Stem Cell Reports
Meeting Report
ISSCR
OPEN ACCESS

Challenges for Computational Stem Cell Biology: A Discussion for the Field

Owen Rackham,[1,*] Patrick Cahan,[2] Nancy Mah,[3] Samantha Morris,[4] John F. Ouyang,[1] Anne L. Plant,[5] Yoshiaki Tanaka,[6,8] and Christine A. Wells[7,*]

[1]Program in Cardiovascular and Metabolic Disorders and Centre for Computational Biology, Duke-NUS Medical School, Singapore
[2]Institute for Cell Engineering, Department of Biomedical Engineering, Department of Molecular Biology and Genetics, Johns Hopkins School of Medicine, Baltimore, MD 21205, USA
[3]Biomedical Data and Bioethics Group, Fraunhofer Institute for Biomedical Engineering (IBMT), Joseph-von-Fraunhofer-Weg 1, 66280 Sulzbach, Germany
[4]Department of Developmental Biology, Department of Genetics, and Center of Regenerative Medicine, Washington University School of Medicine in St. Louis, 660 S. Euclid Avenue, Campus Box 8103, St. Louis, MO 63110, USA
[5]Biosystems and Biomaterials Division, National Institute of Standards and Technology, Gaithersburg MD, 20899 USA
[6]Department of Genetics, Yale Stem Cell Center, Yale School of Medicine, New Haven, CT 06520, USA
[7]Centre for Stem Cell Systems, Faculty of Medicine, Dentistry and Health Sciences, The University of Melbourne, Melbourne, VIC 3010, Australia
[8]Present address: Department of Medicine, Maisonneuve-Rosemont Hospital Research Center, University of Montreal, QC, H1T 2M4, Canada
*Correspondence: owen.rackham@duke-nus.edu.sg (O.R.), wells.c@unimelb.edu.au (C.A.W.)
https://doi.org/10.1016/j.stemcr.2020.12.015

The first meetup for Computational Stem Cell Biologists was held at the 2020 annual meeting of the International Society for Stem Cell Research. The discussions highlighted opportunities and barriers to computational stem cell research that require coordinated action across the stem cell sector.

The 2020 International Society for Stem Cell Research (ISSCR) annual meeting demonstrated just how integral Computational Stem Cell Biology (CSCB) has become in the stem cell laboratory. This meeting highlighted that it is the researchers who are combining experimental and computational techniques who are driving the evolution of the stem cell field. This was exemplified by Aviv Regev's presentation in the first plenary session, where she used rich single-cell atlases to computationally infer developmental trajectories and deduce cell lineage maps. Several plenary speakers including Allon Klein, the recipient of the 2020 Dr. Susan Lim Outstanding Young Investigator Award, showcased a variety of computational approaches to better understand stem cell biology.

The 2020 ISSCR annual meeting also hosted a virtual networking event for researchers interested in CSCB. It attracted >130 attendees interested in developments across the discipline, with 44 of those actively engaging in the discussion. The session was meant to serve as a forum to connect researchers across geographical and disciplinary boundaries, and also to serve as an informal survey to identify the most prominent short-term and long-term questions, issues, and challenges that the field faces (as described in detail below). We were glad to see a large spread in the career stage and expertise of attendees, stretching from PIs *"developing experimental and computational techniques to better understand cell reprogramming"* to post-docs who were *"stem cell biologists who sometimes wish to be computational biologists"* and Ph.D. students who were *"starting in September and would love to learn about computational biology"*. As CSCB participant numbers increase, so has the need to establish the community more formally.

Here we report a summary of the topics and recommendations for development of CSCB within the ISSCR community. The session tackled four main questions designed to stretch the discussion from an understanding of where we are now as a field to where we think we should go next and how we can ensure that we have the right network to achieve this.

Question 1: What Are the Current Challenges for CSCB?

When considering the biggest barriers to CSCB research, the question of how to find, reuse, or combine the vast amount of data already produced by stem cell researchers was a recurring theme. A computational researcher wishing to study a particular stem cell question must search through numerous databases, which lack sufficient information, to systematically (1) identify samples from the cell types of interest; (2) refine the search by properties such as disease status, sex, age/developmental stage, or genetic variants; (3) filter by experimental treatments and conditions; and (4) record genetic modification (e.g., reporter gene constructs or gene editing). Even once the studies have been identified, locating all of the data is difficult, as these may be in multiple resources, or equally problematically, replicated in multiple resources without adequate mapping between them. Reanalysis or meta-analysis of combined studies can yield new insights into a system, but curating data for this purpose remains a difficult task: as one participant commented, *"there has already been a disproportionate expenditure of resources available to generate data describing stem cell models, without also investing in ways to ensure that we can use these data effectively"*—and many in the discussion felt that this is a mission for the wider stem cell society.

Finding the right reference data to classify cell types and differentiation stages obtained in stem cell cultures is

Box 1. Fair Data Principles

Findable: descriptions of how the data were generated and processed are complete, and standardized ontologies or other annotations are used to describe the experimental system.
Accessible: data and metadata are curated in public repositories.
Interoperable: data and metadata are in community-agreed-upon formats.
Reusable: data, metadata, and code are provided. (Wilkinson et al., 2016.)

particularly fraught in the absence of high-quality developmental cell atlases. This was most obvious to the group for key tool development areas such as cell identity/classification, cell fate prediction, and cellular engineering. However, cellular types and states cannot be assigned using a reference if they have not been well characterized previously. "Ground truth" datasets are needed for evaluation of computational tools seeking to model pluripotent networks or predict cell fate change. Dynamical data that are the most useful for prediction are particularly rare. To date, large-scale atlas initiatives, such as ENCODE (Moore et al., 2020) and FANTOM (Forrest et al., 2014), have primarily sampled mature tissue types. We note that the Human Cell Atlas has begun to include developmental stages for some tissues (cf. Bock et al., 2020), as well as a recent developmental atlas (Cao et al., 2020; Domcke et al., 2020). There was strong endorsement from participants for a stem cell atlas project to create suitable reference data from differentiating stem cell lines for comparison across multiple -omic technologies. In some instances the appropriate reference data simply do not exist; these gaps should be recognized and funded accordingly. The CSCB community wanted to see atlas efforts make use of global stem cell collections that are genotyped and phenotyped, such as the HipSci (Streeter et al., 2017) or CiRA Foundation iPS Cell Stock (Umekage et al., 2019). This would serve the dual purpose of building a rich genotype-phenotype catalog associated with publicly available lines.

FAIR data is a principle adopted by major funders and by major consortia (Box 1). But what does FAIR mean for the stem cell field? Reusing data that have been created by individual researchers was identified as both an opportunity and a major challenge for the field. It is hard to find relevant and high-quality examples of specific stem cell or developmental stages. Platforms such as Stemformatics, which focus on data curation of public stem cell transcriptome experiments, apply QC metrics that fail 30% of data reviewed from the public domain. Several such niche resources for relevant data exist (e.g. Stemformatics [Choi et al., 2019] and the human pluripotent stem cell registry [Mah et al., 2020]), but what is needed is interoperability between these resources to aid in data sharing and uptake by the community. Metadata is not interoperable in the stem cell field, and we fail to systematically capture the most relevant information in standardized formats, including naming conventions for stem cell lines. Widespread and coordinated adoption of FAIR data practices could accomplish a level of semantic interoperability that would not only enable researchers to find the right kind of data, but to query catalogs of metadata in a machine-readable way. It would then be possible to find suitable data for compilation into well-characterized reference datasets that could serve as benchmarks for the community.

A second major barrier was "how to define a stem cell or derived cell type." The implicit assumption that cell types are static entities is largely a byproduct of how we observe cellular systems rather than a ground truth of the systems themselves. Traditional notions of cell types defined by morphology and histology need to be reconciled with the signatures derived from computational analysis of -omic data. Computational approaches to labeling groups of cells include (1) annotation with prior knowledge—such as the presence of a validated lineage marker, (2) comparison to reference sets drawn from the reference databases discussed above, and (3) machine-learning-based annotation. Marker-based methods are the most conventional way to define a cell type, and since these markers also can be used for cell sorting and isolation, the marker-based cluster labeling supports integrity with subsequent experiments. However, the employment of canonical markers relies on background knowledge of scientists and is highly variable across laboratories. Furthermore, it has become obvious that cellular phenotypes are more heterogeneous than cell types previously defined by the marker expression. As a result, standard cell ontologies that rely on anatomical parent-child relationships don't adequately capture transdifferentiation or reprogramming, and lineage tracing methods are challenging our traditional view of developmental dynamics. Community input is needed to build a minimum consensus definition of cell types, taking into account cell morphology and functional and molecular criteria, but these should also allow for the discovery and definition of new cell types and intermediate cell states.

Assessing the "equivalency" of stem-cell-derived products is challenging when the end-points of differentiation protocols are poorly defined. Variability within cell lines and between laboratories is a commonly accepted problem, so the temptation is to computationally "normalize" these away from the data signal. Indeed, the distinction between true biological variability due to stochastic behavior and technical measurement uncertainty is not fully

appreciated, and improvements for reporting measurement uncertainty would assist computational modeling of cell types and cell states (Plant et al., 2018). The field would be advanced by access to highly characterized stem cell lines that can be commonly used for more robust interpretation of CSCB data. It was noted by the cell bank curators in the discussion that comprehensive characterization of established cell lines is lacking. In particular, cell banks need infrastructure for collection and harmonization of metadata, including cell biology protocols with sufficient detail to enable comparison of data across laboratories. From a standards perspective there are few studies where laboratory variables such as media, extracellular matrix, passaging methods, etc. have been systematically varied and the outcomes of the stem cell phenotype quantified. The lack of benchmark data on well-characterized cell lines leaves the field with too little information about how to interpret observations on new lines or protocols. The depth of characterization and reporting on engineered cell lines that is provided by the Allen Institute for Cell Science (https://www.allencell.org/methods-for-cells-in-the-lab.html) is an example of good practice. We suggest that the ISSCR could lead a consultation process to agree on the minimum core dataset to describe a stem cell line.

Question 2: Where Are Computational Tools Lacking in Their Applicability to Stem Cell Sciences?

The discussion here took on two related topics. The first was the suitability of current approaches to understanding cell-state dynamics and developmental trajectories. The second was better understanding of the hidden molecular variability inherent to current models, particularly when combining data from different molecular scales or laboratory sources.

Although the development of cell classification tools is a highly subscribed research area, the field is failing to set coherent standards needed to assess identification or labeling of cell clusters. Circular logic, such as using the same analysis framework for cluster prediction and cluster validation, is a common problem. Classifiers are frequently assessed for sensitivity (the ability to find cells of a known class) and specificity (the ability to discriminate between cells of different classes), but the stability of the classifier (whether the same molecular features are chosen) is rarely assessed. This speaks to an inherent bias in the analytical frameworks accepted by many journals, accompanied by poor annotation of code, run parameters, or QC metrics, and results in low reproducibility of the cell type assignment. Consequently, it is hard to apply classifiers generated by different laboratories to new datasets, so benchmarking inevitably favors the newly derived method. Ultimately generating a computationally derived "ground truth" of cell type remains contested ground.

The discussion highlighted the need for computational tools to incorporate information from different modalities to fully understand dynamical systems. Computational approaches for predicting cell lineage are in their infancy and don't deal with gapped data well. That is, despite some recent developments of the monocle algorithm (Trapnell et al., 2014), often they will force cells into a trajectory if there is no option to enable partitioning of unrelated cells, even if those cells should be excluded using a biological rationale. Most trajectory inference tools rely on transcriptomic data alone, which may not have sufficient information to deduce the relationship between single cells. Furthermore, mapping the decision points in networks or signaling pathways necessitates better molecular resolution of dynamic molecular processes—RNA isoforms, RNA turn-over, and protein or chromatin modifications, as well as measurements of molecular history in a single cell. The computational challenge is providing meaningful high-order insight from these combined data. For example, in treating cell identity as a moment along a continuum of cell states within any cell type, classifying a cell type may require reporting a description that summarizes the point-in-time state and history and predicts the potential trajectory of a cell.

The integration/harmonization of datasets is an old problem that needs new solutions. Failing to assess and account for technical differences makes it difficult to compare data from different laboratories and greatly reduces the potential increase in statistical/detection power that might be leveraged from the increased sample representation. In some respects, it is easier to generate more data than solve this problem, although the ideal would be to draw on new and old data types and data learnings. Next generation multimodal atlases such as the HuBMAP consortium (HuBMAP Consortium, 2019) experimentally harmonize data collection and computationally anchor multimodal data onto a reference (e.g. Stuart et al., 2019).

Single-cell biology is inherently noisy, which results in higher uncertainty in data analyses. It is possible to assess the variance in a dataset attributable to technical artifacts, such as batch, and account or correct for this. However, these methods should be used cautiously, as they could remove relevant biological data, or worse, if misused may impose class differences where none exist (see Wells and Choi, 2019 for review). More recently, with the emergence of single-cell transcriptomics, there is a resurgence in newer integration tools (e.g., Seurat [Stuart et al., 2019] and scanaroma [Hie et al., 2019]) that exploit the larger volume of data in terms of the number of single cells. These so-called harmonization approaches are incredibly useful to find coherence in complex data but do so by inevitably treating some genuine biological variation as noise. This hidden variability impacts on our capacity to understand

the links between molecular state and cellular phenotype. Because it is difficult to deconvolute noise from biological signals, considerable methodological care (e.g., in the choice of parameters) should be exercised in analyzing these data, and our computational tools should try to provide some estimate of these uncertainties.

Question 3: What Opportunities Are There for Computational Stem Cell Biology and Biologists?

As stem cell researchers, we have as one of our common focuses a desire to better understand and control how, when, and why stem cells change their behavior. It was interesting to see the different approaches to understanding cell phenotypes suggested throughout the conversation—from imaging to -omics, lineage tracing and pseudotime, chromatin behavior and gene output. Perhaps in line with the earlier thoughts on data integration, it will be important to consider how we combine these to get a full picture of cellular flux and the acquisition of new cell states. One of the characteristics of CSCB is the overriding ambition that the tools and analyses should further our understanding of stem cell systems and where possible facilitate hypothesis generation. Especially as we push toward increasingly multimodal and high-dimensional experiments, outputs must be transformed into summarized formats to be available for human interpretation. The open question of how to summarize the research question, experimental approach, data output, and analyses in computer-readable and human-interpretable forms must be addressed if the stem cell community is to gain knowledge from application of these methods.

Even as the data that we collect becomes more sophisticated, we are looking to advance technologies that allow for collection of dynamic data on large numbers of single cells. The consensus was that we have a responsibility as a community to get the most value from these datasets and use them to go deeper into the underlying biology than has been previously possible. One area in which we expect to see rapid growth is in the application and analysis of spatial technologies. Maps that include cell-cell interactions, for example, which are crucial in guiding cell differentiation, will be a vast improvement over the extrapolation of receptor-ligand interactions inferred from gene expression data. A spatial context will aid development of more nuanced differentiation strategies from a signaling perspective. It will also drive deeper questions on cell behavior during cell differentiation and reprogramming. The arrival of these complex spatial datasets will present new challenges for data integration, particularly the integration of established cell imaging approaches with spatial transcriptomics. Nondestructive live imaging to develop cell tracking will help address the spatial and temporal dimensions currently lacking in multimodal data. As a community, we were excited by the opportunity to bridge computational expertise in cell imaging with expertise in current single-cell genomic technologies. Shared opportunities for stem cell and CSCB researchers might arise from community-led stem cell collections (or atlases) that explore the full expression space (molecular, spatial, and temporal) for human stem cells and their progeny. CSCB researchers must be heard in order to ensure that the experimental design for data collection is aligned with the data analytics required.

There is also a tremendous opportunity to deploy generative machine learning methods in CSCB. Typically, to differentiate or reprogram cells to desired target identities, stem cell biologists screen through many transcription factors, ligands or small molecules, and other regulators to guide cell fate. Even with a restricted panel of candidate factors, guided by prior biological knowledge, these experimental approaches require significant investment of time and resources with no guaranteed outcome. Computational approaches that utilize bulk RNA-seq have already shown promise in this area, for example CellNet (Cahan et al., 2014), Mogrify (Rackham et al., 2016), and SeeSawPred (Hartmann et al., 2018). Moving forward, single-cell analysis, such as pseudo-temporal ordering, can identify regulators of cell identity associated with specific branchpoints, instructing new factor cocktails to direct cells down a specific trajectory. Admittedly, these approaches will not necessarily increase target cell yield if the desired cell type is not captured within the captured population. Nevertheless, engineered cells will explore distinct molecular networks and may even result in new synthetic cell types or cell states. If it is possible to find such newly derived characteristics, it will be transformative in our understanding of how molecular networks dictate cellular phenotype. Generative modeling could offer an opportunity to simulate single-cell differentiation and reprogramming. Such methods have been recently used to predict single-cell perturbation responses (Lotfollahi et al., 2019). The potential of these approaches to predict new factors to improve cell engineering remains to be fully explored. A practical motivation for such an exploration was to leverage reproducible observations into design principles for development of engineered functions.

Question 4: What Makes a Good Set of Skills for Computational Stem Cell Biology and How Can We Help the Community Develop These?

Since the CSCB community is new, it is important that we can identify how to develop the field and make this accessible for new researchers entering the field. One area where there was broad agreement was that CSCB needed to sit firmly across the wet and dry aspects of this discipline. A thorough understanding of biology not only allows for

Box 2. Core Skills for CSCB Researchers

Someone working in this field should expect to be working in a very interdisciplinary way and as a result will need to develop a number of "hard" and "soft" skills, and CSCB PIs will need to facilitate this.

Hard skills: statistics and probability, computer programming (proficiency in at least one language), and developmental and stem cell biology.

Soft skills: communication with biologists from all facets (stem cell, mathematical, computational, and data visualization), interaction with theorists from other disciplines (e.g. systems biology or physics), and project management of collaborations.

PI responsibilities: provide opportunities for dry scientists to learn wet techniques and vice versa, create an interdisciplinary environment where knowledge sharing is encouraged, and find funding for computational projects.

the development of the most useful tools but also enables proper interpretation of any predictions or analyses. As part of this it is also the responsibility of the CSCB community to help better inform other members of the stem cell community about the limitations and pitfalls as well as the potential of the emerging suite of tools that are being developed. The list of available tools continues to grow; https://www.scrna-tools.org/ (Zappia et al., 2018) is an excellent resource to track this. For new users, this choice of tools can be overwhelming. Benchmarking papers can help users select the best-performing tools, although new tools emerge faster than these publications. It may also be helpful for the CSCB community to assemble a "starting guide"—a list of recommended core tools for common analyses in stem cell biology.

Many of the tools that are being developed by the CSCB community can be relatively simple to use, but to do so with a limited understanding of the underlying data science principles can lead to spurious outcomes. For biologists entering the CSCB field, it is important to move beyond treating software as black boxes and understand why an algorithm is used and how parameters are chosen. We agreed that the interdisciplinary nature of CSCB research brings additional responsibility to develop the communications skills needed to guide others in the proper interpretation of results. We were encouraged by the motivation of entry-level researchers to improve their understanding of different computational tools. Taking this further, we recommend building opportunities for the CSCB community to conduct coding workshops for wet lab coworkers. Teaching reinforces the teacher's knowledge and also improves the communication of results as it reveals common misunderstandings that wet lab co-workers may face in interpreting data. With this in mind, Box 2 describes some of the foundational wet-lab and dry-lab skills that CSCB draws from.

It was also felt that even among the computational members of the community there was a need to better understand the statistical aspects of our discipline. A danger with sitting on the cutting edge of technology is that the cost of experimentation competes with the need for statistical rigor. For example, it was felt that many single-cell studies ignore fundamental statistical principles because statistics are not part of the lexicon of the group, nor explicit in the workflow of the tools being used for analysis. It will be our job as a community to not only develop tools but also develop a set of best practices similar to those used in other areas of the stem cell community to ensure that analysis results meet a set of statistical standards. This means that the results can be more easily trusted and will be more likely to be replicable. Indeed, many journal publishers now require more informative descriptions of the statistical methods used. As a community, we should commit to providing computational resources, accompanied by comprehensive tutorials and documentation. This strategy will support users with a range of computational backgrounds and skill levels. In concert, creating a strong community will also be crucial to encourage the development of new skills moving forward.

Take-Home Messages

The inaugural meeting of computational stem cell biologists was a huge success in drawing together researchers from different disciplines and with a variety of use-cases for computational stem cell research. The discussions catalyzed the call to further develop the computational community within the ISSCR and beyond. There was enthusiasm for a CSCB symposium in the near future to allow researchers from both the wet and dry fields to hear about advances in developing and using CSCB tools. There are also plans to provide some training events and primers on important topics to both upskill the stem cell community as a whole and also ensure that the best practices are being used throughout.

In summary the lessons taken away from this meeting can be summarized as follows:

1) There is a clear and pressing need to enforce better standards for data generation and deposition that is the responsibility of the stem cell community as a whole to uphold. The CSCB community will be pressing the ISSCR to develop guidelines that cover this and inform its members on the importance of following these.

2) There is a desire to include stem cells more prominently among the ongoing cell atlas projects and, where possible, see these integrated with existing stem cell resources where extensive characterization of the lines has already taken place.
3) It was felt that the importance of funding and supporting the development of computational tools should gain prominence within the stem cell community. Together with this it was felt that placing a requirement to both develop and use tools in a way that ensures fair and unbiased interpretation should become more a priority for the community. In the same way that as a stem cell society it is our responsibility to identify misuse of stem cells, the CSCB community should be working to identify and avoid the misuse or misinterpretation of computational tools.
4) From a methods development point of view, a lot of emphasis was placed in dealing with multimodal datasets and moving beyond a static view of cell type. As the amount of data collected increases, the possibility to create generative tools will also open new areas of stem cell biology and further increase the need for careful and structured data collection.

We would like to thank the ISSCR for facilitating this meetup and all the participants for their involvement. It is an exciting time for the CSCB community, and we look forward to progressing together.

DECLARATION OF INTERESTS

O.R. is a co-founder, scientific advisory board member, and shareholder of Mogrify Ltd, a cell therapy company. C.W. is an Associate Editor with Stem Cell Reports. A.L.P. is an employee of the US Government; these opinions, recommendations, findings, and conclusions do not necessarily reflect the views or policies of NIST or the United States Government. No other authors have a conflict of interest to declare.

REFERENCES

Bock, C., Boutros, M., Camp, J.G., Clarke, L., Clevers, H., Knoblich, J.A., Liberali, P., Regev, A., Rios, A.C., Stegle, O., et al. (2020). The Organoid Cell Atlas: A Rosetta Stone for Biomedical Discovery and Regenerative Therapy. Zenodo https://doi.org/10.5281/zenodo.4001717.

Cahan, P., Li, H., Morris, S.A., Lummertz da Rocha, E., Daley, G.Q., and Collins, J.J. (2014). CellNet: network biology applied to stem cell engineering. Cell *158*, 903–915.

Cao, J., O'Day, D.R., Pliner, H.A., Kingsley, P.D., Deng, M., Daza, R.M., Zager, M.A., Aldinger, K.A., Blecher-Gonen, R., Zhang, F., et al. (2020). A human cell atlas of fetal gene expression. Science *370*, eaba7721.

Choi, J., Pacheco, C.M., Mosbergen, R., Korn, O., Chen, T., Nagpal, I., Englart, S., Angel, P.W., and Wells, C.A. (2019). Stemformatics: visualize and download curated stem cell data. Nucleic Acids Res. *47* (D1), D841–D846.

Domcke, S., Hill, A.J., Daza, R.M., Cao, J., O'Day, D.R., Pliner, H.A., Aldinger, K.A., Pokholok, D., Zhang, F., Milbank, J.H., et al. (2020). A human cell atlas of fetal chromatin accessibility. Science *370*, eaba7612.

Forrest, A.R.R., Kawaji, H., Rehli, M., Baillie, J.K., de Hoon, M.J.L., Haberle, V., Lassmann, T., Kulakovskiy, I.V., Lizio, M., Itoh, M., et al.; FANTOM Consortium and the RIKEN PMI and CLST (DGT) (2014). A promoter-level mammalian expression atlas. Nature *507*, 462–470.

Hartmann, A., Okawa, S., Zaffaroni, G., and Del Sol, A. (2018). SeesawPred: A Web Application for Predicting Cell-fate Determinants in Cell Differentiation. Sci. Rep. *8*, 13355.

Hie, B., Bryson, B., and Berger, B. (2019). Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. Nat. Biotechnol. *37*, 685–691.

HuBMAP Consortium (2019). The human body at cellular resolution: the NIH Human Biomolecular Atlas Program. Nature *574*, 187–192.

Lotfollahi, M., Wolf, F.A., and Theis, F.J. (2019). scGen predicts single-cell perturbation responses. Nat. Methods *16*, 715–721.

Mah, N., Seltmann, S., Aran, B., Steeg, R., Dewender, J., Bultjer, N., Veiga, A., Stacey, G.N., and Kurtz, A. (2020). Access to stem cell data and registration of pluripotent cell lines: The Human Pluripotent Stem Cell Registry (hPSCreg). Stem Cell Res. (Amst.) *47*, 101887.

Moore, J.E., Purcaro, M.J., Pratt, H.E., Epstein, C.B., Shoresh, N., Adrian, J., Kawli, T., Davis, C.A., Dobin, A., Kaul, R., et al.; ENCODE Project Consortium (2020). Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature *583*, 699–710.

Plant, A.L., Becker, C.A., Hanisch, R.J., Boisvert, R.F., Possolo, A.M., and Elliott, J.T. (2018). How measurement science can improve confidence in research results. PLoS Biol. *16*, e2004299.

Rackham, O.J.L., Firas, J., Fang, H., Oates, M.E., Holmes, M.L., Knaupp, A.S., Suzuki, H., Nefzger, C.M., Daub, C.O., Shin, J.W., et al.; FANTOM Consortium (2016). A predictive computational framework for direct reprogramming between human cell types. Nat. Genet. *48*, 331–335.

Streeter, I., Harrison, P.W., Faulconbridge, A., Flicek, P., Parkinson, H., and Clarke, L.; The HipSci Consortium (2017). The human-induced pluripotent stem cell initiative-data resources for cellular genetics. Nucleic Acids Res. *45* (D1), D691–D697.

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W.M., 3rd, Hao, Y., Stoeckius, M., Smibert, P., and Satija, R. (2019). Comprehensive Integration of Single-Cell Data. Cell *177*, 1888–1902.e21.

Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N.J., Livak, K.J., Mikkelsen, T.S., and Rinn, J.L. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat. Biotechnol. *32*, 381–386.

Umekage, M., Sato, Y., and Takasu, N. (2019). Overview: an iPS cell stock at CiRA. Inflamm. Regen. *39*, 17.

Wells, C.A., and Choi, J. (2019). Transcriptional Profiling of Stem Cells: Moving from Descriptive to Predictive Paradigms. Stem Cell Reports *13*, 237–246.

Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L.B., Bourne, P.E., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data *3*, 160018.

Zappia, L., Phipson, B., and Oshlack, A. (2018). Exploring the single-cell RNA-seq analysis landscape with the scRNA-tools database. PLoS Comput. Biol. *14*, e1006245.