D788–D793 *Nucleic Acids Research, 2018, Vol. 46, Database issue* *Published online 17 October 2017*
doi: 10.1093/nar/gkx921

StemMapper: a curated gene expression database for stem cell lineage analysis

José P. Pinto[1,2,*], Rui S.R. Machado[1,2], Ramiro Magno[2,3], Daniel V. Oliveira[4], Susana Machado[2,3], Raquel P. Andrade[2,3,5], José Bragança[2,3,5], Isabel Duarte[2,3] and Matthias E. Futschik[1,2,6,7,*]

[1]Systems Biology and Bioinformatics Laboratory (SysBioLab), Universidade do Algarve, Faro, 8005-139, Portugal, [2]Center for Biomedical Research (CBMR), Universidade do Algarve, Faro, 8005-139, Portugal, [3]Algarve Biomedical Center (ABC), Campus Gambelas, Ed. 2 - Ala Norte 8005-139, Faro, Portugal, [4]Center for Alzheimer Research, Department of Neurobiology, Care Sciences and Society, Karolinska Institutet, Stockholm 14157, Sweden, [5]Department of Medicine and Biomedical Sciences, Universidade do Algarve 8005-139, Faro, Portugal, [6]Centre of Marine Sciences (CCMAR), Universidade do Algarve, Faro 8005-139, Portugal and [7]School of Biomedical & Healthcare Sciences, Plymouth University Peninsula Schools of Medicine and Dentistry, Plymouth, Devon PL4 8AA, UK

Received August 15, 2017; Revised September 19, 2017; Editorial Decision September 28, 2017; Accepted October 05, 2017

ABSTRACT

Transcriptomic data have become a fundamental resource for stem cell (SC) biologists as well as for a wider research audience studying SC-related processes such as aging, embryonic development and prevalent diseases including cancer, diabetes and neurodegenerative diseases. Access and analysis of the growing amount of freely available transcriptomics datasets for SCs, however, are not trivial tasks. Here, we present StemMapper, a manually curated gene expression database and comprehensive resource for SC research, built on integrated data for different lineages of human and mouse SCs. It is based on careful selection, standardized processing and stringent quality control of relevant transcriptomics datasets to minimize artefacts, and includes currently over 960 transcriptomes covering a broad range of SC types. Each of the integrated datasets was individually inspected and manually curated. StemMapper's user-friendly interface enables fast querying, comparison, and interactive visualization of quality-controlled SC gene expression data in a comprehensive manner. A proof-of-principle analysis discovering novel putative astrocyte/neural SC lineage markers exemplifies the utility of the integrated data resource. We believe that StemMapper can open the way for new insights and advances in SC research by greatly simplifying the access and analysis of SC transcriptomic data. StemMapper is freely accessible at http://stemmapper.sysbiolab.eu.

INTRODUCTION

Stem Cells (SCs) present a unique capacity for self-renewal and differentiation into other cell types. These features have made them the subject of intense research not only in fundamental cell and developmental biology (1–3), but also in a biomedical context in the fields of regenerative medicine (4–6), cancer progression and treatment (7), drug discovery and testing (8,9) and modelling of human diseases (8–10).

While sharing the capacity for self-renewal and generation of differentiated progeny, a large variety of distinct SC types exist. In mammals, SCs can be classified according to their developmental structure or tissue of origin (embryonic versus adult SCs) as well as their potential to differentiate into one or many cell types. An accurate phenotypic classification of SC lineages, however, is a challenging task, particularly for closely related cell types. Commonly, the expression patterns of so-called marker genes have been used to define specific SC populations. Although this system offers a convenient approach for SC classification and purification, it has shortcomings as the obtained cell population may still show considerable heterogeneity (11,12). One of the primary goals of SC research has, therefore, been to determine more comprehensive profiles of SCs for better identification and characterisation of individual lineages. For this purpose, genome-wide expression profiling techniques have been applied extensively. Numerous transcriptomics datasets have been generated for various SC types and deposited in public data repositories. In princi-

*To whom correspondence should be addressed. Tel: + 351 289 800 077; Fax: + 351 289 800 066; Email: jppinto@ualg.pt
Correspondence may also be addressed to Matthias E. Futschik. Tel: +44 1752 586848; Email: matthias.futschik@plymouth.ac.uk

ple, this wealth of data should provide a compelling basis for the dissection of molecular processes underlying cellular identity and lineage. Moreover, the comparison of expression measurements from different experiments could help to identify more robust gene signatures and potential new marker genes. However, individual researchers seeking to explore available gene expression profiles across different SC lineages and experiments face a formidable task. Three major obstacles are prominent: (i) the massive amount of data; (ii) heterogeneity in original data processing and low quality datasets; and (iii) a limited number of ready-to-use and user-friendly analysis tools. Furthermore, it is not trivial for researchers to compare their own SC transcriptomics datasets with existing data in a comprehensive manner.

To help overcome these hurdles, we designed and implemented StemMapper, freely accessible at http://stemmapper.sysbiolab.eu. This resource is a manually curated database of gene expression data, holding information of a diverse range of SC and progenitor cell (PC) types for mouse and human. A main objective of StemMapper is to provide easy access to an integrated compendium of expression data collected from different transcriptomics experiments. StemMapper features (i) quality controlled data processed in a standardized manner; (ii) a clean and simple query interface; (iii) ability to integrate user-specific gene expression (.CEL) files into the analysis; (iv) intuitive visualization tools for gene expression landscape comparison; (v) interactive data analysis; and (vi) straightforward data retrieval.

Here, we briefly describe the data integrated in StemMapper, its pre-processing and curation procedure, how to explore gene expression data using the built-in analysis tools, and list the alternative methods by which information can be accessed and retrieved.

MATERIALS AND METHODS

Data collection and dataset curation

StemMapper currently holds 798 mouse and 166 human transcriptomes, representing a comprehensive coverage of expression profiles for 51 types of murine SCs, PCs and their progeny, as well as 19 types of human SCs, PCs and their progeny (Supplementary Table S1). Transcriptomic datasets for potential integration into StemMapper were collected from NCBI's Gene Expression Omnibus (GEO) (13). To identify relevant GEO data series, the following search criteria were used for the current version of StemMapper: (i) Organism: *Homo sapiens* OR *Mus musculus*; (ii) Platform: Affymetrix Mouse Genome 430 2.0 Array OR Affymetrix Human Genome U133 Plus 2.0 Array, and (iii) Description: 'stem' AND 'cell*'. The particular microarray platforms were chosen based on their broad coverage of SC types and the standardized procedures for sample handling and data pre-processing that helps to improve comparability of independent measurements. Future versions of StemMapper will include additional transcriptomics platforms, as discussed below.

Each retrieved GEO data series was subjected to a detailed curation process. First, each GEO record was inspected to manually classify the samples according to tissue of origin and stem cell lineage. Relevant information regarding treatments, conditions, and (cell surface) markers used for sorting of profiled SC/PC populations was extracted from the GEO records. Extraneous records were removed, and the remaining ones were further processed by a quality control (QC) assessment using the automatic R pipeline AffymetrixQC (14) customized to run locally. Briefly, this pipeline controls for sample quality, signal quality, signal comparability and biases, and array correlation. All samples passing the QC were subsequently processed in a standardized way to enhance comparability and consistency of the integrated data, while reducing potential batch effects related to the study of origin. For this purpose, we created a custom implementation of frozen RMA (fRMA) normalization procedure using the fRMA Tools package (15,16). The frozen parameter vectors were calculated on the basis of the data contained in StemMapper's database. The training data included a diverse set of batches, which were in our case GEO series (GSE) and were selected to cover a broad range of SC types and experimental conditions (e.g. control, treatment and time). A batch size of three microarrays was chosen for the Affymetrix mouse genome platform and a batch size of four microarrays for the Affymetrix human genome platform. In total, 82 batches for mouse (246 arrays) and 17 batches for human (68 arrays) were used to derive the fRMA parameters.

Finally, Principal Component Analysis (PCA) was applied to detect outlier samples, i.e. samples that either segregate outside the major clusters of the same cell lineage, or fall within clusters of other cell types. In this manner, we sought to remove potentially mis-annotated samples that could compromise downstream analyses and the overall quality of data integrated into StemMapper.

Implementation

StemMapper's database was implemented using MySQL. The web interface was developed using JavaScript and JavaServer Faces (JSF) 2.1, a Java-based framework for the development of user interfaces. The PrimeFaces library was used to extend the functionalities of JSF's standard core and tag libraries.

Gene expression analyses, namely the PCA, the generation of heatmaps and processing of user data (.CEL file), are carried out by using R packages from the R/Bioconductor platform. User submitted data are processed using the Bioconductor packages affy (17) and frma (15). Communication between R and the Java components is provided through Rserver. The MySQL database is accessed via the Hibernate library. Plots and heatmaps are drawn using the plotly.js JavaScript charting library.

USER INTERFACE

StemMapper features a user-friendly query interface and a set of interactive visualization and analysis tools that can provide a rapid overview of the data and can serve as a basis for follow-up investigations. They enable the user to: (i) visualize overall gene expression differences between SC types; (ii) compare the expression levels of particular genes of interest across different SC lineages and conditions; (iii)

query and download curated gene expression data, selected based on gene identity, cell type, or both; and (iv) upload gene expression files for direct analysis and comparison with the data integrated in the database.

The query interface is accessible via the *Analysis* tab. The query can be specified and executed in four steps: (1) select datasets; (2) select marker genes; (3) add genes of interest; and (4) analysis (query panel from Figure 1). Question mark buttons are available for each step, containing brief instructions on how to proceed with the selections. In the first step, the user selects the appropriate microarray platform (mouse or human data) and the datasets pertaining to the cell lineage(s) of interest. The second and third steps consist of selecting the genes of interest, for which the expression values will be retrieved from the database to be analysed and visually inspected. To guide the user, we have pre-compiled lists of established marker genes (derived from https://discovery.lifemapsc.com/in-vivo-development/organ-tissues) that can be (optionally) selected in step 2. Additional genes of interest can be directly input in the text box, or searched for (by gene symbol) in the search box enclosed in step 3. Genes can be also removed by deleting their identifiers in the textbox. This enables easy adjustment of the pre-compiled gene lists. In the fourth step, users are given the option to upload their own gene expression (.CEL) file to be processed and analyzed together with the information from the database. Finally, the query can be executed by pressing the 'Submit' button.

Once the queried data is retrieved and processed, two main tabs are generated: *Heatmaps* and *PCA Plot* containing interactive plots for these visualizations, allowing the user to access detailed gene-centered or cell type-centered information. On the *Heatmaps* page, the retrieved expression data for the selected genes and cell types are visualised and the samples are hierarchically clustered. Subsequently, the user can inspect the fRMA processed expression in the *Normalized Expression* sub-tab, or ranked expression values ranging from zero to one in the *Percentile Expression* sub-tab. The latter values are defined by the percentiles of a gene's expression with respect to its expression in all samples included in StemMapper. Thus, it provides a means for assessing how strongly a gene is expressed in a given sample compared to the expression observed in all other samples in StemMapper. The displays are interactive, allowing the user to easily inspect the expression values for genes and samples by moving the pointer over the heatmaps. Additionally, it is possible to zoom into the heatmaps. Clicking on a particular gene track creates a new window that lists summary information related to the gene's expression in the chosen cell types, namely, the median, minimum and maximum expression values, as well as the preferential expression measure (PEM, 18). In our case, the PEM of a gene for a cell type is defined as the difference of its average $\log_2$-transformed expression in the given cell type and its average $\log_2$-transformed expression in all cell types included in the analysis. Thus, PEM can be used to assess the specificity of expression of a gene in a particular cell type. Similarly, clicking on the leaves of the dendrogram opens a window with details about the corresponding sample, together with relevant GEO links. The new window also provides the possibility to download the expression data for all genes from the selected sample.

Since a common task in transcriptomics is the identification of co-regulated genes, an option for correlation analysis is implemented in StemMapper. The calculation of the Pearson correlation between the expression values of a selected gene and the expression values for all other genes in the selected samples can be executed via the corresponding action buttons. A new window subsequently displays a table with calculated correlation coefficients and respective p-values. The rows of the table can be ordered by clicking on the column header.

The *PCA plot* page provides the results of a PCA using the samples' expression profiles based on the queried genes in an interactive manner. Hovering over the points on the plot displays the label of the corresponding samples. Clicking on a data point produces a brief description of the sample together with its GEO links, and the possibility to download (in tab separated format) the expression data for all genes from that sample. Importantly, to facilitate the detection of relevant clusters of samples in the data, StemMapper allows for alternative colouring of the data points based on specific sample features. For instance, data points in the PCA plot can be coloured according to the SC type, tissue of origin, differentiation status, (surface) markers or treatment types. Both the expression heatmaps and the PCA plot can be exported as PNG files.

WORKFLOW EXAMPLE

To illustrate the utility and potential application of StemMapper, we describe the workflow of an exploratory data analysis using our database. Here, our aim is to find novel candidate marker genes from the analysis of a set of established genes associated with the differentiation of neural precursor cells (NPCs) into astrocytes (Figure 1).

Astrocyte differentiation from NPCs

Astrocytes are the most abundant cell type in the mammalian brain. Accordingly, to understand brain development and function, the elucidation of astrocytogenesis is of utmost importance (19). Astrocytes derive from NPCs that can self-renew, or differentiate into neurons, astrocytes or oligodendrocytes. These lineage commitments are typically coordinated by the activity of lineage-specific transcription factors, which in turn are known to be tightly regulated by signalling molecules, such as cytokines (20).

Glial fibrillary acidic protein (Gfap) is considered an astrocyte-specific gene and has been widely used in the study of astrocytogenesis (20). Ito and colleagues set out to discover genes associated with Gfap during astrocyte differentiation. Using genome-wide enhanced circular chromosomal conformation capture (e4C), they found 18 genes to be specifically associated with Gfap and expressed in NPC-derived astrocytes (20).

Finding putative novel astrocyte marker genes (Figure 1)

Steps 1–4. Using StemMapper to prioritize the study of these novel putative neural marker genes we investigated the

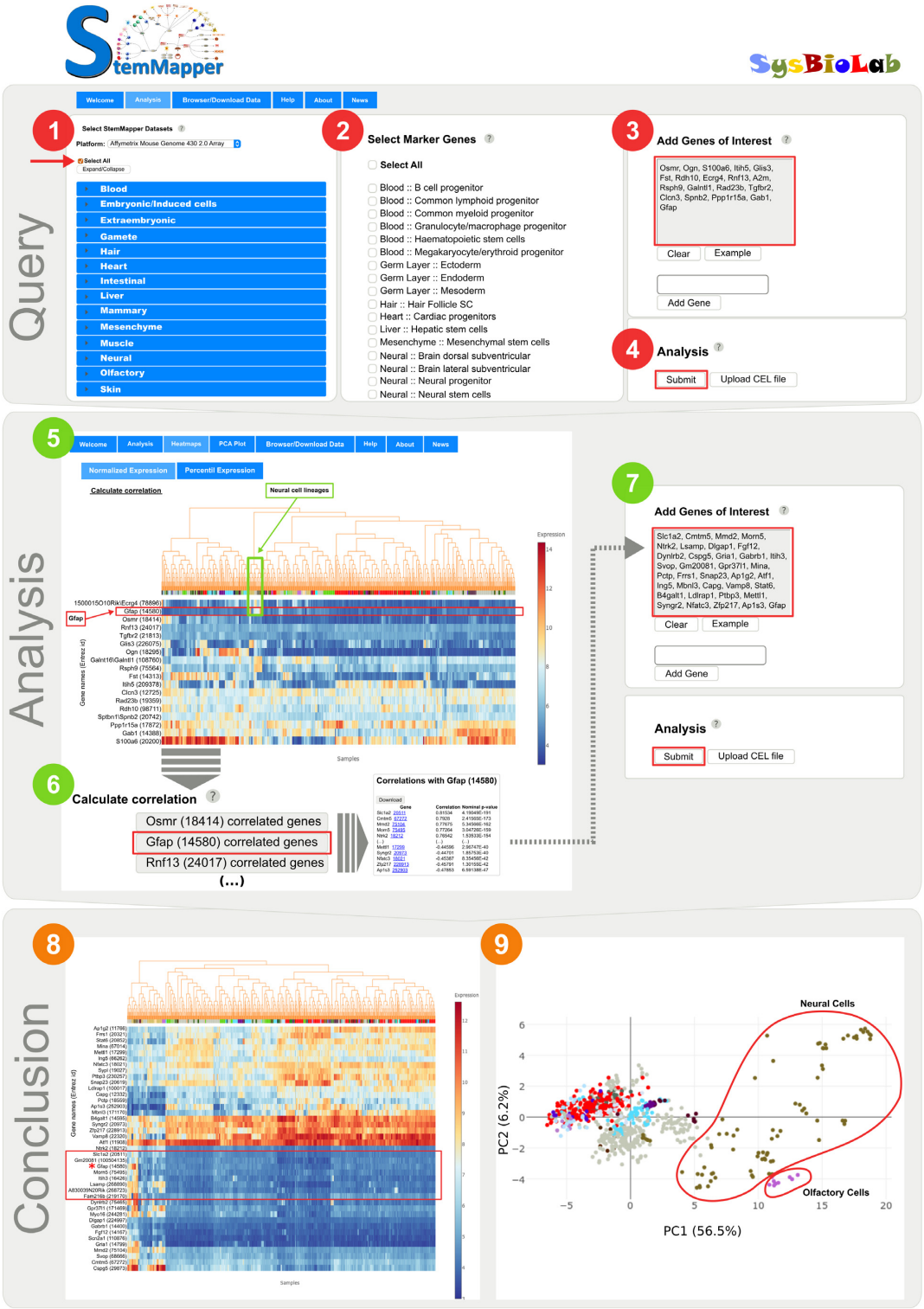

Figure 1. StemMapper workflow using a set of genes associated with the differentiation of neural precursor cells into astrocytes. Query panel (steps 1–4): After selecting all transcriptomic profiles for mouse (step 1), the 18 genes of interest were used as input (step 3) without the optional selection of pre-compiled marker genes (step 2), and the query was executed (step 4). Analysis panel (steps 5–7): Inspection of the produced heatmap (step 5) shows that Gfap is expressed specifically in samples of the neural lineages (highlighted with a green box). Calculation of the correlation with Gfap leads to the identification of 40 genes with strongly correlated or anti-correlated expression patterns (step 6). These genes were then used as new input together with Gfap (step 7). In the newly generated heatmap (step 8) seven genes (highlighted with a red box) displayed high expression in samples of the neural lineage and weak or no expression in most other samples similar to Gfap (indicated by a red star). The PCA plot based on the 40 input genes shows a clear separation of samples of the neural lineage including olfactory (bulb neural stem) cells (step 9). Researchers can then download the processed data to perform follow-up analyses.

expression pattern of Gfap in all cell types, together with the aforementioned 18 genes: Osmr, Ogn, S100a6, Itih5, Glis3, Fst, Rdh10, Ecrg4, Rnf13, A2m, Rsph9, Galntl1, Rad23b, Tgfbr2, Clcn3, Spnb2, Ppp1r15a and Gab1 in all available lineages (Figure 1).

Step 5. Visual inspection of the Normalized Expression Heatmap reveals that, of all the input genes, Gfap shows the most neuron-specific signature, i.e. Gfap is highly expressed specifically in neural lineages corroborated by the large PEM of 6.2 for murine brain samples of the dorsal subventricular zone (as displayed when clicking on Gfap's gene track).

Steps 6–7. For detection of other potential astrocyte markers, we applied the correlation analysis implemented in StemMapper to find genes correlated or anti-correlated with Gfap. The top and bottom 20 genes served then as new input into StemMapper.

Steps 8–9. Examination of the PCA plot shows that the 40 selected genes clearly separates neural cell lineage types (including olfactory bulb stem cells) from other lineages, indicating a strong discriminative potential.

Subsequent inspection of the corresponding heatmap revealed seven genes with expression patterns that were particularly similar to the one of Gfap. Of those, three genes have annotated functions: the well-known CNS-related gene Slc1a2 (Solute carrier family 1 (glial high affinity glutamate transporter), member 2), Itih3 (Inter-alpha trypsin inhibitor, heavy chain 3), and Lsamp (Limbic system-associated membrane protein). More interestingly, we identified four promising candidate marker genes without current functional annotation: Gm20081 (Predicted gene 20081), Morn5 (MORN repeat containing 5), A830039N20Rik (RIKEN cDNA A830039N20 gene), and Fam216b (Family with sequence similarity 216, member B). The latter genes represent attractive objects for further experimental studies in the context of astrocytogenesis given their specific expression in the neural lineage.

The user can subsequently choose to retrieve the relevant data for follow-up analyses, either directly by pressing the 'Download' button presented for each sample (providing the standardized processed data), or by following the GEO links to obtain the raw data from the source.

CONCLUSIONS AND FUTURE DIRECTIONS

SCs have been the focus of intense biomedical research leading to the generation of a vast amount of gene expression data that can be overwhelming for researchers. The salient need to facilitate exploration of the accumulated data has led to the development of StemMapper.

With its described features, StemMapper further complements and extends the repertoire of existing online resources for gene expression in SCs. These include tools such as ImmGen Gene Skyline (21), Codex (22), Gene Expression Commons (23), Expression Atlas (24) and Stemformatics (25), which provide - similarly to StemMapper - access to expression data for a wide range of cell types. Alternatively, online resources have been established to study the gene expression in particular cell lineages, such as, BloodSpot (26), HaemAtlas (27) and Differentiation Map (28) for hematopoietic cells, or for specific types of stem cells such as StemCellDB (29) for human embryonic stem cells.

StemMapper's strength lies in its comprehensive coverage of SC types, allowing the user to easily compare gene expression profiles across a wide range of SC lineages. StemMapper features a collection of quality-controlled and manually curated data, and can function as a 'one stop' resource where researchers can easily check previously reported expression values for their genes of interest, while comparing them across different SC lineages. Users can additionally upload their own gene expression (.CEL) files to visualize their measurements together with previously reported ones, a feature that, as far as we know, is unique to StemMapper.

Like with any other computational tool in biology, certain limitations exist for StemMapper. One of the most prominent is the undertaken selection of a particular transcriptomics platform, i.e. in our case Affymetrix GeneChips, as primary source for gene expression data. While this restriction is likely to enhance the comparability of measurements across different experiments, it led to the exclusion of data generated by other relevant technologies such as RNA-seq. For a future version, we therefore plan to integrate SC expression profiles from other transcriptomics platforms in StemMapper. Although cross-platform data integration is challenging, recent analysis indicate that such integration through appropriate data transformation is feasible (30). This extension of StemMapper will also enable the user to upload transcriptomics data generated by different microarray or sequencing technologies.

Furthermore, we envision the inclusion of single cell expression profiles, as these type of data provide a more precise view of the molecular heterogeneity of SCs which might be masked by the analysis of cell populations (31). In fact, it should be noted that many of the included transcriptomics data sets for SCs do not represent expression profiles of pure cell populations, but rather of cell populations enriched by a particular SC type. For instance, it has been estimated that many purification strategies for hematopoietic SCs only reach a purity rate of up to 50%, despite haematopoiesis being one of the most intensively studied models in SC biology (12). Another enhancement of StemMapper will be the inclusion of marker genes that are more discriminative on transcript level. At the moment, we provide pre-compiled lists of markers that are at least partially based on observed protein abundance and thus might not be optimal for classification of transcriptomic data. Finally, we plan to link StemMapper to other computational resources developed in our group for SC biology analysis, i.e. StemCellNet (32) and StemChecker (33) enabling the user to conduct a multitude of complementary *in silico* analyses within the domain of SC biology.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

J.P.P., I.D. and R.S.R.M. wish to thank the stem cell research community for making their data publicly available. We would also like to thank Jemma Dunn for careful read-

ing of the manuscript and the reviewers for their constructive comments and suggestions how to improve StemMapper.

FUNDING

Portuguese Fundação para a Ciência e Tecnologia (FCT) [SFRH/BPD/96890/2013 to J.P.P., PTDC/BEX-BID/5410/2014 to I.D. and R.P.A., FCT Investigator Grant IF/00881/2013 to M.E.F., UID/BIM/04773/2013 to CBMR, UID/Multi/04326/2013 to CCMAR]; Programa Doutoral ProRegeM—Mecanismos de Doença e Medicina Regenerativa [PD/00117/2012 to R.S.R.M.]. Funding for open access charge: Plymouth University Peninsula Schools of Medicine and Dentistry, Plymouth, Devon, PL4 8AA, UK.

Conflict of interest statement. None declared.

REFERENCES

1. Lund,R.J., Närvä,E. and Lahesmaa,R. (2012) Genetic and epigenetic stability of human pluripotent stem cells. *Nat. Rev. Genet.*, **13**, 732–744.
2. Beck,B. and Blanpain,C. (2013) Unravelling cancer stem cell potential. *Nat. Rev. Cancer*, **13**, 727–738.
3. Clements,W.K. and Traver,D. (2013) Signalling pathways that control vertebrate haematopoietic stem cell specification. *Nat. Rev. Immunol.*, **13**, 336–348.
4. Nelson,T.J., Martinez-Fernandez,A. and Terzic,A. (2010) Induced pluripotent stem cells: developmental biology to regenerative medicine. *Nature Publishing Group*, **7**, 700–710.
5. Tabar,V. and Studer,L. (2014) Pluripotent stem cells in regenerative medicine: challenges and recent progress. *Nat. Rev. Genet.*, **15**, 82–92.
6. Kimbrel,E.A. and Lanza,R. (2016) Pluripotent stem cells: the last 10 years. *Regen. Med.*, **11**, 831–847.
7. Stuckey,D.W. and Shah,K. (2014) Stem cell-based therapies for cancer treatment: separating hope from hype. *Nat. Rev. Cancer*, **14**, 683–691.
8. Sterneckert,J.L., Reinhardt,P. and Schöler,H.R. (2014) Investigating human disease using stem cell models. *Nat. Rev. Genet.*, **15**, 625–639.
9. Chen,I.Y., Matsa,E. and Wu,J.C. (2016) Induced pluripotent stem cells: at the heart of cardiovascular precision medicine. *Nature Publishing Group*, **13**, 333–349.
10. Zhu,H., Lensch,M.W., Cahan,P. and Daley,G.Q. (2011) Investigating monogenic and complex diseases with pluripotent stem cells. *Nat. Rev. Genet.*, **12**, 266–275.
11. Seita,J. and Weissman,I.L. (2010) Hematopoietic stem cell: self-renewal versus differentiation. *WIREs Syst. Biol. Med.*, **2**, 640–653.
12. Wilson,N.K., Kent,D.G., Buettner,F., Shehata,M., Macaulay,I.C., Calero-Nieto,F.J., Sánchez Castillo,M., Oedekoven,C.A., Diamanti,E., Schulte,R. *et al.* (2015) Combined single-cell functional and gene expression analysis resolves heterogeneity within stem cell populations. *Cell Stem Cell*, **16**, 712–724.
13. Barrett,T., Wilhite,S.E., Ledoux,P., Evangelista,C., Kim,I.F., Tomashevsky,M., Marshall,K.A., Phillippy,K.H., Sherman,P.M., Holko,M. *et al.* (2013) NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res.*, **41**, D991–D995.
14. Eijssen,L.M.T., Jaillard,M., Adriaens,M.E., Gaj,S., de Groot,P.J., Müller,M. and Evelo,C.T. (2013) User-friendly solutions for microarray quality control and pre-processing on ArrayAnalysis.org. *Nucleic Acids Res.*, **41**, W71–W76.
15. McCall,M.N., Bolstad,B.M. and Irizarry,R.A. (2010) Frozen robust multiarray analysis (fRMA). *Biostatistics*, **11**, 242–253.
16. McCall,M.N. and Irizarry,R.A. (2011) Thawing frozen robust multi-array analysis (fRMA). *BMC Bioinformatics*, **12**, 369.
17. Gautier,L., Cope,L., Bolstad,B.M. and Irizarry,R.A. (2004) affy—analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics*, **20**, 307–315.
18. Huminiecki,L., Lloyd,A.T. and Wolfe,K.H. (2003) Congruence of tissue expression profiles from Gene Expression Atlas, SAGEmap and TissueInfo databases. *BMC Genomics*, **4**, 31.
19. Namihira,M. and Nakashima,K. (2013) Mechanisms of astrocytogenesis in the mammalian brain. *Curr. Opin. Neurobiol.*, **23**, 921–927.
20. Ito,K., Sanosaka,T., Igarashi,K., Ideta-Otsuka,M., Aizawa,A., Uosaki,Y., Noguchi,A., Arakawa,H., Nakashima,K. and Takizawa,T. (2016) Identification of genes associated with the astrocyte-specific gene Gfap during astrocyte differentiation. *Sci. Rep.*, **6**, 23903.
21. Miller,J.C., Brown,B.D., Shay,T., Gautier,E.L., Jojic,V., Cohain,A., Pandey,G., Leboeuf,M., Elpek,K.G., Helft,J. *et al.* (2012) Deciphering the transcriptional network of the DC lineage. *Nat. Immunol.*, **13**, 888–899.
22. Sánchez-Castillo,M., Ruau,D., Wilkinson,A.C., Ng,F.S.L., Hannah,R., Diamanti,E., Lombard,P., Wilson,N.K. and Gottgens,B. (2015) CODEX: a next-generation sequencing experiment database for the haematopoietic and embryonic stem cell communities. *Nucleic Acids Res.*, **43**, D1117–D1123.
23. Seita,J., Sahoo,D., Rossi,D.J., Bhattacharya,D., Serwold,T., Inlay,M.A., Ehrlich,L.I.R., Fathman,J.W., Dill,D.L. and Weissman,I.L. (2012) Gene expression commons: an open platform for absolute gene expression profiling. *PLoS ONE*, **7**, e40321.
24. Petryszak,R., Keays,M., Tang,Y.A., Fonseca,N.A., Barrera,E., Burdett,T., Füllgrabe,A., Fuentes,A.M.-P., Jupp,S., Koskinen,S. *et al.* (2016) Expression Atlas update—an integrated database of gene and protein expression in humans, animals and plants. *Nucleic Acids Res.*, **44**, D746–D752.
25. 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.
26. Bagger,F.O., Sasivarevic,D., Sohi,S.H., Laursen,L.G., Pundhir,S., Sønderby,C.K., Winther,O., Rapin,N. and Porse,B.T. (2016) BloodSpot: a database of gene expression profiles and transcriptional programs for healthy and malignant haematopoiesis. *Nucleic Acids Res.*, **44**, D917–D924.
27. Watkins,N.A., Gusnanto,A., de Bono,B., De,S., Miranda-Saavedra,D., Hardie,D.L., Angenent,W.G.J., Attwood,A.P., Ellis,P.D., Erber,W. *et al.* (2009) A HaemAtlas: characterizing gene expression in differentiated human blood cells. *Blood*, **113**, e1–e9.
28. Novershtern,N., Subramanian,A., Lawton,L.N., Mak,R.H., Haining,W.N., McConkey,M.E., Habib,N., Yosef,N., Chang,C.Y., Shay,T. *et al.* (2011) Densely interconnected transcriptional circuits control cell states in human hematopoiesis. *Cell*, **144**, 296–309.
29. Mallon,B.S., Chenoweth,J.G., Johnson,K.R., Hamilton,R.S., Tesar,P.J., Yavatkar,A.S., Tyson,L.J., Park,K., Chen,K.G., Fann,Y.C. *et al.* (2013) StemCellDB: the human pluripotent stem cell database at the National Institutes of Health. *Stem Cell Res.*, **10**, 57–66.
30. Lê Cao,K.-A., Rohart,F., McHugh,L., Korn,O. and Wells,C.A. (2014) YuGene: a simple approach to scale gene expression data derived from different platforms for integrated analyses. *Genomics*, **103**, 239–251.
31. Nimmo,R.A., May,G.E. and Enver,T. (2015) Primed and ready: understanding lineage commitment through single cell analysis. *Trends Cell Biol.*, **25**, 459–467.
32. Pinto,J.P., Reddy Kalathur,R.K., Machado,R.S.R., Xavier,J.M., Bragança,J. and Futschik,M.E. (2014) StemCellNet: an interactive platform for network-oriented investigations in stem cell biology. *Nucleic Acids Res.*, **42**, W154–W160.
33. Pinto,J.P., Kalathur,R.K., Oliveira,D.V., Barata,T., Machado,R.S.R., Machado,S., Pacheco-Leyva,I., Duarte,I. and Futschik,M.E. (2015) StemChecker: a web-based tool to discover and explore stemness signatures in gene sets. *Nucleic Acids Res.*, **43**, W72–W77.