in vitro model of totipotency would allow us to study the mechanisms regulating totipotency that are still unknown, without the need to rely on in vivo models that are limited by technical and ethical issues (**Figure 1.6**).

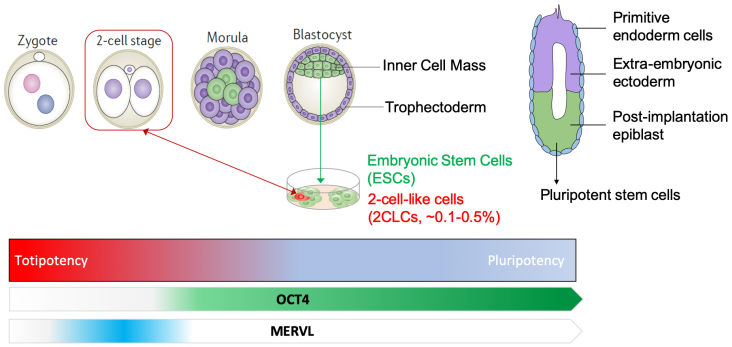


Figure 1.4: **Current model of pluripotency in mouse embryo**. Embryonic stem cells (ESCs) are derived from the inner cell mass of an embryo at the blastocyst stage. A rare subpopulation of ESCs, which comprises less than 0.5% of all ESCs, expresses lower levels of OCT4 and a network of genes, particularly the *Zscan4* gene family that uses the retrotransposon element MERVL as an alternative promoter, that are typically only detected in the 2-cell stage embryo. These ESCs are known as 2-cell embryo-like cells (2CLC) [15]. Due to their transcriptional features, 2CLCs have the ability to contribute to both embryonic and extraembryonic tissues, making them totipotent cells.

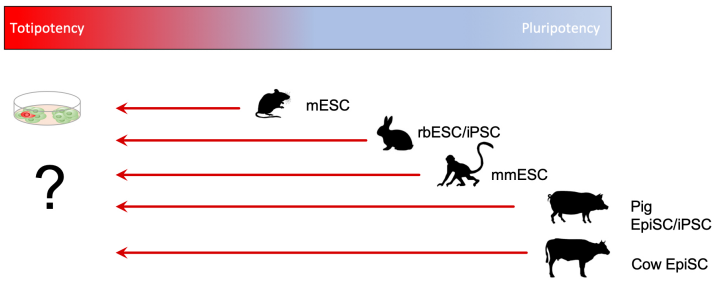


Figure 1.5: **Mammalian models of pluripotency**. In vitro models are available for various mammal species. For certain species, naïve pluripotent cells are utilized, while for others, primed pluripotent cells are used. The naïve state represents the cellular state of the preimplantation blastocyst inner cell mass, while the primed state is representative of the post-implantation epiblast cells[17].

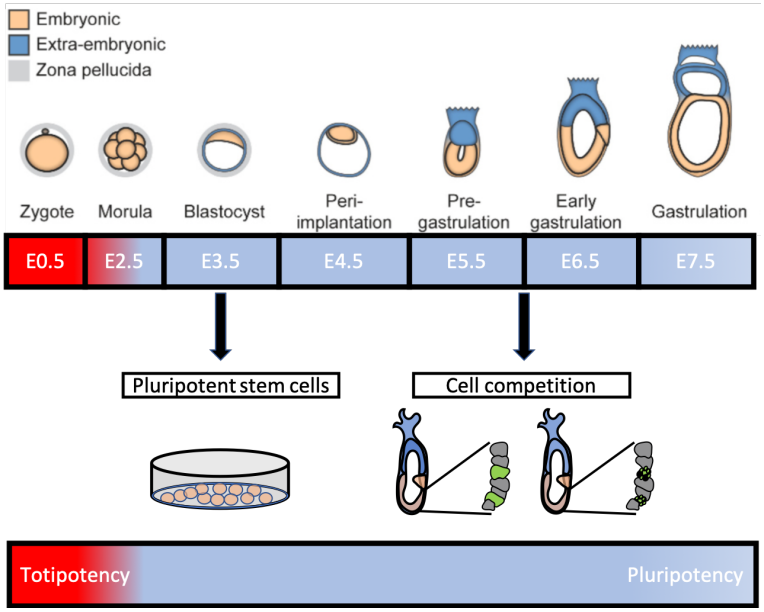


Figure 1.6: **Stages of mouse embryo development during which cell state transitions were studied in my PhD.** Schematic representation of mouse embryo development from the early totipotent stages to gastrulation (top panel). In vitro models of pluripotent cells are typically derived from the inner cell mass of the embryo at the blastocyst stage. During the 24-hours time window before gastrulation, cells with lower quality are eliminated through a cell fitness mechanism called cell competition.

1.5 Modelling human embryo development with organoids

More sophisticated in vitro models are available for studying embryo development during pluripotent stages. These are self-organizing three-dimensional (3D) models generated with stem cells and are called embryonic organoids.
The last few years have seen the emergence of several types of embryonic organoids, each modelling a specific stage or process occurring during embryogenesis. One of these models is called gastruloids, which consist of aggregates of embryonic stem cells that, under defined culture conditions, undergo controlled proliferation, symmetry breaking, and the specification of all three germ layers characteristic of vertebrate embryos [18].
During gastrulation, the epiblast cells invaginate and undergo epithelial to mesenchymal transition (EMT), giving rise to primitive streak (PS) and the formation of mesoderm and endoderm. Primordial germ cells (PGCs), the precursors of sperm and oocytes, also appear during gastrulation. Post-gastrulation development is marked by the initiation of neuronal precursors followed by organogenesis.
Gastruloids provide opportunities to conduct mechanistic studies on specific aspects of human gastrulation, and some of these models recapitulate aspects of post-gastrulation human development. In collaboration with Prof. Dr. Azim Surani and Dr. Jitesh Neupane, I characterized a novel human embryonic organoid model that recapitulates crucial features of human gastrulation and early neurulation.

1.6 Aim of the thesis

The aim of the thesis is to characterize the molecular mechanisms underlying fundamental cell state transitions occurring during embryonic development using in vivo and in vitro systems. To achieve this goal, I have analyzed single-cell RNA sequencing datasets using state-of-the-art algorithms as well as original machine learning methods that I have developed (**Figure 1.7**).
In particular, I focused on two key cell state transitions in early embryonic development. The first transition concerns the epiblast cells in mouse embryos. In my project, I aimed to address the following questions:

1. Do epiblast cells coexist in two states, one with low fitness and the other with high fitness?
2. Are epiblast cells with low fitness eliminated through cell competition?
3. What are the mechanisms that lead cells to transition into a lower fitness state?

To investigate these questions, I analyzed single-cell RNA-seq datasets and developed a tool called MitoHEAR, which quantifies mitochondrial heteroplasmy.
The second transition that I studied is the switch from the totipotency to pluripotency state, which is crucial for the development of an organism as it determines the cell types that a given cell can differentiate into. It has been demonstrated that transitions from a pluripotent to a totipotent-like state are

rare in in vitro cultures of embryonic stem cells in mice and humans [15] [16]. However, we don't know whether such state transitions occur in stem cells of other mammalian species, and whether there is an evoluationary conserved totipotency gene network. I have investigated these questions by analysing single-cell RNA-seq datasets generated from in vitro stem cell models of multiple species, including rabbits, pigs, cows and monkeys. To this aim, I developed a new machine-learning algorithm, called CIARA, to identify rare cell types in single-cell sequencing datasets in an unsupervised manner.

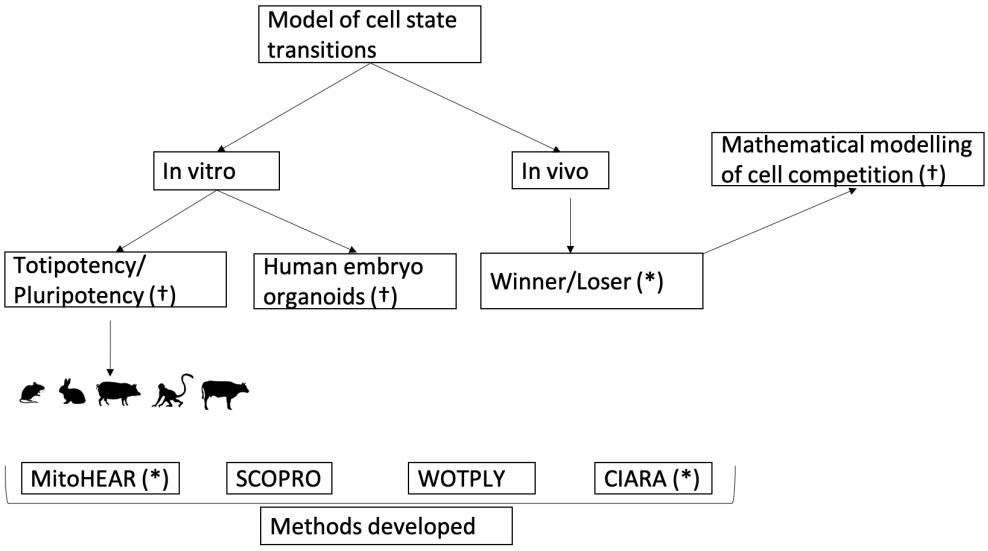


Figure 1.7: **Overview of the PhD projects**.
My thesis focused on key cell state transitions occurring early in embryonic development, in stem cells, and in embryo organoids. To study these transitions, I developed novel computational methods for analyzing single-cell sequencing data. MitoHEAR quantifies heteroplasmy based on single-cell RNA sequencing datasets. SCOPRO facilitates annotation transfer between a reference and query dataset by scoring gene pairs with conserved fold-change across testing and reference datasets. WOTPLY establishes connections between clusters at different time points using transition matrices derived from Waddington-OT analysis. CIARA is a cluster-independent computational tool designed to identify genes likely to serve as markers for rare cell types. For more detailed information on the methods I developed, please refer to **chapter 2**, **chapter 3**, and **chapter 5** of my thesis. Projects marked with an (*) indicate the availability of a published paper, while those marked with a (†) indicate that the manuscript is currently under revision or in preparation.

Chapter 2

Materials and Methods

In this chapter, after introducing single-cell RNA-sequencing, I will describe the algorithms and the computational pipelines that I utilized during my analyses of single-cell sequencing datasets. I will also introduce the original machine learning methods that I developed during my PhD in order to address specific research questions for which an established method was not yet available. In particular, I will briefly introduce new methods to: find rare cells (see section **2.4**); project a query onto a reference dataset while identifying the genes with conserved or non-conserved expression patterns (section **2.7**); visualize the results of a cell lineage relationship analysis (section **2.11**); analyze mitochondrial DNA heteroplasmy from single-cell RNA-sequencing (section **2.12**). A more detailed description of the developed methods will be given in the **third chapter** and in the **fifth chapter** of the thesis.

2.1 Single-cell RNA sequencing

Single-cell RNA sequencing (scRNA-seq) is an experimental technique that allows the profiling of the expression levels of all genes at a single-cell resolution. Hence, it is commonly used to characterize the heterogeneity of cell states in tissues [4]. As such, scRNA-seq is particularly suitable to address the research questions mentioned in the first chapter, which concern the molecular mechanisms underlying cell state transitions. Typical scRNA-seq workflows include single-cell dissociation, single-cell isolation, library construction, and sequencing [19]. The output of sequencing is read data.
Single-cell isolation is performed differently depending on the experimental protocol. While plate-based techniques isolate cells into wells on a plate, droplet-based methods rely on capturing each cell in a microfluidic droplet. The main advantage of plate-based techniques is the high sequencing depth (i.e. more reads per cell). On the contrary droplet-based methods are able to sequence a larger number of cells, at the cost of lower sequencing depth.
Library construction is the process in which the intracellular mRNA is captured, reverse-transcribed to cDNA molecules and amplified. During this step, the mRNA from each cell is labelled with a well- or droplet-specific cellular barcode. Furthermore, many experimental protocols also label captured mRNA molecules with a unique molecular identifier (UMI) [20], before the amplifica-

tion step. The purpose of UMIs is to account for amplification bias, which plays an important role in single-cell RNA sequencing, due to the low amounts of starting material. Furthermore, scRNA-seq techniques can be divided into full-length and tag-sequencing methods. Full-length techniques generate reads from all regions of the transcript, while tag-sequencing techniques are biased towards reads from the 3' or 5' end of the transcript. Full-length methods provide better coverage across transcripts, while tag-sequencing methods are better for detecting short transcripts [21].
Raw data generated by sequencing machines are processed to obtain matrices of molecular counts, representing a quantification of gene expression levels in single cells. Raw data processing pipelines, such as Cell Ranger [22] and Salmon [23], can be used to obtain such gene count matrices. The resulting count matrices have as dimensions the number of barcodes (corresponding to cells) and the number of transcripts (**Figure 2.1**).

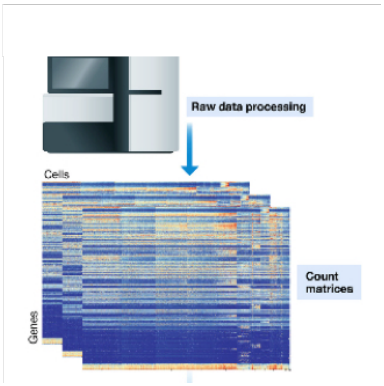


Figure 2.1: **Generation of single-cell RNA sequencing data**.
This schematic picture shows how a single-cell RNA-sequencing dataset is generated, after single-cell capture and library preparation. Sequencing reads are generated, e.g., via an Illumina NGS platform, and then processed with dedicated algorithms (such as CellRanger, etc) that produce count matrices. Such matrices, where features (genes) are listed on the rows and samples (cells) on the columns, represent the starting point of most computational analyses. This image is adapted from [24].

2.2 Quality control and normalization

Once a count matrix has been generated, the first, crucial steps in the analysis are quality control and data normalization. The purpose of quality control is to keep for downstream analysis only good-quality single cells and to discard libraries generated from, e.g., dying cells, doublets, etc.
Cell quality control (QC) is commonly performed based on three QC covariates: the number of molecular counts (count depth), the number of genes detected, and the fraction of counts mapped to mitochondrial genes [25]. Generally, "bad cells" have a low number of molecular counts, a low number of genes detected and a high fraction of counts mapped to mitochondrial genes (**Figure 2.2**).
Count depths can differ between cells for biological as well as for technical rea-

sons, i.e., due to the inherent variability in the steps involved in RNA sequencing. As a result, any differences observed when comparing gene expression between cells based on raw count data may have arisen solely due to sampling effects. To address this issue, data normalization techniques are employed, such as scaling count data to obtain relative gene expression abundances in cells (**Figure 2.2**). These techniques are described in detail in references such as [24] and [26]..
The most commonly used normalization method in scRNA-seq is count depth scaling, also referred to as "counts per million" or CPM normalization. This method, derived from bulk expression analysis, normalizes count data by dividing them by a size factor proportional to the count depth per cell. CPM normalization is widely used and has been implemented in many scRNA-seq analysis packages. Data from full-length protocols may benefit from normalization methods that correct for differences in gene length. One commonly used normalization method for full-length scRNA-seq data is transcript per million (TPM) normalization. This normalization method first normalizes for gene length and later normalizes for count depth.
However, scRNA-seq data normalization is still an open problem, and the most suitable technique might depend on the specific features (such as sparsity) of the dataset one is analysing [27].

2.3 Batch effects

Batch effects are technical, non-biological, differences between samples that can originate from different sources (e.g., changes in temperature, reagents, etc). If these effects are not taken into account, the analysis can lead to incorrect conclusions. Batch effects can be corrected using linear or non-linear approaches. Linear methods work well to correct batch effects between samples that are expected to include cells in the same states, while non-linear methods are suitable for cases where some cell types or states are not shared among datasets (**Figure 2.2**).
In the last few years, many computational techniques for batch effect correction have been developed and tested in systematic benchmarking studies [28] [29].
In my analyses, I have employed Scanorama [30], one of the most popular non-linear methods for batch correction and also one of the best performing methods according to [29].
Scanorama automatically identifies the scRNA-seq datasets, among those given as input, with at least one cell type in common and can leverage those matches for batch correction. Scanorama searches nearest neighbors to identify shared cell types among all pairs of datasets. Mutually linked cells form matches that can be leveraged to correct for batch effects and merge experiments together. The method is robust to different dataset sizes, preserves dataset-specific populations and does not require that all datasets share at least one cell population.

2.4 Features selection and dimensionality reduction

scRNA-seq datasets include the expression profiles of thousands of genes. Typically, many of these genes will not be informative, e.g., because of data sparsity or because the variation of their expression levels is predominantly driven by noise. Feature selection is a processing step during which the genes that are more likely to include biologically relevant signals are filtered to be used in downstream analyses.
A typical approach consists in selecting the highly variable genes [31], i.e., genes whose expression across cells varies more than average (**Figure 2.2**). While selecting highly variable genes can aid in tasks such as identifying cell types via clustering, more specific feature selection strategies should be employed for targeted tasks. For example, standard feature selection methods often miss potential markers of rare cell types (present with a frequency $<1\%$) [32], which can be a significant limitation in many biological contexts. For instance, in developmental studies, it may be important to pinpoint at what stage a particular cell type starts to emerge; in cancer research, researchers may need to search for rare cells that could develop drug resistance [33]; and in stem cell line characterization, researchers may need to search for cell transitions in different pluripotency states [16, 34].
Being able to reliably identify rare cell types was crucial for my projects. Hence, I developed a new method called CIARA (Cluster Independent Algorithm for the Identification of Markers of Rare Cell Types) to identify potential markers of rare cell types. CIARA tests the local enrichment of cells expressing any given gene on k-nearest neighbors graphs and selects those genes that tend to be expressed by a small number of transcriptionally similar cells (**Figure 2.3**). More details about CIARA are included in the manuscript published in *Development*, which is included in the thesis in **chapter 5** (Appendix C).
Feature selection algorithms typically select $\sim 10^3$ genes for downstream analyses. However, for data visualization or other types of analyses, it is necessary to reduce the dimensionality even further. Many dimensionality algorithms have been proposed that embed the expression matrix into a low-dimensional space, which is designed to capture the underlying structure in the data in as few dimensions as possible (**Figure 2.2**).
Reduced dimensions are generated through linear or non-linear combinations of feature space dimensions (gene expression vectors). Principal component analysis (PCA) is a linear approach that generates reduced dimensions by maximizing the captured residual variance in each further dimension [35]. Among non-linear dimensionality reduction methods, the Uniform Manifold Approximation and Projection method (UMAP) is a very common approach [36]. UMAP constructs a high-dimensional graph representation of the data, then optimizes a low-dimensional graph to be as structurally similar as possible to the high-dimensional one. This is achieved by minimizing the binary cross-entropy between the high-dimensional and low-dimensional probability distributions. Since UMAP is a non-linear method, it is able to capture more complex relationships between features in the high-dimensional space compared to linear methods like PCA. However, UMAP has limitations when it comes to the interpretability of the reduced dimensions, while for PCA the new features (principal components)

are simply a linear combination of the old features (genes).
Diffusion maps is another popular non-linear dimensionality reduction technique that is often used in the analysis of scRNA-seq data[37]. In particular, diffusion maps are particularly suitable for reconstructing continuous transcriptional trajectories that represent, for example, differentiation processes [38]. At the core of diffusion maps there is the transition matrix T that approximates the dynamic transitions of cells through stages of the differentiation process. This transition matrix is computed using a nearest neighbor graph whose edge weights have a Gaussian distribution with respect to Euclidean distance in gene expression space; transition probabilities correspond to edge weights. The eigenvectors of T are known as diffusion components.
An alternative to classical visualization on the cell level is partition-based graph abstraction (PAGA) [39]. PAGA generates a graph whose nodes correspond to cell groups, and whose edge weights quantify the connectivity between these groups. The statistical model considers groups as connected if their number of inter-edges exceeds the number of inter-edges expected under random assignment. PAGA is able to preserve both continuous and disconnected structures in data.

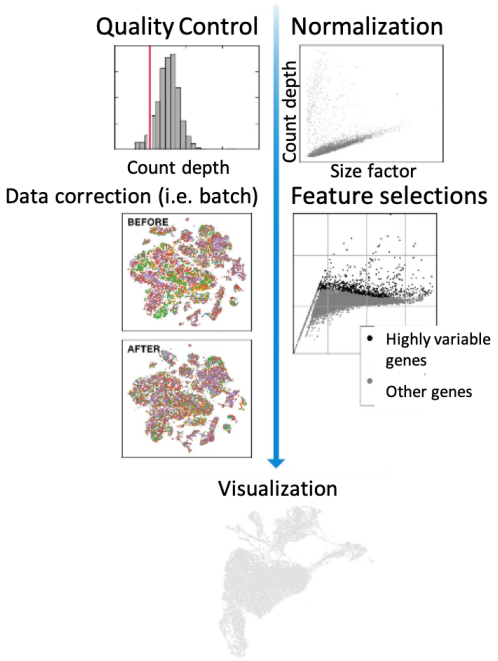


Figure 2.2: **First steps of single-cell RNA sequencing data analysis**. The figure illustrates the first key steps in the analysis of the single-cell RNA-seq datasets. Quality control is a critical step needed to remove cells with poor quality from downstream analysis. Data normalization is then performed, followed by feature selection and batch correction. Typically, a 2D visualization of the data is then produced, using techniques like UMAP. This image is adapted from [24].

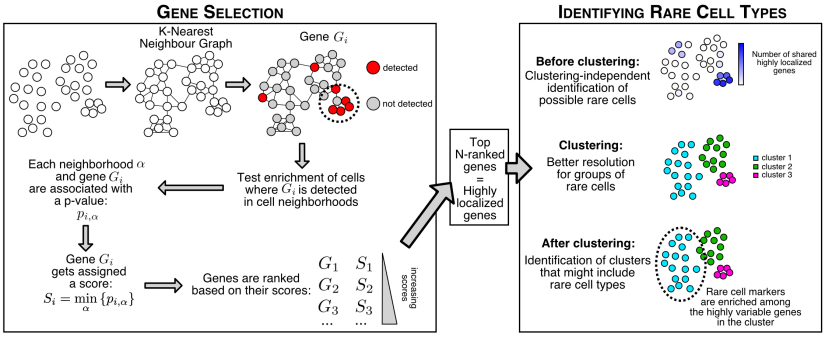


Figure 2.3: **Schematic representation of CIARA.**
CIARA computes a score for each gene based on how cells expressing that gene are distributed on a K-nearest neighbor graph (left panel). Lower scores correspond to genes that are mostly expressed in neighboring cells, i.e., are "highly localized", and hence are more likely to be markers of rare cell types. The right panel summarizes how the top-ranked genes are used to visualize and identify groups of rare cells. CIARA can generate a 2D representation of the data, which shows how many and which of the top selected genes each cell expresses and shares with its neighbors. The "highly localized" genes can be used with standard clustering algorithms to define the group of rare cell types. Importantly, CIARA also provides an unsupervised, quantitative evaluation of whether a cluster in the data may include a rare sub-population of cells: this is done by testing the statistical significance of the overlap between the set of highly variable genes within the cluster and the potential rare cell type markers identified by CIARA. Image adapted from [40], also included in Appendix C from **chapter 5**.

2.5 Cluster analysis

After dimensionality reduction and visualization, the next step is typically the identification of groups (clusters) of cells with similar gene expression profiles, which tend to coincide with the cell types present in the dataset (**Figure 2.7**). In general, there are two different approaches that clustering algorithms can adopt: some algorithms (like hierarchical clustering) [41] aggregate cells based on a distance matrix, while others (such as community detection algorithms) start from a graph-based representation of the data [42].
For the first class of algorithms, cells are assigned to clusters by minimizing intracluster distances or finding dense regions in the reduced expression space. For instance, the popular k-means clustering algorithm divides cells into k clusters by determining cluster centroids and assigning cells to the nearest cluster centroid. Centroid positions are iteratively updated. This approach requires the expected number of clusters as input, which is usually unknown and must be calibrated heuristically.
Community detection methods are graph-partitioning algorithms and thus rely on a graph representation of single-cell data. This graph representation is obtained using a K-nearest neighbors (KNN) graph. In KNN graph, cells are

represented as nodes. Each cell is connected to its K most similar cells, which are typically obtained using euclidean distances on the PCA-reduced expression space. A common community detection method is the Louvain algorithm. Conceptually, the Louvain algorithm detects communities as groups of cells that have more links between them than expected from the number of links the cells have in total. The optimized modularity function includes a resolution parameter, which allows the user to determine the scale of the cluster partition. Tools like clustree [43] can help choose the value of the resolution. Clustree shows the relationships between clusters at multiple resolutions, allowing users to see how clusters change as the resolution increases. This can help assess the relationship between distinct clusters (e.g., clusters representing the same or similar cell types tend to merge at lower resolutions) and can also highlight the presence of "unstable" clusters that only appear at very high resolutions. As an alternative to clustree, other approaches are based on subsetting genes or cells to identify robust clusters.

2.6 Cluster annotation

Once clusters are identified, the next step is to annotate them with a meaningful biological label, which will indicate the type or state of cells in each given cluster. Clustering labelling is done by finding the gene signatures (marker genes) that characterize the clusters (**Figure 2.7**).
The sets of marker genes can be identified by performing differential expression (DE) testing between two groups: the cells in one cluster versus all other cells in the dataset. To this aim, the expression data can be modeled with a negative binomial (NB) distribution, and a fit with a negative binomial generalized linear model for each gene is performed [44] [45] [46]. Other methods are based on the utilization of a zero-inflated negative binomial distribution [47]. The zero-inflated negative binomial distribution is employed to model count data that exhibit a higher proportion of zeros than what would be expected from a standard negative binomial model. This approach is particularly well-suited for data generated through droplet-based methods, which often have low sequencing depth, leading to an excess of zeros.
Other methods for differential expression analysis simply perform a non-parametric test such as a Wilcoxon Rank Sum test.
In the presence of reference datasets, automated cluster annotation is possible by directly comparing the gene expression profiles of annotated reference clusters to individual cells. Tools such as scArches [48], scmap [49], SciBet [50] and Seurat [51] can transfer annotations between a reference and a target dataset. Most of the currently available methods have two main limitations:

- the annotation depends on the cell states included in the reference. If a cell type in the dataset is not included in the reference, then it will still be assigned to the closest reference cell state.
- the set of genes with conserved or non-conserved expression patterns across the reference and the target dataset are not provided as output.

The algorithm SCOPRO (SCOre PROjection), which I developed and introduced in the section below, overcomes both of these limitations by scoring pairs

of genes that exhibit a conserved fold-change across both testing and reference datasets.

2.7 SCOPRO: an algorithm for projecting a query onto a reference dataset

As more and more scRNA-seq datasets are generated, it has become crucial to compare them with existing annotated references. In recent years, several tools have been developed to perform label transfer from a reference to a query dataset. Among the most popular ones are scArches, Seurat, SciBet, and scmap ([48], [51], [50], [49]). However, for most of the methods, the final labeling depends on the composition of the reference dataset, and it is not possible to assess the statistical significance of the assignment. Another limitation is that they do not provide the genes whose expression patterns are conserved or differ between the atlas and the query datasets (**Figure 2.7 f**).
To overcome these limitations, I developed a new algorithm called SCOPRO. First, for each cluster in the reference dataset, SCOPRO builds a network where marker genes are nodes, and an edge is added between a pair of genes if the ratio between their average expression in the cluster is above a given threshold. Then, an analogous network is built for each cluster in the query dataset (**Figure 2.7 a**). Finally, SCOPRO compares the networks obtained in the reference and the query dataset by computing a score ranging from 0 to 1 that quantifies the degree of similarity between each pair of networks, and, thus, between the corresponding clusters in the query and the reference datasets. The score is assigned by computing the percentage of intersection (perINT) [52] of the links in the corresponding networks (**Figure 2.7 b**). A p-value is then estimated by comparing the value of such a score with the score obtained from a randomly shuffled dataset (**Figure 2.7 d, e**).
The score returned by SCOPRO is not dependent on the composition of the reference dataset, in contrast with the scores given by other methods such as Seurat and scmap, or the discrete labeling provided by SciBet. Moreover, the networks that SCOPRO builds allow the identification of the best-conserved genes and those that differ the most between the query and the reference data. A gene A is defined to be conserved between a cluster in the reference and a cluster in the query if the jaccard index between the links from gene A in the two networks is above a defined threshold (**Figure 2.7 c**).

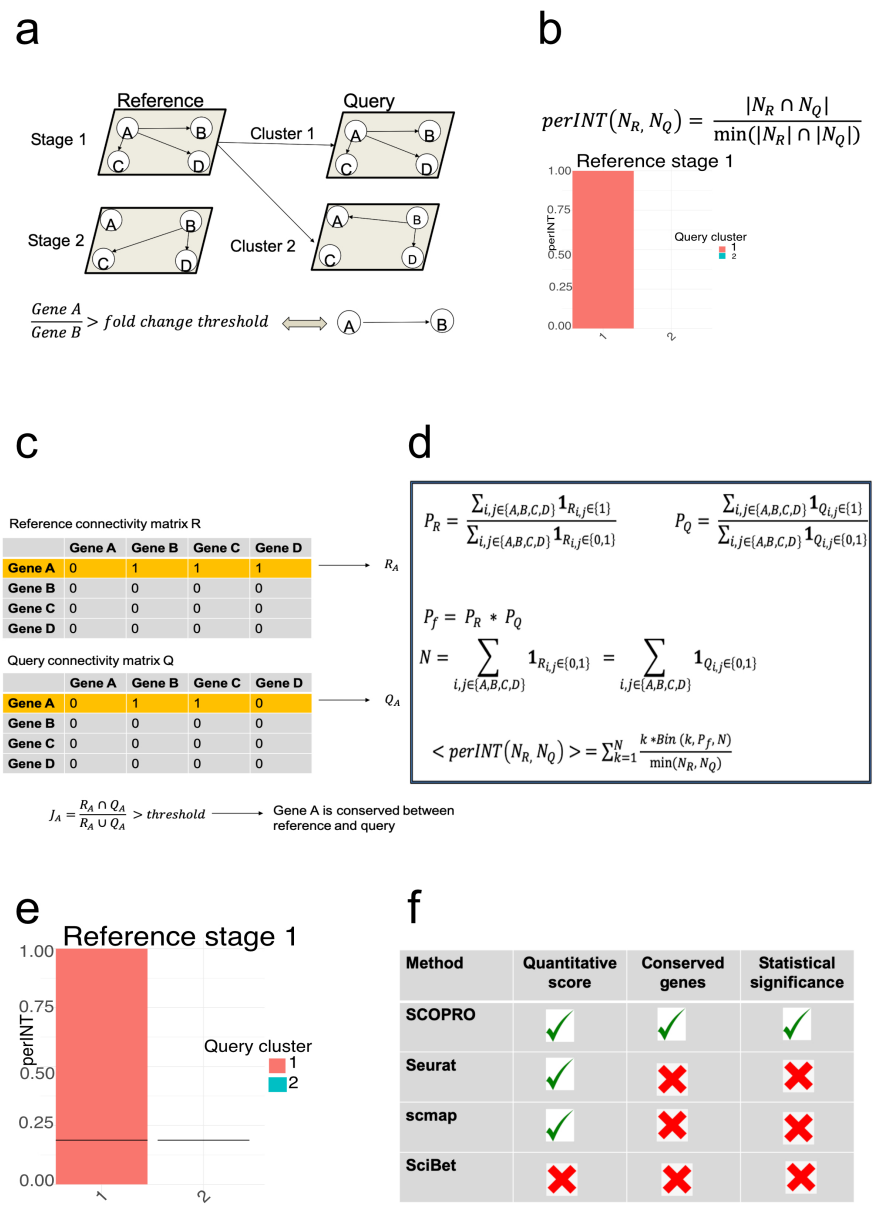


Figure 2.4: **Overview of SCOPRO**

Figure 2.4: **Overview of SCOPRO**.
a)A network is built for each cluster in the reference and the query dataset. SCOPRO builds a network where marker genes are nodes, and an edge is added between a pair of genes if the ratio between their average expression in the cluster is above a given threshold
b) Score (percentage of intersection) given by SCOPRO for the comparison between a cluster in the reference and a cluster in the query.
c) A gene is defined to be conserved between reference and query if the jaccard index between the links from the gene in the two networks is above a defined threshold.
d, e) With SCOPRO is possible to assess the significance of the given score. A p-value is estimated by comparing the value of such a score with the score obtained from a randomly shuffled dataset.
f) Comparison between SCOPRO and previously published methods for projection from reference to query. SCOPRO is the only method that gives as output a quantitative score, a list of conserved genes between the reference and the query and a p-value from which is possible to asses significance.

To illustrate how SCOPRO works, I used a scRNA-seq dataset from mouse embryos as a reference (in vivo dataset ([53]; [54])), which includes blastomeres from 2-cell stage embryos and epiblast cells from embryos at 4.5, 5.5 and 6.5 days post-fertilization. As a query, I used a scRNA-seq dataset from mouse embryonic stem cells [55]. This dataset includes three clusters, respectively with 1101, 153 and 31 cells (**Figure 2.5**).

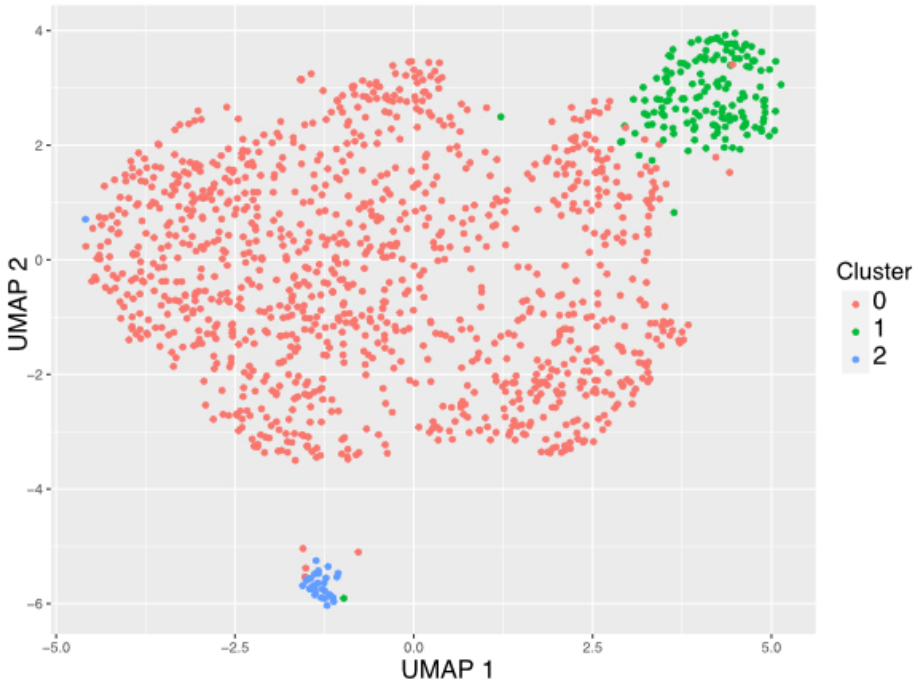


Figure 2.5: **Query dataset**.
As a query for SCOPRO, a scRNA-seq dataset generated from mouse embryonic stem cells [55] is used.

Using SCOPRO, I computed the projection scores of all clusters in the query dataset with respect to the cluster of the 2-cell stage blastomeres present in the reference. This showed that the score for cluster 2 is very high (and above statistical significance, p-value = 0), while the score for cluster 0 and 1 was very low (**Figure 2.6 a**). Conversely, the projection scores computed with respect to the epiblast cluster were not significant for cluster 2 and statistically significant for the clusters 0 and 1 (**Figure 2.6 a**). This suggests that cluster 2 is most similar to 2-cell stage embryos, while the other two clusters in the query dataset resemble epiblast cells found in later stages.
Interestingly, among the conserved markers between cluster 2 and the late 2-cells stage, there were *Zscan4* family genes (**Figure 2.6 b**). Indeed, it is known that in mouse embryonic stem cells, a rare population of cells with typical markers of the late 2-cells stage, including *Zscan4* genes, is present. This population is called 2 cells like cells (2CLC) ([34]; [15]). Therefore, cluster 2 in our query dataset corresponds to the 2CLC.
Among alternative methods, Seurat is able to correctly assign cluster 2 to the late 2-cells stage, while both scmap and SciBet fail to recover the 2CLC identity of cluster 2 because they assign it mainly or only to epiblast 4.5 (**Figure 2.6 c, d, e**). The advantage of the score provided by SCOPRO is shown in the case where we only have epiblast stages (from 4.5 to 6.5) in the reference dataset, but not the late-2 cell stage. In this scenario, SCOPRO correctly assigns a very low score for cluster 2 to each of the epiblast stages. Consistently in all cases, the score is below the random expectation (**Figure 2.6 f**). On the contrary,

Seurat, scmap, and SciBet still assign cluster 2 to epiblast 4.5 and epiblast 5.5, although it is known from the previous analysis that cluster 2 corresponds to 2CLC (**Figure 2.6 g, h, i**). This is an intrinsic limitation of these methods that can assign a final prediction only relative to the stages present in the reference dataset.
SCOPRO is available as an R package on CRAN (`https://CRAN.R-project.org/package=SCOPRO`) and on github (`https://github.com/ScialdoneLab/SCOPRO`).
I used SCOPRO in the project "Single cell dissection of pluripotency states across mammals" (see **chapter 3** section 3.2.2 for a more detailed description of the project).

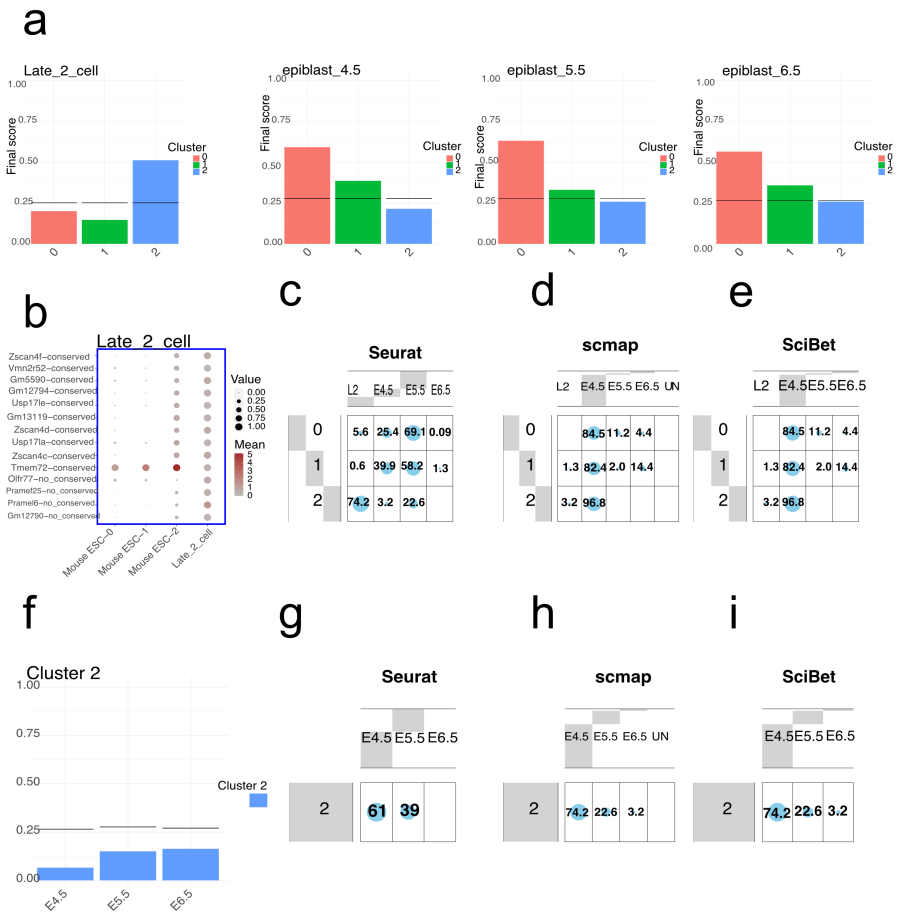


Figure 2.6: **Projection of mouse embryonic stem cells dataset on mouse embryo data.**

Figure 2.6: **Projection of mouse embryonic stem cells dataset on mouse embryo data.**
a) Final score obtained from SCOPRO for a mouse embryonic stem cells dataset ([55]) projected over a dataset from mouse embryo development (in vivo dataset ([53]; [54]) including late 2-cells stage (L2) and epiblast stages from embryonic day 4.5 to 6.5 (E4.5, E5.5 and E6.5) The black line shown the fraction of conserved links that would be expected in a random scenario, where the same amount of links in the network are randomly assigned.
b) List of conserved genes given by SCOPRO between cluster 2 from a mouse embryonic stem cells dataset ([55]) and late 2-cells stage from mouse embryo development (in vivo dataset ([53]).
c) Result obtained by Seurat for mouse embryonic stem cells dataset projected over mouse embryo data (late 2-cells stages and epiblast stages from 4.5 to 6.5). For each cell, Seurat assigned a relative score to each stage in the mouse embryo data. Finally, the cell is assigned to the stage with the highest score.
On the columns there are the clusters from the reference datasets. On the rows there are the clusters from the query. Each number indicates the percentage of cells from the cluster in the query assigned to the corresponding cluster from the reference.
d) Result obtained by scmap for mouse embryonic stem cells dataset projected over mouse embryo data (late 2-cells stages and epiblast stages from 4.5 to 6.5). For each cell, scmap assigned a score to each stage in the mouse embryo data. Finally, the cell is assigned to the stage with the highest score .
e) Result obtained by SciBet for mouse embryonic stem cells dataset projected over mouse embryo data (late 2-cells stages and epiblast stages from 4.5 to 6.5). For each cell, SciBet assigned a discrete label corresponding to one of the stage in the mouse embryo data.
f) Final score obtained by SCOPRO for cluster 2 from mouse embryonic stem cells dataset projected over mouse embryo data where only epiblast stages from 4.5 to 6.5 are present. The score is always below the random expectation for each of the three epiblast stages.
g) Result obtained by Seurat for cluster 2 from mouse embryonic stem cells dataset projected over mouse embryo data with only epiblast stages from 4.5 to 6.5. For each cell, Seurat assigned a relative score to each stage in the mouse embryo data. Finally, the cell is assigned to the stage with the highest score.
h) Result obtained by scmap for cluster 2 from mouse embryonic stem cells dataset projected over mouse embryo data with only epiblast stages from 4.5 to 6.5. For each cell, Seurat assigned a relative score to each stage in the mouse embryo data. Finally, the cell is assigned to the stage with the highest score.
i) Result obtained by SciBet for cluster 2 from mouse embryonic stem cells dataset projected over mouse embryo data with only epiblast stages from 4.5 to 6.5. For each cell, SciBet assigned a discrete label corresponding to one of the stage in the mouse embryo data.
Abbreviated cluster names: "L2" stands for "late 2-cells stage", “E4.5” for “epiblast at the embryonic day 4.5”; "E5.5" for "epiblast at the embryonic day 5.5"; "E6.5" for "epiblast at the embryonic day 6.5"; "UN" for "unassigned".

2.8 Trajectory inference methods

Discrete classification may not be sufficient to describe cellular diversity, as the biological processes underlying this heterogeneity are often continuous [56]. In order to capture transitions between cell identities and branching differentiation processes, trajectory inference (TI) methods are required (**Figure 2.7**). The temporal order of differentiating cells can be intrinsically encoded in their single-cell expression profiles. For methods such as diffusion maps, the one-dimensional variable representing each cell's transcriptional progression towards the final state is referred to as diffusion pseudo time (DPT). DPT measures transitions between cells using diffusion-like random walks [57].
The DPT is computed in the following way. For each cell, the probabilities of transitioning to each other cell in the dataset using random walks of any length on this graph are computed; these walks can be seen as a proxy for the cells' probabilities of differentiating toward different fates. The probabilities for each cell are stored in a vector, and the DPT between two cells is calculated as the Euclidean distance between their two vectors. The developmental progression of each cell is measured in the dataset by computing its DPT with respect to a specified root cell. An alternative to diffusion pseudo time is provided by slingshot [58]. The method works in two steps: first, it identifies lineages (ordered sets of clusters), and then it assigns pseudotime coordinates to individual cells. For this purpose, it makes use of principal curves to draw a path through the gene expression space of each lineage.
While trajectory inference tools typically work with gene expression levels, common single-cell RNA seq protocols allow the identification and quantification of unspliced pre-messenger RNAs (mRNAs) and mature spliced mRNAs for each gene. By using the information of unspliced and spliced mRNA, one can infer the RNA velocity of individual cells. RNA velocity describes the rate of gene expression change for an individual gene at a given time point based on the ratio of its spliced and unspliced mRNA. [59, 60]. Whereas traditional trajectory inference methods reconstruct cellular dynamics given a population of cells of varying maturity, RNA velocity relies on a dynamical model describing splicing dynamics [61].
CellRank [62] combines the robustness of trajectory inference with directional information from RNA velocity, taking into account the gradual and stochastic nature of cellular fate decisions, as well as uncertainty in velocity vectors.
CellRank models state transitions using a Markov chain, where each state in the chain is given by one observed cellular profile, and edge weights denote the probability of transitioning from one cell to another. The resulting matrix of directed transition probabilities is independent of any low-dimensional embedding and reflects both transcriptional similarity and directional information given by RNA velocity. CellRank identifies final states from a given number of macrostates and assigns to each cell a probability of fate towards one of the terminal states.

2.9 Differential expression analysis along trajectory

Once a transcriptional trajectory is identified, it is useful to identify the genes that show dynamic patters of expression along it. Some trajectory inference methods also provide as output the differentially expressed genes along a trajectory. For example, CellRank finds genes whose expression levels change along the trajectory by using the fate probabilities to smooth gene expression trends along lineages.
Lately, specific tools to infer differentially expressed genes along trajectory, such as tradeSEQ [63], have been developed.
TradeSEQ is a framework based on generalized additive models and the negative binomial distribution, which allows for flexible inference of differential gene expression within and between lineages. To achieve this, TradeSEQ infers smooth functions for gene expression measures along pseudotime for each lineage, using generalized additive models.

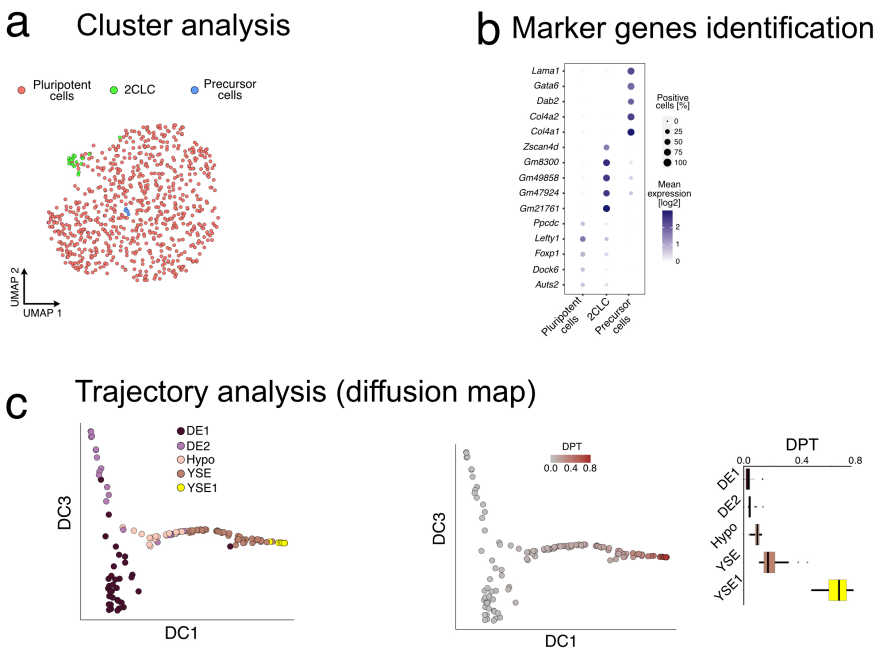


Figure 2.7: **Typical downstream analyses of single-cell RNA sequencing data**.
Once clusters are identified and visualized in a, e.g., UMAP plot (panel **a**), differential expression analysis can be used to detect marker genes for each cluster (panel **b**), which help identify cell types. To capture continuous transitions between cell identities and branching differentiation processes, trajectory inference methods are needed to provide a continuous description of the data. Panel **c** shows the results of a transcriptional trajectory analysis of endodermal cells from a human gastrula [64]. A diffusion map is used to visualize the trajectory in 2D (left panel; DC1 and DC3 are diffusion components 1 and 3, respectively), and the position of the cells along the trajectory is quantified with a diffusion pseudotime (DPT, right panel). The subplots in panels **a** and **b** show mouse embryonic stem cells data from [55]. The subplot in panel **c** shows human embryo data from [64]. Abbreviated cluster names: "DE1" stands for "Definitive endoderm 1"; "DE2" for "Definitive endoderm 2"; "Hypo" for hypoblast; "YSE" for "yolk sac endoderm"; "YSE1" for "yolk sac endoderm 1". DPT stands for "Diffusion pseudo time". I re-analyzed both datasets in [65].

2.10 Time-resolved single-cell RNA sequencing experiments

When working with time-resolved single-cell RNA sequencing datasets, we can access snapshots of single-cell transcriptomic profiles corresponding to different developmental states over time. Usually, the primary objective of the analysis is to reconstruct how cell types and states at different time points are related, specifically in terms of cell lineages.
A popular tool to analyze time-resolved scRNA-seq datasets is Waddington-OT [66]. It uses the notion that cells at any time are drawn from a probability distribution in gene-expression space. It uses scRNA-seq data collected across a time course to infer how these probability distributions evolve over time by using the mathematical approach of optimal transport (OT). This results in the reconstruction of cell lineages and how they evolve through time in the system at hand. For example, I have used Waddington-OT when I analyzed human embryo organoids collected at different time points. In this case the method allowed to link the presence of neural tube, neural crest cells, and neuronal precursors at later time points to neuromesodermal progenitors (NMPs) at a previous time point (see **chapter 3** section 3.2.1 for more information).
In large datasets including multiple cell states, it can be hard to visualize the results of cell lineage inference algorithms. Thus, I developed an algorithm called WOTPLY (Waddington-OT analysis PLot) that produces graphs representing the inferred cell state transitions, starting from a list of transition matrices from time t to time t+1. In the section below, I introduce WOTPLY in more details and show an example of its use.

2.11 WOTPLY: a tool for visualizing cell lineage relationships in time-course scRNA-seq datasets

I developed WOTPLY (Waddington-OT analysis PLot), a tool implemented in a user-friendly R package that allows the visualization of the results of algorithms analyzing cell lineage relationships in time-course scRNA-seq datasets.
For each cell at time t, Waddington-OT algorithm assigns a "bias score" toward every cluster at time t+1. For each cell, the sum of the scores towards all the clusters at time t+1 sum to 1. Each cluster of cells at time t gets assigned a score equal to the average scores of all cells in that cluster. (**Figure 2.8 a**)
For each cluster, the connections between each time point are shown. Each node stands for a cluster and clusters aligned on the same columns are coming from the same time point. The thickness of the edges between clusters at time points t and t+1 represents the average score (**Figure 2.8 b**).
With WOTPLY, it is possible to tune several parameters for customizing the analysis (e.g., to show the origin only for a subset of clusters or visualize only a maximum number of top links between each time point) (**Figure 2.8 c, d**).

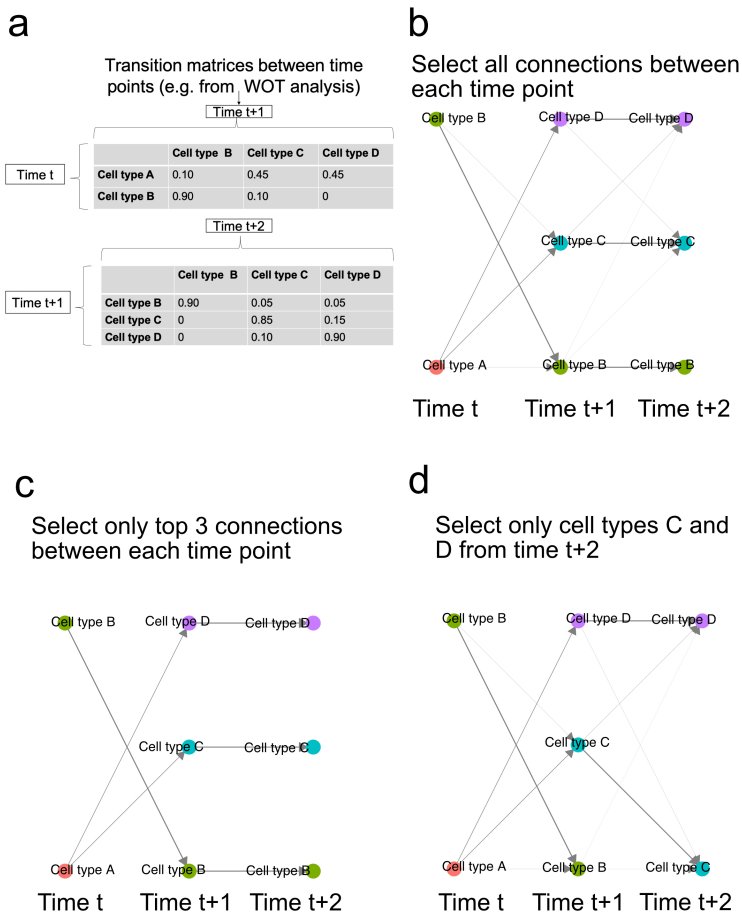


Figure 2.8: **Overview of WOTPLY**.
a) Example of input required by WOTPLY. A list of transition matrices between a set of cells from time t and a set of cells from time t+1 must be provided as input. For instance, these transition matrices can be obtained with the function compute_all_transport_maps from the python package WOT [66].
b) Example of output produced by WOTPLY in the case where the clusters selected from the latest time point (t+2) are Cell type B, Cell type C and Cell type D and all the links are shown.
c) Example of output produced by WOTPLY in the case where the clusters selected from the latest time point (t+2) are Cell type B, Cell type C and Cell type D, and only the top 3 links are shown between each transition from time t to time t+1.
d) Example of output produced by WOTPLY in the case where only clusters C and D from the latest time point (t+2) are selected, and all the links are shown.

As a use case, I ran the algorithm on a published single-cell RNA sequencing dataset (315,000 cells) collected across 18 days of reprogramming mouse embryonic fibroblasts (MEFs) into induced pluripotent stem cells (iPSCs) [66]. Cells gradually adopt either a terminal stromal state or a mesenchymal-to-epithelial transition state. The latter gives rise to populations related to pluripotent, extra-embryonic, and neural cells.
The output of the WOTPLY function displays selected clusters at day 18 ("IPS", "Trophoblast", "Epithelial", "Stromal","Neural") and their connections to clusters at previous time points (day 10, day 12, day 14, day 16 (**Figure 2.9 a**).
Each node is a cluster and nodes aligned on the same column are coming from the same time point. The weight of the links between clusters at time points t and t+1 reflects the weight of the transition probabilities from the input list of transition matrices. To improve visualization, WOTPLY allows tuning the maximum number of links to represent between clusters at time t and clusters at time t+1. Links are sorted according to weight, and only the top links are retained (**Figure 2.9 b**).
WOTPLY is available as R package on CRAN (`https://CRAN.R-project.org/package=WOTPLY`) and on GitHub (`https://github.com/SialdoneLab/WOTPLY`).
WOTPLY is used in the project "Human embryo organoids as a model for peri- and post-gastrulation embryo development" (see **chapter 3** section 3.2.1 for a more detailed description of the project).

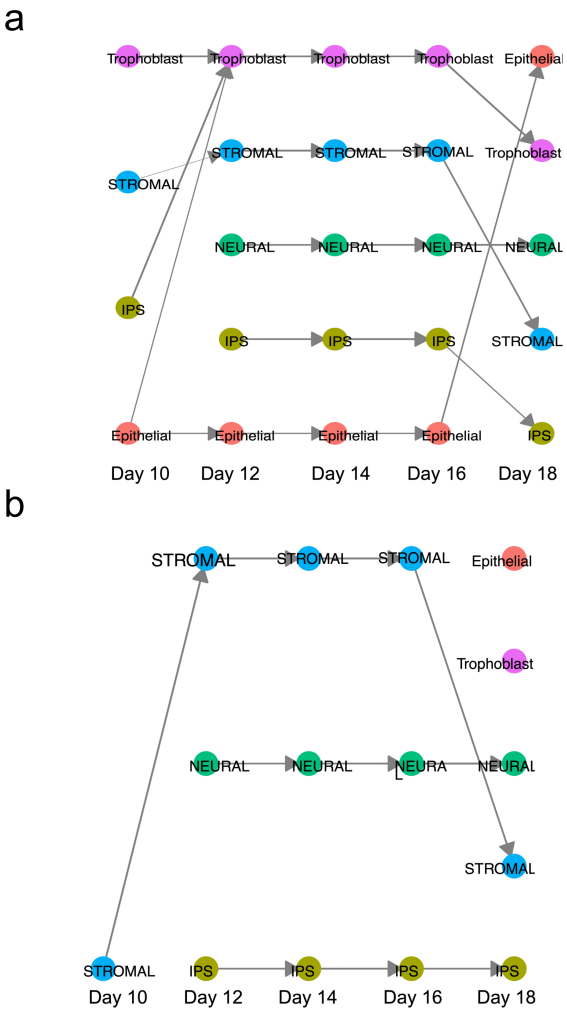


Figure 2.9: **Results of WOTPLY applied to a single-cell RNA sequencing dataset from a time-course of iPS reprogramming** . The dataset is taken from [66], I have processed cells from days 10, 12, 14, 16 and 18 with Waddington-OT and then used WOTPLY to visualize the results. **a)** Cell state transitions visualized with WOTPLY. Each node corresponds to a cluster at a given time point, and each column represents a specific time point from day 10 (first column) to day 18 (last column). In this example, all the links between consecutive time points are shown. **b)** Output from the same analysis shown in panel a, but where only the top 3 links at most (representing the most likely connections) are shown.

2.12 Analysis of mitochondrial DNA heteroplasmy

From scRNA-seq datasets, we can not only analyze the heterogeneity of cells from a transcriptional point of view, but we can also detect and quantify other biologically relevant features, such as the presence of DNA mutations. These include mutations present in the nuclear DNA or mitochondrial DNA [67]. Mitochondria are organelles found in eukaryotic cells that produce energy. They are equipped with their own DNA (mtDNA), and each cell includes multiple mtDNA copies that are not perfectly identical. Such sequence variability is called mtDNA heteroplasmy. Although mtDNA heteroplasmy has important consequences for human health [68] and embryonic development [69], many questions remain regarding how it affects cellular function and how cells regulate it. The mutation rates of mtDNA are estimated to be 10- to 100-fold higher than those of nuclear DNA [70]. This increased mutation rate makes it easier to detect mutations, particularly in samples with low coverage [67].
However, the analysis of mtDNA heteroplasmy requires careful consideration of potential artifacts. One example is the possibility that sequence changes could arise from contaminating nuclear mitochondrial sequences (NUMTS) that are mistakenly assigned to mitochondrial DNA during the read mapping process. Another important source of confounding effect is RNA editing.
Additionally, mtDNA heteroplasmy from scRNA-seq can only be reliably detected at positions of the genome with sufficient coverage.
The existing computational pipelines devised to analyse mtDNA heteroplasmy are limited either in the quality control they carry out or the statistical analyses they perform. I have developed a new computational tool called MitoHEAR (Mitochondrial HEteroplasmy AnalyzeR) implemented in R that allows the quantification of mtDNA heteroplasmy from single-cell RNA-seq data while controlling for potential technical artifacts. Starting from BAM files, MitoHEAR estimates heteroplasmy and offers several options for downstream analyses. For example, statistical tests are provided to investigate the relationship of the mtDNA heteroplasmy with continuous **Figure 2.10 b**) or discrete cell covariate **Figure 2.10 a**). For more details, see **chapter 3** section 3.1.2 and **chapter 5** Appendix B.

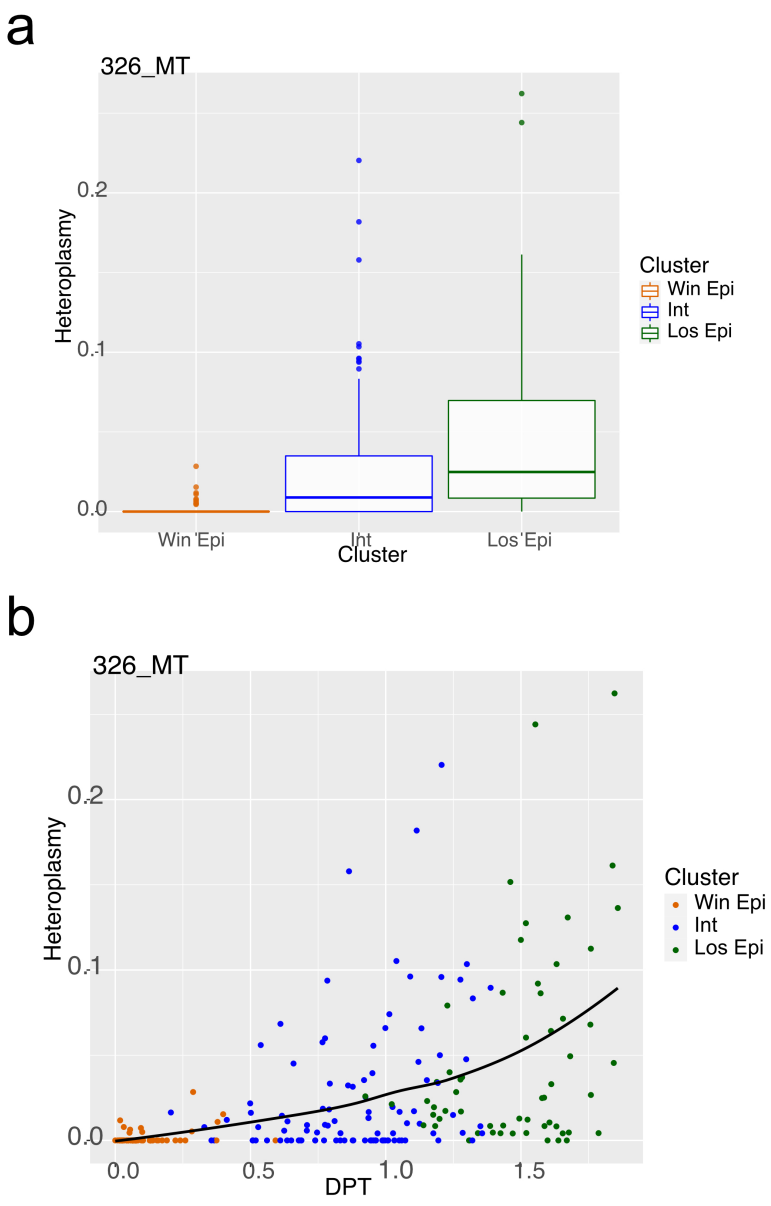


Figure 2.10: **Overview of MitoHEAR**
a) This is an example of a plot generated by MitoHEAR, which shows heteroplasmy values at a specific mtDNA position estimated from single cells found in three transcriptional clusters indicated on the x-axis. The heteroplasmy is defined as one minus the frequency of the most common allele. The scRNA-seq data used is from mouse embryo at the embryonic day 6.5 [71]).
b) This is an example of a figure generated by MitoHEAR where the change in mtDNA heteroplasmy is plotted and analyzed as a function of the diffusion pseudo-time coordinate (DPT) of each cell. Cells are classified into three clusters. The heteroplasmy shows a statistically significant change along the DPT, as indicated by the adjusted p-value reported at the top, which is computed by a generalized additive model fit. Data from [71].
Abbreviated cluster names: "Win Epi" stands for "Winner Epiblast"; "Int" for "Intermediate Epiblast"; "Los Epi" for "Loser Epiblast".

Chapter 3

Summary of contributed articles

This chapter includes a summary of the papers I have contributed to during my PhD.
The following manuscripts are available as published article (see **chapter 5** for detailed informations):

- **Cell competition acts as a purifying selection to eliminate cells with mitochondrial defects during early mouse development**.
- **MitoHEAR: an R package for the estimation and downstream statistical analysis of the mitochondrial DNA heteroplasmy calculated from single-cell datasets**.
- **CIARA: a cluster-independent algorithm for identifying markers of rare cell types from single-cell sequencing data**.

The following manuscripts are under review or currently in preparation and a preprint is not available:

- **Human embryo organoids as a model for peri-and post-gastrulation embryo development**.
- **Single cell dissection of pluripotency states across mammals**.
- **Cell competition as a mechanism to control embryo size at the onset of gastrulation**.

The first section of this chapter is about the papers that are already published, while the second section is about the papers that are not yet published (they are at the moment submitted or under preparation).

3.1 Published papers

3.1.1 Cell competition acts as a purifying selection to eliminate cells with mitochondrial defects during early mouse development

The results reported below are part of the following peer-reviewed publication.

- Ana Lima*, **Gabriele Lubatti***, Jorg Burgstaller, Di Hu, Alistair Green, Aida Di Gregorio, Tamzin Zawadzki, Barbara Pernaute, Elmir Mahammadov, Salvador Perez Montero, Marian Dore, Juan Miguel Sanchez, Sarah Bowling, Margarida Sancho, Mohammed Karimi, David Carling, Nick Jones, Shankar Srinivas, Antonio Scialdone, Tristan A Rodriguez. **Cell competition acts as a purifying selection to eliminate cells with mitochondrial defects during early mouse development**. Nat Metab 3, 1091–1108 (2021). `https://doi.org/10.1038/s42255-021-00422-7`.
  Contributions of the author: I lead the bioinformatics analysis of single cell and bulk RNA seq datasets generated in the study.

*These authors contributed equally

Summary

Cell competition is the process whereby less fit cells get eliminated when cultured together with fitter cells [12]. Although there are many in vitro systems for cell competition, little is known about cell competition in mouse embryos in physiological conditions. In the mouse embryo, 35% of pluripotent embryonic cells are eliminated between embryonic day (E) 5.5 and 6.5 ([11]).

The time window between E5.5 and E6.5 is a crucial time point. During this time window the embryo is preparing to face one of the most important stages of development: gastrulation. It is very important to select only cells of sufficient quality, otherwise, all successive steps of development could be irreversibly compromised. It has been observed that around 35% of epiblast cells die just before gastrulation. It has been speculated that cellular competition might be the reason for that, but it is not know what distinguishes winner from loser epiblast cells and what the function of cell competition might be at this stage of development. To address these questions, we first showed that cell competition occurs at the onset of gastrulation, then we analyzed what sets winner cells apart from loser cells, in terms of transcriptional features and how the transition from loser to winner happens.

In our project, we teamed up with an experimental lab that is expert in cell competition. Our collaborators cultured embryos in two different media: one in which apoptosis was blocked (CI medium) and the other in which apoptosis occurred normally (DMSO medium). Subsequently, we ran scRNA-seq to examine the transcriptional profiles of the cells in these two groups of embryos. Our analysis revealed a cluster that expressed epiblast markers and that was made up only by CI treated cells. This cluster consisted of loser cells that would typically be eliminated in cell competition. To establish a relationship between the winner and loser epiblast cells, we constructed a trajectory using diffusion map analysis from the epiblast cell population. Subsequently, we assigned a "losing score" to each cell and identified genes that were differentially expressed along the trajectory. We used the concepts of diffusion maps and pseudo-time [57] to

transcriptionally identify a "losing state" and quantify the distance of a cell's transcriptome to the losing state. The genes that are more highly expressed at the end of the trajectory, where the losing scores of the cells are higher, are enriched for P53 and its activated target. Among the genes downregulated at the end of the trajectory, there are *Myc* and *mTOR*. This is very well aligned with signatures of losing cells that arise in the in vitro models of cell competition (low *Myc* and *mTOR* expression and high P53) [14], [13].
Another striking enrichment among genes downregulated in the "losing region" are the mitochondrial genes.
There is strong evidence for selection against aberrant mitochondrial function induced by deleterious mtDNA mutations in mammals [72], [73]. Given our observation that cell competition selects against cells with impaired mitochondrial function, we investigated if cell competition could be reducing mtDNA heteroplasmy (frequency of different mtDNA variants) during mouse development.
We therefore tested if mtDNA heteroplasmy is present in our scRNA-seq data and whether this correlates with the losing score of a cell.
To detect mtDNA heterolasmy we used the tool MitoHEAR that we have previously developed. Our analysis revealed that the frequency of specific mtDNA polymorphisms increases with the losing score of epiblast cells. These mtDNA changes occur mainly within mt-Rnr1 and mt-Rnr2 genes and are not dependent on the litter from which the embryos come. Therefore, the presence of these specific mtDNA mutations in the loser cells suggests that cell competition could be contributing to the elimination of deleterious mtDNA mutations during early mouse development.

Contribution

I led the bioinformatics analysis of the single-cell and bulk RNA seq datasets generated in the study. In particular: data processing, quality control and normalization; highly variable genes selection and dimensionality reduction; cell clustering; identification of a single-cell trajectory in the epiblast; differential gene expression analysis along the trajectory; analysis of mitochondrial heteroplasmy from a single-cell RNA-seq dataset.
For the analysis of heteroplasmy in the single-cell RNA sequencing dataset, I developed and ran a pipeline for the identification of mitochondrial DNA mutations from RNA sequencing data that exclude the presence of possible artefacts. Since I inferred mtDNA changes from RNA-seq data, I considered additional potential sources for the sequence changes I observed. One possible source is contamination from NUMTs (nuclear mitochondrial sequences). To address this artefact I considered only the RNA-seq reads that are uniquely mapped to the mitochondrial genome and not to nuclear DNA, and I confirmed that the variants with highest heteroplasmy found in the 'loser' cells were not present in any of the NUMTS that have previously been reported or could be identified using BLAST. Only reads that uniquely mapped to the mtDNA were considered. From these, I obtained allele counts at each mtDNA position with a Phred quality score greater than 33 using the samtools 'mpileup' function. Next, I applied filters to remove cells and mtDNA positions with a low coverage. First, I removed cells with fewer than 2,000 mtDNA positions covered by more than 50 reads. Second, I removed positions having less than 50 reads in more than 50% of cells in each of the three epiblast clusters (winner, intermediate and loser). Starting from these cells and positions, I applied an additional filter to keep only positions with a sufficiently high level of heteroplasmy. To this aim, for

each position with more than 50 reads in a cell, I estimated the heteroplasmy according to equation:

$$H = 1 - f_{max} \tag{3.1}$$

where f_{max} is the frequency of the most common allele. I kept only positions with H>0.01 in at least ten cells.
Finally, using a generalized additive model (GAM) , I identified the positions whose heteroplasmy H changes as a function of the cells' losing score in a statistically significant way. I found a total of eleven significant positions (FDR<0.001), six of them in mt-Rnr1 and five in mt-Rnr2. All of these positions had a higher level of heteroplasmy in loser cells . The results remain substantially unaltered if the Spearman's rank correlation test (as opposed to the GAMs) is used.
For the bulk datasets, I performed differential expression analysis between the two given conditions and the gene ontology enrichment analysis.
I generated all the figures shown in the article related to single-cell/bulk RNA sequencing data analysis.

3.1.2 MitoHEAR: an R package for the estimation and downstream statistical analysis of the mitochondrial DNA heteroplasmy calculated from single-cell datasets

The results reported below are part of the following peer-reviewed publication.

- **Gabriele Lubatti**, Elmir Mahammadov, Antonio Scialdone. (2022). MitoHEAR: an R package for the estimation and downstream statistical analysis of the mitochondrial DNA heteroplasmy calculated from single-cell datasets. Journal of Open Source Software, 7(74), 4265. `https://doi.org/10.21105/joss.04265`.

Summary
Although mtDNA heteroplasmy has important consequences for human health [68] and embryonic development [74], there are still many open questions about how heteroplasmy affects cells' ability to function and how cells keep it under control. With the increasing availability of single-cell data, many questions can begin to be answered. Several single-cell sequencing protocols provide the data needed to estimate mtDNA heteroplasmy, including single-cell DNA-seq, RNA-seq, and ATAC-seq, as well as dedicated protocols like MAESTER [75]. However, it is essential to have efficient and streamlined computational tools that enable researchers to estimate and analyze mtDNA heteroplasmy. Existing packages ([76], [77], [78]) focus only on the first step of quantifying heteroplasmy from BAM files and do not provide any specific tools for further statistical analyses or plotting.
MitoHEAR covers all steps of the analysis in a unique user-friendly package, with highly customizable functions. Starting from BAM files, MitoHEAR applies filters to remove cells and mtDNA positions with low coverage. Starting from these cells and positions, MitoHEAR applies an additional filter to keep only positions with a sufficiently high level of heteroplasmy. MitoHEAR also offers several options for downstream analyses. For example, statistical tests are provided to investigate the relationship between mtDNA heteroplasmy and continuous (i.e., diffusion pseudo time) or discrete (i.e., cluster) cell covariates. Moreover, it includes plotting functions to visualize heteroplasmy and allele frequencies and to perform hierarchical clustering of cells based on heteroplasmy values.
Contribution
I was responsible for the conceptual design and implementation of the algorithm, from conceptualization to the creation of the corresponding R library on CRAN and the development version on GitHub. Additionally, I was responsible for documenting all functions included in the library, creating all figures included in the paper, and developing the package vignette.
Software availability
MitoHEAR is available as R package on CRAN (`https://CRAN.R-project.org/package=MitoHEAR`) and on github (`https://github.com/ScialdoneLab/MitoHEAR`).

3.1.3 CIARA: a cluster-independent algorithm for identifying markers of rare cell types from single-cell sequencing data

The results reported below are part of the following peer-reviewed publication.

- **Gabriele Lubatti**, Marco Stock, Ane Iturbide, Mayra L. Ruiz Tejada Segura, Melina Riepl, Richard Tyser, Anna Danese, Maria Colomé-Tatché, Fabian J. Theis, Shankar Srinivas, Maria-Elena Torres-Padilla, Antonio Scialdone. CIARA: a cluster-independent algorithm for the identification of markers of rare cell types from single-cell sequencing data. Development 2023.`https://doi.org/10.1242/dev.201264`.

Summary
The development of single-cell transcriptomics has facilitated the molecular characterization of cell types in numerous organs and tissues across different organisms. One of the key aims of single-cell studies is to identify rare cell types that bulk techniques are unable to access. However, some types of cells are difficult to identify since they are rare and share markers with more abundant cell types. An example of such cells are primordial germ cells that share markers with cells from the primitive streak([64]; [79]).
The identification of cell types is performed by unsupervised clustering, typically using highly variable genes([24]). While this approach can successfully identify large clusters of well-defined cell types, it often fails to detect small-sized clusters of cells with fewer specific marker genes. Consequently, many algorithms have been developed to specifically detect rare cell types in scRNA-seq data[80] [81] [32]. However, these algorithms are generally more efficient in selecting rare cells with strong markers than in identifying very small cell populations($<$1%) with only a limited number of specific markers. Furthermore, some of these methods have a tendency to overfit and identify a large number of small cell clusters without specific markers. To address these challenges, we developed a novel algorithm called CIARA (Cluster Independent Algorithm for the identification of markers of RAre cell types) that identifies potential marker genes of rare cell types by leveraging their property of being highly expressed in a small number of cells with similar transcriptomic signatures. CIARA ranks genes based on their enrichment in local neighborhoods defined from a K-nearest neighbors (KNN) graph. The top-ranked genes can then be used with standard clustering algorithms to identify groups of rare cell types with high efficiency, requiring the specification of a minimal number of parameters.
We demonstrate how CIARA outperforms existing algorithms for rare cell type identification using scRNA-seq datasets from different organisms and scRNA-seq protocols. We also apply CIARA to detect rare cells in a recently published dataset from a human gastrula ([64]), where we identify several groups of rare cells. Furthermore, we show that CIARA can also be applied to atlas-sized datasets. We process two datasets including approximately $\sim 10^5$ cells with CIARA, leading to the identification of several potential rare populations of cells expressing very specific markers.
The main requirement for CIARA is the definition of a KNN graph, which can be built using any type of data where a notion of distance is defined. Therefore, its applicability is broad, and it can be used to identify rare populations of cells

across multiple data modalities, such as DNA-seq, ATAC-seq, bisulfite sequencing, etc. As proof of principle, we ran CIARA on a paired scRNA/ATAC-seq dataset generated from mouse skin cells with the SHARE-seq protocol.

Contribution

I was responsible for the development and implementation of all functions of the algorithm in R until the creation of the corresponding R library on CRAN and the development version on GitHub. I compared CIARA with previously existing methods. I applied CIARA to several published single-cell datasets and an unpublished mouse embryonic stem cell dataset. To assess the performance of each algorithm, I used the Matthew correlation coefficient (MCC) to measure the agreement between the classification of rare cells and the ground truth classification. MCC is a metric that quantifies the overall agreement between two binary classifications, taking into account both true and false positives and negatives. Furthermore, I contributed to the writing of the article, including drafting the first version and reviewing the text.

Lastly, Melina Riepl was an intern in the lab that I supervised. Under my guidance, she worked on applying CIARA to atlas-sized datasets.

Software availability

CIARA is available as R package on CRAN (`https://CRAN.R-project.org/package=CIARA`) and on github (`https://github.com/ScialdoneLab/CIARA`). The python implementation of CIARA (`https://github.com/ScialdoneLab/CIARA\_python`) was developed by Marco Stock.

3.2 Unpublished papers

3.2.1 Human embryo organoids as a model fsubor peri- and post-gastrulation embryo development

The results reported below are part of the following publication currently under revision in Nature.

- Jitesh Neupane*, **Gabriele Lubatti***, Mayra Luisa Ruiz Tejada Segura, Sabine Dietmann, Antonio Scialdone, Azim Surani. **Human embryo organoids as a model for peri-and post-gastrulation embryo development**.

*These authors contributed equally

Summary

The foundation of the human body plan is established in the first few weeks of development, specifically during gastrulation. At this stage, the epiblast cells invaginate and undergo an epithelial to mesenchymal transition (EMT), forming the primitive streak (PS) and giving rise to mesoderm and endoderm. Additionally, primordial germ cells (PGCs), which are the precursors of sperm and oocytes, also appear during gastrulation. Following gastrulation, neuronal precursors initiate and organogenesis begins.

As human embryos at this stage are highly inaccessible, animal studies and in vitro models using human pluripotent stem cells (hPSCs) have been used to gain insight into early human development. 2D micropatterned embryo models ([82]) and 3D gastruloids ([83]) have been developed, but none of these models fully recapitulate post-gastrulation human development. However, they have provided opportunities for mechanistic studies on specific aspects of human gastrulation.

In this context, we present a novel human embryonic organoids (hEO) model for early human development derived entirely from human embryonic stem cells (hESCs). Our model recapitulates critical features of both human gastrulation and post-gastrulation embryos, including the establishment of the three germ layer derivatives and the specification of hPGCLCs without exogenous bone morphogenic protein (BMP) supplementation. Furthermore, we detected the presence of neural tube, neural crest cells, and neuronal precursors, which are linked to neuromesodermal progenitors (NMPs). These findings demonstrate that our hEOs model is the most complete model of early human embryo development to date and has the potential to yield new insights into still-unanswered questions surrounding human gastrulation and post-gastrulation events.

Contribution

I led the bioinformatics analysis of all the scRNA-seq datasets, which included $\sim$ 60,000 cells sampled from 5 time points during human embryo organoid development. In particular: data alignment and quality control; clustering analysis; batch correction and data visualization; projection of the scRNA-seq datasets from human embryo organoids onto a recently published scRNA-seq data from a human gastrula [64] to investigate similarities and differences between this in vitro embryo model and the real embryo; mapping clusters across time points.

I used the WOT algorithm [66] to find the most likely differentiation trajectories between clusters across time points. For each cell at time t, this algorithm

assigns a "bias score" toward every cluster at time t+1 with the function compute_all_transport_maps. For each cell, the sum of the scores towards all the clusters at time t+1 sum to 1. Each cluster of cells at time t gets assigned a score equal to the average scores of all cells in that cluster. I applied this procedure to identify the most likely precursor cells of the clusters spinal cord, neural tube and neural crest at day 8. Finally, I visualized the results by showing for each cluster the connections between each time point are shown using the R library WOTPLY based on the R library GGally (version 2.1.2) [84] The thickness of the edges between clusters at time points t and t+1 represents the average score, and only the edges with an average score above 0.2 are shown.

3.2.2 Single cell dissection of pluripotency states across mammals

The results reported below are part of the following paper, currently under preparation.

- **Gabriele Lubatti***, Marion Genet*, Maria Elena Torres Padilla, Antonio Scialdone. **Single cell dissection of pluripotency states across mammals**.

*These authors contributed equally

Summary

Pluripotency enables cells to generate all tissues in the adult body, including the germline. Pluripotent cells emerge during early development, but they exist only transiently since they subsequently undergo lineage allocation to generate all the germ layers. Our understanding of the molecular regulatory pathways that sustain pluripotency over the last few decades has come from genetics and developmental biology approaches, but also from studies of embryonic stem cell models in culture, which recapitulate many of the features of their in vivo counterparts. Embryonic stem cells (ESCs) in culture are initially derived from the inner cell mass (ICM) of the mouse blastocyst and can self-renew under appropriate culture conditions [7]. While this is a gold standard in the mouse field, these features are not easily testable in other mammalian species, and therefore the degree to which cells in other species are pluripotent is unclear. A deeper molecular description and understanding of the models is therefore needed to evaluate and fully comprehend the underpinnings of pluripotency in mammalian stem cells. The reprogramming of somatic cells into induced pluripotent stem cells (iPSCs) has opened tremendous opportunities to further manipulate cellular plasticity and, most importantly, has enabled the possibility of generating pluripotent cells from many differentiated cells without the need to generate embryos. In addition to ESCs derived from the ICM, pluripotent stem cell lines have also been derived from the early post-implantation epiblast [85]. In the mouse, these epiblast stem cells (EpiSCs) are epigenetically and transcriptionally different from ESCs.

To map the common and divergent features of pluripotent stem cell models across species, we undertook an in-depth transcriptomics approach to characterize eight different cell lines in five mammalian species. We used ESCs, iPSCs, and EpiSCs of the highest quality, which have been thoroughly characterized and validated for their pluripotency features [86], [87], [88], [89], [90]. Specifically, we focused on the most commonly used pluripotent stem cell models in the mouse, rhesus monkey, rabbit, cow, and pig. We used the 10x Genomics platform for single-cell RNA sequencing and sequenced a total of 125,605 individual cells, which, after stringent quality control, led to 15,615 murine cells, 20,013 rhesus cells, 14,605 rabbit cells, 3,622 bovine cells, and 14,282 porcine cells.

We first investigated the patterns of cell state heterogeneity across stem cell cultures. Next, we directly compared the similarities or differences between the cell clusters in embryo-derived ESCs or EpiSCs against reprogrammed iPSC lines of the same species. Because of the differences in the transcriptional signatures that we uncovered between the embryonic pluripotent stem cell models above,

we quantified the extent to which the different stem cell models recapitulate the properties of their in vivo counterparts. This is especially important considering the known different pluripotency states within the blastocyst across different mammalian species, from which embryo-derived stem cells emerge. Thus, we performed a systematic computational comparison between cellular identities present in in vivo development and all pluripotent stem cell models in culture. Finally, to further investigate the molecular makeup of the studied stem cell lines, we conducted a gene regulatory network (GRN) analysis based on the single-cell RNA sequencing datasets and primarily focused on DNA-binding proteins. Our focus on DNA-binding proteins was based on the understanding that the pluripotency state depends on the action of a gene regulatory network (GRN), which is primarily characterized by transcription factors (TFs) such as NANOG and OCT4.
Our analysis revealed several surprising global differences across species and stem cell lines, both in terms of the number of transcriptional hubs identified and their specific molecular composition.

Contribution

I conducted all the bioinformatics analyses presented in the paper. Specifically, I performed the following tasks: alignment and quality control of single-cell RNA sequencing data; cluster analysis and identification of marker genes; filtering of lower quality cells; projection of stem cell lines data intra-species; projection of stem cell lines data into embryonic data; assignment of zygote genome activation (ZGA) scores to each cell from the in vitro datasets; gene regulatory network analysis.
For the projection of stem cell lines data intra-species, I defined a centroid for each cluster as the mean expression value of highly variable genes across cells in the cluster. The highly variable genes were the union of ESCs and iPSCs HVGs. To compare ESCs and iPSCs, I generated a distance matrix based on Spearman correlation between the cluster centroids. I then computed the Spearman's correlation-based distance between the cells and their respective centroids, and assigned an empirical p-value to each cluster based on a comparison with an empirical distribution. The final empirical p-value for each pair of clusters was defined as the maximum between the empirical p-value assigned to cluster A and the empirical p-value assigned to cluster B. The distance between a pair of clusters was considered significant if the final empirical p-value was smaller or equal to 0.10.
To assign an absolute score between an embryonic stage (species-specific embryo data) and every cluster in the corresponding stem cell lines (our 10x data), I developed the R package SCOPRO, which is available on CRAN. For the zygote genome activation (ZGA) score analysis, I computed ZGA markers using the CIARA function markers_cluster_seurat from the published embryonic dataset of each species. For each species, I filtered out genes without a 1:1 human orthologous name using the orthology search option from gprofiler (`https://biit.cs.ut.ee/gprofiler/orth`). I kept only the genes expressed (> 0) in at least 10% of cells in one or more clusters in ESCs/iPSCs. For each ZGA marker found with the CIARA function, I computed a Z-score for each cell in the in vitro dataset. I kept only ZGA markers with a Z score above 10 in at least 2 cells for downstream analysis. Finally, the ZGA score for each cell was computed as the number of kept ZGA markers expressed at a level above 2 log norm counts.

For the gene regulatory network analysis, I built the network using the GENIE3 function from the R package GENIE3 (version 1.10.0) [91], with the regulators parameter set to the list of mouse transcription factors downloaded from the AnimalTFDB website (`http://bioinfo.life.hust.edu.cn/AnimalTFDB/#!/`). For species other than human, I translated the genes that have a 1:1 human orthologous name using the orthology search option from gprofiler.

3.2.3 Cell competition as a mechanism to control embryo size at the onset of gastrulation

The results reported below are part of the following paper, currently under preparation.

- **Gabriele Lubatti**, Tobias Krauß, Antonio Scialdone. **Cell competition as a mechanism to control embryo size at the onset of gastrulation**.

Summary
In mouse embryos, the time window between embryonic day (E) 5.5 and 6.5 is a crucial point in development, as the embryo is preparing for gastrulation - one of the most important stages of development. During this time window, there is an increase in the proliferation rate (more cells are being produced) but also an increase in the apoptosis rate (more cells are dying) due to cell competition. We explored, through mathematical models, how these two processes might be related and how cellular competition might help embryos control their size (**Figure 3.1**).
Despite the large amount of available data, at different levels (morphological, molecular, scRNA-seq), there are very few attempts to study cell competition in mouse embryo development from a mathematical modeling point of view ([92]). Our model began with a mean-field, ordinary differential equation-based approach to model changes in epiblast cell numbers in the embryo. We assumed that there are two sub-populations of cells in our embryos: the good and the bad. The good cells are the ones that win the cell competition while the bad cells are the losing cells. The bad cells are generated from the good ones with a probability P, which is the fraction of the total number of cells generated from the good cells in a unit of time that will become losers. In the general setting, we considered a different rate of proliferation for unit time for both the winner and loser cells. The number of cells eliminated by the good cells in a unit of time is regulated by the parameter D. To understand the role of cell competition in mouse embryo development, we started with the following system of two ordinary differential equations (ODE):

$$\begin{cases} \frac{dN_w}{dt} = R_w N_w - P N_w \\ \frac{dN_l}{dt} = P N_w + R_l N_l - D N_w N_l \end{cases} \tag{3.2}$$

Where:

- N_w is the number of winner cells.
- N_l is the number of loser cells.
- R_w is the rate of proliferation of the winner cells.
- R_l is the rate of proliferation of the loser cells.
- D represents the number of cells that are eliminated through apoptosis due to cell competition within a given time unit.
- P is the probability of becoming a loser cell.

The parameters of the model are inferred from the scRNA-seq data analysis I performed (i.e. cell cycle analysis and fraction of winner and loser cells).
For a better understanding of the role of fluctuations in the parameters of the model, we switched to a modelling framework that allowed us to study fluctuations, namely stochastic differential equations. Through simulation and analytical results, we demonstrated that coupling together the proliferation and apoptosis rates led to an incoherent feedforward circuit with a noise-buffering effect on the final number of cells.

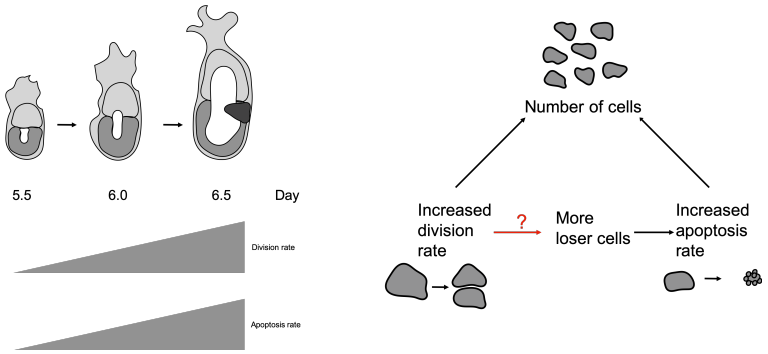


Figure 3.1: **Cell competition might help regulate embryo size at the onset of gastrulation.** The left panel illustrates mouse embryos between 5.5 and 6.5 days of development, when the division rate and the apoptosis rate of epiblast cells both increase The right panel shows a schematics of how the increase of cell division and apoptosis rates have opposite effects on the total number of cells. The red arrow indicates the connection we hypothesise in our model between the emergence of loser cells and the increase in division rate. If such a connection existed, an incoherent feed forward loop emerges, allowing the embryo to limit fluctuations in cell number

Contribution

I designed the mathematical model and led all the analysis in terms of simulation and mathematical modelling. Tobias Krauß was an internship student in the lab that I supervised. Under my supervision, he worked on the mathematical implementation of the stochastic part of the model.

Chapter 4

Discussion

In this chapter, I will provide a summary of the key research questions I have addressed during my PhD and explore potential future directions in the field.

4.1 Analysis of communication between winner and loser cells using spatial transcriptomics and mathematical modelling

The dynamics of the interactions between winner and loser cells are not well understood. In a recent work, it has been shown that embryonic cells lacking the tumour suppressor p53 are super-competitors that eliminate their wild-type neighbours through the direct induction of apoptosis. By combining mathematical modelling and cell-based assays, it has been shown that the elimination of wild-type cells is not through a competition for space or nutrients, but instead is mediated by short-range interactions that are dependent on the local cell neighborhood ([93]). To better characterize these short-range interactions, an intriguing analysis is to identify genes responsible for the communication between winner and loser cells, which ultimately leads to cell competition and subsequent cell death. This task can be accomplished using one of several algorithms ([94], [95]) developed in recent years for the analysis of cell communication in scRNA-seq datasets.
In addition to detecting genes responsible for the communication, a complementary promising approach to better investigate the cell-cell interactions between losers and winners is to look at the spatial distribution of losers/winners with spatial transcriptomics. The information obtained from spatial transcriptomics can be used in a mathematical modelling framework to check what kind of cell-cell interactions can best explain the data.
In order to have better predictions regarding the quantitative role of cell competition in regulating embryo size, we can further extend our model.
An approach would be introducing rate of proliferation and apoptosis rate that are functions of time and not just constant values and/or adopting a Lotka-Volterra model that does not allow the system to indefinitely growth (since in the real biological system there are limits induced by space or resources which

need to be addressed by the model).

4.2 Extend CIARA to spatial transcriptomics and grouping cells together according to top localized markers

CIARA allows identifying the markers of rare cell types in a completely unsupervised way, by selecting genes based on their enrichment in local neighborhoods defined from a K-nearest neighbors (KNN) graph. Then these genes are used as features for clustering cells with standard cluster algorithm. A limitation in cluster analysis is that the number of clusters or a parameter linked to the number of clusters (i.e. resolution) must be provided as input. Although several criteria have been proposed for setting this parameter, to some extent, the choice remains arbitrary.
For CIARA, we are working on developing a method that allows us to define groups of rare cells using directly the list of top rare cell type markers, without doing standard cluster analysis on the dataset. Therefore, we would avoid the limitation intrinsic to standard cluster analysis.
Since the only input required by CIARA are a KNN graph and a normalized count matrix, an interesting scenario for CIARA is to extend the applicability to other omics sequencing techniques. In particular, we are interested in applying it to spatial transcriptomic datasets. When applied to spatial transcriptomics data, CIARA can provide outputs that include both rare cell populations as well as spatially localized groups of cells. This allows for the identification of both infrequent cell types and regions within the spatial context of the tissue where these cell populations are located.

4.3 Identifying similarities and differences between embryo organoids and real embryos with machine learning approaches

In our project, we conducted single-cell RNA sequencing to characterize human embryo organoid samples at various time points. An intriguing perspective that has received limited attention thus far is to explore the differences, not just the similarities, between embryo organoids and real embryos. To accomplish this, we intend to utilize our developed tool, SCOPRO, and refine it to identify genes that are conserved or not conserved between embryo organoids and real embryos. Specifically, by using real embryos as the reference and embryo organoids as the query, we will assess the similarity of cell types between the two systems based on the overall score provided by SCOPRO. Subsequently, for each specific cell type found in both the reference and query, we will examine the set of genes that are conserved and those that are specific to either the reference or query cell type. This analysis will shed light on the similarities and differences at the transcriptional level between embryo organoids and real embryos.

4.4 Retrotransposons expression in pluripotent cell states across mammals

Transposable elements (TEs) are mobile DNA sequences that propagate within the genome and constitute a large fraction of most eukaryotic genomes [96]. In our project, we analyzed the heterogeneity in stem cells from several mammalian species (mouse, rabbit, pig, cow, and monkey) using a count matrix where the features of each cell are genes. An interesting question would be to quantify the expression of transposable elements in the stem cells from rabbit, cow, pig and monkey that we have available and then make a comparison with the TEs expressed in published embryo datasets from the corresponding species. The goal is to investigate whether there are common patterns of retrotransposon expression between species and between in vitro and in vivo models of pluripotency.
TEs are commonly divided into three categories [97]:

1. Retrotransposons with long terminal repeats (LTRs)
2. Retrotransposons with long interspersed nuclear elements (LINEs)
3. Short interspersed nuclear elements (SINEs)

Retrotransposon-associated reads are often discarded in sequencing data analyses because of the uncertainty in attributing ambiguously short sequencing reads to highly repetitive regions of the genome. In recent years, new computational tools were developed to quantify the expression of retrotransposon elements starting from single-cell RNA sequencing data like TEtranscripts [98],SalmonTE [99] and scTE [47] . The first two tools (TEtranscripts and SalmonTE) are designed to work with plate-based methods, but not with droplet based methods. On the contrary scTE is suitable for both droplet and plate-based methods.
It is known that in mouse embryo at the late 2-cells stage, a network of genes (in particular the *Zscan4* genes family) that use the retrotransposon with long terminal repeats MuERV-L- as alternative promoters are activated. MuERV-L and the Zscan4 genes are expressed only in this stage of mouse embryo development [6].
It has been shown that the same pattern (high expression of MuERV-L and Zscan4 genes) is present in the 2-cell embryo like cells (2CLC) from mouse embryonic stem cells [15].
The CIARA algorithm that I have previously developed could be run to verify the presence of rare cell states or types that express specific group of retrotransposons.

Chapter 5

Appendices

5.1 Appendix A: Cell competition acts as a purifying selection to eliminate cells with mitochondrial defects during early mouse development

Core publication as main author

- Ana Lima*, **Gabriele Lubatti***, Jorg Burgstaller, Di Hu, Alistair Green, Aida Di Gregorio, Tamzin Zawadzki, Barbara Pernaute, Elmir Mahammadov, Salvador Perez Montero, Marian Dore, Juan Miguel Sanchez, Sarah Bowling, Margarida Sancho, Mohammed Karimi, David Carling, Nick Jones, Shankar Srinivas, Antonio Scialdone, Tristan A Rodriguez. Cell competition acts as a purifying selection to eliminate cells with mitochondrial defects during early mouse development. Nat Metab 3, 1091–1108 (2021). `https://doi.org/10.1038/s42255-021-00422-7`.

*These authors contributed equally

This is the published version of the article in Nature Metabolism following peer review.

nature metabolism

ARTICLES

https://doi.org/10.1038/s42255-021-00422-7

Check for updates

Cell competition acts as a purifying selection to eliminate cells with mitochondrial defects during early mouse development

Ana Lima[1,2,14], Gabriele Lubatti[3,4,5,14], Jörg Burgstaller[6], Di Hu[7], Alistair P. Green[8], Aida Di Gregorio[1], Tamzin Zawadzki[1], Barbara Pernaute[1,9], Elmir Mahammadov[3,4,5], Salvador Perez-Montero[1], Marian Dore[2], Juan Miguel Sanchez[1,10], Sarah Bowling[1], Margarida Sancho[1], Thomas Kolbe[11,12], Mohammad M. Karimi[2,13], David Carling[2], Nick Jones[8], Shankar Srinivas[7], Antonio Scialdone[3,4,5,15] ✉ and Tristan A. Rodriguez[1,15] ✉

Cell competition is emerging as a quality-control mechanism that eliminates unfit cells in a wide range of settings from development to the adult. However, the nature of the cells normally eliminated by cell competition and what triggers their elimination remains poorly understood. In mice, 35% of epiblast cells are eliminated before gastrulation. Here we show that cells with mitochondrial defects are eliminated by cell competition during early mouse development. Using single-cell transcriptional profiling of eliminated mouse epiblast cells, we identify hallmarks of cell competition and mitochondrial defects. We demonstrate that mitochondrial defects are common to a range of different loser cell types and that manipulating mitochondrial function triggers cell competition. Moreover, we show that in the mouse embryo, cell competition eliminates cells with sequence changes in mt-Rnr1 and mt-Rnr2, and that even non-pathological changes in mitochondrial DNA sequences can induce cell competition. Our results suggest that cell competition is a purifying selection that optimizes mitochondrial performance before gastrulation.

Cell competition is a fitness-sensing mechanism that eliminates cells that, although viable, are less fit than their neighbours. The cells that are eliminated are generically termed losers, while the fitter cells that survive are referred to as winners. Cell competition has been shown to act in a broad range of settings, from the developing embryo to the ageing organisms[1–3]. It has been primarily studied in *Drosophila*, where it was first described in the imaginal wing disc[4]. Since then, it has also been found to be conserved in mammals. In the mouse embryo, 35% of embryonic cells are eliminated between embryonic day (E) 5.5 and E6.5, and strong evidence suggests that this elimination is through cell competition[5–7]. These and other studies identified a number of read-outs of cell competition in the mouse embryo, such as relative low c-MYC expression, a loss of mTOR (mammalian target of rapamycin) signalling, low TEAD transcription factor activity, high P53 expression or elevated levels of ERK phosphorylation[5–9]. Importantly, there is a substantial overlap with the markers of cell competition originally identified in *Drosophila* as well as those found in other cell competition models, such as Madin–Darby canine kidney cells[1–3]. Despite the advance that having these cell competition markers signifies, given that they were primarily identified by using genetic models that rely on over-expression or mutation, we still have little insight into the overarching features of the cells that are eliminated in the physiological context.

Mitochondria, with their diverse cellular functions ranging from determining the bioenergetic output of the cell to regulating its apoptotic response, are strong candidates for determining competitive cell fitness. During early mouse development, mitochondria undergo profound changes in their shape and activity[10]. In the pre-implantation embryo, mitochondria are rounded, fragmented and contain sparse cristae, but after implantation they fuse to form complex networks with mature cristae[11]. The mode of replication of mitochondrial DNA (mtDNA), which encodes vital components of the bioenergetic machinery, also changes during early mouse development. After fertilization, mtDNA replication ceases and its copy number per cell decreases with every division until the post-implantation stages, when mtDNA replication resumes[10]. As the mutation rate of mtDNA is much higher than that of nuclear DNA[12,13], this increased replication most likely leads to an increased mutation load. In fact, inheritable mtDNA-based diseases are reported with a prevalence of 5–15 cases per 100,000 individuals[14,15]. A number of mechanisms have been proposed to reduce this mutation load, such as the bottleneck effect, purifying selection or biased segregation of mtDNA haplotypes[16–21]. However, how these

[1]National Heart and Lung Institute, Imperial College London, London, UK. [2]MRC London Institute of Medical Sciences (LMS), Institute of Clinical Sciences, Imperial College London, London, UK. [3]Institute of Epigenetics and Stem Cells, Helmholtz Zentrum München, Munich, Germany. [4]Institute of Functional Epigenetics, Helmholtz Zentrum München, Neuherberg, Germany. [5]Institute of Computational Biology, Helmholtz Zentrum München, Neuherberg, Germany. [6]Institute of Animal Breeding and Genetics, University of Veterinary Medicine, Vienna, Austria. [7]Department of Physiology, Anatomy and Genetics, University of Oxford, Oxford, UK. [8]EPSRC Centre for the Mathematics of Precision Healthcare, Department of Mathematics, Imperial College London, London, UK. [9]Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Barcelona, Spain. [10]Orchard Therapeutics, London, UK. [11]Biomodels Austria (Biat), University of Veterinary Medicine Vienna, Vienna, Austria. [12]Department IFA-Tulln, University of Natural Resources and Life Sciences, Vienna, Austria. [13]Comprehensive Cancer Centre, School of Cancer & Pharmaceutical Sciences, Faculty of Life Sciences & Medicine, King's College London, London, UK. [14]These authors contributed equally: Ana Lima, Gabriele Lubatti. [15]These authors jointly supervised this work: Antonio Scialdone, Tristan A. Rodriguez. ✉e-mail: antonio.scialdone@helmholtz-muenchen.de; tristan.rodriguez@imperial.ac.uk

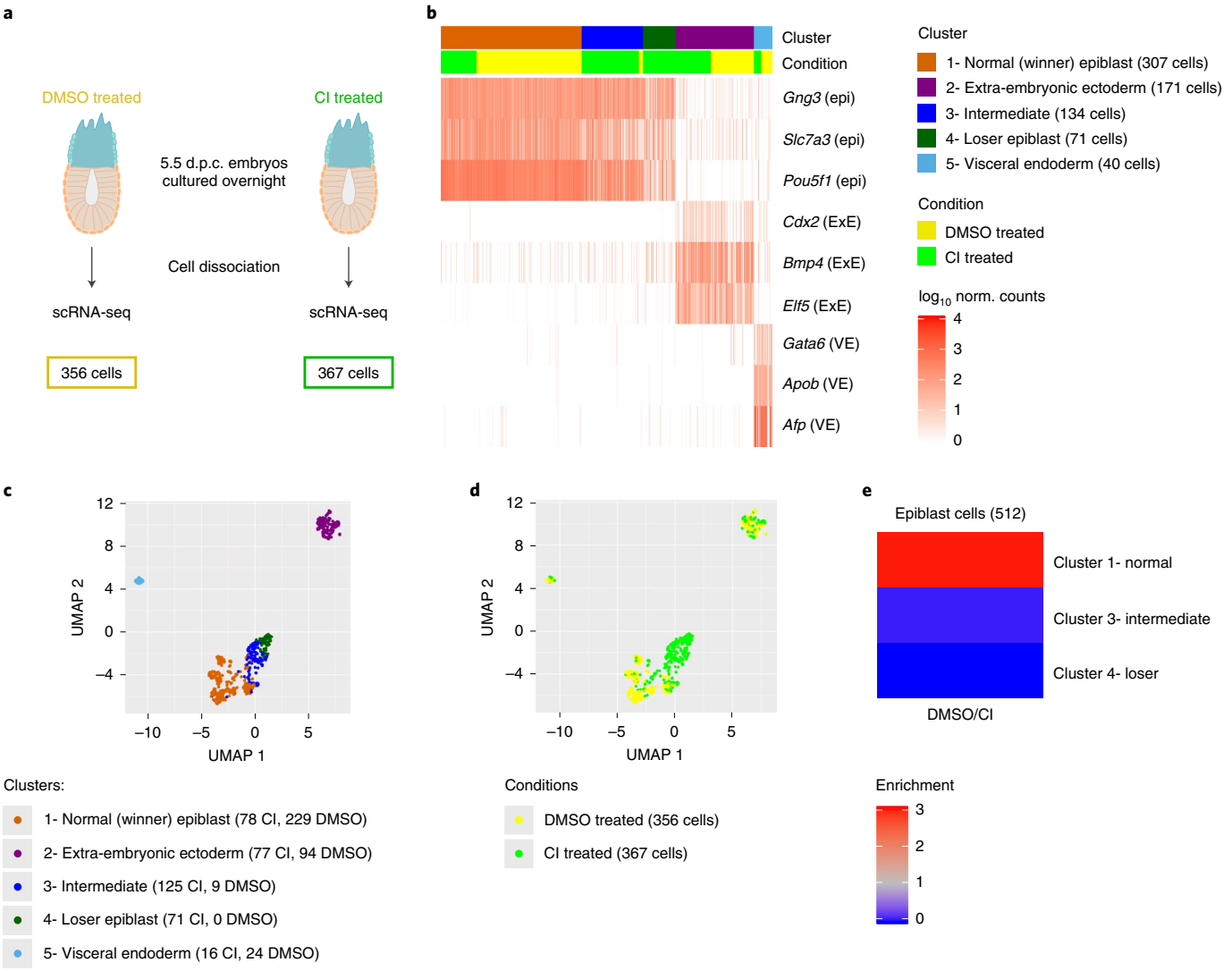


Fig. 1 | Cells eliminated during early mouse embryogenesis have a distinct transcriptional profile. a, Experimental design. The number of cells in the two conditions (DMSO treated and CI treated) refers to the cells that passed the quality control. d.p.c., days post-coitum. **b**, Identification of the clusters according to known gene markers from the different embryonic regions[73]. Three clusters (clusters 1, 3 and 4) show marker genes of the epiblast (Epi), while the remaining clusters correspond to the extra-embryonic visceral endoderm (VE; cluster 5) and extra-embryonic ectoderm (ExE; cluster 2). The epiblast clusters were named 'winner', 'intermediate' and 'loser' on the basis of the relative fraction of cells from CI-treated embryos they include (**e**). **c**,**d**, Uniform manifold approximation projection (UMAP) visualization of the single-cell RNA-seq data, with cells coloured according to cluster (**c**) or condition (**d**). A region made up exclusively by cells from CI-treated embryos emerged. **e**, Ratio between the fraction of cells from DMSO-treated and CI-treated embryos in the three epiblast clusters. While the 'winner' epiblast cluster shows an enrichment of cells from DMSO-treated embryos, the 'intermediate' and the 'loser' epiblast clusters are strongly enriched for cells from CI-treated embryos.

mechanisms act at the molecular and cellular level is still poorly understood.

To understand the nature of the cells eliminated during early mouse post-implantation development, we have analysed their transcriptional profile by single-cell RNA sequencing (scRNA-seq) and found that these cells share a cell competition signature. Analysis of the mis-regulated pathways identified mitochondrial dysfunction as a common feature. Importantly, our studies also found evidence of mtDNA mutations in the eliminated cells. Furthermore, we demonstrate that manipulating mitochondrial activity by either disrupting mitochondrial dynamics or introducing non-pathological mtDNA changes is sufficient to trigger cell competition. Therefore, these results pinpoint mitochondrial performance as a key cellular feature that determines the competitive ability of embryonic cells and suggest that cell competition is acting as a purifying selection during early mammalian development.

Results

Loser cells have a distinct transcriptional profile. We have previously shown that in the early post-implantation mouse embryo about 35% of epiblast cells are eliminated and that these cells are marked by low mTOR signalling[7]. However, we currently do not understand the characteristics of these cells or what triggers their elimination. To answer these questions, we have analysed their transcriptional profile with scRNA-seq. To ensure the eliminated cells can be captured, as we have done before[7], we isolated embryos at E5.5 and cultured them for 16 h in the presence of a caspase inhibitors (CIs) or vehicle (DMSO) (Fig. 1a). Unsupervised clustering of the scRNA-seq data revealed five clusters: two corresponding to extra-embryonic tissues (visceral endoderm and extra-embryonic ectoderm) and three that expressed epiblast marker genes (Fig. 1b,c, Extended Data Fig. 1a–f and Methods). Interestingly, cells from CI-treated and DMSO-treated embryos were unequally distributed

across the three epiblast clusters. In particular, one of these clusters (cluster 4) was only composed of cells from CI-treated embryos (Fig. 1d,e). Also notable is that all epiblast clusters contained cells in the G2/M and S phases of the cell cycle, suggesting they are all cycling (Extended Data Fig. 2a).

The three epiblast clusters are highly connected, as highlighted by a connectivity analysis carried out with PAGA[22] (Extended Data Fig. 2b). Hence, to establish the relationship between these epiblast clusters, we computed a diffusion map[23]. For this, we selected only cells captured from CI-treated embryos, to eliminate possible confounding effects due to the CI (Fig. 2a). However, when all epiblast cells were considered, the results remain unchanged (Extended Data Fig. 2c–e). This analysis identified a trajectory between the three epiblast clusters, with those cells unique to CI-treated embryos falling at one extreme end of the trajectory (corresponding to cluster 4; Fig. 2a) and with those cells present in both DMSO-treated and CI-treated embryos at the other (corresponding to cluster 1; Fig. 2a and Extended Data Fig. 2d).

To further define the identity of the epiblast cells of CI-treated embryos, we analysed the genes differentially expressed along the trajectory (Methods and Extended Data Fig. 3a) using ingenuity pathway analysis (IPA) to characterize gene signatures[24]. Importantly, we found that these differentially expressed genes fell under molecular and cellular function categories associated with cell death and survival, protein synthesis and nucleic acids (Fig. 2b). Analysis of the factors with enriched targets within the genes differentially expressed along the trajectory revealed RICTOR (an mTOR component), TLE3, MYC, MYCN, P53 and IGFR (that is, upstream of mTOR) as the top upstream regulators (Fig. 2c). Breaking down the differentially expressed genes into those downregulated or upregulated along the winner-to-loser trajectory revealed that the targets of RICTOR, MYC, MYCN and IGFR primarily fell within the downregulated genes (Supplementary Tables 1 and 2). P53-activated targets were preferentially upregulated and P53-repressed targets were preferentially downregulated (Extended Data Fig. 3b,c). Moreover, genes related to protein synthesis were primarily found to be downregulated.

The observation that the genes differentially expressed along the trajectory fall into cell death categories, as well as being mTOR, MYC and P53 targets, strongly suggests that cells at each end of the trajectory are the winners and losers of cell competition[5–7]. For this reason, we hereafter refer to those epiblast cells unique to CI-treated embryos as 'loser' epiblast cells and to those at the opposite end of the trajectory as the 'winner' epiblast cells. Those cells lying between these two populations on the trajectory are considered 'intermediate'. Using this knowledge, we can define a diffusion pseudotime (dpt) coordinate[25] originating in the 'winner' cluster that tracks the position of cells along the trajectory and that can be interpreted as a 'losing score'; that is, it quantifies how strong the signature of the 'losing' state is in the transcriptome of a cell (Fig. 2d,e).

In accordance with previous studies[6,8,9], we also found evidence for miss-patterning in the eliminated epiblast cells, as a proportion of these cells co-expressed naïve pluripotency and differentiation markers (Fig. 2f and Extended Data Fig. 3d). To test if loser cells are developmentally delayed or advanced compared to control cells, we projected our data onto a previously published diffusion map that includes epiblast cells from E5.5, E6.25 and E6.5 embryos[26]. We found that all epiblast cells, irrespective of the condition in which the embryos were cultured (that is, treated with DMSO or CI) and of their losing state (that is, that they belonged to the winner, intermediate or loser cluster), mostly overlapped with E6.5 epiblast cells (Extended Data Fig. 3e–g). Cells from the loser cluster were slightly closer to the E6.25 stage than the winner and intermediate cells, as shown by their pseudotime coordinate, but they remain far from the earlier E5.5 stage. This result, combined with the higher expression of some differentiation markers observed in loser cells, suggests that these cells are miss-patterned rather than developmentally delayed.

Loser cells have defects in mitochondrial function. Using IPA, we next analysed the cellular pathways mis-regulated in loser epiblast cells and found that the top two pathways (mitochondrial dysfunction and oxidative phosphorylation (OXPHOS)) were related to mitochondrial function (Fig. 3a,b and Supplementary Tables 1 and 2). For example, we found a downregulation along the winner-to-loser trajectory of the mtDNA-encoded subunits *mt-Nd3* and *mt-Atp6*, of regulators of mitochondrial dynamics such as *Opa1* (optic atrophy 1), as well as of genes involved in mitochondrial membrane and cristae organization such as *Samm50* (Fig. 3c), suggesting that mitochondrial function is impaired in loser cells.

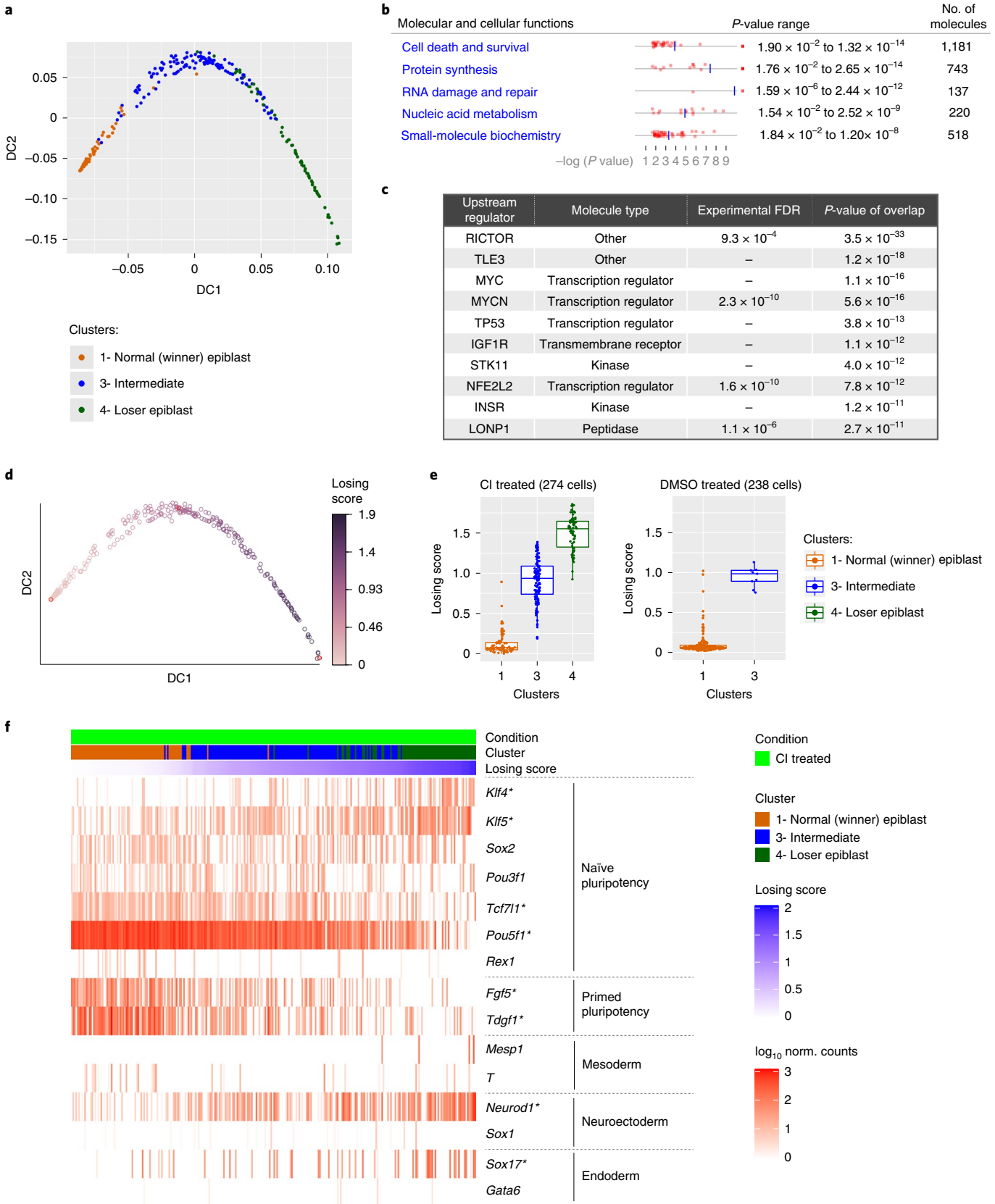
A recent body of evidence has revealed that stress responses, such as the integrated stress response (ISR) or the closely related unfolded protein response (UPR), when triggered in cells with impaired mitochondrial function prompt a transcriptional programme to restore cellular homeostasis[27–29]. We observed that loser epiblast cells displayed a characteristic UPR/ISR signature[30–33] and key regulators of this response, such as *Atf4*, *Ddit3*, *Nfe2l2* (*Nrf2*) and *Foxo3* were all upregulated in these cells (Extended Data Fig. 4a–d). Similarly, *Sesn2*, a target of p53 that controls mTOR activity[34], was also upregulated in loser cells (Extended Data Fig. 4d). These findings support that loser epiblast cells present mitochondrial defects, leading to the activation of a stress response in an attempt to restore cellular homeostasis[35].

To validate the significance of the observed mitochondrial defects, we did two things: First, we asked if the changes in expression of mitochondrial regulators at the mRNA level are also reflected at the protein level. We observed that in CI-treated embryos, loser cells that persist and are marked by low mTOR activity[7] also show significantly lower OPA1 levels (Fig. 3d–f). We also found that DMSO-treated embryos showed strong DDIT3 staining (an UPR/ISR marker also known as CHOP) in the dying cells that accumulate in the pro-amniotic cavity, and that in CI-treated embryos, DDIT3 expression was upregulated in a proportion of epiblast cells (Extended Data Fig. 4e–g). Second, we studied the mitochondrial membrane potential ($\Delta\psi m$), an indication of mitochondrial health, in loser epiblast cells. We observed that while the cells of DMSO-treated embryos showed a high $\Delta\psi m$ that fell within a narrow range, in CI-treated embryos the proportion of cells with a low $\Delta\psi m$ significantly increased (Fig. 3d,g,h). Together, these results

Fig. 2 | A cell competition transcriptional signature is identified in cells eliminated during mouse embryonic development. a, Diffusion map of epiblast cells (only from CI-treated embryos), coloured by cluster. **b**,**c**, IPA of genes differentially expressed along the diffusion trajectory (Extended Data Fig. 3a) generated lists of the top five molecular and cellular functions (**b**) and upstream regulators (**c**) found to be differentially activated in epiblast cells along the diffusion trajectory from winner (cluster 1) to loser status (cluster 4). **d**, Diffusion map of epiblast cells (only from CI-treated embryos) coloured by the dpt coordinate. The winner and the loser clusters were found at the two extremities of the trajectory, hence the dpt coordinate can be interpreted as a 'losing score'. **e**, Losing score of the cells in the three epiblast clusters in CI-treated (left) or DMSO-treated (right) embryos. The losing score of the cells from DMSO-treated embryos was obtained by projecting them on the diffusion map shown in **d** (Methods). **f**, Expression levels in epiblast cells from CI-treated embryos of genes (in rows) that are markers for naïve pluripotency (*Klf4*, *Klf5*, *Sox2*, *Pou3f1*, *Tcf7l1*, *Pou5f1* and *Zfp42 (Rex1)*), primed pluripotency (*Fgf5* and *Tdgf1*), mesoderm (*Mesp1* and *T*), neuroectoderm (*Neurod1* and *Sox1*) and endoderm (*Sox17 and Gata6*). Cells (columns) were sorted by their losing scores. The genes marked with an asterisk were differentially expressed along the trajectory. See Methods for details on statistical analysis.

suggest that loser epiblast cells have impaired mitochondrial activity that triggers a stress response.

Mitochondrial dysfunction is common to different loser cells. To address if mitochondrial defects are a common feature of loser cells eliminated by cell competition, we analysed embryonic stem cells (ESCs) that are defective for bone morphogenetic protein (BMP) signalling (*Bmpr1a*$^{-/-}$) and tetraploid cells (4n)[6]. We first carried out a mass spectrometry analysis using the Metabolon platform and found that metabolites and intermediates of the



Upstream regulator	Molecule type	Experimental FDR	*P*-value of overlap
RICTOR	Other	9.3×10^{-4}	3.5×10^{-33}
TLE3	Other	–	1.2×10^{-18}
MYC	Transcription regulator	–	1.1×10^{-16}
MYCN	Transcription regulator	2.3×10^{-10}	5.6×10^{-16}
TP53	Transcription regulator	–	3.8×10^{-13}
IGF1R	Transmembrane receptor	–	1.1×10^{-12}
STK11	Kinase	–	4.0×10^{-12}
NFE2L2	Transcription regulator	1.6×10^{-10}	7.8×10^{-12}
INSR	Kinase	–	1.2×10^{-11}
LONP1	Peptidase	1.1×10^{-6}	2.7×10^{-11}

tricarboxylic acid (TCA) cycle, such as malate, fumarate, glutamate and α-ketoglutarate are depleted in both *Bmpr1a*$^{-/-}$ and 4n ESCs in differentiation culture conditions (Fig. 4a). Next, we performed an extracellular flux Seahorse analysis of *Bmpr1a*$^{-/-}$ ESCs to measure their glycolytic and OXPHOS rates. We observed that when these cells are maintained in pluripotency culture conditions that are not permissive for cell competition[6], they exhibit a higher OXPHOS rate than control cells (Extended Data Fig. 5a,b). In contrast, when *Bmpr1a*$^{-/-}$ cells are induced to differentiate, this phenotype is reversed, with mutant cells showing lower ATP generated through OXPHOS and a higher glycolytic capacity than controls (Fig. 4b–e and Extended Data Fig. 5c,d). This suggests that after differentiation *Bmpr1a*$^{-/-}$ cells are unable to sustain proper OXPHOS activity.

To further test the possibility that defective mouse ESCs (mESCs) have impaired mitochondrial function, we assessed their Δψm. We found that whilst *Bmpr1a*$^{-/-}$ and 4n cells had a similar Δψm to control cells in pluripotency conditions (Extended Data Fig. 5e,f), following differentiation both these cell types presented a loss of Δψm, irrespective of whether they were separate or co-cultured with wild-type cells (Fig. 4f,g). This reduction in Δψm is not due to excessive mitochondrial reactive oxygen species (ROS) production or to a lower mitochondrial mass within mutant cells because, as for example, *Bmpr1a*$^{-/-}$ cells had lower ROS levels and similar TOMM20 and mt-CO1 expression to control cells (Fig. 4h–j and Extended Data Fig. 5g). The fact that the loss of Δψm and lower OXPHOS activity can be observed even when loser cells are cultured separately suggests that the mitochondrial dysfunction phenotype is an inherent property of loser cells and not a response to them being out-competed. These results also indicate that the mitochondrial defects are directly linked to the emergence of the loser status: In conditions that are not permissive for cell competition (pluripotency), mutant cells do not show defective mitochondrial function, but when they are switched to differentiation conditions that allow for cell competition, they display impaired mitochondrial function.

To further explore the relationship between mitochondrial activity and the competitive ability of the cell, we analysed the Δψm of BMP-defective cells that are null for p53 (*Bmpr1a*$^{-/-}$;*p53*$^{-/-}$ ESCs), as these are not eliminated by wild-type cells[7]. Remarkably, we observed that mutating *p53* in *Bmpr1a*$^{-/-}$ cells not only rescues the loss of Δψm of these cells, but also causes hyperpolarization of their mitochondria (Fig. 4k). These results suggest a role for P53 in regulating mitochondrial activity of ESCs and strongly support a pivotal role for mitochondrial activity in cell competition.

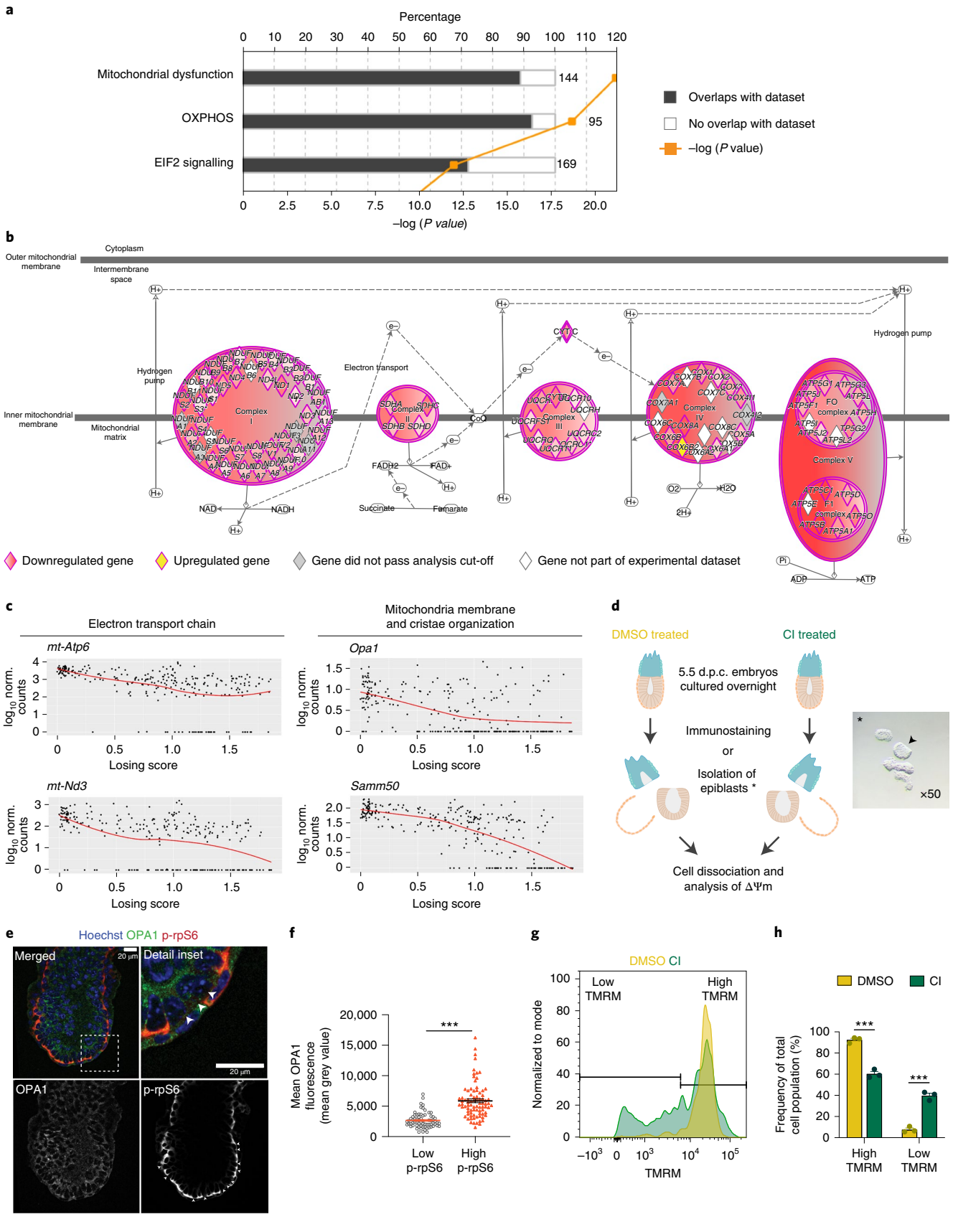
Impaired mitochondrial function triggers cell competition. The mitochondrial defects observed in loser cells led us to ask if disrupting mitochondrial activity alone is sufficient to trigger cell competition. During the onset of differentiation, mitochondrial shape changes substantially. In pluripotent cells, mitochondria have a round and fragmented shape, but after differentiation they fuse and become elongated, forming complex networks[10]. Given that this change in shape correlates with when cell competition occurs, we tested if disrupting mitochondrial dynamics is sufficient to induce cell competition. MFN1 and MFN2 regulate mitochondrial fusion and DRP1/DNM1L controls their fission[36–38]. We generated ESCs null for mitofusin 2 (*Mfn2*$^{-/-}$), which have enlarged globular mitochondria, and ESCs null for dynamin-related protein 1 (*Drp1*$^{-/-}$), which show hyper-elongated mitochondria (Fig. 5a). We first tested the competitive ability of *Mfn2*$^{-/-}$ ESCs in pluripotency conditions, which we have previously found not to induce out-competing in *Bmpr1a*$^{-/-}$ or 4n cells[6]. Interestingly, we found that although *Mfn2*$^{-/-}$ cells grow similarly to wild-type cells in separate cultures, they were out-competed in co-culture (Fig. 5b). Analogously, the *Drp1* mutant cells did not grow significantly slower than wild-type cells when cultured separately in differentiation-inducing conditions, but they were out-competed by wild-type cells in co-culture (Fig. 5c). The observation that disrupting mitochondrial dynamics can induce cell competition even in pluripotency culture conditions, suggests that mitochondrial activity is a dominant parameter determining the competitive ability of the cell.

To establish how disruption of mitochondrial fusion and fission affects mitochondrial performance, we compared the Δψm, respiration rates and mitochondrial ATP production of *Mfn2*$^{-/-}$ and *Drp1*$^{-/-}$ ESCs to those of wild-type cells (Fig. 5d–g). We found that whilst *Mfn2*$^{-/-}$ and *Drp1*$^{-/-}$ ESCs had lower Δψm than control cells (Fig. 5d,f), *Mfn2*$^{-/-}$ ESCs had lower maximal respiration rates but similar basal respiration and ATP production to controls, and *Drp1*$^{-/-}$ ESCs showed similar respiration and ATP production to controls (Fig. 5e,g). This suggests that ATP production or respiration rates alone do not determine the relative competitive ability of ESCs.

Besides mitochondrial dysfunction, another prominent signature of loser cells found in vivo was the UPR/ISR (Extended Data Fig. 4). Because the loss of *Drp1* has been associated with activation of the UPR[39–41], we investigated if the *Drp1*$^{-/-}$ loser cells also showed evidence for the activation of the UPR/ISR. We observed that *Drp1*$^{-/-}$ cells show higher expression of ATF4 and phosphorylated eukaryotic initiation factor 2α (p-eIF2α) than their wild-type counterparts, which is indicative of UPR/ISR activation (Fig. 5h)[39–41]. Another feature previously described following loss of *Drp1* is the proteolytic cleavage of OPA1, where short isoforms (S-OPA1) are accumulated in detriment of the long isoforms (L-OPA1)[39]. When we analysed the expression of OPA1 in wild-type and *Drp1*$^{-/-}$ cells, we observed that while wild-type cells retained L-OPA1 expression, loser cells predominantly expressed the S-OPA1 isoforms and

Fig. 3 | Cells eliminated during early mouse embryogenesis have mitochondrial defects. a, Top canonical pathways, identified by IPA, mis-regulated in loser cells in comparison to normal epiblast cells. The numbers at the end of each bar refer to total amount of genes involved in that pathway. The percentage refers to the number of genes found mis-regulated in loser cells relative to the number total genes within each pathway. **b**, Details of changes in the OXPHOS pathway identified in **a**. Circular and oval shapes represent each of the electron transport chain (ETC) complexes (complexes I to V). Diamond shapes represent subunits of each ETC complex. Downregulated genes in loser cells are coloured in shades of red. Darker shades correspond to lower false discovery rate (FDR) values. *Cox6b2*, in yellow, was upregulated in loser cells. Grey denotes genes that were not differentially expressed between loser and winner cells (FDR > 0.01). White denotes genes from the Knowledge Base that were not tested (for example, because they were not detected in our dataset). **c**, Expression levels of mitochondrial genes as a function of the losing score of cells. **d**, Experimental design adopted to assess mitochondrial function in **e**–**h**. The asterisk indicates a representative micrograph of one of the isolated epiblasts (arrow) used for Δψm analysis after embryo microdissection. **e**, Representative immunohistochemistry of OPA1 in E6.5 embryos where cell death was inhibited (CI treated), quantified in **f**. Loser cells were identified by low mTOR activation (low p-rpS6; arrowheads). Scale bar, 20 µm. **f**, Quantification of OPA1 fluorescence in normal epiblast cells and loser cells. $N = 6$ embryos with a minimum of 8 cells analysed per condition. **g**, Representative histogram of flow cytometry analysis of tetramethylrhodamine methyl ester (TMRM) probe, indicative of Δψm, in epiblast cells from embryos where cell death was allowed (DMSO treated) or inhibited (CI treated), quantified in **h**. **h**, Frequency of epiblast cells with high or low TMRM fluorescence, according to the range defined in **g** from embryos where cell competition was allowed (DMSO treated) or inhibited (CI treated). Data were obtained from three independent experiments and are shown as the mean ± s.e.m. (**g** and **h**). Twelve embryos per condition were pooled for each experiment. See Methods for details on statistical analysis.

displayed almost no expression of L-OPA1 (Fig. 5i). This defect has been associated with mito-ribosomal stalling, a phenotype that can be replicated by treating cells with actinonin (Extended Data Fig. 6)[42]. To test if the shift in isoform expression observed in *Drp1*$^{-/-}$ ESCs is due to aberrant mitochondrial translation, we treated cells with doxycycline, which inhibits translation in mitochondria[43], and



observed that this was sufficient to partially rescue L-OPA1 expression (Fig. 5j). This rescue together with the evidence for UPR/ISR activation suggests that *Drp1*$^{-/-}$ cells display defects in mitochondrial translation.

Loser epiblast cells accumulate mtDNA mutations. There is strong evidence for selection against aberrant mitochondrial function induced by deleterious mtDNA mutations in mammals[21,44–47]. Given our observation that cell competition selects against cells with impaired mitochondrial function, we asked if cell competition could be reducing mtDNA heteroplasmy (frequency of different mtDNA variants) during mouse development. It has been recently shown that scRNA-seq can be used to reliably identify mtDNA variants, although with a lower statistical power compared to more direct approaches, like mtDNA sequencing[48]. We therefore tested if mtDNA heteroplasmy is present in our scRNA-seq data and whether this correlates with the losing score of a cell. Our analysis revealed that the frequency of specific mtDNA polymorphisms increased with the losing score of epiblast cells (Fig. 6a), and such mtDNA changes occurred within *mt-Rnr1* and *mt-Rnr2* (Fig. 6b–h and Extended Data Fig. 7a–e). Moreover, these changes were not dependent on the litter from which the embryo came from (Extended Data Fig. 7f–k). As it was formally possible that these loser-specific sequence changes could originate from contaminating nuclear mitochondrial sequences (NUMTS) or from RNA editing, we performed several controls to confirm that mtDNA polymorphisms are the most likely source of these changes (Methods). For example, we considered only the RNA-seq reads that are uniquely mapped to the mitochondrial genome and not to nuclear DNA, and we confirmed that the variants with highest heteroplasmy found in the 'loser' cells were not present in any of the NUMTS that have previously been reported or could be identified using BLAST. Moreover, we verified that the observed sequence changes were not compatible with canonical RNA editing (Methods). It is worth noting that the sequence changes we detected in *mt-Rnr1* and *mt-Rnr2* strongly co-occurred in the same cell, with those closest together having the highest probability of coexisting (Fig. 6i and Extended Data Fig. 7l). This is suggestive of mtDNA replication errors that could be 'scarring' the mtDNA, disrupting the function of *mt-Rnr1* (12S rRNA) and *mt-Rnr2* (16S rRNA) and causing the loser phenotype. Importantly, the presence of these specific mtDNA mutations in the loser cells suggests that cell competition could be contributing to the elimination of deleterious mtDNA mutations during early mouse development. Of note, we only report mtDNA variants detected in regions of the genome with high sequencing coverage (Extended Data Fig. 7m); therefore, the presence of other variations in mtDNA sequences between winner and loser cells cannot be excluded.

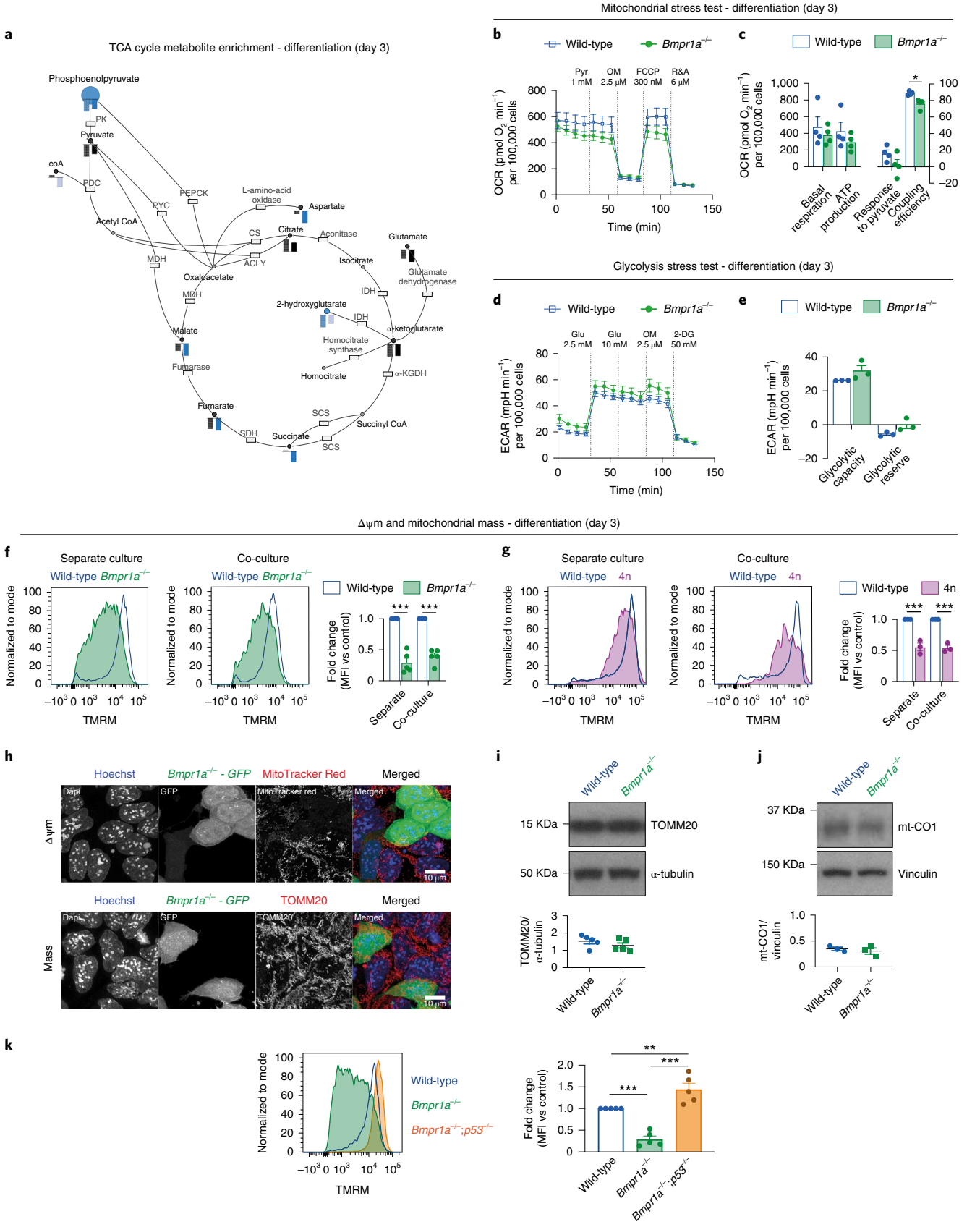
mtDNA sequence determines the competitive ability of a cell. To explore this possibility further, we analysed if alterations in mtDNA can induce cell competition by testing the competitive ability of ESCs with non-pathological differences in mtDNA sequence. For this we compared the relative competitive ability of ESCs that shared the same nuclear genome background but differed in their mitochondrial genomes by a small number of non-pathological sequence changes. We derived ESCs from hybrid mouse strains that we had previously engineered to have a common nuclear C57BL/6N background, but mtDNAs from different wild-caught mice[16]. Each wild-derived mtDNA variant (or haplotype) contains a specific number of single-nucleotide polymorphisms (SNPs) that lead to a small number of amino acid changes when compared to the C57BL/6N mtDNA haplotype. Furthermore, these haplotypes (BG, HB and ST) can be ranked according to their genetic distance from the C57BL/6N mtDNA (Fig. 7a and Extended Data Fig. 8a). Characterization of the isolated ESCs revealed that they have a range of heteroplasmy (mix of wild-derived and C57BL/6N mtDNAs) that is stable over several passages (Extended Data Fig. 8b). Importantly, these different mtDNA haplotypes and different levels of heteroplasmy do not alter cell size, cell granularity, mitochondrial mass or mitochondrial dynamics, nor do they substantially impact the cell's $\Delta\psi$m (Extended Data Fig. 8c–f).

When we tested the competitive ability of these ESCs with different mtDNA content, in pluripotency culture conditions, we observed that cells carrying the mtDNAs that were most distant from the C57BL/6N mtDNA, such as the HB (100%), the HB (24%) and the ST (46%) ESCs could all out-compete the C57BL/6N line (Fig. 7b,c and Extended Data Fig. 8g). Similarly, when we tested the HB (24%) line against the BG (99%) or the BG (95%) lines (which have mtDNAs more closely related to the C57BL/6N mtDNA), we found that cells with the HB haplotype could also out-compete these ESCs (Fig. 7d and Extended Data Fig. 8h). In contrast, we observed that the HB (24%) ESCs were unable to out-compete their homoplasmic counterparts, HB cells (100%) or the ST cells (46%) that carry the most distant mtDNA variant from C57BL/6N (Fig. 7e and Extended Data Fig. 8i). These results tell us three things: First, non-pathological differences in mtDNA sequence can trigger cell competition. Second, a competitive advantage can be conferred by only a small proportion of mtDNA content, as indicated by our finding that HB (24%) behave as winners. Finally, these findings suggest that the phylogenetic proximity between mtDNA variants can potentially determine their competitive cell fitness.

To characterize the mode of competition between cells with different mtDNA, we focused on the HB (24%) and the BG (95%) ESCs. Analysis of these cell lines revealed that specifically when co-cultured, the BG (95%) cells displayed high levels of apoptosis (Fig. 7f), indicating that they are out-competed through their

Fig. 4 | Mitochondrial defects are a common feature of cells eliminated by cell competition. a, Metabolic enrichment analysis of the TCA cycle and intermediate metabolites obtained using Metabolon platform for defective cells (*Bmpr1a*$^{-/-}$, left bar; 4n, right bar), in comparison to wild-type cells during differentiation. Bars indicate compound levels relative to wild-type cells. Blue bars indicate compounds that were significantly altered ($P < 0.05$), and light-blue bars indicate compounds that were almost significantly altered ($0.05 \leq P \leq 0.1$). Black bars indicate compounds that were altered although not statistically significant in comparison to the levels found in wild-type cells. The enzymes on the pathway are represented as boxes and labelled by their canonical names. **b**–**e**, Metabolic flux analysis of wild-type and BMP-defective cells during differentiating conditions. Analysis of OCR as a measure of mitochondrial function (mitochondrial stress test; **b**). Details of metabolic parameters found changed from the analysis of the mitochondrial stress test (**c**). Analysis of extracellular acidification rate (ECAR) as a measure of glycolytic function (glycolysis stress test; **d**). Details of metabolic parameters found changed from the analysis of the glycolysis stress test (**e**). **f**,**g**, $\Delta\psi$m in defective mESCs undergoing differentiation in separate or co-culture conditions. Representative histograms of TMRM fluorescence and quantification for wild-type and *Bmpr1a*$^{-/-}$ (**f**) and wild-type and 4n (**g**) cells. **h**, Representative micrographs of wild-type and *Bmpr1a*$^{-/-}$ cells co-cultured during differentiation and stained for a reporter of $\Delta\psi$m (MitoTracker Red; top) or mitochondrial mass (TOMM20; bottom). Nuclei were stained with Hoechst. Scale bar, 10 µm. **i**,**j**, Western blot analysis of mitochondrial mass markers TOMM20 (**i**) and mt-CO1 (**j**) for wild-type and *Bmpr1a*$^{-/-}$ cells during differentiation. **k**, Analysis of $\Delta\psi$m for wild-type, *Bmpr1a*$^{-/-}$ and *Bmpr1a*$^{-/-}$;*p53*$^{-/-}$ cells during differentiation. Representative histogram of TMRM fluorescence and quantification. Data are the mean ± s.e.m. Extracellular flux Seahorse data were obtained from three (**d** and **e**) or four (**b** and **c**) independent experiments, with five replicates per cell type in each assay. The remaining data were obtained from three (**g** and **j**) or five (**a**,**f**,**i** and **k**) independent experiments. See Methods for details on statistical analysis. MFI, mean fluorescence intensity.

elimination. To gain further insight, we performed bulk RNA-seq of these cells in separate and co-culture conditions (Extended Data Fig. 8j) and analysed the differentially expressed genes by gene-set enrichment analysis (GSEA). We found that in separate culture the most notable features that distinguished BG (95%) from HB (24%) cells were a downregulation of genes involved in OXPHOS



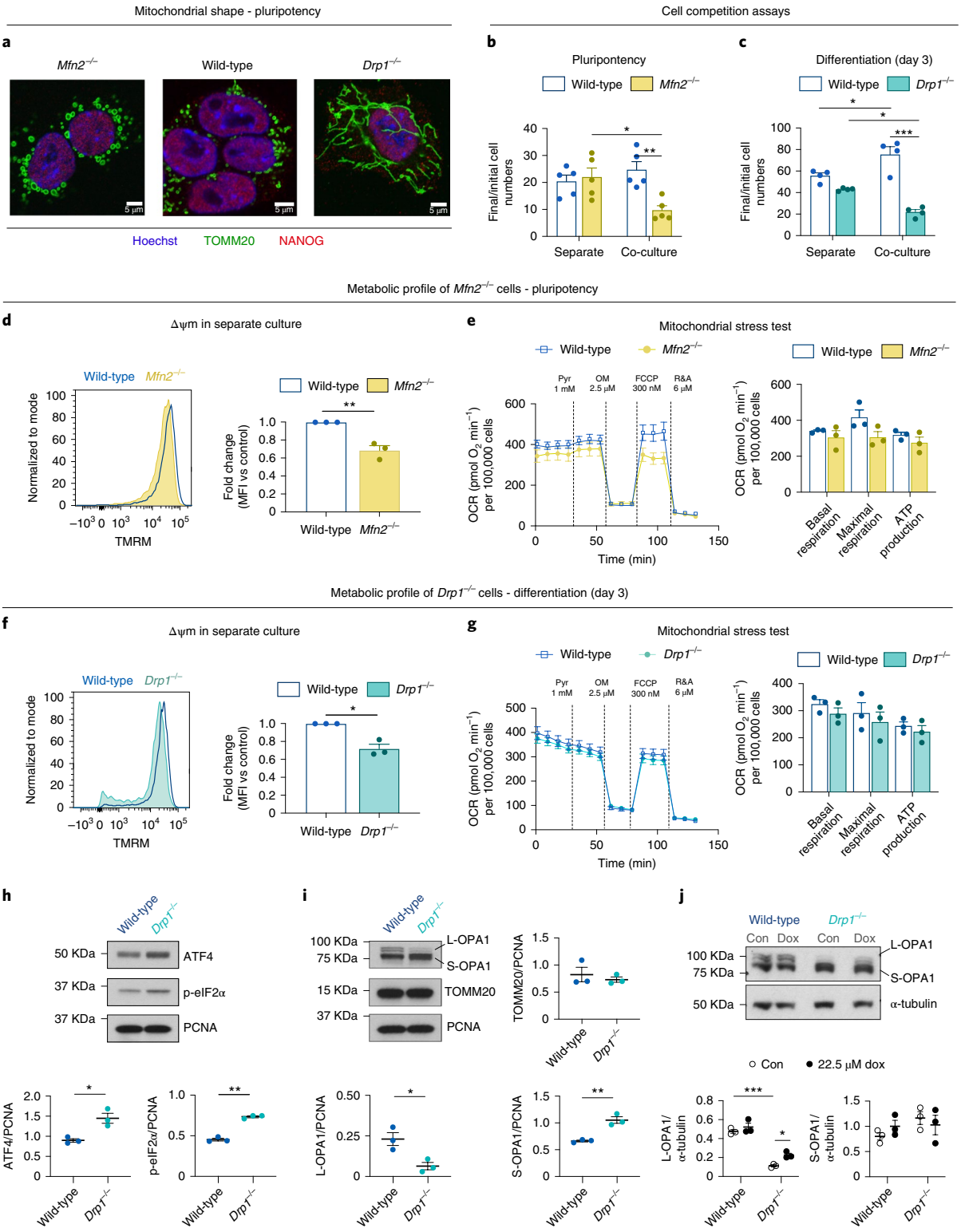


Fig. 5 | Manipulating mitochondria biology is sufficient to trigger cell competition. a, Representative micrographs of wild-type, *Mfn2*$^{-/-}$ and *Drp1*$^{-/-}$ mESCs showing alterations in mitochondrial morphology in mutant cells. TOMM20 was used as a mitochondrial marker and NANOG as a pluripotency marker. Nuclei were stained with Hoechst. Scale bar, 5 μm. **b**,**c**, Cell competition assays between wild-type mESCs and cells with altered morphology: *Mfn2*$^{-/-}$ during pluripotency (**b**) and *Drp1*$^{-/-}$ during differentiation (**c**). The ratio of final/initial cell numbers cultured separately or in co-culture is shown. **d**–**j**, Metabolic profile of *Mfn2*$^{-/-}$ and *Drp1*$^{-/-}$ mESCs. Analysis of mitochondrial Δψm for wild-type and *Mfn2*$^{-/-}$ cells cultured separately during pluripotency (**d**) and for wild-type and *Drp1*$^{-/-}$ mESCs $^{-/-}$ during differentiation in a separate culture (**f**). Metabolic flux analysis of wild-type and *Mfn2*$^{-/-}$ mESCs cultured separately during pluripotency (**e**) and for wild-type and *Drp1*$^{-/-}$ undergoing differentiation in separate cultures (**g**). Data were collected from three independent experiments. **h**–**j**, Western blot analysis of markers of UPR and mitochondrial markers in wild-type and *Drp1*$^{-/-}$ during differentiation in separate culture. Cells were treated with doxycycline (Dox, 22.5 μM) or vehicle (Con) from day 1 of differentiation and samples were collected on day 3 (**j**). Data are the mean ± s.e.m. of three (**d**–**j**), four (**c**) or five (**b**) independent experiments. See Methods for details on statistical analysis.

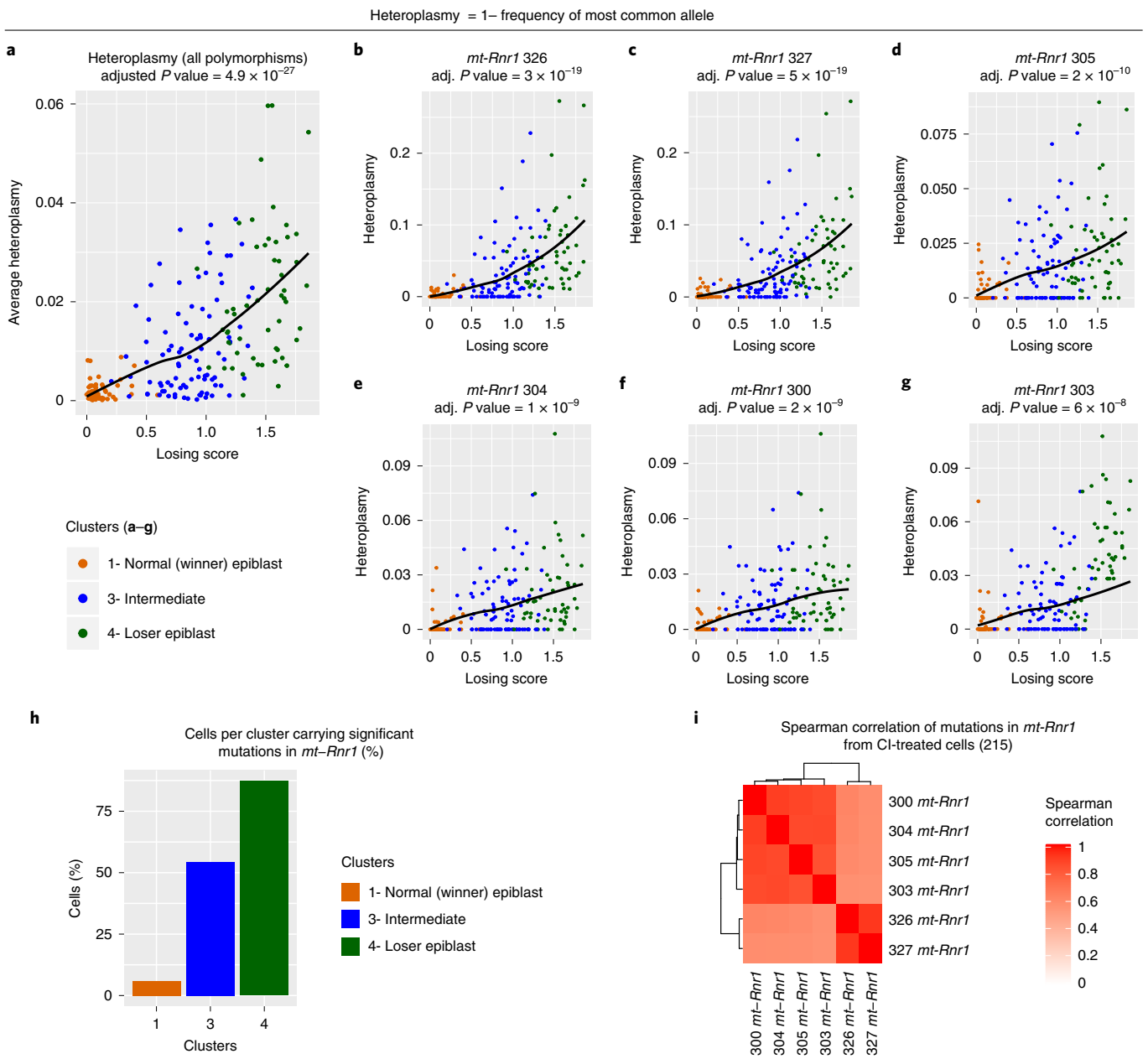


Fig. 6 | Intermediate and loser epiblast cells accumulate polymorphisms in mtDNA sequences. a–g, mtDNA heteroplasmy (plotted as heteroplasmy = 1 minus the frequency of most common allele) in epiblast cells from CI-treated embryos. Average heteroplasmy (considering all 11 polymorphisms that had a statistically significant dependence on the losing score; Methods) as a function of the losing scores of the cells (**a**). mtDNA heteroplasmy for six positions within *mt-Rnr1* (**b–g**). The heteroplasmy at all these positions, as well as the average heteroplasmy, increased with the losing scores of the cells in a statistically significant way (the adjusted *P* value estimated via a generalized linear model is indicated at the top of each plot). **h**, The bar plot indicates the fraction of epiblast cells in each of the clusters indicated on the *x* axis (winner, intermediate and loser) that carries a mean heteroplasmy (computed on the six positions within the *mt-Rnr1* indicated in **b–g**) greater than 0.01. This shows that the level of mtDNA heteroplasmy in *mt-Rnr1* is strongly associated with the loser status of the cells, as ~55% and ~87% of cells in the intermediate and the loser clusters, respectively, had heteroplasmic sequences in this gene compared to only ~5% of cells in the winner cluster. **i**, Spearman's correlation coefficient between the mtDNA heteroplasmy at the six positions shown in **b–g**. See Methods for details on statistical analysis.

and an upregulation of those associated with cytokine activity (Fig. 7g). Interestingly, in the co-culture condition, in addition to these signatures, BG (95%) cells revealed a downregulation in signature markers of MYC activity and mTOR signalling (Fig. 7h), whose downregulation is a known read-out of loser status during cell competition in the embryo[5–7] (Fig. 2c).

To test if the downregulation of genes involved in OXPHOS was also reflected at the functional level, we compared oxygen consumption rates (OCRs) and mitochondrial ATP generation in HB (100%), HB (24%), BG (95%) and C57BL/6N ESCs. We found that the winner cells HB (100%) and HB (24%) had higher basal respiration, higher maximal respiration and higher mitochondrial ATP production than the loser BG (95%) and C57BL/6N ESCs (Extended Data Fig. 9). These data indicate that the mtDNA differences that exist between winner and loser cells are sufficient to affect their mitochondrial performance and this ultimately determines their competitive ability. However, the observation that differentiating *Drp1*$^{-/-}$ ESCs are eliminated by cell competition but do not show differences in respiration rates or

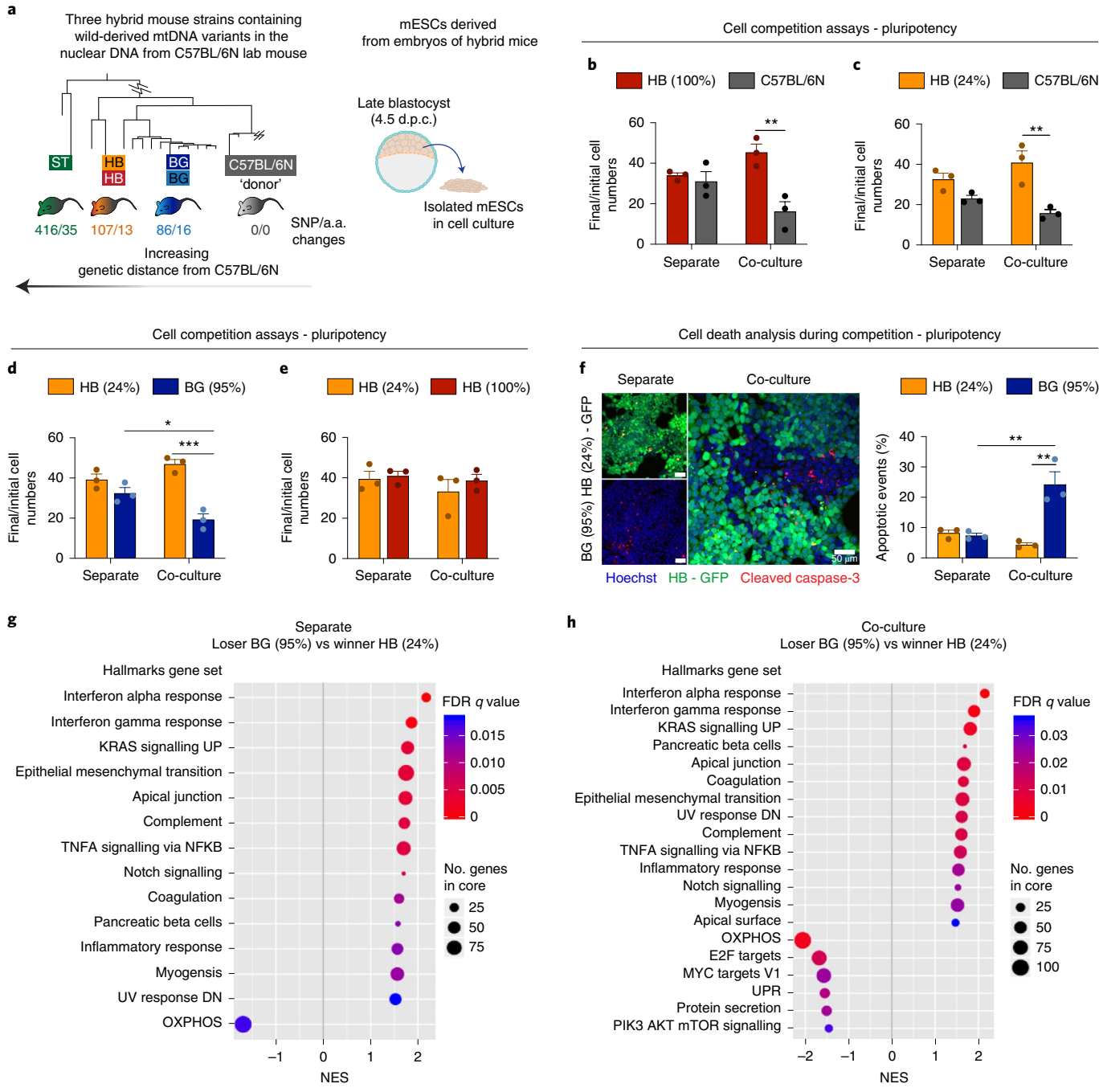


Fig. 7 | Changes in mtDNA sequence can determine the competitive ability of a cell. a, Derivation of mESCs from hybrid mouse strains, generated elsewhere by Burgstaller and colleagues. Neighbour-joining phylogenetic analysis of mtDNA from wild-derived and C57BL/6N mouse strains that were used to generate hybrid mice (adapted from a previous study[16]), illustrates the genetic distance of the mtDNA from wild-derived mouse strains to the C57BL/6N laboratory mouse. The number of SNPs and amino acid (a.a.) changes from the wild-derived to laboratory mouse strain is shown. mESCs were derived from embryos of hybrid mice, containing the nuclear background of a C57BL/6N laboratory mouse and mtDNA from three possible wild-derived strains (BG, HB or ST). **b**–**e**, Cell competition assays between cells derived from the embryos of hybrid mice performed in pluripotency maintenance conditions. The ratio of final/initial cell numbers in a separate culture or co-culture is shown. **f**, Representative micrographs of cleaved caspase-3 staining and quantification of the percentage of apoptotic events in winners HB (24%) and loser BG (95%) mESCs maintained in pluripotency and cultured in separate or co-culture conditions. **g**,**h**, GSEA of differentially expressed genes from bulk RNA-seq in loser BG (95%) compared to winner HB (24%) mESCs maintained in pluripotency and cultured in separate (**g**) or co-culture (**h**) conditions. Gene sets that show positive normalized enrichment scores (NESs) are enriched in loser cells, while gene sets that show negative NESs are depleted in loser cells. Data were obtained from four independent experiments (**g**,**h**). Remaining data are the mean ± s.e.m. of three independent experiments (**b**–**f**). See Methods for details on statistical analysis.

mitochondrial ATP production (Fig. 5c,g), suggests that respiration or ATP production rates alone are unlikely to be the mitochondrial parameters that control competitive cell fitness.

The finding that the genes downregulated in BG (95%) cells when co-cultured with HB (24%) cells fell under functional categories relating to mitochondrial function (Extended Data Fig. 10a)

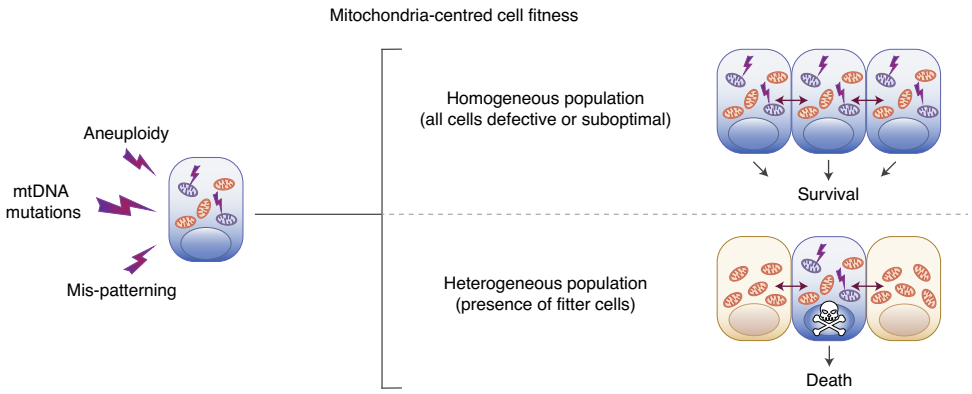


Fig. 8 | Model of cell competition. Summary of the main findings of the study. A range of cellular defects, such as aneuploidy, mis-patterning or mtDNA mutations, cause alterations in mitochondrial function, affecting the relative fitness of cells. The cells with suboptimal mitochondrial activity survive in a homogeneous population but are eliminated by cell competition in the presence of fitter cells.

led us to analyse the degree of overlap between these genes and the genes differentially expressed along the winner-to-loser trajectory in the embryo. We observed a significant overlap in mis-regulated genes (Extended Data Fig. 10b), as well as in the functional components that these genes can be categorized into (Extended Data Fig. 10c). This further highlights the importance of relative mitochondrial activity for determining the competitive ability of embryonic cells.

Discussion

The emerging role of cell competition as a regulator of cell fitness in a wide range of cellular contexts, from the developing embryo to the ageing tissue[1–3], has highlighted the importance of understanding which cell types are normally eliminated by this process. With the aim of understanding this question, we analysed the transcriptional identity of the cells eliminated in the early mouse embryo. We found that they not only present a cell competition signature but also have impaired mitochondrial function and are marked by sequence changes in mt-Rnr1 and mt-Rnr2. Starting from these results, we leveraged in vitro models of cell competition to show that: (1) mitochondrial function is impaired in loser cells eliminated by cell competition, and (2) differences in mitochondrial activity are sufficient to trigger cell competition in ESCs. Overall, this points to mitochondrial performance as a key determinant of the competitive ability of cells during early mammalian embryonic development. One implication of our findings is that a range of different types of defects, such as mis-patterning, karyotypic abnormalities or mtDNA mutations, lead to dysfunctional mitochondria at the onset of differentiation and that ultimately it is their impaired mitochondrial function that triggers cell competition, inducing their elimination (Fig. 8).

Embryos are exposed to different microenvironments in vivo and when cultured ex vivo. Similarly, ESCs also experience a different microenvironment to epiblast cells in the embryo. These different microenvironments could potentially affect the selective pressure and hence the transcriptional signature of loser cells. However, there are two reasons why we think that the loser cell signatures identified here are conserved across systems. First, the transcriptional profile of our epiblast cells from cultured embryos is very similar to that of epiblast cells from freshly isolated embryos (Extended Data Fig. 3e–g). Second, the loser signature identified here is enriched for targets of P53 and depleted for mTOR and c-MYC targets. Given that these are regulators of cell competition identified by us and others in the embryo and in ESCs[5–7], it suggests that the same pathways are inducing loser cell elimination in in vivo, ex vivo and in ESC models of cell competition.

It is well known that the successful development of the embryo can be influenced by the quality of its mitochondrial pool[10]. Moreover, divergence from normal mitochondrial function during embryogenesis can either be lethal or lead to the development of mitochondrial disorders[49]. Deleterious mtDNA mutations are a common cause of mitochondrial diseases and, during development, selection against mutant mtDNA has been described to occur through at least two mechanisms: the bottleneck effect and the intracellular purifying selection. The bottleneck effect is associated specifically with the unequal segregation of mtDNAs during primordial germ cell specification, for example, as seen in the human embryo[50]. In contrast to this, purifying selection, as the name implies, allows for selection against deleterious mtDNAs and has been proposed to take place during both development and postnatal life[51]. Importantly, purifying selection has been found to occur at the molecule and organelle levels, as well as at the cellular level[52]. Our findings indicate that purifying selection can occur not only at the intracellular level but also at the intercellular level (cell non-autonomously). We show that epiblast cells can sense their relative mitochondrial activity and that those cells with mtDNA mutations or with lower or aberrant mitochondrial function are eliminated. By selecting those cells with the most favourable mitochondrial performance, cell competition would not only prevent cells with mitochondrial defects from contributing to the germline or future embryo, but also ensure optimization of the bioenergetic performance of the epiblast, therefore contributing to the synchronization of growth during early development.

Cell competition has been studied in a variety of organisms, from *Drosophila* to mammals, and it is likely that multiple different mechanisms fall under its broad umbrella[1–3]. Despite this, there is considerable interest in understanding if there could be any common feature in at least some of the contexts where cell competition has been described. The first demonstration of cell competition in *Drosophila* was made by inducing clones carrying mutations in the ribosomal gene *Minute*[4] and this has become one of the primary models to study this process. Our finding that during normal early mouse development cell competition eliminates cells carrying mutations in *mt-Rnr1* and *mt-Rnr2* transcripts, demonstrates that in the physiological context mutations in ribosomal genes also trigger cell competition. While we identified 11 mutations specific to loser cells, we cannot exclude the presence of additional variants differentiating winners from losers in those positions that did not have sufficient coverage in our RNA-seq data. Our observation that mis-patterned and karyotypically abnormal cells show impaired mitochondrial activity indicates that during early mouse development different types of defects impair mitochondrial function and

trigger cell competition. Interestingly, mtDNA genes are amongst the top mis-regulated factors identified during cell competition in mouse skin[53]. In the *Drosophila* wing disc oxidative stress, a general consequence of dysfunctional mitochondria, underlies the out-competing of *Minute* and *Mahj* mutant cells[54]. Similarly, in Madin–Darby canine kidney cells, a loss of $\Delta\psi m$ occurs during the out-competing of RasV12 mutant cells and is key for their extrusion[55]. These observations raise the possibility that differences in mitochondrial activity may be a key determinant of competitive cell fitness in a wide range of systems. Unravelling which mitochondrial features lead to cellular differences that can be sensed between cells during cell competition and if these are conserved in human systems will be key not only for understanding this process, but also to open up the possibility for future therapeutic avenues in the diagnosis or prevention of mitochondrial diseases.

Methods

Animals. Mice were maintained and treated in accordance with the Home Office's Animals (Scientific Procedures) Act 1986 and covered by the Home Office project licence PBBEBDCDA. All mice were housed on a 10–14-h light–dark cycle with access to water and food ad libitum. All mice were housed within individually ventilated cages. Temperature was maintained between 21–24 °C and humidity between 45–65%. Mattings were generally set up in the afternoon. Noon on the day of finding a vaginal plug was designated as E0.5. Embryo dissection was performed at appropriate time points in M2 medium (Sigma), using Dumont no.5 forceps (11251-10, FST). No distinction was made between male and female embryos during the analysis.

Cell lines, cell culture routine and drug treatments. E14 mESCs (RRID: CVCL_C320), kindly provided by A. Smith from Cambridge University, were used as wild-type control tdTomato-labelled or unlabelled cells. GFP-labelled or unlabelled cells defective for BMP signalling (*Bmpr1a*$^{-/-}$), tetraploid cells (4n) and *Bmp1a*$^{-/-}$ null for *p53* (*Bmpr1a*$^{-/-}$;*p53*$^{-/-}$) are described elsewhere[6,7]. *Drp1*$^{-/-}$ or *Mfn2*$^{-/-}$ cells were generated by CRISPR mutagenesis. Cells with different mtDNA content in the same nuclear background were derived from embryos of hybrid mice, generated elsewhere[16].

Cells were maintained at pluripotency and cultured at 37 °C in 5% CO_2 in 25-cm^2 flasks (Nunc) coated with 0.1% gelatin (Sigma) in DPBS. Growth medium (ES medium) consisted of GMEM supplemented with 10% FCS, 1 mM sodium pyruvate, 2 mM L-glutamine, 1× minimum essential medium non-essential amino acids and 0.1 mM β-mercaptoethanol (all from Gibco) and 0.1% leukaemia inhibitory factor (LIF, produced and tested in the laboratory). Cells derived from hybrid mice (C57BL/6N nuclear background) were maintained on 0.2% LIF. The growth medium was changed daily, and cells were split every 3 d.

To manipulate mitochondrial translation during differentiation, wild-type and *Drp1*$^{-/-}$ mESCs were treated with doxycycline (22.5 µM), from day 1 to day 3 of culture, or with actinonin (150 µM), for 6 h on day 3 of culture in N2B27 medium ('Differentiation and cell competition assays'). As the control condition, cells were treated with vehicle. Samples were collected on day 3 of differentiation for western blot analysis.

CRISPR mutagenesis. *Drp1* and *Mfn2* knockout ESCs were generated by CRISPR–Cas9-mediated deletion of *Drp1* exon 2 and *Mfn2* exon 3, respectively. sgRNA guides flanking *Drp1* exon 2 or *Mfn2* exon 3 were cloned into the PX459 vector (Addgene)[56]: *Drp1* exon 2 upstream sgRNA: 5′ TGGAACGGTCACAGCTGCAC 3′; *Drp1* exon 2 downstream sgRNA: 5′ TGGTCGCTGAGTTTGAGGCC 3′; *Mfn2* upstream sgRNA: 5′ GTGGTATGACCAATCCCAGA 3′; *Mfn2* downstream sgRNA: 5′ GGCCGGCCACTCTGCACCTT 3′. E14 ESCs were co-transfected with 1 µg of each sgRNA expression using Lipofectamine 2000 (Invitrogen) according to the manufacturer's instructions. As the control, E14 ESCs were transfected in parallel with an equal amount of empty PX459 plasmid. Following 6 d of puromycin selection, single colonies were picked from both *Drp1* sgRNA and ESCs transfected with empty vector and screened for mutations. *Drp1* exon 2 deletion was confirmed by PCR genotyping using the following primers: Drp1_genot F: 5′ GGATACCCCAAGATTTCTGGA 3′; Drp1_genot R: 5′ AGTCAGGTAATCGGGAGGAAA 3′, followed by Sanger sequencing. *Mfn2* exon 3 deletion was confirmed by PCR genotyping using the following primers: Mfn2_genot F: 5′ CAGCCCAGACATTGTTGCTTA 3′; Mfn2_genot R: 5′ AGCTGCCTCTCAGGAAATGAG 3′, followed by Sanger sequencing.

Derivation of mouse embryonic stem cells from hybrid mouse strains. The derivation of new mESC lines was adapted from work by Czechanski et al.[57]. Cells were derived from embryos of hybrid mouse strains BG, HB and ST. These contain the mtDNA of C57BL/6N (BL6) laboratory mouse and mtDNA variants from wild-caught mice[16].

Embryos were isolated at E2.5 (morula stage) and cultured in four-well plates (Nunc, Thermo Scientific) containing KSOM medium (Millipore) plus two inhibitors (KSOM + 2i): 1 µM MEK inhibitor PDO325901 (Sigma-Aldrich) and 3 µM GSK-3 inhibitor CHIR9902 (Cayman Chemicals) for 2 d at 37 °C in a 5% CO_2 incubator. To reduce evaporation, the area surrounding the wells was filled with DPBS. Embryos were further cultured in fresh 4-well plates containing N2B27 + 2i + LIF medium: N2B27 medium supplemented with 1 µM MEK inhibitor PDO325901 and 3 µM GSK-3 inhibitor and 0.1% LIF for up to 3 d until reaching the blastocyst stage. Each embryo was then transferred to a well of a 96-well plate coated with 0.1% gelatin in DPBS and containing 150 µl of N2B27 + 2i + LIF medium per well. In these conditions, the embryos should attach to the wells allowing the epiblast to form an outgrowth. This plate was then incubated at 37 °C in a 5% CO_2 incubator for 3 to 7 d until ESC-like colonies start to develop from the epiblast outgrowth. Cells were passaged by dissociation with Accutase (Sigma) and seeded in gradually increasing growth surface areas (48-well, 24-well and 12-well plates; T12.5 and T25 flasks), until new cell lines were established. At this stage, cells were weaned from N2B27 + 2i + LIF medium and then routinely cultured in ESC medium.

These new cell lines were then subjected to characterization by flow cytometry (cell size, granularity and mitochondrial $\Delta\psi m$) and amplification refractory mutation system (ARMS)–qPCR assay[16] to determine heteroplasmy.

Heteroplasmy quantification by ARMS–qPCR assay. Every qPCR run consisted of the consensus and an ARMS assay.

Consensus assay. CO2-F: TCTTATATGGCCTACCCATTCCAA, CO2-R: GGAAA ACAATTATTAGTGTGTGATCATG, CO2-FAM: 6FAM-TTGGTCTACAAGAC GCCACATCCCCT-BHQ-1

(amplicon length: 103 bp)

ARMS assays. 16SrRNA2340/Staudach-f: AAACCAACATATCTCATTGACCgAA (haplotype ST), *16SrRNA*2340(3)G-f: AATCAACATATCTTATTGACCaAG (haplotype C57BL/6N), *16SrRNA*2340(3)A-f: AATCAACATATCTTATTGACCgAA (haplotypes BG and HB), *16SrRNA*2458-r: CAC CAT TGG GAT GTC CTG ATC, 16SrRNA-FAM: FAM-CAA TTA GGG TTT ACG ACC TCG ATG TT-BHQ-1.

Lower-case letters indicate the intentional mismatch (ARMS), underlined letters indicate SNP-specific bases (amplicon length: 142 bp for BG and HB; 143 bp for ST).

Master-mixes for triplicate qPCR reactions contained 1× buffer B2 (Solis BioDyne), 4.5 mM $MgCl_2$, 200 µM of the four deoxynucleotides (dNTPs, Solis BioDyne), 0.7 units HOT FIREPol DNA polymerase (Solis BioDyne), 300 nM of each primer and 100 nM hydrolysis probe. For each reaction, 12 µl of master-mix and 3 µl DNA were transferred in triplicates to 384-well PCR plates (Life Technologies) using the automated pipetting system epMotion 5075TMX (Eppendorf). Amplification was performed on the ViiA 7 Real-Time PCR System using the ViiA 7 software v1.1 (Life Technologies). DNA denaturation and enzyme activation were performed for 15 min at 95 °C. DNA was amplified over 40 cycles consisting of 95 °C for 20 s, 58 °C for 20 s and 72 °C for 40 s for all assays.

The standard curve method was applied. Amplification efficiencies were determined for each run separately by DNA dilution series consisting of DNA from mice harbouring the respective analysed mtDNA. Typical results were: slope = −3.462, −3.461, −3.576 and −3.668; mean efficiency = 0.95, 0.94, 0.90 and 0.87; and *y* intercept = 32.4, 33.8, 34.5 and 31.9; for the consensus, C57BL/6N, HB and BG, and ST assays, respectively (Supplementary Figs. 1–4). Coefficient of correlation was ≥0.99 in all assays in all runs. All target samples were within the linear interval of the standard curves. To test for specificity, in each run, a negative control sample, that is, a DNA sample of a mouse harbouring the mtDNA of the non-analysed type in the heteroplasmic mouse (C57BL/6N or the respective wild-derived mtDNA) was measured. All assays could discriminate between C57BL/6N and wild-derived mouse mtDNA at a minimum level of >1%. Target sample DNA was tested for inhibition by dilution in Tris-EDTA buffer (pH 8.0).

For the calculation of mtDNA heteroplasmy, the assay detecting the minor allele (C57BL/6N or wild-derived mice, <50%) was always used. If both specific assays gave values >50% (that can happen at around 50% heteroplasmy), the mean value of both assays was taken. All qPCR runs contained no template controls for all assays; these were negative in 100% of analyses.

ARMS–qPCR standard curves and detection limit. mtDNA heteroplasmy was quantified by ARMS–qPCR, an established method in the field[16,19,58–63]. Calibration curves were created with a dilution series of DNA that showed a 100% match with the respective assay. Therefore, for all assays, necessarily divergent dilution series had to be used. The amount of DNA between the dilution series can diverge and thus values were plotted as arbitrary units. Supplementary Fig. 1 shows the standard curve produced for the consensus assay (detecting *mt-Co2* as a measure of total mtDNA) and Supplementary Figs. 2–4 show standard curves produced for specific mtDNA variants (laboratory mouse mtDNA, C57BL/6N; wild-derived mice mtDNAs BG, HB and ST).

SNP-specific quantification of mtDNA. To test the SNP-specific quantification of mtDNA, mixtures of match and mismatch DNA were analysed in triplicates. All assays could discriminate between C57BL/6N and wild-derived mouse mtDNA (and vice versa) at a minimum level of 1%, as shown by the ARMS–qPCR typical false-positive signal with the 100% mismatch DNA (detection limit, in all assays below 0.3%). The results and amplification plots for the specific quantification of HB and BG wild-derived mouse mtDNA from C57BL/6N mtDNA are shown in Supplementary Fig. 5 and Supplementary Table 9. The results and amplification plots for the specific quantification of ST wild-derived mouse mtDNA from C57BL/6N mtDNA are available in Supplementary Fig. 6 and Supplementary Table 10. Average values of the triplicate values are shown.

Embryo experiments. Early mouse embryos were isolated at E5.5 (from pregnant CD1 females, purchased from Charles River). Following dissection from the decidua, embryos were cultured overnight in poor N2B27 medium (same formulation as N2B27 medium but supplemented with 0.5× B27 supplement and 0.5× N2 supplement) with pan-CIs (100 µM, Z-VAD-FMK, FMK001, R&D Systems) or an equal volume of vehicle (DMSO) as the control. On the next morning, embryos were processed for scRNA-seq or functional validation (Δψm analysis and immunohistochemistry for markers of loser cells).

For the scRNA-seq and Δψm analysis, embryos were dissociated into a single-cell suspension. Briefly, up to 12 embryos were dissociated in 600 µl Accutase (A6964, Sigma) over 12 min at 37 °C, with tapping of the tube at 2-min intervals. Accutase was then neutralized with an equal volume of FCS, cells were spun down and stained with TMRM (for Δψm analysis) or directly resuspended in 300 µl DPBS with 1% FCS (for single-cell sorting and RNA-seq). Sytox Blue (1:1,000 dilution, S34857, Thermo Fisher Scientific) was used for viability staining.

Differentiation and cell competition assays. Cell competition assays between wild-type cells and *Bmpr1a*$^{-/-}$, 4n or *Drp1*$^{-/-}$ cells were performed in differentiating conditions. Cells were seeded onto fibronectin-coated plates (1:100, Merck) in DPBS for 1 h at 37 °C and grown in N2B27 medium to promote the differentiation of mESCs into a stage resembling the post-implantation epiblast, as cell competition was previously shown to occur in these conditions[6]. N2B27 medium consisted of 1:1 DMEM/F12 nutrient mixture and Neurobasal medium supplemented with N2 (1×) and B27 (1×) supplements, 2 mM L-glutamine and 0.1 mM β-mercaptoethanol (all from Gibco). Cell competition assays between wild-type and *Mfn2*$^{-/-}$ cells and between mESCs with different mtDNA content were performed in conditions of pluripotency maintenance (ESC medium).

Cells were either seeded separately or mixed for co-cultures at a 50:50 ratio, onto 12-well plates, at a density of 8×10^4 cells per well, except for assays between wild-type and *Mfn2*$^{-/-}$ mESCs, where 3.2×10^5 cells were seeded per well. The growth of cells was followed daily and compared between separate cultures or co-cultures, to control for cell-intrinsic growth differences, until the fourth day of culture. Viable cells were counted daily using a Vi-CELL XR Analyser (Beckman Coulter), and proportions of each cell type in co-cultures were determined using an LSR II Flow Cytometer (BD Bioscience), based on the fluorescent tag of the ubiquitously expressed GFP or TdTomato in one of the cell populations.

Metabolomic analysis. The metabolic profile was obtained using the Metabolon Platform (Metabolon). Each sample consisted of five biological replicates. For each replicate, 1×10^7 cells were spun down and snap frozen in liquid nitrogen. Pellets from five independent experiments for each condition were analysed by Metabolon using a combination of ultra-high performance liquid chromatography–tandem mass spectroscopy (UHPLC–MS/MS) and gas chromatography–mass spectroscopy (GC–MS). Compounds were identified by comparison to library entries of purified standards based on the retention time/index, mass-to-charge ratio (*m/z*) and chromatographic data (including MS/MS spectral data) on all molecules present in the library. Samples were normalized to protein content measured by Bradford assay. Statistical analysis was performed using Welch's two-sample *t*-test and statistical significance was defined as a P value ≤ 0.05.

Seahorse analysis. The metabolic function of cells was assessed by extracellular flux analysis using Seahorse XF24 (Agilent Technologies). For assays ran during pluripotency, cells were seeded, on the day before the assay, onto 0.1% gelatin-coated plates (Sigma) in 300 µl of ESC medium. All cell types were seeded at 5×10^4 cells per well, except for *Bmpr1a*$^{-/-}$ cells, which were seeded at 5×10^4 cells per well. For assays ran during differentiation, cells were seeded, 3 d before the assay, onto fibronectin-coated plates (1:100 dilution; Merck) in 300 µl of N2B27 medium. All cell types were seeded at 2.4×10^4 cells per well, except for *Bmpr1a*$^{-/-}$ cells, which were seeded at 3.2×10^4 cells per well.

On the day of the assay, cells were carefully washed twice with assay medium and then left with a final volume of 600 µl per well. The plate was then equilibrated on a non-CO_2 incubator at 37 °C for 30 min. The assay medium consisted of unbuffered DMEM (D5030, Sigma) that was supplemented on the day of the assay according to the test performed. For the OCR measurements, the assay medium was supplemented with 0.5 g l^{-1} glucose (Sigma) and 2 mM L-glutamine (Life Technologies), while for the ECAR measurements, the medium was supplemented with 1 mM sodium pyruvate and 2 mM L-glutamine (both from Life Technologies) at pH 7.4 and 37 °C.

The protocol for the assay consisted of four baseline measurements and three measurements after each compound addition. Compounds (all from Sigma) used in OCR and ECAR assays were prepared in the supplemented assay medium. For the OCR assay, the following compounds were added: 1 mM pyruvate, 2.5 µM oligomycin, 300 nM carbonyl cyanide-4-(trifluoromethoxy) phenylhydrazone and a mixture of rotenone and antimycin A at 6 µM each (R&A). For the ECAR assay, the following compounds were added: 2.5 mM and 10 mM of glucose, 2.5 µM of oligomycin and 50 mM of 2-deoxyglucose.

Each of the experiments was performed three times, with five biological replicates of each cell type. For background correction measurements, four wells were left without cells (A1, B4, C3 and D6). ECAR and OCR measurements were performed on the same plate. The assay parameters for both tests were calculated following the Seahorse assay report generator (Agilent Technologies).

At the end of the assay, cells were fixed and stained with Hoechst. Both OCR and ECAR were normalized to cell number, determined by manual cell counts using Fiji software. The normalization of the data was processed on Wave Desktop software (Agilent Technologies) and data were exported to Prism 8 (GraphPad) for statistical analysis.

Analysis of mitochondrial membrane potential and reactive oxygen species. For TMRM staining in single cells from early mouse epiblasts, embryos were dissected at E5.5 and cultured overnight in the presence or absence of CIs. On the following morning, to avoid misleading readings, epiblasts were isolated initially by an enzymatic treatment with 2.5% pancreatin, 0.5% trypsin and 0.5% polyvinylpyrrolidone (PVP40), all from Sigma-Aldrich, to remove the visceral endoderm. Embryos were treated for 8 min at 4 °C, followed by 2 min at room temperature (RT). The visceral endoderm was then peeled with the forceps and the extra-embryonic ectoderm was removed to isolate the epiblasts. Twelve epiblasts were pooled per 600 µl of Accutase (Sigma-Aldrich) for dissociation into single cells before staining. The reaction was stopped with an equal volume of FCS and cells were subjected to TMRM staining. Cells were incubated in 200 µl of 10 nM Nernstian probe TMRM perchlorate (T5428, Sigma), prepared in N2B27 medium. After incubation for 15 min at 37 °C, cells were pelleted again and resuspended in flow cytometry (FC) buffer (3% FCS in DPBS). Sytox Blue (1:1,000 dilution; S34857, Thermo Fisher Scientific) was used as viability staining.

Quantitative analysis of Δψm and mitochondrial ROS was performed by flow cytometry. Cells were grown in pluripotency or differentiating conditions. Cells were dissociated and pelleted to obtain 2×10^5 cells per sample for the staining procedure. For TMRM staining in mESCs, 2×10^5 cells of each cell line were resuspended in 200 µl of 10 nM TMRM (T5428, Sigma), prepared in N2B27 medium. Cells were incubated at 37 °C for 15 min, and then resuspended in FC buffer (3% FCS in DPBS). For the analysis of mitochondrial ROS, cells were grown in differentiating conditions and stained on the third day of culture. Briefly, 2×10^5 cells of each cell line were resuspended in 200 µl of a 5-µM solution of MitoSOX (M36008, Invitrogen) prepared in N2B27 medium. Cells were incubated at 37 °C for 15 min, and then resuspended in FC buffer. Sytox Blue was used for viability staining.

Cell suspensions stained with TMRM or MitoSOX were analysed in a BD LSR II flow cytometer operated through FACSDiva software (Becton Dickinson Biosciences). For TMRM fluorescence detection, the yellow laser was adjusted for excitation at $\lambda = 562$ nm, capturing the emission light at $\lambda = 585$ nm for TMRM. MitoSOX fluorescence was analysed with the violet laser adjusted for excitation at $\lambda = 405$ nm, capturing the emission light at $\lambda = 610$ nm. In the case of GFP-labelled cell lines, for GFP fluorescence detection, the blue laser was adjusted for excitation at $\lambda = 488$ nm, capturing the emission light at $\lambda = 525$ nm. Results were analysed in FlowJo Software v9 or v10.0.7r2. See the FACS gating strategy in Supplementary Fig. 7.

Qualitative analysis of Δψm was performed by confocal microscopy. Wild-type and *Bmpr1a*$^{-/-}$ cells were grown in fibronectin-coated glass coverslips. On the third day of differentiation, cells were loaded with a 200-nM MitoTracker Red probe (Life Technologies), prepared in N2B27 medium, for 15 min at 37 °C. Cells were then washed with DPBS and fixed with 3.7% formaldehyde for subsequent immunocytochemical staining of total mitochondrial mass, with TOMM20 antibody.

Immunofluorescence. Cells were washed with DPBS and fixed with 3.7% formaldehyde (Sigma) in N2B27, for 15 min at 37 °C. Permeabilization of the cell membranes was performed with 0.4% Triton X-100 in DPBS (DPBS-Tx), at RT with agitation. The blocking step with 5% BSA in DPBS-Tx 0.1% was performed for 30 min, at RT with agitation. Mitochondria were labelled with TOMM20 antibody (1:100 dilution; Santa Cruz Biotechnologies). Dead cells were labelled with cleaved caspase-3 antibody (1:400 dilution; CST-9664), and NANOG antibody was used to mark pluripotent cells (1:100 dilution; eBioscience). Secondary antibodies were Alexa Fluor 488 and 568 (1:600 dilution; Invitrogen). Primary antibody incubation was performed overnight at 4 °C and secondary antibody incubation was done for 45 min, together with Hoechst to stain nuclei (1:1,000 dilution; Thermo Scientific) at RT and protected from light. In both cases, antibodies were diluted in blocking solution. Three 10-min washes with DPBS-Tx 0.1% were performed between each critical step and before mounting with Vectashield medium (Vector Laboratories).

Samples were imaged with a Zeiss LSM 780 confocal microscope and processed with Fiji[64]. Mitochondrial stainings were imaged with a ×63/1.4 oil objective. For samples stained with TOMM20 antibody and MitoTracker Red, *z*-stacks were acquired and processed for deconvolution using Huygens software (Scientific Volume Imaging; https://svi.nl/). Samples stained with cleaved caspase-3 were imaged with a ×20/0.8 air objective. Imaging and deconvolution analysis were performed with support and advice from S. Rothery from the Facility for Imaging by Light Microscopy (FILM) at Imperial College London.

Embryo immunofluorescence staining for p-rpS6, OPA1 and DDIT3 (CHOP) markers was performed as follows. Cultured embryos were fixed in 4% paraformaldehyde in DPBS containing 0.01% Triton and 0.1% Tween 20 for 20 min at RT. Permeabilization of the membranes was performed for 10 min in DPBS with 0.5% Triton. Embryos were blocked in 5% BSA in DPBS with 0.25% Triton during 45 min. Incubation with primary antibodies (CHOP (1:500 dilution; CST, 2895), OPA1 (1:100 dilution; BD Biosciences, 612606) and p-rpS6 (1:200 dilution; CST, 5364)) was completed overnight at 4 °C in 2.5% BSA in DPBS with 0.125% Triton. The following morning, hybridization with secondary antibodies Alexa Fluor 568 and Alexa Fluor 488 (Invitrogen, diluted at 1:600 in DPBS with 2.5% BSA and 0.125% Triton) was performed next for 1 h at RT. Hoechst was also added to this mixture to stain nuclei (1:1,000 dilution; Invitrogen). Three 10-min washes with filtered DPBS-Tx 0.1% were performed between each critical step. All steps included gentle agitation.

Embryos were imaged in embryo dishes (Nunc) in a drop of Vectashield using a Zeiss LSM 780 confocal microscope at ×40/1.3 oil objective.

Further details about image acquisition and processing are specified in Supplementary Table 8.

Western blotting. Cells were washed in DPBS and lysed with Laemmli lysis buffer (0.05 M Tris-HCl at pH 6.8, 1% SDS, 10% glycerol and 0.1% β-mercaptoethanol in distilled water). Total protein quantification was done using BCA assay (Thermo Scientific) and samples (15 µg of protein per lane) were loaded into 12% Bis-Tris protein gels (Bio-Rad). Resolved proteins were transferred into nitrocellulose membranes (GE Healthcare). The following primary antibodies, prepared in TBS-0.1% Tween containing 5% BSA were incubated overnight at 4 °C with gentle agitation: rabbit anti-TOMM20 (1:1,000 dilution; CST, 42406), mouse anti-ATPB (1:1,000 dilution; Abcam, ab14730), rabbit anti-α-tubulin (1:1,000 dilution; CST, 2144), mouse anti-mt-CO1 (1:2,000 dilution; Abcam, ab14705), rabbit anti-DRP1 (1:1,000 dilution; CST, 8570), mouse anti-MFN1 (1:1,000 dilution; Abcam, ab57602), mouse anti-MFN2 (1:500 dilution; Abcam, ab56889), mouse anti-vinculin (1:1,000 dilution; Sigma, V9131), mouse anti-OPA1 (1:1,000 dilution; BD Biosciences, 612606), rabbit anti-ATF4 (1:1,000 dilution; CST, 11815), rabbit anti-PCNA (1:5,000 dilution; Abcam, ab18197) and rabbit anti-p-eIF2α (Ser51; 1:1,000 dilution; CST, 9721). On the following morning, HRP-conjugated secondary antibodies (1:5,000 dilution; sc-2004 and sc-2005, Santa Cruz), prepared in TBS-0.1% Tween containing 5% milk (Sigma) were incubated for 1 h at RT under gentle agitation. Membranes were developed with ECL reagents (Promega) and mounted in cassettes for time-controlled exposure to film (GE Healthcare).

Bulk and single-cell RNA sequencing. For bulk RNA-seq in the competitive scenario between cells with different mtDNA, HB (24%) and BG (95%) mESCs were grown separately or in co-culture. On the third day of culture, cells were dissociated and subjected to FACS to separate the cell populations in co-culture according to their GFP label. Propidium iodine (1:1,000 dilution; 81845, Sigma) was used for viability staining. See the FACS gating strategy in Supplementary Fig. 8. To control for eventual transcriptional changes due to the FACS process, a mixture of the two separate populations was subjected to the same procedure as the co-cultured samples. Total RNA isolation was then carried out using RNA extraction Kit (RNeasy Mini Kit, QIAGEN). PolyA selection/enrichment was the method adopted for library preparation, using the NEB Ultra II RNA Prep Kit. Single-end 50-bp libraries were sequenced on an Illumina HiSeq 2500. Raw base call files were converted to fastq files using Illumina's bcl2fastq (v2.1.7). Reads were aligned to the mouse genome (mm9) using Tophat2 (v2.0.11)[65] with default parameters. Mapped reads that fell on genes were counted using featureCounts from Rsubread package[66]. Generated count data were then used to identify differentially expressed genes using DESeq2 (ref. [67]). Genes with very low read counts were excluded. Finally, GSEA was performed using GSEA software[68,69] on a pre-ranked list generated by DESeq2.

To investigate the nature of cells eliminated by cell competition during early mouse embryogenesis by means of scRNA-seq, early mouse embryos were dissected at E5.5 and cultured overnight in the presence or absence of CIs. The next morning, embryos were dissociated with Accutase and subjected to single-cell sorting into 384-well plates. Total RNA isolation was then carried out using an RNA extraction Kit (RNeasy Mini Kit, QIAGEN). scRNA-seq was performed using the Smart-seq2 protocol[70]. PolyA selection/enrichment with Ultra II Kit (NEB) was the method adopted for library preparation.

Data processing, quality control and normalization. We performed transcript quantification in our scRNA-seq data by running Salmon (v0.8.2)[71] in the quasi-mapping-based mode. First, a transcriptome index was created from the mouse reference (version GRCm38.p4) and ERCC spike-in sequences. Then, the quantification step was carried out with the 'quant' function, correcting for the sequence-specific biases ('--seqBias' flag) and the fragment-level GC biases ('--gcBias' flag). Finally, the transcript-level abundances were aggregated to gene-level counts. On the resulting raw count matrix including 1,495 cells, we applied a quality-control check to exclude poor quality cells from downstream analyses.

For quality control, we applied the following criteria: identification of cells that had a $\log_{10}$ total number of reads equal to or greater than 4, a fraction of mapped reads equal to or greater than 0.8, a number of genes with an expression level above ten reads per million equal to or greater than 3,000 and a fraction of reads mapped to endogenous genes equal to or greater than 0.5. This resulted in the selection of 723 cells, which were kept for downstream analyses. Transcripts per million (TPM) normalization (as estimated by Salmon) was used.

Highly variable genes and dimensionality reduction. To identify highly variable genes (HVGs), we first fitted a mean and total variance trend using the R function 'trendVar' and then the variance was decomposed into biological and technical components with the R function 'decomposeVar'; both functions are included in the package 'scran' (v1.6.9)[72].

We considered HVGs those with a biological component that was significantly greater than zero at an FDR (Benjamini–Hochberg method) of 0.05. Then, we applied further filtering steps by keeping only genes that had an average expression greater to or equal than 10 TPM and were significantly correlated with one another (function 'correlatePairs' in 'scran' package, FDR < 0.05). This yielded 1,921 genes, which were used to calculate a distance matrix between cells defined as $\sqrt{(1-\rho)/2}$, where ρ is the Spearman's correlation coefficient between cells. A two-dimensional representation of the data was obtained with the UMAP package (v0.2.0.0; https://cran.r-project.org/web/packages/umap/index.html) using the distance matrix as input.

Cell clustering and connectivity analysis. To classify cells into different clusters, we ran hierarchical clustering on the distance matrix (see above; 'hclust' function in R with ward.D2 aggregation method) followed by the dynamic hybrid cut algorithm ('cutreeDynamic' function in R package 'dynamicTreeCut' (https://CRAN.R-project.org/package=dynamicTreeCut) v1.63.1, with the hybrid method, using a minimum cluster size of 35 cells and a 'deepSplit' parameter equal to 0), which identified five clusters. Cells from different batches were well mixed across these five clusters (Extended Data Fig. 1), suggesting that the batch effect was negligible. The identity of the five clusters was established based on the expression of known marker genes of epiblasts, visceral endoderm and extra-embryonic ectoderm, which were identified in a previous study[73]. The expression levels of some of the top markers are plotted in Fig. 1b.

We performed a robustness analysis on the clustering by exploring in detail how the choices of genes, clustering parameters and algorithms affect the identity and the number of clusters. First, we quantified the cluster robustness by calculating Pearson's gamma and the average silhouette width obtained with 100 random subsets of 60% of the HVGs and different values of the deepSplit parameter. While the robustness at a deepSplit value of 0 and 1 was similar, for greater values of deepSplit (corresponding to less conservative clustering), the robustness rapidly declined (Extended Data Fig. 1e). The clustering with deepSplit value of 0 and 1 (the more robust choices) yielded very similar results, the only difference being the splitting of the intermediate cluster in two subclusters (Extended Data Fig. 1f).

In addition to this, we also used Louvain clustering on the HVGs (resolution = 0.3, k = 20 with 20 principal components), which again produced very similar clusters.

We quantified the connectivity between the clusters (using only CI-treated cells) with PAGA[22] implemented in the Python library scanpy (v1.4.7)[74]. The analysis revealed that the three epiblast clusters were connected with each other, whereas the two extra-embryonic tissues (visceral endoderm and extra-embryonic ectoderm) were isolated (Extended Data Fig. 2b).

Identification of a single-cell trajectory in the epiblast. We calculated a diffusion map ('DiffusionMap' function in the R package 'destiny' (v2.6.2)[23]) on the distance defined above on the epiblast cells from CI-treated embryos. The pseudotime coordinate was computed with the 'DPT' function with the root cell in the winner epiblast cluster (identified by the function 'tips' in the 'destiny' package). Such pseudotime coordinates can be interpreted as a 'losing score' for all the epiblast cells from the CI-treated embryos.

We estimated the losing scores of the epiblast cells from DMSO-treated embryos by projecting such data onto the diffusion map previously calculated (function 'dm_predict' in the destiny package). Finally, for each of the projected cells, we assigned the losing score as the average of the losing scores of the ten closest neighbours in the original diffusion map (detected with the function 'projection-dist' in the destiny package).

While for the clustering and the trajectory analysis we used the HVGs computed from the whole dataset, we verified that all results concerning the separation between winner and loser epiblast cells (for example, clusters and losing score) remain unaffected if the HVGs are calculated using only the epiblast cells.

Mapping of data from epiblast cells onto published datasets. We compared the transcriptional profile of epiblasts from embryos cultured in DMSO and CI with that of epiblasts collected from freshly isolated embryos at different stages.

To do this, we considered a dataset published previously[26], which includes epiblast cells from embryos at the stages E5.5 (102 cells), E6.25 (130 cells) and E6.5 (288 cells). A diffusion map and a diffusion pseudotime coordinate were computed with these cells following the same procedure described above (Extended Data Fig. 3e,f). Then, we projected epiblast cells from CI-treated and DMSO-treated embryos and we assigned to them a diffusion pseudotime coordinate as described above (Extended Data Fig. 3g).

Differential gene expression analysis along the trajectory. To identify the genes that were differentially expressed along the trajectory, we first kept only genes that had more than 15 TPM in more than ten cells (this list of genes is provided in Supplementary Table 4); then, we obtained the log-transformed expression levels of these genes (adding 1 as a pseudo-count to avoid infinities) as a function of the losing score and we fitted a generalized additive model (GAM) to them (R function 'gam' from 'GAM' package version 1.16.). We used the ANOVA test for parametric effects provided by the 'gam' function to estimate a P value for each tested gene. This yielded a list of 5,311 differentially expressed genes (FDR < 0.01).

Next, we looked for groups of differentially expressed genes that shared similar expression patterns along the trajectory. To this aim, similarly to what we did when clustering cells, we calculated a correlation-based distance matrix between genes, defined as $\sqrt{(1-\rho)/2}$, where ρ is the Spearman's correlation coefficient between genes. Hierarchical clustering was then applied to this matrix ('hclust' function in R, with the 'ward.D2' method) followed by the dynamic hybrid cut algorithm ('dynamicTreeCut' package) to define clusters ('cutreeDynamic' function in R with the hybrid method and a minimum cluster size of 100 genes and a deepSplit parameter equal to 0). This resulted in the definition of four clusters, including three clusters of genes that decreased along the trajectory (merged together for the Gene Ontology enrichment and the IPA analysis) and one cluster of increasing genes (Extended Data Fig. 3a). IPA (QIAGEN; https://www.qiagenbioinformatics.com/products/ingenuity-pathway-analysis/), was run on all genes differentially expressed (FDR < 0.01) along the trajectory from winner to loser cells (Figs. 2a–d and 3a–c), using all the tested genes as a background (Supplementary Table 4). This software generated networks, canonical pathways and functional analysis. The list of decreasing/increasing genes is provided in Supplementary Tables 1 and 2. The pathways found as mis-regulated in Fig. 3 were: mitochondrial dysfunction, $-\log_{10}(P\text{ value}) = 21.1$; OXPHOS, $-\log_{10}(P\text{ value}) = 18.6$; EIF2 signalling, $-\log_{10}(P\text{ value}) = 11.9$. FDRs for the genes shown in Fig. 3b range from 1.25×10^{-51} (for *Atp5b*) to 5.42×10^{-3} (for *Ndufa11*). *Cox6b2* was found to be upregulated in loser cells (FDR = 2.69×10^{-13}).

Analysis of heteroplasmy in a single-cell RNA-seq dataset. We used STAR (v2.7)[75] to align the transcriptome of the epiblast cells from CI-treated embryos (274) to the mouse reference genome (mm10). Only reads that uniquely mapped to the mtDNA were considered. From these, we obtained allele counts at each mtDNA position with a Phred quality score greater than 33 using the samtools 'mpileup' function.

Next, we applied filters to remove cells and mtDNA positions with a low coverage. First, we removed cells with fewer than 2,000 mtDNA positions covered by more than 50 reads. Second, we removed positions having less than 50 reads in more than 50% of cells in each of the three epiblast clusters (winner, intermediate and loser). These two filters resulted in 259 cells and 5,192 mtDNA positions (covered by ~700 reads per cell on average) being considered for further analyses.

Starting from these cells and positions, we applied an additional filter to keep only positions with a sufficiently high level of heteroplasmy. To this aim, for each position with more than 50 reads in a cell, we estimated the heteroplasmy according to equation (1):

$$H = 1 - f_{max} \quad (1)$$

where f_{max} is the frequency of the most common allele. We kept only positions with $H > 0.01$ in at least ten cells.

Finally, using GAMs (see above), we identified the positions whose heteroplasmy H changes as a function of the cells' losing score in a statistically significant way. We found a total of eleven significant positions (FDR < 0.001), six of them in *mt-Rnr1* and five in *mt-Rnr2*. All of these positions had a higher level of heteroplasmy in loser cells (Fig. 6b–g and Extended Data Fig. 7a-e). The results remain substantially unaltered if the Spearman's rank correlation test (as opposed to the GAMs) is used.

The most common substitutions observed in each position were: *mt-Rnr1* 300 A-to-C; *mt-Rnr1* 303 T-to-G; *mt-Rnr1* 304 T-to-G; *mt-Rnr1* 305 C-to-G; *mt-Rnr1* 326 A-to-G; *mt-Rnr1* 327 C-to-G; *mt-Rnr2* 2,031 T-to-G; *mt-Rnr2* 2,074 C-to-G; *mt-Rnr2* 2,077 A-to-C; *mt-Rnr2* 2,079 C-to-T; *mt-Rnr2* 2,081 A-to-G.

For the bar plot shown in Fig. 6h and the correlation heat maps in Fig. 6i and Extended Data Fig. 7l, we took into account only cells that covered with more than 50 reads all the significant positions in the *mt-Rnr1* gene (215 cells; Fig. 6h,i) or in both the *mt-Rnr1* and *mt-Rnr2* genes (214 cells; Extended Data Fig. 7l).

As a negative control, we repeated the analysis described above using the ERCC spike-ins added to each cell. As expected, none of the positions were statistically significant, which suggests that our procedure is robust against sequence errors introduced during PCR amplification.

We also performed the mtDNA heteroplasmy analysis in cells from the visceral endoderm and the extra-embryonic ectoderm in both DMSO and CI conditions; none of these cells had a mtDNA heteroplasmy higher than 0.01 in the 11 significant positions identified within *mt-Rnr1* and *mt-Rnr2* in loser epiblast cells, and the reference allele was always the most common. This reinforces the hypothesis that such variants are specific to loser epiblast cells and are not resulting from contamination.

To test the reliability of our heteroplasmy estimations, we used RNA-seq data from two of the mtDNA cell lines (BG and HB; Fig. 7) for which the heteroplasmy was measured also by ARMS–qPCR. To do so, first we downloaded the fasta files of the two mtDNA cell lines from https://www.ncbi.nlm.nih.gov/nuccore/KC663619.1/ and https://www.ncbi.nlm.nih.gov/nuccore/KC663620.1/, then we identified the mtDNA positions that differed from the BL6 reference genome. Finally, on these different positions, the heteroplasmy, H, was computed as explained above. The values of heteroplasmy we found with our computational analysis were very close to those estimated by ARMS–qPCR: for HB (24%), ~17% from RNA-seq data versus ~24% measured by ARMS–qPCR; and for BG (95%), ~93% from RNA-seq data versus ~95% measured by ARMS–qPCR.

Because we are inferring mtDNA changes from RNA-seq data, we also considered additional potential sources for the sequence changes we observed. Specifically, one possible source is contamination from NUMTs. However, a NUMT contamination is very unlikely for the following reasons: (1) we considered only reads that uniquely mapped to the mitochondrial genome; (2) the variants with the highest heteroplasmy identified in 'loser' cells (*mt-Rnr1* 326 and 327) were not present in any of the NUMTs previously reported[76] or those that we identified using *blastn* (also taking into account the SNPs of the mouse strain we used); (3) the variants detected were exclusively found in 'loser' epiblast cells, and they were not detected in any other cell type from the same embryos, that is, neither in 'winner' epiblast cells nor in cells from extra-embryonic tissues; (4) we estimated that if the variants with the strongest heteroplasmy (that is, *mtRnr-1* 326 and 327) were present on a NUMT, in order for them to reach an heteroplasmy of ~20% (Fig. 6b,c), the NUMT would have to be expressed at high levels, comparable to or even higher than many mitochondrial genes.

Another possible cause of the sequence changes is RNA editing. However, the majority of the changes that we found (see above) are not compatible with the canonical RNA editing in Metazoans, which consists of A-to-I (which would be read as A-to-G in RNA-seq) and C-to-U[77].

Common features of scRNA-seq and bulk RNA-seq datasets. Differential expression analysis between the co-cultured winner HB (24%) and loser BG (95%) cell lines was performed using the package EdgeR (v3.20.9)[78].

Batches were specified in the argument of the function 'model.matrix'. We fitted a quasi-likelihood negative binomial generalized log-linear model (with the function 'glmQLFit') to the genes that were filtered by the function filterByExpr (with default parameters). These genes were used as background for the gene enrichment analysis.

We set an FDR of 0.001 as a threshold for significance. The enrichment analysis for both the scRNA-seq and bulk RNA-seq datasets were performed using the tool g:Profiler[79]. The list of upregulated, downregulated and background genes related to the differential expression analysis for the bulk RNA-seq dataset is provided in the Supplementary Tables 5–7.

Quantification, statistical analysis and reproducibility. The quantification of the DDIT3 and OPA1 expression in embryos was performed using two distinct methods. DDIT3 expression was quantified by counting the number of epiblast cells with positive staining in the embryos of each group. The expression of OPA1 was quantified on Fiji software as the mean fluorescence across a ten-pixel-width line drawn on the basal cytoplasm of each cell with high or low p-rpS6 fluorescence intensity, as specified in a previous study[7]. A minimum of eight cells were quantified per condition (high versus low mTOR activity) in each embryo. Six embryos treated with CI were analysed. Mean values of OPA1 fluorescence for each epiblast cell were pooled on the same graph.

Flow cytometry data were analysed with FlowJo Software v9 or v10.0.7r2.

Western blot quantification was performed using Image Studio Lite v5.2.5 (LI-COR). Protein expression levels were normalized to loading controls vinculin, α-tubulin or PCNA.

Normalization of data from metabolic flux analysis with Seahorse was performed using Wave Desktop software (Agilent Technologies) and data were exported to Prism v8 (GraphPad) for statistical analysis.

All box plots show the lower quartile (Q1, 25th percentile), the median (Q2, 50th percentile) and the upper quartile (Q3, 75th percentile). Box length refers to interquartile range (IQR, Q3 − Q1). The upper whisker marks the minimum between the maximum value in the dataset and 1.5 times the IQR from Q3 (Q3 + 1.5 × IQR), while the lower whisker marks the maximum between the minimum value in the dataset and the IQR times 1.5 from Q1 (Q1 − 1.5 × IQR). Outliers are shown outside the interval defined by box and whiskers as individual points.

The micrographs shown on Fig. 3d represents one of the micro-dissected embryo epiblasts used for the experiment presented in Fig. 3g,h. The representative confocal microscopy images shown in Fig. 4h are from confocal imaging deconvolution performed from one experiment following reproducibility of observations from previous independent experiments.

The statistical analysis of the results generated in wet-lab experiments was performed using GraphPad Prism v8.0.0 for Mac (GraphPad Software). Data were

tested for normality using the Shapiro–Wilk normality test. Two-tailed parametric or non-parametric statistical tests were applied accordingly. Statistical significance was considered with a confidence interval of 0.05%; $^{*}P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$.

Here we specify details about the statistical test and multiple-comparisons test (when applicable) used for each experiment. The statistical significance of IPA analysis shown in Figs. 2b,c and 3a,b was calculated with A right-tailed Fisher's exact test ($P < 0.05$). Data presented in Figs. 3f and 4i and Extended Data Fig. 4g were analysed by Mann–Whitney test. Data shown in Fig. 4b–e and Extended Data Fig. 5a–d were analysed by an unpaired *t*-test or Mann–Whitney *U* test. Data shown in Figs. 4j and 5e,g–i were analysed with an unpaired *t*-test. A one-sample *t*-test was used to analyse data presented in Fig. 5d,f. Figure 4k and Extended Data Fig. 8b–f show data analysed by one-way ANOVA, followed by Holm–Sidak's multiple-comparisons test. Data presented in Figs. 3h, 4f,g, 5b,c,j and 7b–f and Extended Data Figs. 5e–g, 6b,c and 8g–i were analysed by two-way ANOVA, followed by Holm–Sidak's multiple-comparisons test. The statistical analysis of data from Extended Data Fig. 9 was carried out with one-way ANOVA or Kruskal–Wallis test, followed by Holm–Sidak's or Dunn's multiple-comparison test, respectively. ANOVA for parametric effects on a GAM fit was used test statistical significance for data presented in Fig. 6a–g and Extended Data Fig. 7a–e. The adjusted *P* values (indicated at the top of each plot) were computed using the Benjamini–Hochberg method. The correlation coefficients shown in Fig. 6i and Extended Data Figs. 2e, 3d and 7l were calculated with Spearman's rank correlation rho test (two-sided test, 0.95 confidence level). Data shown in Extended Data Figs. 3b,c, 4a and 10b,c were analysed with Fisher's exact test (two-sided). Gene enrichment analysis shown in Extended Data Fig. 10 was tested for statistical significance with a cumulative hypergeometric test. *P* values were adjusted for multiple comparisons using the g:Profiler algorithm g:SCS (https://doi.org/10.1093/nar/gkm226). Finally, the statistical analysis on data presented in Supplementary Tables 5 and 6 was performed using empirical Bayes quasi-likelihood *F* tests. *P* values were adjusted for multiple comparisons using the Benjamini–Hochberg method.

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

Data availability

Data were analysed with standard programmes and packages, as detailed above. All relevant data are included in the paper and/or its Supplementary Information files. RNA-seq raw data as well as processed data are available through ArrayExpress, under accession numbers E-MTAB-8640, for scRNA-seq data, and E-MTAB-8692, for bulk RNA-seq data. Source data are provided with this paper.

Code availability

Code used to generate the figures in the paper is available at https://github.com/ScialdoneLab/Cell-Competition-Paper-Figures/.

Received: 24 February 2020; Accepted: 2 June 2021;
Published online: 12 July 2021

References

1. Bowling, S., Lawlor, K. & Rodriguez, T. A. Cell competition: the winners and losers of fitness selection. *Development* **146**, dev167486 (2019).
2. Diaz-Diaz, C. & Torres, M. Insights into the quantitative and dynamic aspects of cell competition. *Curr. Opin. Cell Biol.* **60**, 68–74 (2019).
3. Madan, E., Gogna, R. & Moreno, E. Cell competition in development: information from flies and vertebrates. *Curr. Opin. Cell Biol.* **55**, 150–157 (2018).
4. Morata, G. & Ripoll, P. Minutes: mutants of *Drosophila* autonomously affecting cell division rate. *Dev. Biol.* **42**, 211–221 (1975).
5. Claveria, C., Giovinazzo, G., Sierra, R. & Torres, M. Myc-driven endogenous cell competition in the early mammalian embryo. *Nature* **500**, 39–44 (2013).
6. Sancho, M. et al. Competitive interactions eliminate unfit embryonic stem cells at the onset of differentiation. *Dev. Cell* **26**, 19–30 (2013).
7. Bowling, S. et al. P53 and mTOR signalling determine fitness selection through cell competition during early mouse embryonic development. *Nat. Commun.* **9**, 1763 (2018).
8. Diaz-Diaz, C. et al. Pluripotency surveillance by myc-driven competitive elimination of differentiating cells. *Dev. Cell* **42**, 585–599 (2017).
9. Hashimoto, M. & Sasaki, H. Epiblast formation by TEAD-YAP-dependent expression of pluripotency factors and competitive elimination of unspecified cells. *Dev. Cell* **50**, 139–154 (2019).
10. Lima, A., Burgstaller, J., Sanchez-Nieto, J. M. & Rodriguez, T. A. The mitochondria and the regulation of cell fitness during early mammalian development. *Curr. Top. Dev. Biol.* **128**, 339–363 (2018).
11. Zhou, W. et al. HIF1α induced switch from bivalent to exclusively glycolytic metabolism during ESC-to-EpiSC/hESC transition. *EMBO J.* **31**, 2103–2116 (2012).
12. Khrapko, K. et al. Mitochondrial mutational spectra in human cells and tissues. *Proc. Natl Acad. Sci. USA* **94**, 13798–13803 (1997).
13. Allio, R., Donega, S., Galtier, N. & Nabholz, B. Large variation in the ratio of mitochondrial to nuclear mutation rate across animals: implications for genetic diversity and the use of mitochondrial DNA as a molecular marker. *Mol. Biol. Evol.* **34**, 2762–2772 (2017).
14. Burgstaller, J. P., Johnston, I. G. & Poulton, J. Mitochondrial DNA disease and developmental implications for reproductive strategies. *Mol. Hum. Reprod.* **21**, 11–22 (2015).
15. Gorman, G. S. et al. Mitochondrial diseases. *Nat. Rev. Dis. Prim.* **2**, 16080 (2016).
16. Burgstaller, J. P. et al. MtDNA segregation in heteroplasmic tissues is common in vivo and modulated by haplotype differences and developmental stage. *Cell Rep.* **7**, 2031–2041 (2014).
17. Johnston, I. G. et al. Stochastic modelling, Bayesian inference, and new in vivo measurements elucidate the debated mtDNA bottleneck mechanism. *eLife* **4**, e07464 (2015).
18. Latorre-Pellicer, A. et al. Regulation of mother-to-offspring transmission of mtDNA heteroplasmy. *Cell Metab.* **30**, 1120–1130 (2019).
19. Lee, H. S. et al. Rapid mitochondrial DNA segregation in primate preimplantation embryos precedes somatic and germline bottleneck. *Cell Rep.* **1**, 506–515 (2012).
20. Zhang, H., Burr, S. P. & Chinnery, P. F. The mitochondrial DNA genetic bottleneck: inheritance and beyond. *Essays Biochem.* **62**, 225–234 (2018).
21. Sharpley, M. S. et al. Heteroplasmy of mouse mtDNA is genetically unstable and results in altered behavior and cognition. *Cell* **151**, 333–343 (2012).
22. Wolf, F. A. et al. PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biol.* **20**, 59 (2019).
23. Angerer, P. et al. destiny: diffusion maps for large-scale single-cell data in R. *Bioinformatics* **32**, 1241–1243 (2016).
24. Kramer, A., Green, J., Pollard, J. Jr. & Tugendreich, S. Causal analysis approaches in ingenuity pathway analysis. *Bioinformatics* **30**, 523–530 (2014).
25. Haghverdi, L., Buttner, M., Wolf, F. A., Buettner, F. & Theis, F. J. Diffusion pseudotime robustly reconstructs lineage branching. *Nat. Methods* **13**, 845–848 (2016).
26. Cheng, S. et al. Single-cell RNA-seq reveals cellular heterogeneity of pluripotency transition and X chromosome dynamics during early mouse development. *Cell Rep.* **26**, 2593–2607 (2019).
27. Topf, U., Wrobel, L. & Chacinska, A. Chatty mitochondria: keeping balance in cellular protein homeostasis. *Trends Cell Biol.* **26**, 577–586 (2016).
28. Melber, A. & Haynes, C. M. UPR^{mt} regulation and output: a stress response mediated by mitochondrial–nuclear communication. *Cell Res.* **28**, 281–295 (2018).
29. Munch, C. The different axes of the mammalian mitochondrial unfolded protein response. *BMC Biol.* **16**, 81 (2018).
30. Zhao, Q. et al. A mitochondrial specific stress response in mammalian cells. *EMBO J.* **21**, 4411–4419 (2002).
31. Nargund, A. M., Pellegrino, M. W., Fiorese, C. J., Baker, B. M. & Haynes, C. M. Mitochondrial import efficiency of ATFS-1 regulates mitochondrial UPR activation. *Science* **337**, 587–590 (2012).
32. Quiros, P. M., Mottis, A. & Auwerx, J. Mitonuclear communication in homeostasis and stress. *Nat. Rev. Mol. Cell Biol.* **17**, 213–226 (2016).
33. Mouchiroud, L. et al. The NAD^{+}/sirtuin pathway modulates longevity through activation of mitochondrial UPR and FOXO signaling. *Cell* **154**, 430–441 (2013).
34. Saveljeva, S. et al. Endoplasmic reticulum stress-mediated induction of SESTRIN 2 potentiates cell survival. *Oncotarget* **7**, 12254–12266 (2016).
35. Yun, J. & Finkel, T. Mitohormesis. *Cell Metab.* **19**, 757–766 (2014).
36. Chen, H. et al. Mitofusins Mfn1 and Mfn2 coordinately regulate mitochondrial fusion and are essential for embryonic development. *J. Cell Biol.* **160**, 189–200 (2003).
37. Prudent, J. & McBride, H. M. The mitochondria–endoplasmic reticulum contact sites: a signalling platform for cell death. *Curr. Opin. Cell Biol.* **47**, 52–63 (2017).
38. Smirnova, E., Griparic, L., Shurland, D. L. & van der Bliek, A. M. Dynamin-related protein Drp1 is required for mitochondrial division in mammalian cells. *Mol. Biol. Cell* **12**, 2245–2256 (2001).
39. Favaro, G. et al. DRP1-mediated mitochondrial shape controls calcium homeostasis and muscle mass. *Nat. Commun.* **10**, 2576 (2019).
40. Quiros, P. M. et al. Multi-omics analysis identifies ATF4 as a key regulator of the mitochondrial stress response in mammals. *J. Cell Biol.* **216**, 2027–2045 (2017).
41. Restelli, L. M. et al. Neuronal mitochondrial dysfunction activates the integrated stress response to induce fibroblast growth factor 21. *Cell Rep.* **24**, 1407–1414 (2018).
42. Richter, U. et al. A mitochondrial ribosomal and RNA decay pathway blocks cell proliferation. *Curr. Biol.* **23**, 535–541 (2013).

43. Moullan, N. et al. Tetracyclines disturb mitochondrial function across eukaryotic models: a call for caution in biomedical research. *Cell Rep.* **10**, 1681–1691 (2015).
44. Kauppila, J. H. K. et al. A phenotype-driven approach to generate mouse models with pathogenic mtDNA mutations causing mitochondrial disease. *Cell Rep.* **16**, 2980–2990 (2016).
45. Fan, W. et al. A mouse model of mitochondrial disease reveals germline selection against severe mtDNA mutations. *Science* **319**, 958–962 (2008).
46. Stewart, J. B. et al. Strong purifying selection in transmission of mammalian mitochondrial DNA. *PLoS Biol.* **6**, e10 (2008).
47. Freyer, C. et al. Variation in germline mtDNA heteroplasmy is determined prenatally but modified during subsequent transmission. *Nat. Genet.* **44**, 1282–1285 (2012).
48. Ludwig, L. S. et al. Lineage tracing in humans enabled by mitochondrial mutations and single-cell genomics. *Cell* **176**, 1325–1339 (2019).
49. Chinnery, P. F. & Hudson, G. Mitochondrial genetics. *Br. Med. Bull.* **106**, 135–159 (2013).
50. Floros, V. I. et al. Segregation of mitochondrial DNA heteroplasmy through a developmental genetic bottleneck in human embryos. *Nat. Cell Biol.* **20**, 144–151 (2018).
51. Burr, S. P., Pezet, M. & Chinnery, P. F. Mitochondrial DNA heteroplasmy and purifying selection in the mammalian female germline. *Dev. Growth Differ.* **60**, 21–32 (2018).
52. Rajasimha, H. K., Chinnery, P. F. & Samuels, D. C. Selection against pathogenic mtDNA mutations in a stem cell population leads to the loss of the 3243 A > G mutation in blood. *Am. J. Hum. Genet.* **82**, 333–343 (2008).
53. Ellis, S. J. et al. Distinct modes of cell competition shape mammalian tissue morphogenesis. *Nature* **569**, 497–502 (2019).
54. Kucinski, I., Dinan, M., Kolahgar, G. & Piddini, E. Chronic activation of JNK JAK/STAT and oxidative stress signalling causes the loser cell status. *Nat. Commun.* **8**, 136 (2017).
55. Kon, S. et al. Cell competition with normal epithelial cells promotes apical extrusion of transformed cells through metabolic changes. *Nat. Cell Biol.* **19**, 530–541 (2017).
56. Ran, F. A. et al. Genome engineering using the CRISPR–Cas9 system. *Nat. Protoc.* **8**, 2281–2308 (2013).
57. Czechanski, A. et al. Derivation and characterization of mouse embryonic stem cells from permissive and nonpermissive strains. *Nat. Protoc.* **9**, 559–574 (2014).
58. Burgstaller, J. P. et al. Large-scale genetic analysis reveals mammalian mtDNA heteroplasmy dynamics and variance increase through lifetimes and generations. *Nat. Commun.* **9**, 2488 (2018).
59. Burgstaller, J. P., Schinogl, P., Dinnyes, A., Muller, M. & Steinborn, R. Mitochondrial DNA heteroplasmy in ovine fetuses and sheep cloned by somatic cell nuclear transfer. *BMC Dev. Biol.* **7**, 141 (2007).
60. Kang, E. et al. Mitochondrial replacement in human oocytes carrying pathogenic mitochondrial DNA mutations. *Nature* **540**, 270–275 (2016).
61. Yahata, N., Boda, H. & Hata, R. Elimination of mutant mtDNA by an optimized mpTALEN restores differentiation capacities of heteroplasmic MELAS-iPSCs. *Mol. Ther. Methods Clin. Dev.* **20**, 54–68 (2021).
62. Venegas, V. & Halberg, M.C. Quantification of mtDNA mutation heteroplasmy (ARMS–qPCR). *Methods Mol. Biol.* **837**, 313–326 (2012).
63. Machado, T. S. et al. Real-time PCR quantification of heteroplasmy in a mouse model with mitochondrial DNA of C57BL/6 and NZB/BINJ strains. *PLoS ONE* **10**, e0133650 (2015).
64. Schindelin, J. et al. Fiji: an open-source platform for biological-image analysis. *Nat. Methods* **9**, 676–682 (2012).
65. Kim, D. et al. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14**, R36 (2013).
66. Liao, Y., Smyth, G. K. & Shi, W. The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. *Nucleic Acids Res.* **47**, e47 (2019).
67. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
68. Mootha, V. K. et al. PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.* **34**, 267–273 (2003).
69. Subramanian, A. et al. Gene-set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 (2005).
70. Picelli, S. et al. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat. Methods* **10**, 1096–1098 (2013).
71. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **14**, 417–419 (2017).
72. Lun, A. T., 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).
73. Scialdone, A. et al. Resolving early mesoderm diversification through single-cell expression profiling. *Nature* **535**, 289–293 (2016).
74. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).
75. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
76. Calabrese, F. M., Simone, D. & Attimonelli, M. Primates and mouse NumtS in the UCSC Genome Browser. *BMC Bioinformatics* **13**, S15 (2012).
77. Lukes, J., Kaur, B. & Speijer, D. RNA editing in mitochondria and plastids: weird and widespread. *Trends Genet.* **37**, 99–102 (2021).
78. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
79. Reimand, J., Arak, T. & Vilo, J. g:Profiler—a web server for functional interpretation of gene lists (2011 update). *Nucleic Acids Res.* **39**, W307–W315 (2011).
80. Zappia, L. & Oshlack, A. Clustering trees: a visualization for evaluating clusterings at multiple resolutions. *GigaScience* **7**, giy083 (2018).
81. Scialdone, A. et al. Computational assignment of cell-cycle stage from single-cell transcriptome data. *Methods* **85**, 54–61 (2015).

Acknowledgements

We thank S. Rothery for guidance and advice with confocal microscopy. The FILM at Imperial College London is supported in part by funding from the Wellcome Trust (grant no. 104931/Z/14/Z) and BBSRC (grant no. BB/L015129/1). We thank J. Elliot and B. Patel from the LMS/NIHR Imperial Biomedical Research Centre Flow Cytometry Facility for support. We are thankful to G. Chennell and A. Sardini for guidance and support with Seahorse experiments. Research in the laboratory of T.A.R. was supported by the MRC project grant (MR/P018467/1) and the BBSRC project grant (BB/S008284/1) and by the British Heart Foundation (BHF) PhD studentships (FS/14/62/31288 and FS/17/64/33476). Work in the laboratory of A.S. is funded by the Helmholtz Association. A.L. was funded by a BHF centre of excellence PhD studentship. S.S. was funded through Wellcome awards 103788/Z/14/Z and 108438/Z/15/Z.

Author contributions

A.L. performed most of the experimental wet-lab work. J.B. and A.L. derived heteroplasmic mESC lines. J.B. performed heteroplasmy measurements in heteroplasmic mESCs. B.P. generated *Mfn2*$^{-/-}$ and *Drp1*$^{-/-}$ mESCs, and J.M.S. conducted characterization of mitochondria shape and pluripotency status. S.P.-M. participated in the metabolic characterization of *Drp1*$^{-/-}$ cells. D.H. performed embryo dissections, treatments and cell dissociation before scRNA-seq experiments. G.L. did the bioinformatic analysis of scRNA-seq data. E.M., N.J. and A.P.G. participated in the analysis of mtDNA heteroplasmy. A.D.G. performed the metabolomic studies using the Metabolon platform and participated in embryo dissections and immunohistochemistry stainings for validation of results obtained by scRNA-seq. T.K. collected the embryos for derivation of the mESCs with different mtDNA content. M.D. and M.K. performed the bioinformatic analysis of bulk RNA-seq experiments. N.J., S.S. and D.C. participated in the design of experimental work and analysis of results. A.L., G.L., A.S and T.A.R. interpreted results and wrote the paper. T.A.R. and A.S. directed and designed the research.

Competing interests

The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s42255-021-00422-7.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42255-021-00422-7.

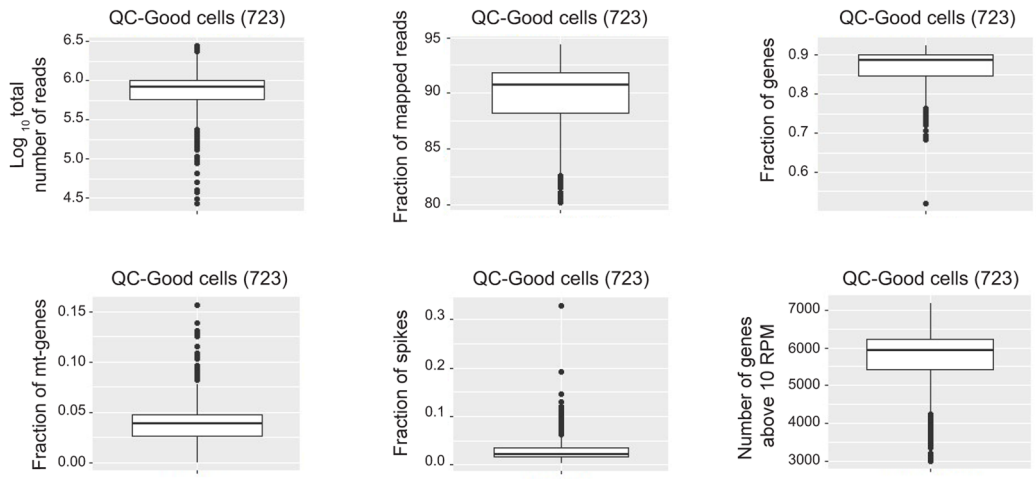
Correspondence and requests for materials should be addressed to A.S. or T.A.R.

Peer review information *Nature Metabolism* thanks Anna-Katerina Hadjantonakis and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editors: Christoph Schmitt; Elena Bellafante.

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

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

a



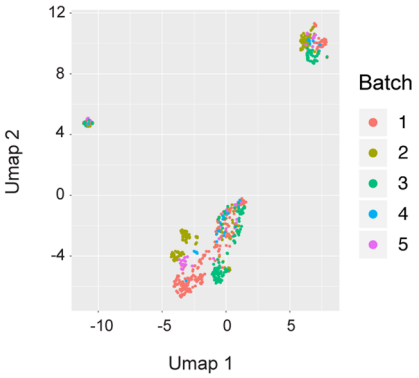
b

Condition\Batch	1	2	3	4	5
CI-treated	136	105	86	16	24
DMSO	132	110	78	15	21

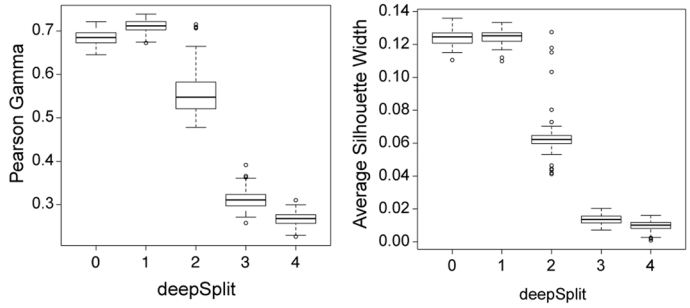
c

Cluster/Batch	1	2	3	4	5
1	147	81	57	7	15
2	45	65	44	7	10
3	46	34	35	13	6
4	23	18	21	2	7
5	7	17	7	2	7

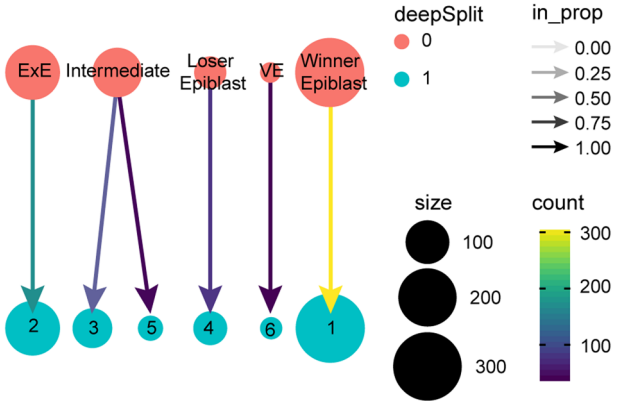
d



e

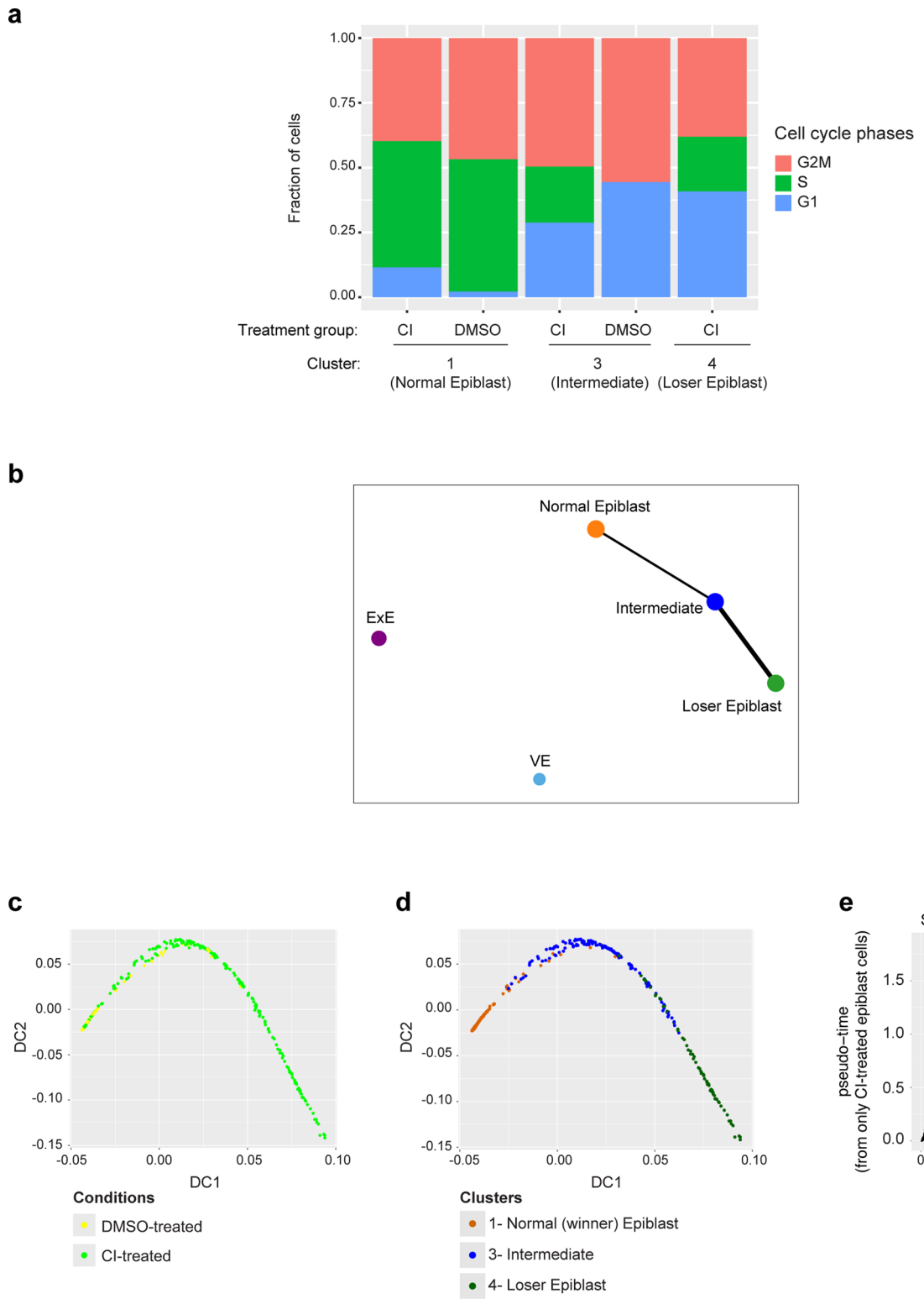


f

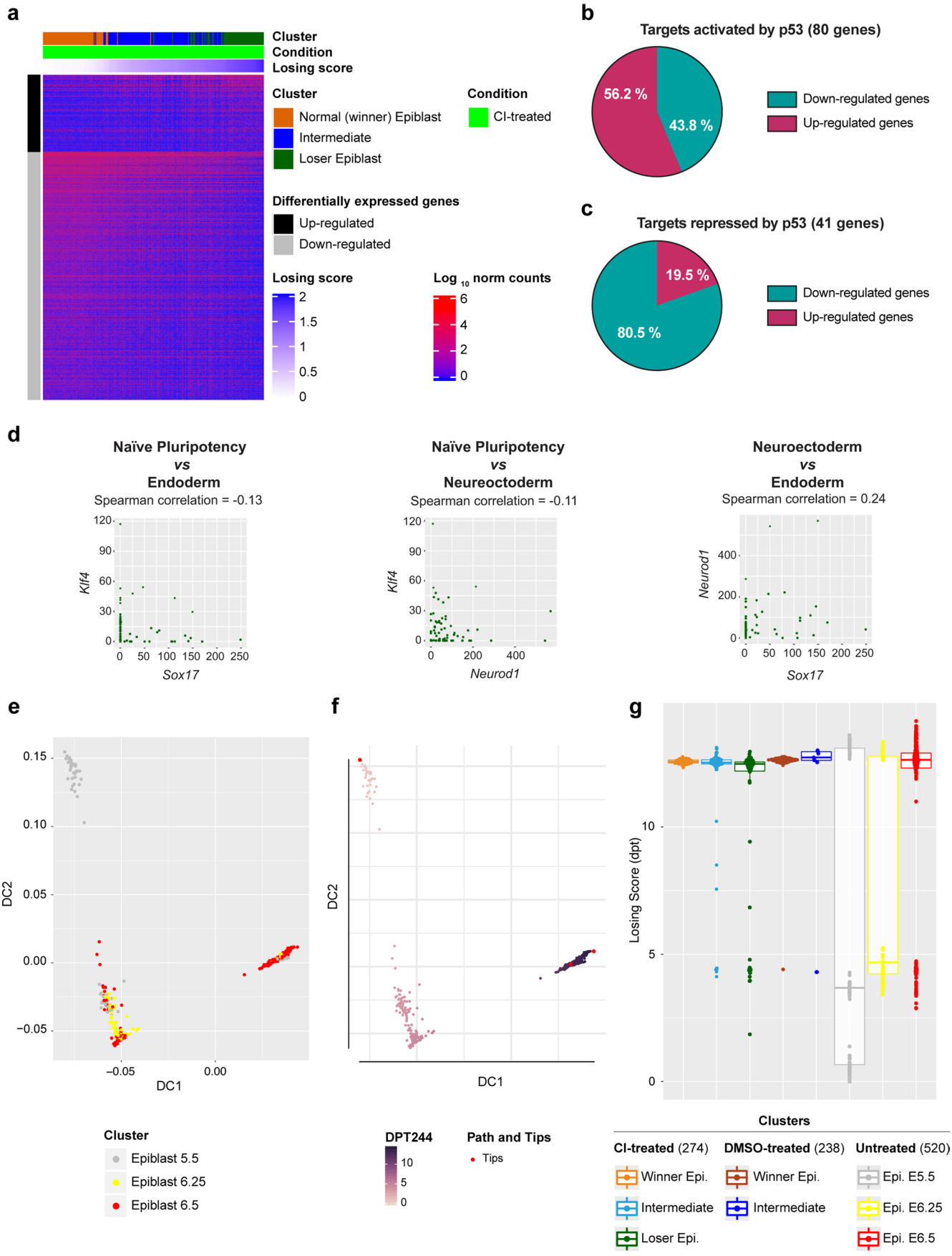


Extended Data Fig. 1 | See next page for caption.

Extended Data Fig. 1 | Quality controls of scRNA-seq and clustering robustness analysis. a, Selection criteria for quality control (QC) of all cells. A total of 723 passed the quality control (723 good quality cells) and were considered for downstream analysis. All these parameters were computed for each cell. Log_{10} total number of reads (top left): log_{10} of the sum of the number of reads that were processed in every cell; Fraction of mapped reads (top central): number of reads that are confidentially mapped to the reference genome divided by total number of reads that were processed for each cell. This number is automatically provided by Salmon v0.8.2; Fraction of genes (top right): number of reads mapped to endogenous genes divided by the total sum of reads that were processed; Fraction of mt-genes (bottom left): number of reads mapped to mitochondrial genes divided by the total sum of reads that were processed; Fraction of spikes (bottom central): number of reads mapped to ERCC spike-ins divided by the total sum of reads that were processed; Number of genes above 10 RPM (bottom right): number of genes with expression level above 10 reads per million. **b**, Number of good quality cells in each condition (rows) and batch (columns). **c**, Number of good quality cells per cluster (rows) and batch (columns). **d**, UMAP plot of the data with cells coloured by batch. In each batch there is a balanced distribution of cells in the two conditions and across the five clusters. **e**, The Pearson's gamma (left panel) and the Average Silhouette Width (right panel) was calculated for each set of clusters obtained with 100 random subsamples of 60% of highly variable genes and different values of the deepSplit parameter (see Methods). The most robust clusters correspond to deepSplit values of 0 and 1. **f**, The changes in composition and number of clusters between the clustering obtained with deepSplit 0 (top) and 1 (bottom) are shown using the library 'clustree'[80]. See methods for details on statistical analysis.



Extended Data Fig. 2 | Cell cycle analysis and cluster connectivity. a, Cell cycle analysis of epiblast cells from clusters 1, 3 and 4. Cell cycle phase was predicted with cyclone algorithm[81] and shows that there are cells in S and G2M phase also in the loser and intermediate clusters. **b**, PAGA plot showing the connectivity of the five clusters of cells from CI-treated embryos. **c-d**, Diffusion map analysis in all epiblast cells (from DMSO and CI-treated embryos): cells are coloured according to the condition (**c**) and to the cluster (**d**). **e**, The pseudotime coordinate of the CI-treated epiblast cells obtained from the diffusion map including all epiblast cells correlates extremely well (with the pseudo-time coordinate obtained in the diffusion map calculated only from CI-treated epiblast cells (Fig. 2a). See methods for details on statistical analysis.



Extended Data Fig. 3 | See next page for caption.

Extended Data Fig. 3 | Analysis on epiblast cells from DMSO and CI-treated embryos. a, Heatmap showing the expression pattern of all genes differentially expressed along the trajectory from winning to losing cells in Fig. 2d. **b-c**, Overlap of genes differentially expressed along the trajectory joining winning and losing epiblast cells in CI-treated embryos (Fig. 2a and panel d) and genes targeted by p53. Pie charts show the percentage of genes up- or down-regulated in loser cells within the group of target genes that are activated (**b**) or repressed (**c**) by p53. There is an enrichment of activated/repressed targets among genes upregulated/downregulated in losing cells respectively (p-value=1E-4). The list of p53 targets is taken from[58]. **d**, Scatter plots of the expression levels of different marker genes plotted against each other in loser epiblast cells (cluster 4). Loser cells have higher expression of pluripotency markers as well as higher expression of some lineage-specific markers and the co-expression of these markers is only weakly correlated - the Spearman's correlation coefficient is shown. **e-g** Our scRNA-seq data from epiblast cells is projected on top of previously published data from epiblast collected from freshly isolated embryos at different stages (E5.5, E6.25 and E6.5; data from[26]). First, a diffusion map (**e**) and a pseudotime coordinate (**f**) is computed for the epiblast cells from freshly isolated embryos. Then, a pseudotime coordinate is estimated for our data after projecting it onto the diffusion map. Panel **g** shows the pseudotime coordinates for both datasets, split by stage, treatment and cluster. See methods for details on statistical analysis.

b

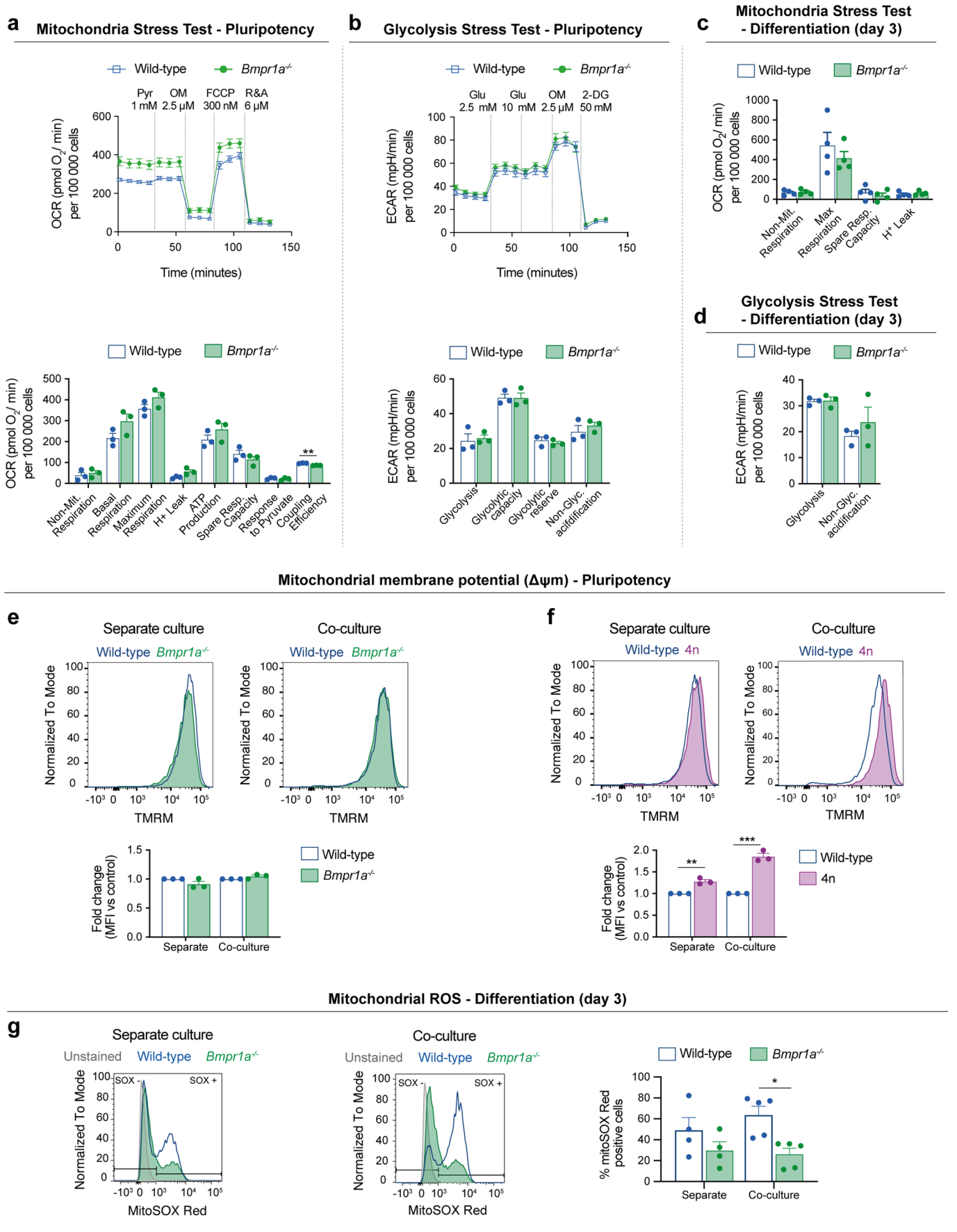
Gene	FDR	Rank
Ddit3	4.63E-39	2
Atf3	6.08E-27	22
Atf4	2.14E-23	31
Foxo3	2.69E-22	37
Ppp1r15a	8.33E-18	68
Eif2ak3	7.17E-13	150
Nfe2l2	1.55E-10	207
Gdf15	5.53E-08	333

c

Gene	FDR	Rank
Mthfd1l	2.54E-35	147
Hspe1	8.71E-34	164
Cat	2.44E-30	219
Hspd1	6.93E-13	1262
Sod2	1.25E-10	1551
Hsph1	4.48E-10	1655
Lonp1	1.08E-06	2348
Eif2a	1.49E-06	2382
Mthfd2	1.31E-05	2693
Hspa4	2.84E-05	2790
Cth	2.53E-03	3677
Nrf1	2.86E-03	3698

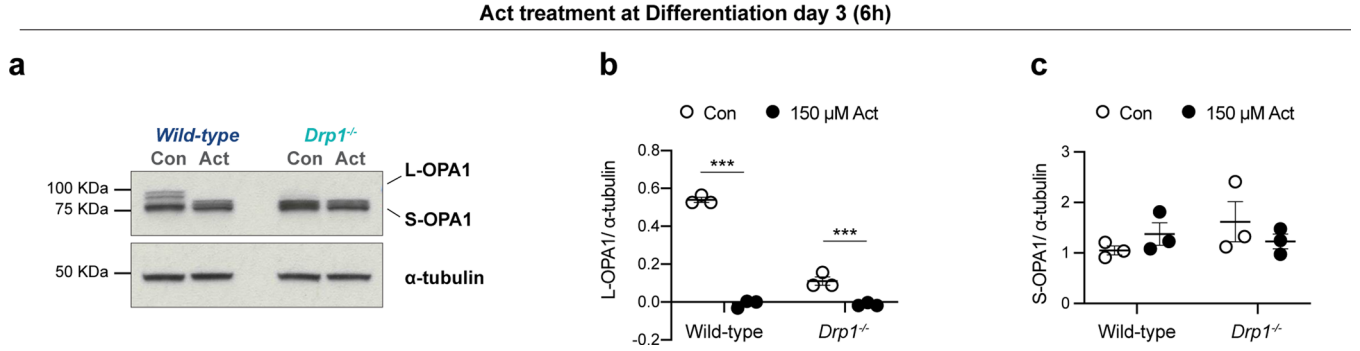
Extended Data Fig. 4 | See next page for caption.

Extended Data Fig. 4 | Cells eliminated during early mouse embryogenesis have activated stress responses. a, Overlap of genes differentially expressed along the trajectory joining winning and losing epiblast cells in CI-treated embryos (Fig. 2a and Extended Data Fig. 3a) and genes related to the unfolded protein response and integrated protein response pathways (UPR_ISR, see Supplementary Table 3). From the 32 genes related to the UPR & ISR pathways, 12 are down-regulated in loser cells, 8 genes are up-regulated in loser cells, and 12 genes are not differentially expressed between loser and winner cells. There is a statistically significant enrichment of UPR&ISR genes among the up-regulated genes in loser cells (odds ratio=3.0, p-value=0.012). The intersection between UPR-ISR genes and the down regulated genes is not significant (odds ratio=1.2, p value=0.69). **b-c**, List of genes from UPR-ISR pathways that are statistically significantly up-regulated (**b**) or down-regulated (**c**) in loser cells. **d**, Scatterplots with the expression levels of genes involved in stress responses in epiblast cells from CI-treated embryos as a function of cells' losing score. **e**, Experimental design with the approach taken to validate the expression of the stress response marker DDIT3 in epiblast cells from DMSO or CI-treated embryos. **f**, Representative micrographs of DMSO (upper panel) or CI-treated embryos (100 μM, lower panel) stained for DDIT3, quantified in (**g**). Nuclei are labelled with Hoechst. In control embryos (DMSO-treated), dying cells in the cavity show very high DDIT3 expression (arrow), while live cells in the epiblast of the CI-treated embryos show more modest levels of DDIT3 expression (arrowheads). Scale bar = 20 μm. **g**, Quantification of the percentage of epiblast cells with nuclear DDIT3 expression. N = 10 DMSO and N = 9 CI-treated embryos. Data shown as mean ± SEM. See methods for details on statistical analysis.

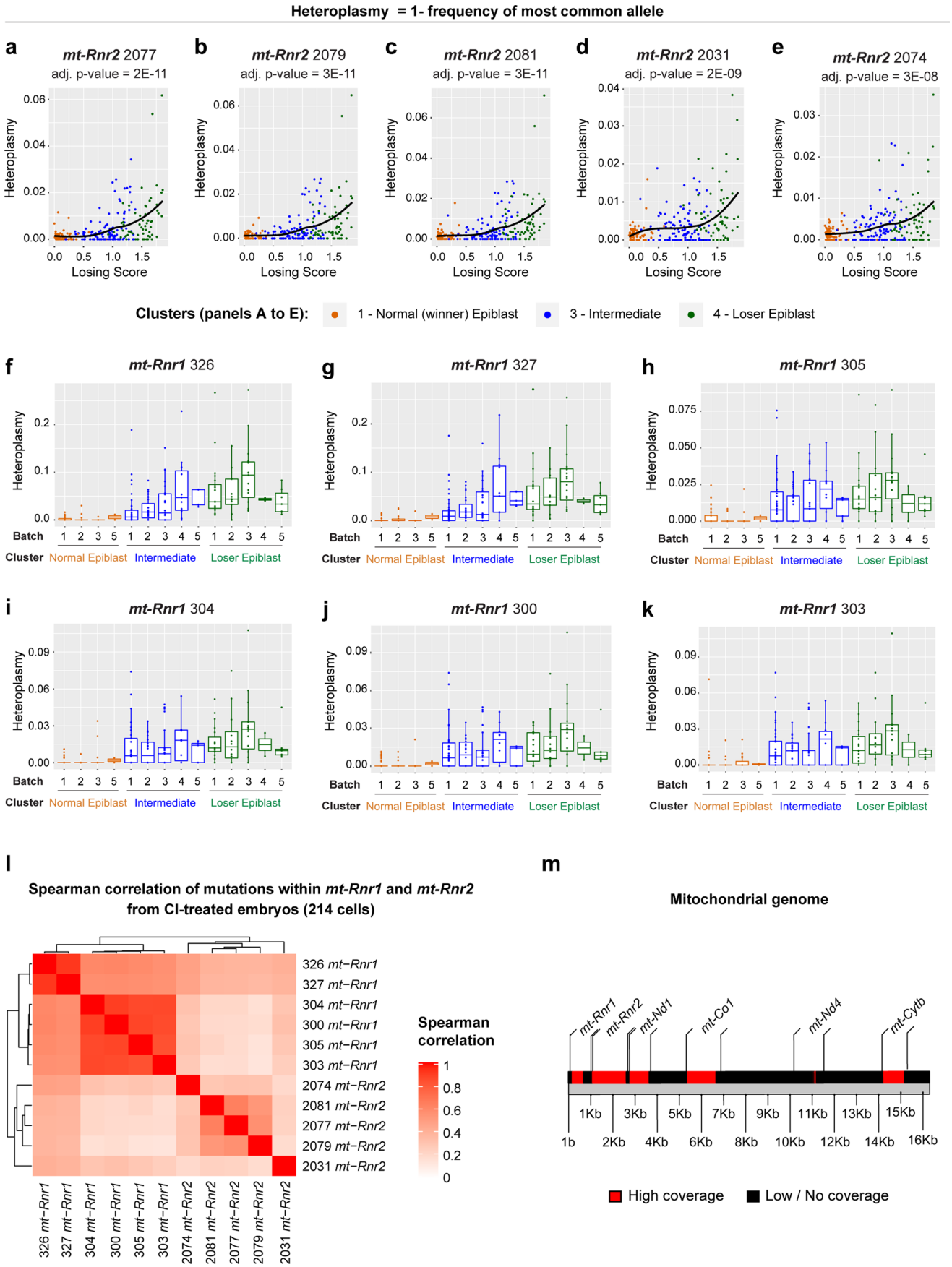


Extended Data Fig. 5 | See next page for caption.

Extended Data Fig. 5 | Mitochondrial function in wild-type, *Bmpr1a*$^{-/-}$ and 4n mESCs. a-d, Metabolic flux analysis of wild-type and *Bmpr1a*$^{-/-}$ mESCs. OCR profile and metabolic parameters assessed during the mitochondria stress test performed in pluripotency conditions (**a**). ECAR profile and metabolic parameters assessed during the glycolysis stress test performed in pluripotency conditions (**b**). Metabolic parameters from the mitochondria stress test found to be similar between wild-type and *Bmpr1a*$^{-/-}$ mESCs during differentiation – day 3 (**c**). Metabolic parameters from the glycolysis stress test found to be similar between wild-type and *Bmpr1a*$^{-/-}$ mESCs during differentiation – day 3 (**d**). Data obtained from 3 (**a,b**) or 5 (**c,d**) independent experiments, with 5 replicates per cell type in each assay. **e-f**, Analysis of mitochondrial membrane potential ($\Delta\psi$m) in defective mESCs maintained in pluripotency conditions, in separate or co-culture. Representative histograms of TMRM fluorescence and quantification for wild-type and *Bmpr1a*$^{-/-}$ (**e**) and wild-type and 4n (**f**). **g**, Analysis of mitochondrial ROS in wild-type and *Bmpr1a*$^{-/-}$ mESCs undergoing differentiation in separate or co-culture: representative histograms of mitoSOX Red fluorescence and quantification of the percentage of mitoSOX positive cells. Data shown as mean ± SEM from 3 (**e-f**) or 5 (**g**) independent experiments. See methods for details on statistical analysis.

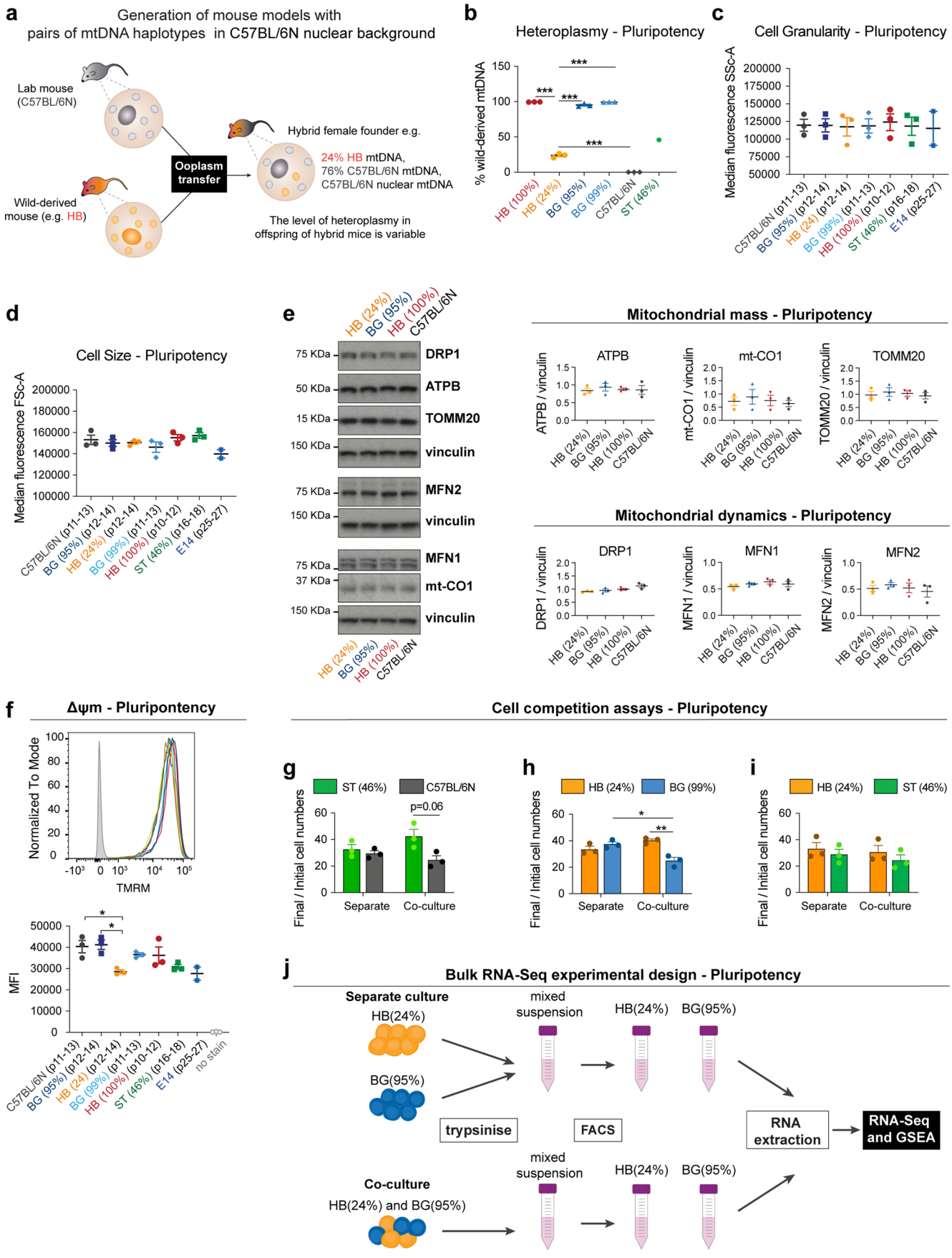


Extended Data Fig. 6 | Effect of actinonin in OPA1 expression in wild-type and *Drp1*$^{-/-}$ cells. a, Western blot analysis of OPA1 expression in wild-type and *Drp1*$^{-/-}$ cells treated with actinonin (Act, 150 μM) during 6 hours on the third day of differentiation, quantified in (**b-c**). **b-c**, Expression levels of L-OPA1 (**b**) and S-OPA1 (**c**) relative to α-tubulin. Data shown as mean ± SEM of 3 independent experiments.



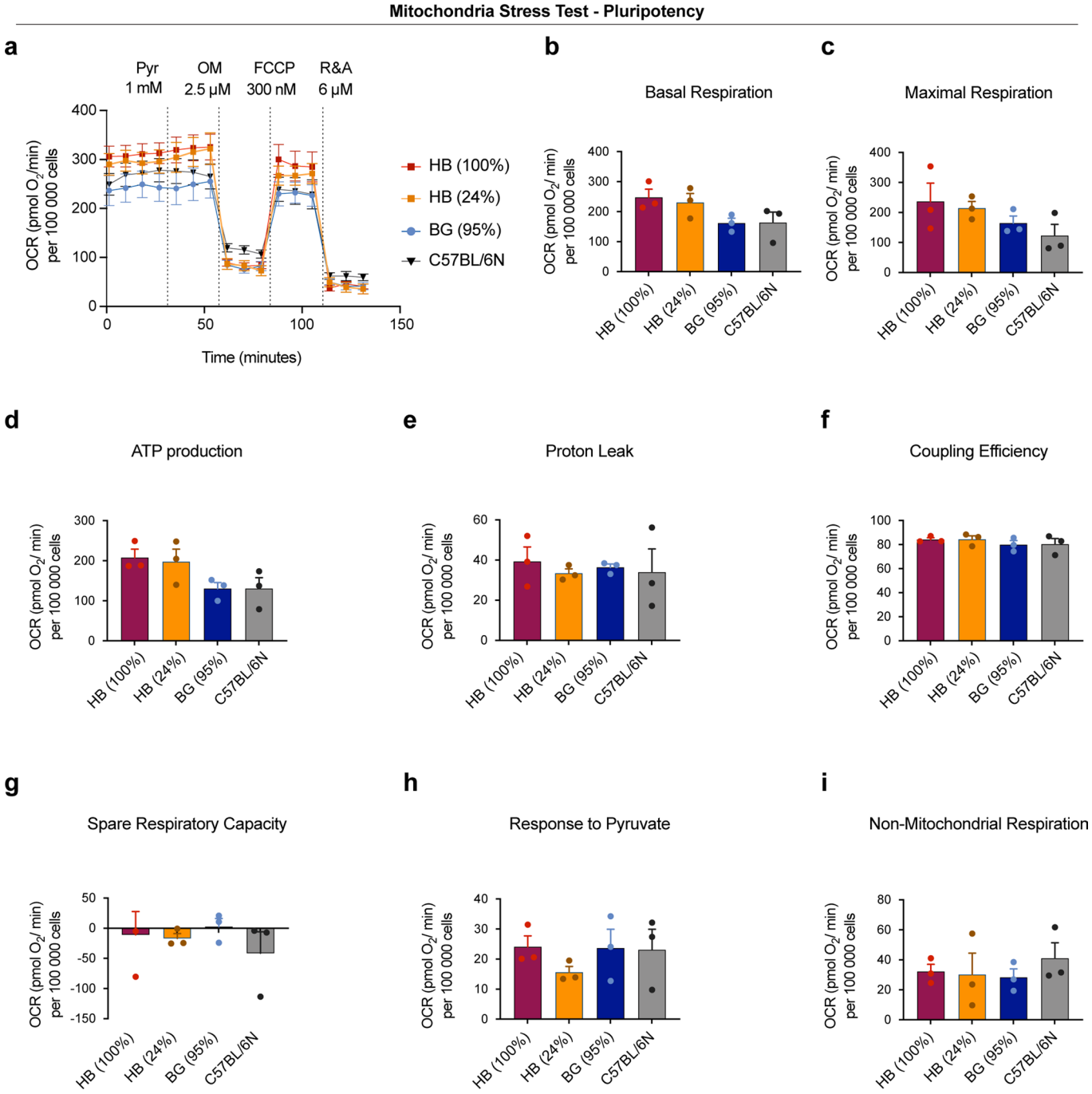
Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | Analysis of SNPs in mtDNA in epiblast cells. a-e, mtDNA heteroplasmy (plotted as Heteroplasmy = 1- frequency of most common allele) in epiblast cells from CI-treated embryos for five positions within the *mt-Rnr2* gene. All these positions have an heteroplasmy that increases with the cells' losing scores in a statistically significant way - the adjusted p-values are indicated at the top of each plot. **f-k**, The variation in the heteroplasmy across the CI-treated cells is not due to a batch effect for the 6 significant positions within the *mt-Rnr1* gene. The number of cells analysed per cluster (and batch) is as follows: number of cells in Normal Epiblast :42 (1),16 (2),18 (3),0 (4),2 (5); number of cells in Intermediate: 42 (1), 28 (2), 28(3), 12 (4), 5 (5); number of cells in Loser Epiblast: 22 (1), 15(2), 20 (3), 2 (4), 7 (5). **l**, Correlation between the mtDNA heteroplasmy at all the statistically significant positions, six within the gene mt-*Rnr1* and five within the gene mt-*Rnr2*. **m**, Schematic representation of the mitochondrial genome showing in red the positions that passed our filtering based on coverage and were considered for the heteroplasmy analysis. Only the genes that include these positions are indicated. See methods for details on statistical analysis.



Extended Data Fig. 8 | See next page for caption.

Extended Data Fig. 8 | Changes in mtDNA sequence are enough to trigger cell competition. a, Illustration of the process of derivation of the mESCs lines from mice that are hybrid between the wild-caught strains (BG, HB or ST) and the lab mouse (C57BL/6N). These hybrid mice were generated elsewhere[16] by ooplasmic transfer: the zygote of a C57BL/6N mouse was injected with ooplasm from a wild-caught mouse (orange, HB pictured). Therefore, these hybrid mice contain the nuclear background of the C57BL/6N strain and the mtDNA of wild-caught strain and potentially C57BL/6N mtDNA (heteroplasmic mice strains). mESCs lines were derived from the hybrid mice and characterised. **b-f**, Characterisation of the derived cell lines by flow cytometry, during pluripotency, in comparison to the wild-type cell line used in previous experiments (E14, 129/Ola background). Heteroplasmy analysis of the derived mESC lines from the hybrid mice, indicating the percentage of wild-derived mtDNA (**b**). Cell granularity (internal complexity) given as median fluorescence intensity of SSc-A laser (**c**). Cell size given as median fluorescence intensity of FSc-A laser (**d**). Analysis of the expression of mitochondrial markers: representative western blot and quantification of markers of mitochondrial mass (ATPB, mt-CO1 and TOMM20) and mitochondrial dynamics (DRP1, MFN1and MFN2), relative to vinculin, in cells derived from hybrid mice (**e**). **f**, Representative histograms and quantification of median TMRM fluorescence, indicative of $\Delta\psi m$, for the hybrid cell lines derived, in comparison to the wild-type cell line used in previous experiments (E14, 129/Ola background). **g-i**, Cell competition assays between hybrid cell lines maintained in pluripotency culture conditions. The ratio of final/initial cell numbers in separate or co-culture is shown. **j**, Experimental design for RNA-Seq and gene set enrichment analysis (GSEA). The isolation of RNA from winner HB(24%) and loser BG(95%) cells was performed after three days in separate or co-culture conditions, once cells have been subjected to FACS to isolate the two populations form mixed cultures. Data shown as mean ± SEM of 3 independent experiments. See methods for details on statistical analysis.

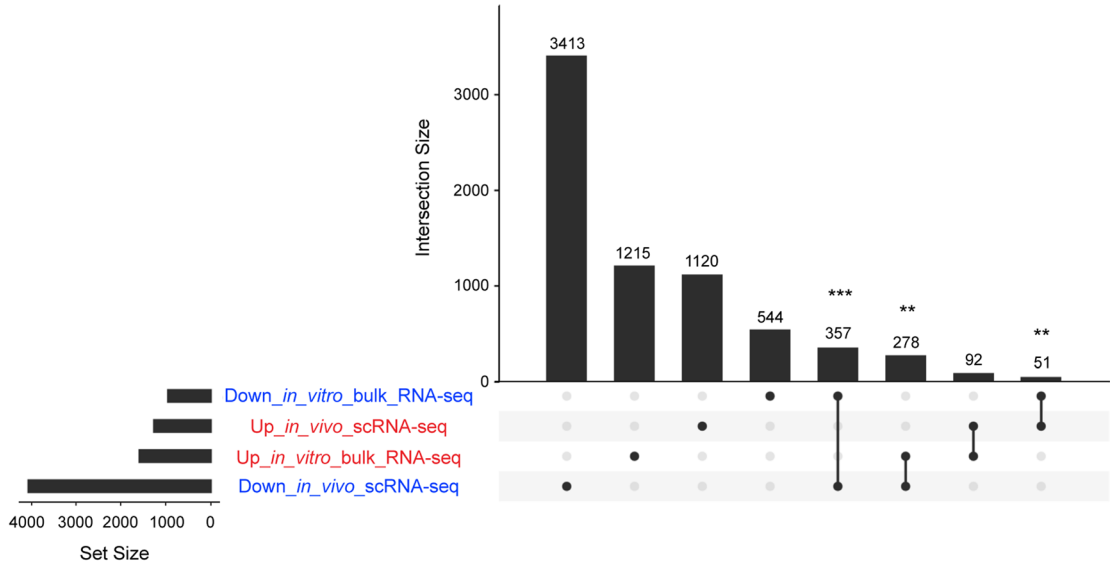


Extended Data Fig. 9 | Metabolic flux analysis of the cells with different mtDNA variants: HB(100%), HB(24%), BG(95%) and C57BL/6N. a, OCR profile during mitochondria stress test performed in pluripotency maintenance conditions. **b-i**, Metabolic parameters assessed during the during the mitochondria stress test performed in pluripotency conditions. Data obtained from 3 independent experiments, with 5 replicates per cell type in each assay. Error bars represent SEM. See methods for details on statistical analysis.

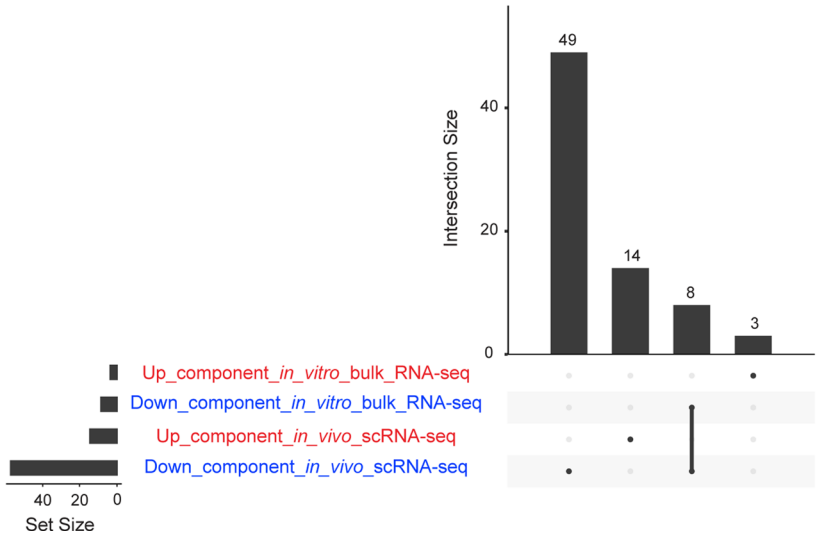
a

Source	Term	Adjusted p-value
GO:CC	mitochondrial protein complex	5.91E-05
GO:CC	inner mitochondrial membrane protein complex	8.84E-04
GO:CC	mitochondrial inner membrane	8.93E-04
GO:CC	mitochondrial respirasome	2.44E-03
GO:CC	respiratory chain complex	3.89E-03
GO:CC	respirasome	6.50E-03
GO:CC	mitochondrial part	1.06E-02
GO:CC	organelle inner membrane	4.65E-02
KEGG	oxidative phosphorylation	7.71E-04
KEGG	Huntington disease	2.35E-03
WP	electron transport chain	1.26E-03

b



c



Extended Data Fig. 10 | See next page for caption.

Extended Data Fig. 10 | Common features of scRNA-seq and bulk RNA-seq datasets. a, Terms significantly enriched among genes downregulated in BG(95%) (loser) ESCs *in vitro* when co-cultured with HB(24%) cells. The loss of mitochondrial activity emerges as a common feature between loser cells *in vivo* and *in vitro*. The gene enrichment analysis was performed using g-profiler tool (see Methods) and p-values were adjusted for multiple comparisons using the g:Profiler algorithm g:SCS (10.1093/nar/gkm226). **b**, Intersection between differentially expressed genes along the trajectory from winning to losing epiblast cells ('in_vivo_scRNA-seq'; Fig. 2a and Extended Data Fig. 3a, and genes differentially expressed between co-cultured HB(24%) (winner) and BG(95%) (loser) ESCs ('in_vitro_bulk_RNA-seq'). 'Up' and 'Down' here refer to genes up- or down-regulated in loser cells. For the intersection between down-regulated genes from scRNA-seq (*in vivo*) and down-regulated genes from bulk RNA-seq (*in vitro*): p-value, 1.71E-12; odds ratio 1.80. For the intersection between down-regulated genes from scRNA-seq (*in vivo*) and up-regulated genes from bulk RNA-seq (*in vitro*): p-value, 5.20E-3; odds ratio 0.67. For the intersection between up-regulated genes from scRNA-seq (*in vivo*) and down-regulated genes from bulk RNA-seq (*in vitro*): p-value, 4.87E-3; odds ratio 0.80. The intersection between up-regulated genes from sc-RNA-seq (*in vivo*) and up-regulated genes from bulk RNA-Seq (*in vitro*) is not statistically significant: p-value: 0.30, odds ratio 1.14. **c**, Intersection between the significantly enriched terms in genes upregulated or downregulated in loser cells in the epiblast of CI-treated embryos ('*in_vivo*_scRNA-Seq') or in our *in vitro* model of competition between co-cultured HB(24%) (winner) and BG(95%) (loser) ESCs ('*in_vitro*_bulk_RNA-seq'). All the terms enriched among downregulated genes *in vitro* are also enriched *in vivo*. See methods for details on statistical analysis.

nature research

Corresponding author(s): Antonio Scialdone & Tristan Rodriguez

Last updated by author(s): May 24, 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	Vi-CELL XR Software (XR v2.04, Beckman Coulter) was used for automated viable cell counts and FACSDiva software for Windows was used for flow cytometry data collection. Excel v16 (currently v16.49) for Mac OS v10 was used for general data compilation prior to statistical analysis.
Data analysis	The software used for RNA-seq data analysis was Salmon v0.8.2, STAR v2.7, samtools v1.11, R studio (https://rstudio.com), Python v3.8.5 and IPA v01-13 (Qiagen). Specific packages used for RNA-seq data analysis are scran v1.6.9, UMAP v0.2.0.0, dynamicTreeCut v1.63.1, Seurat v4.0.1, destiny v2.6.2, GAM v1.16, EdgeR v3.20.9 and scanpy v1.4.7. Microscopy imaging analysis was performed with Fiji. Statistical analysis was performed with GraphPad Prism v8. Western Blot band intensity determined with Image Studio Lite v5.0 (LI-COR). All these were run on Mac OS v10. Flow cytometry analysis data performed with FlowJo v9 & v10.0.7r2 and confocal imaging deconvolution on Huygens (Scientific Volume Imaging) for Windows. Seahorse data normalization was done with Wave Desktop v2.6 (Agilent) run on Windows 10. Data was then plotted subjected to statistical analysis with GraphPad Prism v8.

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

Authors can confirm that all relevant data are included in the paper and/ or its supplementary information files. Source data for Figures 2-5,7 and for Extended Data

Figures 4-6, 8-9 are provided as Excel files with the paper. Due to big size of files, source data for Figure 6 and Extended Data Figure 7 are available from https://drive.google.com/drive/folders/1hSQ_otFYUtxT1t8rpN2sMCDMIH6FIcnp. RNA-seq raw as well as processed data are available through ArrayExpress, accession numbers E-MTAB-8640, for scRNA-seq data, and E-MTAB-8692, for bulk RNA-seq data.

All the code used for generating the figures in the paper is available at https://github.com/ScialdoneLab/Cell-Competition-Paper-Figures.

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 calculated based on our experience with similar previous experiments.
Data exclusions	No data were excluded.
Replication	Experimental observations were reproducible, as indicated by individual data points plotted and sample size number disclosed in figure legends.
Randomization	For cell culture experiments, plate wells were randomly assigned between treatment and control groups within each cell type. For experiments performed in mouse embryos, embryos from different litters were pooled and randomly assigned to control or treatment groups.
Blinding	If n> 4 investigators were not blinded to group allocation but instead groups were blinded during data collection and analysis analysis. With smaller sample numbers blinding was not feasible as the order of sample collection was determined by the well the cells came from and this order was easy to remember. Additionally we relied on unbiased measurements of quantitative parameters.

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

Rabbit anti-ATF-4 (CST-11815, Cell Signaling Technology; RRID:AB_2616025);
Goat anti-rabbit Alexa Fluor 568 (A-11011, Invitrogen; RRID: AB_143157);
Goat anti-mouse Alexa Fluor 488(A-11001, Invitrogen; RRID: AB_2534069).
Mouse anti-ATPB (ab14730, Abcam RRID:AB_301438);
Mouse anti-CHOP (CST-2895, Cell Signaling Technology; RRID:AB_2089254);
Rabbit anti-Cleaved caspase-3 (CST-9664, Cell Signaling Technology; RRID:AB_2070042);
Mouse anti-mt-CO1 (ab14705, Abcam; RRID: AB_2084810);
Rabbit anti-DRP1 (CST-8570, Cell Signaling Technology; RRID: AB_10950498);
Rabbit anti-peIF2alpha(Ser51) (CST-9721, Cell Signaling Technology; RRID:AB_330951);
Mouse anti-MFN1 (ab57602, Abcam; RRID: AB_2142624);
Mouse anti-MFN2 (ab56889, Abcam; RRID: AB_2142629);
Goat anti-mouse HRP conjugated (sc-2005, Santa Cruz Biotechnology; RRID: AB_631736);
Rat anti-NANOG (14-5761-80, eBioScience; RRID: AB_763613);
Mouse anti-OPA1 (612606, BD Biosciences; RRID: AB_399888);
Rabbit anti-PCNA (ab18197, Abcam; RRID:AB_444313);
Rabbit anti-prpS6 (CST-5364, Cell Signaling Technology; RRID: AB_10694233);

Goat anti-rabbit HRP conjugated (sc-2004, Santa Cruz Biotechnology; RRID: AB_631746);
Rabbit anti-TOMM20 (sc-11415, Santa Cruz Biotechnology; RRID: AB_2207533);
Rabbit anti-TOMM20 (CST-42406, Cell Signaling Technology; RRID: AB_2687663);
Rabbit anti-α-Tubulin (CST- 2144, Cell Signaling Technology; RRID: AB_2210548)
Mouse anti-Vinculin (V9131, Sigma-Aldrich; RRID: AB_477629).

Validation

All antibodies used were commercially available and therefore validated by the companies. We have performed our own validation in the lab by including samples that were not probed with primary antibody (incubated with secondary antibody only) alongside the complete staining or blotting protocol. Additionally, we compared target protein expression patterns with the ones seen in relevant published literature.

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)

All cell lines used were mouse embryonic stem cells (mESCs). E14 cells, the wild-type mESCs (RRID: CVCL_C320), were a gift from Prof A. Smith (Cambridge). Tetraploid cells (4n), mESCs null for Bmpr1a , mESCs null for both Bmpr1a and p53 are described elsewhere: Di-Gregorio et al., 2007, Sancho et al., 2013; and Bowling et al., 2018, respectively. Cells null for Mfn2 and cells null for Drp1 are described in this manuscript.

Authentication

Species authentication was performed by PCR or RNA seq.

Mycoplasma contamination

All cell lines tested negative for mycoplasma contamination.

Commonly misidentified lines
(See ICLAC register)

No cell lines used in this study were found listed in the database of known misidentified cell lines.

Animals and other organisms

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

Laboratory animals

For scRNA-seq and validation experiments, pregnant CD1 mice purchased from Charles River were 6-10 weeks old. Embryos were dissected at embryonic day 5.5 (E5.5).
For the derivation of hybrid mESC lines, embryos at morula stage (E2.5) were isolated from hybrid mouse strains generated elsewhere (Burgstaller et al., 2014). These contain the mtDNA of C57BL/6N lab mouse and mtDNA variants from wild-caught mice.

Wild animals

No wild animals were used in this study.

Field-collected samples

This study did not involve field-collected samples.

Ethics oversight

All animal work was done in accordance with the Home Office's Animals (Scientific Procedures) Act 1986 and covered by the Home Office project license PBBEBDCDA.

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

For a flow cytometry performed in epiblast cells isolated from embryos, embryos were dissected ar E5.5 from pregnant CD1 mouse females and cultures overnight in N2B27 poor media with pan-caspase inhibitors (100 µM, Z-VAD-FMK, FMK001, R&D Systems, USA) or equal volume of vehicle (DMSO) as control. On the following morning, to avoid misleading readings, epiblasts were isolated initially by an enzymatic treatment with of 2.5% pancreatin, 0.5% trypsin and 0.5% polyvinylpyrrolidone (PVP40) - all from Sigma-Aldrich - to remove the visceral endoderm (VE). Embryos were treated during 8 min at 4ºC, followed by 2 min at RT. The VE was then pealed with the forceps and the extraembryonic ectoderm removed to isolate the epiblasts. Twelve embryo epiblasts were pooled per treatment condition and dissociated into single cells with 600 µL Acccutase (A6964, Sigma, UK) during 12 min at 37ºC, tapping the tube every two minutes. Accutase was then neutralised with equal volume of FCS, cells span down and stained with 10 nM of the TMRM (T5428, Sigma, UK) prepared in N2B27 media. After incubating for 15 min at 37ºC, cells were pelleted again and re-suspended in 3% FCS in DPBS. Sytox blue (1:1000, S34857, ThermoFisher Scientific, UK), was used as viability staining.

Quantitative analysis of mitochondrial membrane potential (Δψm) and mitochondrial ROS was performed by flow cytometry. Cells were grown in pluripotency or differentiating conditions, dissociated and pelleted to obtain 2E05 cells per sample for the staining procedure. For TMRM staining in mESCs, 2E05 cells of each cell line were resuspended in 200 µL of 10 nM TMRM (T5428, Sigma, UK), prepared in N2B27 media. Cells were incubated at 37ºC for 15 min, and then resuspended in FC buffer (3% FCS in DPBS). For the analysis of mitochondrial ROS, cells were grown in differentiating conditions and stained on the third day of culture. Briefly, 2E05 cells of each cell line were resuspended in 200 µL of 5 µM solution of MitoSOX (M36008, Invitrogen, UK) prepared in N2B27 media. Cells were incubated at 37ºC for 15 min, and then resuspended in FC buffer. Sytox blue was used as viability staining.
Stained cell suspensions with TMRM or MitoSOX were analysed in BD LSRII flow cytometer operated through FACSDiva software (Becton Dickinson Biosciences, UK). For TMRM fluorescence detection the yellow laser was adjusted for excitation at λ=562 nm, capturing the emission light at λ=585 nm for TMRM. MitoSOX fluorescence was analysed with the violet laser adjusted for excitation at λ=405 nm, capturing the emission light at λ=610 nm. In the case of GFP-labelled cell lines, for GFP fluorescence detection the blue laser was adjusted for excitation at λ=488 nm, capturing the emission light at λ=525 nm.

Instrument

BD LSRII cell analyser (Becton Dickinson Biosciences, UK) for cell competition assays, TMRM and MitoSOX staining analysis or BD FACSAria III cell sorter (Becton Dickinson Biosciences, UK) for sorting experiments.

Software

FACSDiva software (Becton Dickinson Biosciences, UK)

Cell population abundance

For cell TMRM and MitoSOX staining analysis, cell population abundance is described in the manuscript figures and corresponding source data. This was determined with FlowJo software. The proportion of cells sorted prior to bulk RNA-seq was determined with FACSDiva software. For scRNA-seq samples, post-sort fractions were determined by computational analysis as described in the Methods.

Gating strategy

The gating strategy is exemplified in the Supplementary Information file, both for experiments using mitochondrial dyes (TMRM or MitoSOX) and sorting of cells within a mixed population based on GFP label prior to bulk RNA collection and sequencing. Briefly, cell debris were excluded on FSC-A vs SCC-A plots. Consequentially, single cells were isolated both with FSC and SSC laser plots and, from the single cell populations, only live cells were considered, based on viability staining applied (Sytox Blue of propidium iodine). Positive staining signal was considered to be equal or above the magnitude of 10^3 on the logarithmic scale of fluorescence intensity for the relevant fluorophore (Sytox Blue shown in example: live cells remain unstained). Flow cytometry data for TMRM and MitoSOX is presented in the form of univariate histogram plots, with x-axis re-labeled with fluorophore name. As plots show more than one sample, data is presented normalised to mode. When distinction between high and low TMRM or MitoSOX positive and negative levels was made, the threshold cut-off value was defined at the magnitude of 5.3^3 and 10^3, respectively, on the logarithmic scale of fluorescence intensity.

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

5.2 Appendix B: MitoHEAR: an R package for the estimation and downstream statistical analysis of the mitochondrial DNA heteroplasmy calculated from single-cell datasets

Core publication as main author

- **Gabriele Lubatti**, Elmir Mahammadov, Antonio Scialdone. (2022). MitoHEAR: an R package for the estimation and downstream statistical analysis of the mitochondrial DNA heteroplasmy 2 calculated from single-cell datasets. Journal of Open Source Software, 7(74), 4265. `https://doi.org/10.21105/joss.04265`.

This is a published version of the article in the Journal of Open Source Software following peer review.

The Journal of Open Source Software

MitoHE R: an R package for the estimation and downstream statistical analysis of the mitochondrial DN heteroplasmy calculated from single-cell datasets

Gabriele Lubatti[*1,2,3], **Elmir Mahammadov**[†1,2,3], **and ntonio Scialdone**[1,2,3]

1 Institute of Epigenetics and Stem Cells, Helmholtz Zentrum München, Munich, Germany **2** Institute of Functional Epigenetics, Helmholtz Zentrum München, Neuherberg, Germany **3** Institute of Computational Biology, Helmholtz Zentrum München, Neuherberg, Germany Corresponding author

DOI: 10.21105/joss.04265

Software

- Review
- Repository
- rchive

Editor: Charlotte Soneson

Reviewers:

- @BatoolMM
- @juanvillada

Submitted: 21 February 2022
Published: 01 June 2022

Summary

Eukaryotic cells rely on mitochondria: organelles that are equipped with their own DN (mtDN) to produce the energy they need. Each cell includes multiple mtDN copies that are not perfectly identical but have differences in their sequence; such sequence variability is called heteroplasmy. mtDN heteroplasmy has been associated with diseases (Nissanka & Moraes, 2020), which can affect cellular fitness and have an impact on cellular competition (Lima et al., 2021). Several single-cell sequencing protocols provide the data to estimate mtDN heteroplasmy, including single-cell DN -seq, RN -seq, and T C-seq, in addition to dedicated protocols like M ESTER (Miller et al., 2022). Here, we provide MitoHE R (Mitochondrial HEteroplasmy nalyzeR), a user-friendly software package written in R that allows this estimation as well as downstream statistical analysis of the mtDN heteroplasmy calculated from single-cell datasets. MitoHE R takes as input B M files, computes the frequency of each allele and, starting from these, estimates the mtDN heteroplasmy at each covered position for each cell.
The analysis parameters (e.g., the filtering of the mtDN positions based on read quality and coverage) are easily tuneable. Moreover, statistical tests are available to explore the dependency of the mtDN heteroplasmy on continuous or discrete cell covariates (e.g., culture conditions, differentiation states, etc.), as extensively shown in the included detailed tutorials.

Statement of need

lthough mtDN heteroplasmy has important consequences on human health (Stewart & Chinnery, 2015) and embryonic development (Floros et al., 2019), there are still many open questions on how heteroplasmy affects cells' ability to function and how cells keep it under control. With the increasing availability of single-cell data, many questions can begin to be answered. Still, it is essential to have efficient and streamlined computational tools that enable researchers to estimate and analyse mtDN heteroplasmy. Existing packages (Calabrese et al., 2014; Huang & Huang, 2021; Prashant et al., 2021) focus only on the first step of quantifying heteroplasmy from B M files, and do not provide any specific tools for further statistical analyses or plotting. MitoHE R covers all steps of the analysis in a unique user-friendly package, with highly customisable functions. Starting from B M files, MitoHE R estimates heteroplasmy and offers several options for downstream analyses. For example, statistical tests are provided to investigate the relationship of the mtDN heteroplasmy with continuous or

*first author
†co-author

discrete cell covariates. Moreover, it includes plotting functions to visualise heteroplasmy and allele frequencies and to perform hierarchical clustering of cells based on heteroplasmy values.

Key functions

The two main functions of `MitoHEAR` are:

1. `get_raw_counts_allele`: parallelised function that relies on Rsamtools and generates the raw counts matrix starting from B M files, with cells as rows and bases with the four possible alleles as columns.
2. `get_heteroplasmy`: Starting from the output of `get_raw_counts_allele`, this function computes the matrix with heteroplasmy values (defined as 1 minus the frequency of the most common allele) and the matrix with allele frequency values, for all the cells and bases that pass a filtering procedure.

mong the downstream analyses implemented in the package are:

- Several statistical tests (e.g., Wilcoxon rank-sum test) for the identification of the mtDN positions with the most different levels of heteroplasmy between discrete groups of cells or along a trajectory of cells (i.e., cells sorted according to a diffusion pseudo-time) (**Figure 1** and **Figure 2**).
- Plotting functions for the visualisation of heteroplasmy and the corresponding allele frequency values among cells.
- Unsupervised hierarchical clustering of cells based on a distance matrix defined from the angular distance of allele frequencies that could be relevant for lineage tracing analysis (Ludwig et al., 2019) (**Figure 3**).

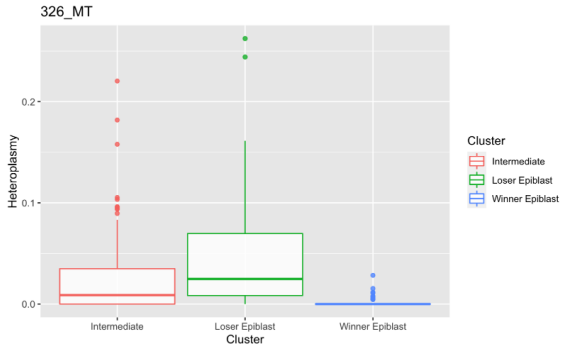


Figure 1: Example of an output plot generated by MitoHE R showing heteroplasmy values at a given position estimated from single cells in three clusters indicated on the x-axis. Data from Lima et al. (2021).

Figure 2: Example of an output figure generated by MitoHE R where the heteroplasmy is plotted as a function of the pseudo-time coordinate of each cell. Cells are classified into three clusters. The heteroplasmy shows a statistically significant change along the pseudo-time, as indicated by the adjusted p-value reported at the top, which is computed by a generalised additive model fit. Data from Lima et al. (2021).

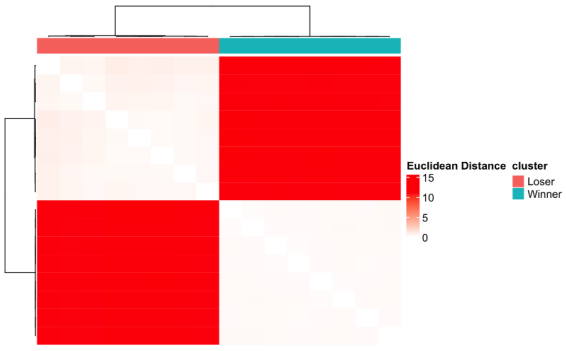


Figure 3: Unsupervised hierarchical clustering of cells based on a distance matrix defined from the angular distance of allele frequencies. The data shown is bulk RN -seq mouse data from two mtDN cell lines labelled *Los r* and *Winn r*. Data from Lima et al. (2021).

The package has been used in a recently published paper (Lima et al., 2021), where we revealed that cells with higher levels of heteroplasmy are eliminated by cell competition in mouse embryos and are characterised by specific gene expression patterns.

References

Calabrese, C., Simone, D., Diroma, M. ., Santorsola, M., Guttà, C., Gasparre, G., Picardi, E., Pesole, G., & ttimonelli, M. (2014). MToolBox: a highly automated pipeline for heteroplasmy annotation and prioritization analysis of human mitochondrial variants in high-throughput sequencing. *Bioinform tics*, *30*(21), 3115–3117. https://doi.org/10.1093/bioinformatics/btu483

Floros, V., Pyle, ., Dietmann, S., Wei, W., Tang, W., Irie, N., Payne, B., Capalbo, ., Noli, L., Coxhead, J., Hudson, G., Crosier, M., Strahl, H., Khalaf, Y., Saitou, M., Ilic, D., Surani, M., & Chinnery, P. (2019). Segregation of mitochondrial DN heteroplasmy through a developmental genetic bottleneck in human embryos. *N ture Cell Biology*. https://doi.org/10.1038/s41556-017-0017-8

Huang, X., & Huang, Y. (2021). Cellsnp-lite: an efficient tool for genotyping single cells. *Bioinform tics*. https://doi.org/10.1093/bioinformatics/btab358

Lima, ., Lubatti, G., Burgstaller, J., Hu, D., Green, ., Gregorio, . D., Zawadzki, T., Pernaute, B., Mahammadov, E., Dore, M., Sanchez, J. M., Bowling, S., Sancho, M., Karimi, M., Carling, D., Jones, N., Srinivas, S., Scialdone, ., & Rodriguez, T. . (2021). Cell competition acts as a purifying selection to eliminate cells with mitochondrial defects during early mouse development. *N ture Met bolism*. https://doi.org/10.1038/s42255-021-00422-7

Ludwig, L. S., Lareau, C. ., Ulirsch, J. C., Christian, E., Muus, C., Li, L. H., Pelka, K., Ge, W., Oren, Y., Brack, ., Law, T., Rodman, C., Chen, J. H., Boland, G. M., Hacohen, N., Rozenblatt-Rosen, O., ryee, M. J., Buenrostro, J. D., Regev, ., & Sankaran, V. G. (2019). Lineage tracing in humans enabled by mitochondrial mutations and single-cell genomics. *Cell*, *176*(6), 1325–1339.e22. https://doi.org/10.1016/j.cell.2019.01.022

Miller, T. E., Lareau, C. ., Verga, J. ., Ssozi, D., Ludwig, L. S., Farran, C. E., Griffin, G. K., Lane, . ., Bernstein, B. E., Sankaran, V. G., & van Galen, P. (2022). Mitochondrial variant enrichment from high-throughput single-cell RN -seq resolves clonal populations. *N ture Biotechnology*. https://doi.org/10.1038/s41587-022-01210-8

Nissanka, N., & Moraes, C. T. (2020). Mitochondrial DN heteroplasmy in disease and targeted nuclease-based therapeutic approaches. *EMBO Reports*, *21*(3), e49612. https://doi.org/10.15252/embr.201949612

Prashant, N., lomran, N., Chen, Y., Liu, H., Bousounis, P., Movassagh, M., Edwards, N., & Horvath, . (2021). SCReadCounts: Estimation of cell-level SNVs expression from scRN -seq data. *BMC Genomics*. https://doi.org/10.1186/s12864-021-07974-8

Stewart, J., & Chinnery, P. (2015). The dynamics of mitochondrial DN heteroplasmy: Implications for human health and disease. *N ture Reviews Genetics*. https://doi.org/10.1038/nrg3966

5.3 Appendix C: CIARA: a cluster-independent algorithm for the identification of markers of rare cell types from single-cell sequencing data

Core publication as main author

- **Gabriele Lubatti**, Marco Stock, Ane Iturbide, Mayra L. Ruiz Tejada Segura, Melina Riepl, Richard Tyser, Anna Danese, Maria Colomé-Tatché, Fabian J. Theis, Shankar Srinivas, Maria-Elena Torres-Padilla, Antonio Scialdone. CIARA: a cluster-independent algorithm for the identification of markers of rare cell types from single-cell sequencing data. Development 2023.`https://doi.org/10.1242/dev.201264`.

This is a published version of the article in Development following peer review.

© 2023. Published by The Company of Biologists Ltd | Development (2023) 150, dev201264. doi:10.1242/dev.201264

TECHNIQUES AND RESOURCES

RESEARCH ARTICLE

CIARA: a cluster-independent algorithm for identifying markers of rare cell types from single-cell sequencing data

Gabriele Lubatti[1,2,3], Marco Stock[1,2,3,4,*], Ane Iturbide[1,*], Mayra L. Ruiz Tejada Segura[1,2,3], Melina Riepl[1,2,3], Richard C. V. Tyser[5], Anna Danese[6], Maria Colomé-Tatché[3,7], Fabian J. Theis[3,8], Shankar Srinivas[9], Maria-Elena Torres-Padilla[1,10] and Antonio Scialdone[1,2,3,‡]

ABSTRACT

A powerful feature of single-cell genomics is the possibility of identifying cell types from their molecular profiles. In particular, identifying novel rare cell types and their marker genes is a key potential of single-cell RNA sequencing. Standard clustering approaches perform well in identifying relatively abundant cell types, but tend to miss rarer cell types. Here, we have developed CIARA (Cluster Independent Algorithm for the identification of markers of RAre cell types), a cluster-independent computational tool designed to select genes that are likely to be markers of rare cell types. Genes selected by CIARA are subsequently integrated with common clustering algorithms to single out groups of rare cell types. CIARA outperforms existing methods for rare cell type detection, and we use it to find previously uncharacterized rare populations of cells in a human gastrula and among mouse embryonic stem cells treated with retinoic acid. Moreover, CIARA can be applied more generally to any type of single-cell omic data, thus allowing the identification of rare cells across multiple data modalities. We provide implementations of CIARA in user-friendly packages available in R and Python.

KEY WORDS: Computational method, Rare cell types, Single-cell sequencing

INTRODUCTION

The development of single-cell omics technologies has allowed the molecular characterization of cell types in a large number of organs and tissues in many different organisms. One goal of single-cell studies is the identification of rare cell types, which bulk techniques are not able to access. Characterization of rare cells is fundamentally important in many biological contexts: for example, during development, to pin down the stage at which a given cell type starts to emerge; when studying cancer, to look for rare cells that might develop drug resistance (Emert et al., 2021); or for the characterization of stem cell lines, searching for cell transitions in different pluripotency states (Taubenschmid-Stowers et al., 2022; Rodriguez-Terrones et al., 2018).

In particular, transcriptional profiling obtained with single-cell RNA sequencing (scRNA-seq) enables the identification of rare cells and their marker genes. Some types of cells can be challenging to identify because, in addition to being rare, they have overlapping markers with other, more abundant cell types. This is the case, for instance, for primordial germ cells, which share markers with cells from the primitive streak (Tyser et al., 2021b; Pijuan-Sala et al., 2019).

Cell type identification is carried out by performing unsupervised clustering, which is typically done using highly variable genes (Luecken and Theis, 2019). Although this strategy is usually successful at identifying large clusters of distinct cell types, it often fails to detect small-sized clusters of cells with fewer specific marker genes.

For this reason, many algorithms that are specifically designed to detect rare cell types in scRNA-seq data have been devised. Some algorithms (e.g. CellSIUS; Wegmann et al., 2019) rely on an existing cluster annotation or assign a rareness score to each of the cells using a sketching technique to measure the density around them (such as FiRE; Jindal et al., 2018). Others, e.g. GiniClust (Dong and Yuan, 2020) and RaceID (Herman et al., 2018), work in a cluster-independent way to identify rare cells and/or their markers.

These methods generally work well in selecting rare cells with strong markers, but they are less efficient in identifying very small cell populations (<1%) with a limited number of specific markers. Moreover, some of these methods tend to overfit and identify a large number of small cell clusters without specific markers.

Here, we developed a novel algorithm called CIARA (Cluster Independent Algorithm for the identification of markers of RAre cell types) that identifies potential marker genes of rare cell types by exploiting their property of being highly expressed in a small number of cells with similar transcriptomic signatures. To achieve this, CIARA ranks genes based on their enrichment in local neighborhoods defined from a K-nearest neighbors (KNN) graph. The top-ranked genes can then be used with standard clustering algorithms to identify groups of rare cell types with high efficiency, requiring the specification of a minimal number of parameters.

We show how CIARA outperforms existing algorithms for rare cell type identification on scRNA-seq datasets generated from different organisms and from different protocols. Moreover, we use CIARA to detect rare cells in a new scRNA-seq dataset of mouse embryonic stem cells (mESCs) treated with retinoic acid and in a recently published dataset from a human gastrula (Tyser et al.,

[1]Institute of Epigenetics and Stem Cells, Helmholtz Munich, D-81377 Munich, Germany. [2]Institute of Functional Epigenetics, Helmholtz Munich, D-85764 Neuherberg, Germany. [3]Institute of Computational Biology, Helmholtz Munich, D-85764 Neuherberg, Germany. [4]TUM School of Life Sciences Weihenstephan, Technical University of Munich, D-85354 Freising, Germany. [5]Wellcome-MRC Cambridge Stem Cell Institute, University of Cambridge, Cambridge CB2 0AW, UK. [6]Biomedical Center Munich (BMC), Physiological Genomics, Faculty of Medicine, Ludwig Maximilians University, D-82152 Munich, Germany. [7]Biomedical Center (BMC), Physiological Chemistry, Faculty of Medicine, Ludwig Maximilians University, D-82152 Munich, Germany. [8]Department of Mathematics, Technical University of Munich, D-85748 Munich, Germany. [9]Department of Physiology, Anatomy and Genetics, University of Oxford, Oxford OX1 3PT, UK. [10]Faculty of Biology, Ludwig-Maximilians University, D-82152 Munich, Germany.
*These authors contributed equally to this work

‡Author for correspondence (antonio.scialdone@helmholtz-muenchen.de)

S.S., 0000-0001-5726-7791; A.S., 0000-0002-4956-2843

Handling Editor: Samantha Morris
Received 5 September 2022; Accepted 25 April 2023

Development (2023) 150, dev201264. doi:10.1242/dev.201264

2021b), where we find several groups of rare cells. Finally, we demonstrate how CIARA can be applied to other types of single-cell omic datasets and can identify rare cells across multiple data modalities.

CIARA is available as R and Python packages, and the scripts to perform all analyses are freely accessible in GitHub.

RESULTS

Overview of CIARA

The input of CIARA is a normalized gene count matrix and a KNN graph that can be computed with a standard approach (Luecken and Theis, 2019; Fig. 1A, left; Materials and Methods). Because rare cell type markers are only expressed in a small number of cells, we restrict the set of genes analyzed to those that are expressed above a threshold in a limited number of cells (by default, more than 1 normalized log-count in 20 cells at most).

If a gene were a marker of a rare cell type, then there would be at least one cell neighborhood in which there is an enrichment of cells expressing the gene. Conversely, if a gene is not a marker of a rare cell type, but its changes in expression are driven by noise, we would expect it to be detected in cells that are randomly scattered across the KNN graph. In this case, the number of cells where the gene is detected in any given neighborhood follows a hypergeometric distribution (see Materials and Methods).

Starting from these observations, CIARA performs a one-tailed Fisher's test to verify whether the number of cells in which the gene is detected is enriched or not in the neighborhoods of all cells defined from the KNN graph. If a gene shows a significant enrichment (P<0.001 by default) in at least one neighborhood, then it is assigned a score equal to the minimum P-value across all neighborhoods (Fig. 1A, left); if the enrichment never reaches statistical significance, then it is assigned a score equal to 1.

All tested genes are then ranked by increasing scores: the genes with lower scores are those that are most likely to be markers of rare cell types. Such a ranked list is given in the output by CIARA (Fig. 1A, left).

CIARA can also generate a 2D representation of the data [e.g. with uniform manifold approximation and projection (UMAP); McInnes and Healy, 2018 preprint], which shows how many and which of the top selected genes each cell expresses and shares with

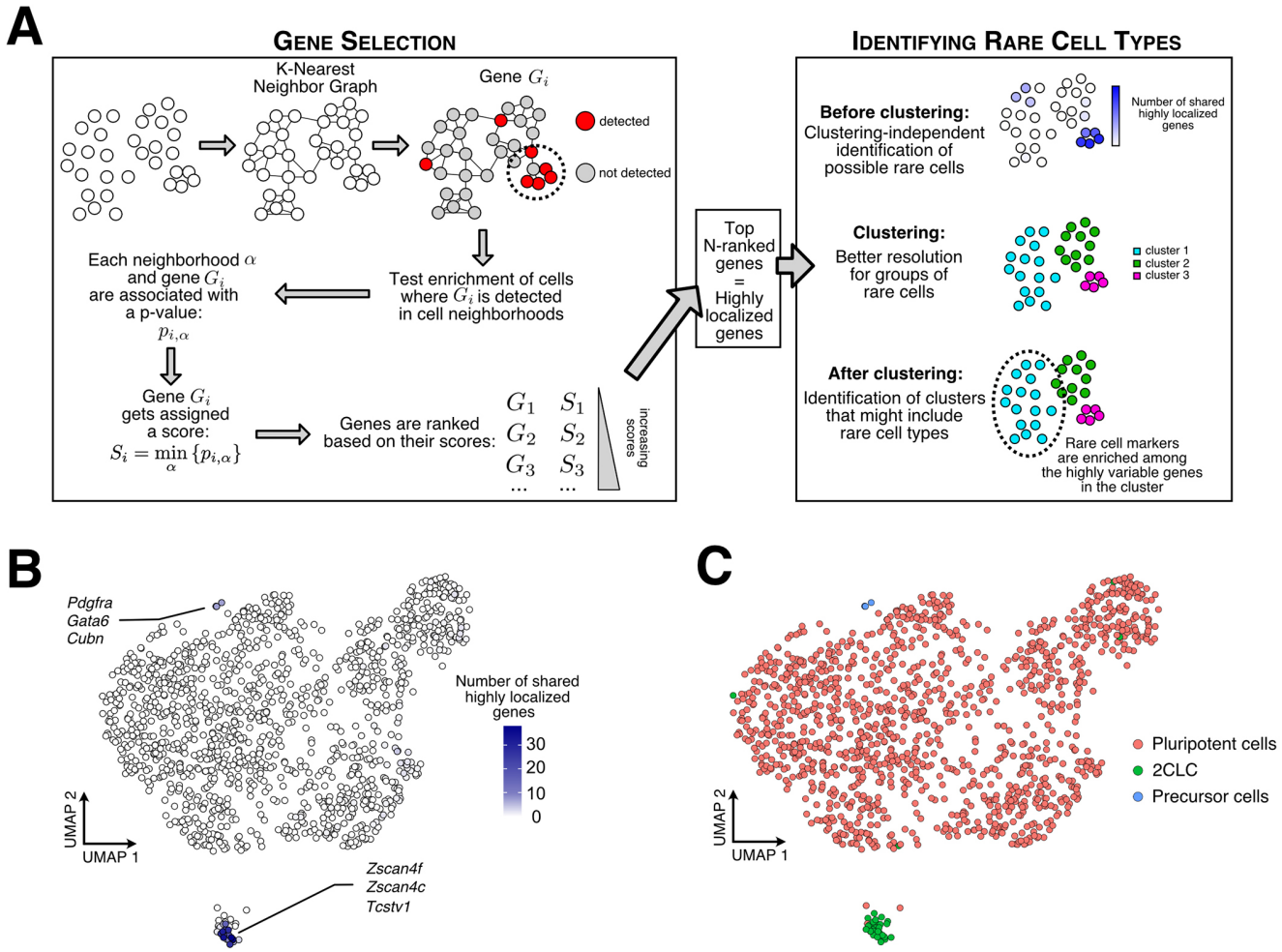


Fig. 1. Schematic representation and example of application of CIARA. (A) Left: CIARA computes a score for each gene based on how cells expressing that gene are distributed on a K-nearest neighbor graph. Lower scores correspond to genes that are mostly expressed in neighboring cells, i.e. are 'highly localized' and hence are more likely to be markers of rare cell types. Right: Summary of how the top-ranked genes are used to visualize and identify groups of rare cells. (B) UMAP representation of a previously published mESC dataset (n=1285 cells; Iturbide et al., 2021). The different shades of blue indicate the number of genes among the top 100 selected by CIARA that each given cell and its neighbors express. Nearby groups of darker-colored cells are more likely to represent rare cell types. (C) UMAP representation of the same dataset shown in B, with cells colored according to their cluster.

Development (2023) 150, dev201264. doi:10.1242/dev.201264

its neighbors (Fig. 1A, right; Fig. 1B). Such a plot is also available in an interactive format, where the names of the genes are displayed (see examples in Figs S7-S9). The significant genes can then be used with standard clustering algorithms to define the group of rare cell types, either on the whole dataset or within specific clusters that were previously defined in the data (Fig. 1A, right; Materials and Methods).

Importantly, CIARA also provides an unsupervised, quantitative evaluation of whether a cluster in the data may include a rare sub-population of cells: this is done by testing the statistical significance of the overlap between the set of highly variable genes within the cluster and the potential rare cell type markers identified by CIARA (Fig. 1A, right).

To showcase how CIARA works, we applied it to a previously published scRNA-seq dataset from mESCs (Iturbide et al., 2021) that includes a rare population of 2-cell-like cells (2CLC) (Macfarlan et al., 2012). The 2CLCs represent an *in vitro* model of totipotent-like cells and are typically present at a <1% frequency in mESC cultures. When applied to this dataset, CIARA found well-known 2CLC markers such as *Zscan4f* and *Zscan4c* among the top 15 ranked genes (Fig. 1B; Figs S1A, S7). By contrast, the genes that were ranked low were detected in a small number of cells that are not close on the KNN graph and, thus, are unlikely to represent any specific cell type (Fig. S1B). In addition, the top-ranked genes also included *Pdgfra* and *Gata6*, which are known markers of primitive endoderm cells, and thus expressed in differentiating cells (Iturbide et al., 2021; Wamaitha et al., 2015). A UMAP plot shows that the markers from 2CLCs and from those cells undergoing differentiation are expressed in two small groups of cells (Fig. 1B). Indeed, when the data was clustered with the 2475 genes selected by CIARA, we found three clusters: in addition to the largest cluster made of pluripotent cells, one cluster represents 2CLCs (18 cells, ∼2% of total), and the other includes four differentiating precursor cells (0.3% of the total; Fig. 1C; Fig. S1C,D). We also sub-sampled this mESC dataset to include fewer 2CLCs, and found that 2CLC markers are enriched among the genes selected by CIARA even when only three 2CLCs are present in the dataset (P<0.05, Fisher's exact test).

Furthermore, CIARA can also be applied to atlas-sized datasets (Fig. S2; Materials and Methods). To show this, we processed two datasets including ∼10^5 cells with CIARA, which led to the identification of several potential rare populations of cells expressing very specific markers (Table S7).

CIARA outperforms existing methods for rare cell type identification

We tested the performance of CIARA against several existing methods currently available to detect rare cell types from scRNA-seq datasets: GiniClust (Tsoucas and Yuan, 2018; Dong and Yuan, 2020), CellSIUS (Wegmann et al., 2019), FiRE (Jindal et al., 2018), RaceID (Herman et al., 2018) and GapClust (Fa et al., 2021). All these methods provide clusters of rare cells as output. As for CIARA, a list of rare cell type markers is also provided by GiniClust and CellSIUS. CellSIUS requires data partitioning in clusters as input, whereas all other algorithms do not. The features of the algorithms are summarized in Fig. 2A. We evaluated performance by quantifying the agreement between the classification of rare cells obtained with each method and the ground truth classification using Matthew's correlation coefficient (MCC; see Materials and Methods).

To make the comparison as fair as possible and minimize the effects of confounding factors due to, for example, the use of sub-optimal parameter settings, we ran a first series of tests on the datasets included in the papers where the alternative algorithms were introduced. The results of this benchmarking analysis are illustrated in Fig. 2B, and they show that CIARA generally outperforms the other algorithms (see also Materials and Methods; Fig. S3A-G). Specifically, CIARA tends to find fewer false positives, requires less manual curation (e.g. a manual merging of clusters), and can robustly detect extremely rare cell types (e.g. with n=3; see Materials and Methods).

We also ran all algorithms on a recently published scRNA-seq dataset that comprises 1195 cells from a human gastrula (Tyser et al., 2021b). A small population of seven primordial germ cells (PGCs) was identified within this dataset, which is marked by the expression of previously known PGC markers (*NANOS3*, *NANOG*, *DPPA5, SOX17*). However, PGCs have markers in common with other cell types, such as *SOX17* and *ETV4* (marking endodermal cells), which complicates the identification of PGCs with unsupervised methods.

CIARA detected a cluster including all seven PGCs, achieving an MCC=1 (see Materials and Methods). Conversely, all other algorithms achieved a lower MCC value (Fig. 2C; Fig. S3H,I; Materials and Methods).

In addition to the algorithms mentioned above, we ran three more algorithms on the human gastrula dataset: singleCellHayStack (Vandenbon and Diez, 2020), SAM (Tarashansky et al., 2019) and Triku (M Ascensión et al., 2022). Although not specifically designed for detecting rare cell types, these algorithms find genes that have a non-random distribution of expression values across cells. These approaches offer a valid alternative to standard differential expression analysis methods, but they tend to miss rare cell markers, as is seen with PGC markers (Fig. 2D; Materials and Methods).

Overall, these analyses show that CIARA performs better than alternative algorithms with respect to detecting rare cells in several published datasets, also in the most challenging situations when the rare cells share marker genes with more abundant cell types.

CIARA detects small changes in cell type composition in mESCs after retinoic acid treatment

Time-course scRNA-seq experiments are particularly suitable to the study of systems undergoing cell differentiation or reprogramming in order to capture and characterize cell types as they emerge, *in vivo* and *in vitro* (Griffiths et al., 2018).

In a recent study (Iturbide et al., 2021), we showed that low doses of retinoids induce the reprogramming of mESCs into 2CLCs, a cell type that resembles totipotent cells (Rodriguez-Terrones et al., 2018; Macfarlan et al., 2012). In particular, by performing a time-course experiment with scRNA-seq, we found that, although transcriptional changes are small within the first 12 h of treatment with retinoic acid (RA), after 48 h the cell type composition shows major changes: (1) the relative abundance of 2CLCs increases by ∼41% (from 2.6% to 44%) and (2) a small cluster of differentiating precursor cells (3%) is present.

However, when these transcriptional and cellular composition changes start to emerge and how long the RA treatment must be to produce any effects on cell fate decisions is unknown. Thus, we generated a new scRNA-seq dataset from mESCs following a 24 h RA treatment (Fig. 3A), and we analyzed the dataset with CIARA to determine changes in cell type composition.

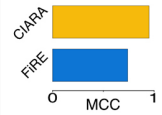
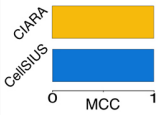
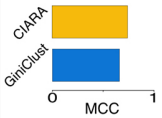
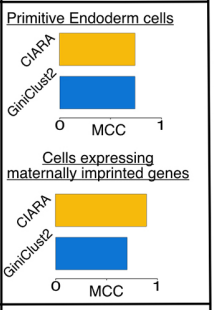
We first applied standard quality-control thresholds, which led to the selection of 766 good-quality cells (Fig. S4A-D; Materials and Methods). A UMAP plot showing which cells express the

Development (2023) 150, dev201264. doi:10.1242/dev.201264

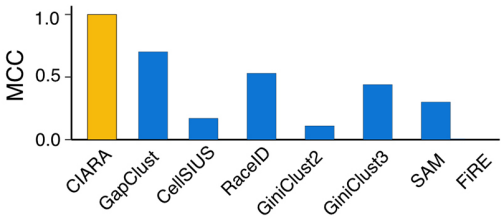
A

Method	Designed for rare cells identification	No need for pre-defined clusters	Output	
			Marker genes	Cells
CIARA	✓	✓	✓	✓
FiRE	✓	✓	✗	✓
CellSIUS	✓	✗	✓	✓
RaceID	✓	✓	✗	✓
GiniClust	✓	✓	✓	✓
GapClust	✓	✓	✗	✓
SingleCellHaystack	✗	✓	✓	✗
SAM	✗	✓	✓	✓
Triku	✗	✓	✓	✗

B

Method	Dataset	Comparison with CIARA
FiRE	1580 cells including a rare population of 40 Jurkat cells	
CellSIUS	3984 human cells including 3 H1437 cells and 6 Jurkat cells	
GiniClust	472 cells including glioblastoma primary tumour cells and a rare group of 16 oligodendrocytes	
GiniClust2	278 mouse embryonic stem cells with rare populations of primitive endoderm (9 cells) and cells expressing maternally imprinted genes (8 cells)	
RaceID	317 murine intestinal epithelial cells. RaceID identifies 4 rare cell types sub-divided in multiple clusters.	CIARA identifies the 4 rare cell types found by RaceID, plus an additional cluster of 3 Tuft cells (Fig. S2)

C



D

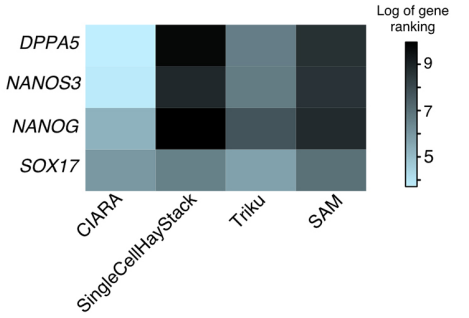


Fig. 2. CIARA outperforms existing methods for detecting rare cell types. (A) Table listing the methods for rare cell type identification that we benchmarked CIARA against. Specific features of each approach are indicated. (B) Table summarizing the data sets and results of the benchmarking analysis. The last column shows the values of the Matthews Correlation Coefficient (MCC) computed between the group of rare cells identified by each method and the ground truth. (C) MCC computed for the PGC group of cells present in the human gastrula data (Tyser et al., 2021b). (D) Heatmap showing the ranking (in natural log scale) of four PGC markers (rows) obtained by four methods (columns).

significant genes identified by CIARA (Fig. 3B) highlights the presence of two small groups of cells: one expressing well-known markers of 2CLCs, such as *Zscan4f*, *Zscan4c* and *Arg2*, and the other expressing markers of differentiating precursor cells, such as *Pdgfra*. Indeed, clustering with the genes selected by CIARA detected three different clusters (Fig. 3C). The largest cluster (744, ~97% of the total) included pluripotent cells expressing, for example, *Zfp42* (also known as *Rex1*) and *Sox2*; the intermediate cluster of 18 cells (~2%) corresponded to 2CLCs (marked by, for example, *Zscan4d*); and the smallest cluster of four cells (<1%) was marked by a distinct set of genes including differentiation markers such as *Gata4* and *Gata6* (Fig. 3D; Tables S1-S3).

A comparison with previously published datasets (Iturbide et al., 2021) confirmed that this small cluster includes four precursor cells that are compatible with those found at 0 h and 48 h of treatment (see Materials and Methods). These results indicate that, during the first 48 h of RA treatment, the cell types present within the mESC culture (as determined by their transcriptional features) remain the same, but their relative abundance changes only after more than 24 h of treatment (Fig. 3E). This is in agreement with the previously published quantification of 2CLCs by fluorescence-activated cell sorting (FACS), showing that the percentage of 2CLCs does not increase significantly after 24 h of RA treatment (Iturbide et al., 2021).

CIARA enables the discovery of rare cell types in a human gastrula dataset

Single-cell analyses are fundamental to mapping embryonic development and the first stages of cell differentiation. One of the milestones of embryo development is gastrulation, during which a single set of pluripotent cells (the epiblast) differentiates into three germ layers (endoderm, mesoderm and ectoderm), which later form the various organs.

Single-cell transcriptomics has contributed to revealing the steps of cell type diversification during gastrulation in several organisms (Briggs et al., 2018; Wagner et al., 2018; Nowotschin et al., 2019; Pijuan-Sala et al., 2019; Bergmann et al., 2022), including humans, with a recently published single-cell characterization of a human gastrula (Tyser et al., 2021b). A clustering analysis of this dataset revealed the presence of 11 main cell populations, some of which could be split into sub-clusters, representing, for example, different types of blood and endodermal cells (Tyser et al., 2021b).

Development (2023) 150, dev201264. doi:10.1242/dev.201264

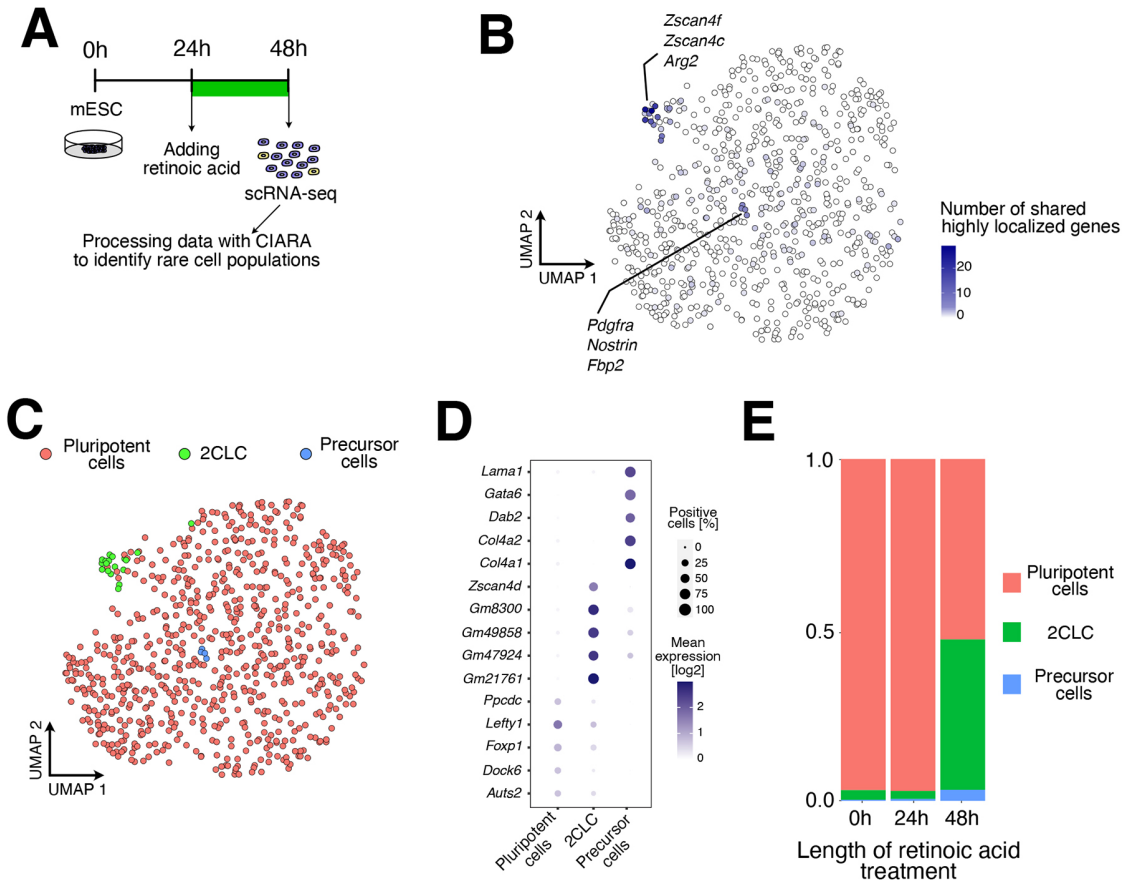


Fig. 3. CIARA identifies rare populations of totipotent-like and differentiating cells among mESCs treated with retinoic acid. (A) We treated mESCs with retinoic acid for 24 h before collecting and processing them for scRNA-seq. (B) UMAP representation of the mESC dataset (*n*=766 cells) indicating the number of highly localized genes expressed by each cell and shared with their neighbors. (C) Same UMAP representation as in B, with cells colored by cluster. (D) Top marker genes of the clusters found in the mESC data. The markers for the clusters were detected with the 'FindMarkers' function (with parameter only.pos=T) from Seurat (version 4.0.5). Only markers with adjusted *P*-value (based on the Bonferroni correction) below or equal to 0.05 were considered for downstream analysis. Finally, for each cluster, only unique markers (i.e. that are not included among the markers of other clusters) were kept. *Gm8300*, Eif1ad8. (E) Cell type composition changes in mESC datasets after 0 h, 24 h and 48h-long RA treatment. The datasets with 0 h (*n*=1285 cells) and 48h-long treatment (*n*=1867 cells) are taken from Iturbide et al. (2021).

Sub-dissection of the sample enabled the cells to be annotated based on their anatomical location: hence, cells could be identified as originating from the embryonic disk (rostral or caudal portions) or the extra-embryonic yolk sac.

Using CIARA, we performed an unsupervised analysis of this human gastrula dataset in order to search for rare cell types. In addition to the PGCs described above (Fig. 2C,D), we found two small populations in the yolk sac endoderm (YSE) and the megakaryocyte-erythroid progenitor (MEP) clusters (Fig. 4A,B; Fig. S9).

The small YSE sub-cluster of 11 cells, which we named YSE1, expressed very specific markers, including, for example, members of the SERPIN family genes such as *SERPIND1* and *SERPINC1* (Fig. 4C; Fig. S5B; Table S4). These genes are known to be expressed in the adult kidney and liver (Heit et al., 2013), which is consistent with the functions that the yolk sac plays during early development (Ross and Boroviak, 2020). Interestingly, by running CIARA on an scRNA-seq dataset from mouse embryos at embryonic day (E) 7.75 to E8.25 (Tyser et al., 2021a), we found a sub-cluster of 21 endodermal cells that share the same transcriptional profile as the YSE1 cluster in the human embryo (see Materials and Methods; Fig. S5D; Table S6). This observation indicates that YSE1 is a relatively rare endodermal sub-population present in human and mouse embryos.

A diffusion map and pseudo-time analysis of the human endoderm cluster revealed that YSE1 is more transcriptionally distinct from the embryonic endoderm populations (represented by the definitive endoderm clusters) than the rest of the YSE cluster (Fig. 4D; Fig. S5C). Furthermore, all cells included in YSE1 derived from the yolk sac region, whereas the rest of the YSE cluster also included cells from the embryonic disk and were annotated as rostral or caudal (Fig. 4E). This transcriptional signature and separation in cell origin suggest that YSE1 represents a yolk sac endoderm population located further away from the embryonic–extra-embryonic boundary and, therefore, potentially closer to the forming blood islands where primitive erythropoiesis occurs (Tyser et al., 2021b). In support of this hypothesis, one of the markers of YSE1 was transferrin (*TF*), a protein iron carrier required for erythropoiesis (Richard and Verdier, 2020), the receptors of which, *TFRC* and *TFR2*, are expressed by erythroblasts (Fig. S5E).

The second population of rare cells detected by CIARA was in the MEP cluster, which we named MEP1 (Fig. 4B). This cluster comprised 13 cells with a distinct transcriptional signature characterized by high levels of markers such as *PPBP*, *ITGA2B*

Development (2023) 150, dev201264. doi:10.1242/dev.201264

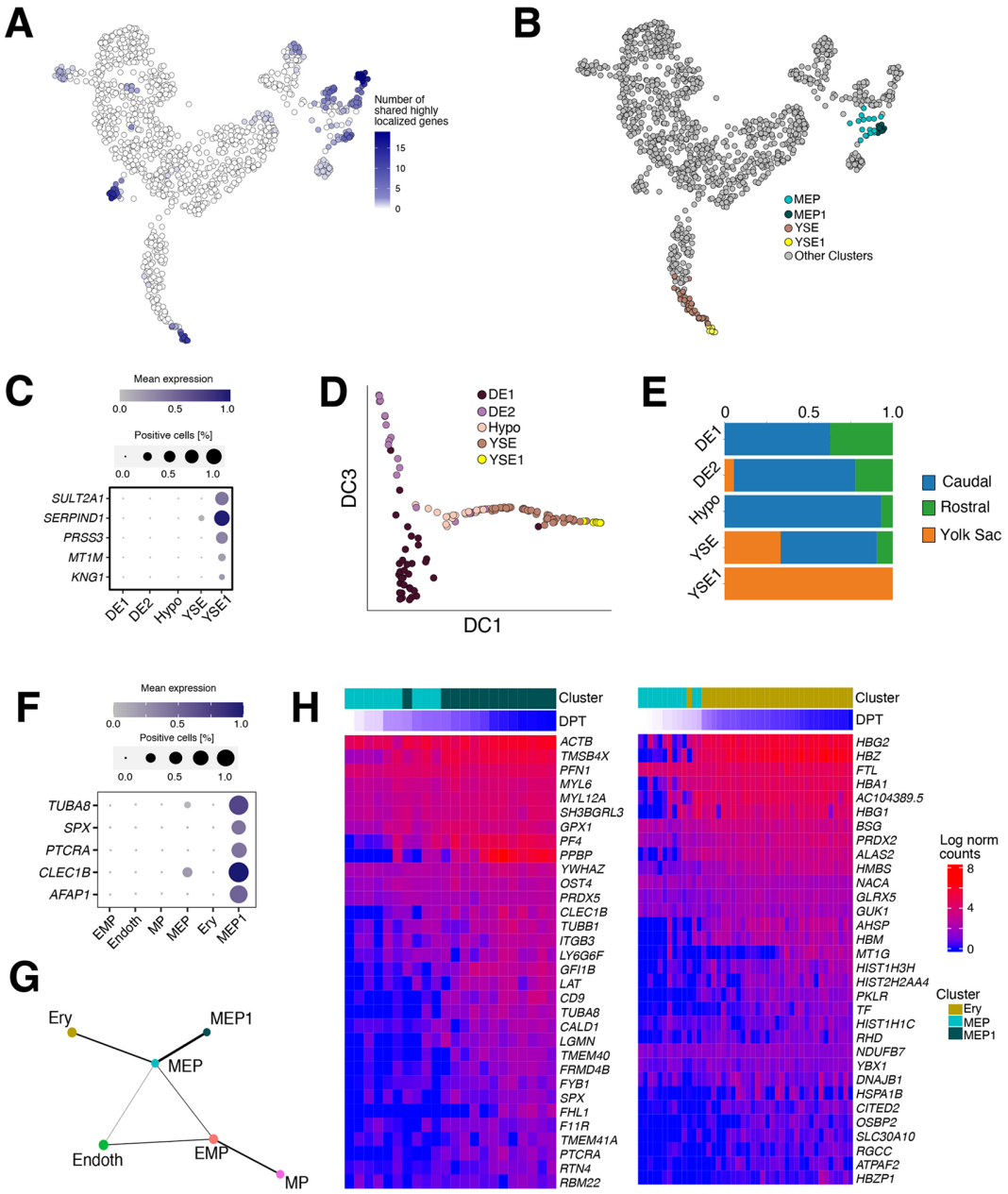


Fig. 4. CIARA identifies previously uncharacterized rare populations of cells in a human gastrula dataset. (A) UMAP representation of the human gastrula dataset (*n*=1195 cells; Tyser et al., 2021b) showing the number of shared highly localized genes in each cell. (B) Same UMAP representation as in A, with cells colored according to the clusters they belong to. The sub-clusters highlighted, YSE and MEP, are those in which CIARA finds new rare cell populations (YSE1 and MEP1). (C) Top marker genes of the YSE1 rare cell population. Mean expression levels are normalized by the maximum within each cluster. The markers for the clusters were detected with the 'FindMarkers' function (with parameter only.pos=T) from Seurat (version 4.0.5). Only markers with adjusted *P*-value (based on the Bonferroni correction) below or equal to 0.05 were considered for downstream analysis. Finally, for each cluster, only unique markers (i.e. that are not included among the markers of other clusters) were kept. (D) Diffusion components 1 and 3 (DC1, DC3) of the endodermal cells (*n*=135 cells). (E) Stacked bar plot showing the distribution of the anatomical origin of cells in each cluster. (F) Top marker genes of the MEP1 rare cell population, identified as explained above. Mean expression levels are normalized by the maximum within each cluster. (G) Graphical representation of the connectivity between the clusters of blood cells (*n*=143 cells) estimated with PAGA (Wolf et al., 2019). (H) Top differentially expressed genes along the differentiation trajectories joining MEP and MEP1 (left) or MEP and Ery (right). The trajectory analysis was performed using the function slingshot (with start.clus='MEP' and reducedDim equal to the diffusion map provided in the original human gastrula paper) from the R library slingshot version 1.6.1 (Street et al., 2018). To identify differentially expressed genes along the differentiation trajectories joining MEP and MEP1 or MEP and Ery, the functions 'fitGAM' and 'startVsEndTest' (with parameter lineage equal to TRUE) from the R package tradeSeq version 1.2.1 (Van den Berge et al., 2020) were used. HIST1H3H, H3C10; HIST2H2AA4, H2AC19; HIST1H1C, H1-2. DE, definitive endoderm; EMP, erythro-myeloid progenitors; Endoth, endothelium; Ery, erythroblasts; Hypo, hypoblast; MEP, megakaryocyte-erythroid progenitors; MP, myeloid progenitors; YSE, yolk sac endoderm.

Development (2023) 150, dev201264. doi:10.1242/dev.201264

and *GP1BB* (Fig. 4F; Fig. S5F; Table S5). Based on the expression of these and other markers (*LAT*, *CLEC1B*, *TREML1*, *RAB27B*; Pijuan-Sala et al., 2019), we identified these cells as megakaryocytes, a population of cells reported to be present in the early human embryo (Ivanovs et al., 2017), but not transcriptionally defined. This conclusion is also supported by the analysis of the differentiation trajectories within the blood clusters (Fig. 4G,H; Materials and Methods). Specifically, we found a branching event where the MEP cluster splits into the MEP1 cluster (when megakaryocytes markers are upregulated) and erythroblasts (Fig. 4G), allowing us to identify genes marking the differentiation between these two cell types (Fig. 4H).

An analogous rare population of megakaryocytes with the same transcriptional signature was also identified in mice at a later developmental stage (see Materials and Methods), which is consistent with human hematopoiesis starting earlier than in mice, as suggested by other analyses (Tyser et al., 2021b).

CIARA identifies rare cells across multiple single-cell data modalities

So far, we have shown applications of CIARA to scRNA-seq datasets. However, the main requirement of CIARA is the definition of a KNN graph, which can be built with any type of data where a notion of distance is defined. Hence, its applicability is very broad; in particular, CIARA can be applied to any type of single-cell omic datasets, such as DNA-seq, assay for transposase-accessible chromatin with sequencing (ATAC-seq), bisulfite sequencing, etc. Such wide applicability could be used, for instance, to identify rare populations of cells across multiple data modalities.

As proof of principle, we ran CIARA on a paired scRNA/ATAC-seq dataset generated from 34,774 mouse skin cells with the SHARE-seq protocol (Ma et al., 2020). Running CIARA on each modality provided a list of cells that had at least one localized feature shared with its neighbors, which represent candidates for rare cell types (Fig. S6). With the scATAC-seq modality, we performed the analysis using different sets of features: peaks, genes or enhancers (Danese et al., 2021).

We then computed the overlaps of these lists of cells obtained from the scRNA-seq and the scATAC-seq modalities, and found that they are statistically significant (Fisher's exact test, *P*-values were all less than ∼5e−5; see Materials and Methods), regardless of the features used in the analysis of the scATAC-seq dataset (Fig. S6D; Materials and Methods). One example of a potential rare population that CIARA found is a group of seven cells in the endothelial cluster, which emerged both in the scRNA-seq and the scATAC-seq modalities (Fig. S6A-C).

Overall, this result suggests that rare cell types can be identified across multiple modalities with CIARA, which could help validate the presence of rare cells and find genes or enhancers regulating their emergence.

DISCUSSION

We have developed a new algorithm, CIARA, that can identify potential marker genes of rare populations of cells in scRNA-seq data. Starting from a KNN graph, CIARA compares the number of cells in which a gene is detected in each K-neighborhood with the value expected from a hypergeometric distribution and then combines the results across all neighborhoods to provide a 'score' for each gene. Lower scores indicate a tendency of a gene to be detected only in small groups of cells with similar transcriptomes, which suggests that the gene is a potential marker of a rare cell type.

These marker genes can then be used to find rare cell types by exploring the data in a cluster-independent manner or in combination with standard clustering algorithms. This results in the identification of groups of rare cell types that are typically missed when following common strategies involving gene selection based on, for example, high variability. In the implementation we presented, CIARA identifies marker genes that are detected in small cell populations only. However, the algorithm can be generalized to find marker genes that are expressed in multiple cell populations (Materials and Methods; Fig. S1E).

The use of an exact probability distribution to compute the scores and the lack of a requirement for pre-defined clusters (similar to recently published methods to compare cell type abundance across conditions; Dann et al., 2022) imply that, to run CIARA, only few parameters need to be specified and that it can scale to atlas-size datasets (Fig. S2).

Both R and Python are standard choices for scRNA-seq data analysis in the scientific community. Hence, we made CIARA available as R and Python packages: the R package is available from CRAN, and the Python package can be downloaded from GitHub. Both packages can be easily integrated with standard analysis pipelines based on, for example, Seurat (Hao et al., 2021) and Scanpy (Wolf et al., 2018) (see 'Code Availability' in the Materials and Methods section for details).

The identification of rare cell types is an important task in single-cell omic data analysis; thus, in the last few years, many algorithms to identify rare cell types have been developed.

We performed a comprehensive benchmarking of CIARA against five algorithms for rare cell detection and three algorithms for gene selection. We showed that CIARA outperforms all these algorithms with respect to the identification of rare cells and their markers (Fig. 2), as it is able to cope with extremely rare populations (down to approximately three cells in the datasets we analyzed here) that might be specified by a limited number of markers. For example, CIARA was the only algorithm able to identify in an unsupervised manner a group of seven primordial germ cells in an scRNA-seq dataset from a human gastrula (Fig. 2C,D; Fig. S5A; Materials and Methods).

To demonstrate CIARA's capabilities further, we applied it to two datasets.

The first dataset was a newly generated scRNA-seq data from mESCs treated with retinoic acid for 24 h. In addition to a cluster of reprogrammed 2CLCs, CIARA identified a small group of four differentiating, precursor cells. By comparing these results with a previously published study (Iturbide et al., 2021), we found that after 24 h of retinoic acid treatment, the same cell types as after a 48h-long treatment are present, even though the relative abundance is different (Fig. 3).

The second dataset included cells from a gastrulating human embryo (Tyser et al., 2021b). CIARA identified in an unsupervised manner two previously uncharacterized rare cell populations. One rare cell group was composed of endodermal cells from the yolk sac, likely located in a region distant from the embryonic disk and potentially closer to differentiating blood cells. The other group of rare cells represents megakaryocytes, and their identification allowed us to reconstruct the transcriptional changes during primitive blood differentiation (Fig. 4).

These applications exemplify two general tasks for which CIARA can be employed: first, the detection of small changes in cell type composition over a time-course experiment; second, the characterization of a system in which new cell types are just emerging, to pinpoint the first transcriptional steps that accompany cellular fate decisions.

Development (2023) 150, dev201264. doi:10.1242/dev.201264

CIARA is a powerful method to identify and characterize rare cell types, and its main requirement is the definition of a KNN graph. Hence, it is applicable to any single-cell dataset, such as ATAC-seq (Fig. S6). In particular, the application of CIARA to multi-omic datasets allows the identification of rare cells across multiple modalities, which could lead to a more in-depth characterization of rare cell types as they differentiate.

MATERIALS AND METHODS

CIARA algorithm

Gene selection

CIARA starts from a normalized gene count matrix and KNN graph, which can be built with standard approaches available in the Seurat (Hao et al., 2021) or Scanpy (Wolf et al., 2018) libraries. Given its goal to find potential markers of rare cells, CIARA performs a filtering step to select only genes that are expressed above a threshold value in a relatively small number of cells. All thresholds can be set manually; otherwise, default values will apply for the following parameters: threshold expression value, threshold=1; minimum number of cells, *n*_cells_low=3. The user needs to specify the maximum number of cells in which a gene can be detected (*n*_cells_high). Unless specified otherwise, we used *n*_cells_high=20 in all the analyses we performed. It might be useful to increase this value in experiments with higher sensitivities, where genes tend to be detected in a larger number of cells. Although we found that the default parameters work well with all the datasets we analyzed, we verified that the results are robust to parameter changes (see 'Robustness analysis' section).

For the genes that pass this filtering step, CIARA carries out a one-sided Fisher's exact test to check whether there is a statistically significant enrichment of cells expressing the gene in each neighborhood (formed by a cell and its KNN). This is done with the function 'fisher.test' in R, with the option 'alternative=greater'. By default, the result of the test is considered statistically significant if the unadjusted *P*-value is less than 0.001.

All the genes that show statistically significant enrichment in at least one neighborhood have expression patterns that are highly localized and are considered potential markers of rare cell types. These highly localized genes are assigned a score equal to the minimum *P*-value obtained across all neighborhoods (Fig. 1A). Such a score is used to rank the genes, with smaller scores being associated with genes that are more strongly enriched in at least one neighborhood. If a gene is not enriched in any neighborhood, it gets assigned a score equal to 1.

The gene selection procedure can be generalized to select marker genes that are expressed in multiple cell types. To achieve this, first, only the top 10% (default value) genes with the largest interdecile range are considered. Then, the expression levels are binarized by assigning a value of 1 to the 20 cells (default value) with the highest expression values and a value of 0 to all other cells. Finally, the standard procedure of CIARA is run to identify the genes with local enrichments of '1' values on the KNN graph.

The genes selected by such a procedure (implemented in the function 'get_background_full' with the option 'extend_binarization=TRUE') will include markers that have higher levels of expression in potential rare cell types but are also expressed in other cell types in the dataset. For example, by using this function on the mESC dataset, we were able to detect 2CLC markers such as *Tmem72*, which are ubiquitously expressed in the dataset but have higher levels in 2CLCs (see Fig. S1E).

Identifying rare cells

The highly localized genes (having a score <1, see above) are used by CIARA to identify groups of rare cell types following two main strategies.

The first is clustering independent, and consists of counting for every single cell the number of highly localized genes expressed in that cell and in its KNN: the larger this number, the more likely it is that the cell is part of a group of rare cells. The results of this analysis are reported in a 2D representation of the data, such as a UMAP plot (see Figs 1B, 3B and 4A), where each cell is colored based on the number of highly localized genes expressed and shared across the KNN. These 2D plots are also available in an interactive html format; hovering the mouse cursor over any cell reveals the names of the top highly localized genes expressed and shared across the KNN (see Figs S7-S9).

The second strategy for rare cell type identification is based on utilizing standard clustering algorithms with the highly localized genes selected by CIARA. In the R version of CIARA, clustering is done with the Louvain algorithm on the first 30 principal components as a default value (defined from the top 2000 highly variable genes) with the functions 'FindNeighbors' and 'FindClusters' from the R library Seurat version 4.0.5.

The clustering can involve the entire dataset or only part of it. In particular, given an existing partition of the data, CIARA can verify which clusters are more likely to include groups of rare cell types by testing the enrichment of highly localized genes among the top 100 (default value) highly variable genes within each cluster (Fisher's test, $P<0.001$ and odds ratio greater than 1). Clusters that show a significant enrichment are then sub-clustered with the same algorithm as specified above.

Marker gene identification

The markers for the clusters identified by CIARA are detected with the 'FindMarkers' function (with parameter only.pos=T) from Seurat (version 4.0.5). Only markers with adjusted *P*-value (based on the Bonferroni correction) less than or equal to 0.05 are considered for downstream analysis. Finally, for each cluster, only unique markers (i.e. those not included among the markers of other clusters) are kept.

Unless otherwise specified, in the balloon plots showing marker genes expression the size of the dots is determined by the fraction of cells with log norm counts above 1 (function 'NormalizeData' from R library Seurat).

Analysis of previously published datasets for method benchmarking

Below, we briefly describe the datasets we used for the benchmarking analysis shown in Fig. 2B. To evaluate the performance of each algorithm, we quantified the agreement between the classification of rare cells obtained with each method and the ground truth classification using the MCC. MCC is a metric that quantifies the overall agreement between two binary classifications, taking into account both true and false positives and negatives. MCC values range from −1 to 1, where 1 indicates a perfect agreement between clustering and the ground truth, 0 means the clustering is as good as a random guess, and −1 indicates no overlap between the clustering and the ground truth. MCC is computed with the function 'mcc' from the R library mltools version 0.3.5 (https://CRAN.R-project.org/package=mltools). The MCC values shown in Fig. 2B,C for each algorithm represent the maximum values obtained across all clusters.

In all the datasets analyzed with CIARA, the normalized count matrix was obtained with the function 'NormalizeData' (with parameter normalization.method=LogNormalize) and the KNN graph was built with the function 'FindNeighbors' (on the first 30 principal components built from the top 2000 highly variable genes). Both functions are from Seurat version 4.0.5.

293T and Jurkat cells (Fig. 2B)

This dataset of 1580 cells comprises 293T and Jurkat cells in a known proportion, with the Jurkat cells being the rare population (40 cells, ∼2.5% of total cells). This dataset was previously analyzed using FiRE (Jindal et al., 2018).

Here, CIARA identified 2077 highly localized genes. By clustering the data with these genes, we found two clusters (resolution 0.1, k.param equal to 5 and number of principal components equal to 30), one of which corresponded to Jurkat cells, based on the markers expressed.

CIARA outperforms FiRE (MCC values are 0.95 and 0.74, respectively; see Fig. 2B), based on fewer false positives (four cells) compared with those detected by FiRE (32 cells).

Mixture of eight human cell lines (Fig. 2B)

This dataset includes 3984 cells, and was previously analyzed using CellSIUS (Wegmann et al., 2019). Two rare populations of H1437 and Jurkat cells (three and six cells, respectively) are present and marked in the dataset.

Development (2023) 150, dev201264. doi:10.1242/dev.201264

Here, CIARA identified 3704 highly localized genes. By clustering the data with these genes, we identified nine clusters (resolution 0.1, k.param equal to 3, and number of principal components equal to 30). Two of these clusters could be identified as H1437 and Jurkat cells based on their markers. Hence, CIARA could identify both of these rare cell types, achieving the same performance as CellSIUS (MCC equal to 1 for both methods; Fig. 2B).

Glioblastoma (GBM) primary tumors (Fig. 2B)

This dataset includes 472 cells, and was previously analyzed using GiniClust (Jiang et al., 2016). It includes a small group of 16 oligodendrocytes, which are defined as the cells co-expressing the four marker genes *CLDN11*, *MBP*, *PLP1* and *KLK6* (Jiang et al., 2016). CIARA identified 68 highly localized genes. By clustering the data with these genes, we identified 13 clusters (resolution 0.1, k.param equal to 3 and number of principal components equal to 30), one of which corresponded to oligodendrocytes.

Differentiating mESCs at day 4 after LIF withdrawal (Fig. 2B)

This dataset includes 278 mESCs that are differentiating after LIF removal, and was previously analyzed using GiniClust2 (Tsoucas and Yuan, 2018). On day 4 after LIF removal, two small clusters of cells (nine and eight cells) were detected by GiniClust2, which, based on their markers, were identified as cells differentiating towards primitive endoderm (PrE cells; markers: *Col4a1*, *Col4a2*, *Lama1*, *Lama2* and *Ctsl*) and cells expressing maternally imprinted genes (*Rhox6*, *Rhox9* and *Sct*).

CIARA identified 287 highly localized genes. By clustering the data with these genes, we identified three clusters (resolution 0.3, k.param equal to 5 and number of principal components equal to 30). Two of these clusters expressed the same markers as the rare cells identified by GiniClust2 (Fig. S3A-C). Although in this dataset we lack a 'ground truth' for the rare cells, we defined a set of 'bona fide' clusters based on the co-expression of the marker genes mentioned above, and we computed the MCC values of GiniClust2 and CIARA using these clusters as reference. The two methods had the same MCC score for the cluster of differentiating cells, but CIARA achieved a higher MCC value on the set of cells expressing maternally imprinted genes (Fig. 2B).

Murine intestinal epithelial cells (Fig. 2B)

This dataset includes 317 cells and was previously analyzed with RaceID (Grün et al., 2016) (see vignette https://cran.r-project.org/web/packages/RaceID/vignettes/RaceID.html). Here, four rare cell types (enterocytes, goblet cells, Paneth cells and enteroendocrine cells) were found after manually merging multiple clusters expressing similar marker genes (Grün et al., 2016). CIARA identified 1514 highly localized genes. By clustering the data with these genes, we found eight clusters (resolution 0.2, k.param equal to 3, and number of principal components equal to 20), four of which correspond to the rare cell types that RaceID found. Additionally, one of the clusters found by CIARA (cluster number 7) expressed markers of Tuft cells (Fig. S3F).

The markers for the dataset (using the clusters defined with CIARA, see Fig. S3D-F) were identified as specified above.

To investigate the relationship between the six smallest clusters (≤25 cells) detected by CIARA (2, 3, 4, 5, 6 and 7) and the original cluster partition obtained with RaceID, a plot was generated with the function 'clustree' from the R package clustree version 0.4.4 (Zappia and Oshlack, 2018; Fig. S3G).

Among these clusters identified by CIARA, cluster 2 corresponds to goblet cells (marked by *Clca3*), cluster 3 to enterocytes (marked by *Apoa1*), cluster 4 to Paneth cells (marked by *Defa24*), cluster 5 to enteroendocrine cells (marked by *Chgb*) and cluster 7 to Tuft cells (Herman et al., 2018). The markers used to label the clusters from 2 to 5 are described in Fig. S2B from Grün et al. (2016), where the data were published. The clustree plot in Fig. S3G shows that each of the above rare cell types identified by CIARA are split between several clusters with RaceID. Cluster 6 (4 cells) shows a very clear transcriptional profile and corresponds to a cell type not previously described (Fig. S3E).

Identification of PGCs from a human gastrula dataset

We analyzed a previously published human gastrula dataset from Tyser et al. (2021b) using CIARA and the other seven algorithms we tested in Fig. 2. Among the 1195 cells of this dataset, there is a small population of seven PGCs, which were identified by Tyser et al. (2021b) in a supervised way (i.e. by using the co-expression of known PGC markers such as *NANOS3*, *NANOG* and *DPPA5*). We describe below how we ran the algorithms and tested their ability to find PGCs.

CIARA found 2917 highly localized genes in the whole dataset. By clustering the data with these genes, the seven PGCs are always identified as a single cluster over a wide range of resolutions (Fig. S5A).

GiniClust2 and GiniClust3 pipelines were used following the documentation available from https://github.com/dtsoucas/GiniClust2 and https://github.com/rdong08/GiniClust3 with default values for all parameters. Note that the gene selection based on the Gini index tends to miss PGC markers owing to their low average expression values (Fig. S3H,I).

For CellSIUS, we used the R package available from https://github.com/Novartis/CellSIUS/. We decreased the value of the 'min_*n*_cells' parameter from its default value 10 to 5 (given that there are only seven PGCs in the data), whereas default values were used for the other parameters.

The FiRE R package is available from https://github.com/princethewinner/FiRE. Using the default threshold on the FiRE score (i.e. 1.5*interquartile range+third quantile), no rare cells were identified. Hence, we chose a less stringent threshold of 0.5*interquartile range+third quantile. Because FiRE does not provide clusters of cells as output, for the MCC computation we considered the rare cells identified by FiRE in the 'Primitive Streak' cluster as PGCs.

The analysis with RaceID 3 was performed with standard parameters using the R package https://github.com/dgrun/RaceID3_StemID2_package.

For the analysis with GapClust, we used the implementation available from the GitHub repository https://github.com/fabotao/GapClust with default parameters.

The SingleCellHaystack algorithm is implemented in the R package available from https://github.com/alexisvdb/singleCellHaystack. Default values were used for all parameters, and the algorithm was run on the first 30 principal components.

Analysis with SAM was performed with default values of all parameters from the Python package https://github.com/atarashansky/self-assembling-manifold/tree/master.

The Triku algorithm is implemented from the Python package available from the website https://triku.readthedocs.io/en/latest/. This website also includes a tutorial that we followed to perform our analysis. For gene filtering, we ran the function pp.filter_genes from Scanpy (version 1.8.0) with min_cells=3 instead of the default value equal to 10 (given that the number of PGCs is less than 10).

SingleCellHaystack, SAM and Triku return a ranked list of 'most informative' genes having a non-random distribution of expression values across cells. We verified whether the top 1000 genes selected by these three algorithms were enriched with PGC markers by running a Fisher's test (R function 'fisher.test' with alternative='two.sided') using as background all the genes with normalized expression above 0.5 in more than six cells. None of the tested methods showed a statistically significant enrichment of PGC markers, apart from CIARA ($P=8\times10^{-4}$). The data normalization was done with the function 'NormalizeData' from Seurat with parameter normalization.method='LogNormalize'.

The PGC markers were detected with the 'FindMarkers' function (with parameter only.pos=T) from Seurat using a threshold for the Bonferroni-adjusted *P*-value of 0.05 and excluding all genes that were also markers of other non-PGC clusters.

mESCs experiment

Cell culture

Cells were grown in a medium containing DMEM-GlutaMAX-I, 15% fetal bovine serum, 0.1 mM 2-β-mercaptoethanol, non-essential amino acids, penicillin and streptomycin and 2× LIF over gelatin-coated plates. The medium was supplemented with 2i (3 µM CHIR99021 and 1 µM PD0324901, Miltenyi Biotec) for maintenance and expansion. The 2i was

Development (2023) 150, dev201264. doi:10.1242/dev.201264

removed 24 h before the addition of RA as described by Iturbide et al. (2021).

scRNA-seq

Cells were collected after RA treatment and sorted for live single cells by FACS. Cells were then counted and tested for viability with an automated cell counter. Five thousand cells of the sample were then input into the 10x Genomics 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 (Melsted et al., 2021). First, the 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 performed 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).

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 (Lun et al., 2019). For this, a lower threshold of 1000 UMI counts per barcode was considered. Afterward, quality control was performed using the Scanpy library. Cells having more than 10% counts mapped to mitochondrial genes or fewer than 1000 detected genes were removed. After quality control, 766 cells were kept for downstream analysis (Fig. S4A-D).

Analysis with CIARA

CIARA identified 2475 highly localized genes in this dataset. We ran cluster analysis on these genes with the 'FindNeighbors' (on the 30 top principal components and with k.param equal to 3) and 'FindClusters' functions (with resolution 0.1), which gave three clusters.

The marker genes of these clusters (see Tables S1-S3) were detected with the 'FindMarkers' function (with parameter only.pos=T) from Seurat. Only markers with an adjusted *P*-value based on Bonferroni correction below or equal to 0.05 (for 2CLCs and precursor cells) or with a *P*-value below 0.05 (for pluripotent cells) are considered for downstream analysis. Moreover, for each cluster, only unique markers (e.g. those not included in the marker list of other clusters) were kept.

Based on the lists of marker genes, the three clusters could be identified as pluripotent cells, 2CLCs and precursor cells (Fig. 3B-D).

Comparison with previously published mESC data

We compared the clusters found in our mESC dataset with those in the previously published mESC datasets after a 0 h and 48 h RA treatment (Iturbide et al., 2021). The dataset at 0 h was re-analyzed with CIARA, which identified 3302 highly localized genes. Using these genes, we performed clustering with the functions 'FindNeighbors' (on the top 30 principal components with k.param=5) and 'FindClusters' with resolution 0.1, which gave three clusters. Based on their markers (found with the procedure described above), these clusters could be identified as pluripotent cells (1245 cells), 2CLCs (36 cells) and precursor cells (four cells; Fig. S1). These same clusters were identified in the dataset at 48 h by Iturbide et al. (2021).

We assessed the statistical significance of the intersection between the markers of the three clusters found at 0 h, 24 h and 48 h by using a Fisher's test (with the 'fisher.test' function from the R package stats, with 'alternative=two.sided').

The intersections between the markers of the precursor cells clusters at 24 h versus 48 h ($P=7\times10^{-48}$) and at 24 h versus 0 h ($P=10^{-31}$) were both statistically significant.

Similarly, the markers of the 2CLC cluster had a significant overlap at 24 h versus 48 h ($P=9\times10^{-102}$) and at 24 h versus 0 h ($P=6\times10^{-83}$).

Finally, also the intersections between the markers of pluripotent cells at 24 h versus 0 h ($P=0.0001$) and at 24 h versus 48 h ($P=2\times10^{-91}$) were statistically significant.

Identifying rare cell types in the human gastrula dataset

First, we tested the enrichment of the 2917 highly localized genes found by CIARA among the top 100 highly variable genes (HVGs) within each of the clusters provided by Tyser et al. (2021b) (as described above in the 'Identifying rare cells' section).

We found a statistically significant overlap in the endoderm (Endo; $P=4\times10^{-5}$) and the hemato-endothelial progenitors (HEP; $P=4\times10^{-5}$) clusters. Then, we sub-clustered the Endo and HEP clusters using their HVGs (for the Endo cluster: resolution=0.2, k.param=5, top 30 principal components; for the HEP cluster: resolution=0.6, k.param=5, top 30 principal components). The two smallest clusters found in the Endo and HEP clusters are denoted as YSE1 and MEP1, respectively, and they were not described by Tyser et al. (2021b).

Marker analysis

The markers for the human gastrula were detected with the 'FindMarkers' function (with parameter only.pos=T) from Seurat, with the same criteria described above. The analysis was run separately using all sub-clusters reported by Tyser et al. (2021b) for the Endo cluster (including the new rare cluster found by CIARA, YSE1) and the HEP cluster (including MEP1 found by CIARA).

Trajectory and PAGA analysis

For the cells in the Endo sub-clusters (i.e. DE1, DE2, YSE, Hypoblast and YSE1), a diffusion map was computed from the normalized count matrix with the top 2000 highly variable genes (using the 'NormalizeData' and 'FindVariableFeatures' functions from Seurat) with the function 'DiffusionMap' from the R package destiny version 3.2.0 (Angerer et al., 2016). The diffusion pseudotime was computed using the function 'DPT' from the same package.

For the cells in the HEP sub-clusters (EMP, HE, MP, MEP, MEP1) and the erythroblast cluster, trajectory analysis was performed using the function 'slingshot' (with start.clus='MEP' and reducedDim equal to the diffusion map provided in the original human gastrula paper) from the R library slingshot version 1.6.1 (Street et al., 2018).

To identify differentially expressed genes along the differentiation trajectories joining MEP and MEP1 or MEP and erythroblasts, the functions 'fitGAM' and 'startVsEndTest' (with parameter lineage equal to TRUE) from the R package tradeSeq version 1.2.1 (Van den Berge et al., 2020) were used.

To estimate the connectivity between clusters, we performed an analysis with PAGA (Wolf et al., 2019) (functions tl.paga and pl.paga from Scanpy).

Comparison with published mouse datasets

We analyzed with CIARA a previously published dataset from mouse embryos at E7.75-E8.25 (Tyser et al., 2021a). This dataset of 665 cells included two small endodermal clusters. CIARA found 1700 highly localized genes (with *n*_cells_high=30); using these genes for clustering (resolution=0.2, k.param=5 and number of principal components equal to 30), we identified three clusters (Fig. S5D), one of which was a small sub-cluster of 21 cells in the endodermal cluster labeled as 'En2'. The markers of this sub-cluster (found with the procedure described above; Table S6) had a statistically significant overlap with the markers of the YSE1 cluster in the human gastrula ($P=0.0009$, two-sided Fisher's test; only mouse genes with a 1:1 human ortholog were considered, see below).

Pijuan-Sala et al. (2019) identified a cluster of megakaryocytes in mouse embryos. We tested the statistical significance of the overlap between the markers of these cells in mouse (from 'source data Fig. 3f' in Pijuan-Sala et al., 2019) and the markers of MEP1 cluster from the human gastrula using a two-sided Fisher's test, and obtained a *P*-value of 9×10^{-7}.

The genes in the two mouse datasets (Tyser et al., 2021a; Pijuan-Sala et al., 2019) were converted into the corresponding human orthologous

Development (2023) 150, dev201264. doi:10.1242/dev.201264

name if there was a 1:1 correspondence between the mouse and the human gene name, using g:Profiler (Raudvere et al., 2019).

Analysis of single-cell transcriptomic atlases

The mouse gastrulation atlas dataset (Pijuan-Sala et al., 2019) includes 116,312 cells. CIARA identified 3197 highly localized genes with parameters: threshold=1, *n*_cells_low=3, *n*_cells_high=20. The run time with the Python CIARA package was ∼3 h with eight 3.0 GHz cores.

The scRNA-seq dataset generated from human peripheral blood mononuclear cells (Zheng et al., 2017) includes 68,579 cells. CIARA identified 4207 highly localized genes with parameters: threshold=1, *n*_cells_low=3, *n*_cells_high=100). The run time with the Python CIARA package was ∼1.8 h with eight 3.0 GHz cores.

Analysis of the SHARE-seq dataset

The raw RNA and ATAC peak count matrices (mm10) were downloaded from Gene Expression Omnibus (GSE140203). The processing of the data was done using Scanpy 1.9.1 and epiScanpy 0.4.0. Further filtering of the count matrices was applied. For the RNA count matrix, cell barcodes containing fewer than 200 genes and genes present in fewer than three cells were filtered out, resulting in 40,780 cells×21,317 genes. For the ATAC peak count matrix, we binarized the counts and then filtered out barcodes with fewer than 1000 peaks as well as peaks present in fewer than 20 barcodes. We obtained a filtered count matrix of 34,166 cells×338,975 peaks.

Additionally, the ATAC gene and enhancer-based count matrix were built using the fragment file and the list of valid barcodes available on Gene Expression Omnibus as well as gene coordinates from GENCODE (release M1) and enhancers coordinates from the EnhancerAtlas 2.0 (Gao and Qian, 2020). The enhancer count matrix was binarized, and barcodes with fewer than 1000 peaks as well as enhancers present in fewer than ten barcodes were filtered out, resulting in 34,614 cells×420,475 enhancers.

Further processing was carried out identically for both RNA and ATAC count matrices. We normalized the data such that the library size had the same number of total counts per cell by dividing each cell by the total counts of all genes. The normalized counts were then log transformed. To build the KNN graph, we used 30 PCs and a number of neighbors of 15.

CIARA identified 639 highly localized genes for the RNA count matrix, 596 highly localized features for the ATAC gene-based count matrix, 17,652 highly localized features for the ATAC peak-based count matrix, and 8205 highly localized features for the ATAC enhancers-based count matrix.

Robustness analysis

We performed several robustness tests to verify how changes in the parameters affect the results obtained by CIARA.

Gene filtering is the first step in the CIARA algorithm, and is performed based on threshold values for the gene expression levels and the number of cells in which a gene is detected. To test CIARA's robustness relative to changes in these thresholds, we re-ran CIARA on all the datasets analyzed in this study using more/less stringent thresholds on the expression values (2 or 0.5 normalized log-count, instead of the default value of 1) and the maximum number of cells in which a gene is detected (10 or 30 cells, instead of the default value of 20). To compare the results, we computed the Pearson's correlation coefficients between the number of shared highly localized genes in each cell (which mark candidate rare cell types; see above and Figs 1B, 3B and 4A) obtained with the different settings (including the default one). In all cases, we obtained a statistically significant value of correlation (all *P*-values were less than ∼2.5e−20), indicating that, overall, the results are robust to changes in threshold values.

Another key step in CIARA is the building of the KNN graph, which requires the specification of the number of nearest neighbors, K, the number of highly variable genes, and the distance metric. We assessed the robustness of CIARA's results to changes in all of these parameters using two datasets: the mESC dataset (Fig. 1B,C) and the human gastrula dataset (Figs 2C,D and 4), in which the presence of rare cells is well documented [i.e. the 2CLCs and the precursor cells in the mESC dataset (Iturbide et al., 2021); and the PGCs in the human gastrula dataset (Tyser et al., 2021b)], but they mostly go undetected with existing methods (Fig. 2C,D).

First, we ran CIARA on the mESC dataset with different values of K (3, 5, 10) and of the expression threshold (0.5, 1, 2 log-counts). In each run, we verified with a Fisher's exact test whether the lists of marker genes of the two rare populations present in this dataset were enriched or not among the genes selected by CIARA. All the statistical tests run with the markers of both rare cell types were statistically significant (P<0.01), except for one parameter combination (K=10 and expression threshold=2. This suggests that CIARA is overall robust to changes in K and the expression threshold, but that increasing the number of neighbours and using a more stringent expression threshold can generally impair the identification of rare cell type markers.

Finally, we ran CIARA on the human gastrula dataset, choosing different numbers of highly variable genes (top 1000, 2000 or 3000), different values of K (3, 5, 10, 15, 20), and distance metrics (Euclidean or cosine distance). CIARA identified the cluster of seven PGCs and their markers with any of these combinations of parameters. Moreover, we also ran CIARA on the KNN graph generated after removing all PGC markers from the set of highly variable genes; even in this case, the PGC cluster was identified by clustering the data using the genes selected by CIARA. Taken together, these results suggest that CIARA is robust with respect to changes in parameters and can successfully identify very rare cells even when their markers are absent from the genes used to build the KNN graph.

Code availability

The code used to generate the figures in this paper is available at https://github.com/ScialdoneLab/CIARA. In this repository, there are also additional examples of applications of CIARA.

CIARA is available both in R (https://CRAN.R-project.org/package=CIARA) and Python (https://github.com/ScialdoneLab/CIARA_python). Both packages can be easily integrated with standard analysis pipelines based on, for example, Seurat (Hao et al., 2021) and Scanpy (Wolf et al., 2018).

Acknowledgements

We thank members of the Scialdone lab for discussions and feedback on the manuscript. We thank I. de la Rosa Velazquez and the Genomics Facility of Helmholtz Munich for sequencing, and M. Genet for advice.

Competing interests

F.J.T. consults for Immunai Inc., Singularity Bio B.V., CytoReason Ltd and Omniscope Ltd, and has ownership interest in Dermagnostix GmbH and Cellarity. The other authors declare no competing interests.

Author contributions

Conceptualization: G.L., A.S.; Methodology: G.L.; Software: G.L., M.S.; Validation: A.I., R.C.V.T., A.D., M.C.-T., S.S., M.-E.T.-P.; Formal analysis: G.L., M.S., M.R., R.C.V.T., A.D., M.C.-T., S.S., M.-E.T.-P.; Resources: A.I.; Data curation: G.L., A.I., M.L.R.T.S.; Writing - original draft: G.L., A.S.; Writing - review & editing: G.L., R.C.V.T., M.C.-T., F.J.T., S.S., M.-E.T.-P., A.S.; Visualization: G.L.; Supervision: A.S.; Project administration: A.S.; Funding acquisition: F.J.T., A.S.

Funding

Work in the Scialdone lab is funded by the Helmholtz Association. Work in the Torres-Padilla laboratory is funded by the Helmholtz Association, Helmholtz Zentrum München Small Molecule projects (Developmental projects) and the Deutsche Forschungsgemeinschaft (German Research Foundation; CRC 1064). A.I. was a recipient of a long-term European Molecular Biology Organization fellowship (ALTF 383-2016). G.L. was funded by the Bundesministerium für Bildung und Forschung project MechML (01IS18053A). M.S. was supported by the Helmholtz Association under the joint research school 'Munich School for Data Science – MUDS' and by an Add-on Fellowship for Interdisciplinary Life Science from the Joachim Herz Stiftung. A.D. was funded by the Deutsche Forschungsgemeinschaft (DFG STR 1385/5-1).

Data availability

Raw data for the mouse embryonic stem cells scRNA-seq dataset are available through ArrayExpress, under accession number E-MTAB-11610.

Peer review history

The peer review history is available online at https://journals.biologists.com/dev/lookup/doi/10.1242/dev.201264.reviewer-comments.pdf.

References

Angerer, P., Haghverdi, L., Büttner, 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. doi:10.1093/bioinformatics/btv715

Bergmann, S., Penfold, C. A., Slatery, E., Siriwardena, D., Drummer, C., Clark, S., Strawbridge, S. E., Kishimoto, K., Vickers, A., Tewary, M. et al. (2022). Spatial profiling of early primate gastrulation in utero. *Nature* **609**, 136-143. doi:10.1038/s41586-022-04953-1

Briggs, J. A., Weinreb, C., Wagner, D. E., Megason, S., Peshkin, L., Kirschner, M. W. and Klein, A. M. (2018). The dynamics of gene expression in vertebrate embryogenesis at single-cell resolution. *Science* **360**, eaar5780. doi:10.1126/science.aar5780

Danese, A., Richter, M. L., Chaichoompu, K., Fischer, D. S., Theis, F. J. and Colomé-Tatché, M. (2021). EpiScanpy: integrated single-cell epigenomic analysis. *Nat. Commun.* **12**, 5228. doi:10.1038/s41467-021-25131-3

Dann, E., Henderson, N. C., Teichmann, S. A., Morgan, M. D. and Marioni, J. C. (2022). Differential abundance testing on single-cell data using k-nearest neighbor graphs. *Nat. Biotechnol.* **40**, 245-253. doi:10.1038/s41587-021-01033-z

Dong, R. and Yuan, G.-C. (2020). GiniClust3: a fast and memory-efficient tool for rare cell type identification. *BMC Bioinformatics* **21**, 158. doi:10.1186/s12859-020-3482-1

Emert, B. L., Cote, C. J., Torre, E. A., Dardani, I. P., Jiang, C. L., Jain, N., Shaffer, S. M. and Raj, A. (2021). Variability within rare cell states enables multiple paths toward drug resistance. *Nat. Biotechnol.* **39**, 865-876. doi:10.1038/s41587-021-00837-3

Fa, B., Wei, T., Zhou, Y., Johnston, L., Yuan, X., Ma, Y., Zhang, Y. and Yu, Z. (2021). GapClust is a light-weight approach distinguishing rare cells from voluminous single cell expression profiles. *Nat. Commun.* **12**, 4197. doi:10.1038/s41467-021-24489-8

Gao, T. and Qian, J. (2020). EnhancerAtlas 2.0: an updated resource with enhancer annotation in 586 tissue/cell types across nine species. *Nucleic Acids Res.* **48**, D58-D64. doi:10.1093/nar/gkz980

Griffiths, J. A., Scialdone, A. and Marioni, J. C. (2018). Using single-cell genomics to understand developmental processes and cell fate decisions. *Mol. Syst. Biol.* **14**, e8046. doi:10.15252/msb.20178046

Grün, D., Muraro, M. J., Boisset, J.-C., Wiebrands, K., Lyubimova, A., Dharmadhikari, G., van den Born, M., van Es, J., Jansen, E., Clevers, H. et al. (2016). De novo prediction of stem cell identity using single-cell transcriptome data. *Cell Stem Cell* **19**, 266-277. doi:10.1016/j.stem.2016.05.010

Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W. M., Zheng, S., Butler, A., Lee, M. J., Wilk, A. J., Darby, C., Zager, M. et al. (2021). Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573-3587.e29. doi:10.1016/j.cell.2021.04.048

Heit, C., Jackson, B. C., McAndrews, M., Wright, M. W., Thompson, D. C., Silverman, G. A., Nebert, D. W. and Vasiliou, V. (2013). Update of the human and mouse SERPIN gene superfamily. *Hum. Genomics* **7**, 22. doi:10.1186/1479-7364-7-22

Herman, J. S., Sagar, J. S. and Grün, D. (2018). FateID infers cell fate bias in multipotent progenitors from single-cell RNA-seq data. *Nat. Methods* **15**, 379-386. doi:10.1038/nmeth.4662

Iturbide, A., Ruiz Tejada Segura, M. L., Noll, C., Schorpp, K., Rothenaigner, I., Ruiz-Morales, E. R., Lubatti, G., Agami, A., Hadian, K., Scialdone, A. et al. (2021). Retinoic acid signaling is critical during the totipotency window in early mammalian development. *Nat. Struct. Mol. Biol.* **28**, 521-532. doi:10.1038/s41594-021-00590-w

Ivanovs, A., Rybtsov, S., Ng, E. S., Stanley, E. G., Elefanty, A. G. and Medvinsky, A. (2017). Human haematopoietic stem cell development: from the embryo to the dish. *Development* **144**, 2323-2337. doi:10.1242/dev.134866

Jiang, L., Chen, H., Pinello, L. and Yuan, G.-C. (2016). GiniClust: detecting rare cell types from single-cell gene expression data with gini index. *Genome Biol.* **17**, 144. doi:10.1186/s13059-016-1010-4

Jindal, A., Gupta, P., Jayadeva, P. and Sengupta, D. (2018). Discovery of rare cells from voluminous single cell expression data. *Nat. Commun.* **9**, 4719. doi:10.1038/s41467-018-07234-6

Luecken, M. D. and Theis, F. J. (2019). Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol. Syst. Biol.* **15**, e8746. doi:10.15252/msb.20188746

Lun, A. T. L., Participants in the 1st Human Cell Atlas Jamboree, Riesenfeld, S., Andrews, T., Dao, T. P., Gomes, T. and Marioni, J. C. (2019). EmptyDrops: distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data. *Genome Biol.* **20**, 63. doi:10.1186/s13059-019-1662-y

M Ascensión, A., Ibáñez-Solé, O., Inza, I., Izeta, A. and Araúzo-Bravo, M. J. (2022). Triku: a feature selection method based on nearest neighbors for single-cell data. *GigaScience* **11**, giac017. doi:10.1093/gigascience/giac017

Ma, S., Zhang, B., LaFave, L. M., Earl, A. S., Chiang, Z., Hu, Y., Ding, J., Brack, A., Kartha, V. K., Tay, T. et al. (2020). Chromatin potential identified by shared single-cell profiling of RNA and chromatin. *Cell* **183**, 1103-16.e20. doi:10.1016/j.cell.2020.09.056

Macfarlan, T. S., Gifford, W. D., Driscoll, S., Lettieri, K., Rowe, H. M., Bonanomi, D., Firth, A., Singer, O., Trono, D. and Pfaff, S. L. (2012). Embryonic stem cell potency fluctuates with endogenous retrovirus activity. *Nature* **487**, 57-63. doi:10.1038/nature11244

Mcinnes, L. and Healy, J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *ArXiv [Stat.ML]. arXiv*. http://arxiv.org/abs/1802.03426.

Melsted, P., Booeshaghi, A. S., Liu, L., Gao, F., Lu, L., Min, K. H., da Veiga Beltrame, E., Hjörleifsson, K. E., Gehring, J. and Pachter, L. (2021). Modular, efficient and constant-memory single-cell RNA-seq preprocessing. *Nat. Biotechnol.* **39**, 813-818. doi:10.1038/s41587-021-00870-2

Nowotschin, S., Setty, M., Kuo, Y.-Y., Liu, V., Garg, V., Sharma, R., Simon, C. S., Saiz, N., Gardner, R., Boutet, S. C. et al. (2019). The emergent landscape of the mouse gut endoderm at single-cell resolution. *Nature* **569**, 361-367. doi:10.1038/s41586-019-1127-1

Pijuan-Sala, B., Griffiths, J. A., Guibentif, C., Hiscock, T. W., Jawaid, W., Calero-Nieto, F. J., Mulas, C., Ibarra-Soria, X., Tyser, R. C. V., Ho, D. L. L. et al. (2019). A single-cell molecular map of mouse gastrulation and early organogenesis. *Nature* **566**, 490-495. doi:10.1038/s41586-019-0933-9

Raudvere, U., Kolberg, L., Kuzmin, I., Arak, T., Adler, P., Peterson, H. and Vilo, J. (2019). G:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 Update). *Nucleic Acids Res.* 47, W191-W198. doi:10.1093/nar/gkz369

Richard, C. and Verdier, F. (2020). Transferrin receptors in erythropoiesis. *Int. J. Mol. Sci.* **21**, 9713. doi:10.3390/ijms21249713

Rodriguez-Terrones, D., Gaume, X., Ishiuchi, T., Weiss, A., Kopp, A., Kruse, K., Penning, A., Vaquerizas, J. M., Brino, L. and Torres-Padilla, M.-E. (2018). A molecular roadmap for the emergence of early-embryonic-like cells in culture. *Nat. Genet.* **50**, 106-119. doi:10.1038/s41588-017-0016-5

Ross, C. and Boroviak, T. E. (2020). Origin and function of the yolk sac in primate embryogenesis. *Nat. Commun.* **11**, 3760. doi:10.1038/s41467-020-17575-w

Street, K., Risso, D., Fletcher, R. B., Das, D., Ngai, J., Yosef, N., Purdom, E. and Dudoit, S. (2018). Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics* **19**, 477. doi:10.1186/s12864-018-4772-0

Tarashansky, A. J., Xue, Y., Li, P., Quake, S. R. and Wang, B. (2019). Self-assembling manifolds in single-cell RNA sequencing data. *eLife* **8**, e48994. doi:10.7554/eLife.48994

Taubenschmid-Stowers, J., Rostovskaya, M., Santos, F., Ljung, S., Argelaguet, R., Krueger, F., Nichols, J. and Reik, W. (2022). 8C-like cells capture the human zygotic genome activation program in vitro. *Cell Stem Cell* **29**, 449-459.e6. doi:10.1016/j.stem.2022.01.014

Tsoucas, D. and Yuan, G.-C. (2018). GiniClust2: a cluster-aware, weighted ensemble clustering method for cell-type detection. *Genome Biol.* **19**, 58. doi:10.1186/s13059-018-1431-3

Tyser, R. C. V., Ibarra-Soria, X., McDole, K., Jayaram, S. A., Godwin, J., Van Den Brand, T. A. H., Miranda, A. M. A., Scialdone, A., Keller, P. J., Marioni, J. C., et al. (2021a). Characterization of a common progenitor pool of the epicardium and myocardium. *Science* **371**, eabb2986. doi:10.1126/science.abb2986

Tyser, R. C. V., Mahammadov, E., Nakanoh, S., Vallier, L., Scialdone, A. and Srinivas, S. (2021b). Single-cell transcriptomic characterization of a gastrulating human embryo. *Nature* **600**, 285. doi:10.1038/s41586-021-04158-y

Van Den Berge, K., Roux de Bézieux, H., Street, K., Saelens, W., Cannoodt, R., Saeys, Y., Dudoit, S. and Clement, L. (2020). Trajectory-based differential expression analysis for single-cell sequencing data. *Nat. Commun.* **11**, 1201. doi:10.1038/s41467-020-14766-3

Vandenbon, A. and Diez, D. (2020). A clustering-independent method for finding differentially expressed genes in single-cell transcriptome data. *Nat. Commun.* **11**, 4318. doi:10.1038/s41467-020-17900-3

Wagner, D. E., Weinreb, C., Collins, Z. M., Briggs, J. A., Megason, S. G. and Klein, A. M. (2018). Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. *Science* **360**, 981-987. doi:10.1126/science.aar4362

Wamaitha, S. E., del Valle, I., Cho, L. T. Y., Wei, Y., Fogarty, N. M. E., Blakeley, P., Sherwood, R. I., Ji, H. and Niakan, K. K. (2015). Gata6 potently initiates reprograming of pluripotent and differentiated cells to extraembryonic endoderm stem cells. *Genes Dev.* **29**, 1239-1255. doi:10.1101/gad.257071.114

Wegmann, R., Neri, M., Schuierer, S., Bilican, B., Hartkopf, H., Nigsch, F., Mapa, F., Waldt, A., Cuttat, R., Salick, M. R. et al. (2019). CellSIUS provides sensitive and specific detection of rare cell populations from complex single-cell RNA-seq data. *Genome Biol.* **20**, 142. doi:10.1186/s13059-019-1739-7

Wolf, F. A., Angerer, P. and Theis, F. J. (2018). SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15. doi:10.1186/s13059-017-1382-0

Wolf, F. A., Hamey, F. K., Plass, M., Solana, J., Dahlin, J. S., Göttgens, B., Rajewsky, N., Simon, L. and Theis, F. J. (2019). PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biol.* **20**, 59. doi:10.1186/s13059-019-1663-x

Zappia, L. and Oshlack, A. (2018). Clustering trees: a visualization for evaluating clusterings at multiple resolutions. *GigaScience* **7**, giy083. doi:10.1093/gigascience/giy083

Zheng, G. X. Y., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., Ziraldo, S. B., Wheeler, T. D., Mcdermott, G. P., Zhu, J. et al. (2017). Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049. doi:10.1038/ncomms14049

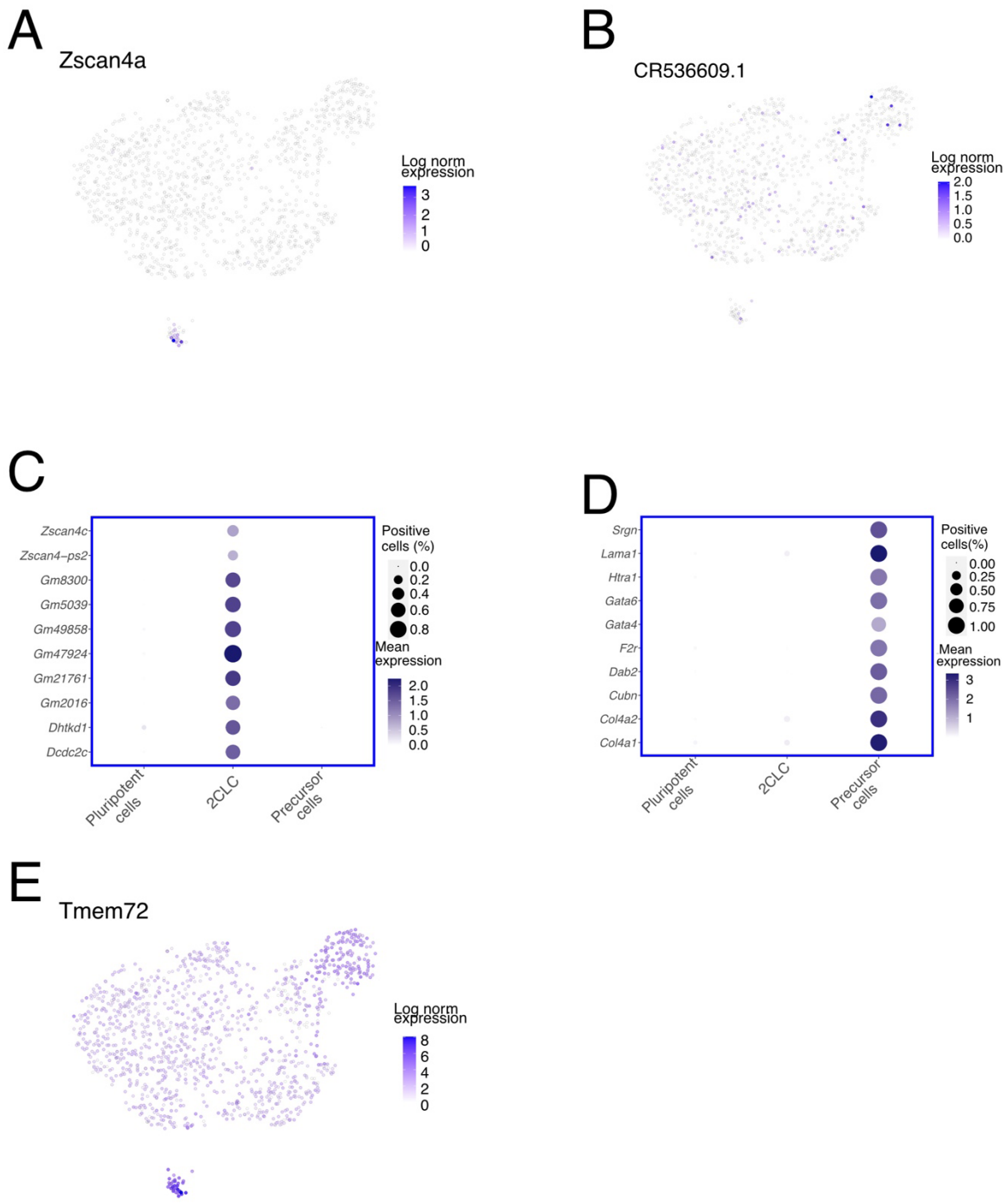


Fig. S1. Application of CIARA to a scRNA-seq data from mouse ESC, taken from (Iturbide et al. 2021). A, UMAP representation of the dataset (N=1285 cells) with the expression pattern of a highly localized gene found by CIARA, *Zscan4a.* **B**, Same UMAP representation as in (a) with the expression pattern of a gene not selected by CIARA, *CR536609.*1. Top marker genes of the 2CLC (panel **C**) and precursor cell (panel **D**) populations. The size of the dot is given by the fraction of cells with log norm counts above 1 (function NormalizeData from R library Seurat).The markers for the clusters are detected with the FindMarkers function (with parameter only.pos = T) from seurat (version 4.0.5).

Only markers with adjusted p-value (based on the Bonferroni correction) below or equal to 0.05 are considered for downstream analysis. Finally, for each cluster, only unique markers (i.e., that are not included among the markers of other clusters) are kept **E,** UMAP representation of the dataset with the expression pattern of a gene, *Tmem72*, identified by CIARA when using the option "extend_binarization=TRUE", designed to detect markers of rare cells expressed by multiple cell types.

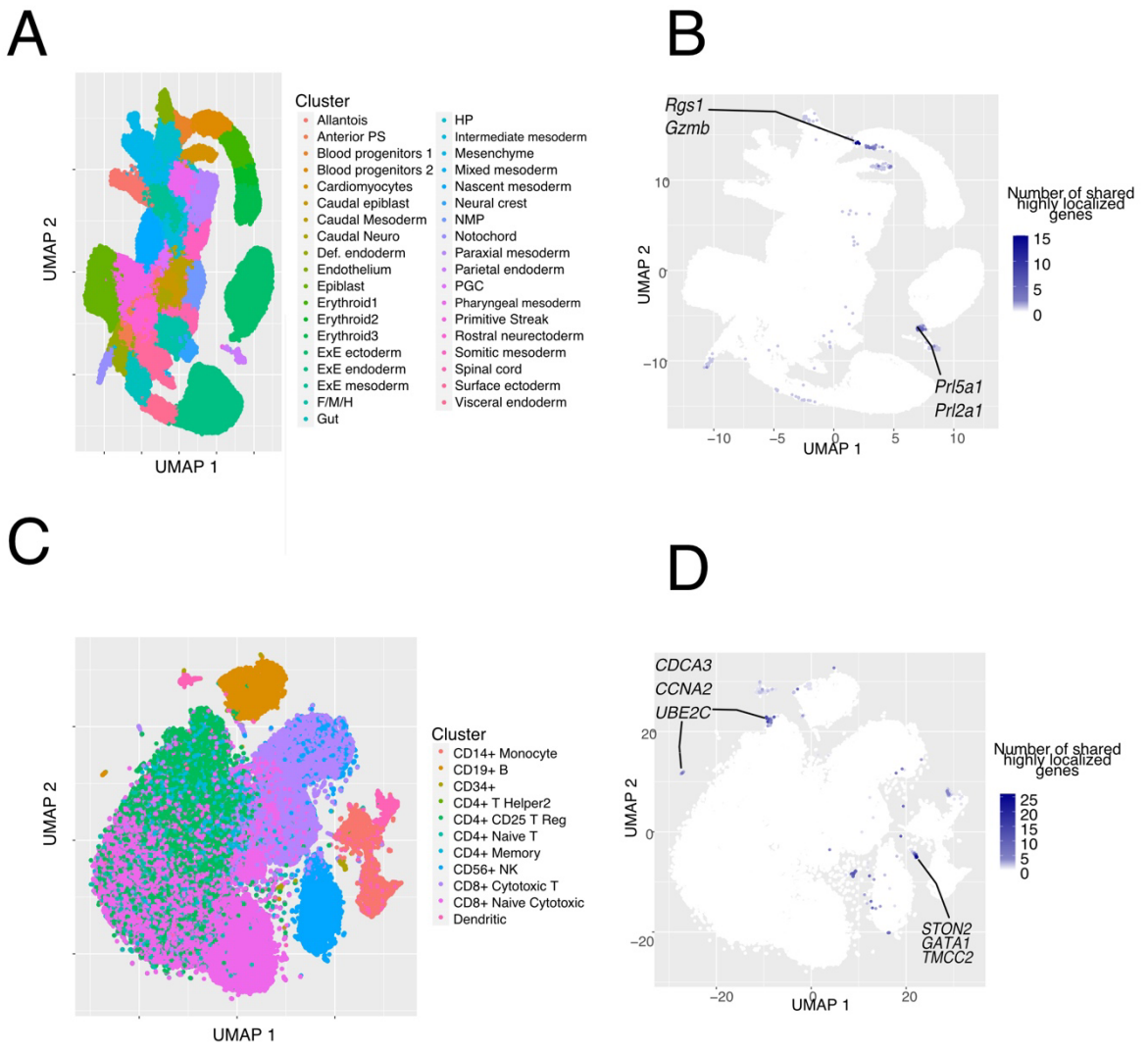


Fig. S2. **Application of CIARA to two single-cell transcriptomic atlases. A,** UMAP representation with cluster annotation of the mouse gastrulation atlas (Pijuan-Sala et al. 2019), which includes 116,312 cells. Abbreviated cluster names: “Anterior PS stands” for “Anterior Primitive Streak”; “Caudal neuro” stands for “Caudal neurectoderm”, “FMH” stands for “Forebrain/Midbrain/Hindbrain”, HP stands for “Haematoendothelial progenitors”. **B,** UMAP representation of the mouse gastrulation atlas, with colors indicating the number of highly localized genes expressed by each cell and shared with their neighbours. Cells with a large number of shared highly localized genes are more likely to represent rare populations of cells with a specific transcriptomic signature. In this dataset, the strongest signal comes from two sub-population of cells in the Blood Progenitor 1 cluster, and the Extra-Embryonic Ectoderm cluster. The first population of cells expresses markers like *Rgs1* and *Gzmb*, and the second *Prl5a1* and *Prl5a2* (see the full list of markers in Table S7). **C,** UMAP

representation with cluster annotation of a scRNA-seq dataset generated from human peripheral blood mononuclear cells (68,579 cells), taken from (Zheng et al. 2017). Abbreviated cluster names: “CD4+ Naive T” stands for “CD4+/CD45RA+/CD25- Naive T”; “CD4+ Memory” stands for “CD4+ CD45RO+ Memory”; “CD8+ Naive Cytotoxic” stands for “CD8+/CD45RA+ Naive Cytotoxic”. **D,** UMAP representation of the human blood cell datasets, with colors indicating the number of highly localized genes expressed by each cell and shared with their neighbours. When applied to the whole dataset, CIARA identifies a small cluster of megakaryocytes (expressing *STON2*, *GATA1,* and *TMCC2*, also highlighted in the original publication (Zheng et al. 2017)), in addition to other possibly rare cell types, for example, in the clusters of CD19+ B cells and of CD4+/CD25+ regulatory T cells, which express high levels of cell-cycle related genes (such as *CCNA2*, *CDCA3*, and *UBE2C*).

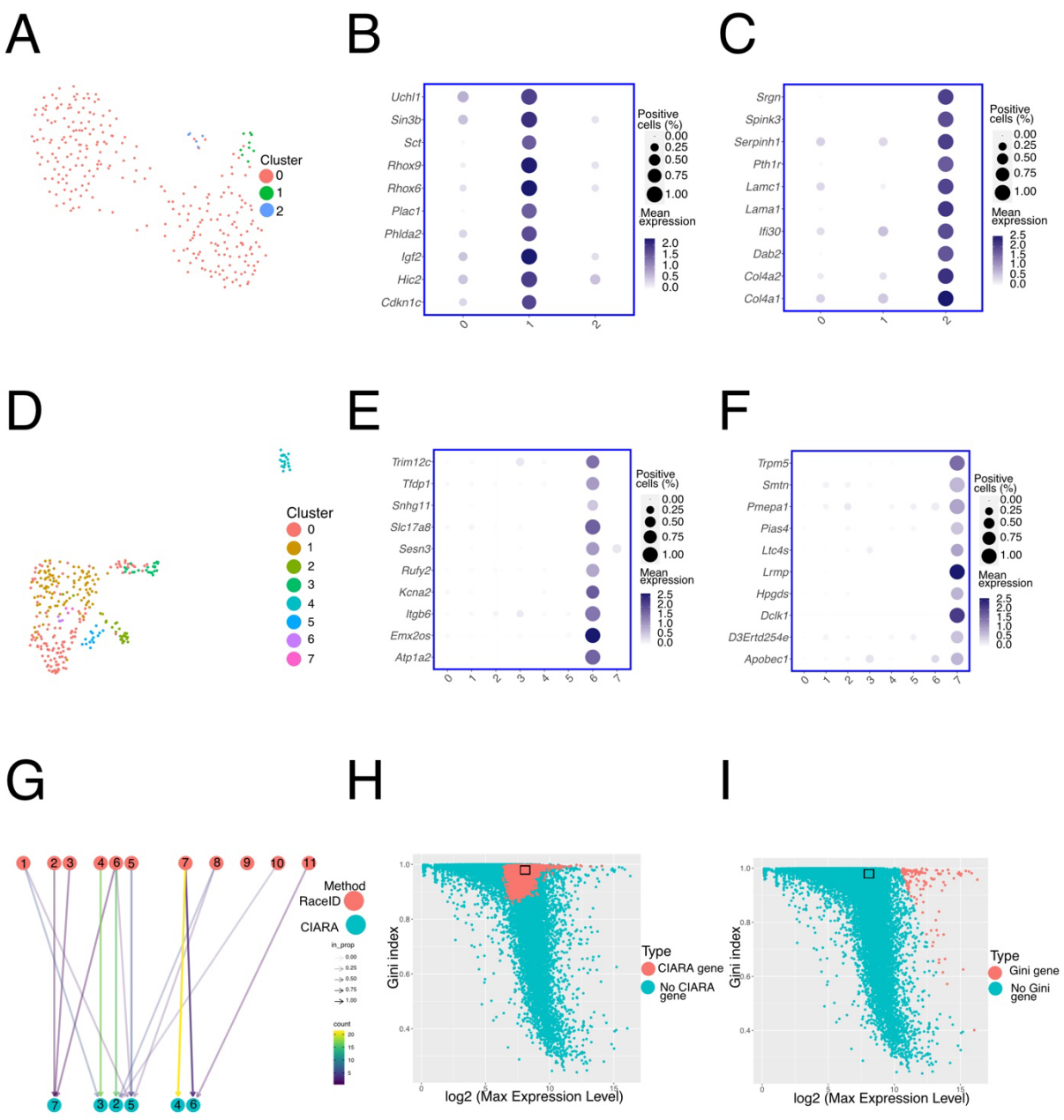


Fig. S3. Analyses of published datasets with CIARA and alternative algorithms. A, UMAP representation with the cluster partition found by CIARA in the mouse ESCs scRNA-seq dataset analyzed in the GiniClust2 paper (Tsoucas and Yuan 2018). The dataset includes 278 cells. **B**, **C**, Top marker genes of the two rare populations detected by CIARA in the dataset shown in (a). The size of the dot is given by the fraction of cells with log norm counts above 1 (function NormalizeData from R library Seurat). The markers for the clusters are detected with the FindMarkers function (with parameter only.pos = T) from seurat (version 4.0.5). Only markers with adjusted p-value (based on the Bonferroni correction) below or equal to 0.05 are considered for downstream analysis. Finally, for each cluster, only unique markers (ie, that are not included among the markers of other clusters) are kept D**,** UMAP representation of the murine intestinal epithelial cell dataset analyzed in RaceID vignette (https://cran.r-project.org/web/packages/RaceID/vignettes/RaceID.html), with the cluster partition found by CIARA. The dataset has 317 cells. **E, F**, Top marker genes of the two smallest populations (respectively 4 and 3 cells in clusters 6 and 7) detected by CIARA. The size of the dot is

given by the fraction of cells with log norm counts above 0.5 (see Method). In particular, cluster 7 expresses typical markers of Tuft cells (Herman, Sagar, and Grün 2018). The markers for the clusters are detected with the FindMarkers function (with parameter only.pos = T) from seurat (version 4.0.5). Only markers with adjusted p-value (based on the Bonferroni correction) below or equal to 0.05 are considered for downstream analysis. Finally, for each cluster, only unique markers (ie, that are not included among the markers of other clusters) are kept. **G,** clustree ploty to investigate the relationship between rare clusters found by CIARA (<= 25 cells, top circles) and the original clusters provided by RaceID (bottom circles; Cluster 2 corresponds to Goblet cells (*Clca3* as marker), cluster 3 to enterocytes (*Apoa1* as marker), cluster 4 to Paneth cells (*Defa24* as marker), cluster 5 to enteroendocrine cells (*Chgb* as marker), cluster 7 to Tuft cells). The clusters found with CIARA correspond to single cell types, while these are split into several clusters with RaceID. **H, I**, Scatterplots of the Gini index as function of the mean expression values for all the genes in the human gastrula dataset (Tyser, Mahammadov, et al. 2021). The red circles mark the genes selected by CIARA (panel h) or GiniClust2 (panel i). The black rectangle in the two panels indicate where some of the strongest markers of Primordial Germ Cells are (*NANOS3, NANOG, SOX17* and *DPPA5*).

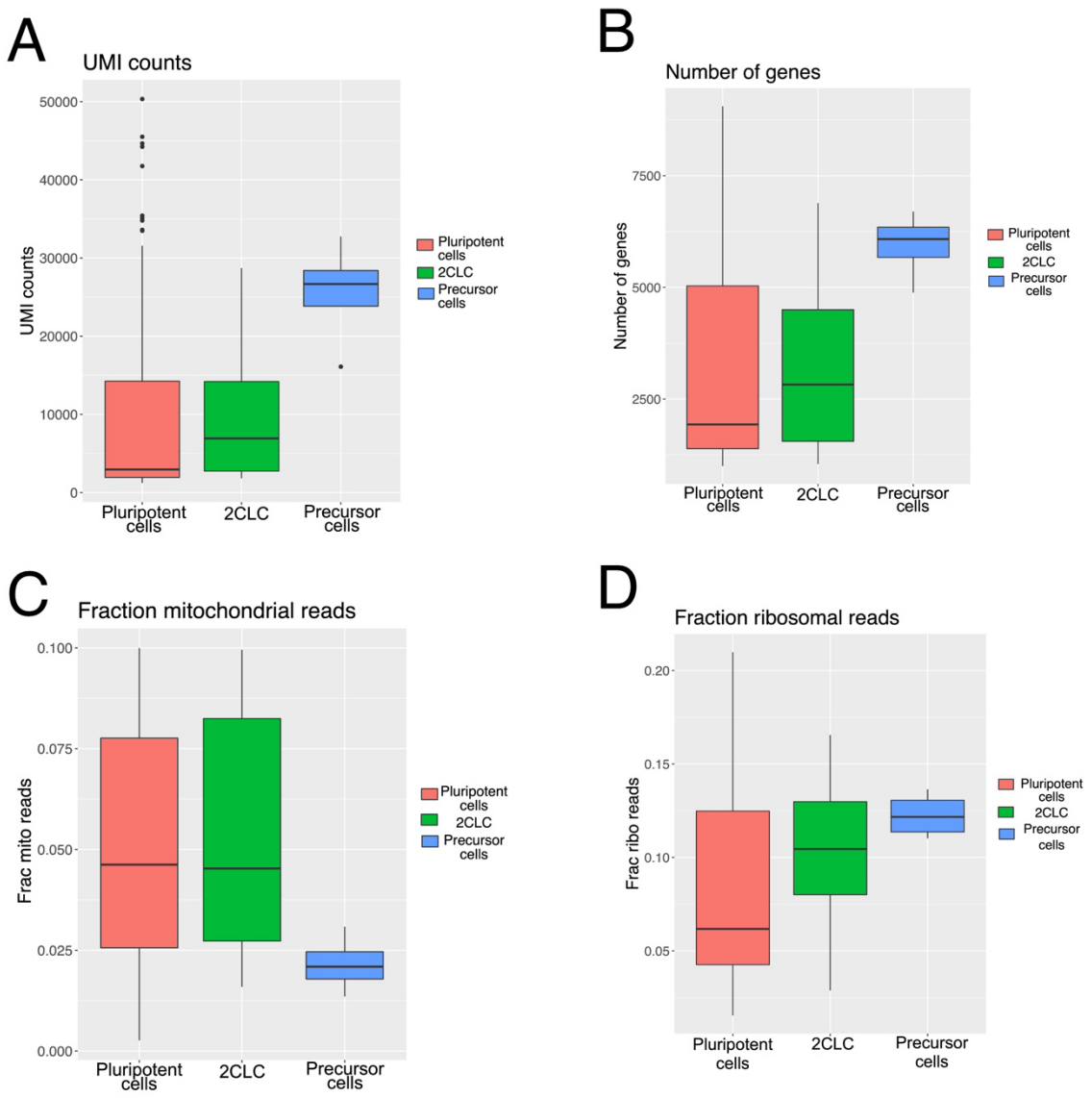


Fig. S4. Analysis of a scRNA-seq dataset from mouse ESCs treated with retinoic acid (RA) for 24h. The dataset includes 766 cells. Boxplot of UMI counts (**A**), number of expressed genes (**B**), fraction of mitochondrial (**C**) and ribosomal reads (**D**). All box plots show the lower quartile (Q1, 25th percentile), the median (Q2, 50th percentile) and the upper quartile (Q3, 75th percentile). Box length refers to interquartile range (IQR, Q3 − Q1). The upper whisker marks the minimum between the maximum value in the dataset and 1.5 times the IQR from Q3 (Q3 + 1.5 × IQR), while the lower whisker marks the maximum between the minimum value in the dataset and the IQR times 1.5 from Q1 (Q1 − 1.5 × IQR). Outliers are shown outside the interval defined by box and whiskers as individual points.

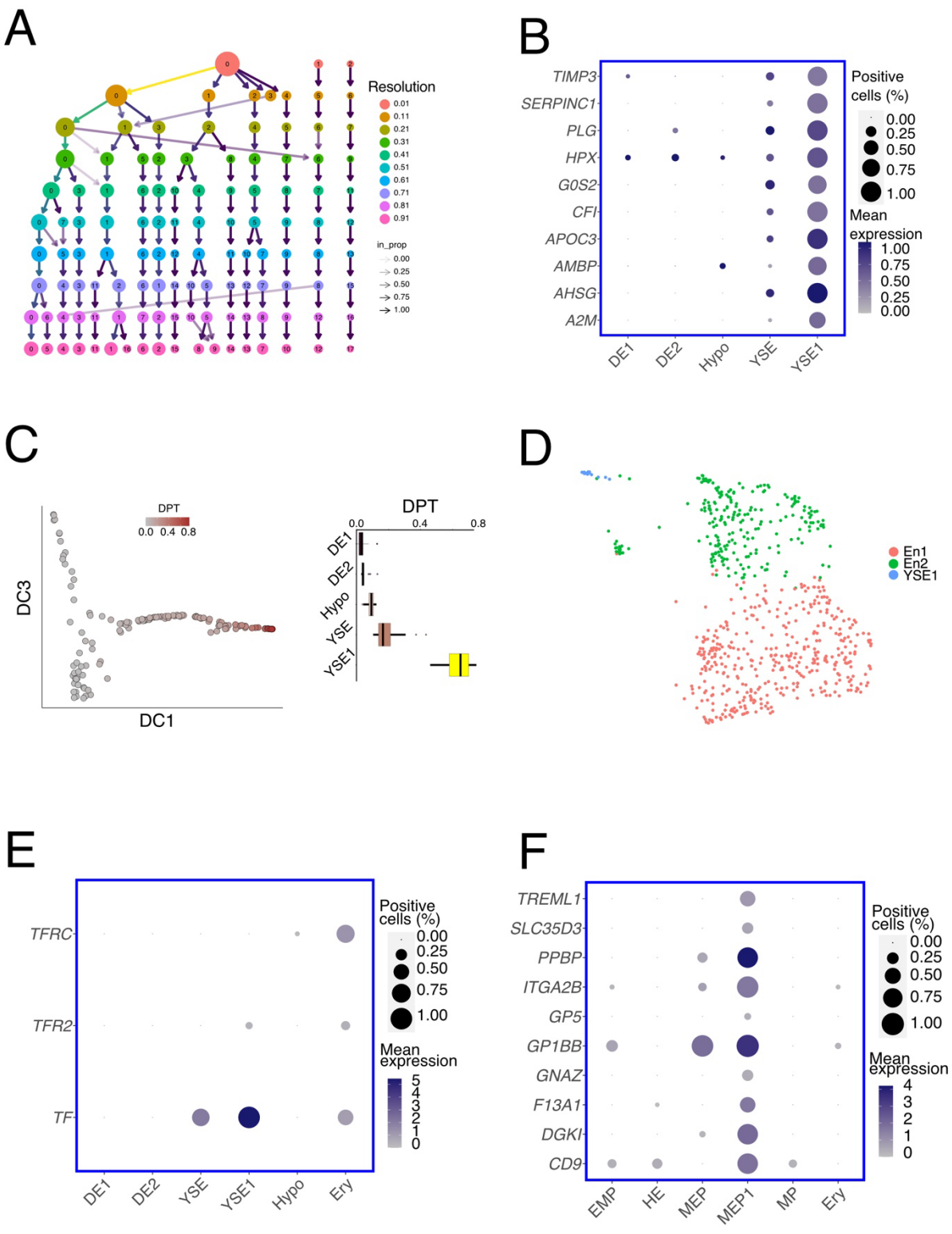
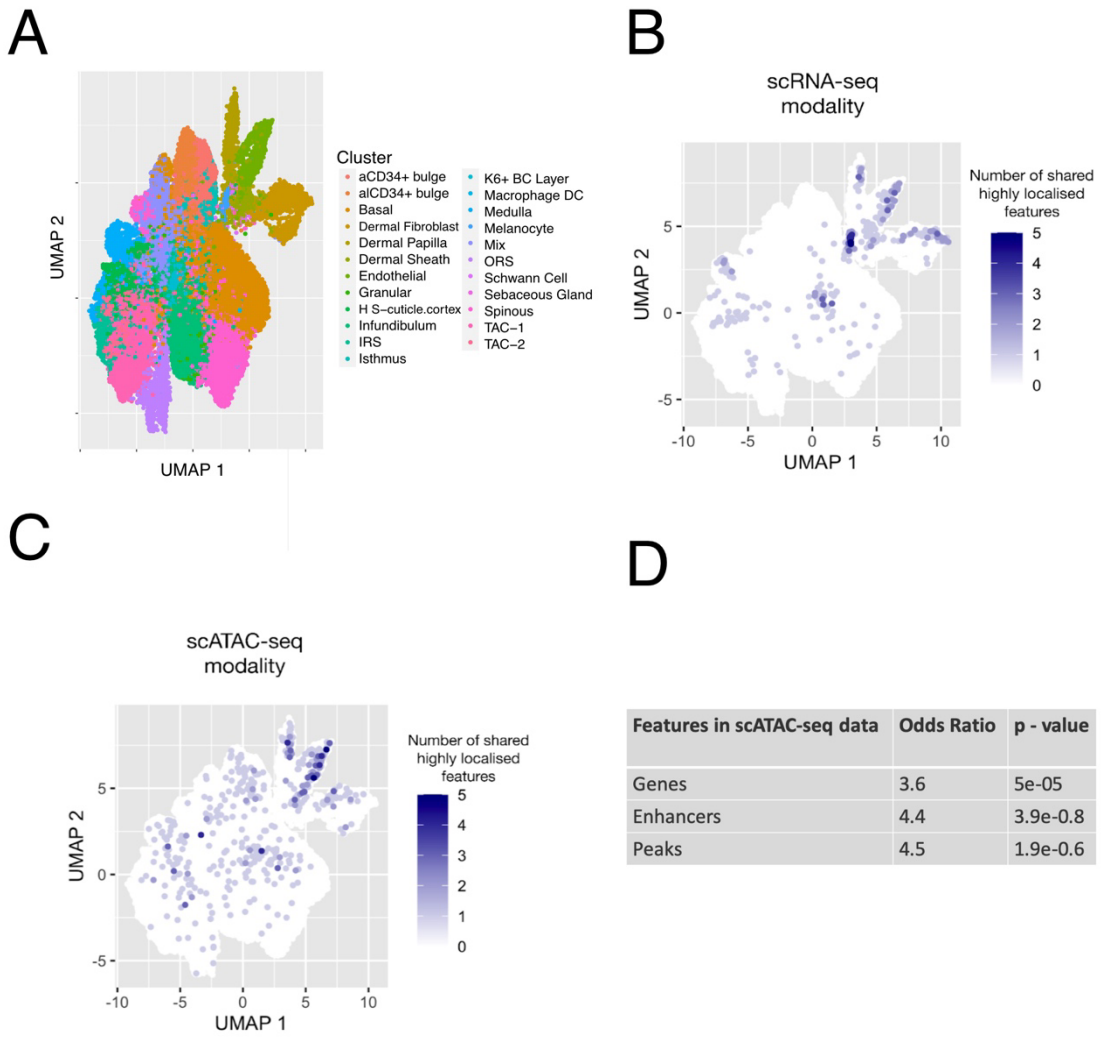


Fig. S5. Analysis of a human gastrula dataset (Tyser, Mahammadov, et al. 2021). A, clustree plot (Zappia and Oshlack 2018) showing the relationship of the clusters found with a Louvain algorithm for different values of resolution ranging between 0.01 up to 1. The genes used for clustering are those selected by CIARA. The primordial germ cells cluster (last column on the right) remains unaltered at all values of resolutions.

B, Extended list of top markers of the YSE1 cluster. Mean expression levels are normalized by the maximum within each cluster, and the size of the dot is given by the fraction of cells with log norm counts above 1 (function NormalizeData from R library Seurat). The markers for the clusters are detected with the FindMarkers function (with parameter only.pos = T) from seurat (version 4.0.5).

Only markers with adjusted p-value (based on the Bonferroni correction) below or equal to 0.05 are considered for downstream analysis. Finally, for each cluster, only unique markers (ie, that are not included among the markers of other clusters) are kept. **C**, The left panel shows the diffusion components 1 and 3 (DC1 and DC3) of endodermal cells (135 cells). Cells are colored based on the corresponding value of the diffusion pseudo-time (DPT). The DPT values of cells in each cluster are shown as boxplots in the right panel.
D, UMAP representation of the mouse endoderm dataset (Tyser, Ibarra-Soria, et al. 2021) with the cluster partition found by CIARA. The dataset has 665 cells. **E**, Balloon plot of transferrin (*TF*) and its two receptors *TFRC* and *TFR2* among endoderm and erythroblasts. The size of the dot is given by the fraction of cells with log norm counts above 1 (see Methods). The markers for the clusters are detected with the FindMarkers function (with parameter only.pos = T) from seurat (version 4.0.5). Only markers with adjusted p-value (based on the Bonferroni correction) below or equal to 0.05 are considered for downstream analysis. Finally, for each cluster, only unique markers (ie, that are not included among the markers of other clusters) are kept. **F**, Extended list of the top markers of the MEP1. The size of the dot is given by the fraction of cells with log norm counts above 1 (function NormalizeData from R library Seurat). The markers for the clusters are identified as specified above.



Features in scATAC-seq data	Odds Ratio	p - value
Genes	3.6	5e-05
Enhancers	4.4	3.9e-0.8
Peaks	4.5	1.9e-0.6

Fig. S6. CIARA identifies potential rare cell types across different modalities in a single-cell multiomic dataset. A, UMAP representation obtained from the RNA-seq modality of the SHARE-seq dataset generated from N=34,774 mouse skin cells (Ma et al. 2020). Cells are colored according to the clusters they belong to, as defined in (Ma et al. 2020). Abbreviated cluster names: “aCD34+ bulge” stands for “ahighCD34+ bulge”; **“**alCD34+ bulge” stands for “alowCD34+ bulge”; “HS−cuticle.cortex” stands for “Hair Shaft −cuticle.cortex”; “K6+ BC Layer” stands for “K6+ Bulge Companion Layer”. **B,** Same UMAP representation as in (a), with colors representing the number of shared highly localized features estimated by CIARA from the RNA-seq modality. **C,** Same UMAP representation as in (a), with colors representing the number of shared highly localized features estimated by CIARA from the ATAC-seq modality, using genes as features (see Methods). **D,** Table summarizing the results of the Fisher’s exact tests used to verify the statistical significance of the overlap between the potential rare cells identified in the RNA-seq and the ATAC-seq modality. The Fisher’s test is done with the function "fisher.test" in R, with the option "alternative=greater".

Fig. S7. Interactive plot for mouse ESCs treated with RA at 0h

Click here to download Fig. S7

Fig. S8. Interactive plot for mouse ESCs treated with RA at 24h

Click here to download Fig. S8

Fig. S9. Interactive plot for human gastrula dataset

Click here to download Fig. S9

Table S1. Markers of precursor cells found in mESC treated with RA for 24h

Click here to download Table S1

Table S2. Markers of 2CLC found in mESC treated with RA for 24h

Click here to download Table S2

Table S3. Markers of pluripotent cells found in mESC treated with RA for 24h

Click here to download Table S3

Table S4. Markers of YSE1 cluster in the human gastrula

Click here to download Table S4

Table S5. Markers of MEP1 cluster in the human gastrula

Click here to download Table S5

Table S6. Markers of sub-cluster of mouse endodermal cells equivalent to YSE1

Click here to download Table S6

Table S7. Markers localized in the rare cells in the blood progenitors 1 and extra-embryonic ectoderm clusters in the mouse gastrula atlas dataset

Click here to download Table S7

References

Herman, Josip S., Sagar, and Dominic Grün. 2018. "FateID Infers Cell Fate Bias in Multipotent Progenitors from Single-Cell RNA-Seq Data." *Nature Methods* 15 (5): 379–86.

Iturbide, Ane, Mayra L. Ruiz Tejada Segura, Camille Noll, Kenji Schorpp, Ina Rothenaigner, Elias R. Ruiz-Morales, Gabriele Lubatti, et al. 2021. "Retinoic Acid Signaling Is Critical during the Totipotency Window in Early Mammalian Development." *Nature Structural & Molecular Biology* 28 (6): 521–32.

Ma, Sai, Bing Zhang, Lindsay M. LaFave, Andrew S. Earl, Zachary Chiang, Yan Hu, Jiarui Ding, et al. 2020. "Chromatin Potential Identified by Shared Single-Cell Profiling of RNA and Chromatin." *Cell* 183 (4): 1103–16.e20.

Pijuan-Sala, Blanca, Jonathan A. Griffiths, Carolina Guibentif, Tom W. Hiscock, Wajid Jawaid, Fernando J. Calero-Nieto, Carla Mulas, et al. 2019. "A Single-Cell Molecular Map of Mouse Gastrulation and Early Organogenesis." *Nature* 566 (7745): 490–95.

Tsoucas, Daphne, and Guo-Cheng Yuan. 2018. "GiniClust2: A Cluster-Aware, Weighted Ensemble Clustering Method for Cell-Type Detection." *Genome Biology* 19: 58.

Tyser, Richard C. V., Ximena Ibarra-Soria, Katie McDole, Satish Arcot Jayaram, Jonathan Godwin, Teun A. H. van den Brand, Antonio M. A. Miranda, et al. 2021. "Characterization of a Common Progenitor Pool of the Epicardium and Myocardium." *Science* 371 (6533). https://doi.org/10.1126/science.abb2986.

Tyser, Richard C. V., Elmir Mahammadov, Shota Nakanoh, Ludovic Vallier, Antonio Scialdone, and Shankar Srinivas. 2021. "Single-Cell Transcriptomic Characterization of a Gastrulating Human Embryo." *Nature* 600 (November): 285.

Zappia, Luke, and Alicia Oshlack. 2018. "Clustering Trees: A Visualization for Evaluating Clusterings at Multiple Resolutions." *GigaScience* 7 (7). https://doi.org/10.1093/gigascience/giy083.

Zheng, Grace X. Y., Jessica M. Terry, Phillip Belgrader, Paul Ryvkin, Zachary W. Bent, Ryan Wilson, Solongo B. Ziraldo, et al. 2017. "Massively Parallel Digital Transcriptional Profiling of Single Cells." *Nature Communications* 8 (January): 14049.

5.4 Appendix D: Retinoic acid signaling is critical during the totipotency window in early mammalian development

Further manuscript as co-author

- Ane Iturbide, Mayra L Ruiz Tejada Segura, Camille Noll, Kenji Schorpp, Ina Rothenaigner, Elias R Ruiz-Morales, **Gabriele Lubatti**, Ahmed Agami, Kamyar Hadian, Antonio Scialdone, Maria-Elena Torres-Padilla. Retinoic acid signaling is critical during the totipotency window in early mammalian development. Nat Struct Mol Biol 28, 521–532 (2021). `https://doi.org/10.1038/s41594-021-00590-w`.

Contributions of the author: I led the trajectory analysis shown in the paper in figure 4 panels c, d, e, f and g.

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 '2C' 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;

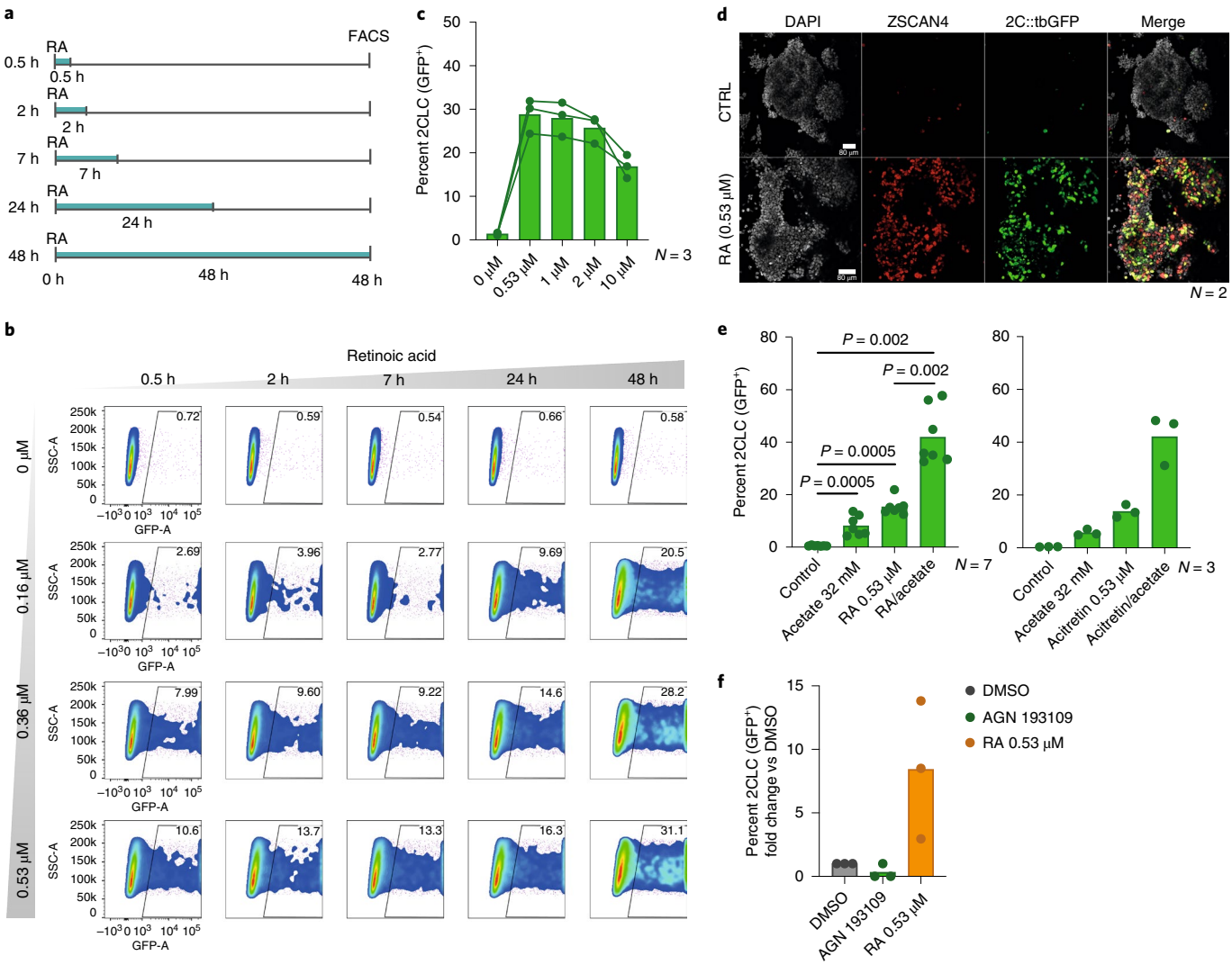


Fig. 1 | Low concentrations of RA robustly induce 2CLCs. a, Experimental design. Embryonic stem (ES) cells were treated with a range of RA concentrations for different time periods. 2CLC induction was measured by FACS, 48 h after treatment. **b**, Representative scatter plot for the experiment in **a**, 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

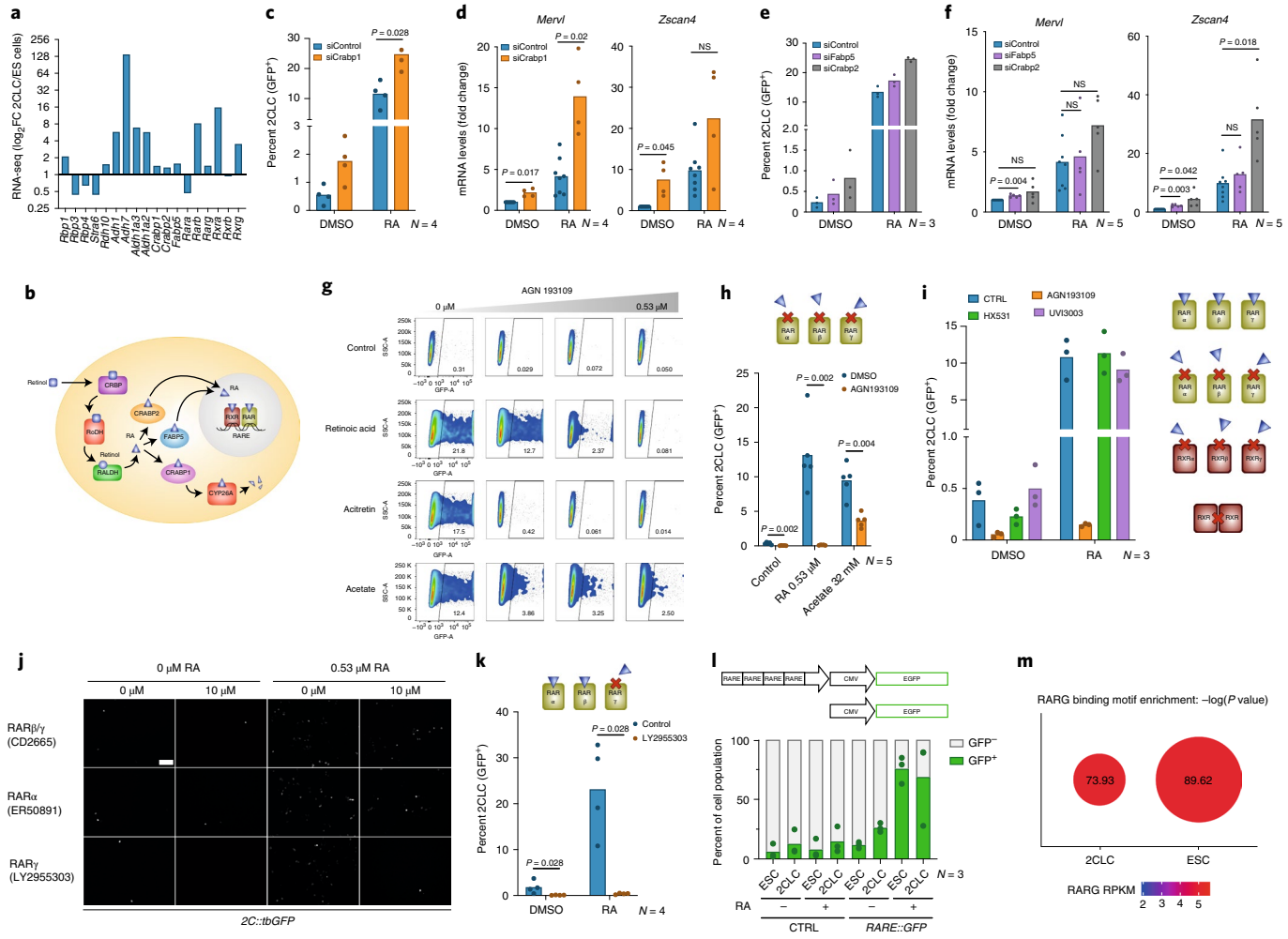


Fig. 2 | RARγ is required for 2CLC emergence. a, Expression levels (log_2FC) (FC, fold change) of selected RA-pathway-related genes in 2CLCs and ES cells (ESCs) based on RNA-seq data ($N = 2$, from ref. [16]). **b**, Schematic of the RA pathway. **c**, Induction of 2CLCs upon siRNA for *Crabp1* and RA 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* 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 $P < 2.2 \times 10^{-16}$). 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).

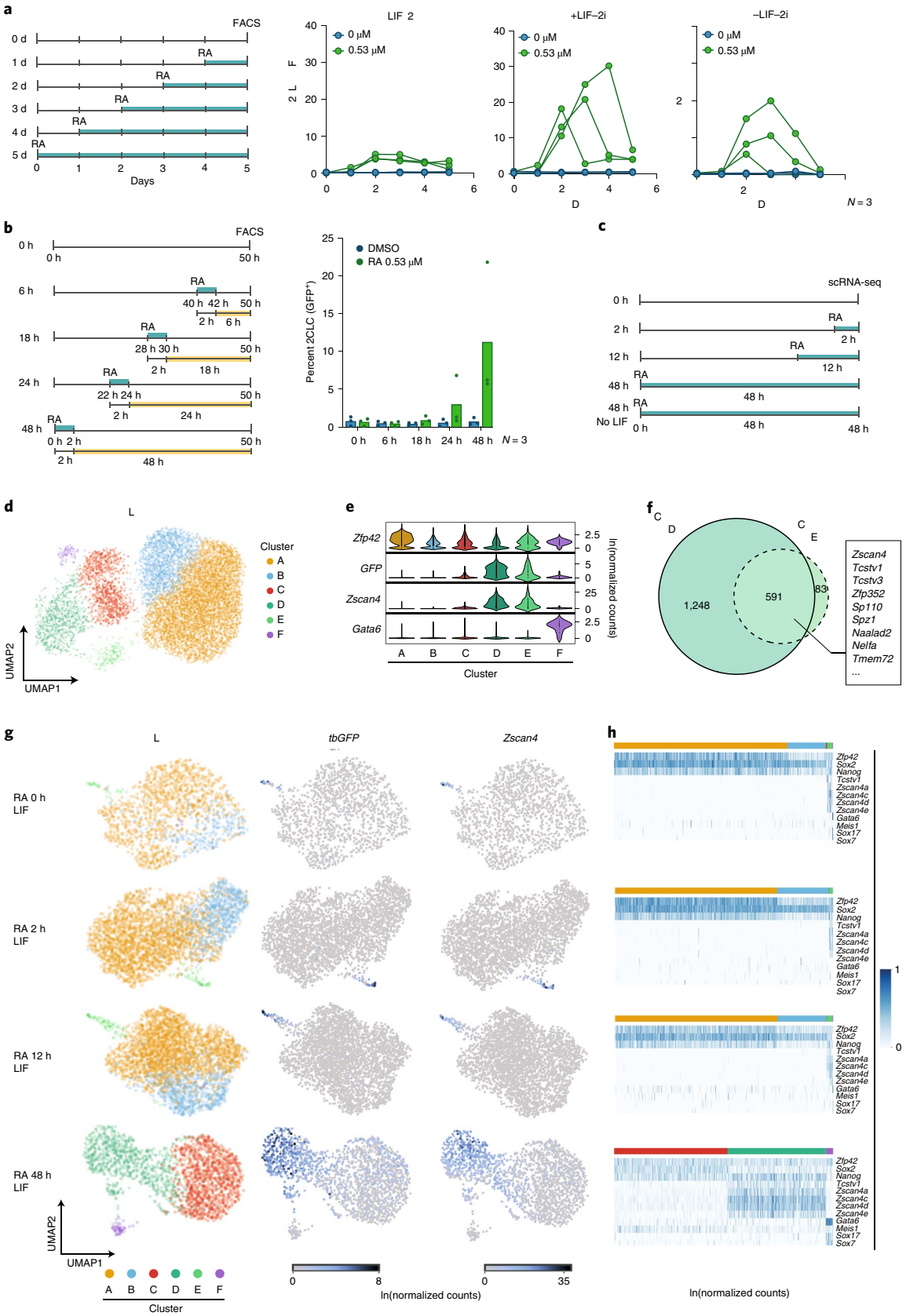
In addition to the 2CLC clusters, the two clusters comprising pluripotent ES cells exhibiting high and medium levels of *Rex1* and *Nanog* (clusters A and B) were consistently present across early timepoints (0, 2 and 12 h) and represented the majority of the cells at these timepoints (Fig. 3g). Specifically, 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.



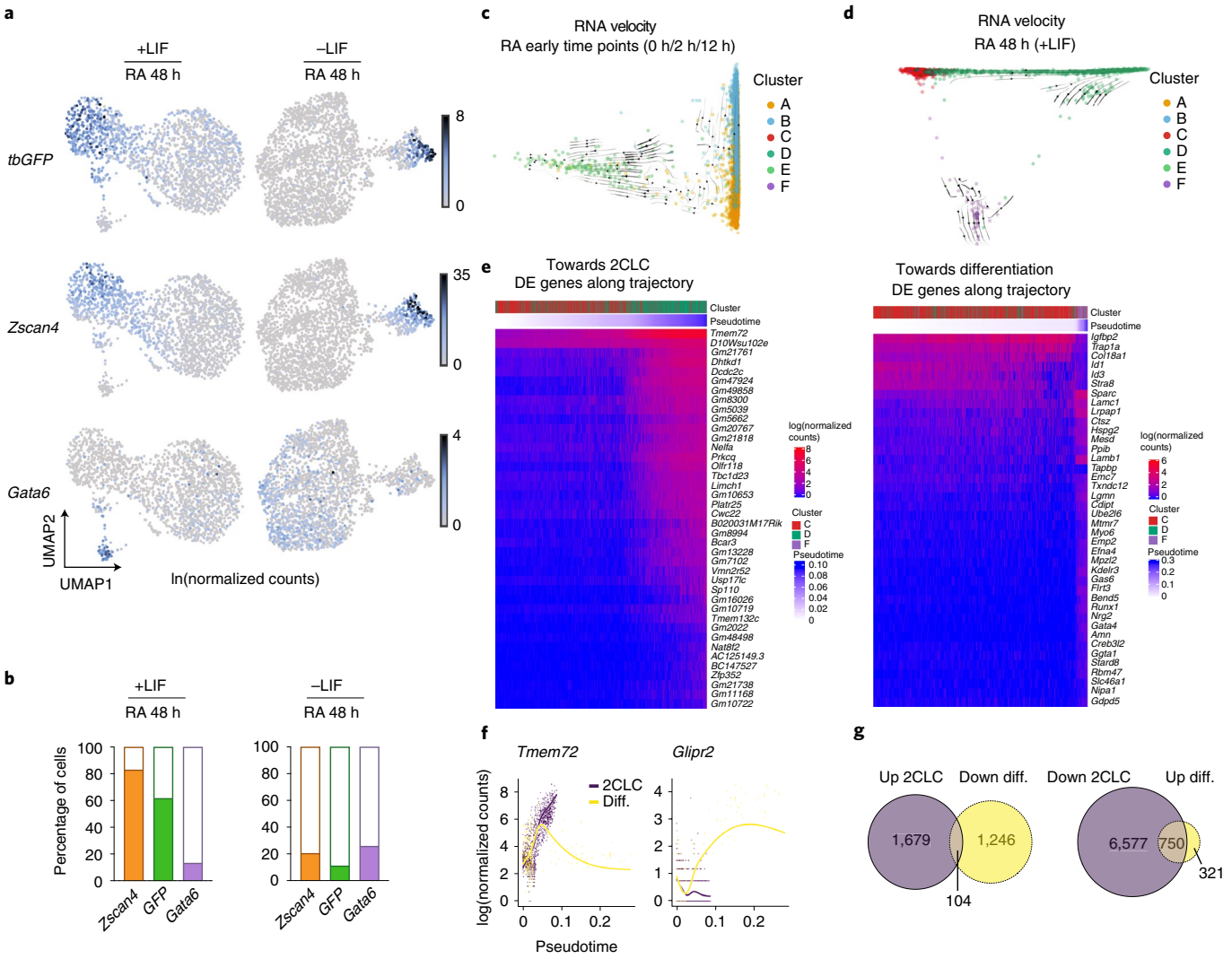


Fig. 4 | RA-reprogrammed 2CLCs differ from differentiating cells. a, UMAP plots of cells treated with RA for 48 h with LIF (left column) or without LIF (right column). Rows from top to bottom are colored by expression of *tbGFP* (MERVL), *Zscan4* (marking 2CLCs) and *Gata6* (marking differentiating cells). **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.

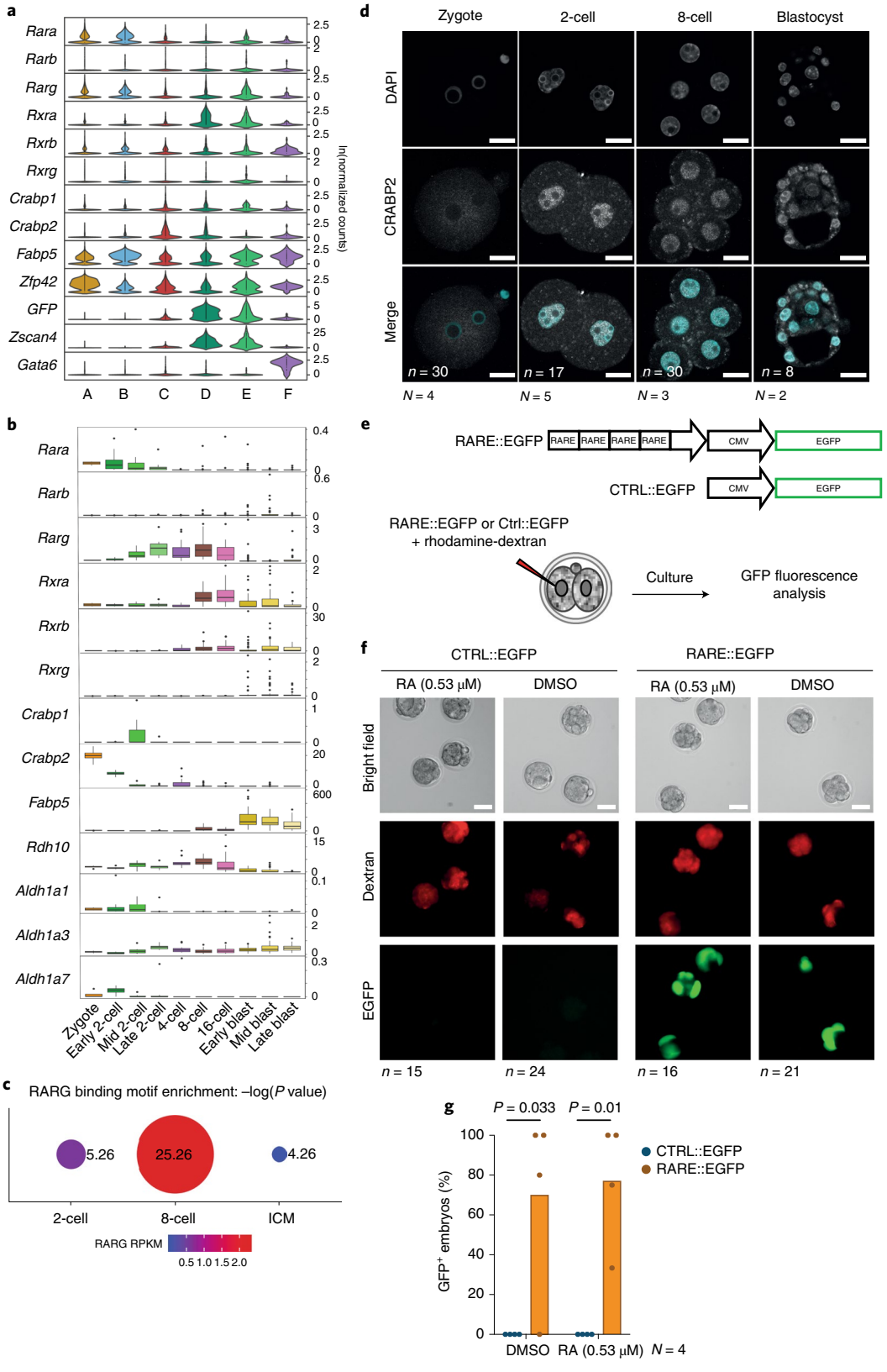
Blocking ZGA with a general RNA PolII inhibitor results in most embryos arresting at the 2-cell stage[39], similarly to the phenotype observed upon LY2955303 treatment. Thus, we next addressed if inhibiting RARγ affects ZGA by analyzing MERVL expression—a key ZGA marker—in embryos treated with LY2955303. 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 ($\log_2FC>1$ and $\log_2FC<-1$, respectively; $P_{adj}<0.05$) (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 ($P=2.2\times10^{-16}$, Fisher test) and over-represented in the downregulated genes ($P=2.723\times10^{-11}$, 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, $-\log_{10}(P\ value)$. **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



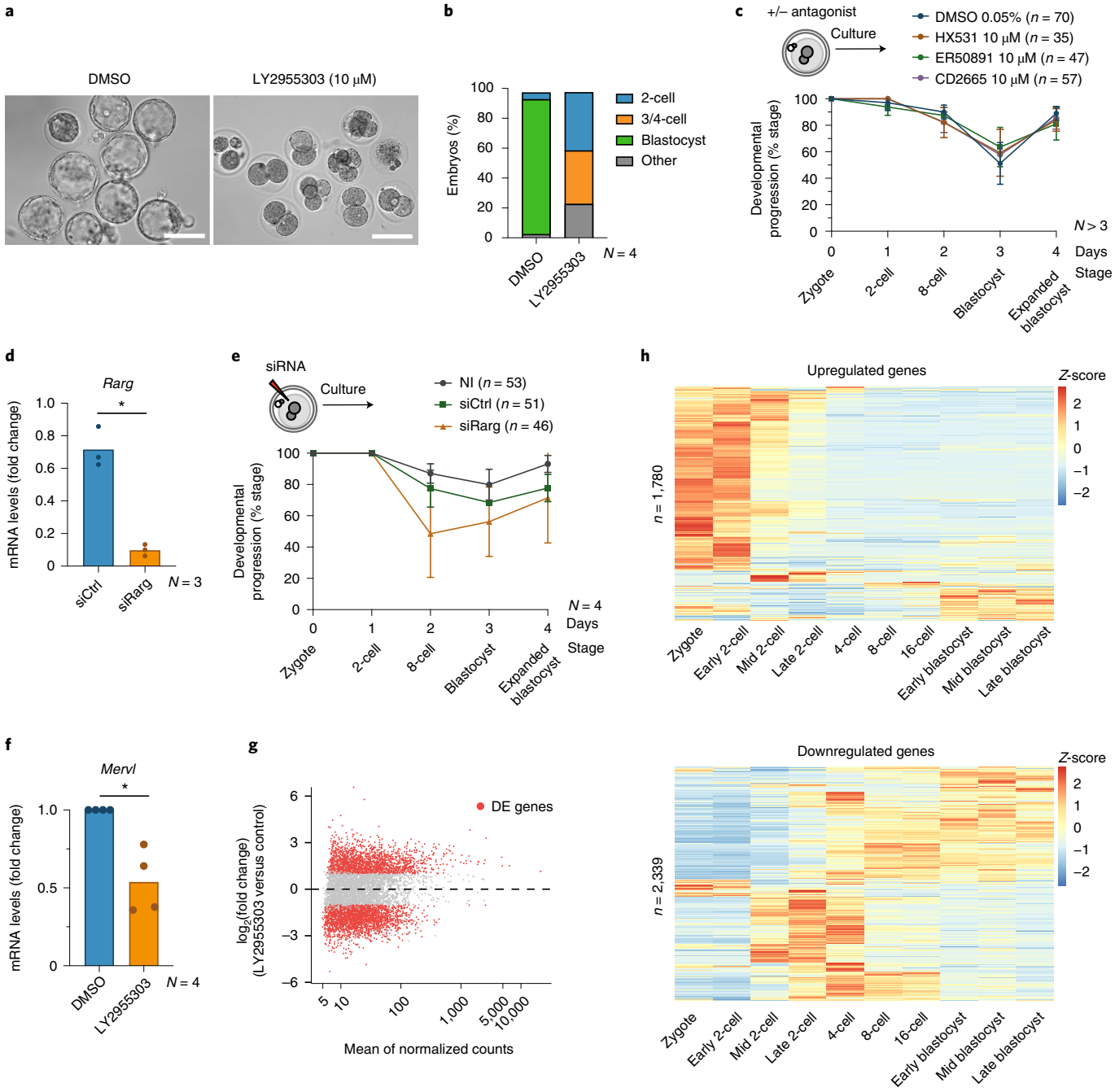


Fig. 6 | Perturbing RA signaling in the early mouse embryo affects developmental progression. a, Phase-contrast images of representative embryos treated with the RARγ antagonist LY2955303 or control DMSO. $N = 4$. Scale bars, 100 μm. **b**, Developmental progression (in percentage) of control (DMSO, $n = 51$) 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).

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, $\log_2$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 ≥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 MervI 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.

5.5 Appendix E: Letter of approval from publisher

5.5.1 Appendix A: Cell competition acts as a purifying selection to eliminate cells with mitochondrial defects during early mouse development

https://www.nature.com/natmetab/editorial-policies/self-archiving-and-license-to-publish

Self archiving and license to publish

(...)

Creative commons licences

Nature Portfolio open access and hybrid journals

Open access articles in Nature Portfolio Journals are published under a CC BY license (Creative Commons Attribution 4.0 International License). Under Creative Commons, authors retain copyright in their articles. The CC BY license is the most open licence available and considered the industry 'gold standard' for open access. It allows for maximum dissemination and re-use of open access materials and is preferred by many research funding bodies. Under this license, users are free to share (copy, distribute and transmit) and remix (adapt) the contribution including for commercial purposes, providing they attribute the contribution in the manner specified by the author or licensor (read full legal code). All Springer Nature journals with an open access option offer intergovernmental organisation (IGO) versions of Creative Commons licences on request, where required by the author's employer.

Authors are advised to check their funder's open access requirements, to ensure compliance. For more information about open access licensing, please see "OA licensing and copyright" on our journal open access policies page on SpringerNature.com.

The Nature Portfolio hybrid journals are Transformative Journals and offer a gold open access option. Please see our external announcement here.

5.5.2 Appendix B: MitoHEAR: an R package for the estimation and downstream statistical analysis of the mitochondrial DNA heteroplasmy calculated from single-cell datasets

https://joss.theoj.org/about#content_license

Content Licensing & Open Access

JOSS is a diamond/platinum open access journal. Copyright of JOSS papers is retained by submitting authors and accepted papers are subject to a Creative Commons Attribution 4.0 International License.

Any code snippets included in JOSS papers are subject to the MIT license regardless of the license of the submitted software package under review.

Any use of the JOSS logo is licensed CC BY 4.0. See the joss/logo directory in the digital-assetsrepository for more information about it.

5.5.3 Appendix C: CIARA: a cluster-independent algorithm for the identification of markers of rare cell types from single-cell sequencing data

RE: DEV201264 use the publication for further qualifications

devprodeds <devprodeds@biologists.com>
Tue 5/30/2023 1:25 PM

To: Gabriele Lubatti <Gabriele.Lubatti@helmholtz-munich.de>

Cc: Antonio Scialdone, Dr. <Antonio.Scialdone@helmholtz-munich.de>

Dear Gabriele
I can confirm that we allow the use of your paper in your PhD thesis. For more details see the self-archiving section on our webpage relating to re-use.
https://journals.biologists.com/dev/pages/rights-permissions
We ask that you provide a link to the article online
Please get in touch if you need any further information.
Best wishes
Lindsay
(Production Editor, Development)

From: Gabriele Lubatti <Gabriele.Lubatti@helmholtz-munich.de>
Sent: 30 May 2023 11:12
To: devprodeds <devprodeds@biologists.com>
Cc: Antonio Scialdone, Dr. <Antonio.Scialdone@helmholtz-munich.de>
Subject: DEV201264 use the publication for further qualifications

Good Morning,
I am writing regarding the paper DEV201264 that has been recently accepted in Development (https://journals.biologists.com/dev/article-abstract/doi/10.1242/dev.201264/310178/CIARA-a-cluster-independent-algorithm-for?redirectedFrom=fulltext)
For my PhD thesis I would need the following document related to the paper:

Include in your dissertation the part of the email that confirms that you are allowed to use the publication for further qualifications. As this happens all the time, publishers include this sentence somewhere.

Could you provide me this information?
Thanks a lot and best wishes,

Bibliography

[1] C. Mulas, A. Chaigne, A. Smith, and K. J. Chalut, "Cell state transitions: definitions and challenges," *Development*, vol. 148, no. 20, p. dev199950, 2021.

[2] A. Hima Bindu and B. Srilatha, "Potency of various types of stem cells and their transplantation," *J Stem Cell Res Ther*, vol. 1, no. 3, p. 115, 2011.

[3] C. Trapnell, "Defining cell types and states with single-cell genomics," *Genome research*, vol. 25, no. 10, pp. 1491–1498, 2015.

[4] J. A. Griffiths, A. Scialdone, and J. C. Marioni, "Using single-cell genomics to understand developmental processes and cell fate decisions," *Molecular systems biology*, vol. 14, no. 4, p. e8046, 2018.

[5] A. K. Tarkowski, "Experiments on the development of isolated blastomeres of mouse eggs," *Nature*, vol. 184, no. 4695, pp. 1286–1287, 1959.

[6] F. Lu and Y. Zhang, "Cell totipotency: molecular features, induction, and maintenance," *National science review*, vol. 2, no. 2, pp. 217–225, 2015.

[7] M. J. Evans and M. H. Kaufman, "Establishment in culture of pluripotential cells from mouse embryos," *nature*, vol. 292, no. 5819, pp. 154–156, 1981.

[8] I. Wilmut, A. E. Schnieke, J. McWhir, A. J. Kind, and K. H. Campbell, "Viable offspring derived from fetal and adult mammalian cells," *Nature*, vol. 385, no. 6619, pp. 810–813, 1997.

[9] A. Scialdone and N. Rivron, "In preprints: improving and interrogating embryo models," *Development*, vol. 149, no. 23, p. dev201404, 2022.

[10] C. Xiao, M. Grzonka, C. Meyer-Gerards, M. Mack, R. Figge, and H. Bazzi, "Gradual centriole maturation associates with the mitotic surveillance pathway in mouse development," *EMBO reports*, vol. 22, no. 2, p. e51127, 2021.

[11] S. Bowling, "Di, gregorio, a., sancho, m., pozzi, s., aarts, m., signore, m., schneider, md, martinez-barbera, jp, gil, j., and rodríguez, ta, p53 and mtor signalling determine fitness selection through cell competition during early mouse embryonic development," *Nat. Commun*, vol. 9, p. 1763, 2018.

[12] S. Bowling, K. Lawlor, and T. A. Rodríguez, "Cell competition: the winners and losers of fitness selection," *Development*, vol. 146, no. 13, p. dev167486, 2019.

[13] M. Sancho, A. Di-Gregorio, N. George, S. Pozzi, J. M. Sánchez, B. Pernaute, and T. A. Rodríguez, "Competitive interactions eliminate unfit embryonic stem cells at the onset of differentiation," *Developmental cell*, vol. 26, no. 1, pp. 19–30, 2013.

[14] C. Clavería, G. Giovinazzo, R. Sierra, and M. Torres, "Myc-driven endogenous cell competition in the early mammalian embryo," *Nature*, vol. 500, no. 7460, pp. 39–44, 2013.

[15] T. S. Macfarlan, W. D. Gifford, S. Driscoll, K. Lettieri, H. M. Rowe, D. Bonanomi, A. Firth, O. Singer, D. Trono, and S. L. Pfaff, "Embryonic stem cell potency fluctuates with endogenous retrovirus activity," *Nature*, vol. 487, no. 7405, pp. 57–63, 2012.

[16] J. Taubenschmid-Stowers, M. Rostovskaya, F. Santos, S. Ljung, R. Argelaguet, F. Krueger, J. Nichols, and W. Reik, "8c-like cells capture the human zygotic genome activation program in vitro," *Cell Stem Cell*, vol. 29, no. 3, pp. 449–459, 2022.

[17] S. Takahashi, S. Kobayashi, and I. Hiratani, "Epigenetic differences between naïve and primed pluripotent stem cells," *Cellular and Molecular Life Sciences*, vol. 75, pp. 1191–1203, 2018.

[18] A. M. Arias, Y. Marikawa, and N. Moris, "Gastruloids: Pluripotent stem cell models of mammalian gastrulation and embryo engineering," *Developmental Biology*, vol. 488, pp. 35–46, 2022.

[19] C. Ziegenhain, B. Vieth, S. Parekh, B. Reinius, A. Guillaumet-Adkins, M. Smets, H. Leonhardt, H. Heyn, I. Hellmann, and W. Enard, "Comparative analysis of single-cell rna sequencing methods," *Molecular cell*, vol. 65, no. 4, pp. 631–643, 2017.

[20] V. Svensson, K. N. Natarajan, L.-H. Ly, R. J. Miragaia, C. Labalette, I. C. Macaulay, A. Cvejic, and S. A. Teichmann, "Power analysis of single-cell rna-sequencing experiments," *Nature methods*, vol. 14, no. 4, pp. 381–387, 2017.

[21] F. Ma, B. K. Fuqua, Y. Hasin, C. Yukhtman, C. D. Vulpe, A. J. Lusis, and M. Pellegrini, "A comparison between whole transcript and 3'rna sequencing methods using kapa and lexogen library preparation methods," *BMC genomics*, vol. 20, no. 1, pp. 1–12, 2019.

[22] G. X. Zheng, J. M. Terry, P. Belgrader, P. Ryvkin, Z. W. Bent, R. Wilson, S. B. Ziraldo, T. D. Wheeler, G. P. McDermott, J. Zhu, *et al.*, "Massively parallel digital transcriptional profiling of single cells," *Nature communications*, vol. 8, no. 1, pp. 1–12, 2017.

[23] R. Patro, G. Duggal, M. Love, R. Irizarry, and C. Kingsford, "Salmon provides fast and bias-aware quantification of transcript expression.. 2017;(4): 417-419|," *Nat Methods*, vol. 14.

[24] M. D. Luecken and F. J. Theis, "Current best practices in single-cell rna-seq analysis: a tutorial," *Molecular systems biology*, vol. 15, no. 6, p. e8746, 2019.

[25] T. Ilicic, J. K. Kim, A. A. Kolodziejczyk, F. O. Bagger, D. J. McCarthy, J. C. Marioni, and S. A. Teichmann, "Classification of low quality cells from single-cell rna-seq data," *Genome biology*, vol. 17, no. 1, pp. 1–15, 2016.

[26] M. Dillies, A. Rau, J. Aubert, C. Hennequet-Antier, M. Jeanmougin, N. Servant, C. Keime, G. Marot, D. Castel, J. Estelle, *et al.*, "Laloã «d, le gall c, schaã «ffer b, le crom s, guedj m, jaffrézic f. a comprehensive evaluation of normalization methods for illumina high-throughput rna sequencing data analysis," *Brief Bioinform*, vol. 14, no. 6, p. 671, 2013.

[27] C. A. Vallejos, D. Risso, A. Scialdone, S. Dudoit, and J. C. Marioni, "Normalizing single-cell rna sequencing data: challenges and opportunities," *Nature methods*, vol. 14, no. 6, pp. 565–571, 2017.

[28] M. Büttner, Z. Miao, F. A. Wolf, S. A. Teichmann, and F. J. Theis, "A test metric for assessing single-cell rna-seq batch correction," *Nature methods*, vol. 16, no. 1, pp. 43–49, 2019.

[29] M. D. Luecken, M. Büttner, K. Chaichoompu, A. Danese, M. Interlandi, M. F. Müller, D. C. Strobl, L. Zappia, M. Dugas, M. Colomé-Tatché, *et al.*, "Benchmarking atlas-level data integration in single-cell genomics," *Nature methods*, vol. 19, no. 1, pp. 41–50, 2022.

[30] B. Hie, B. Bryson, and B. Berger, "Efficient integration of heterogeneous single-cell transcriptomes using scanorama," *Nature biotechnology*, vol. 37, no. 6, pp. 685–691, 2019.

[31] T. Andrews, M. T. Andrews, I. RColorBrewer, R. Suggests, S. biocViews RNASeq, and G. Transcriptomics, "Package 'm3drop'," *Methods*, vol. 10, pp. 1093–1095, 2016.

[32] L. Jiang, H. Chen, L. Pinello, and G.-C. Yuan, "Giniclust: detecting rare cell types from single-cell gene expression data with gini index," *Genome biology*, vol. 17, no. 1, pp. 1–13, 2016.

[33] B. L. Emert, C. J. Cote, E. A. Torre, I. P. Dardani, C. L. Jiang, N. Jain, S. M. Shaffer, and A. Raj, "Variability within rare cell states enables multiple paths toward drug resistance," *Nature biotechnology*, vol. 39, no. 7, pp. 865–876, 2021.

[34] D. Rodriguez-Terrones, X. Gaume, T. Ishiuchi, A. Weiss, A. Kopp, K. Kruse, A. Penning, J. M. Vaquerizas, L. Brino, and M.-E. Torres-Padilla, "A molecular roadmap for the emergence of early-embryonic-like cells in culture," *Nature genetics*, vol. 50, no. 1, pp. 106–119, 2018.

[35] L. KPFRS, "Edinburgh dublin philos. mag," *J. Sci*, vol. 2, p. 559, 1901.

[36] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," *arXiv preprint arXiv:1802.03426*, 2018.

[37] R. R. Coifman, S. Lafon, A. B. Lee, M. Maggioni, B. Nadler, F. Warner, and S. W. Zucker, "Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps," *Proceedings of the national academy of sciences*, vol. 102, no. 21, pp. 7426–7431, 2005.

[38] H. Bastos, B. Lassalle, A. Chicheportiche, L. Riou, J. Testart, I. Allemand, and P. Fouchet, "Flow cytometric characterization of viable meiotic and postmeiotic cells by hoechst 33342 in mouse spermatogenesis," *Cytometry Part A: the journal of the International Society for Analytical Cytology*, vol. 65, no. 1, pp. 40–49, 2005.

[39] F. A. Wolf, F. K. Hamey, M. Plass, J. Solana, J. S. Dahlin, B. Göttgens, N. Rajewsky, L. Simon, and F. J. Theis, "Paga: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells," *Genome biology*, vol. 20, no. 1, pp. 1–9, 2019.

[40] G. Lubatti, M. Stock, A. Iturbide, M. L. Ruiz Tejada Segura, M. Riepl, R. C. Tyser, A. Danese, M. Colomé-Tatché, F. J. Theis, S. Srinivas, *et al.*, "Ciara: a cluster-independent algorithm for identifying markers of rare cell types from single-cell sequencing data," *Development*, vol. 150, no. 11, 2023.

[41] T. Kim, I. R. Chen, Y. Lin, A. Y.-Y. Wang, J. Y. H. Yang, and P. Yang, "Impact of similarity metrics on single-cell rna-seq data clustering," *Briefings in bioinformatics*, vol. 20, no. 6, pp. 2316–2326, 2019.

[42] S. N. Morgan and K. K. Govender, "Conceptualizing loyalty in the south african mobile telecommunications industry," *Global Journal of Management And Business Research*, 2017.

[43] L. Zappia and A. Oshlack, "Clustering trees: a visualization for evaluating clusterings at multiple resolutions," *Gigascience*, vol. 7, no. 7, p. giy083, 2018.

[44] X. Yu, H. Liu, K. A. Hamel, M. G. Morvan, S. Yu, J. Leff, Z. Guan, J. M. Braz, and A. I. Basbaum, "Dorsal root ganglion macrophages contribute to both the initiation and persistence of neuropathic pain," *Nature communications*, vol. 11, no. 1, pp. 1–12, 2020.

[45] M. I. Love, W. Huber, and S. Anders, "Moderated estimation of fold change and dispersion for rna-seq data with deseq2," *Genome biology*, vol. 15, no. 12, pp. 1–21, 2014.

[46] M. D. Robinson, D. J. McCarthy, and G. K. Smyth, "edger: a bioconductor package for differential expression analysis of digital gene expression data," *bioinformatics*, vol. 26, no. 1, pp. 139–140, 2010.

[47] Z. He, Y. Pan, F. Shao, and H. Wang, "Identifying differentially expressed genes of zero inflated single cell rna sequencing data using mixed model score tests," *Frontiers in genetics*, vol. 12, p. 616686, 2021.

[48] M. Lotfollahi, M. Naghipourfar, M. D. Luecken, M. Khajavi, M. Büttner, M. Wagenstetter, Ž. Avsec, A. Gayoso, N. Yosef, M. Interlandi, *et al.*, "Mapping single-cell data to reference atlases by transfer learning," *Nature biotechnology*, vol. 40, no. 1, pp. 121–130, 2022.

[49] V. Y. Kiselev, A. Yiu, and M. Hemberg, "scmap: projection of single-cell rna-seq data across data sets," *Nature methods*, vol. 15, no. 5, pp. 359–362, 2018.

[50] C. Li, B. Liu, B. Kang, Z. Liu, Y. Liu, C. Chen, X. Ren, and Z. Zhang, "Scibet as a portable and fast single cell type identifier," *Nature communications*, vol. 11, no. 1, pp. 1–8, 2020.

[51] T. Stuart, A. Butler, P. Hoffman, C. Hafemeister, E. Papalexi, W. M. Mauck III, Y. Hao, M. Stoeckius, P. Smibert, and R. Satija, "Comprehensive integration of single-cell data," *Cell*, vol. 177, no. 7, pp. 1888–1902, 2019.

[52] Y. Kang, D. Thieffry, and L. Cantini, "Evaluating the reproducibility of single-cell gene regulatory network inference algorithms," *Frontiers in genetics*, p. 362, 2021.

[53] Q. Deng, D. Ramsköld, B. Reinius, and R. Sandberg, "Single-cell rna-seq reveals dynamic, random monoallelic gene expression in mammalian cells," *Science*, vol. 343, no. 6167, pp. 193–196, 2014.

[54] H. Mohammed, I. Hernando-Herraez, A. Savino, A. Scialdone, I. Macaulay, C. Mulas, T. Chandra, T. Voet, W. Dean, J. Nichols, *et al.*, "Single-cell landscape of transcriptional heterogeneity and cell fate decisions during mouse early gastrulation. cell rep 20: 1215–1228," *J. CELREP*, vol. 9, 2017.

[55] A. Iturbide, M. L. Ruiz Tejada Segura, C. Noll, K. Schorpp, I. Rothenaigner, E. R. Ruiz-Morales, G. Lubatti, A. Agami, K. Hadian, A. Scialdone, *et al.*, "Retinoic acid signaling is critical during the totipotency window in early mammalian development," *Nature Structural & Molecular Biology*, vol. 28, no. 6, pp. 521–532, 2021.

[56] A. Tanay and A. Regev, "Scaling single-cell genomics from phenomenology to mechanism," *Nature*, vol. 541, no. 7637, pp. 331–338, 2017.

[57] L. Haghverdi, M. Büttner, F. A. Wolf, F. Buettner, and F. J. Theis, "Diffusion pseudotime robustly reconstructs lineage branching," *Nature methods*, vol. 13, no. 10, pp. 845–848, 2016.

[58] K. Street, D. Risso, R. B. Fletcher, D. Das, J. Ngai, N. Yosef, E. Purdom, and S. Dudoit, "Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics," *BMC genomics*, vol. 19, no. 1, pp. 1–16, 2018.

[59] V. Bergen, M. Lange, S. Peidli, F. A. Wolf, and F. J. Theis, "Generalizing rna velocity to transient cell states through dynamical modeling," *Nature biotechnology*, vol. 38, no. 12, pp. 1408–1414, 2020.

[60] G. La Manno, R. Soldatov, A. Zeisel, E. Braun, H. Hochgerner, V. Petukhov, K. Lidschreiber, M. E. Kastriti, P. Lönnerberg, A. Furlan, *et al.*, "Rna velocity of single cells," *Nature*, vol. 560, no. 7719, pp. 494–498, 2018.

[61] P. Weiler, K. Van den Berge, K. Street, and S. Tiberi, "A guide to trajectory inference and rna velocity," in *Single Cell Transcriptomics: Methods and Protocols*, pp. 269–292, Springer, 2022.

[62] M. Lange, V. Bergen, M. Klein, M. Setty, B. Reuter, M. Bakhti, H. Lickert, M. Ansari, J. Schniering, H. B. Schiller, *et al.*, "Cellrank for directed single-cell fate mapping," *Nature methods*, vol. 19, no. 2, pp. 159–170, 2022.

[63] K. Van den Berge, H. Roux de Bézieux, K. Street, W. Saelens, R. Cannoodt, Y. Saeys, S. Dudoit, and L. Clement, "Trajectory-based differential expression analysis for single-cell sequencing data," *Nature communications*, vol. 11, no. 1, pp. 1–13, 2020.

[64] R. C. Tyser, E. Mahammadov, S. Nakanoh, L. Vallier, A. Scialdone, and S. Srinivas, "Single-cell transcriptomic characterization of a gastrulating human embryo," *Nature*, vol. 600, no. 7888, pp. 285–289, 2021.

[65] G. Lubatti, M. Stock, A. Iturbe, M. L. R. T. Segura, R. Tyser, F. J. Theis, S. Srinivas, M.-E. Torres-Padilla, and A. Scialdone, "Ciara: a cluster-independent algorithm for the identification of markers of rare cell types from single-cell rna seq data," *bioRxiv*, 2022.

[66] G. Schiebinger, J. Shu, M. Tabaka, B. Cleary, V. Subramanian, A. Solomon, J. Gould, S. Liu, S. Lin, P. Berube, *et al.*, "Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming," *Cell*, vol. 176, no. 4, pp. 928–943, 2019.

[67] L. S. Ludwig, C. A. Lareau, J. C. Ulirsch, E. Christian, C. Muus, L. H. Li, K. Pelka, W. Ge, Y. Oren, A. Brack, *et al.*, "Lineage tracing in humans enabled by mitochondrial mutations and single-cell genomics," *Cell*, vol. 176, no. 6, pp. 1325–1339, 2019.

[68] J. B. Stewart and P. F. Chinnery, "The dynamics of mitochondrial dna heteroplasmy: implications for human health and disease," *Nature Reviews Genetics*, vol. 16, no. 9, pp. 530–542, 2015.

[69] V. I. Floros, A. Pyle, S. Dietmann, W. Wei, W. C. Tang, N. Irie, B. Payne, A. Capalbo, L. Noli, J. Coxhead, *et al.*, "Segregation of mitochondrial dna heteroplasmy through a developmental genetic bottleneck in human embryos," *Nature cell biology*, vol. 20, no. 2, pp. 144–151, 2018.

[70] T. Biezuner, A. Spiro, O. Raz, S. Amir, L. Milo, R. Adar, N. Chapal-Ilani, V. Berman, Y. Fried, E. Ainbinder, *et al.*, "A generic, cost-effective, and scalable cell lineage analysis platform," *Genome research*, vol. 26, no. 11, pp. 1588–1599, 2016.

[71] A. Lima, G. Lubatti, J. Burgstaller, D. Hu, A. P. Green, A. Di Gregorio, T. Zawadzki, B. Pernaute, E. Mahammadov, S. Perez-Montero, *et al.*, "Cell competition acts as a purifying selection to eliminate cells with mitochondrial defects during early mouse development," *Nature metabolism*, vol. 3, no. 8, pp. 1091–1108, 2021.

[72] M. S. Sharpley, C. Marciniak, K. Eckel-Mahan, M. McManus, M. Crimi, K. Waymire, C. S. Lin, S. Masubuchi, N. Friend, M. Koike, *et al.*, "Heteroplasmy of mouse mtdna is genetically unstable and results in altered behavior and cognition," *Cell*, vol. 151, no. 2, pp. 333–343, 2012.

[73] J. H. Kauppila, H. L. Baines, A. Bratic, M.-L. Simard, C. Freyer, A. Mourier, C. Stamp, R. Filograna, N.-G. Larsson, L. C. Greaves, *et al.*, "A phenotype-driven approach to generate mouse models with pathogenic mtdna mutations causing mitochondrial disease," *Cell reports*, vol. 16, no. 11, pp. 2980–2990, 2016.

[74] M. Ginsburg, M. Snow, and A. McLAREN, "Primordial germ cells in the mouse embryo during gastrulation," *Development*, vol. 110, no. 2, pp. 521–528, 1990.

[75] T. E. Miller, C. A. Lareau, J. A. Verga, E. A. DePasquale, V. Liu, D. Ssozi, K. Sandor, Y. Yin, L. S. Ludwig, C. A. El Farran, *et al.*, "Mitochondrial variant enrichment from high-throughput single-cell rna sequencing resolves clonal populations," *Nature Biotechnology*, pp. 1–5, 2022.

[76] C. Calabrese, D. Simone, M. A. Diroma, M. Santorsola, C. Gutta, G. Gasparre, E. Picardi, G. Pesole, and M. Attimonelli, "Mtoolbox: a highly automated pipeline for heteroplasmy annotation and prioritization analysis of human mitochondrial variants in high-throughput sequencing," *Bioinformatics*, vol. 30, no. 21, pp. 3115–3117, 2014.

[77] X. Huang and Y. Huang, "Cellsnp-lite: an efficient tool for genotyping single cells," *Bioinformatics*, vol. 37, no. 23, pp. 4569–4571, 2021.

[78] N. Prashant, N. Alomran, Y. Chen, H. Liu, P. Bousounis, M. Movassagh, N. Edwards, and A. Horvath, "Screadcounts: Estimation of cell-level snvs expression from scrna-seq data," *BMC genomics*, vol. 22, no. 1, pp. 1–9, 2021.

[79] B. Pijuan-Sala, J. A. Griffiths, C. Guibentif, T. W. Hiscock, W. Jawaid, F. J. Calero-Nieto, C. Mulas, X. Ibarra-Soria, R. C. Tyser, D. L. L. Ho, *et al.*, "A single-cell molecular map of mouse gastrulation and early organogenesis," *Nature*, vol. 566, no. 7745, pp. 490–495, 2019.

[80] R. Wegmann, M. Neri, S. Schuierer, B. Bilican, H. Hartkopf, F. Nigsch, F. Mapa, A. Waldt, R. Cuttat, M. R. Salick, *et al.*, "Cellsius provides sensitive and specific detection of rare cell populations from complex single-cell rna-seq data," *Genome biology*, vol. 20, no. 1, pp. 1–21, 2019.

[81] A. Jindal, P. Gupta, D. Sengupta, *et al.*, "Discovery of rare cells from voluminous single cell expression data," *Nature communications*, vol. 9, no. 1, pp. 1–9, 2018.

[82] A. Warmflash, B. Sorre, F. Etoc, E. D. Siggia, and A. H. Brivanlou, "A method to recapitulate early embryonic spatial patterning in human embryonic stem cells," *Nature methods*, vol. 11, no. 8, pp. 847–854, 2014.

[83] M. Simunovic, J. J. Metzger, F. Etoc, A. Yoney, A. Ruzo, I. Martyn, G. Croft, D. S. You, A. H. Brivanlou, and E. D. Siggia, "A 3d model of a human epiblast reveals bmp4-driven symmetry breaking," *Nature cell biology*, vol. 21, no. 7, pp. 900–910, 2019.

[84] B. Schloerke, J. Crowley, D. Cook, F. Briatte, M. Marbach, E. Thoen, A. Elberg, and J. Larmarange, "Ggally: Extension to 'ggplot2'," *R package version*, vol. 1, no. 0, 2018.

[85] I. G. M. Brons, L. E. Smithers, M. W. Trotter, P. Rugg-Gunn, B. Sun, S. M. Chuva de Sousa Lopes, S. K. Howlett, A. Clarkson, L. Ahrlund-Richter, R. A. Pedersen, *et al.*, "Derivation of pluripotent epiblast stem cells from mammalian embryos," *Nature*, vol. 448, no. 7150, pp. 191–195, 2007.

[86] S. Velychko, K. Adachi, K.-P. Kim, Y. Hou, C. M. MacCarthy, G. Wu, and H. R. Schöler, "Excluding oct4 from yamanaka cocktail unleashes the developmental potential of ipscs," *Cell stem cell*, vol. 25, no. 6, pp. 737–753, 2019.

[87] P. Osteil, A. Moulin, C. Santamaria, T. Joly, L. Jouneau, M. Aubry, Y. Tapponnier, C. Archilla, B. Schmaltz-Panneau, J. Lecardonnel, *et al.*, "A panel of embryonic stem cell lines reveals the variety and dynamic of pluripotent states in rabbits," *Stem cell reports*, vol. 7, no. 3, pp. 383–398, 2016.

[88] Y. Tapponnier, M. Afanassieff, I. Aksoy, M. Aubry, A. Moulin, L. Medjani, W. Bouchereau, C. Mayère, P. Osteil, J. Nurse-Francis, *et al.*, "Reprogramming of rabbit induced pluripotent stem cells toward epiblast and chimeric competency using krüppel-like factors," *Stem cell research*, vol. 24, pp. 106–117, 2017.

[89] M. Kinoshita, T. Kobayashi, B. Planells, D. Klisch, D. Spindlow, H. Masaki, S. Bornelöv, G. G. Stirparo, H. Matsunari, A. Uchikura, *et al.*, "Pluripotent stem cells related to embryonic disc exhibit common self-renewal requirements in diverse livestock species," *Development*, vol. 148, no. 23, p. dev199901, 2021.

[90] A. Rodríguez, C. Allegrucci, and R. Alberio, "Modulation of pluripotency in the porcine embryo and ips cells," *PloS one*, vol. 7, no. 11, p. e49079, 2012.

[91] V. A. Huynh-Thu, A. Irrthum, L. Wehenkel, and P. Geurts, "Inferring regulatory networks from expression data using tree-based methods," *PloS one*, vol. 5, no. 9, p. e12776, 2010.

[92] S. Nishikawa, A. Takamatsu, S. Ohsawa, and T. Igaki, "Mathematical model for cell competition: Predator–prey interactions at the interface between two groups of cells in monolayer tissue," *Journal of theoretical biology*, vol. 404, pp. 40–50, 2016.

[93] S. P. Montero, S. Bowling, R. Pérez-Carrasco, and T. A. Rodriguez, "Levels of p53 expression determine the competitive ability of embryonic stem cells during the onset of differentiation," *bioRxiv*, 2022.

[94] M. Efremova, M. Vento-Tormo, S. A. Teichmann, and R. Vento-Tormo, "Cellphonedb: inferring cell–cell communication from combined expression of multi-subunit ligand–receptor complexes," *Nature protocols*, vol. 15, no. 4, pp. 1484–1506, 2020.

[95] N. Browaeys, "Modeling intercellular communication by linking ligands to target genes," *Nat Methods*, no. 17.

[96] J. N. Wells and C. Feschotte, "A field guide to eukaryotic transposable elements," *Annual review of genetics*, vol. 54, p. 539, 2020.

[97] V. V. Kapitonov and J. Jurka, "A universal classification of eukaryotic transposable elements implemented in repbase," *Nature Reviews Genetics*, vol. 9, no. 5, pp. 411–412, 2008.

[98] Y. Jin, O. H. Tam, E. Paniagua, and M. Hammell, "Tetranscripts: a package for including transposable elements in differential expression analysis of rna-seq datasets," *Bioinformatics*, vol. 31, no. 22, pp. 3593–3599, 2015.

[99] H.-H. Jeong, H. K. Yalamanchili, C. Guo, J. M. Shulman, and Z. Liu, "An ultra-fast and scalable quantification pipeline for transposable elements from next generation sequencing data," in *PACIFIC SYMPOSIUM ON BIOCOMPUTING 2018: Proceedings of the Pacific Symposium*, pp. 168–179, World Scientific, 2018.