Gene regulatory networks inference in the pigs embryos from scRNAseq and scMulti-omics

Adrien Dufour, Yoann Bailly, Patrick Manceau, Stéphane Ferchaud, Sylvain Foissac, Jérome Artus, Marie-José Mercat, Hervé Acloque

Gene regulatory networks inference in the pigs embryos from scRNAseq and scMulti-omics

Adrien Dufour[1], Yoann Bailly[2], Patrick Manceau[2], Stéphane Ferchaud[2], Sylvain Foissac[3], Jérome Artus[4], Marie-José Mercat[5] and Hervé Acloque[1]

1 Paris-Saclay University, INRAE, AgroParisTech, GABI, Jouy-en-Josas, France ; 2 INRAE, GenESI, La Gouvanière, Rouillé, France ; 3 Université de Toulouse, INRAE, ENVT, GenPhySE, Chemin de Borde Rouge, 31326 Castanet-Tolosan, France ; 4 INSERM U1310, Paris Saclay University, Villejuif, France ; 5 IFIP - Institut du Porc, Paris, France
*herve.acloque@inrae.fr

INTRODUCTION

The single-cell approach allowed the study of tissue at the cell scale and allowed the identification of the different cell populations within. The regulatory genomics studies from those data have then been widely developed in the last years with methods such as SCENIC1. Recently the emergence of new methods of single-cell multi-omics (scRNAseq and scATACseq combined) allowed new possibilities.

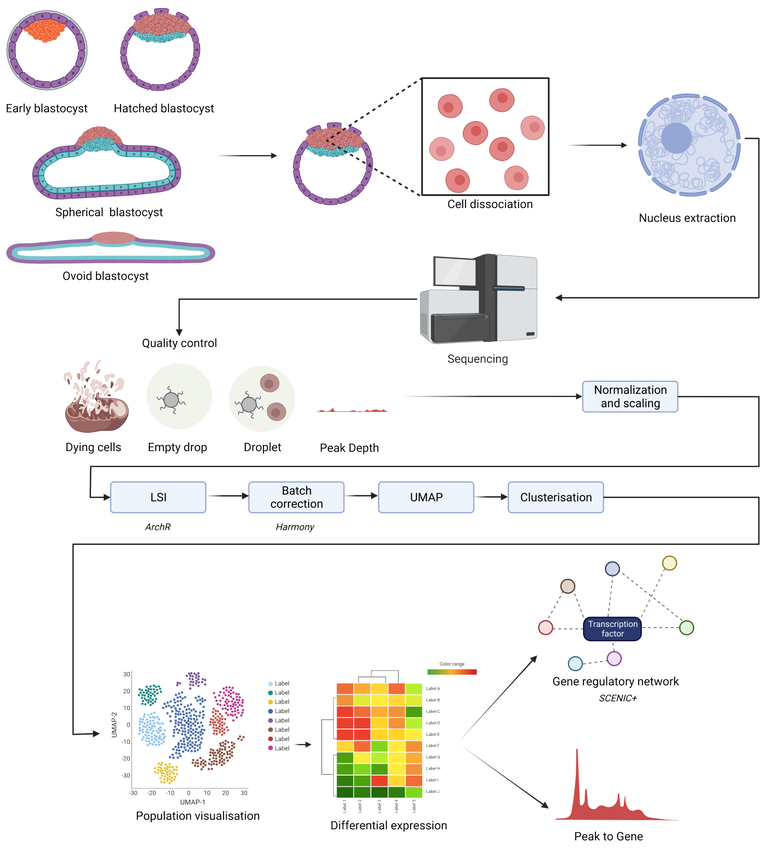
I will take the example of my PhD work on pig pre-implantation embryos. To better understand the pig embryos we have produced a large dataset of single-cell RNAseq and single-cell multi-omics at different embryonic states (early (E5), expanded (E7), spherical (E9) and ovoid blastocysts (E11)) using the Chromium 10x Genomics technology. Comparisons will be performed with similar publications in pigs and humans.

METHODS

scRNAseq datasets were cleaned, filtered and represent a total of 35,000 cells. Pre-processing steps including normalisation, batch removal and principal component analysis were performed for each state using Seurat [1] and Harmony [2]. Identification of the different clusters was made from known markers from the literature. Differential expression analysis was performed on those lineages using the Seurat ConservedMarker function. Regulon inference was performed using SCENIC [3] with a pig-based database : the transcription factor motif association was performed by orthology between human and pig genes. Afterwards, a pig genome ranking was calculated by scoring putative regulatory regions for the presence of homotypic clusters of motif instances. Those analyse has currently available as a preprint [4].

Single-cell multiomics datasets (coupled RNAseq & ATACseq) at different embryonic states (hatched, spherical and ovoid blastocysts) were also produced. Quality controls and pre-processing steps (peak calling, normalisation, batch correction) were completed through the ArchR package [5] and five us 18,000 cells. Then the cluster assignment was performed with an alignment on our previous scRNAseq datasets using the ArchR function. Then an eRegulon inference was performed using SCENIC+ [6]. To confirm those regulons a motif enrichment analysis and a motif footprint research were performed with ArchR.

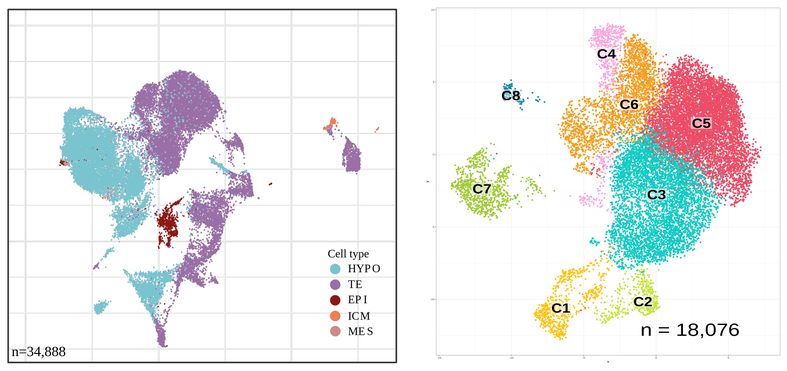
Meta-analysis The SCENIC pipeline was applied to two other single-cell RNAseq datasets of pre-implantation embryos in the pig [7] and in humans [8]. The identified transcription factor was compared with a upsetplot and a Jaccard index was calculated for each common regulons.



Schema of preprocessing of single-cell multiomics data

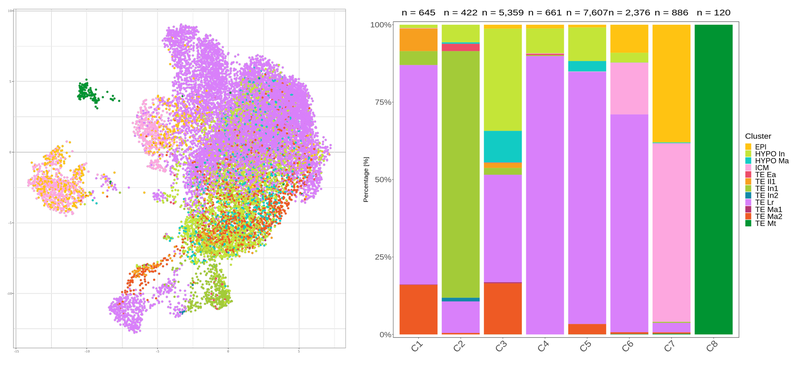
RESULTS

I Identification of cells population



UMAP of cells from scRNAseq (left) and scMultiomics (right)

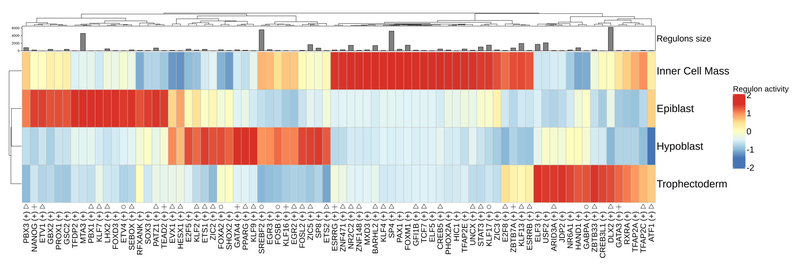
In our scRNAseq datasets, we identified a trophectoderm (TE, in violet), epiblast (EPI, in red) and hypoblast (HYPO, in cyan). We also identified time-specific populations with inner cell mass (ICM in orange) at the stage (E5) and mesendoderm (MES in dark red) at the stage (E11). Those populations have then been refined into more precise subpopulations. The multiomics clusterisation provides 8 clusters, where we identified hypoblast (C3, C4, C5, C6), trophectoderm (C1, C2, C8) and epiblast (C7). We can observe that hypoblast cells have replaced trophectoderm as hypoblast represent 45% of the cells in scRNAseq and 88% in scMultiomics.



UMAP of cells labels assignation from scRNAseq datasets (left) Histogramm of repartition of cells attribution in each scMulitomics cluster (right)

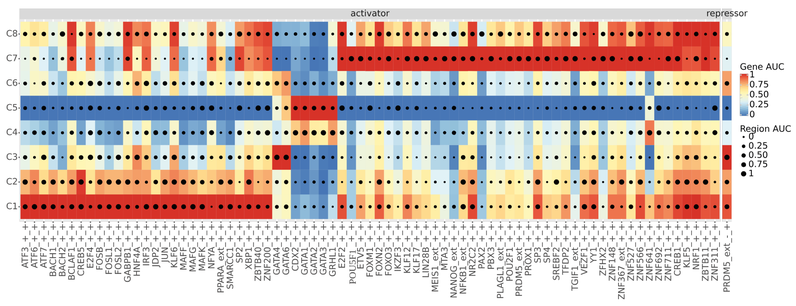
We observe that cluster C7 have been successfully attributed to the EPI/ICM populations, the cluster C8 to the TE Mt populations. Clusters C1 has been mainly assigned to the TE Lr and cluster C2 to the TE In1. Identifying of the hypoblast populations remains challenging as only some cells have been assigned to those clusters. Visualisation of known marker genes of these populations helps us assign cluster C3-6 to hypoblast even though those clusters have been assigned to the TE Lr populations but have a low assignation score.

II Gene regulatory network



Heatmap of the regulons activity by cell populations in scRNAseq data

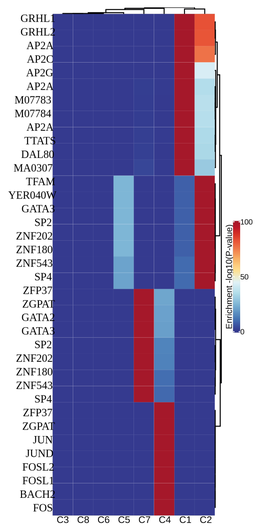
We discovered nearly 300 regulons in the analysis of our cells. We kept the twenty regulons for each population. Of those 38 were familiar with Zhi et al. dataset, 7 with Meistermann et al. datasets and 5 in both studies. We found transcription factors that have been described in the literature (*NANOG*, *GATA4*, *GATA3*). We also identified new candidates like *ETV1* (EPI), *KLF17* (ICM), *FOSB* (HYPO), *GAPBA* (TE).

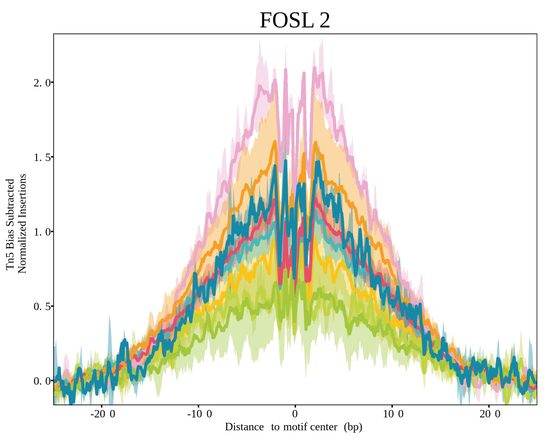


Heatmap of the eRegulons activity by cell populations in scMultiomics

We discovered nearly 100 regulons in the analysis of our cells. We found that most of our previous transcription factors have been retrieved (65% of similarity) with a better capture of zing-finger TF and literature known TF like *POU5F1* or *CDX2* that was not captured previously.

RESULTS

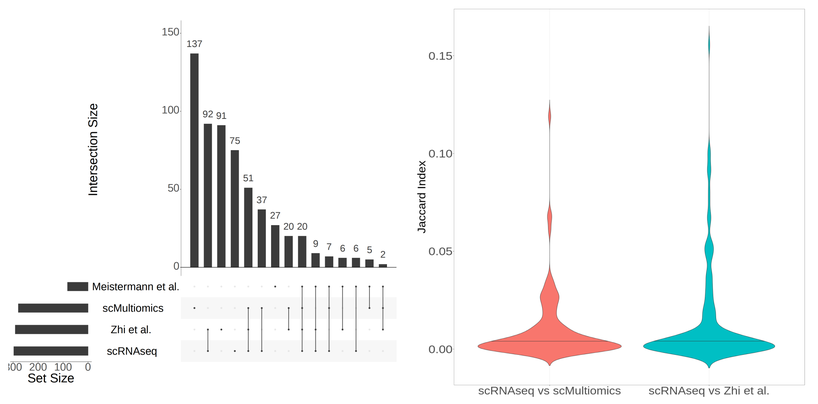




Heatmap of motif enrichment in our clusters (left) Footprint of the Fosl2 motif in our clusters

Visualisation of motif enrichment shows the high specificity of some of our identified regulons like *GRHL2* (TE), *FOSL2* (HYPO) and *NANOG* (EPI). We have also observed the footprint of some TF, validating the cluster specificity of some TF and showing the relative heterogeneity of the print for the different clusters of the same lineage.

III Regulon meta-analysis



Upset plot of commons TF between 4 different analyses (on the left) Violin plot of Jaccard index score comparing regulons composition across commons TF between regulons and eRegulons in our datasets (red) and between our scRNAseq datasets and Zhi et al. datasets (blue)

On the 585 TF identified in the four different single-cell analyses, we observe 243 TF identified in at least two pig's embryos datasets and 52 TF between the three studies in the pigs' embryos. However, let us look at the similarity of target genes on common regulons with the Jaccard index. We can see that neither sequencing technologies nor regulons vs eRegulons approach most of the regulons show that most have a low composition similarity.

CONCLUSION

Deconvolution of scMultiomics on scRNAseq datasets works well in high-quality cells and enough discriminant between clusters in the reference datasets.

Regulons approach shows similar TF identification and regulon activities between the approach and datasets. Unfortunately, gene regulons compositions slightly differ between datasets and approaches

The addition of the open regions from Multiomics helps us confirm those regulons by identifying enriched motifs and footprints for some of the identified regulons

REFERENCES

1. Butler, et al. 2018, Nat Biotechnol 36:411-420
2. Korsunsky, et al. 2019, Nat Methods 16:1289-1296
3. Aibar, et al. 2017, Nat Methods 14:1083-1086
4. Dufour, et al. 2023, Preprint
5. Granja, et al. 2021, Nat Genet 53:403-411
6. González-Blas, et al. 2022, Preprint
7. Stuart, et al. 2021, Nat Methods 18:1333-1341
8. Zhi, et al. 2022, Cell Res 32:383-400