© 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

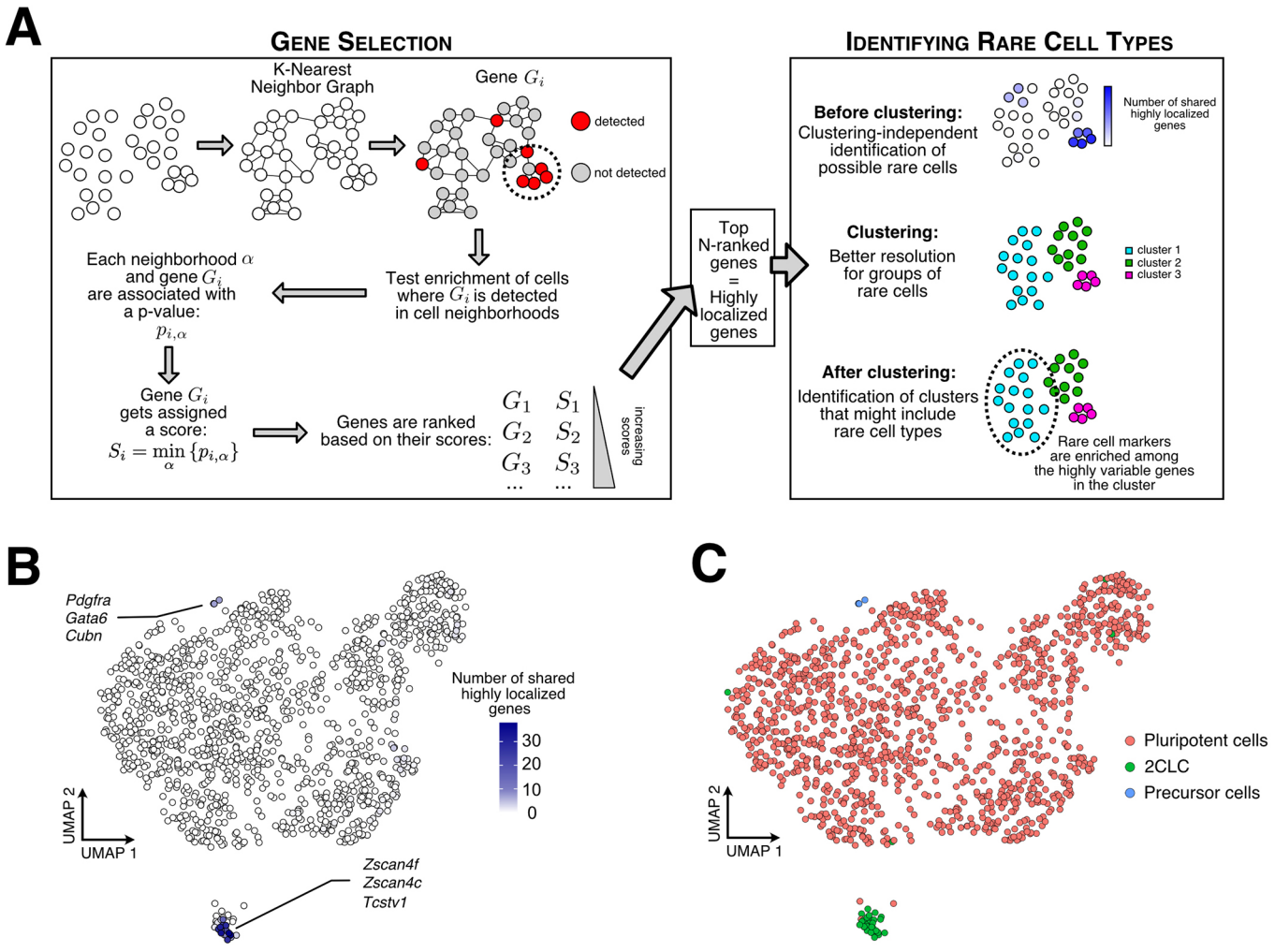


Fig. 1. Schematic representation and example of application of CIARA. (A) Left: CIARA computes a score for each gene based on how cells expressing that gene are distributed on a K-nearest neighbor graph. Lower scores correspond to genes that are mostly expressed in neighboring cells, i.e. are 'highly 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.

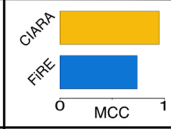
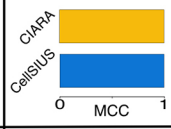
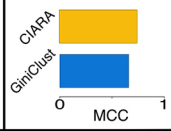
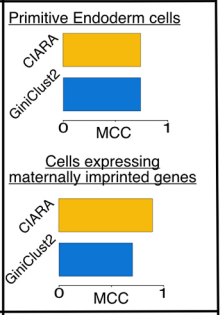
We first applied standard quality-control thresholds, which led to the selection of 766 good-quality cells (Fig. S4A-D; Materials and Methods). A UMAP plot showing which cells express the

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

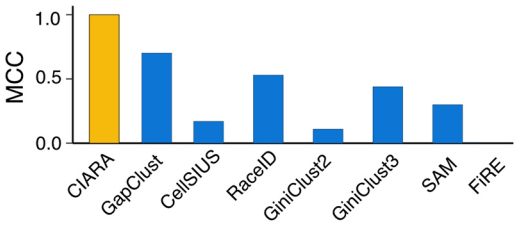
A

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

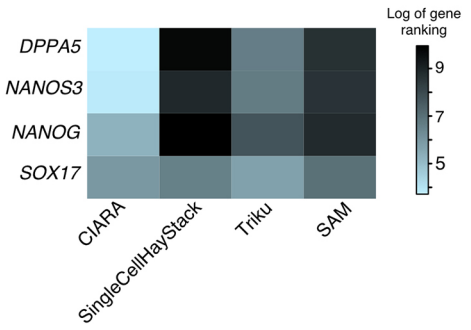


Fig. 2. CIARA outperforms existing methods for detecting rare cell types. (A) Table listing the methods for rare cell type identification that we benchmarked CIARA against. Specific features of each approach are indicated. (B) Table summarizing the data sets and results of the benchmarking analysis. 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

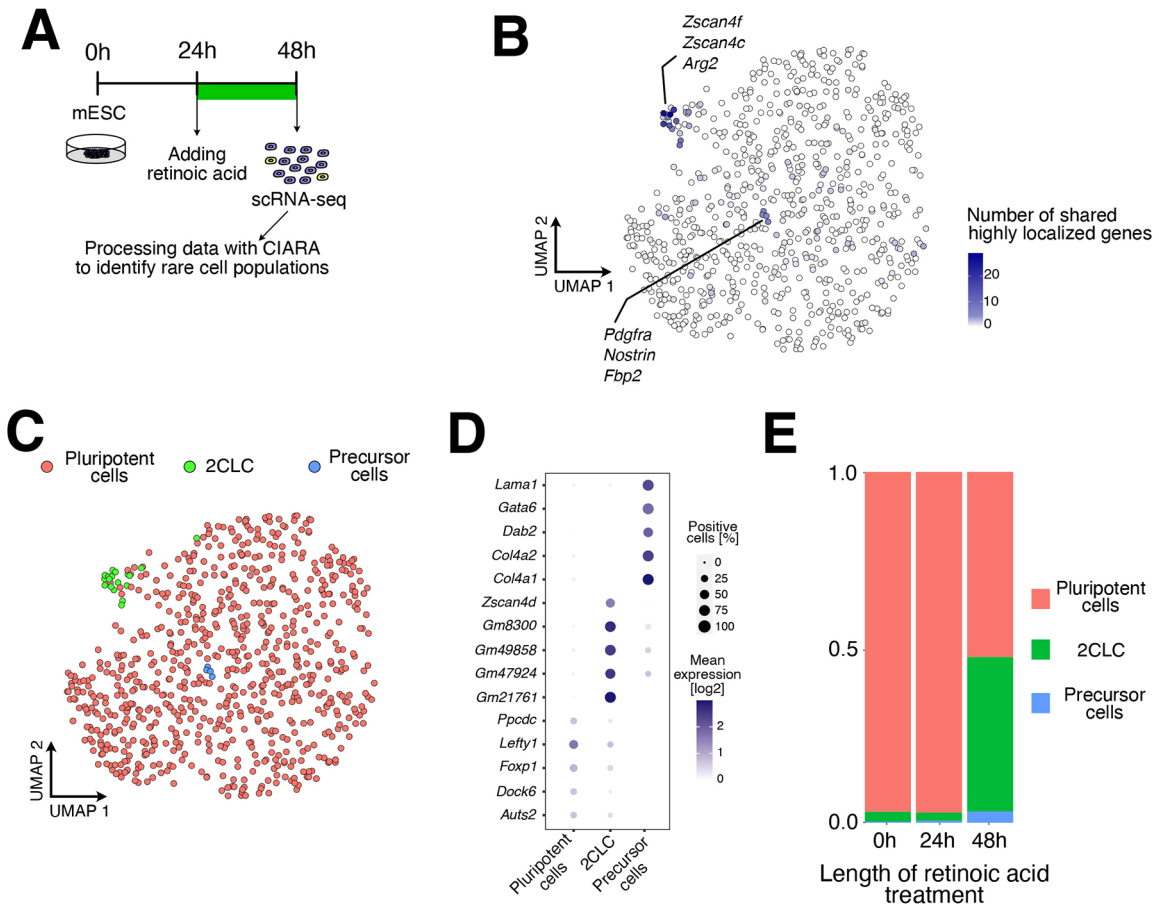


Fig. 3. CIARA identifies rare populations of totipotent-like and differentiating cells among mESCs treated with retinoic acid. (A) We treated mESCs with retinoic acid for 24 h before collecting and processing them for scRNA-seq. (B) UMAP representation of the mESC dataset (*n*=766 cells) indicating the 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

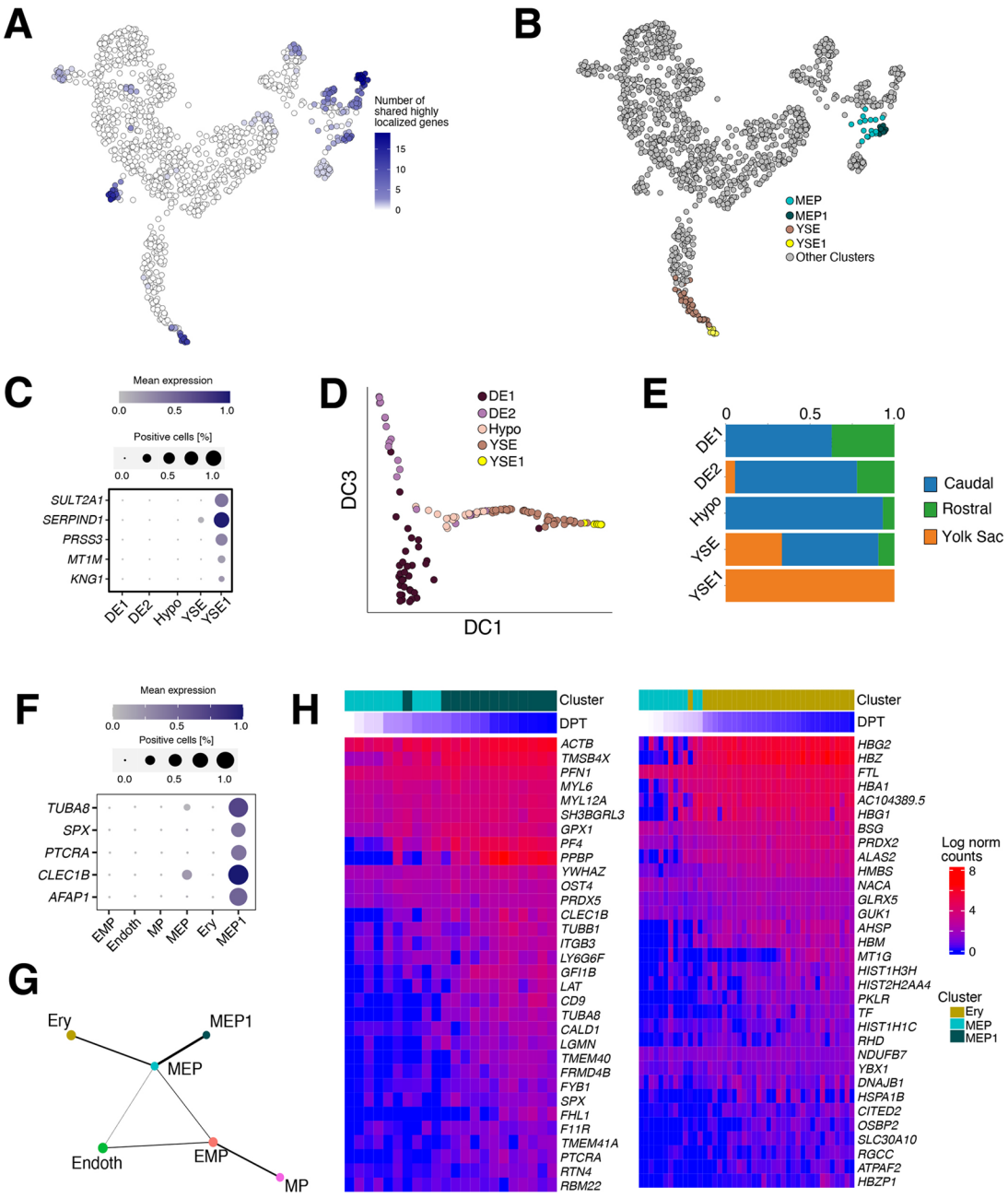


Fig. 4. CIARA identifies previously uncharacterized rare populations of cells in a human gastrula dataset. (A) UMAP representation of the human gastrula dataset (*n*=1195 cells; Tyser et al., 2021b) showing the number of shared highly localized genes in each cell. (B) Same UMAP representation as in A, 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