

Ontology challenges for the stem cell community: towards integrative data mining in the Stemformatics atlas

Chris Pacheco Rivera¹, Rowland Mosbergen¹, Othmar Korn²
Tyrone Chen¹, Isha Nagpal¹ and Christine A. Wells^{1,3}

¹ Centre for Stem Cell Systems, Department of Anatomy and Neuroscience, MDHS, The University of Melbourne, 30 Royal Parade Parkville, Melbourne, VIC 3010, Australia

² Australian Institute for Bioengineering and Nanotechnology, The University of Queensland, Building 75 Cnr College Rd & Cooper Rd, Brisbane, QLD 4072, Australia

³ The Walter and Eliza Hall Research Institute, 1G Royal Parade, Parkville, Melbourne, VIC 3010, Australia

ABSTRACT

Stemformatics (www.stemformatics.org) is a web-based pocket dictionary targeted to stem cell biologists with limited knowledge in bioinformatics. It holds a growing collection of manually-curated and high-quality public stem cell datasets. It allows easy visualisation and comparison of gene expression profiles across different platforms from different laboratory sources in mouse and human. Stemformatics hosts >344 public datasets, with >7060 human and >1853 mouse samples.

We have a large set of curated data, primarily transcriptome, including microarray and RNAseq, as well as unconventional “omics” platforms such as ChIPSeq, miRNA, proteomics, and metabolomics data. Stem cell metadata fall into two broad categories – (1) the description of endogenous stem cells, isolated using cell surface proteins and characterised on their originating tissue or developmental stage. (2) in vitro derived cells, including a variety of reprogrammed, as well as directed differentiation protocols aimed at recapitulating a specific class of cell.

Here, we review the challenges of adapting ontology standards to fit a stem cell framework and implementation in Stemformatics. Our aim is to develop a stem cell ontology that can describe different cell types and provide information of their biological background. Stemformatics has started to standardize specific naming conventions to differentiate several cell types. Building a dictionary of stem cell types and their integration into existing ontology resources will be included in the near future.

Annotation of samples metadata is a difficult task when it involves description of synthetic cells whose provenance is hard to capture using existing anatomical ontologies. Induced pluripotent stem cells do not have a developmental equivalent, because these are artificially transformed from mature cell types, such as a skin or blood biopsy. Equally problematic is the description of samples in intermediate states (mid-reprogramming, or mid-differentiation) as these include cell states that have not been defined before and do not have a developmental or anatomical equivalent. Our ontologies must capture information about the source, manipulation, characterisation of

the starting materials, as well as any transformation to a new cell type in the laboratory.

Stemformatics hosts a large amount of primary data, which leads to challenges in data aggregation and downstream analysis if sample annotations are not well standardised. Dealing with several related cell types and cell lines increases the complexity of this problem. Historically, we have had several annotators with different backgrounds who have inadvertently introduced inconsistencies because of a lack of standardised ontology, the rapid pace of change in the field, and a lack of appropriate resources to cross check new samples against existing ontologies.

Large-scale gene expression profiling approaches are used by the stem cell community to for the purposes of bench-marking cell types, defining stem cell states, and characterising molecular networks including predictions of cell-cell and molecular relationships. Deep mining of Stemformatics datasets have facilitated the identification of novel cell types, resolved questions about phenotype similarities between stromal subpopulations, and identified genes involved in maintenance of pluripotency and the differentiation to embryonic lineages.

Stemformatics facilitates data visualisation including interactive graphs like Yugene, where the ranking of all samples can be visualised across a single gene. Furthermore, the Rohart Mesenchymal Stromal Cells (MSC) test is an example of using well-curated data and metadata to create an algorithm to classify stem cells behaving like MSCs.

* To whom correspondence should be addressed:
chris.pacheco@unimelb.edu.au