Stem Cell Reports
Review
OPEN ACCESS

Transcriptional Profiling of Stem Cells: Moving from Descriptive to Predictive Paradigms

Christine A. Wells[1,*] and Jarny Choi[1]
[1]Centre for Stem Cell Systems, Faculty of Medicine, Dentistry and Health Sciences, University of Melbourne, Parkville 3010, Australia
*Correspondence: wells.c@unimelb.edu.au
https://doi.org/10.1016/j.stemcr.2019.07.008

Transcriptional profiling is a powerful tool commonly used to benchmark stem cells and their differentiated progeny. As the wealth of stem cell data builds in public repositories, we highlight common data traps, and review approaches to combine and mine this data for new cell classification and cell prediction tools. We touch on future trends for stem cell profiling, such as single-cell profiling, long-read sequencing, and improved methods for measuring molecular modifications on chromatin and RNA that bring new challenges and opportunities for stem cell analysis.

In vitro culturing and differentiation are necessary steps in the derivation and study of human stem cells. A fundamental challenge with this study design is that stem and progenitor cells propagated *in vitro* suffer from an identity crisis: to meet the definitions of a stem cell—the capacity to self-renew and the capacity to differentiate to appropriate cell and tissue lineages—researchers must alter the state of that cell by differentiating it. *In-vitro*-derived cells and structures rely on a variety of tests to assess their equivalency to developmental or tissue structures. Integration into tissues and organs may be the gold standard for function, but this is neither possible nor desirable as a routine assay for human stem cell lines. Molecular assays rely on comparative benchmarking of *in-vitro*- and *in-vivo*-derived cells as a functional surrogate: this is where transcriptional profiling is most commonly adopted. These studies aim to compare and contrast the *in vitro* cells with their *in vivo* counterparts, find key transcriptional drivers of a cell type, and sometimes make predictions about cell fates. Unfortunately, this is also an area where we observe frequent misuses of data.

Molecular profiling is an informative companion for functional pluripotency or differentiation assays (e.g., Polanco et al., 2013), and the most commonly profiled molecule is RNA. This is largely driven by the scalability and reliability of sequencing technologies (Figure 1), and the availability of reference genomes to annotate a fragment of expressed sequence to a gene. RNA sequencing (RNA-seq) is inexpensive and quantitative across a large linear range. By providing a catalog of genes active in a cell, RNA-seq also infers the proteins and pathways available to a cell (Kolle et al., 2011; Tonge et al., 2014). Similar systems-scale methods have been developed for proteomic profiling (Rigbolt et al., 2011). Chromatin history via protein binding or histone modifications have been measured using chromatin immunoprecipitation, and chromatin accessibility with the assay for transposase accessible chromatin sequencing (Knaupp et al., 2017; Lee et al., 2014). Even subsets of molecules, such as noncoding or microRNA have been used to benchmark stem-like properties of cells (Clancy et al., 2014; De Rie et al., 2017).

There is, however, a disturbing trend in the use of systems-scale data in the stem cell sciences: studies that benchmark a stem cell-derived phenotype against an *in vivo* counterpart often draw on a small number of public exemplars, with little attention paid to how well the cells that are being used as the standard have been characterized. Despite broad adoption, big-data studies of stem cells can lack reproducibility between laboratories, requiring computational interventions to harmonize data (Volpato et al., 2018): these frequently rely on "black box" methods or third-party analyses, and consequently interpretability of 'omics data can be poor (Figure 2). Data transformation can be co-opted into proving a hypothesis before the comparison is even made. Equally problematic is using profiling experiments as a check-box exercise to reinforce cell type similarity rather than genuinely evaluate the quality of the derived material. A lack of adequate benchmarking leads to iterative science, a missed opportunity to evaluate gaps in developmental patterning or other factors that might otherwise lead to improvements in derivation protocols.

Poor data husbandry is demonstrated by the high fail rate of public stem cell data that is curated by the Stemformatics resource. Stemformatics is a web-based platform that hosts >450 curated public stem cell datasets (>14,000 samples) using FAIR data principles to allow the rapid review of genes and cell types by the stem cell community (Choi et al., 2018; Wells et al., 2013). Stemformatics fails >30% of published data, most commonly because of confounded experimental design or poor data quality, but in over 8% of reviewed studies the underlying primary data were missing, and this was frequently associated with partial data deposition (e.g., control samples only) to obtain an accession number. Failure to provide primary data, obscured sample annotations, poor experimental design, and inappropriate data transformation all contribute to poor reproducibility between studies and increase suspicion of the underlying methodologies. It also highlights a serious problem in the way the sector reviews systems-scale data.

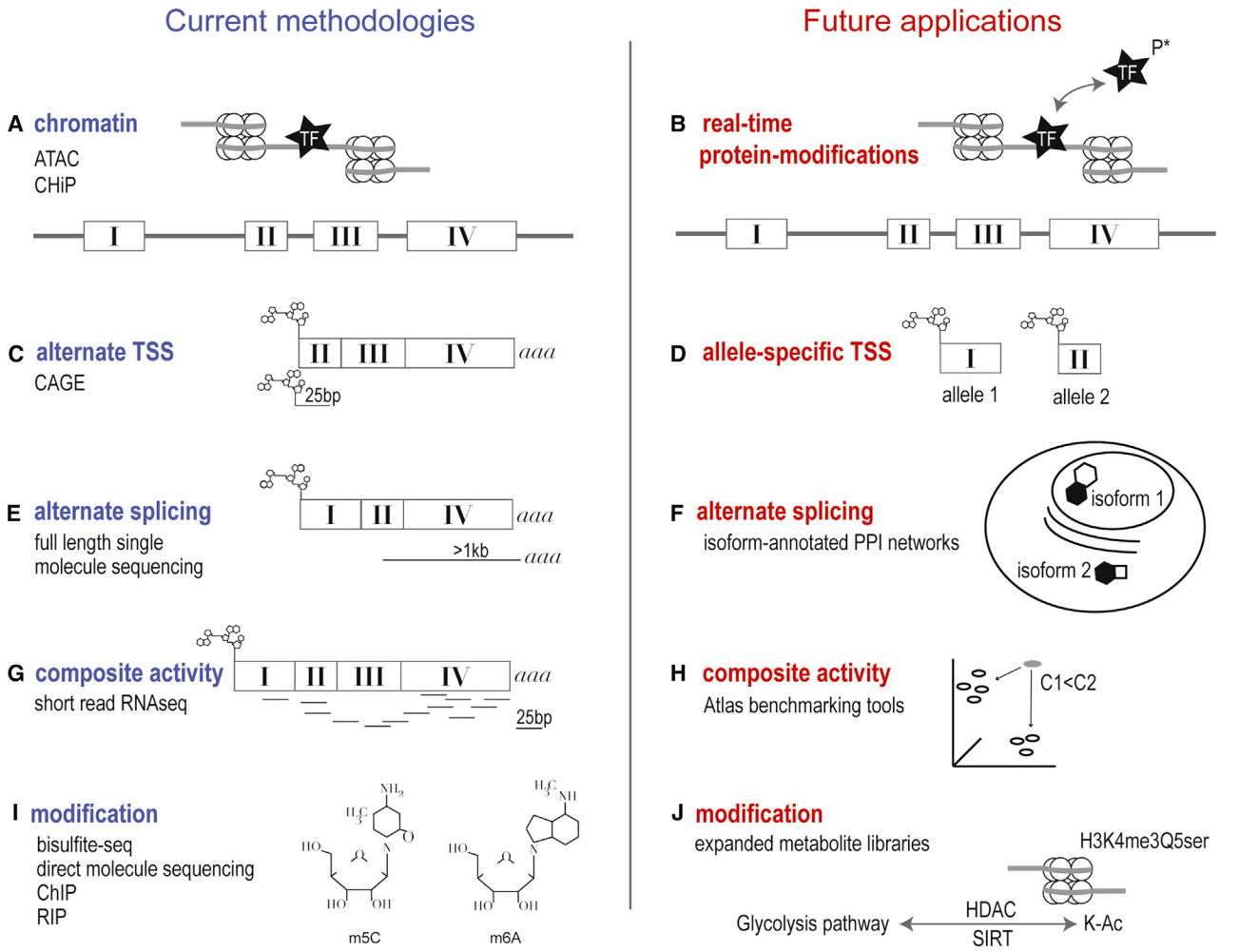


Figure 1. Future Platforms for Molecular Profiling of Stem Cells

(A and B) Current platforms for stem cell profiling include (A) assays of chromatin modifications using chromatin immunoprecipitation (ChIP) and chromatin accessibility using the assay for transposase accessible chromatin sequencing (ATAC). Future modifications (B) will involve real-time measurements of the dynamics of protein phosphorylation during transcriptional programs.

(C and D) (C) Transcription start sites (TSS) are currently measured by capped analysis of gene expression (CAGE), which relies on capture of the methyl-G mRNA cap. Future platforms (D) in single cells will allow discrimination of allelic differences in transcription initiation.

(E and F) (E) Alternate splicing is currently predicted by computational alignment of short sequencing reads across exon boundaries, but these are poor at resolving unique transcripts and commonly result in consensus transcripts. Long-read sequencing, stretching over 1 kb or more are now evolving to explore transcript isoforms. The next iteration of alternate splicing (F) will be computational, moving from gene-centric to isoform-centric interaction networks and enabling the annotation of higher-resolution stem cell pathways.

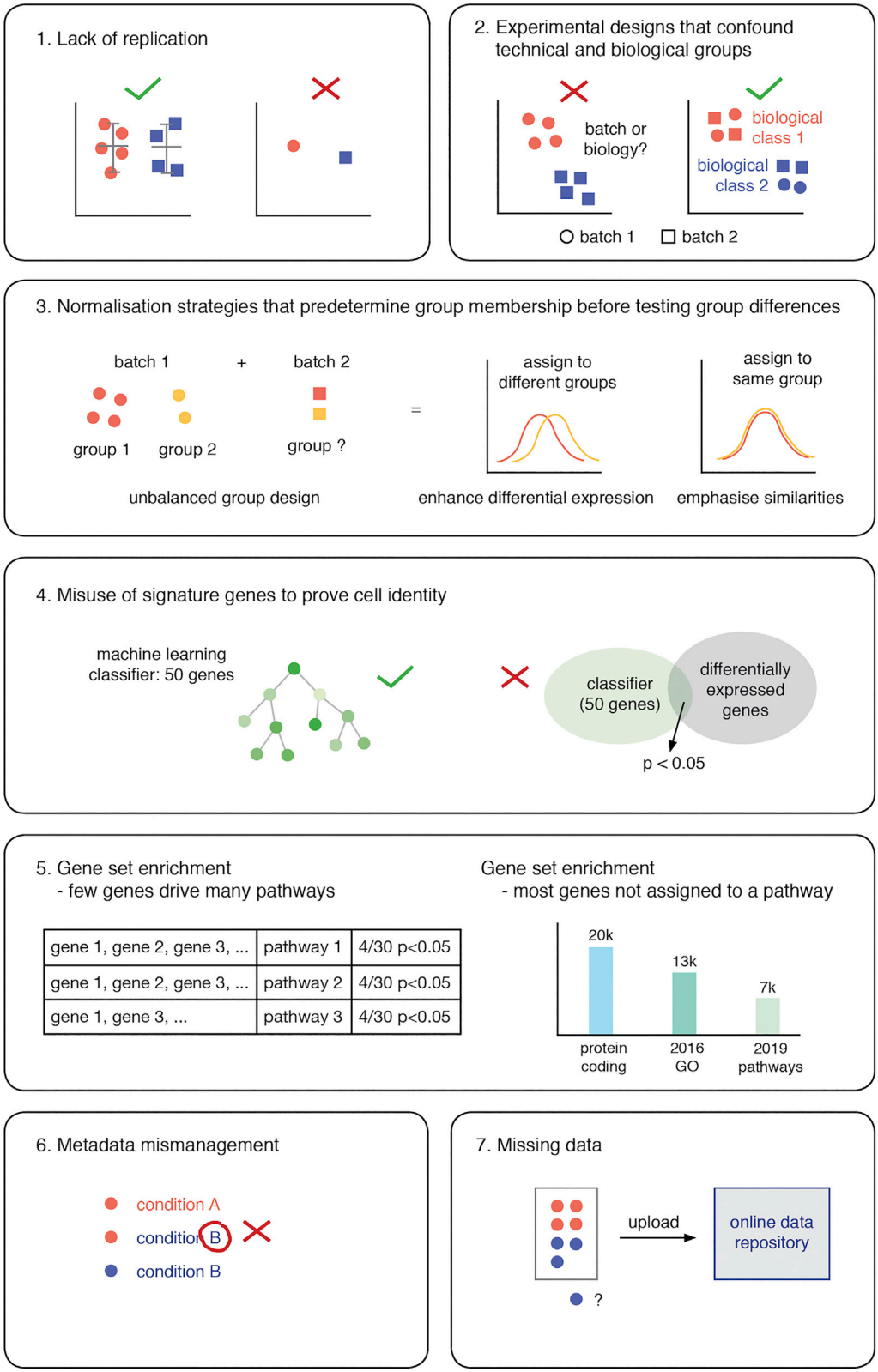
(G and H) (G) Short-read RNA-seq is the most widely adopted method of measuring transcriptional activity from a locus. Future applications of RNA-seq (H) will be the compilation of gold standard transcriptional atlases that allow users to upload and benchmark their own data.

(I) Current methods for measuring nucleotide modifications involve bisulfite DNA sequencing to convert unmethylated-cytosine to uracil, or antibody-based immunoprecipitation methods that bind methylated adenosine or variants of methylated cytosine on RNA (RNA immunoprecipitation [RIP]) or DNA (ChIP).

(J) Future methods will expand the repertoire of metabolites capable of modifying chromatin proteins or RNA, building more immediate linkages between the cell transcriptome and metabolome.

Curated transcriptome atlases such as Stemformatics, and EBI's Expression Atlas (Papatheodorou et al., 2018) therefore offer an important resource—providing user-friendly platforms for stem cell researchers to interrogate relevant studies, a review process that promotes confidence in the underlying data and the means for authors to unambiguously share results with reviewers and readers. These efforts are more than a catalog—important insights for

STEMCR 1352



(legend on next page)

stem cell identity, pluripotency, or reprogramming have come from analysis of integrated expression datasets (e.g., Hussein et al., 2014; Liu et al., 2017), of sufficient size for the application of machine-learning algorithms to build cell-type classifiers (Nguyen et al., 2018), and predictive reprogramming tools (e.g., Rackham et al., 2016).

Integrated Expression Atlases for Categorization of Cell Types

Machine-learning methods for cell classification require much bigger sample sizes than are available from most individual experimental series. Instead, classification tools build on data integration of multiple datasets to train and test the classification model. Examples include Plurinet (Müller et al., 2008) or the Rohart test (Rohart et al., 2016), used to classify specific cell classes, and others that identify stem cell-derived lineages include CellNet (Cahan et al., 2014), KeyGenes (Roost et al., 2015), and CellScore (Mah et al., 2018). By taking an integrated approach, these resources form part of a useful toolkit for profiling pluripotent or tissue-resident cells.

The first, Plurinet, profiled ~150 cell lines on a single microarray platform. The Loring laboratory obtained pluripotent stem cell samples from dozens of laboratories, comparing these with stromal cell or differentiating populations (Müller et al., 2008). A comparative clustering approach allowed the team to develop a manually curated protein-protein interaction map of approximately 300 genes that included key pluripotency transcription factors POU5F1, SOX2, LIN28, and DNMT3B. PluriTest has since been iterated using non-negative matrix factorization to categorize samples into "pluripotent" or "novel" categories (Müller et al., 2011). PluriTest has had widespread uptake in the last decade, not least because of an enabling, community-focused interface that allows users to easily upload and benchmark their own data with the PluriTest

Figure 2. Seven Deadly Sins of Data Analysis

1. Replication. Technical replication measures the reliability of the platform but are not informative in a statistical analysis of biological group differences. These statistical tests broadly assess whether the variance in expression between two groups is greater than the variance expected within that group. Therefore, well-designed studies provide enough replication to properly assess biological variability. In a stem cell context, this would include profiling of multiple stem cell lines, rather than the same line multiple times. As the novelty of scRNA-seq studies dissipates, and the cost of running the experiments decrease, a group of cells from a single individual will no longer be considered sufficient replication of a model. 2. Experimental design: confounding "batch" and "biology." There is no bioinformatic way to separate experimental variable from biological variable if biological groups have been "batched" separately. When groups are batched in this way biological signal can be confounded by experimental variables such as RNA kit, amplification method, platform differences, or even sequencing date. This accounts for >10% of datasets reviewed and failed by the Stemformatics pipeline. 3. Normalization strategies that predetermine group membership before testing group differences. Data integration needs careful consideration. Normalization strategies that preassign groups that are "similar" or "different" will adjust expression values to harmonize members of a group, and this is particularly problematic if the study design is unbalanced (for example, if a study is comparing in-house data with a subset of exemplar samples from an external dataset). This can result in a self-fulfilling prophecy that samples expected to be similar share patterns of expression (for example, group close to one another on a principal-component analysis), because those similarities are enforced by the normalization strategy. Likewise, differences between the groups should be expected to be exaggerated (Nygaard et al., 2016). 4. Misuse of signature genes to prove cell identity. Stem cell researchers may be most comfortable when using antibodies to visualize expression of individual genes in a cell, particularly as methods such as flow cytometry allow us to classify cells based on positive and negative molecular gates. There is an entire literature that spuriously claims that stromal cells are pluripotent because of an anti-OCT4 antibody signal in the cultured cells (Warthemann et al., 2012; Xu et al., 2015). However, computational predictors of cell identity do not work in the same way. The output of a machine-learning classifier is a vector of gene expression, in which the presence or absence of any single molecule is not able to substitute for the whole. These types of classifiers cannot be validated using single-gene PCR measurements or antibody staining, but rather require application of the whole signature for accurate classification. Validation of these signatures rely on their application to new datasets, and continuous assessment of the stability of the signature, and the false-positive/false-negative rates as it is applied to new data. 5. Gene set enrichment: few genes drive many pathways. Results of a gene set enrichment analysis should be interpreted with care, because it is easy to find many gene sets being enriched with low p values due to a small number of the same genes occurring in multiple sets (Mar et al., 2011). This can lead to a false interpretation that many gene sets drive a process, whereas they may be passengers in another process, which confounds the analysis (Venet et al., 2011). 6. Metadata mismanagement. A crucial yet often overlooked aspect of transcriptome data generation is the importance of metadata management. Mislabeled samples or errors in data entry can commonly occur without being detected at all, leading to potentially erroneous conclusions on the data. "Sample swaps" are common in the public databases, in which samples have been clearly assigned to an incorrect group. It is thus very important to give due consideration to this issue from the beginning of the experimental design phase. 7. Missing data. Unfortunately, it is not uncommon to find published studies where some of the crucial raw data are missing from the public repositories where they should reside. A frequent scenario is deposition of partial information (e.g., control samples only) to obtain an accession number. This accounts for more than one-third of the publications reviewed and rejected by the Stemformatics platform. Regardless of the underlying intension behind such missing data, this highlights a serious flaw in the current system of reviews carried out by journals.

algorithm, with a report that is easy for a user to interpret. However, PluriTest is not equivalent to a pluripotency pathway, and a number of genes included in the classification matrix remain functionally unannotated. These have not yet had a role in pluripotency described aside from correlated expression in pluripotent stem cells. This is most problematic if isolated expression of any PluriTest gene is used as evidence of pluripotency (see Figure 2).

Related, but for stromal cell populations, the mixOmics and Stemformatics teams developed a robust transcriptional classifier to address the vexed question of how similar mesenchymal stromal cells (MSCs) are to other stromal cell types, including fibroblasts. They also addressed whether MSCs from different tissues share any comment attributes (Rohart et al., 2016, 2017). By curating hundreds of public stromal cell microarray datasets they identified a 16-gene classifier that discriminates MSCs from dozens of other cell types, including adult stem/progenitor cells, and terminally differentiated muscle, blood, neural, and vascular cell types and fibroblasts. The classifier includes several genes commonly used to phenotype or prospectively enrich for MSCs by flow cytometry, including VCAM1, PDGFRB, ITGA11, and CCDC80, but also identified a number of genes involved in modification of the extracellular matrix via proteoglycan synthesis and catabolism. A web interface allows users to explore the underlying data and benchmark their own data using the classifier (Choi et al., 2018).

Similarly, this user-friendliness for the stem cell community has been a key part of the successful CellNet model. A lineage predictor for cell differentiation/reprogramming studies, CellNet initially integrated data from public microarray studies (Cahan et al., 2014). The platform focused on transcription factors that form tissue-specific gene regulatory networks (GRNs). Its original implementation allowed users to upload microarray data via a web interface, but recent upgrades to the underlying CellNet data to RNA-seq platforms means that potential users need to implement the code themselves (Radley et al., 2017). KeyGenes (Roost et al., 2015) and CellScore (Mah et al., 2018) also provide algorithms to benchmark transcriptome data against an atlas of human fetal tissues (KeyGenes) or a reprogramming score that accounts for distance between parental and final cell types (CellScore). Like CellNet, these require the user to implement code locally, which may represent a major barrier to small stem cell laboratories.

Each of these classification tools takes descriptive expression data to build reproducible classifiers that predict the identity of the cell or tissue being profiled. These are all excellent benchmarking tools that deserve to be more widely adopted but are nevertheless limited in their scope. They are constrained by the parameters of the original model—a pluripotency test, for example, may reliably predict a pluripotent phenotype, but not discriminate between different pluripotent states, and so would not be applicable to test naive pluripotent stem cells.

Nor can PluriTest predict whether a cell type is capable of commitment to a specific lineage. The Stemformatics MSC test cannot predict whether the profiled MSCs have any clinical efficacy, and CellNet, KeyGenes, and CellScore are not designed to predict which factors will drive the reprogramming process, although these are the forerunners to curated networks that will more completely identify factors required for differentiated states (Kinney et al., 2019).

Moving from Descriptive to Predictive Methods

Moving from descriptive to predictive analyses requires transcriptome platforms to fulfill their promise of being an engine for hypothesis generation. This, in turn, requires careful experimental design and a focus on mechanism driven by specific sets of molecules. An example is the use of temporal profiling to predict that differences in HOXA patterning during early mesoderm commitment is a key step in the derivation of mature CD34+ progenitor cells from pluripotent stem cells (Ng et al., 2016). Targeted transcriptome analysis developed a hypothesis that was non-obvious by looking at purely descriptive differences between the end-stage cells and relied on a working knowledge of embryonic hematopoiesis. The hypothesis was tested by manipulation of WNT signals in early differentiation stages to reinstate HOXA9 expression, successfully resolving a missing step in the derivation of blood progenitors from pluripotent stem cells. Approaching 'omics data with a knowledge-based framework is key for such hypothesis generation, and this is not easily outsourced to a bioinformatics pipeline, but requires close collaboration between data analysts and biological specialists.

Mathematical models can successfully theorize and test some aspects of cell signaling, but these are generally constrained by scale to interactions between a small number of genes. For example, Boolean models that transform expression information into binary values (on or off: 0 or 1) have successfully recapitulated the oscillating relationship between pluripotency factors POU5F1 and NANOG (Chickarmane et al., 2006) (reviewed by Herberg and Roeder, 2015). In small-scale studies, the correlation of gene expression with one another or with key experimental variables can build useful gene-phenotype relationships (Langfelder and Horvath, 2008). By abstracting key relationships, such as coexpression into Boolean values, a summary of the network is possible—such as early literature-based networks of key pluripotent transcription factors that can predict stem cell responses to experimental perturbation, including the conditions that support reprogramming to naive pluripotent states (Dunn et al., 2014, 2019).

Cell Mogrify offers an early proof-of-principle of the possibilities of predictive frameworks for stem cell research (Rackham et al., 2016). The hypothesis tested by the Mogrify team was that the transcriptional regulators necessary for cell reprogramming or transdifferentiation could be predicted by comparing GRNs of starting and desired cell populations. Mogrify exploited the increased resolution of proximal promoters and enhancers afforded by capped analysis of gene expression in the FANTOM5 promoter atlas to footprint lineage transcription factors and build highly specific GRNs (Forrest et al., 2014; Rackham et al., 2016). An online tool draws on the FANTOM promoter atlas to provide users with a set of transcription factors predicted to reprogram one cell type to another.

What are the lessons for curators of pluripotency or lineage transcriptome atlases looking to transition from descriptive classification of cell types to predictive models? Straightforward and accessible methods that assist stem cell researchers benchmark cells and assays must be a priority. Curation efforts to refine stem cell and developmental pathways are also needed. This will enhance computational predictions of the drivers of cell lineage, or cell function to fuel the hypothesis generation by the stem cell community.

Single-Cell Profiling: The Collision between Predictive and Descriptive Bioinformatics

Molecular profiles of individual cells offer an exciting opportunity to find new cell types or stage differentiation trajectories. We gain new insights into molecular heterogeneity by profiling thousands of cells in a tissue or a dish and using this to infer phenotypic heterogeneity. The scale of these new methods turns the analysis pipeline on its head—instead of predefining biological classes to compare, cell identity becomes a post hoc analysis challenge. The opportunity to derive molecular networks at the resolution of a single cell has the potential to transform predictive computational methodology. There are, however, a few caveats to consider when interpreting single-cell transcriptome data (scRNA-seq).

The first, is that the data are necessarily sparse. This is a consequence of limitations of RNA capture in the initial cDNA library construction steps, which create gaps in the transcriptome that cannot be resolved with more sequencing reads. It also reflects the labile nature of mRNA, an intermediate that necessarily has a half-life in a cell, depending on rates of transcription, and catabolic steps associated with translation or degradation.

In combination, these factors mean that absence of a transcript in any single-cell library does not mean that the gene is not active in that cell. Rather, examination of changes in the turnover of RNA present new insights into transcriptional regulation in different cell types, or even across transitioning cell states. For example, RNA Velocity (La Manno et al., 2018) exploits changes in RNA processing to predict the next stage of cell differentiation and order single cells into developmental trajectories. Similarly, the assessment of parameters such as transcriptional regulators offers new opportunities to build predictive models from single-cell data. The relationships between groups of cells can be constructed from the transcription factors that regulate coexpressed genes, as for example implemented in SCENIC (van den Oord et al., 2017). Different components of the same GRN might be represented in discrete RNA-seq datasets, such that relationships are revealed not by the genes that are common, but by inferred GRN membership.

Single-cell profiling does offer new perspectives on the dynamic range of gene expression, by sheer force of numbers capturing all possible transcriptional states for a given gene. This can be inferred from the distribution of gene expression across groups of cells, rather than absolute or average expression values. This approach provides valuable additional information to the more traditional Boolean GRNs: populations of cells can be examined for steady-state transcriptional dynamics, such as that observed in pluripotent cells under expected cell-cycle flux. In contrast, transcriptional heterogeneity might be associated with exit from pluripotency. Narrow distributions versus broad distributions are here interpreted as kinetic values in gene expression—representing "slow," "intermediate," or "fast-switching" groups (Lin et al., 2018b). Indeed, this expectation that two cells may share a cell identity but be in different molecular states reflects a new level of sophistication in the interpretation of scRNA-seq data. Cell identity may be stable under different environmental conditions, whereas cell state may be expected to fluctuate even under homeostatic conditions.

Indeed, assigning each cell an identity is the second challenge of scRNA-seq methodologies. Even if profiling relatively well-defined cell populations, for example peripheral blood mononuclear cells, the clustering and classifying of cell classes can be unreliable. In the absence of a "ground truth," for example, when profiling *in-vitro*-derived cells and tissues, even understanding how many discrete cell types, or cell states, should be present is a computational challenge. Different clustering methods derive different cell groups from the same data (Freytag et al., 2018), providing a note of caution for those relying on clustering methods to map expected cell types, or define new cell classes. Nevertheless, this also provides an opportunity for improved cell classification methods that will benefit not only single-cell data series, but also more traditional transcriptome methods implemented in an atlas context.

A relevant example is the landmark study by Petropoulos et al. (2016), which profiled 1,500+ individual cells isolated from human preimplantation embryos. The resulting data showed divergence of trophectoderm, primitive

endoderm, and epiblast lineages at a broad level, but also showed that it was difficult to obtain a finer scale of differentiation. This highlights the basic problem of obtaining samples at sufficient resolution from the embryonic stages that might be equivalent to propagated stem cells, leading to a lack of data that can serve as benchmarks for new stem cell types. Reanalysis of the Petropoulos data using previous knowledge of trophectoderm genes has resulted in reannotation of a large number of cells to extraembryonic tissues (Stirparo et al., 2018). Attempts to combine additional embryonic datasets (Boroviak et al., 2018; Stirparo et al., 2018) have also resulted in the reclassification of individual cells; however, rare and transitioning cell types remain difficult to unambiguously identify, and a surprisingly high proportion of cells remain unclassified.

Integration of scRNA-seq libraries comes with a new set of technical challenges to overcome. Substantive methodological and analytical improvements are needed to accurately identify libraries that result from two cells not one, or genuine differences in transcriptome depth between cell types versus technical differences in library capture and sequence depth. Genetic differences and experimental batch are unavoidably confounded. Harmonization methods inevitably reward large groups of cells and penalise rare cell types that are seen in one experiment but not the other. Nevertheless, we remain optimistic that with increased depth of data, biological signal will be the overriding emergent property of combined datasets, providing new insights into the bounds of cell identity, cell activation, homeostatic flux, and cell state transitions.

Future Opportunities and Challenges

As the data generated by the stem cell community become more sophisticated, and increasingly at a higher cell or molecular resolution, we expect that the approaches for benchmarking and analyzing *in-vitro*-derived cell types will also improve. We anticipate increased reliance on exemplar atlases, improved methods for data integration and comparison, and, in doing so, moving from anecdotal comparisons to generalizable and reproducible observations. The era of single-cell profiling has just begun, and it is already clear that new methods for lineage tracing (Biddy et al., 2018; Lin et al., 2018a), as well as integrated perturbation, chromatin profiling, and scRNA-seq (Dixit et al., 2016), will address many questions about the molecular program of differentiation, reprogramming, or other cell state transitions.

We predict that future methods that allow for exploration of single-molecule modifications will drive new relationships between the metabolites of a cell and the metabolic modification of RNA or chromatin proteins. While riboswitches—interactions between metabolites and RNA that control stability, splicing, and translation have been described in lower eukaryotes (Caron et al., 2012; Donovan et al., 2018)—the metabolite-transcriptome axis in higher eukaryotes has focused on modification of proteins that regulate RNA stability and splicing (Galván-Peña et al., 2019). Likewise, histone deacetylases have dual roles in the modification of metabolic pathways, such as glycolysis (reviewed in Shakespear et al., 2018). More recently, observations of metabolites such as serotonin acting on histone proteins indicate a role for metabolic processes to impact on chromatin and, concomitantly, on gene expression (Farrelly et al., 2019). New methods for rapid profiling of the phosphoproteome offer opportunities to measure temporally sensitive phosphorylation on chromatin, and the impact of these on transcription factor stabilization and turnover, as well as transcription elongation and termination (Engholm-Keller et al., 2019). These preliminary studies foreshadow opportunities to target the role of specific metabolites as short- and long-term modifiers of transcriptional programs.

Improvements to long-read sequencing will lead to improved molecular resolution in stem cell and developmental pathways—moving from generic to specific isoforms, interactions, and cell partitions (reviewed in Arzalluz-Luque Á and Conesa, 2018). Computationally, data integration will increasingly use variable-selection methodologies to identify key molecular features, rather than brute-force data merges (Singh et al., 2019); gene-centric analysis methods will need to become isoform-centric; data curation will remain a key component of the stem cell atlas, and pathway curation will become an increasingly important area of research. The type of data obtained from multi-omic profiling at the single-cell level will necessarily drive new analytical approaches beyond classical linear regression models, allowing for nonlinear relationships between chromatin, transcriptome, proteome, and metabolome; and the connectome (Boisset et al., 2018)—the partnership between cells in a niche—will become an important future focus.

AUTHOR CONTRIBUTIONS

C.A.W. and J.C. shared discussions, coauthored, and edited the manuscript.

ACKNOWLEDGMENTS

C.A.W. is funded by the Australian Research Council (FT150100330) and Stem Cells Australia (SR110001002). J.C. is funded by the University of Melbourne Centre for Stem Cell Systems and philanthropic support from the JEM research foundation.

REFERENCES

Arzalluz-Luque, Á., and Conesa, A. (2018). Single-cell RNAseq for the study of isoforms-how is that possible? Genome Biol. https://doi.org/10.1186/s13059-018-1496-z.

Biddy, B.A., Kong, W., Kamimoto, K., Guo, C., Waye, S.E., Sun, T., and Morris, S.A. (2018). Single-cell mapping of lineage and identity in direct reprogramming. Nature *564*, 219–224.

Boisset, J.C., Vivié, J., Grün, D., Muraro, M.J., Lyubimova, A., and Van Oudenaarden, A. (2018). Mapping the physical network of cellular interactions. Nat. Methods *15*, 547–553.

Boroviak, T., Stirparo, G.G., Dietmann, S., Hernando-Herraez, I., Mohammed, H., Reik, W., Smith, A., Sasaki, E., Nichols, J., and Bertone, P. (2018). Single cell transcriptome analysis of human, marmoset and mouse embryos reveals common and divergent features of preimplantation development. Development *145*. https://doi.org/10.1242/dev.167833.

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.

Caron, M.-P., Bastet, L., Lussier, A., Simoneau-Roy, M., Masse, E., and Lafontaine, D.A. (2012). Dual-acting riboswitch control of translation initiation and mRNA decay. Proc. Natl. Acad. Sci. U S A *109*, E3444–E3453.

Chickarmane, V., Troein, C., Nuber, U.A., Sauro, H.M., and Peterson, C. (2006). Transcriptional dynamics of the embryonic stem cell switch. PLoS Comput. Biol. *2*, e123, Public Library of Science.

Choi, J., Pacheco, C.M., Mosbergen, R., Korn, O., Chen, T., Nagpal, I., Englart, S., Angel, P.W., and Wells, C.A. (2018). Stemformatics: visualize and download curated stem cell data. Nucleic Acids Res., 2–7. Oxford University Press. https://doi.org/10.1093/nar/gky1064.

Clancy, J.L., Patel, H.R., Hussein, S.M.I., Tonge, P.D., Cloonan, N., Corso, A.J., Li, M., Lee, D.S., Shin, J.Y., Wong, J.J.L., et al. (2014). Small RNA changes en route to distinct cellular states of induced pluripotency. Nat. Commun. *5*, 5522.

Dixit, A., Parnas, O., Li, B., Chen, J., Fulco, C.P., Jerby-Arnon, L., Marjanovic, N.D., Dionne, D., Burks, T., Raychowdhury, R., et al. (2016). Perturb-seq: dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. Cell *167*, 1853–1866.e17.

Donovan, P.D., Holland, L.M., Lombardi, L., Coughlan, A.Y., Higgins, D.G., Wolfe, K.H., and Butler, G. (2018). TPP riboswitch-dependent regulation of an ancient thiamin transporter in *Candida*. PLoS Genet. *14*. https://doi.org/10.1371/journal.pgen.1007429.

Dunn, S., Li, M.A., Carbognin, E., Smith, A., and Martello, G. (2019). A common molecular logic determines embryonic stem cell self-renewal and reprogramming. EMBO J. *38*, e100003.

Dunn, S.J., Martello, G., Yordanov, B., Emmott, S., and Smith, A.G. (2014). Defining an essential transcription factor program for naïve pluripotency. Science https://doi.org/10.1126/science.1248882.

Engholm-Keller, K., Waardenberg, A.J., Müller, J.A., Wark, J.R., Fernando, R.N., Arthur, J.W., Robinson, P.J., Dietrich, D., Schoch, S., and Graham, M.E. (2019). The temporal profile of activity-dependent presynaptic phospho-signalling reveals longlasting patterns of poststimulus regulation. PLoS Biol. *17*, e3000170. https://doi.org/10.1371/journal.pbio.3000170.

Farrelly, L.A., Thompson, R.E., Zhao, S., Lepack, A.E., Lyu, Y., Bhanu, N.V., Zhang, B., Loh, Y.-H.E., Ramakrishnan, A., Vadodaria, K.C., et al. (2019). Histone serotonylation is a permissive modification that enhances TFIID binding to H3K4me3. Nature *567*, 535–539.

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. (2014). A promoter-level mammalian expression atlas. Nature *507*, 462–470.

Freytag, S., Tian, L., Lönnstedt, I., Ng, M., and Bahlo, M. (2018). Comparison of clustering tools in R for medium-sized 10x Genomics single-cell RNA-sequencing data. F1000Res. *7*, 1297.

Galván-Peña, S., Carroll, R.G., Newman, C., Hinchy, E.C., Palsson-McDermott, E., Robinson, E.K., Covarrubias, S., Nadin, A., James, A.M., Haneklaus, M., et al. (2019). Malonylation of GAPDH is an inflammatory signal in macrophages. Nat. Commun. *10*. https://doi.org/10.1038/s41467-018-08187-6.

Herberg, M., and Roeder, I. (2015). Computational modelling of embryonic stem-cell fate control. Development *142*, 2250–2260.

Hussein, S.M.I., Puri, M.C., Tonge, P.D., Benevento, M., Corso, A.J., Clancy, J.L., Mosbergen, R., Li, M., Lee, D.S., Cloonan, N., et al. (2014). Genome-wide characterization of the routes to pluripotency. Nature *516*, 198–206.

Kinney, M.A., Vo, L.T., Frame, J.M., Barragan, J., Conway, A.J., Li, S., Wong, K.-K., Collins, J.J., Cahan, P., North, T.E., et al. (2019). A systems biology pipeline identifies regulatory networks for stem cell engineering. Nat. Biotechnol. *37*, 810–818.

Knaupp, A.S., Buckberry, S., Pflueger, J., Lim, S.M., Ford, E., Larcombe, M.R., Rossello, F.J., de Mendoza, A., Alaei, S., Firas, J., et al. (2017). Transient and permanent reconfiguration of chromatin and transcription factor occupancy drive reprogramming. Cell Stem Cell *21*, 834–845.e6.

Kolle, G., Shepherd, J.L., Gardiner, B., Kassahn, K.S., Cloonan, N., Wood, D.L.A., Nourbakhsh, E., Taylor, D.F., Wani, S., Chy, H.S., et al. (2011). Deep-transcriptome and ribonome sequencing redefines the molecular networks of pluripotency and the extracellular space in human embryonic stem cells. Genome Res. *21*, 2014–2025.

Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics https://doi.org/10.1186/1471-2105-9-559.

Lee, D.S., Shin, J.Y., Tonge, P.D., Puri, M.C., Lee, S., Park, H., Lee, W.C., Hussein, S.M.I., Bleazard, T., Yun, J.Y., et al. (2014). An epigenomic roadmap to induced pluripotency reveals DNA methylation as a reprogramming modulator. Nat. Commun. *5*. https://doi.org/10.1038/ncomms6619.

Lin, D.S., Kan, A., Gao, J., Crampin, E.J., Hodgkin, P.D., and Naik, S.H. (2018a). DiSNE movie visualization and assessment of clonal kinetics reveal multiple trajectories of dendritic cell development. Cell Rep. *22*, 2601–2614.

Lin, Y.T., Hufton, P.G., Lee, E.J., and Potoyan, D.A. (2018b). A stochastic and dynamical view of pluripotency in mouse embryonic stem cells. PLoS Comput. Biol. *14*, e1006000.

Liu, X., Nefzger, C.M., Rossello, F.J., Chen, J., Knaupp, A.S., Firas, J., Ford, E., Pflueger, J., Paynter, J.M., Chy, H.S., et al. (2017).

Comprehensive characterization of distinct states of human naive pluripotency generated by reprogramming. Nat. Methods *14*, 1055–1062. https://doi.org/10.1038/nmeth.4436.

Mah, N., Taškova, K., El Amrani, K., Hariharan, K., and Andrade-navarro, M.A. (2018). Evaluating cell identity from transcription profiles. BioRxiv, 250431. https://doi.org/10.1101/250431.

La Manno, G., Soldatov, R., Zeisel, A., Braun, E., Hochgerner, H., Petukhov, V., Lidschreiber, K., Kastriti, M.E., Lönnerberg, P., Furlan, A., et al. (2018). RNA velocity of single cells. Nature, 494–498. https://doi.org/10.1038/s41586-018-0414-6.

Mar, J.C., Matigian, N.A., Quackenbush, J., and Wells, C.A. (2011). Attract: a method for identifying core pathways that define cellular phenotypes. PLoS One *6*. https://doi.org/10.1371/journal.pone.0025445.

Müller, F.J., Laurent, L.C., Kostka, D., Ulitsky, I., Williams, R., Lu, C., Park, I.H., Rao, M.S., Shamir, R., Schwartz, P.H., et al. (2008). Regulatory networks define phenotypic classes of human stem cell lines. Nature *455*, 401–405.

Müller, F.J., Schuldt, B.M., Williams, R., Mason, D., Altun, G., Papapetrou, E.P., Danner, S., Goldmann, J.E., Herbst, A., Schmidt, N.O., et al. (2011). A bioinformatic assay for pluripotency in human cells. Nat. Methods *8*, 315–317.

Ng, E.S., Azzola, L., Bruveris, F.F., Calvanese, V., Phipson, B., Vlahos, K., Hirst, C., Jokubaitis, V.J., Yu, Q.C., Maksimovic, J., et al. (2016). Differentiation of human embryonic stem cells to HOXA+ hemogenic vasculature that resembles the aorta-gonad-mesonephros. Nat. Biotechnol. *34*, 1168–1179. https://doi.org/10.1038/nbt.3702.

Nguyen, Q.H., Lukowski, S.W., Chiu, H.S., Senabouth, A., Bruxner, T.J.C., Christ, A.N., Palpant, N.J., and Powell, J.E. (2018). Single-cell RNA-seq of human induced pluripotent stem cells reveals cellular heterogeneity and cell state transitions between subpopulations. Genome Res. *28*, 1053–1066. Cold Spring Harbor Laboratory Press. https://doi.org/10.1101/gr.223925.117.

Nygaard, V., Rødland, E.A., and Hovig, E. (2016). Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses. Biostatistics *17*, 29–39. https://doi.org/10.1093/biostatistics/kxv027.

van den Oord, J., Atak, Z.K., Geurts, P., Aerts, S., Huynh-Thu, V.A., Moerman, T., González-Blas, C.B., Rambow, F., Aerts, J., Imrichova, H., et al. (2017). SCENIC: single-cell regulatory network inference and clustering. Nat. Methods *14*, 1083–1086. https://doi.org/10.1038/nmeth.4463.

Papatheodorou, I., Fonseca, N.A., Keays, M., Tang, Y.A., Barrera, E., Bazant, W., Burke, M., Füllgrabe, A., Fuentes, A.M.P., George, N., et al. (2018). Expression Atlas: gene and protein expression across multiple studies and organisms. Nucleic Acids Res. *46*, D246–D251. https://doi.org/10.1093/nar/gkx1158.

Petropoulos, S., Deng, Q., Panula, S.P., Codeluppi, S., Reyes, A.P., Linnarsson, S., Sandberg, R., and Lanner, F. (2016). Single-cell RNA-seq reveals lineage and X-chromosome dynamics in human preimplantation embryos. Cell *165*, 1012–1026. https://doi.org/10.1016/j.cell.2016.03.023.

Polanco, J.C., Ho, M.S.H., Wang, B., Zhou, Q., Wolvetang, E., Mason, E., Wells, C.A., Kolle, G., Grimmond, S.M., Bertoncello, I., et al. (2013). Identification of unsafe human induced pluripotent stem cell lines using a robust surrogate assay for pluripotency. Stem Cells *31*, 1498–1510.

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

Radley, A.H., Schwab, R.M., Tan, Y., Kim, J., Lo, E.K.W., and Cahan, P. (2017). Assessment of engineered cells using CellNet and RNA-seq. Nat. Protoc. *12*, 1089–1102. https://doi.org/10.1038/nprot.2017.022.

De Rie, D., Abugessaisa, I., Alam, T., Arner, E., Arner, P., Ashoor, H., Åström, G., Babina, M., Bertin, N., Burroughs, A.M., et al. (2017). An integrated expression atlas of miRNAs and their promoters in human and mouse. Nat. Biotechnol. *35*, 872–878. https://doi.org/10.1038/nbt.3947.

Rigbolt, K.T.G., Prokhorova, T.A., Akimov, V., Henningsen, J., Johansen, P.T., Kratchmarova, I., Kassem, M., Mann, M., Olsen, J.V., and Blagoev, B. (2011). System-wide temporal characterization of the proteome and phosphoproteome of human embryonic stem cell differentiation. Sci. Signal. *4*. https://doi.org/10.1126/scisignal.2001570.

Rohart, F., Mason, E.A., Matigian, N., Mosbergen, R., Korn, O., Chen, T., Butcher, S., Patel, J., Atkinson, K., Khosrotehrani, K., et al. (2016). A molecular classification of human mesenchymal stromal cells. PeerJ *4*, e1845. https://doi.org/10.7717/peerj.1845.

Rohart, F., Gautier, B., Singh, A., and Lê Cao, K.-A. (2017). mixOmics: an R package for 'omics feature selection and multiple data integration. PLoS Comput. Biol. *13*, e1005752, Public Library of Science.

Roost, M.S., Van Iperen, L., Ariyurek, Y., Buermans, H.P., Arindrarto, W., Devalla, H.D., Passier, R., Mummery, C.L., Carlotti, F., De Koning, E.J.P., et al. (2015). KeyGenes, a tool to probe tissue differentiation using a human fetal transcriptional atlas. Stem Cell Reports *4*, 1112–1124. https://doi.org/10.1016/j.stemcr.2015.05.002.

Shakespear, M.R., Iyer, A., Cheng, C.Y., Das Gupta, K., Singhal, A., Fairlie, D.P., and Sweet, M.J. (2018). Lysine deacetylases and regulated glycolysis in macrophages. Trends Immunol., 473–488. https://doi.org/10.1016/j.it.2018.02.009.

Singh, A., Shannon, C.P., Gautier, B., Rohart, F., Vacher, M., Tebbutt, S.J., and Lê Cao, K.-A. (2019). DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays. Bioinformatics https://doi.org/10.1093/bioinformatics/bty1054.

Stirparo, G.G., Boroviak, T., Guo, G., Nichols, J., Smith, A., and Bertone, P. (2018). Integrated analysis of single-cell embryo data yields a unified transcriptome signature for the human pre-implantation epiblast. Development *145*. https://doi.org/10.1242/dev.158501.

Tonge, P.D., Corso, A.J., Monetti, C., Hussein, S.M.I., Puri, M.C., Michael, I.P., Li, M., Lee, D.S., Mar, J.C., Cloonan, N., et al. (2014). Divergent reprogramming routes lead to alternative stem-cell states. Nature *516*, 192–197. https://doi.org/10.1038/nature14047.

Venet, D., Dumont, J.E., and Detours, V. (2011). Most random gene expression signatures are significantly associated with breast cancer outcome. PLoS Comput. Biol. *7*. https://doi.org/10.1371/journal.pcbi.1002240.

Volpato, V., Smith, J., Sandor, C., Ried, J.S., Baud, A., Handel, A., Newey, S.E., Wessely, F., Attar, M., Whiteley, E., et al. (2018). Reproducibility of molecular phenotypes after long-term differentiation to human iPSC-derived neurons: a multi-site omics study. Stem Cell Reports *11*, 897–911. https://doi.org/10.1016/j.stemcr.2018.08.013.

Warthemann, R., Eildermann, K., Debowski, K., and Behr, R. (2012). False-positive antibody signals for the pluripotency factor OCT4A (POU5F1) in testis-derived cells may lead to erroneous data and misinterpretations. Mol. Hum. Reprod. *18*, 605–612. https://doi.org/10.1093/molehr/gas032.

Wells, C.A., Mosbergen, R., Korn, O., Choi, J., Seidenman, N., Matigian, N.A., Vitale, A.M., and Shepherd, J. (2013). Stemformatics: visualisation and sharing of stem cell gene expression. Stem Cell Res. *10*, 387–395. https://doi.org/10.1016/j.scr.2012.12.003.

Xu, G., Yang, L., Zhang, W., and Wei, X. (2015). All the tested human somatic cells express both Oct4A and its pseudogenes but express Oct4A at much lower levels compared with its pseudogenes and human embryonic stem cells. Stem Cells Dev. *24*, 1546–1557. https://doi.org/10.1089/scd.2014.0552.