J Biosci (2020)45:53
DOI: 10.1007/s12038-020-00027-4

PSCRIdb: A database of regulatory interactions and networks of pluripotent stem cell lines

KRISHNENDU BANERJEE, TANMOY JANA, ZHUMUR GHOSH and SUDIPTO SAHA*

Division of Bioinformatics, Bose Institute, Kolkata, India

**Corresponding author (Email, ssaha4@gmail.com, ssaha4@jcbose.ac.in)*

MS received 14 January 2019; accepted 12 March 2020

Pluripotency in stem cells is regulated by a complex network between the transcription factors, signaling molecules, mRNAs, and epigenetic regulators like non-coding RNAs. Different pluripotent stem cell (PSC) lines were isolated and characterized to study the regulatory network topology to understand the mechanism that control developmental potential of pluripotent cells. PSCRIdb is a manually curated database of regulatory interactions including protein–protein, protein–DNA, gene–gene, and miRNA–mRNA interactions in mouse and human pluripotent stem cells including embryonic stem cells and embryonic carcinoma cells. At present, 22 different mouse and human pluripotent stem-cell-line-specific regulatory interactions are compiled in the database. Detailed information of the four types of interaction data are presented in tabular format and graphical network view in Cytoscape layout. The database is available at *http://bicresources.jcbose.ac.in/ssaha4/pscridb*. The database contains 3037 entries of experimentally validated molecular interactions that can be useful for systematic study of pluripotency integrating multi-omics data. In summary, the database can be a useful resource for identification of regulatory networks present in different pluripotent stem cell lines.

Keywords. Cytoscape; Graphical; multi-omics; network; regulatory interactions; systematic

Abbreviations: PSC, Pluripotent stem cell; ChIP-chip/seq, Chromatin immunoprecipitation followed by microarrays/sequencing; AP-MS, Affinity-purification followed by mass-spectrometry; IP-MS, Immunoprecipitation followed by mass-spectrometry; ChIP-PET, Chromatinb immunoprecipitation followed by paired-end tag sequencing; LOI, Loss of imprinting; PIRN, Pluripotent integrated regulatory network

1. Introduction

Pluripotent stem cells (PSCs), like embryonic stem cells and epiblast stem cells, possess two hallmark properties to self-renew and differentiate into any other cell types, and thus have great potential for regenerative therapy (Avior *et al.* 2016). Many studies were carried out with pluripotent stem cells using high-throughput and low-throughput techniques including chromatin immunoprecipitation followed by microarray or sequencing (ChIP-chip/seq), mass spectrometry (MS), RNA sequencing and luciferase reporter assay (Boyer *et al.* 2005; Chen *et al.* 2008; Chaerkady *et al.* 2010; Gu *et al.* 2016). Loh *et al.* studied E14 cell line to map the binding sites of Oct-4 and Nanog in the mouse ES cell genome using ChIP-PET method and reported high-confidence binding sites for Oct-4 and Nanog, respectively (Loh *et al.* 2006). Similar kind of work was carried out in mouse J1 ES cells and target promoters of core transcription factors Oct-4, Sox2, Klf4, and extended transcriptional factors like c-Myc, Nanog, Dax1, Rex1, Zpf281, and Nac1 using ChIP-chip methods were identified (Kim *et al.* 2008). An Oct-4-centered protein interaction network was studied in ZHBTc4 embryonic stem cell using affinity purification followed by mass spectrometry (AP-MS). This study leads to purification and characterization of Oct-4 interactome including transcription factors and chromatin-modifying complexes (van den Berg *et al.* 2010). MicroRNA-centric regulatory network in pluripotency

Electronic supplementary material: The online version of this article (https://doi.org/10.1007/s12038-020-00027-4) contains supplementary material, which is available to authorized users.

Published online: 25 March 2020

and cell-fate decision was reported in human pluripotent H9 cell line using luciferase reporter assay (Xu *et al.* 2009). Kolodziejczyk *et al.* quantified cell-to-cell heterogeneity using single-cell RNA sequencing data that improved the understanding of the dynamics of pluripotency regulatory networks and cell-fate decision (Kolodziejczyk *et al.* 2015). Recently, pluripotent-cell-line-specific research was used to study different genetically admixtured pluripotent cell lines and analyze genomic ancestries of a population (Tofoli *et al.* 2016). Besides, genetic and epigenetic abnormalities like loss of imprinting (LOI) using multiple pluripotent stem cell lines had been beneficial to understand developmental disorders, and there hold great therapeutic potential (Bar *et al.* 2017). Overall, pluripotent-cell-line-specific data were generated for different types of molecular regulatory interactions for understanding the developmental and disease biology.

There were a few dedicated pluripotency networks constructed and several databases developed for stem cells. *In silico* studies like PluriNetWork (Som *et al.* 2010) and PluriNet (Muller *et al.* 2008) provide information about the regulatory mechanism underlying the pluripotency of mouse and human, respectively, using gene expression data. Xu *et al.* constructed a regulatory network of self-renewal, differentiation, and lineage-specific commitment in mouse embryonic stem cells (Xu *et al.* 2014). Furthermore, specific databases were developed for compiling pluripotency-related molecular information. A database, named StemMapper (Pinto *et al.* 2018), provides information of gene expression data of diverse stem cells and progenitor cell (PC) types for mouse and human. FunGenES (Schulz *et al.* 2009) is also a transcriptomic database to study the gene expression changes of mouse embryonic stem cells during differentiation. Other databases like StemCellDB (Mallon *et al.* 2013) and StemBase (Sandie *et al.* 2009) have integrated similar high-content published data of mouse and human pluripotent stem cells. The databases like StemMapper, FunGenES, and StemBase have focused on transcriptomic regulatory relationship, whereas the database called Embryonic Stem Cell Atlas from Pluripotency Evidence (ESCAPE) has integrated multi-omic data of mouse and human embryonic stem cells (Xu *et al.* 2013). This database provides information of high-throughput datasets of both expression and interaction profiles of DNA, RNA, miRNA and protein using transcriptomics, chromatin immunoprecipitation with sequencing, inhibitory RNA screening, proteomics, and phosphoproteomics. It was observed that most of the dedicated databases for stem cells were compiled from high-throughput transcriptomic data, whereas very few databases contain interaction information.

Here, we present PSCRIdb, a database comprising of pluripotent stem cell regulatory interactions compiling protein–protein, protein–DNA, gene–gene, and miRNA–mRNA interactions in the mouse and human pluripotent cells. The database provides information of high-throughput followed by experimentally validated low-throughput studies of 22 different mouse and human embryonic cell lines. The data in the four classes of interaction are presented in graphical networks in Cytoscape layout and tabular format in graphical interface. At present, induced pluripotent stem cells (iPSCs) data have not be the focus in our database. iPSCs are derived from somatic cells reprogramming by over-expression of variable sets of a few transcription factors (Takahashi and Yamanaka 2006). iPSCs could possess significant differences from embryonic state at various molecular levels as a result of various methods used for reprogramming. Studies have uncovered differential gene expression signatures between ESCs and iPSC lines generated from different species and in different reprogramming experiments (Chin *et al.* 2009). In summary, PSCRIdb is a compilation of pluripotent stem cell-line-specific molecular interactions and network which may be useful in the study of integrated pluripotency regulatory network in cell line.

2. Methods

2.1 *Implementation and data access*

The PSCRIdb database is implemented using MySQL, HTML, CSS, JavaScript and PHP for building a user-friendly Web interface. Cytoscape version 3.6.1 was used for network visualization.

2.2 *Data sources*

2.2.1 *Literature survey and data collection:* The data was manually curated for regulatory interaction of pluripotent stem cells from PubMed. An extensive literature mining was done using Advance search option of PubMed employing appropriate keywords such as ‘Embryonic Stem Cell’, ‘Pluripotent Stem Cell’, ‘Protein-Protein Interaction’, ‘Proteomics’, ‘ChIP-Seq/chip’, and ‘miRNA targets’. Almost 3000 articles were mined for data curation. Thereafter, the protein–protein, protein–DNA, gene–gene, and miRNA–mRNA

interaction information was manually curated from both mouse and human pluripotent stem cell studies. At present, 22 different mouse and human embryonic and embryonal carcinoma cell lines specific data have been curated from searched literatures. Among 22 different cell lines, a significant amount of data is available from NT2, AB2.2, E14, J1, mESC, R1, ZHBTc4 cell lines and very little amount of data is available from H1, D3, ESC2, F9, H9, hESC, HUES-6, KH2, KhES-1, RUES2, RW4, V6.5, P19, and mEpiSC cell lines in our datasets.

2.2.2 *Database contents:* Experimentally validated mouse and human embryonic stem cell data were collected to construct the database. About 3037 entries were compiled in the database containing protein–protein, protein–DNA, gene–gene, and miRNA–mRNA interactions. Protein–protein interaction datasets were curated from proteomic based studies including affinity-purification followed by mass-spectrometry (AP-MS), immunoprecipitation followed by mass spectrometry (IP-MS) and co-immunoprecipitation followed by Western blot studies. Protein–DNA interaction information were extracted from Chromatin immunoprecipitation followed by microarrays/sequencing (ChIP-chip/Seq) followed by qRT-PCR studies. Gene–gene interaction data were curated from siRNA or shRNA mediated knockdown or knockout followed by qRT-PCR studies. miRNA–mRNA interaction datasets were extracted from luciferase reporter assay, transfection of miRNA mimics or inhibitor followed by qRT-PCR studies. The statistics of cell line specific four different interaction data is presented in table 1.

Table 1. Statistics of cell-line-specific molecular interactions in PSCRIdb

Cell line	Protein–protein interaction	Protein–DNA interaction	Gene–gene interaction	miRNA–mRNA interaction
AB2.2	97	–	–	–
D3	8	–	21	–
E14	5	198	178	–
E14Tg2a.4	122	–	–	6
ESC2	–	–	–	9
F9	1	–	–	–
H1	–	–	24	9
H9	-	31	41	27
hESC**	–	–	–	26
HUES-6	–	2	–	–
J1	118	177	–	–
KH2	37	–	–	–
KhES-1	–	–	13	–
mESC**	5	24	51	192
NT2	717	–	–	84
P19	1	–	11	–
R1	125	118	52	–
RUES2	–	1	9	–
RW4	2	–	–	–
V6.5	–	5	40	
mEpiSC	–	7	–	–
ZHBTc4	434	9	–	–
TOTAL	1672	572	440	353

**Generic terms of mouse and human embryonic stem cells. As the cell line name information is not available.

3. Results

3.1 *User interface*

3.1.1 *Search options:* The PSCRIdb provides Web-based search, browsing, and download options to users without any restriction with login and password. Users can search the database using proper keywords, e.g., Gene Name (Pou5f1, Nanog, Sox2), Protein Name (Oct-4, NANOG, SOX-2), miRNA (mmu-miR-294-3p, hsa-miR-320b), Organism (Mus musculus, Homo sapiens), and Cell line (E14, D3, H9). Users can only select one interaction option out of four, i.e., protein–protein, protein–DNA, gene–gene or miRNA–mRNA interaction as shown in figure 1a.

3.1.2 *Information of the output page:* The output page provides information about all four interaction datasets collected by manual curation. The protein–protein interaction datasets provide detail information of bait–prey proteins, organism, cell lines, method, and PubMed ID, as shown in figure 1b. In protein–DNA interaction, the output shows protein name, target gene, cell line, methods both high-throughputs followed by low-throughput validation studies, expression level (up or down), and PubMed IDs as shown in figure 1c. Similar information is also available in gene–gene interaction network as shown in figure 1d. The miRNA–mRNA interaction datasets provide information of miRNAs, target mRNAs, cell lines, ectopic cell lines, experimental conditions, methods, and expression level as shown in figure 1e. The total number of records found for each input query is provided on top of the output page. Users can download the result file of search input as comma separated value (.csv) format by clicking on ‘Download result (CSV format)’. The UniProt IDs of proteins, RefSeq version of the genes,

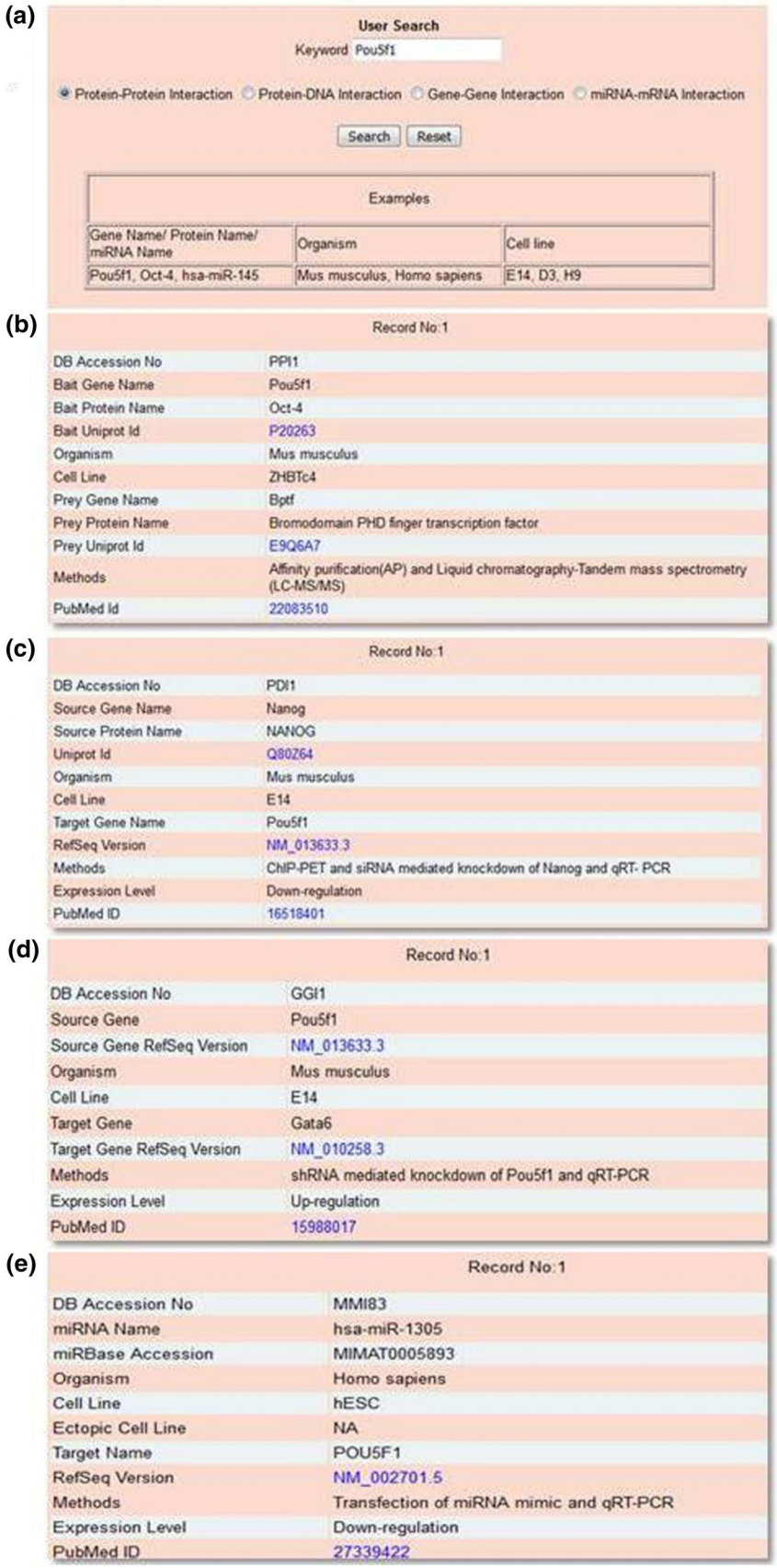

Figure 1. Snapshots of (**a**) search and output pages: (**b**) protein–protein, (**c**) protein–DNA, (**d**) gene–gene, and (**e**) miRNA–mRNA Interaction.

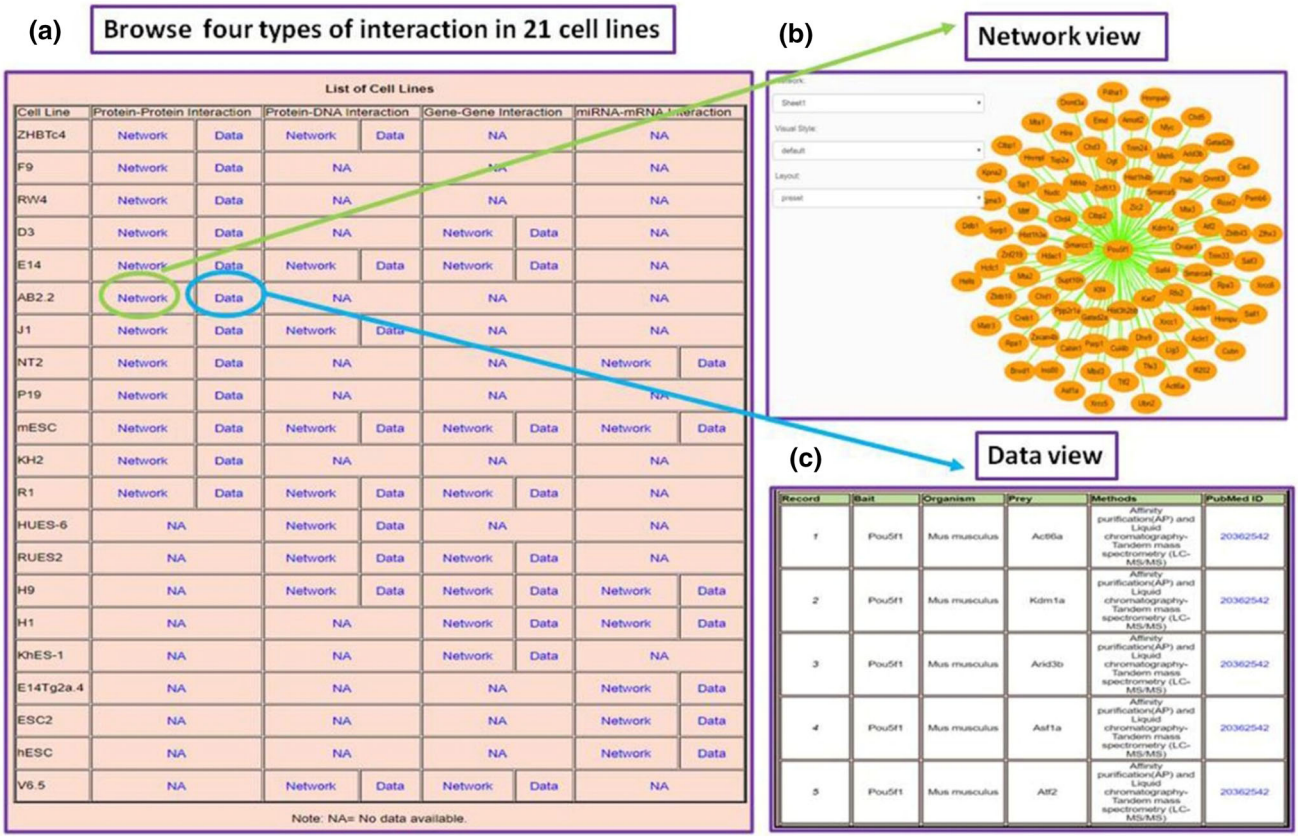

Figure 2. Browse page of PSCRIdb showing (**a**) information of regulatory interaction of different cell lines in both (**b**) network and (**c**) tabular format. The green arrow directs the page into Cytoscapic layout format and blue arrow directs the page into tabular format of data.

and miRBase accession of the miRNAs are hyperlinked to the corresponding UniProt, RefSeq, and miRBase pages. The PubMed IDs of the reference articles are also hyperlinked to the respective PubMed articles. Two networks compiling protein–DNA and both protein–DNA and miRNA–mRNA interaction data of Pou5f1 and Nanog have been constructed using Cytoscape. The networks were termed as gene regulatory (supplementary figure 1) and integrated regulatory network (supplementary figure 2) containing 86 and 169 interactions respectively. Thus, it can be concluded from these two networks, that integrating different layer of interactions enriches the biological understanding of the cellular process.

3.1.3 *Information of the browse page:* The browse page of PSCRIdb provides regulatory network information on 22 different mouse/human embryonic cell lines. Users can browse four different interactions in graphical networks in Cytoscape layout and tabular format as shown in figure 2 (a, b and c). Cytoscape version 3.6.1 was used to construct the networks. The visualization style of networks is in default mode but users can choose 17 different style modes. As well as the networks are in present layout and user can use 8 different options accordingly. Users can download four different interaction data in .csv format from 'Download' page.

3.2 *Comparision of PSCRIdb with other existing stem cell databases*

Databases like StemMapper (Takahashi and Yamanaka 2006), FunGeneES (Tofoli *et al.* 2016), StemCellDB (van den Berg *et al.* 2010), StemBase (Xu *et al.* 2014), and ESCAPE (Xu *et al.* 2013) integrate high content transcriptomic, SNP, proteomics and phosphoproteomics data. Among all of these databases, ESCAPE only contains regulatory interaction information from high-throughput studies. ESCAPE contains large set of predicted miRNA–target interaction information. FunGeneES (Tofoli *et al.* 2016), StemCellDB (van den Berg *et al.* 2010), StemBase (Xu *et al.* 2014), and ESCAPE (Xu *et al.* 2013) have compiled data from 3 mouse embryonic stem cell lines, 29 human pluripotent stem cell lines (including 21 human ESC and 8 hiPSCs), 21 embryonic and adult

Table 2. Comparision of PSCRIdb with other existing stem cell databases

	Data type					
Database	Omics	Interaction	Organism	Cell type	Cell lines	References
ESCAPE	Proteomics, phosphoproteomics,transcriptomics, histone modification	Protein–protein, protein–DNA and miRNA–mRNA predicted and experimental interaction	Mouse, Human	Embryonic stem cells	7 (2 hESC and 5 mESC)	Xu *et al.* (2013)
FunGenES	Transcriptomics	NA	Mouse	Embryonic stem cells	3 (mESC)	Schulz *et al.* (2009)
PSCRIdb	NA	Protein–protein, protein–DNA, gene-gene and miRNA–-mRNA experimentally validated interaction	Mouse, Human	Embryonic stem cells, Embryonic carcinoma cell	21 (12 mESC, 6 hESC, 2 mECC, 1 hECC, 1mEpiSC)	–
StemBase	Transcriptomics	NA	Mouse, Human and Rat	Embryonic, Adult stem cells and their derivatives 21	21 (8 mESC; 1 mECC; 1 hHSC; 1 mN-HSC; 1 Mouse Myoblasts; 1 Human Embryonic myoblasts; 1hESM; 1 hEKC; 2 mEF; 1m EFDC; 1 mTSC; 1 Mouse Stromal cells; 1 Human Epithelial Cells)	Sandie *et al.* (2009)
StemCellDB	Transcriptomics, single nucleotide polymorphism (SNP), array-based comparative genomic hybridization (aCGH), miRNA array and DNA methylation	NA	Human	Embryonic, induced pluripotent stem cells	29 (21 hESC and 8 hiPSC)	Mallon *et al.* (2013)
StemMapper	Transcriptomics	NA	Mouse, Human	Embryonic, Adult stem cells	NA	Pinto *et al.* (2018)

h/mESC: human/mouse embryonic stem cells; h/mECC: human/mouse embryonic carcinoma cells; hHSC: human hematopoietic stem cells; mN-HSC: mouse non-hematopoietic stem cells; hESM: human embryonic skeletal myoblasts; hEKC: human embryonic kidney cells; mEF: mouse embryonic fibroblasts; mEFDC: mouse embryonic fibroblast-derived cells; mTSC: mouse trophoblast cells; hiPSC: human-induced pluripotent stem cells, mEpiSC: mouse epiblast stem cells.

NA: Not available.

stem cells (including 8 mESCs, 1 mECCs, 12 adult stem cells and their derivatives), and 7 embryonic stem cell lines (including 2 hESC and 5 mESCs) respectively. Compared to above-mentioned databases, PSCRIdb integrates cell-line-specific, experimentally validated high-throughput followed low-throughput interaction datasets, e.g. protein–protein, protein–DNA, gene–gene, and miRNA–mRNA interactions. PSCRIdb contains data of 22 pluripotent stem cell lines (including 12 mESC, 6 hESC, 2 mECC, 1 hECC, and 1 mEpiSC). Further, ESCAPE integrates total 1197942 entries consisting both large amounts of predicted as well as high-throughput experimental datasets. However, PSCRIdb compiles 3037 entries containing only experimentally validated high-throughputdatasets followed by low- throughput datasets. PSCRIdb contains much more cell-line-specific information compared to ESCAPE. The comparison of PSCRIdb with other existing stem-cell-related databases is presented in table 2.

4. Discussion

PSCs have been the focus of intense biomedical research that has led to the generation of huge amount of expression and molecular interaction data. The need to compile the published data has led to the development of PSCRIdb. As majority of the existing stem cell databases have focused on expression data, PSCRIdb only focuses on molecular interaction datasets including protein–protein, protein–DNA, gene–gene, and miRNA–mRNA in human and mouse pluripotent stem cell lines. PSCRIdb's strength lies on its coverage of 22 different pluripotent-stem-cell line specific information that can be searched in user-friendly graphical interface. The database can be used to find active regulatory sub-networks present in pluripotent stem cells and system-level analysis of lineage commitment.

As the database has compiled datasets from embryonic stem and embryonic carcinoma cells, in the future we plan to add datasets from epiblast stem cells to provide more complete and holistic information of PSCs. PSCRIdb will be updated in every 6 months to incorporate more experimentally validated molecular interactions data. The database will be active for minimum 5 years.

4.1 *Availability of data and materials*

Database homepage: *http://bicresources.jcbose.ac.in/ssaha4/pscridb*. These data are freely available without restrictions for use by academics.

Acknowledgements

The authors acknowledge Division of Bioinformatics, Bose Institute, Kolkata, for providing the infrastructure. This work was supported by Indian Council of Medical Research (ICMR), India (ICMR-BIC/12(30)/2012).

References

Avior Y, Sagi I and Benvenisty N 2016 Pluripotent stem cells in disease modelling and drug discovery. *Nat. Rev. Mol. Cell Biol*. **17** 170–182

Bar S, Schachter M, Eldar-Geva T and Benvenisty N 2017 Large-scale analysis of loss of imprinting in human pluripotent stem cells. *Cell Rep*. **19** 957–968

Boyer LA, Lee TI, Cole MF, Johnstone SE, Levine SS, Zucker JP, Guenther MG, Kumar RM, *et al.* 2005 Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell* **122** 947–956

Chaerkady R, Kerr CL, Kandasamy K, Marimuthu A, Gearhart JD and Pandey A 2010 Comparative proteomics of human embryonic stem cells and embryonal carcinoma cells. *Proteomics* **10** 1359–1373

Chen X, Xu H, Yuan P, Fang F, Huss M, Vega VB, Wong E, Orlov YL, *et al.* 2008 Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell* **133** 1106–1117

Chin MH, Mason MJ, Xie W, Volinia S, Singer M, Peterson C, Ambartsumyan G, Aimiuwu O, *et al.* 2009 Induced pluripotent stem cells and embryonic stem cells are distinguished by gene expression signatures. *Cell Stem Cell* **5** 111–123

Gu KL, Zhang Q, Yan Y, Li TT, Duan FF, Hao J, Wang XW, Shi M, *et al.* 2016 Pluripotency-associated mir-290/302 family of micrornas promote the dismantling of naive pluripotency. *Cell Res*. **26** 350–366

Kim J, Chu J, Shen X, Wang J and Orkin SH 2008 An extended transcriptional network for pluripotency of embryonic stem cells. *Cell* **132** 1049–1061

Kolodziejczyk AA, Kim JK, Tsang JC, Ilicic T, Henriksson J, Natarajan KN, Tuck AC, Gao X, *et al.* 2015 Single cell rna-sequencing of pluripotent states unlocks modular transcriptional variation. *Cell Stem Cell* **17** 471–485

Loh YH, Wu Q, Chew JL, Vega VB, Zhang W, Chen X, Bourque G, George J, *et al.* 2006 The oct4 and nanog transcription network regulates pluripotency in mouse embryonic stem cells. *Nat. Genet*. **38** 431–440

Mallon BS, Chenoweth JG, Johnson KR, Hamilton RS, Tesar PJ, Yavatkar AS, Tyson LJ, Park K, *et al.* 2013 Stemcelldb: The human pluripotent stem cell database at the national institutes of health. *Stem Cell Res*. **1** 57–66

Muller FJ, Laurent LC, Kostka D, Ulitsky I, Williams R, Lu C, Park IH, *et al.* 2008 Regulatory networks define

phenotypic classes of human stem cell lines. *Nature* **455** 401–405

Pinto JP, Machado RSR, Magno R, Oliveira DV, Machado S, Andrade RP, Braganca J, Duarte I and Futschik ME 2018 Stemmapper: A curated gene expression database for stem cell lineage analysis. *Nucleic Acids Res*. **46** D788–D793

Sandie R, Palidwor GA, Huska MR, Porter CJ, Krzyzanowski PM, Muro EM, Perez-Iratxeta C and Andrade-Navarro MA 2009 Recent developments in stembase: A tool to study gene expression in human and murine stem cells. *BMC Research Notes* **2** 39

Schulz H, Kolde R, Adler P, Aksoy I, Anastassiadis K, Bader M, Billon N, Boeuf H, *et al.* 2009 The fungenes database: A genomics resource for mouse embryonic stem cell differentiation. *PLoS One* **4** e6804

Som A, Harder C, Greber B, Siatkowski M, Paudel Y, Warsow G, Cap C, Scholer H and Fuellen G 2010 The plurinetwork: An electronic representation of the network underlying pluripotency in mouse, and its applications. *PLoS One* **5** e15165

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

Tofoli FA, Dasso M, Morato-Marques M, Nunes K, Pereira LA, da Silva GS, Fonseca SA, Costas RM, *et al.* 2016 Increasing the genetic admixture of available lines of human pluripotent stem cells. *Sci. Rep*. **6** 34699

van den Berg DL, Snoek T, Mullin NP, Yates A, Bezstarosti K, Demmers J, Chambers I and Poot RA 2010 An oct4-centered protein interaction network in embryonic stem cells. *Cell Stem Cell* **6** 369–381

Xu H, Ang YS, Sevilla A, Lemischka IR and Ma'ayan A 2014 Construction and validation of a regulatory network for pluripotency and self-renewal of mouse embryonic stem cells. *PLoS Comput. Biol*. **10** e1003777

Xu H, Baroukh C, Dannenfelser R, Chen EY, Tan CM, Kou Y, Kim YE, Lemischka IR and Ma'ayan A 2013 Escape: Database for integrating high-content published data collected from human and mouse embryonic stem cells. *Database* https://doi.org/10.1093/database/bat045

Xu N, Papagiannakopoulos T, Pan G, Thomson JA and Kosik KS 2009 Microrna-145 regulates oct4, sox2, and klf4 and represses pluripotency in human embryonic stem cells. *Cell* **137** 647–658

Corresponding editor: BJ RAO