SOMITE

Somite Therapeutics

Digital twins for ML-guided cell replacement therapy

Last edited: Dec. 10th, 2023

1. Introduction.
2. Motivation: discovering and optimizing in vitro differentiation protocols for CRT.
3. Digital twins: concept and data sources.
4. Progress in digital twin construction.
5. Analytical steps in digital twin construction by scRNA-Seq.
6. Putting digital twins to work for CRT protocol discovery: case studies.
7. Iterative GD and DP protocol optimization guided by machine-learning.
8. References.

1. Introduction.

Cell replacement therapy (CRT) is a therapeutic strategy that involves the transplantation of functional cells or tissues to replace or repair damaged or diseased cells in the body (*1*, *2*). Generation of cells for CRT requires multi-factorial protocols that can vary in efficiency, scalability, and robustness (*2*, *3*). CRT protocols span a large combinatorial search space, which can be optimized using data-driven approaches supported by formal methods of machine learning (ML) (*4*, *5*).

A central data concept for CRT protocol optimization is the digital twin of embryo development and in vitro differentiation. A digital twin is a computational replica of a biological entity, which guides decision-making (*6*, *7*). For CRT, digital twins are constructed from data-rich sources on embryos and from cells at stages of differentiation in a dish, primarily single-cell RNA sequencing augmented with extensive auxiliary information (*8–14*). Embryo digital twins serve two purposes: **(1)** they establish strong priors on the protocol search space, and **(2)** they establish a ground truth for the identity and composition of cells at every stage of the protocols. These together enable: rapidly identifying novel protocols for generating new cell types; discovering new regulators of cell differentiation; and carrying out rapid CRT protocol optimization cycles.

This white paper discusses the state-of-the-art in digital twin construction, distills the general use-cases, surveys case-studies from the literature on data-driven CRT protocol optimization, and identifies immediate priorities that position digital twins as a pivotal tool in the future of CRT.

2. Motivation: discovering and optimizing *in vitro* differentiation protocols for CRT.

The concept of using stem cells to generate cells for therapeutic purposes has its roots in the mid 20th century. It involves solving (at least) two challenges: (1) generating pluripotent stem cells; and then (2) converting them into desired mature cell types at scale, with high purity, and reliably. The first challenge of generating stem cells is now sufficiently solved (*15*). The second class problem must be solved anew for every target cell type. There are hundreds of distinct cell types, and many have specialized sub-sets, meaning that there are likely thousands of mature end-points, each requiring a distinct protocol (*2*, *3*, *16*). "Digital twins" represent an analytical framework for addressing this challenge systematically as discussed here. We note also that there exist additional challenges in delivering cell types and ensuring successful engraftment, which are not discussed in this white paper.

Generating stem cells. Although not the focus of this white paper, we briefly mention key milestones leading to availability of stem cells as a resource for CRT. The idea of harnessing stem cells for therapy arose in the mid-20th with successes in transplanting hematopoietic stem cells and skin grafts. The idea of generating new cells from CRT was galvanized by the advent of embryonic stem cell research in 1980s and 1990s: the first isolation of embryonic stem cells from mice was achieved in 1981 (*17*), and in the late 1990s there was a significant breakthrough with the successful generation of human embryonic stem cells (hESCs) (*18*). In 2006, Shinya Yamanaka showed that adult cells could be reprogrammed into induced pluripotent stem cells (iPSCs) (*19*), offering a way to create patient-specific cells and to establish banks of stem cells. These breakthroughs established the raw substrate for cell replacement therapy.

Stem cell differentiation into mature cell types. Turning human embryonic stem cells (hESCs) and induced pluripotent stem cells (iPSCs) into specific mature cell types for cell replacement therapy involves a two major strategies (*2*, *3*): (1) Guided differentiation (GD), mimicking natural developmental processes that would normally occur in embryos as cells progress from the early pluripotent state to a given tissue. This can include adding specific growth factors and chemicals to the culture medium, sequentially at defined doses for defined periods of time. And, (2) Direct programming (DP), genetically manipulating cells in order to directly transition cells into the final desired state.

In both cases (GD and DP), the goal for CRT is to define a robust protocol that leads to the desired terminal cell types at large scale, with minimal contaminants of unwanted cell types, and with high reproducibility (*3*). A “protocol” is an instruction set fully specifying the sequence of manipulations on cells that would be carried out in a GMP facility to generate cells for therapeutic purposes.

Challenges in stem cell differentiation protocols. There are around 200 major cell types in the human body, and many of these have highly specialized subsets that vary between and within tissues. Several examples illustrate the scale of the challenge:

- There are multiple neuronal subsets, and current estimates from single cell transcriptomic profiling suggest that these may number in the thousands.
- Goblet cells secrete mucins, and have very different mucin production and auxiliary gene expression between different mucosal epithelial linings (airway, gut, intestine).
- Fibroblasts are cells that support the mesenchymal extracellular matrix, and vary considerably between niches.
- Epidermal cells (keratinocytes) give rise to skin with different mechanical properties and thickness between different body regions (e.g. the palm vs the back), and they retain their identity after grafting.
- Muscle satellite stem cells partially retain regional identity after transplantation from one body region to the other.

These illustrative examples make clear that building a protocol to generate the correct cell types has many potential end-points, and requires the ability to distinguish between functional (desirable) and unwanted CRT products. That is - the overarching challenge for CRT is to generate the right cell type, with minimal unwanted cell products (i.e. at high purity), in large amounts, with minimal variation due to experimental process or cell-of-origin (i.e. robustly), and at scale.

Several representative iPSC differentiation protocols makes clear the challenge and opportunities of harnessing digital twins:

- A first protocol to differentiate iPSCs into muscle satellite stem cells (MSCs) generated on-target cells with 25% purity (75% off-target cells), and the cells have an immature (fetal) identity. Subsequent optimization using data underlying the “digital twin” (discussed below) increased purity to 50%, and then 80% with further protocol optimization (*20*).
- A first protocol to differentiate iPSCs into enameloblasts (cells that regenerate the tooth enamel) was only made possible after analyzing data underlying the “digital twin” (discussed below) (*21*).

- Attempts to differentiate hematopoietic stem cells for bone marrow transplantation using GD protocols gives rise to embryonic-like (yolk sac primitive/definitive) cells with inability to engraft, and the generation of mature cells currently requires simultaneous exogenous expression of seven transgenes by DP, which raises safety and efficiency concerns (*22*, *23*).
- Differentiation of iPSCs into beta cells for treatment of diabetic patients initially gave rise to a large fraction of insulin-negative cells. The identity of these cells and their subsequent depletion to increase purity of the resulting beta cell islets was enabled by analysis via scRNA-Seq, using data underlying the "digital twin" (*24*).
- Attempts to generate hepatocytes (liver cells) from iPSCs have not yet been successful (*25*), despite a clear clinical for liver transplants.

3. Digital twins: concept and data sources.

Digital twins offer a data-driven approach to the task of protocol discovery and optimization. The digital twin of an embryo is a data-rich representation of the composition and organization of an embryo as it transitions from pluripotency to the formation of mature cell types. The digital twin of *in vitro* differentiation is a similar representation of cells grown in a dish as they transition from pluripotency under a defined GD or DP protocol. Such representations identify likely treatments, and offer a ground-truth against which to compare existing protocols.

The ideal entity. A digital twin should represent a description of embryos and cells that generate hypotheses for guiding cells to mature end-states and for comparing generated cells to those found in embryos. A "complete" digital twin is currently an idealization that is not yet possible to reach: a full (but impractical) catalog of molecular composition would include transcriptome (mRNA), proteome, protein modifications (e.g. phospho-proteome), lipids, metabolites, genome epigenetic state, genome 3D organization, mechanical forces acting on cells, extracellular matrix composition and organization, the 3D organization of a tissue, and the milieu of extra-cellular signaling ligands that act on cells. It would also allow querying relationships between these and provide predictions for dynamics and the effect of treatments on cells. In practice, the amount of data that can be generated falls short of such a complete description, but the current state-of-the-art is already sufficient to empower protocol optimization [multiple examples – (*21*, *26*–*30*)]. Technical advances continue to extend the data modalities available for digital twin construction (*31*, *32*).

Realization of digital twin embryos. Today, there are multiple data sources that should support digital twin embryo construction:

1. **Single cell genomics**: single cell RNA sequencing (scRNA-Seq) gives information on the genes that are expressed in every cell in the embryo over time. scRNA-Seq is the most robust method to produce high-quality data on the state of cells. scRNA-Seq also enables establishing dynamic relationships between embryonic cell states, by capturing cells at transitional stages that form a continuum of states. scRNA-Seq data has been used to generate “single cell atlases” of embryo development. These atlases establish a backbone for digital twins. Other single cell information can be used to augment scRNA-Seq, including information on chromatin accessibility (scATAC), and partial information on protein expression. These data alone do not form a “digital twin” as they require extensive expert domain knowledge to interpret. An important consideration in digital twin construction is the need for dense time-series, which can track dynamic changes in cells from pluripotency to the formation of mature cell types (*9*, *10*, *13*, *33*).
2. **Pathway databases:** The academic literature has been curated into several widely-used databases that identify gene sets associated with specific molecular functions and signaling pathways. These include GO, REACTOME, KEGG. These databases alone have been used to empower inference of pathway activity from gene expression in scRNA-Seq atlases, because signaling ligands acting on cells will leave transcriptional signatures unique to different ligands that can be interpreted using these databases (*34*). Similar approaches have used these databases to infer metabolic state from scRNA-Seq data (*35*). These databases enable some limited interpretation of scRNA-Seq databases, although with considerable false-positives because the nature of database construction was never intended for inference.
3. **Gene expression databases:** Public repositories of gene expression data (GEO (*36*), EBI Expression Atlas (*37*)) are large collections of unrelated studies, each associated with a publication. These studies make a crucial link between expert domain knowledge (in the papers) and quantitative changes in gene expression (in the repositories). Such information is ripe for mining in order to guide interpretation and prediction on scRNA-Seq atlases.
4. **Academic literature in developmental and stem cell biology:** Beyond repositories, the whole gamut of academic text on developmental biology and in vitro stem cell differentiation contains decades of knowledge on regulators of development. This large text resource is ripe for systematic mining by large language models (LLMs) with training for specific tasks stem cell and developmental biology (*38*), including (a) identifying progenitor cell states from genes expressed; (b) identifying signaling pathways that have been demonstrated to regulate the fate of progenitors and associating them with specific outcomes; (c) identifying the natural micro-environment of progenitor

cells including metabolites and mechanical cues; (d) generating bibliographies supporting the LLM outputs.

4. Progress in digital twin construction.
To date, there has been progress on (1) single cell genomic representation of embryos, and (2) utilizing pathway databases in the academic community. There are still major gaps in (1) that are discussed below. Data sources (3) Gene Expression databases, and (4) Academic journal papers have not yet been systematically integrated into digital embryo efforts. The fact that domain expert knowledge has until recently been the primary guide in establishing iPSC differentiation protocols argues very strongly that these data sources represent a major untapped resource, which can dramatically accelerate CRT protocol discovery and optimization. Their integration with scRNA-Seq atlases will provide an empirical, relevant digital twin.

Single cell genomic representation of embryos. The first time-series digital twin scRNA-Seq representations of vertebrate embryo development were published in 2018, initially in zebrafish (*10*, *39*) and frogs (*9*), which spanned the first day of life and tracked cells from pluripotency up to the formation of tens of tissues. Representations of mammalian embryos have followed with mouse in 2019 onwards (*13*, *40*, *41*). The initial efforts collected information on 10,000s of cells across whole embryos, and now more recent efforts have collected information on millions of cells in both zebrafish (*42*) and mouse (*43*). Collection of data on human development has necessarily been more challenging due to ethical considerations, and constructing time-series remains particularly challenging because it is not possible to specifically target the collection of samples at consistent time points (*33*). A first atlas covering some stages of human fetal development has been published (*44*).

Single cell genomic representation of isolated tissues. As embryo development proceeds the number of tissues and their complexity makes whole-organism analyses both cost-prohibitive and technically difficult due to specific considerations in dissociating and analyzing different tissues. As a result, efforts focusing on a single tissue or cell type benefit from targeted analyses. These efforts follow the same exact concepts as for a whole embryo, but manage to efficiently collect information on just a subset of cells. There are many tissue-specific human developmental datasets available in public repositories (*45–48*), although gaps exist for many tissues.

Outlook for digital twin data generation. Available resources provide an actionable starting point for protocol optimization, but they are still patchy. They do not cover all stages of development, do not cover all tissues, and do not capture natural variation. As

we work to build digital twin databases that will further empower CRT protocol development, we anticipate the following three priorities for data collection:

- **Human digital twins by scRNA-Seq:** ongoing effort will be needed to build up data on human development for specific tissues of interest, making use of ethical tissue sources. Recognizing human genetic diversity and natural variation, a strong digital twin resource should collect data on tens to hundreds of samples, so as to control for and leverage natural variation in the human population. Academic community efforts will serve this purpose in part (*33*), and we anticipate a need to generate our own data focusing intensely on somite-derived tissues through in-house and sponsored research agreements.
- **Multi-species digital twins:** There is likely significant value in cross-species comparisons because of the ability to collect embryonic tissues from experimental model systems with high time resolution, systematically and reproducibly. These include mouse, rabbit, pig, primates and more distant vertebrates. Data is available for some of these organisms, albeit not spanning the full range of time points needed to trace somite-derived tissues. For many practical CRT applications it will be necessary to move towards much longer time series covering later stages of tissue maturation. Data generation in some cases is extremely accessible (e.g. zebrafish and mice). Putting data from these organisms to work will also require establishing a computational platform to simultaneously query information from orthologous tissues across species.
- **"5D" digital twins:** as spatially-resolved transcriptomics improves, considerable efforts should be made to generate 5D digital twins - resolved across space (3D) and time (the fourth dimension) and natural variation (the fifth dimension). This will be a major effort given the need to collect and analyze tens to hundreds of high-quality sections per time point. Such efforts will become accessible within the next five years.

5. Analytical steps in digital twin construction by scRNA-Seq.
Digital twin construction from scRNA-Seq time series involves a series of computational steps that initially require manual supervision and curation.

Dimensionality reduction and embedding. The first step involves data-filtering and quality-control following standard scRNA-Seq data hygiene practices (*49*, *50*). Subsequently, unsupervised approaches are used for low-dimensional embedding and clustering of scRNA-Seq data at each time point. These approaches can use linear methods (Principal Component Analysis), or multi-layered encoding (variational auto-encoders) that explicitly model noise in scRNA-Seq data.

Cell type annotation. Cell transcriptomes are annotated such that each single cell transcriptome is associated with a particular tissue and cell type. The process of annotation initially involves expert input by developmental biologists, and in some cases it can require input from experts that bring decades of expertise in specific tissues. As data accumulate, this knowledge can be encoded efficiently by transfer learning between data sets and applying pre-trained classifiers (*51*, *52*).

Time-series integration. While cells are annotated, a next task is to integrate information from multiple time-points together to form a time-series tracing the changes in cell state over time. Such a time-series can be approximately represented by a tree that is rooted in the pluripotent state. At present, time-series construction is carried out after data from each time point is separately processed and annotated. The relationship between time points modeled can be described using heuristic distance metrics, or by explicitly defining a Transition Map between time points that attempts to model dynamic transition probabilities between states (*26*, *53*).

Manifold isolation. The construction of the digital twin embryo building on scRNA-Seq data alone can be complicated by additional sources of variation that are partially or completely independent of cell type differentiation. Cells in all tissues can exist in different cell cycle phases, and express phase-specific genes that alter the distance between cells in the embedding space. In addition, the same cell types can differentiate in different parts of the embryo and thus express genes that vary in a spatially-defined manner (“polytopy”). The same cells can also appear at different times, leading to asynchrony that can lead to strong similarities between cells across time points. These competing processes can lead to co-clustering or co-embedding of cells from different locations or timing, leading to cell state representations that artificially combine gene expression. Such problems are partly overcome by semi-supervised approaches that enforce cell cycle-, spatial-, and timing- aware embeddings (*50*, *54*, *55*). At present, methods for cell cycle deconvolution have been developed, building on databases trained on cells in different cell cycle phases (*56*, *57*). Similar approaches can be considered for other sources of variation by developing appropriate databases for anterior-posterior gradients in embryos, although these databases are still lacking.

Extrinsic environmental signature discovery. A goal for digital twins is to identify likely changes in cues that can be used to guide cells towards desired differentiated outcomes. Databases of transcriptional responses to signaling, along with information on the receptors expressed by cells and their cognate ligands, can be used to generate hypotheses for the signals acting in cells in the embryo, which may be useful candidates for treatment in vitro. Similarly, variation in genes associated with cell metabolism can serve to resolve changes in metabolic environments. It is possible that changes in

mechanical environment similarly lead to stereotyped gene expression responses, although such databases are still lacking. Several methods have been developed to identify likely signals. These can be improved by generating tissue-specific databases by treating tissues and stem cell derivatives with known signaling ligands at varying doses. Such a database has been generated for one tissue (*58*), and the process can be generalized.

Integration with literature and annotated gene expression repositories. Further interpretation of scRNA-Seq data sets is carried out in light of the gamut of prior work. Gene expression repositories represent a rich resource that links changes in gene expression to specific experimental perturbations, across tens of thousands of studies. In the past few years, manual curation of these data sets has enabled interpretation of cell states in scRNA-Seq data [e.g. as in (*59*)]. Such analyses will benefit dramatically from the use of large language models (LLMs) that can prioritize and then execute relevant comparisons of scRNA-Seq atlases (*60*).

Outlook and priorities. As we work to build digital twin databases that will accelerate CRT protocol development, we anticipate the following priorities for computational methods and platforms:

- Digital twins are ripe for the establishment of generative models, which can predict the next step of differentiation in one tissue and organism by training on data from other tissues and organisms. Such generative models will serve as a platform for trouble-shooting *in vitro* differentiation protocols.
- Given the commonality of development across vertebrates, a unified data framework that encompasses digital twins across all organisms, time-series and tissues will allow sharing of annotations, and bridging inevitable gaps in human databases.
- The establishment of databases of signaling, metabolic and mechanical responses across several tissues will establish the substrate for supervised ML of signaling cues in digital twin embryos.
- Integrating aforementioned data sources (1)-(4). This can be realized through LLMs trained on GEO metadata, large text repositories on stem cell and developmental biology publications, the REACTOME/KEGG databases and with the ability to evaluate scRNA-Seq embryo atlases. LLMs have already begun to find use in biological data analysis [reviewed here (*61*)] but so far focused on sequence-prediction tasks, or on annotation of scRNA-Seq data based on transfer from other scRNA-Seq data sets (*62*). A specific challenge is to build on expert knowledge embodied in literature to guide GD protocol optimization prioritization of pathway activity and prediction of relevant perturbations. Such prioritization currently relies on expert knowledge or piece-wise analyses using

curated databases of signaling responses, which lack cell type-specific and developmental context. We expect the prioritization task to benefit from use of large language models (LLMs) that can relate scRNA-Seq atlases to past work.

6. Putting digital twins to work for CRT protocol discovery: case studies. Several common steps used to date are reviewed in (*29*), with specific examples discussed below:

Optimizing muscle satellite stem cell differentiation. Satellite cells (SCs) are a population of cells able to regenerate damaged skeletal muscles. Patient derived SCs lose their regenerative capacity when amplified in culture, motivating the need for SCs derived from iPSCs for CRT. Initial protocols based on expert knowledge of the embryo yielded SCs from iPSCs with a purity of 25% (*20*). Computational analysis of digital twin scRNA-Seq revealed signatures of ligand-mediated signaling with different pathways from those used in the established protocol. A resulting optimized protocol using these predictions generates cultures containing up to ~75% human SCs. These SCs are functional as they can regenerate injured muscles in mice and restore force production as compared to uninjected controls. This work demonstrates the utility of digital twins for optimization and establish a working CRT protocol for muscle stem cell therapy with potential application to diseases such as Duchenne Muscular Dystrophy.

Developing protocols for brown adipose tissue differentiation. This case study illustrates protocol discovery via a digital twin. Brown adipocytes (BAs) are a potential source of cells for treating metabolic diseases, including type 2 diabetes. In recent work, a protocol was developed to differentiate iPSCs through paraxial mesoderm progenitors, into BAs. To optimize protocols for BA production, a digital embryo scRNA-Seq time-series was first generated for relevant tissues in mouse embryos at gestational days E13.5, E14.5, and E15.5. This mapping identified a previously unrecognized population of BA precursors expressing the transcription factor GATA6. Armed with this knowledge, iPSC differentiation protocols could target generation of this intermediate state, and successfully identified conditions to generate these cells from paraxial mesoderm precursors differentiated in vitro from hPSCs by modulating the signaling pathways identified in the digital embryo scRNA-Seq data. These precursors could in turn be efficiently converted into functional brown adipocytes which can respond to adrenergic stimuli by increasing their metabolism resulting in heat production.

Developing protocols for tooth enamel differentiation. This case study illustrates a second example of protocol discovery via a digital twin. Tooth enamel is gradually damaged or partially lost in over 90% of adults and cannot be regenerated due to a lack of a specialized cell type - the ameloblast - which disappears once teeth erupt (i.e. emerge). By establishing an scRNA-seq and spatially-resolved atlas of the developing

human tooth and using it to prioritize signaling pathways, Ref. (*21*) were able to generate human ameloblasts *in vitro* from iPSCs. They showed that the resulting AMs matured to give rise to mineralized structure in vivo, suggesting a therapeutic strategy for restoring enamel.

Towards protocols for regenerating visceral organs with iPSC derivatives. This case study illustrates a third example of protocol discovery via a digital twin. Visceral organs, such as the lungs, stomach and liver, are derived from the fetal foregut through a series of inductive interactions between the definitive endoderm (DE) and the surrounding splanchnic mesoderm (SM). To correctly generate such tissues in vitro, a goal is to generate SM subtypes from human pluripotent stem cells (hPSCs), which has been elusive. Ref. (*27*) used scRNA-Seq to generate a high-resolution cell state map of the embryonic mouse foregut. This digital twin analysis identified a diversity of SM cell types that develop in close register with different organ-specific epithelia. Leveraging this analysis, they were able to prioritize protocols that generated different SM subtypes from iPSCs.

7. Iterative GD and DP protocol optimization guided by machine-learning.
The previous sections have focused particularly on generation of a digital twin of natural embryo development, as a constraint and ground-truth for on iPSC differentiation protocols. Building such a twin then requires the ability to carry out rapid rounds of protocol optimization.

DP protocol optimization. Direct programming protocols rely on combinatorial expression of transgenes that result in iPSCs entering a final differentiated state. DP protocols are extremely amenable to very high-throughput cycles of optimization by introducing random pools of transgenes (typically transcription factors, or TFs) into cells, and then carrying out scRNA-Seq combined with perturbation detection at the single cell level (PERTURB-SEQ) in order to identify particular combinations of TFs that drive cells towards desired outcomes. Because at least tens of thousands of TF combinations can be evaluated in parallel, it becomes possible to carry out iterative learning (*63*). Thus, DP protocol optimization is extremely well suited for ML-guided iterative learning. A downside is that DP protocols have not yet been demonstrated to have high efficiency, scalability and to yield functional products, and the introduction of multiple transgenes may potentially demand further stringency in consideration to safety.

GD protocol optimization. GD protocols require the application of sequential treatments to cells (metabolic conditions, media, growth factors, cytokines and small molecules). As such they cannot be simply evaluated in massively-parallel pooled formats. The use of industry-standard liquid-handling robotics can be used to establish large-scale

plate-based screens, where every well in a plate represents a different condition for optimization. For such approaches, there is a trade-off between the speed of iteration and the number of conditions that can be practically realized. Overall speed is accelerated by breaking down full GD protocols into steps that cover 1-2 days of differentiation, and then optimizing each step in isolation. The use of fluorescent reporters integrated into iPSC lines enables rapid read-out of changes to protocol efficiency as a function of defined conditions. The choice of reporters are informed by the embryo digital twin.With minimal effort, hundreds of conditions can be evaluated in each iteration. The use of combinatorial labeling can increase this into the thousands with simple experimental designs (*64*), and these may be turned into massively-parallel approaches by split-and-pool of iPSC cultured encapsulated into transferable micro-particles [combining (*64*) and (*65*)]. Formal machine-learning here offers approaches to systematically design and improve experiments by defining appropriate (1) cost functions (reflecting protocol yield and efficiency); and (2) policy function for exploit-explore of new conditions in subsequent iterations. Such an approach implementing Gaussian Process regression has been realized at a small scale to optimize of iPSC differentiation into retinal pigmented epithelium (RPE) (*5*).

8. References.

1. M. X. Doss, C. I. Koehler, C. Gissel, J. Hescheler, A. Sachinidis, Embryonic stem cells: a promising tool for cell replacement therapy. *J. Cell. Mol. Med.* **8**, 465–473 (2004).

2. S. Yamanaka, Pluripotent Stem Cell-Based Cell Therapy-Promise and Challenges. *Cell Stem Cell*. **27**, 523–531 (2020).

3. A. Aijaz, M. Li, D. Smith, D. Khong, C. LeBlon, O. S. Fenton, R. M. Olabisi, S. Libutti, J. Tischfield, M. V. Maus, R. Deans, R. N. Barcia, D. G. Anderson, J. Ritz, R. Preti, B. Parekkadan, Biomanufacturing for clinically advanced cell therapies. *Nat Biomed Eng*. **2**, 362–376 (2018).

4. M. Ashraf, M. Khalilitousi, Z. Laksman, Applying Machine Learning to Stem Cell Culture and Differentiation. *Curr Protoc*. **1**, e261 (2021).

5. G. N. Kanda, T. Tsuzuki, M. Terada, N. Sakai, N. Motozawa, T. Masuda, M. Nishida, C. T. Watanabe, T. Higashi, S. A. Horiguchi, T. Kudo, M. Kamei, G. A. Sunagawa, K. Matsukuma, T. Sakurada, Y. Ozawa, M. Takahashi, K. Takahashi, T. Natsume, Robotic search for optimal cell culture in regenerative medicine. *Elife*. **11** (2022), doi:10.7554/eLife.77007.

6. M. Singh, E. Fuenmayor, E. P. Hinchy, Y. Qiao, N. Murray, D. Devine, Digital Twin: Origin to Future. *Applied System Innovation*. **4**, 36 (2021).

7. F. Tao, B. Xiao, Q. Qi, J. Cheng, P. Ji, Digital twin modeling. *Journal of Manufacturing Systems*. **64**, 372–389 (2022).

8. Y. Xu, T. Zhang, Q. Zhou, M. Hu, Y. Qi, Y. Xue, Y. Nie, L. Wang, Z. Bao, W. Shi, A single-cell transcriptome atlas profiles early organogenesis in human embryos. *Nat. Cell Biol.* **25**, 604–615 (2023).

9. J. A. Briggs, C. Weinreb, D. E. Wagner, S. Megason, L. Peshkin, M. W. Kirschner, A. M. Klein, The dynamics of gene expression in vertebrate embryogenesis at single-cell resolution. *Science*. **360** (2018), doi:10.1126/science.aar5780.

10. D. E. Wagner, C. Weinreb, Z. M. Collins, J. A. Briggs, S. G. Megason, A. M. Klein, Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. *Science*. **360**, 981–987 (2018).

11. I. Imaz-Rosshandler, C. Rode, C. Guibentif, L. T. G. Harland, M.-L. N. Ton, P. Dhapola, D. Keitley, R. Argelaguet, F. J. Calero-Nieto, J. Nichols, J. C. Marioni, M. F. T. R. de Bruijn, B. Göttgens, Tracking early mammalian organogenesis - prediction and validation of differentiation trajectories at whole organism scale. *Development* (2023), doi:10.1242/dev.201867.

12. M.-L. N. Ton, D. Keitley, B. Theeuwes, C. Guibentif, J. Ahnfelt-Rønne, T. K. Andreassen, F. J. Calero-Nieto, I. Imaz-Rosshandler, B. Pijuan-Sala, J. Nichols, È. Benito-Gutiérrez, J. C. Marioni, B. Göttgens, An atlas of rabbit development as a model for single-cell comparative genomics. *Nat. Cell Biol.* **25**, 1061–1072 (2023).

13. B. Pijuan-Sala, N. K. Wilson, J. Xia, X. Hou, R. L. Hannah, S. Kinston, F. J. Calero-Nieto, O. Poirion, S. Preissl, F. Liu, B. Göttgens, Single-cell chromatin accessibility maps reveal regulatory programs driving early mouse organogenesis. *Nat. Cell Biol.* **22**, 487–497 (2020).

14. G. Peng, G. Cui, J. Ke, N. Jing, Using Single-Cell and Spatial Transcriptomes to Understand Stem Cell Lineage Specification During Early Embryo Development. *Annu. Rev. Genomics Hum. Genet.* **21**, 163–181 (2020).

15. Y. Shi, H. Inoue, J. C. Wu, S. Yamanaka, Induced pluripotent stem cell technology: a decade of progress. *Nat. Rev. Drug Discov.* **16**, 115–130 (2017).

16. D. M. Hoang, P. T. Pham, T. Q. Bach, A. T. L. Ngo, Q. T. Nguyen, T. T. K. Phan, G. H. Nguyen, P. T. T. Le, V. T. Hoang, N. R. Forsyth, M. Heke, L. T. Nguyen, Stem cell-based therapy for human diseases. *Signal Transduct Target Ther*. **7**, 272 (2022).

17. M. J. Evans, M. H. Kaufman, Establishment in culture of pluripotential cells from mouse embryos. *Nature*. **292**, 154–156 (1981).

18. J. Yu, M. A. Vodyanik, K. Smuga-Otto, J. Antosiewicz-Bourget, J. L. Frane, S. Tian, J.

Nie, G. A. Jonsdottir, V. Ruotti, R. Stewart, I. I. Slukvin, J. A. Thomson, Induced pluripotent stem cell lines derived from human somatic cells. *Science*. **318**, 1917–1920 (2007).

19. K. Takahashi, S. Yamanaka, Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell*. **126**, 663–676 (2006).

20. Z. Al Tanoury, J. Rao, O. Tassy, B. Gobert, S. Gapon, J.-M. Garnier, E. Wagner, A. Hick, A. Hall, E. Gussoni, O. Pourquié, Differentiation of the human PAX7-positive myogenic precursors/satellite cell lineage in vitro. *Development*. **147** (2020), doi:10.1242/dev.187344.

21. A. Alghadeer, S. Hanson-Drury, A. P. Patni, D. D. Ehnes, Y. T. Zhao, Z. Li, A. Phal, T. Vincent, Y. C. Lim, D. O'Day, C. H. Spurrell, A. A. Gogate, H. Zhang, A. Devi, Y. Wang, L. Starita, D. Doherty, I. A. Glass, J. Shendure, B. S. Freedman, D. Baker, M. C. Regier, J. Mathieu, H. Ruohola-Baker, Single-cell census of human tooth development enables generation of human enamel. *Dev. Cell*. **58**, 2163–2180.e9 (2023).

22. S. Demirci, A. Leonard, J. F. Tisdale, Hematopoietic stem cells from pluripotent stem cells: Clinical potential, challenges, and future perspectives. *Stem Cells Transl. Med.* **9**, 1549–1557 (2020).

23. I. Rao, L. Crisafulli, M. Paulis, F. Ficara, Hematopoietic Cells from Pluripotent Stem Cells: Hope and Promise for the Treatment of Inherited Blood Disorders. *Cells*. **11** (2022), doi:10.3390/cells11030557.

24. A. Veres, A. L. Faust, H. L. Bushnell, E. N. Engquist, J. H.-R. Kenty, G. Harb, Y.-C. Poh, E. Sintov, M. Gürtler, F. W. Pagliuca, Q. P. Peterson, D. A. Melton, Charting cellular identity during human in vitro β-cell differentiation. *Nature*. **569**, 368–373 (2019).

25. T. Tricot, C. M. Verfaillie, M. Kumar, Current Status and Challenges of Human Induced Pluripotent Stem Cell-Derived Liver Models in Drug Discovery. *Cells*. **11** (2022), doi:10.3390/cells11030442.

26. S.-W. Wang, M. J. Herriges, K. Hurley, D. N. Kotton, A. M. Klein, CoSpar identifies early cell fate biases from single-cell transcriptomic and lineage information. *Nat. Biotechnol.* **40**, 1066–1074 (2022).

27. L. Han, P. Chaturvedi, K. Kishimoto, H. Koike, T. Nasr, K. Iwasawa, K. Giesbrecht, P. C. Witcher, A. Eicher, L. Haines, Y. Lee, J. M. Shannon, M. Morimoto, J. M. Wells, T. Takebe, A. M. Zorn, Single cell transcriptomics identifies a signaling network coordinating endoderm and mesoderm diversification during foregut organogenesis. *Nat. Commun.* **11**, 4158 (2020).

28. K. Marzec-Schmidt, N. Ghosheh, S. R. Stahlschmidt, B. Küppers-Munther, J.

Synnergren, B. Ulfenborg, Artificial intelligence supports automated characterization of differentiated human pluripotent stem cells. *bioRxiv* (2023), p. 2023.01.08.523148.

29. S. Shen, Y. Sun, M. Matsumoto, W. J. Shim, E. Sinniah, S. B. Wilson, T. Werner, Z. Wu, S. T. Bradford, J. Hudson, M. H. Little, J. Powell, Q. Nguyen, N. J. Palpant, Integrating single-cell genomics pipelines to discover mechanisms of stem cell differentiation. *Trends Mol. Med.* **27**, 1135–1158 (2021).

30. R. Yasui, K. Sekine, K. Yamaguchi, Y. Furukawa, H. Taniguchi, Robust parameter design of human induced pluripotent stem cell differentiation protocols defines lineage-specific induction of anterior-posterior gut tube endodermal cells. *Stem Cells*. **39**, 429–442 (2021).

31. S. R. Srivatsan, M. C. Regier, E. Barkan, J. M. Franks, J. S. Packer, P. Grosjean, M. Duran, S. Saxton, J. J. Ladd, M. Spielmann, C. Lois, P. D. Lampe, J. Shendure, K. R. Stevens, C. Trapnell, Embryo-scale, single-cell spatial transcriptomics. *Science*. **373**, 111–117 (2021).

32. Y. Wang, P. Yuan, Z. Yan, M. Yang, Y. Huo, Y. Nie, X. Zhu, J. Qiao, L. Yan, Single-cell multiomics sequencing reveals the functional regulatory landscape of early embryos. *Nat. Commun.* **12**, 1247 (2021).

33. M. Haniffa, D. Taylor, S. Linnarsson, B. J. Aronow, G. D. Bader, R. A. Barker, P. G. Camara, J. G. Camp, A. Chédotal, A. Copp, H. C. Etchevers, P. Giacobini, B. Göttgens, G. Guo, A. Hupalowska, K. R. James, E. Kirby, A. Kriegstein, J. Lundeberg, J. C. Marioni, K. B. Meyer, K. K. Niakan, M. Nilsson, B. Olabi, D. Pe'er, A. Regev, J. Rood, O. Rozenblatt-Rosen, R. Satija, S. A. Teichmann, B. Treutlein, R. Vento-Tormo, S. Webb, Human Cell Atlas Developmental Biological Network, A roadmap for the Human Developmental Cell Atlas. *Nature*. **597**, 196–205 (2021).

34. E. Armingol, A. Officer, O. Harismendy, N. E. Lewis, Deciphering cell-cell interactions and communication from gene expression. *Nat. Rev. Genet.* **22**, 71–88 (2021).

35. J. Cosgrove, A.-M. Lyne, I. Rodriguez, V. Cabeli, C. Conrad, S. Tenreira-Bento, E. Tubeuf, E. Russo, F. Tabarin, Y. Belloucif, S. Maleki-Toyserkani, S. Reed, F. Monaco, A. Ager, C. Lobry, P. Bousso, P. J. Fernández-Marcos, H. Isambert, R. J. Argüello, L. Perié, Metabolically Primed Multipotent Hematopoietic Progenitors Fuel Innate Immunity. *bioRxiv* (2023), p. 2023.01.24.525166.

36. Home - GEO - NCBI, (available at https://www.ncbi.nlm.nih.gov/geo/).

37. Expression Atlas, (available at https://www.ebi.ac.uk/gxa/home).

38. I. Beltagy, K. Lo, A. Cohan, SciBERT: A Pretrained Language Model for Scientific Text. *arXiv [cs.CL]* (2019), (available at http://arxiv.org/abs/1903.10676).

39. J. A. Farrell, Y. Wang, S. J. Riesenfeld, K. Shekhar, A. Regev, A. F. Schier, Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis. *Science*. **360** (2018), doi:10.1126/science.aar3131.

40. M. Mittnenzweig, Y. Mayshar, S. Cheng, R. Ben-Yair, R. Hadas, Y. Rais, E. Chomsky, N. Reines, A. Uzonyi, L. Lumerman, A. Lifshitz, Z. Mukamel, A.-H. Orenbuch, A. Tanay, Y. Stelzer, A single-embryo, single-cell time-resolved model for mouse gastrulation. *Cell*. **184**, 2825–2842.e22 (2021).

41. J. Cao, M. Spielmann, X. Qiu, X. Huang, D. M. Ibrahim, A. J. Hill, F. Zhang, S. Mundlos, L. Christiansen, F. J. Steemers, C. Trapnell, J. Shendure, The single-cell transcriptional landscape of mammalian organogenesis. *Nature*. **566**, 496–502 (2019).

42. L. M. Saunders, S. R. Srivatsan, M. Duran, M. W. Dorrity, B. Ewing, T. H. Linbo, J. Shendure, D. W. Raible, C. B. Moens, D. Kimelman, C. Trapnell, Embryo-scale reverse genetics at single-cell resolution. *Nature*. **623**, 782–791 (2023).

43. X. Huang, J. Henck, C. Qiu, V. K. A. Sreenivasan, S. Balachandran, O. V. Amarie, M. Hrabě de Angelis, R. Y. Behncke, W.-L. Chan, A. Despang, D. E. Dickel, M. Duran, A. Feuchtinger, H. Fuchs, V. Gailus-Durner, N. Haag, R. Hägerling, N. Hansmeier, F. Hennig, C. Marshall, S. Rajderkar, A. Ringel, M. Robson, L. M. Saunders, P. da Silva-Buttkus, N. Spielmann, S. R. Srivatsan, S. Ulferts, L. Wittler, Y. Zhu, V. M. Kalscheuer, D. M. Ibrahim, I. Kurth, U. Kornak, A. Visel, L. A. Pennacchio, D. R. Beier, C. Trapnell, J. Cao, J. Shendure, M. Spielmann, Single-cell, whole-embryo phenotyping of mammalian developmental disorders. *Nature*. **623**, 772–781 (2023).

44. J. Cao, D. R. O'Day, H. A. Pliner, P. D. Kingsley, M. Deng, R. M. Daza, M. A. Zager, K. A. Aldinger, R. Blecher-Gonen, F. Zhang, M. Spielmann, J. Palis, D. Doherty, F. J. Steemers, I. A. Glass, C. Trapnell, J. Shendure, A human cell atlas of fetal gene expression. *Science*. **370** (2020), doi:10.1126/science.aba7721.

45. M. Li, X. Zhang, K. S. Ang, J. Ling, R. Sethi, N. Y. S. Lee, F. Ginhoux, J. Chen, DISCO: a database of Deeply Integrated human Single-Cell Omics data. *Nucleic Acids Res.* **50**, D596–D602 (2022).

46. Single cell portal, (available at https://singlecell.broadinstitute.org/single_cell).

47. Single Cell Expression Atlas, (available at https://www.ebi.ac.uk/gxa/sc/home).

48. UCSC Cell Browser, (available at https://cells.ucsc.edu).

49. M. D. Luecken, F. J. Theis, Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol. Syst. Biol.* **15**, e8746 (2019).

50. P. V. Kharchenko, The triumphs and limitations of computational methods for

scRNA-seq. *Nat. Methods*. **18**, 723–732 (2021).

51. A. Ianevski, A. K. Giri, T. Aittokallio, Fully-automated and ultra-fast cell-type identification using specific marker combinations from single-cell transcriptomic data. *Nat. Commun.* **13**, 1246 (2022).

52. G. Pasquini, J. E. Rojo Arias, P. Schäfer, V. Busskamp, Automated methods for cell type annotation on scRNA-seq data. *Comput. Struct. Biotechnol. J.* **19**, 961–969 (2021).

53. G. Schiebinger, J. Shu, M. Tabaka, B. Cleary, V. Subramanian, A. Solomon, J. Gould, S. Liu, S. Lin, P. Berube, L. Lee, J. Chen, J. Brumbaugh, P. Rigollet, K. Hochedlinger, R. Jaenisch, A. Regev, E. S. Lander, Optimal-Transport Analysis of Single-Cell Gene Expression Identifies Developmental Trajectories in Reprogramming. *Cell*. **176**, 928–943.e22 (2019).

54. A. Wagner, A. Regev, N. Yosef, Revealing the vectors of cellular identity with single-cell genomics. *Nat. Biotechnol.* **34**, 1145–1160 (2016).

55. S. Domcke, J. Shendure, A reference cell tree will serve science better than a reference cell atlas. *Cell*. **186**, 1103–1114 (2023).

56. M. Barron, J. Li, Identifying and removing the cell-cycle effect from single-cell RNA-Sequencing data. *Sci. Rep.* **6**, 33892 (2016).

57. A. Scialdone, K. N. Natarajan, L. R. Saraiva, V. Proserpio, S. A. Teichmann, O. Stegle, J. C. Marioni, F. Buettner, Computational assignment of cell-cycle stage from single-cell transcriptome data. *Methods*. **85**, 54–61 (2015).

58. K. B. Mccauley, K. Kukreja, A. B. Jaffe, A. M. Klein, A map of signaling responses in the human airway epithelium. *bioRxiv* (2022), doi:10.1101/2022.12.21.521460.

59. D. Zemmour, R. Zilionis, E. Kiner, A. M. Klein, D. Mathis, C. Benoist, Single-cell gene expression reveals a landscape of regulatory T cell phenotypes shaped by the TCR. *Nat. Immunol.* **19**, 291–301 (2018).

60. M. M. Hassan, A. Knipper, S. K. K. Santu, ChatGPT as your Personal Data Scientist. *arXiv [cs.CL]* (2023), , doi:10.48550/ARXIV.2305.13657.

61. S. Batzoglou, Large Language Models in Molecular Biology. *Towards Data Science* (2023), (available at https://towardsdatascience.com/large-language-models-in-molecular-biology-9eb6b65d8a30).

62. H. Cui, C. Wang, H. Maan, K. Pang, F. Luo, B. Wang, scGPT: Towards Building a Foundation Model for Single-Cell Multi-omics Using Generative AI. *bioRxiv* (2023), p.

2023.04.30.538439.

63. J. Joung, S. Ma, T. Tay, K. R. Geiger-Schuller, P. C. Kirchgatterer, V. K. Verdine, B. Guo, M. A. Arias-Garcia, W. E. Allen, A. Singh, O. Kuksenko, O. O. Abudayyeh, J. S. Gootenberg, Z. Fu, R. K. Macrae, J. D. Buenrostro, A. Regev, F. Zhang, A transcription factor atlas of directed differentiation. *Cell*. **186**, 209–229.e26 (2023).

64. G. H. T. Yeo, L. Lin, C. Y. Qi, M. Cha, D. K. Gifford, R. I. Sherwood, A Multiplexed Barcodelet Single-Cell RNA-Seq Approach Elucidates Combinatorial Signaling Pathways that Drive ESC Differentiation. *Cell Stem Cell*. **26**, 938–950.e6 (2020).

65. D. Bavli, X. Sun, C. Kozulin, D. Ennis, A. Motzik, A. Biran, S. Brielle, A. Alajem, E. Meshorer, A. Buxboim, O. Ram, CloneSeq: A highly sensitive analysis platform for the characterization of 3D-cultured single-cell-derived clones. *Dev. Cell*. **56**, 1804–1817.e7 (2021).