Stem cell biology

https://doi.org/10.1038/s41592-024-02494-1

Building a molecular reference map of the human embryo

Rina C. Sakata & Marta N. Shahbazi

Two independent studies provide comprehensive human embryo reference maps by integrating multiple human embryo single-cell RNA sequencing (scRNA-seq) datasets. These references are instrumental in advancing cell type annotation and benchmarking stem cells and stem cell–based embryo models.

How do we define a cell type? Historically, these have been characterized based on phenotypic features, such as morphology and location within a tissue[1]. The development of single-cell transcriptome sequencing (scRNA-seq) techniques has revolutionized cell type profiling, enabling researchers to move beyond limited descriptions to more holistic characterizations of cell types and states. In recent years, we have seen the development of integrated reference datasets for several species and tissues, along with tools for efficient cell type annotation[2,3]. However, a similar resource for human embryonic development has so far been lacking. Now, two recent studies by Zhao et al. and Proks et al., published in this issue of *Nature Methods*, fill this gap by generating integrated reference maps for human embryos and user-friendly tools for cell type identification[4,5].

The logistical, technical and ethical challenges of studying the human embryo have led to the development of various stem cell–based embryo models[6]. To validate and benchmark these models for their applications, it is crucial to compare them against corresponding stages of human embryonic development. Although morphological and functional comparisons are essential, transcriptome-level comparisons provide a global and comprehensive unbiased tool for cell type annotation. However, the comparison of scRNA-seq data with data from human embryos presents challenges. As human embryo transcriptomic datasets remain scarce, combining them is crucial; yet downloading, storing and integrating large datasets across multiple studies can be both resource and labor intensive.

To address the challenges posed by human embryonic data analyses, Zhao et al. constructed an integrated reference map encompassing six human embryo datasets, spanning from the zygote to the gastrula stages (0–19 days after fertilization). This transcriptomic atlas was further enriched with non-human primate data to facilitate the characterization of cell types of the gastrula stage, for which data from human embryos are sparse, and allow the identification of divergent gene expression patterns. Leveraging this comprehensive resource, the authors developed the 'Early Embryogenesis Prediction Tool' that accurately predicts cell types across different embryonic stages and datasets. Moreover, the tool allows the projection of query datasets against the reference and the annotation of the predicted cell types through a user-friendly online platform. The authors plan to expand the reference as more datasets become available in future.

In parallel, Proks et al. provide an alternative reference tool based on a set of deep-learning techniques. They compiled 13 mouse and 6 human scRNA-seq datasets from pre-implantation embryos and developed a machine-learning cell-type classifier to annotate cell types in an unbiased fashion. Crucially, this classifier can predict the identity of cells from pre-implantation human embryos previously annotated as 'unknown'. Moreover, the authors went beyond integration and annotation and decoded the logic behind the neural network and the genes used by the model to accurately identify cell types. Interestingly, they found that the classifier uses both canonical and non-canonical markers for cell type classification. Proks et al. also provide an online platform for visualizing their reference model.

Both articles highlight the importance of an integrated reference dataset for examining the similarities between cell types within the embryo and the stem cell–based embryo models — but, crucially, also their differences (Fig. 1). For instance, using data from blastoids, stem cell-based embryo models of the pre-implantation blastocyst, Zhao et al. identified cells mapping to post-implantation identities, with different proportions depending on the protocol used. These cells may be misclassified when limited reference datasets are used. The authors also characterized transcriptional differences between cells of blastoids and blastocyst, even if they clustered as similar cell types. Importantly, they suggest that these molecular differences could have functional implications. Along similar lines, the data from Proks et al. shows that in contrast to what has been previously suggested, blastoids do not contain inner cell mass-like cells, which in embryos are the precursors of the embryonic epiblast lineage.

Currently, human embryo scRNA-seq datasets focus primarily on unperturbed samples. However, as the field matures, we can expect a proliferation of datasets from perturbed conditions. Comprehensive reference maps of healthy development will be invaluable for deciphering changes in gene expression, molecular pathways and the emergence of off-target cells in disease models. However, in the context of the human embryo, defining 'healthy' cells may be challenging because of the inherent variability of in vitro-fertilized human embryos and the high incidence of chromosomal abnormalities during the early stages of human development[7]. This variability needs to be carefully considered when creating new reference datasets to minimize potential biases.

With the emergence of distinct reference maps and their associated tools, the research community now faces choices in their usage, as each resource comes with unique strengths. Regardless, as the number of stem cell-based human embryo models continues to rapidly grow, these new tools are an invaluable resource for building more accurate stem cell-based embryo models and generating hypotheses that need to be functionally tested using stem cells.

Despite the undeniable value of these two reference tools, they still have their limitations. The accuracy of the predictions is

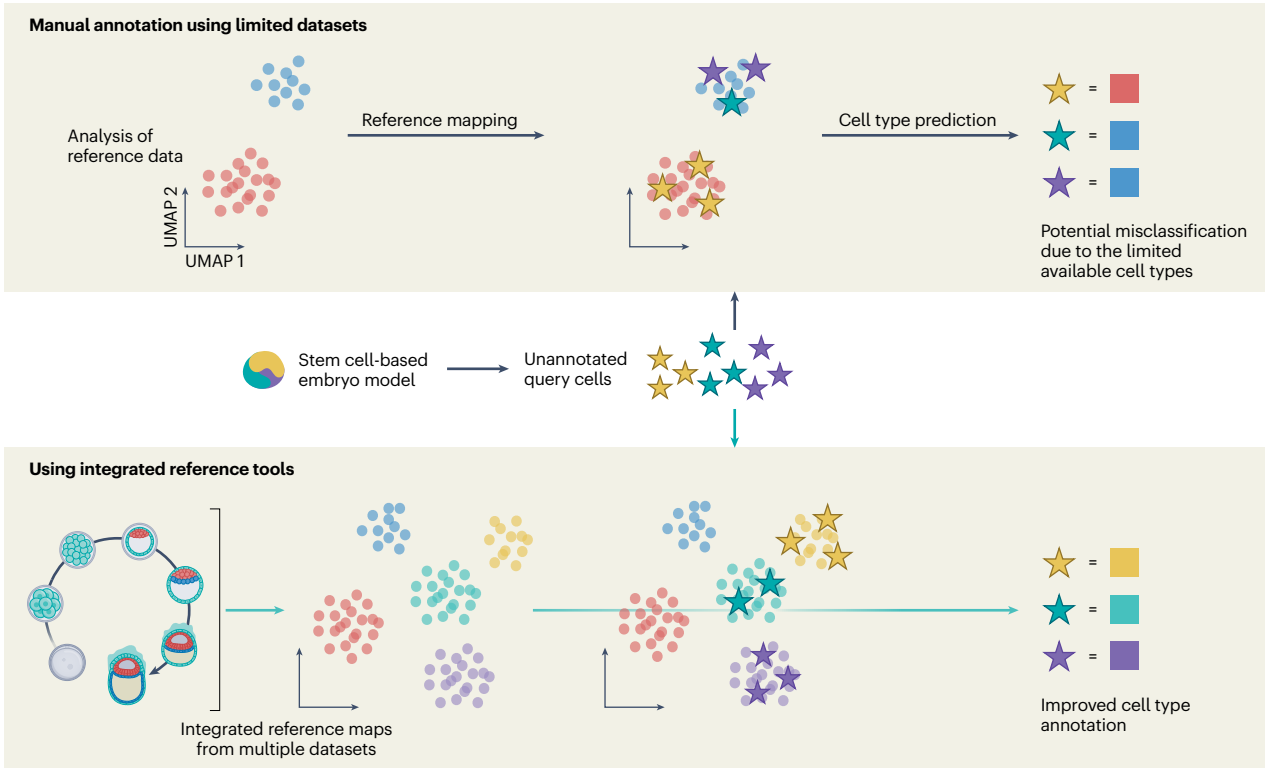

Fig. 1 | Potential limitations of annotating cell types based on limited reference datasets compared with the use of integrated reference tools. Different cell types are shown in different colors. Unannotated query cells are denoted by stars. UMAP, uniform manifold approximation and projection.

heavily influenced by the quality and quantity of data used to build the reference maps. Although multiple datasets have been integrated for a more comprehensive reference, users still need to be aware of the potential for false positive predictions, particularly for rare cell types with limited representation in the reference, such as primordial germ cells, the progenitors of the gametes. Throughout history, technology has influenced our approach to cell type annotation. For example, recent developments in single-cell proteomics[8] and single-cell morphometrics[9] are reviving the histological classifications of cell types. Thus, as both studies acknowledge, expanding the reference dataset (for example, with missing developmental stages), incorporating spatial information and integrating data from targeted methods for specific cell types will be key to continue improving the characterization of cellular diversity across early human development.

Rina C. Sakata & Marta N. Shahbazi
MRC Laboratory of Molecular Biology, Cambridge Biomedical Campus, Cambridge, UK.
e-mail: mshahbazi@mrc-lmb.cam.ac.uk

Published online: 14 November 2024

References

1. Mulas, C., Chaigne, A., Smith, A. & Chalut, K. J. *Development* **148**, dev199950 (2021).
2. Regev, A. et al. *eLife* **6**, e27041 (2017).
3. Tabula Sapiens, C. et al. *Science* **376**, eabl4896 (2022).
4. Proks, M., Salehin, N. & Brickman, J. M. *Nat. Methods* https://doi.org/10.1038/s41592-024-02511-3 (2024).
5. Zhao, C. et al. *Nat. Methods* https://doi.org/10.1038/s41592-024-02493-2 (2024).
6. Shahbazi, M. N. & Pasque, V. *Cell Stem Cell* **31**, 1398–1418 (2024).
7. van Echten-Arends, J. et al. *Hum. Reprod. Update* **17**, 620–627 (2011).
8. Bennett, H. M., Stephenson, W., Rose, C. M. & Darmanis, S. *Nat. Methods* **20**, 363–374 (2023).
9. Andrews, T. G. R., Pönisch, W., Paluch, E. K., Steventon, B. J. & Benito-Gutierrez, E. *Development* **148**, dev199430 (2021).

Acknowledgements

We are grateful to L. Schwarz for reading the article and providing suggestions. R.C.S. is funded by the Funai Foundation for Information Technology. The laboratory of M.N.S. is funded by the Medical Research Council as part of UK Research and Innovation (grant reference MRC, MC_UP_1201/24), and the Engineering and Physical Sciences Research Council (Horizon Europe guarantee funding EP/X023044/1).

Competing interests

The authors declare no competing interests.