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

2022/08/15

Ping-Han Hsieh

Comparison

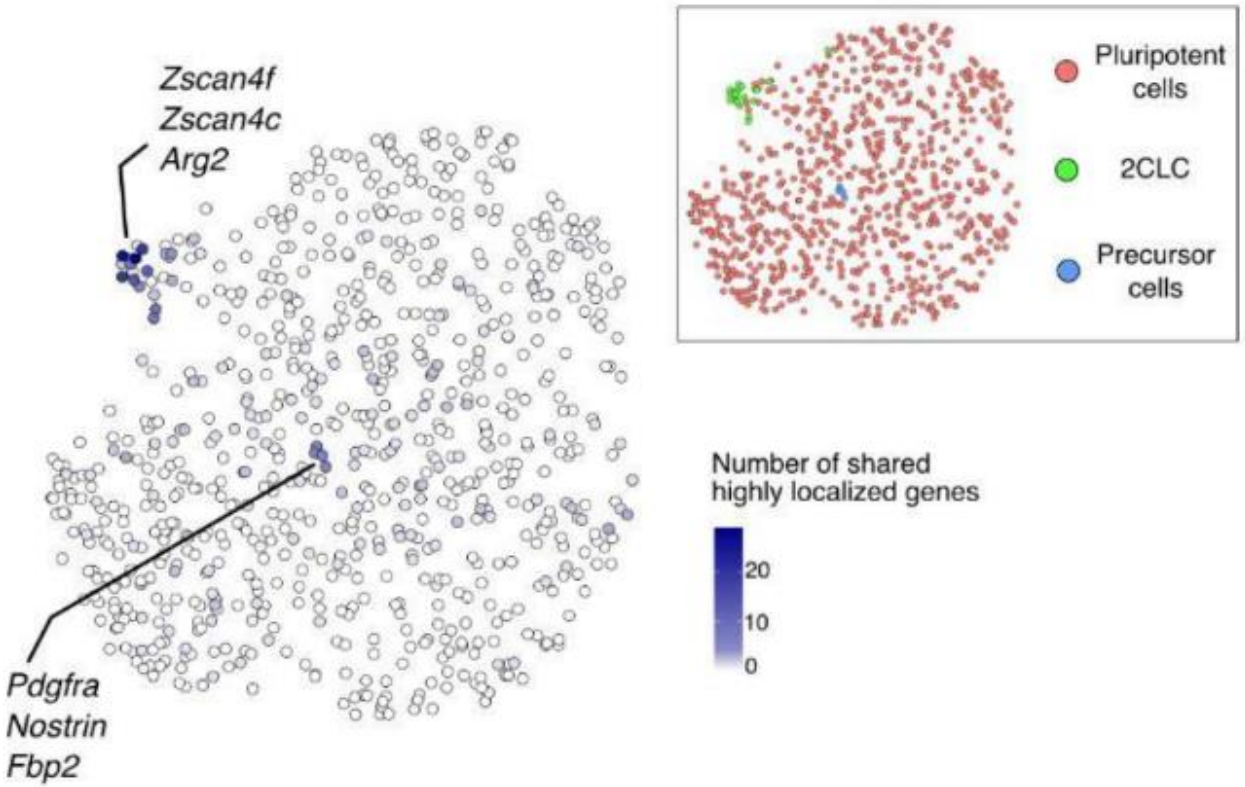
- Compare all algorithms on scRNA-Seq data from human gastrula (1,195 cells).
  - 7 primordial germ cells (PGCs) is included
    - Have markers in common with other cell types (SOX17, ETV4), makes it more difficult to identify with unsupervised methods

Biological Insights (1)

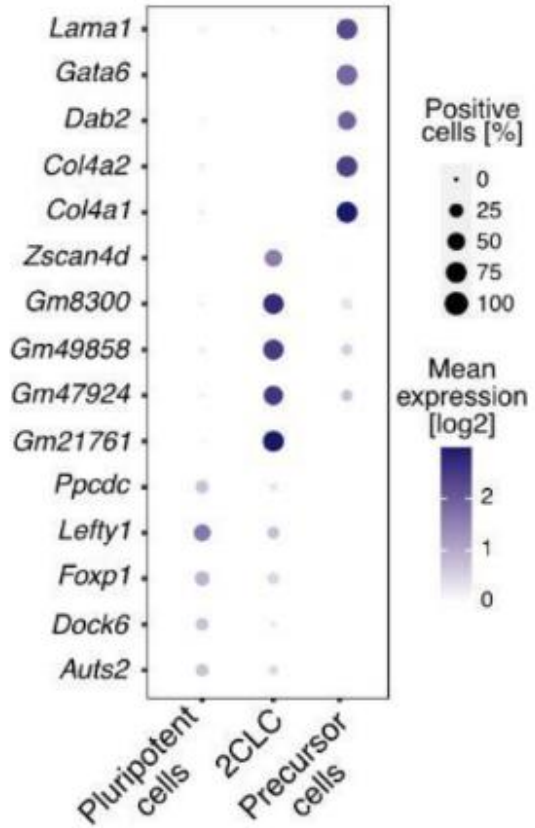
- Previous study showed that a 48h-long treatment with low doses of retinoic acid (RA) induces the reprogramming of mouse embryonic stem cells (mESCs)
- It is not known how long the RA treatment must be to produce any effects on cell fate decisions.
- Generated a new scRNA-Seq dataset from mouse embryonic stem cells (mESCs) following a 24h low doses of retinoic acid (RA) treatment.
  - CIARA detected a small group of 4 cells (<1%) marked by a distinct set of genes (*Gata4, Gata6*)
  - This small cluster with 4 precursor cells are compatible with those found at 0h and 48h of treatment.
  - Indicate that during the first 48h of RA treatment, the same cell types are present.

Biological Insights (1)

UMAP representation



Top marker genes



Biological Insights (1)

- Analysis
  - Identify 2475 highly localized genes.
  - Cluster analysis with FindNeighbors (k=3) and FindClusters (resolution 0.1) function, which gave 3 clusters.
  - Marker genes were detected with the FindMarkers function. Only markers with Bonferroni corrected p-value velow or equal to 0.05 are considered.
  - Three clusters are identified as Pluripotent Cells, 2-cell-like cells and Precursor Cells.

  - Compared the clusters with previously published mESC datasets after 0h and 48h RA treatment. Re-analyzed with CIARA, identified 3302 highly localized genes.
  - Performed cluster analysis with FindNeighbors (k=5) and FindClusters (resolution 0.1).

  - Assess the statistical significance of the intersection between the markers of the three clusters found at 0h, 24h, and 48h using Fisher's exact test (significant for all pairs).

Biological Insights (2)

- scRNA-seq dataset from a human gastrula.
  - In addition to PGCs
  - CIARA identified two small populations in the Yolk Sac Endoderm (YSE) and the Megakaryocyte-Erythroid Progenitors (MEP) clusters.
  - Small YSE subcluster (YSE1)
    - 11 cells with very specific markers (*SERPIND1, SERPINC1*, known to be expressed in the adult kidney and liver) * yolk sac functions during early development.
    - Small cluster of 21 endodermal cells with the same transcriptional profile in mouse embryos at the E7.75-E8.25 stage. * YSE1 is a relatively rare endodermal sub-population
    - YSE1 is more transcriptionally distinct from the embryonic endoderm populations than the reset of the YSE cluster (Figure 2f)
    - YSE includes cells from the embryonic disk, cells included in YSE1 only come from the yolk sac region (might be those located further away from the embryonic disk, possibly closer to the forming blood islands where primitive erythropoiesis occurs).

Biological Insights (2)

- scRNA-seq dataset from a human gastrula.
  - Megakaryocyte-Erythroid Progenitors (MEP) clusters (MEP1).
    - 13 cells with distinct transcriptional signature characterized by high levels of markers (*PPBP, ITGA2B* and *GP1BB*). Based on these markers, we could identify these cells as megakaryocytes.
    - Supported by differentiation trajectories analysis within the blood clusters.
      - Branch event where the MEP cluster splits into the MEP1 cluster and erythroblasts (Figure 2i)
      - Identify genes marking the differentiation between the two cells (Figure 2j)

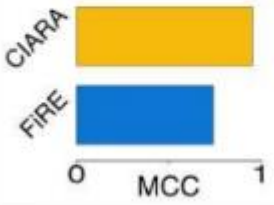
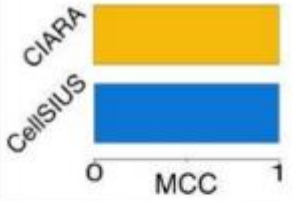
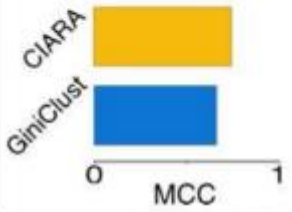
Biological Insights (2)

- Rare Cell Types Analysis
  - Test the enrichment of the 2917 highly localized genes found by CIARA among the top 100 highly variable genes within each of the clusters provided in previous research.
  - Found significant overlap in the Endoderm (Endo), Haemato-Endothelial Progenitors (HEP) cluster. Two smallest clusters found in the Endo and HEP clusters are denoted as YSE1 and MEP1.

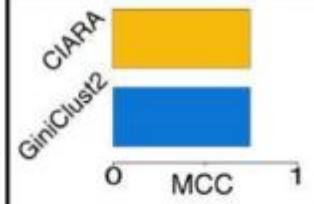
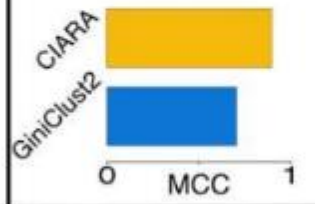
Main Analysis Workflow

1. Run CIARA to identify highly localized genes.
2. Perform cluster analysis using
    1. Enrichment analysis with highly variable genes (HVG) of known clusters*.
    2. FindNeighbors and FindClusters functions.
3. Perform marker genes analysis using FindMarkers function.
    - Only unique and significant markers are considered for downstream analysis.
4. Perform additional analysis based on the clusters or markers.

Comparison with Other Methods

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	

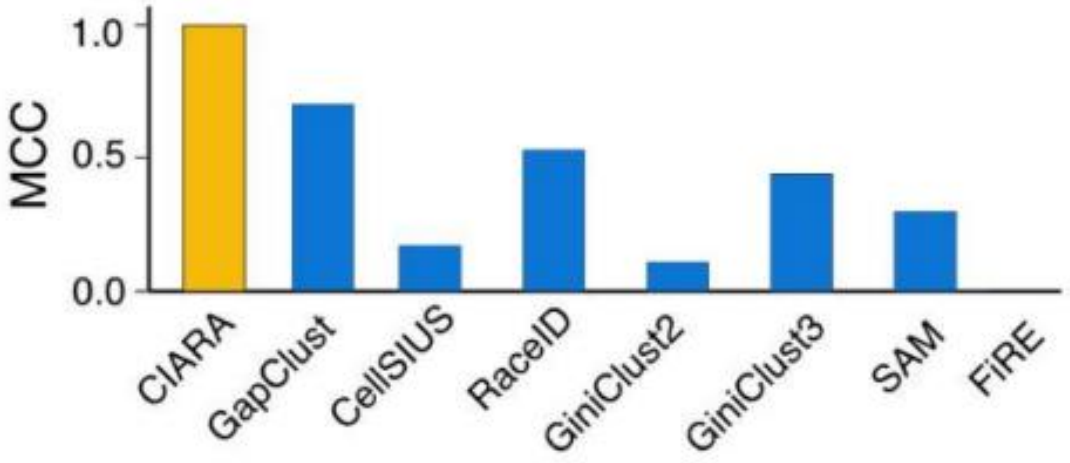
1. identify 2 clusters, FP: 4 cells vs 32 cells
2. Identify 9 clusters (2)
3. Identify 13 clusters (1)

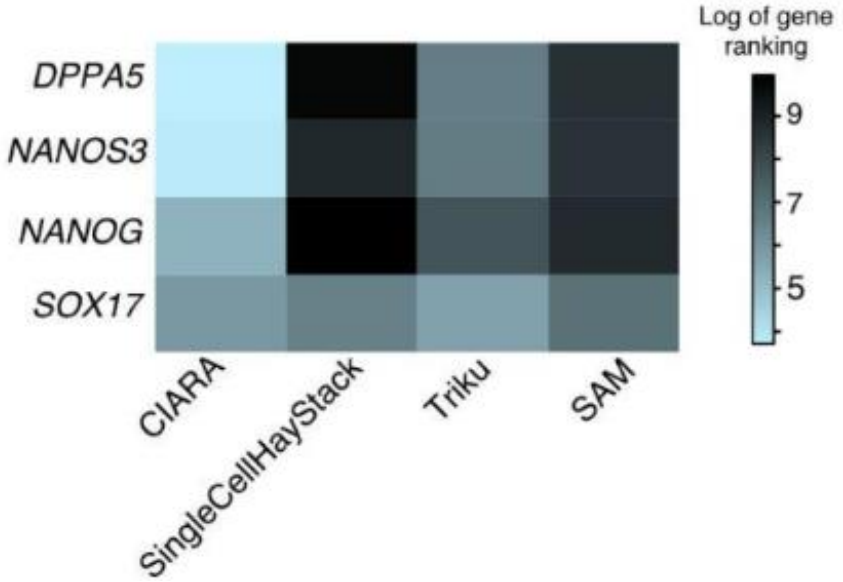
Method	Dataset	Comparison with CIARA
GiniClust2	278 mouse embryonic stem cells with rare populations of primitive endoderm (9 cells) and cells expressing maternally imprinted genes (8 cells)	Primitive Endoderm cells  Cells expressing maternally imprinted genes 
RaceID	317 murine intestinal epithelial 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 (Supp. Fig. 2)

4. identify 3 clusters (2)
5. Identify 8 clusters (5)

Human Gastrula (Primordial Germ Cell)

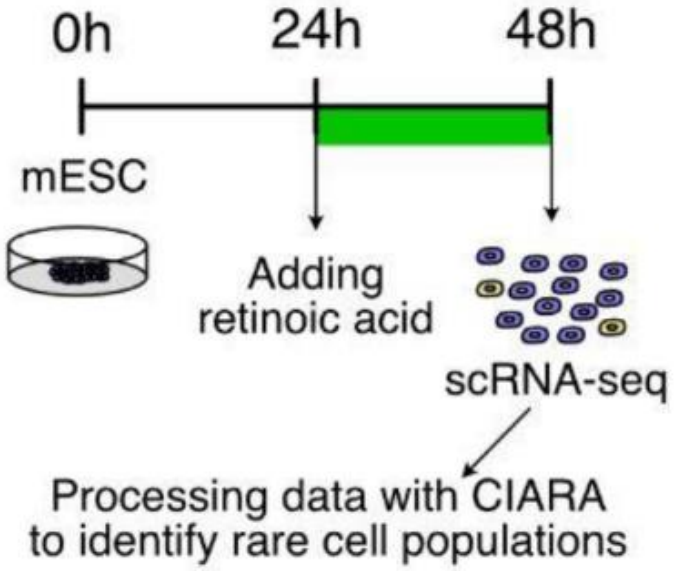
- External dataset to compare with all other methods.
- Validation data: small population of seven Primordial Germ Cells (PGCs) identified in previous study using supervised way.





Mouse Embryonic Stem Cell (1)

- Previous study showed that a 48h-long treatment with low doses of retinoic acid (RA) induces the reprogramming of mouse embryonic stem cells (mESCs)
- It is not known how long the RA treatment must be to produce any effects on cell fate decisions.
- Generated a new scRNA-Seq dataset from mouse embryonic stem cells (mESCs) following a 24h low doses of retinoic acid (RA) treatment (Page 12).

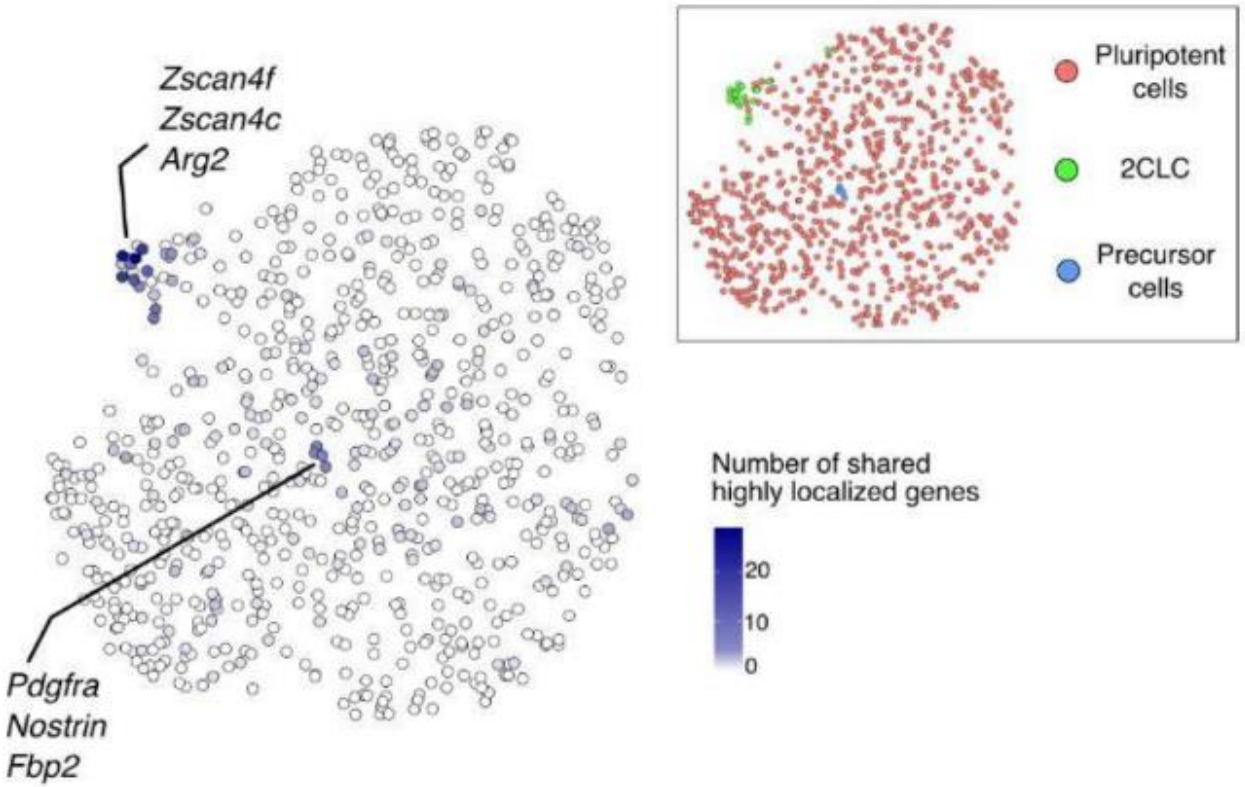


Mouse Embryonic Stem Cell (2)

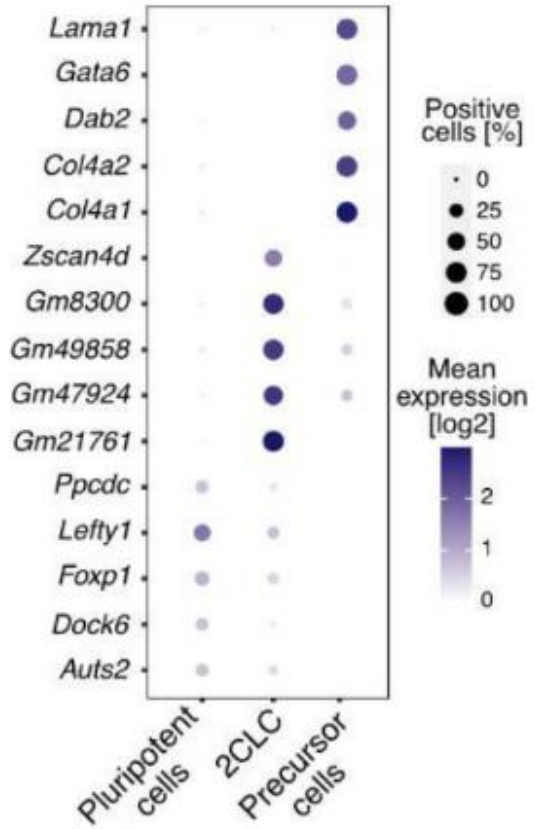
- Analysis Result
  - CIARA identified:
    - (00h) 3302 highly localized genes.
    - (24h) 2475 highly localized genes .
    - (48h) dataset was not reanalyzed.
  - Cluster analysis:
    - (00h) Pluripotent Cells, 2CLC and Precursor cells clusters based on markers.
    - (24h) Pluripotent Cells, 2CLC and Precursor cells clusters based on markers.
    - (48h) Pluripotent Cells, 2CLC and Precursor cells clusters from previous study.
    - Precursor cells clusters contains 4 cells, with differentiation markers (*Gata4, Gata6*)
    - Fisher exact test suggests significant intersection between the cell types across datasets.
- Conclusions
  - During the first 48h of RA treatment, the same cell types are present.

Mouse Embryonic Stem Cell (3)

UMAP representation



Top marker genes

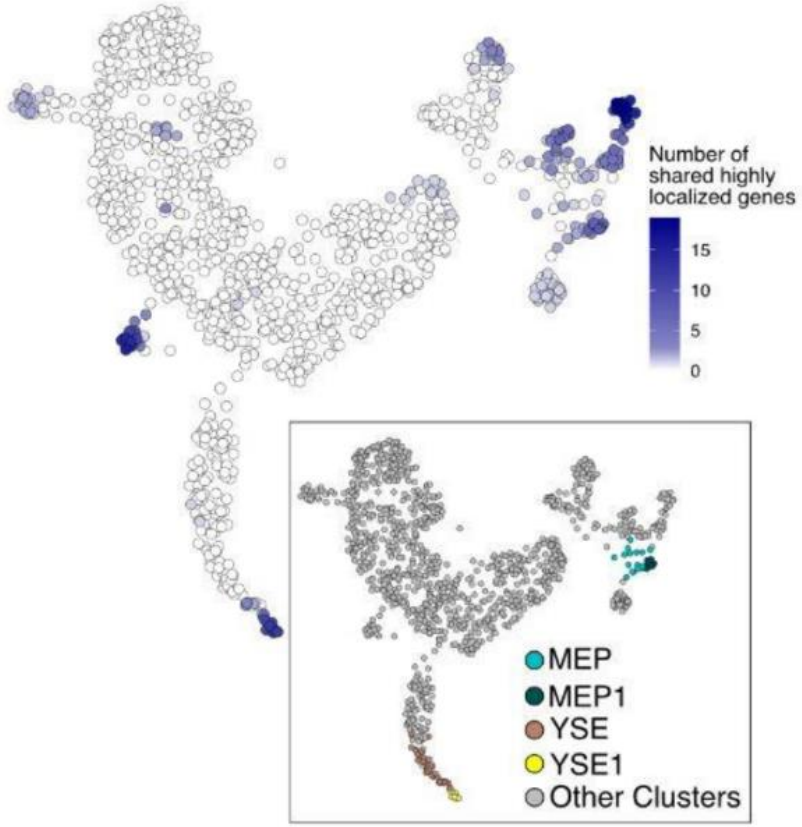


Human Gastrula (Rare Cell Types) (1)

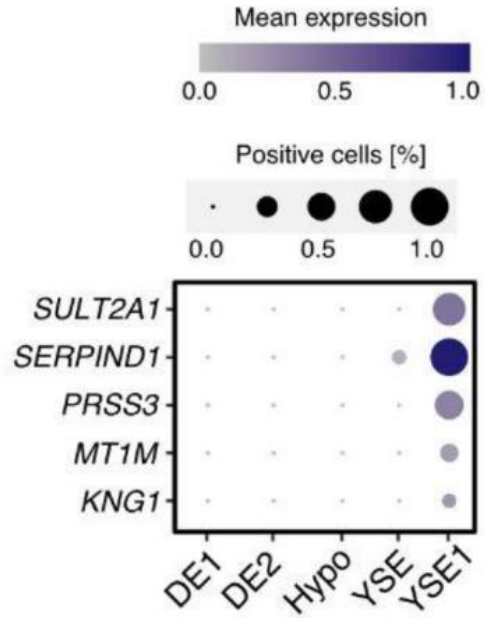
- Analysis Result
  - CIARA Identify 2917 highly localized genes.
  - Significant overlap with the HVGs in Endoderm (Endo) and Haemato-Endothelial Progenitors (HEP) clusters.
  - Subclustered the Endo and HEP into YSE1 and MEP1.
  - Marker analysis:
    - Small YSE subcluster (YSE1)
      - 11 cells with very specific markers (*SERPIND1, SERPINC1*, known to be expressed in the adult kidney and liver)
    - Small Megakaryocyte-Erythroid Progenitors subcluster (MEP1)
      - 13 cells with distinct transcriptional signature characterized by high levels of markers (*PPBP, ITGA2B* and *GP1BB*).

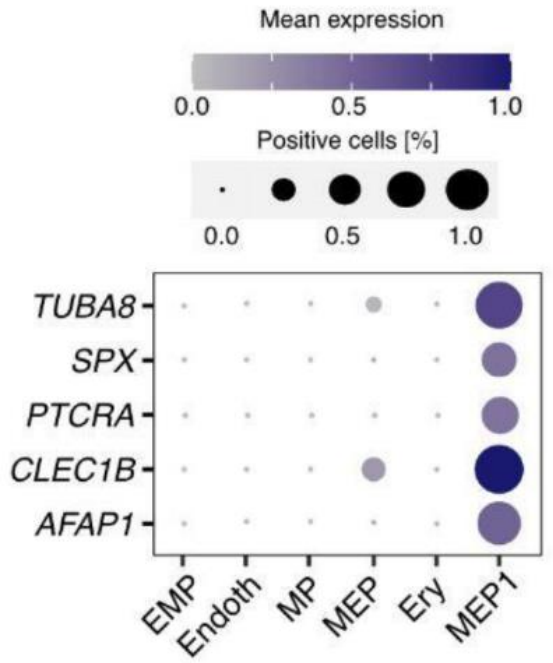
Human Gastrula (Rare Cell Types) (2)

UMAP representation



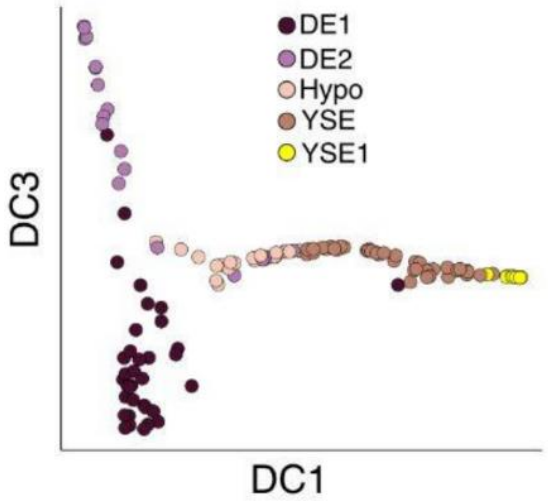
Top marker genes

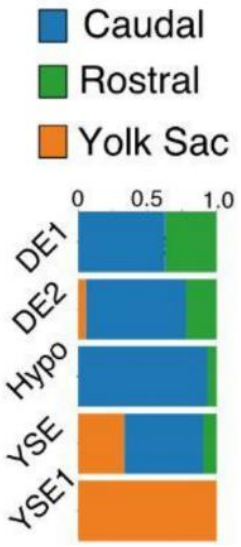


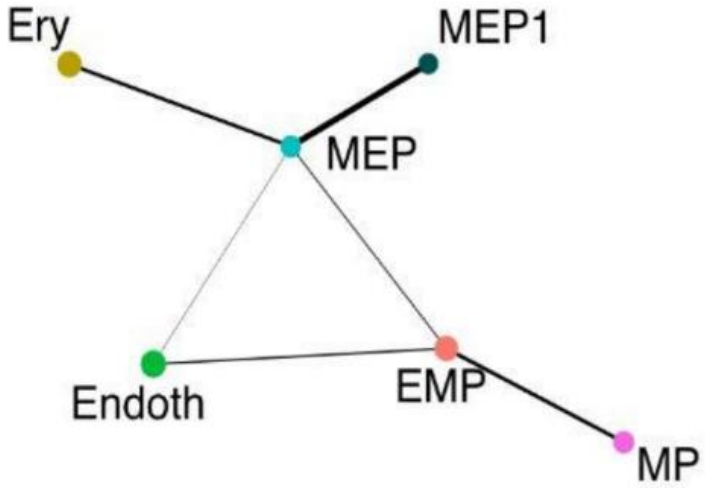


Human Gastrula (Rare Cell Types) (3)

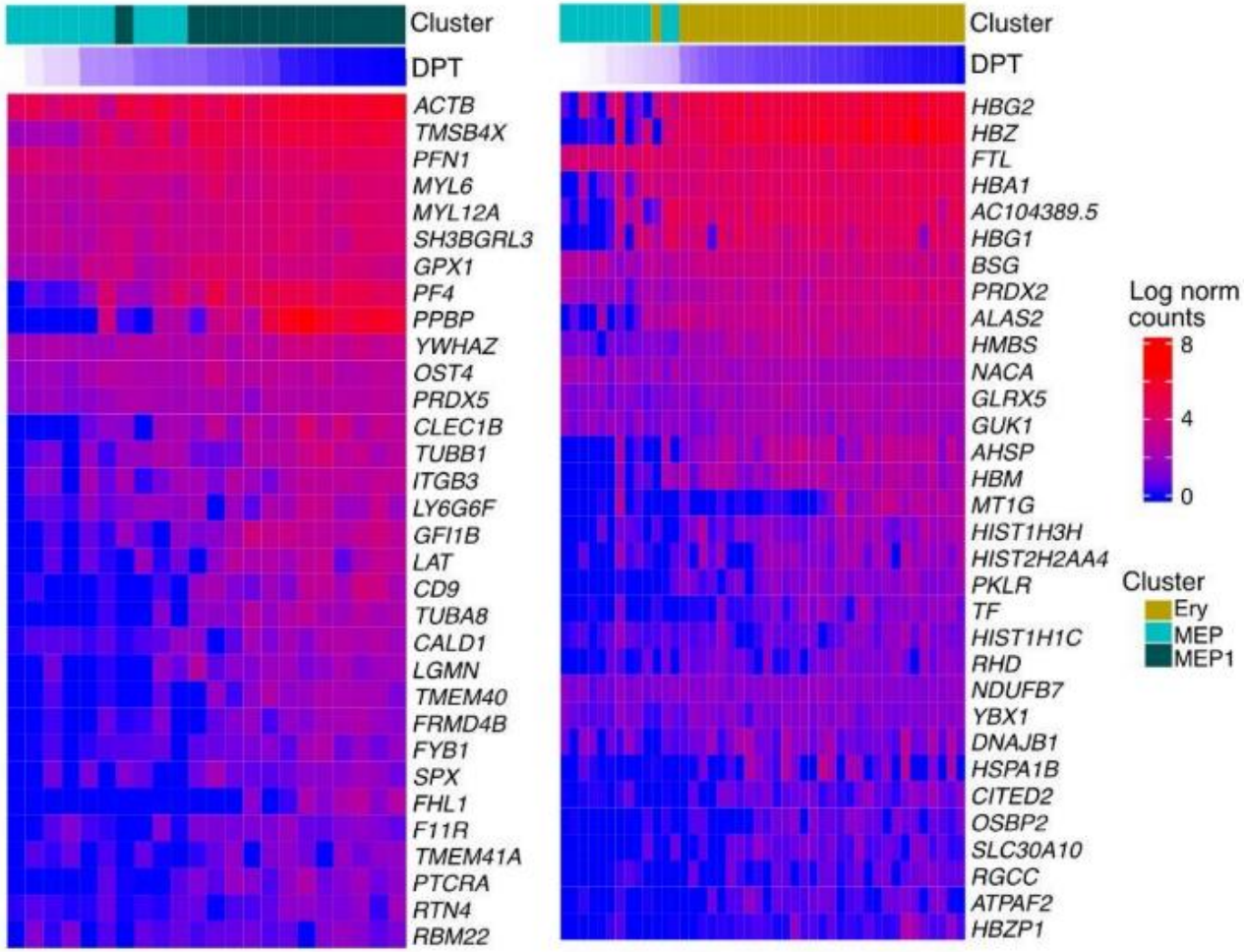
- Trajectory
  - Diffusion map for Endo sub-clusters (DE1, DE2, YSE, Hypoblast, YSE1)
  - Trajectory analysis for HEP sub-clusters (EMP, HE, MP, MEP, MEP1)







Human Gastrula (Rare Cell Types) (4)



Human Gastrula (Rare Cell Types) (5)

- Conclusion
  - Identify branch event where the MEP cluster splits into the MEP1 cluster and erythroblasts.
  - Identify genes marking the differentiation between the two cells.