**CIARA: a cluster independent algorithm for the identification of rare cell types from single cell RNA seq data**

Gabriele Lubatti,1,2,3 Marco Stock1,2,3 and Antonio Scialdone1,2,3

1Institute 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

**Abstract**

Single cell RNA sequencing (scRNA-seq) offers the opportunity to have access to the transcriptome of very high numbers of cells. A common task is to identify clusters of cells based on transcriptome similarities. Standard clustering approaches perform well in identifying common cell types but tend to miss rare cell types. Here we propose CIARA (Cluster Independent Algorithm for the identification of RAre cell types), a cluster independent approach that selects genes localized in a small number of neighboring cells from high dimensional PCA space. The selected genes are then easily integrated with common clustering algorithms to identify rare clusters. We showed that CIARA outperforms existing methods for rare cell type detection, is scalable, fast and allows the identification of new rare cell subtypes from human gastrula and mouse embryonic stem cells. The concept of CIARA is implemented in the user-friendly package, which is both available for R and python.

**Introduction**

In many single-cell studies, the goal is to characterize the heterogeneity of cell types/states based on their transcriptome.

While standard clustering algorithms can identify groups of cells consisting of hundreds or thousands, they usually fail to detect the more rare populations, including few tens of cells. Detection of rare cell types might be very important in a broad variety of context including primordial germ cells (PGCs) to early embryonic like cells.  
For this reason, ad hoc algorithms have been proposed to search for rare cell populations. Some (CellSIIUS6, RaceID8) detect rare cell types starting from an existing cluster annotation or by assigning a rareness score to each of the cells using a sketching technique to measure the density around them (FiRE7). Others (GiniClust2/316,5, singleCellHaystack9, SAM18) rank genes in a cluster independent way, based on their distribution across cells and then this set of genes can be used as features for cluster analysis. While the methods generally work well in selecting rare cells with strong markers signature, they have some drawbacks in identifying small populations whose markers are not expressed at high level, thus very specific to the population. The main reason lies in the several parameters required by the different approaches whose default values tend to miss weaker signal. Many of the methods that rank genes in a cluster independent way give an high rank to both markers of rare cell types and genes that are just expressed in few random cells. This is corrected with approaches that involve a new set of parameters and assumptions.

We devised a new method called CIARA (Cluster Independent Algorithm for the identification of RAre cell types) that identifies potential markers of rare cell types looking at genes whose expression is confined in small regions of the expression space. CIARA is based on the idea that rare cell populations are characterized by a set of markers that are expressed only in the rare cell types but not in all the other cells. Using the fisher test and relying on a k-nearest neighbors matrix given as input, CIARA is able to distinguish between genes expressed in few neighbor cells and genes expressed in few cells far from each other in the expression space. Only the first set of genes are relevant for rare cell types detection.

Below, we show how CIARA outperforms existing algorithms in different settings including different scRNA-seq protocols  
We will apply CIARA to a new dataset of mouse ES cells treated with retinoic acid, where we'll show how it successfully identifies a small group of 4 cells that is beginning to differentiate. Moreover, when applied to a unique dataset from a gastrulating human embryo, CIARA identifies rare populations corresponding to megakaryocytes and a subset of endoderm cells which appear also in mouse at a later stage.

**Results**

**Overview of CIARA**

Rare cell populations are characterized by a set of markers that are expressed only in the rare cell types but not in all the other cells. The idea of CIARA is to identify potential markers of rare cells with a 4-step approach relying on the p value given by the fisher test and on the position of cells given by the k-nearest neighbors matrix.

The input consists of a normalized count matrix and a k nearest neighbor matrix (knn matrix), previously defined using a standard approach (euclidean distance between cells in the high dimensional PCA space built using as features HVGs- see method section) (**fig 1 a**).

Then for each of the genes previously identified we binarize the normalized expression using the median as threshold. In this way each cell will have value 1 (original expression above the median) or 0 (original expression below the median) (**fig 1 a**). Using the knn matrix provided as input, for each of the cells we define a local region, made up by the cell and its knn.

The distribution of 1 and 0 within the local region follows a hypergeometric distribution (see method section).

A marker of rare cells will be only or mainly expressed in a small number of neighboring cells (so there will be one or more local regions enriched in 1), while a non-marker will be expressed in cells that are not neighbors (so there will not be any enriched local region in 1) (**fig 1 a**).

To test the enrichment in 1 a fisher test is performed. If the gene is enriched in at least one local region, then the p value from the fisher test is assigned to the local region. The final p value is defined as the minimum value between all the local p values (see method). If a gene is not enriched in any local region, then it will have a default final p value equal to 1 (**fig 1 a**).

Since potential rare cell markers will only be expressed in a small number of cells, we reduce the set of genes for which we apply CIARA, keeping only those whose expression is above a fixed expression threshold in fewer cells than a chosen maximum cell threshold (**fig 1 b**).

As output, the CIARA algorithm returns the list of all tested genes, ranked according to increasing p value given by the fisher test (see method section) (**fig 1 b**).

Each of the tested genes will have a p value between 0 (highly localized in a small number of neighboring cells) and 1 (not localized).

We used CIARA on published mouse embryonic stem cell (mESC) dataset4. In mESCs a rare population of cells (less than%1 of the total) called 2cells-like cells (2clc) is present3.

Running the CIARA algorithm reveals that the gene with the lowest p value is Zscan4-ps3, a known marker of 2clc3,20 while among the genes not enriched (p value equal 1) there is AC147512.2 (**fig1 b**).

It is also possible to visualize where the top rare cell type markers are located in the umap plot (**fig 1 b**). This plot can also be provided in an interactive way.

The two main features of CIARA are that 1) the identification of rare cell type markers is done without relying on cluster annotation and the results from CIARA can be used to better refine the cluster annotation and 2) having a known theoretical distribution for testing the enrichment of the local regions (hypergeometric distribution) allows to extremely reduce the number of parameters in the algorithm (see method).

Since a common task in scRNA sequencing is to perform cluster analysis on cells, we propose a way to integrate the result of CIARA with existing standard cluster algorithms for better identification of rare cell types. The step-a is to perform a cluster analysis with a default approach (i.e. Louvain/Leiden algorithm provided by Seurat19/Scanpy21 using highly variable genes as features) (**fig 1 c**). In step-b, we re-run the same algorithm using genes with low p value as features (**fig 1 c**). This step allows to distinguish rare cell types from common cell types in a completely unsupervised way.

Any cluster that arises at step-b with a number of cells below a chosen threshold will become an independent cluster in the original partition at step-a (s **fig 1 c**).

Starting from the partition of step 1, CIARA can be used to provide an objective criterion for running sub-clustering. Only for clusters where the intersection between the top 100 HVGs and genes with low p value is significant, a sub-cluster approach is performed using a default approach, as shown in step 1 (**fig 1 c**).

This additional sub-cluster approach can be useful for detecting small populations of cells that will not arise in step-b since the cells share many markers with a common cell type.

**CIARA outperforms existing method**

We test the performance of CIARA in comparison with other methods that were specifically developed for the identification of rare cell types. In particular we focus on GiniClust2/316,5, CellSIUS 6, FiRE7 , RaceID8, singleCellHaystack9 and SAM18 .The last two tools are cluster independent algorithm for marker genes identification without a specific focus on rare cell types.

The methods differ in terms of input requirements and output structure.

CellSIUS is the only method that requires cluster partition as input. All of the methods provide a clustering on cells as output, apart from singleCellHaystack. It just returns a list of marker genes with non-random expression patterns in high dimensional PCA space. GiniClust2/3, CellSIUS and CIARA also provide a list of rare cell type markers as output (**fig 2 a**).

We test the performance of CIARA, by running it on datasets that were part of original papers of the above mentioned methods.

In particular we first focus on two datasets for which a ground truth assignment is available.

The first is a scRNA sequencing data comprising 293T and Jurkat cells (1580 cells) (**6**). The proportion of Jurkat cells is at 2.5%. This dataset was included in the original FiRE publication (**fig 2 b**).

The CIARA approach consistently identified the two populations of cells (**fig 2 c**).

We further quantify the performance of CIARA and FiRE in comparison to the ground truth with the Matthew correlation coefficient (MCC). The MCC is a metric that quantifies the overall agreement between two binary classifications, taking into account both true and false positives and negatives. The MCC value ranges from − 1 to 1, where 1 means a perfect agreement between a clustering and the reference, 0 means the clustering is as good as a random guess, and − 1 means a total disagreement between a clustering and the reference.

The performance of CIARA (MCC 0.95) is strikingly better than FiRE (MCC 0.74) (**fig 2 d**). With CIARA the number of false positives (4 cells) is much lower than the false positives detected by FiRE (32 cells (**sup figure 1 e**)).

The second scRNA sequencing dataset is from the original CellSIUS paper **(7)** and it includes 8 human cell lines with known cellular composition. The analyzed dataset has 3984 cells and the cell lines with the smallest number of cells are H1437 and Jurkat, with 3 and 6 cells respectively (**fig 2 e**). Our CIARA algorithm is perfectly able to identify the two small clusters and it reach the same performance of CellSIUS (MCC equal to 1 for both methods) (**fig 2 f g**).

We extensively tested CIARA on several other dataset for which the ground truth is not available, published in original papers of rare cell type methods (**fig 2 h**). We define the rare cells as the set of cells with positive expression of markers used in the original paper to assign the identity to the rare cell type.

To assess the performance of CIARA, we compare the Jaccard indexes between rare cluster from CIARA/original methods and rare cells.

CIARA was able to recover the 2clc cluster (3 cells) from a dataset of 2484 mouse ESCs13 and from a dataset obtained of glioblastoma (GBM) primary tumors (576 cells) 14 to identify a rare cluster of 9 cells originating from tumor MGH31.

Both of these two datasets were part of the GiniClust paper17.

While for the first dataset the Jaccard index from the two methods is the same from, for the second dataset CIARA (Jaccard index 0,56) performs better than GiniClust (Jaccard index 0,44).

The CIARA algorithm was then run on a dataset associated with leukemia inhibitory factor (LIF) withdrawal-induced mouse embryonic stem cell (mESC) differentiation. In particular we focused on the cells assayed on day 4 post-LIF withdrawal13, as done in the GiniClust2 paper. In the paper, two rare cell types were detected.

The first rare cell types expressed known primitive endoderm (PrEn) markers (Col4a1, Col4a2, Lama1, Lama2, and Ctsl) and the second has maternally imprinted genes as markers (Rhox6, Rhox9, and Sct).

Regarding the first rare population, CIARA and GiniClust2 achieved the same performance, while for the second small population CIARA (Jaccard index 0.80) outperforms GiniClust2 (Jaccard index 0.56).

Finally we test the CIARA algorithm on the murine intestinal epithelial cells dataset (317 cells)15 that is part of RaceID vignette. RaceID identifies 8 rare clusters between 3 and 27 cells while CIARA selected 6 small clusters between 3 and 25 cells (**sup figure 2 a**).

The CIARA approach is able to detect all the cell types showed in (**15**) (goblet cells, enterocytes, paneth cells and enteroendocrine cells, see method section) and assign them to a unique cluster (sometimes in the original paper the same cell type was split in more clusters like for goblet cells).

In addition, CIARA identifies two more small clusters (3 and 4 cells) that are well characterized according to the transcriptome (**sup figure 2 b, c**).

The 3-cells cluster expressed typical markers of Tuft cells (Dclk1, Lrmp, Trpm5)22.

Tuft cells constitute a minor fraction (0.4 %) of the adult mouse intestinal epithelium23.

These results suggest that RaceID, tends very often to over-cluster and to identify small clusters that are not well defined according to markers.

We then test the performance of CIARA and the other methods on a scRNA seq data from a human embryo at the gastrula stage10.

The smallest cluster in terms of number of cells provided in the original paper is the primordial germ cell cluster (7 cells) (**fig 2 i**). In the original paper it was possible to identify this cluster only with a supervised approach starting from previously known PGC markers.

Identification of PGCs is particularly difficult because some of the known markers are also expressed in other cell types (e.g. SOX17) (**sup figure 1 b**).

The second smallest cluster provided in the original cluster annotation is the Axial Mesoderm (24 cells) (**fig 2 i**).

We wanted to test if the different methods are able to detect the PGC cluster.

We define cells with a positive expression of 3 previously known PGC markers (NANOS3, NANOG, DPPA5) as PGCs. This criterion leads to a total of 7 PGCs.

We assume that the PGCs are detected by the different methods if the Jaccard index between the rare cell type and one of the clusters provided by the method is greater or equal to 0.5 (for CIARA, RaceID, CellSIUS, GiniClust2 and GiniClust3). For FiRE, which assigns to each cell 0 (no rare type) or 1 (rare type), if the PGCs are enriched in 1. For singleCellHaystack and SAM, which returns a ranked list of genes, if the top 1000 genes are enriched in markers of PGCs (see method).

By following the above criteria, CIARA is the only method available to identify the PGCs cluster (**fig 2 j**).

Running CIARA on the human gastrula data set and then consequently performing cluster analysis (Louvain algorithm) using localized genes from CIARA as features easily identified the PGCs cluster in a completely unsupervised way (**fig 2 j**). The identification of the PGC cluster is extremely robust and not depending on the resolution value fixed for running the Louvain clustering algorithm (**sup figure 1 a**).

In addition to the PGC cluster, the Axial Mesoderm cluster arises as an independent cluster (**fig 2 j**).

GiniClust3 allows to identify both common and rare cell types, using the Gini index and Fano factor, respectively, through a cluster-aware, weighted ensemble clustering technique. The Gini index describes how much the gene expression is unequally distributed across cells and is always between 0 (equally distributed) and 1 (high inequality in distribution and therefore a potential marker of rare cells).

The Gini index correlates with the maximum value of gene expression. The way the authors suggest to correct this bias (two-step curve fitting strategy with smooth curve through all data points by LOESS) tends to keep genes with higher maximum value compared to the genes selected with our CIARA approach. Therefore, relevant markers with low or medium expression may not be selected among the genes with high Gini index and the final clustering will be affected (**sup figure 1 c, d**).

This is the case for PGC markers NANOS, DPPA5 and SOX17. They have a log2 max value around 8.5 and are not among the genes selected by the Gini index (129 in total).

A sub-optimal selection of genes with high Gini index may lead to an excessive high number of clusters, many of them not well defined according to markers.

CellSIUS is a two-step approach consisting of an initial coarse clustering step followed by CellSIUS. CellSIUS performs sub-clustering using a set of genes that show bi-modal distribution within the cluster as features. The mean expression in the second mode is then compared to this gene’s mean expression level outside the cluster.

To define cluster specific genes, we need to introduce additional parameters. Standard values may miss weaker signals (as for PGC markers). On the other hand, lowering the threshold can lead to an excessive number of clusters not well defined according to marker genes.

In the human gastrula data set, markers of PGCs (NANOS2, NANOG, SOX17, DPP5A) are not selected as genes with bimodal distribution because the min fc parameter (minimum difference in mean (log2) between the two modes of the gene expression distribution) in the CellSIUS function is always below 1.5 and the default value is 2. Manually lowering the value of min fc parameter to 1 allows to recover PGCs as a cluster.

FiRE assigns a rareness score to each of the cells using a sketching technique to measure the density around them. Rare cells are cells with low density.

FiRE marks a cell as rare if its FiRE score is ≥q3 + 1.5 × IQR, where q3 and IQR denote the third quartile and the interquartile range (75th percentile−25th percentile).

It has the advantage of not relying on clusters, but finding the optimal threshold to label a cell as rare may need to be fine-tuned for different datasets. The default value leads to no cells labeled as rare for the human gastrula dataset. Lowering the threshold (to 0.5XIQR) allows to have some cells labeled as rare (124 cells in total), but none of them is a PGC.

RaceID identifies rare cell types using a two-step procedure. The first step is K-means clustering to detect large clusters and then outliers are detected within each cluster. Outlier genes are detected based on the variance of expression level. It often tends to over-cluster also because single outliers can be assigned to an individual cluster. The 7 PGCs cells with the final race ID partition are split in 5 different clusters.

SingleCellHaystack detects differential expressed (DE) genes without relying on cluster analysis. SingleCellHaystack uses Kullback-Leibler Divergence to find genes that are expressed in subsets of cells that are non-randomly positioned in a PCA multi-dimensional space or 2D representations.

It assigns significance to genes in comparison with randomization, where expression levels are randomly shuffled over all cells. This correction introduces additional parameters (as for the Gini index correction) and the default values may lead to missing genes expressed only in a small number of cells. Another limitation of SingleCellHaystack is that it provides only a list of genes with non-random patterns in PCA space but it does not provide a way to integrate this information in order to perform cluster analysis on cells.

Indeed, PGC markers NANOS3, NANOG and DPPA5 are not among the top 1000 genes detected by SingleCellHaystack (respectively they appear in 10023th, 29191th and 23606th position, while CIARA ranks them at the 64th, 275th and 57th position).

Similarly to SingleCellHaystack, the algorithm SAM18 (Self-Assembling Manifold) does not require any a priori clustering information and it gives as output a list of ranked genes based on an iterative process that rescales the gene expression matrix to emphasize genes that are spatially variable along the intrinsic manifold of the data. This approach, while it offers a valid alternative to standard differential expression analysis methods without relying on cluster analysis, tends to miss signal of rare cells. PGC markers NANOS3, NANOG and DPPA5 are not among the top 1000 genes detected by SingleCellHaystack (respectively they appear in 6325th 8401th and 6790th position).

Overall the CIARA algorithm achieved equal or better performance on several published dataset analyzed in the original publications of GiniClus/2, CellSIUS, FiRE and RaceID and it was the only approach able to detect PGCs.

**CIARA is scalable and fast**

The implementation of the CIARA algorithm in both R and python is highly parallelized in order to minimize the execution time. We test the running time of CIARA (R version) with dataset of several sizes from a few hundred to some thousand cells. The time increase with the size of the dataset and is above 15 minutes for about four thousand cells (**Sup figure 6 a**). Applying CIARA for a big dataset of 68 k peripheral blood mononuclear cells (PBMCs)11 results in a running time of more than 10h. To overcome this limitation, we introduce the option to run CIARA in the approximation mode. The approximation mode focuses on the local region of the cells where the binarized expression of the gene is 1, instead of considering every possible local region (see method). The results obtained with and without approximation are highly consistent, as measured by the Jaccard index between highly localized genes from the two options (**sup figure 7 a**).

With this approximation the running time is extremely reduced and shrinks to less than 4 minutes for the 68K PBMCs (**sup figure 6 a**).

We focus on the results from the PBMCs dataset. In the original paper11 the authorsperformed unsupervised clustering of the cells and annotated the clusters based on previously known markers. The smallest cluster is megakaryocytes (labeled CD34+ in the original paper, 262 cells, 0.3% of the entire dataset) (**sup figure 6 b**).

Consistently among the top highly localized genes from CIARA there are several known markers of megakaryocytes (PF4, GATA1, HBG2) (**sup figure 6 c d e**).

**CIARA allows identification of 4 differentiating cells in mouse embryonic stem cells**

We apply cluster analysis based on CIARA on the mouse embryonic stem cell (mESC) dataset treated with retinoic acid for 24h (766 cells) (**fig 3 a**), after having regressed out the number of umi counts and fraction of mitochondrial reads (see sup method) (**sup figure 3 a b c d**). Retinoic acid is a known inducer of 2-cells like cells4, a rare population of cells presented in mouse ESCs with transcriptional features close to cells at the late two cells stage in mouse embryos20,21.

CIARA identified 3 clusters in total with respectively 744, 18 and 4 cells (**fig 3 b**).

The small cluster of 18 cells has Zscan4 family genes (Zscan4c, Zscan4d, Zscan4b) among its top markers (**fig 3 c**). These are known markers of 2clc3,20.

The third small cluster is made up of only 4 cells. Among the top markers there are both Gata4 and Gata6 (**fig 3 d**). These are two markers of differentiation, suggesting these 4 cells are on the way to exit pluripotency and start differentiating.

To further validate this interpretation, we test the enrichment between the markers of the 4 cells cluster and the markers of the differentiating cells identified in a recently published paper from a mouse embryonic stem cell (mESC) dataset treated with retinoic acid for 48h (**4**).

The intersection reveals to be significantly enriched (**fig 3 e, see method**).

We also apply the other rare cell type detection methods to the mESC dataset treated with retinoic acid for 24h to see if they also were able to recover these two rare subtypes (**sup figure 4 a**). To assess if a are cell types was detected or not, we adopted the same criteria used for PGCs.

The 2clcs (18 cells) are defined as the cells positive for Gm21761, Gm8300, Gm47924 and Gm49858.

The differentiating cells (8 cells) are the ones positive for Col4a1, Col4a2, Lama1 and Dab2.

The four top markers were selected starting from all the markers of the specific cluster with a two-step filtering procedure. First, we keep only the genes with a median above zero in the cluster of interest but equal to zero in all the others. We then sort this reduced list of markers according to the mean expression within the cluster of interest. The top 4 genes with higher mean value are the one used for defining the rare cell type.

FiRE, RaceID, singleCellHaystack and SAM were able to detect the 2clc cluster.

CellSIUS was the only one able to identify the 4-differentiating cells cluster.

GiniClust2/3 did not detect any of the two clusters.

**CIARA enables discovery of new rare cell types in human gastrula**

We apply CIARA on a recently published dataset from a human embryo at the gastrula stage (**10**). This dataset offers a unique opportunity to access the transcriptome profiles of in vivo human gastrulating cells. Running CIARA and consequently cluster analysis using highly localized genes as features allows to identify the PGCs cluster in a completely unsupervised way, as reported above.

Given the uniqueness of the data involved, we want to extract as much information as possible and therefore we asked if other rare cell subtypes, not identified in the original paper, are present in the dataset.

For this task, we follow clustering approach implemented in CIARA (see method) starting from the original annotation provided in the paper.

This results in additional sub-cluster arising from the original endoderm and hemogenic endothelial progenitor clusters, in addition to the identification of the PGC cluster previously described (**fig 4 a b**).

The three new endoderm sub-clusters are made up of respectively 79, 45 and 11 cells.

The first cluster is a mixture of yolk sac endoderm and hypoblast in the original paper annotation, the second is a mixture of DE(NP) and DE (P) and the third is only made up by yolk sac endoderm cells (**sup figure 5 a**).

The smallest cluster shows a very clear transcriptional profile. Among the top markers there are SERPIN family genes (SERPIND1, SERPINC1, SERPINF2) (**fig 4 c**). The diffusion map from endoderm cells reveals that the 11 rare cells are located together at the extreme right side of the trajectory (**fig 4 d**).

We wonder if the same rare endoderm sub type is also present in mouse data. We run CIARA on the endoderm cluster (665 cells) from a scRNA sequencing dataset from mouse at the stage E7.75 E8.25 (post gastrula stage) (**12**). A small cluster with 21 cells, characterized by a very specific transcriptional profile, arises (**fig 4 e**). Interestingly many markers of this small cluster are the same as the human subtype endoderm cluster (in particular SERPINC1, SERPINF2) (**fig 4 f**). The intersection between the two lists of markers is statistically significantly enriched (**fig 4 g**).

Therefore, with CIARA we were able to detect a new endoderm rare subtype at the gastrulation stage that is also present in mice, although in humans we observe it at the gastrulation stage while in mice it is observed at a later stage.

CIARA reveals the presence of 4 additional sub-clusters from the hemogenic endothelial progenitors cluster, made up of respectively 33, 27, 23 and 15 cells. The first cluster correspond mainly to hemogenic endothelium in the original paper annotation, the second is a mixture of myeloid and erythro-myeloid progenitors,the third is mainly erythro-myeloid progenitors and the fourth and the fifth are entirely blood progenitors (**sup figure 5 b**).

The smallest cluster shows a very clear transcriptional profile (**fig 4 h**). The diffusion map from hemogenic endothelial progenitors reveals that the 15 rare cells are located together at a branching point of the trajectory, nicely suggesting that they constitute a new subtype arising at the gastrula stage (**fig 4 i**). A closer look at the top markers indicates that the cells of this rare subtype are megakaryocytes, revealing that hematopoiesis is well undergoing in the human embryo at the gastrula stage. This is a different behavior to what has been observed in mouse, where these rare population (corresponding to 0.1% of the total number of cells) arises at later stages (E6.5-E8.5)24.

We also apply the other rare cell type detection methods to the human gastrula dataset to see if they also were able to recover these two rare subtypes (**sup figure 4 b**). To assess if a rare cell type is detected or not, we adopted the same criteria used for mouse ESC analysis.

The rare cells from the endoderm cluster (13 cells) are defined as the one positive for APOC3, SERPIND1, AMBP and SERPINC1.

The rare cells from hemogenic endothelial progenitors cluster (15 cells) are defined as the ones positive for PPBP, DGKI, HPSE and LTBP1.

FiRE and CellSIUS were the only two able to identify the two rare small clusters.

GiniClust3 was able to detect only endoderm sub types, but not the megakaryocyte. RaceID didn’t detect any of the two rare cell types. In addition, both methods provided an excessive number of final clusters (more than 40), many of them are not well defined according to markers.

SingleCellHaystack and SAM were not able to detect any of the two clusters. This is aligned with the main idea behind these tools (Identifying marker genes with nonrandom expression pattern without relying on cluster annotation, but also without any specific focus towards rare cell types markers.

**Discussion**

Single cell RNA sequencing (scRNA-seq) offers the opportunity to have access to the transcriptome of very high numbers of cells. A common task is to identify clusters of cells based on transcriptome similarities. Standard clustering approaches perform well in identifying common cell types but tend to miss rare cell types. Here we propose CIARA, a cluster independent algorithm that selects genes localized in a small number of neighboring cells from high dimensional PCA space. The selected genes are then easily integrated with common clustering algorithms to identify rare clusters. We showed that CIARA outperforms existing methods for rare cell types detection.

It has the advantage of using a theoretical distribution (hypergeometric) for distinguishing between lowly expressed genes with and without a random pattern. This allows to reduce the number of parameters and consequently reducing the dependence of the algorithm to their default values. Another key point is that the selection of rare cell type markers is done using only a normalized count matrix and a knn matrix as input, without any dependence from cluster partition.

Among common methods for rare cell type detection there are GiniClust2/3, CellSIUS, FiRE, RaceID SinglecellHaystack and SAM.

CIARA was the only method able to identify primordial germ cells cluster (7 cells) from human gastrula data, it easily identified rare cell types from several published datasets also analyzed in the original papers of GiniClust/2, CellSIUS, FiRE, RaceID and allows the identification of two new rare subtypes from endoderm and hemogenic endothelial progenitors in the human gastrula and a group of 4 differentiating cells in mESCs after 24h RA treatment.

Highly parallelized implementation of CIARA allows it to be fast and extremely scalable with the size of the dataset, requiring less than 5 minutes of running time for 68K dataset.

GiniClust2/3 uses the Gini index (a numeric score between 0 and 1) to select relevant rare cell markers, with potential markers having an index close to 1. The Gini index strongly correlates to the maximum expression of the genes. The correction proposed tends to select genes with higher expression than highly localized genes from CIARA and may lead to missing weaker signals (like for PGC markers). A sub-optimal selection of high Gini genes may also cause over-clustering problems, dividing or missing relevant rare cell types (2clc and differentiating cells in mESCs and rare subtypes from hemogenic endothelial progenitors in the human gastrula dataset)

A similar problem in terms of high number of final clusters is common with RaceID, as it identifies more than 40 clusters in human gastrula data (the 7 PGCs are splitted across 5 clusters) and most of them are not well defined according to markers.

RaceID identifies potentially rare cells using outlier genes based on the variance of expression levels. As previously pointed out, this feature selection method is not ideal for capturing rare cell type markers.

singleCellHaystack detects DE genes without relying on cluster analysis. It uses the Kullback-Leibler Divergence to find genes that are expressed in subsets of cells that are non-randomly positioned in a PCA multi-dimensional space or 2D representations. It assigns a significance to genes in comparison with randomization, where expression levels are randomly shuffled over all cells. This correction, as for the Gini Index, may lead to missing genes expressed only in a small number of cells. Indeed, it misses both the two rare sub types and PGCs from the human gastrula dataset and the 4-differentiating cells from the mESCs.

SAM (Self-Assembling Manifold) groups the cells randomly and looks for genes with different expression patterns between the groups. It then uses the resulting information to re-classify the cells and start the process over again, taking the new groups and finding differences between them. While this approach may be beneficial for selecting markers of normal size population of cells, with the advantage of not relying on cluster information, on the other hand markers of rare cell types may be missed and not ranked among the top genes.

CellSIUS and FiRE are the two methods that perform better, after CIARA. They are able to identify the two rare subtypes in human gastrula.

CellSIUS is the only method together with CIARA that detects the 4 differentiating clusters.

FiRE assigns a rareness score to each of the cells using a Sketching technique to measure the density around them. It has the advantage to not rely on clusters, but finding the optimal threshold to label a cell may need to be fine-tuned for different datasets. Default values lead to no rare cells (for Human Gastrula) or to an excessive number of false positives (human cell lines dataset in original FiRE paper).

CellSIUS strongly relies on prior cluster information and this can be an issued especially if there is no clear global structure in the data (as in mESC after 24h RA). It performs sub-cluster using sets of genes as features that shows bi modal distribution within the cluster.

The mean expression in the second mode is then compared to this gene’s mean expression level outside.

To define cluster specific genes, we need to introduce additional parameters. Standard values may miss weaker signals (as for PGCs in human gastrula). On the other hand, lowering the threshold will lead to an excessive number of clusters not well defined according to markers.

**Method section**

**CIARA pipeline**

The input data are the normalized count matrix (using NormalizeData from the R package Seurat19 version 4.0.5) and the K nearest neighbors matrix using the euclidean distance from high dimensional PCA space built on highly variable genes (30 components as a default value, using the FindNeighbors function from the R library Seurat).

CIARA is applied only to lowly expressed genes whose normalized expression is above *threshold* in at least *n\_cells\_low* and in less than *n\_cells\_high* (standard parameter: *threshold*=1, *n\_cells\_low*=3 and *n\_cells\_high*=20).

The expression of each gene across cells is binarized according to the median value.

The cell and its KNN constitute a local region.

For each local region the enrichment in 1 is tested with a fisher test, assuming a hypergeometric distribution as the null distribution (using the fisher.test function from the R package stats, with the parameter ‘alternative’ equal to ‘greater’).

We define a local region enriched in 1 if the fisher test has as output an odds ratio greater or equal to 2 and a p value smaller than 0.001.

For each gene a final p value is also given as an output. If the gene is enriched in at least one local region, then the final p value is the minimum p value given by the fisher test obtained from all the enriched local regions. If the gene is not enriched in any region, then the p value is set to 1 by default.

All the genes that have at least one enriched local region are considered potential markers of rare cell types and kept for downstream analysis. These genes are defined highly localized genes.

**CIARA and cluster analysis**

We proposed the following approach based on 4 steps for integrating CIARA in standard cluster analysis workflow.

The starting point is a general cluster partition of the dataset that can describe the common cell types that are present. In the R version of CIARA global cluster analysis is done on the first 30 PCA components as a default value (defined from top 2000HVGs) with the functions FindNeighbors and FindClusters (from the R library Seurat). An already available partition of the dataset (obtained with other clustering algorithms) can also be used as a starting point.

The second step consist of running the previous clustering algorithm (functions FindNeighbors on top 30 PCA components as a default value and FindClusters) but this time the high dimensional PCA space is built starting from highly localized genes given by CIARA.

Any cluster with a size below a certain threshold (default value is 20) arises as an independent cluster from the global partition of step 1 (step 3).

Finally, in order to detect rare subtypes that do not arise with step 2, a sub cluster analysis is performed starting from only the clusters where the intersection between the top 100 HVGs and the highly localized genes from CIARA is enriched (p value less than 0.05 and odds ratio greater than 1 with the R function fisher.test, using all the genes tested for CIARA as background). The sub cluster analysis is run on the first 30 PCA components as default value (defined from top 2000HVGs) with the functions FindNeighbors and FindClusters (from the R library Seurat). The concept of CIARA is implemented in the user-friendly CIARA package (Cluster Independent Algorithm for the identification of RAre cell types), which is both available for R and python. For more info about the R package including several completely reproducible tutorials see https://github.com/ScialdoneLab/CIARA

The python package integrates into scanpy’s analysis pipeline using the AnnData format. More details on the python package including a reproducible tutorial on how to use the package can be found at [https://github.com/ScialdoneLab/CIARA\_python](https://github.com/ScialdoneLab/IDENTOM_python).

**Analysis of published dataset with CIARA**

**Single-cell RNA-seq dataset from mouse embryonic stem cells (1285 cells)**4.

Running CIARA with standard parameters results in 3302 highly localized genes.

**Single-cell RNA-seq dataset from human embryo at the gastrula stage (1195 cells)**10.

Running CIARA with standard parameters results in 2917 highly localized genes.

As a global cluster analysis, we used the annotation provided in the original paper (step 1)

As for step 2, we run the cluster analysis with the FindNeighbors (on top 30 PCA components with k.param equal to 5) and FindClusters functions (with resolution 0.01).

As for step 4, the sub cluster analysis for the endoderm cluster was run with the FindNeighbors (on top 30 PCA components with k.param equal to 5) and FindClusters functions (with resolution 0.2).

The sub cluster analysis for hemogenic endothelial progenitors was run with the FindNeighbors (on top 30 PCA components with k.param equal to 5) and FindClusters functions (with resolution 0.6).

**Endoderm cluster (665 cells)12 from scRNA-seq dataset from mouse at the stage E7.75 E8.25 (post gastrula stage)**.

Running CIARA (with *n\_cells\_high* equal to 30) results in 1700 highly localized genes.

As for step 2, we run the cluster analysis with the FindNeighbors (on top 30 PCA components with k.param equal to 5) and FindClusters functions (with resolution 0.1).

To assess the significance of the intersection between the markers of the small endoderm cluster from mouse12 and the rare population from endoderm in human gastrula10, a fisher test is performed (using the fisher.test function from the R package stats, with the parameter ‘alternative’ equal to ‘two.sided’).

The markers for the dataset are detected with FindMarkers function (with parameter only.pos = T) from R library Seurat19(version 4.0.5). Only markers with adjusted p value (based on bonferroni correction) below or equal to 0.05 are considered for downstream analysis. Finally, for each cluster, only the markers that are not included in the markers list of other clusters are kept.

For both datasets, the cluster partition used for marker analysis is the result provided by the cluster analysis approach used by CIARA.

**Single-cell RNA-seq dataset associated with leukemia inhibitory factor (LIF) withdrawal-induced mouse embryonic stem cell (mESC) differentiation13**.

We analyzed a subset of 2509 cells obtained from the Day 0 stage (as already done in 17) and a subset from day 4 post-LIF withdrawal (as done in 16).

Running CIARA with the standard parameters for data from Day0 results in 3017 highly localized genes.

As for step 2 of the cluster analysis, we run the cluster analysis with the FindNeighbors (on top 30 PCA components with k.param equal to 3) and FindClusters functions (with resolution 0.1).

Running CIARA with standard parameters for data from Day4 (278 cells) results in 287 highly localized genes.

As for step 2 of the cluster analysis, we run the cluster analysis with the FindNeighbors (on top 30 PCA components with k.param equal to 5) and FindClusters functions (with resolution 0.3).

**Single-cell RNA-seq dataset obtained from glioblastoma (GBM) primary tumors (576 cells)**14.

This dataset was also studied in the GiniClust paper (**17**) and we perform the same filtering step as done in the original publication.

Running CIARA with the standard parameters results in 68 highly localized genes.

As for step 2 of the cluster analysis, we run the cluster analysis with the FindNeighbors (on top 30 PCA components with k.param equal to 3) and FindClusters functions (with resolution 0.1).

**Single-cell RNA-seq data comprising 293T and Jurkat cells mixed in vitro (1580 cells)**6**.**

Running CIARAwith standard parameters results in 2077 highly localized genes.

As for step 2 of the cluster analysis, we run the cluster analysis with the FindNeighbors (on top 30 PCA components with k.param equal to 3) and FindClusters functions (with resolution 0.1).

This dataset was already analyzed in the original FiRE paper6.

**Single cell RNA-seq dataset with known cellular composition generated from mixtures of eight human cell lines (3984 cells)**7.

Running CIARAwith standard parameters results in 3704 highly localized genes.

As for step 2 of the cluster analysis, we run the cluster analysis with the FindNeighbors (on top 30 PCA components with k.param equal to 3) and FindClusters functions (with resolution 0.1).

This dataset was already analyzed in the original CellSIUS paper7.

**Single-cell RNA seq data from murine intestinal epithelial cells (317 cells) (14), already analyzed by RaceID**15**.**

Running CIARA with standard parameters results in 1514 highly localized genes.

As for step 2 of the cluster analysis, we run the cluster analysis with the FindNeighbors (on top 20 PCA components with k.param equal to 3) and FindClusters functions (with resolution 0.2).

Since many of the marker genes mentioned in 14are expressed at a level greater than zero by almost all the cells, we adopt a slightly different approach in comparison to what we do with the other published dataset to assess if a cell type is found with CIARA.

We assume that a cell type is detected by CIARA if the cell type marker reported in the original paper appears in the list of markers of a cluster.

Markers analysis is done with FindMarkers function (with parameter only.pos = T) from R library Seurat19(version 4.0.5). Only markers with adjusted p value (based on bonferroni correction) below or equal to 0.05 are considered for downstream analysis. Finally, for each cluster, only the markers that are not included in the markers list of other clusters are kept.

The cluster partition used for marker analysis is the result provided by the cluster analysis approach used by CIARA on the murine intestinal epithelial cells

**Single-cell RNA seq data RNA-seq data containing ~68 k peripheral blood mononuclear cells**11 (already analyzed by 16 and 6).

Running CIARA(with *n\_cells\_high* equal to 100) results in 4207 highly localized genes.

A summary of all the analyzed datasets with parameters used for running CIARA and cluster analysis in step b is shown in **supplementary figure 7 b**.

**Analysis of mouse ESC after 24h RA**

Cells that pass the QC filter are kept for down-stream analysis (766 cells). Global cluster analysis is done on the first 30 PCA components (defined from top 2000HVGs) with the function FindClusters (from R library Seurat) with a resolution of 0.1.

This results in two clusters (0 and 1) that differ only in the quality of cells (cluster 0 has very low level of UMI counts, number of genes expressed and high level of fraction of mitochondrial reads).

Therefore the cluster analysis was repeated, regressing out the number of umi counts and the fraction of mitochondrial and ribosomal reads (using the parameter vars.to.regress in the function ScaleData from the R library Seurat). This leads to a unique final cluster.

Running CIARA results in 2475 highly localized genes.

As for step 2, we run the cluster analysis with the FindNeighbors (on top 30 PCA components with k.param equal to 3) and FindClusters functions (with resolution 0.1).

To assess the significance of the intersection between the markers of the 4 cells cluster and the markers of the differentiating cells identified in mESC dataset treated with retinoic acid for 48h4, a fisher test is performed (using the fisher.test function from the R package stats, with the parameter ‘alternative’ equal to ‘two.sided’).

The markers for the two mESCs dataset treated with retinoic acid (RA) for 2h and 48 hours are detected with FindMarkers function (with parameter only.pos = T) from R library Seurat19(version 4.0.5). Only markers with adjusted p value (based on bonferroni correction) below or equal to 0.05 are considered for downstream analysis. Finally, for each cluster, only the markers that are not included in the markers list of other clusters are kept.

For mESCs dataset treated with RA for 24h, the cluster partition used for marker analysis is the result provided by the cluster analysis approach used by CIARA.

For mESCs dataset treated with RA for 48h, the cluster partition used for marker analysis is the one provides in4.

**Analysis of human gastrula and mouse ESC after 24h RA datasets with previously published methods**

GiniClust216 and GiniClust35 pipelines were used following the standard procedure (<https://github.com/dtsoucas/GiniClust2> and <https://github.com/rdong08/GiniClust3>), for the human gastrula dataset and for mouse ESCs RA after 24h dataset.

CellSIUS analysis7 is based on the corresponding R package (<https://github.com/Novartis/CellSIUS/>), setting the “min\_n\_cells” parameter for the CellSIUS function equal to 5 for the human gastrula data and equal to 3 for the mouse ESC data.

FiRE6 was run with the standard parameters provided in the R package (<https://github.com/princethewinner/FiRE>) for mouse ESC data and with a threshold on the FiRE score of 0.5\*interquartile range + third quantile for human gastrula data for human gastrula (with the standard parameters no cells were labeled as rare).

RaceID 38 analysis was performed setting the number of clusters equal to 5 in the clustexp function from the related R package (<https://github.com/dgrun/RaceID3_StemID2_package>) for mouse ESC and with the standard parameters for the human gastrula data.

SingleCellHaystack9 was in the standard mode on PCA space with 30 dimensions following the corresponding R package (<https://github.com/alexisvdb/singleCellHaystack>).

A rare cell type is defined as the set of cells positive for four top markers. The four top markers were selected starting from all the markers of the specific cluster with a two-step filtering procedure. First, we keep only genes with a median above zero in the cluster of interest but equal to zero in all the others. We then sort this reduced list of markers according to the mean expression within the cluster of interest. The top 4 genes with higher mean value are the ones used for defining the rare cell type.

We assume that a rare cell type detected with CIARA is also found with another method if the Jaccard index between the rare cell type and one of the clusters provided by the method is greater or equal to 0.5 (for RaceID, CellSIUS and GiniClust3).

For FiRE, which assigns each cell 0 (no rare type) or 1 (rare type), there is agreement with the CIARA method if a rare cell type is enriched in 1 (according to a Fisher test with the R function fisher.test).

For SingleCellHaystack and SAM, which return a ranked list of genes according to their expression pattern in high dimensional space, there is agreement with the CIARA algorithm if the markers of the rare cell type clusters are enriched in the top 1000 SingleCellHaystack/SAM genes.

The marker of rare cell type clusters is done with FindMarkers function (with parameter only.pos = T) from R library Seurat19(version 4.0.5). Only markers with adjusted p value (based on bonferroni correction) below or equal to 0.05 are considered for downstream analysis. Finally, for each cluster, only the markers that are not included in the markers list of other clusters are kept.

For both mouse ESCs and human gastrula dataset, the cluster partition used for marker analysis is the result provided by the four steps cluster analysis approach used by CIARA.

**References**

1) Haghverdi, L., Lun, A., Morgan, M. *et al.* Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat Biotechnol* **36,**421–427 (2018). https://doi.org/10.1038/nbt.4091

2)Brandani et al, Plos One, 2013 https://doi.org/10.1371/journal.pone.0065617

3) Macfarlan, T., Gifford, W., Driscoll, S. *et al.* Embryonic stem cell potency fluctuates with endogenous retrovirus activity. *Nature* **487,**57–63 (2012). https://doi.org/10.1038/nature11244

4) Iturbide, A., Ruiz Tejeda Segura, M.L., Noll, C. *et al.* 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

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

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

7) Wegmann, R., Neri, M., Schuierer, S. *et al.* CellSIUS provides sensitive and specific detection of rare cell populations from complex single-cell RNA-seq data. *Genome Biol***20,**142 (2019). https://doi.org/10.1186/s13059-019-1739-7

8) Herman, J., Sagar & Grün, D. FateID infers cell fate bias in multipotent progenitors from single-cell RNA-seq data. *Nat Methods* **15,**379–386 (2018).

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

10) Tyser, R.C.V., Mahammadov, E., Nakanoh, S. *et al.* Single-cell transcriptomic characterization of a gastrulating human embryo. *Nature* (2021). https://doi.org/10.1038/s41586-021-04158-y

11) Zheng, G., Terry, J., Belgrader, P. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nat Commun* **8,**14049 (2017). <https://doi.org/10.1038/ncomms14049>

12) Tyser RCV, Ibarra-Soria X, McDole K, Arcot Jayaram S, Godwin J, van den Brand TAH, Miranda AMA, Scialdone A, Keller PJ, Marioni JC, Srinivas S. Characterization of a common progenitor pool of the epicardium and myocardium. Science. 2021 Mar 5;371(6533):eabb2986. doi: 10.1126/science.abb2986. Epub 2021 Jan 7. PMID: 33414188.

13) Klein AM, Mazutis L, Akartuna I, Tallapragada N, Veres A, Li V, Peshkin L, Weitz DA, Kirschner MW. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell. 2015 May 21;161(5):1187-1201. doi: 10.1016/j.cell.2015.04.044. PMID: 26000487; PMCID: PMC4441768.

14) Patel AP, Tirosh I, Trombetta JJ, Shalek AK, Gillespie SM, Wakimoto H, Cahill DP, Nahed BV, Curry WT, Martuza RL, Louis DN, Rozenblatt-Rosen O, Suvà ML, Regev A, Bernstein BE. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. Science. 2014 Jun 20;344(6190):1396-401. doi: 10.1126/science.1254257. Epub 2014 Jun 12. PMID: 24925914; PMCID: PMC4123637.

15) <https://doi.org/10.1016/j.stem.2016.05.010>

16) Tsoucas, D., Yuan, GC. GiniClust2: a cluster-aware, weighted ensemble clustering method for cell-type detection. *Genome Biol* **19,**58 (2018). <https://doi.org/10.1186/s13059-018-1431-3>

17) Jiang, L., Chen, H., Pinello, L. *et al.* GiniClust: detecting rare cell types from single-cell gene expression data with Gini index. *Genome Biol* **17,**144 (2016). https://doi.org/10.1186/s13059-016-1010-4

18) eLife 2019;8:e48994 DOI: [10.7554/eLife.48994](https://doi.org/10.7554/eLife.48994)

19) Hao and Hao et al. Integrated analysis of multimodal single-cell data. Cell

  (2021) [Seurat V4]

20) Rodriguez-Terrones, D. et al. A molecular roadmap for the emergence of early-embryonic-like cells in culture. *Nat. Genet.* **50**, 106–119 (2018).

21) Wolf, F., Angerer, P. & Theis, F. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol* **19,**15 (2018). <https://doi.org/10.1186/s13059-017-1382-0>

22) Dye FS, Larraufie P, Kay R, Darwish T, Rievaj J, Goldspink DA, Meek CL, Middleton SJ, Hardwick RH, Roberts GP, Percival-Alwyn JL, Vaughan T, Ferraro F, Challis BG, O'Rahilly S, Groves M, Gribble FM, Reimann F. Characterisation of proguanylin expressing cells in the intestine - evidence for constitutive luminal secretion. Sci Rep. 2019 Oct 30;9(1):15574. doi: 10.1038/s41598-019-52049-0. PMID: 31666564; PMCID: PMC6821700.

23) Gerbe, F., Legraverend, C. & Jay, P. The intestinal epithelium tuft cells: specification and function. *Cell. Mol. Life Sci.* **69,**2907–2917 (2012). <https://doi.org/10.1007/s00018-012-0984-7>

24) Pijuan-Sala, B., Griffiths, J.A., Guibentif, C. *et al.* A single-cell molecular map of mouse gastrulation and early organogenesis. *Nature* **566,**490–495 (2019). https://doi.org/10.1038/s41586-019-0933-9