# Regulatory networks define phenotypic classes of human stem cell lines

**Franz-Josef Müller**[1,2], **Louise C. Laurent**[1,3], **Dennis Kostka**[4,‡], **Igor Ulitsky**[5], **Roy Williams**[6], **Christina Lu**[1], **In-Hyun Park**[7], **Mahendra S. Rao**[8,9], **Ron Shamir**[5], **Philip H. Schwartz**[10,11], **Nils O. Schmidt**[12], and **Jeanne F. Loring**[1,6]

1*Center for Regenerative Medicine, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, CA 92037, USA*

2*Center for Psychiatry, ZIP-Kiel, University Hospital Schleswig Holstein, Niemannsweg 147, D-24105 Kiel, Germany*

3*University of California, San Diego, Department of Reproductive Medicine, 200 West Arbor Drive San Diego, CA 92035, USA*

4*Department of Computational Molecular Biology, Max Planck Institute for Molecular Genetics, Ihnestrasse 63-73, D-14195 Berlin, Germany*

5*School of Computer Science, Tel Aviv University, Tel Aviv 69978, Israel*

6*The Burnham Institute for Medical Research, 10901 N Torrey Pines Road, La Jolla CA 92037, USA*

7*Division of Pediatric Hematology/Oncology, Children's Hospital Boston and Dana Farber Cancer Institute, Boston, Massachusetts 02115, USA*

8*Invitrogen Co, 3705 Executive Way, Frederick, MD 21704, USA*

9*Center for Stem Cell Biology, Buck Institute on Aging, 8001 Redwood Blvd, Novato, CA, 94945, USA*

10*Center for Neuroscience Research, Children's Hospital of Orange County Research Institute, 455 South Main Street, Orange, CA 92868, USA*

11*Developmental Biology Center, University of California, Irvine, 4205 McGaugh Hall, Irvine, CA 92697. USA*

12*Department for Neurosurgery University Medical Center Hamburg-Eppendorf, Martinistrasse 52, D-20246 Hamburg, Germany*

## Abstract

Stem cells are defined as self-renewing cell populations that can differentiate into multiple distinct cell types. However, hundreds of different human cell lines from embryonic, fetal, and adult sources have been called stem cells, even though they range from pluripotent cells, typified by embryonic stem cells, which are capable of virtually unlimited proliferation and differentiation, to adult stem cell lines, which can generate a far more limited repertoire of differentiated cell types. The rapid increase in reports of new sources of stem cells and their anticipated value to regenerative medicine[1, 2] have highlighted the need for a general, reproducible method for classification of these

cells[3]. We report here the creation and analysis of a database of global gene expression profiles ("Stem Cell Matrix") that enables the classification of cultured human stem cells in the context of a wide variety of pluripotent, multipotent, and differentiated cell types. Using an unsupervised clustering method[4, 5] to categorize a collection of ~150 cell samples, we discovered that pluripotent stem cell lines group together, while other cell types, including brain-derived neural stem cell lines, are very diverse. Using further bioinformatic analysis[6] we uncovered a protein-protein network ("PluriNet") that is shared by the pluripotent cells (embryonic stem cells, embryonal carcinomas, and induced pluripotent cells). Analysis of published data showed that the PluriNet appears to be a common characteristic of pluripotent cells, including mouse ES and iPS cells and human oocytes. Our results offer a new strategy for classifying stem cells and support the idea that pluripotence and self-renewal are under tight control by specific molecular networks.

---

Cultured cell populations are traditionally classified as having the qualities of stem cells by their expression of immunocytochemical or PCR markers.[7] This approach can often be misleading if these markers are used to categorize novel stem cell preparations or predict inherent multi- or pluripotent features.[8] To develop a more robust classification system, we created a framework for identifying putative novel stem cell preparations by their whole genome mRNA expression phenotypes (Figure 1). The core reference dataset, which we call the Stem Cell Matrix, includes cultures of human cells that have been reported to have either stem cell or progenitor qualities, including human embryonic stem cells, mesenchymal stem cells, and neural stem cells. To provide the context in which to place the stem cells, we included non-stem cell samples such as fibroblasts and differentiated embryonic stem cell derivatives. To avoid biasing the classification methods, it was critical that we designate the input cell types with terminology that carried as little preconception about their identity as possible. Our nomenclature ("Source Code") has two components: the first is the tissue or cultured cell line of origin. The second term captures a description of the culture itself. Supplementary Tables 1 – 8 summarize the descriptions of the core samples and their assigned Source Codes.

To sort the cell types we used an unsupervised machine learning approach to cluster transcriptional profiles of the cell preparations into stable distinct groups. Sparse nonnegative matrix factorization (sNMF) was adjusted for this task by implementing a bootstrapping algorithm to find the most stable groupings (see also Supplementary Discussion 1).[4, 5] The stability of the clustering[9] indicated that the dataset most likely contained about twelve different types of samples (Figure 2; Supplementary Method 2). The composition of the stable clusters revealed both predictable and unpredicted groupings of *a priori* designations (Figure 2 and Supplementary Figure 1). The twenty samples identified as undifferentiated human pluripotent stem cell (PSC) preparations were grouped together in one dominant cluster (Figure 2, Cluster 1) and one secondary cluster (Figure 2, Cluster 5). Sixty-two of the samples were brain-derived cells that were described as neural stem or progenitor cells based on their source, culture methods and classical markers. Most of the designated neural stem cells were distributed among multiple clusters, indicating a great deal of diversity in neural stem cell preparations. But one group of the brain-derived lines, those derived from surgical specimens from living patients (HANSE cells, see below), remained together throughout the iterative clusterings (Figure 2, Cluster 6; Supplementary Figure 3; Supplementary Method 1). The HANSE cell group consisted of transcriptional profiles that were derived from neurosurgical specimens following published protocols for multipotent neural progenitor derivation and propagation.[10, 11] These cells expressed markers that are commonly used to identify neural stem cells[12] (see Supplementary Figure 4), but the clustering clearly separated them from the other samples that had been derived from postmortem brains of prematurely born infants (see Figure 2).[10,11]

We tested the ability of our dataset to categorize additional preparations by adding 66 samples comprising new cultures derived from PSC lines that were already in the matrix, preparations that were not yet included (but their presumptive cell type was already represented), or new cell types. We chose two new types of cells: a differentiated cell type (umbilical vein endothelial cells [HUVEC]) and a recently developed new source of pluripotent cells, induced pluripotent stem cells[13-16] (iPSC, Supplementary Table 9). iPSCs have been generated from somatic cells, including adult fibroblasts, by genetic manipulation of certain transcription factors.[13, 15-17] We re-computed clustering results including the test dataset (Supplementary Table 10). All of the HUVEC samples clustered together and formed a distinct group. Most of the additional PSC lines (human ES cells [embryonic PSC; ePSC] and iPSCs) from several different labs were placed into a context that contained solely PSC lines. The three additional germ cell tumor lines clustered together with the tumor-derived pluripotent stem cell (tPSC) line 2102Ep and samples of three human ES cell lines: BG01v[18], Hues7[19], and Hues13[19]. BG01v is an established aneuploid variant line and the two Hues lines were aneuploid variants of the originally euploid lines (not shown).

We used a combination of analysis tools to explore the basis of the unsupervised classification of the samples in the core dataset. Gene Set Analysis[3] (GSA) is a means to identify the underlying themes in transcriptional data in terms of their biological relevance.

GSA uses lists of genes[5] that are related in some way; the common criterion is that the relationships among the genes in the lists are supported by empirical evidence.[20] GSA highlighted numerous significant differences among the computationally defined categories. (See Supplementary Figure 2, Supplementary Table 11 and Supplementary Online Materials).

While GSA is valuable for discovering specific differences among sample groups, it is limited to curated gene lists and cannot be used to discover new regulatory networks. The MATISSE algorithm[6] (http://acgt.cs.tau.ac.il/matisse) takes predefined protein-protein interactions (e.g. from yeast-two-hybrid screens) and seeks connected subnetworks that manifest high similarity in sample subsets. The modified version used in this analysis is capable of extracting sub-networks that are co-expressed in many samples but also significantly up- or down-regulated in a specific sample cluster. Since the PSC preparations were consistently clustered together we used MATISSE to look for distinctive molecular networks that might be associated with the unique PSC qualities of pluripotence and self-renewal. A Nanog-associated regulatory network has been outlined in mouse embryonic PSC,[21] and we looked for the elements of this network in human PSCs using our unbiased algorithm. We found that the algorithm predicts that human PSC possess a similar NANOG-linked network (Figure 3a; elements labelled in red). However, we also discovered that the human NANOG network appears to be integrated as a small component of a much larger protein-protein interaction network that is up-regulated in human PSCs (Figure 3). Remarkably, this PSC-specific network (termed <u>Pluri</u>potency associated <u>Net</u>work, PluriNet) contains key regulators that are involved in the control of cell cycle, DNA replication, DNA repair, DNA methylation, SUMOylation, RNA processing, histone modification and nucleosome positioning (see also Supplementary Discussion 2 and www.openstemcellwiki.org). Many of the genes in the PluriNet have been linked to embryogenesis, tumorigenesis, and aging (Figure 3c and Supplementary Figure 6). We further explored the hypothesis that pluripotency is closely linked to PluriNet expression by analyzing published gene expression datasets from human oocytes, various types of PSCs, and murine embryos (see Table 1 for a summary of our findings in various model systems). Analysis of a microarray dataset[22] that spans development from murine oocytes to the late blastocyst stage revealed that the PluriNet expression is dynamic and up-regulated during early mammalian embryogenesis (Table1; Supplementary Figures 7 - 9).[23] Also, our preliminary analyses indicate that the PluriNet is strongly up-regulated in mouse PSCs, mouse iPSCs, and mouse epiblast-derived stem cells[24] when compared to somatic cells. Therefore the PluriNet may be

useful as a biologically inspired gauge for classifying both murine and human PSC phenotypes (Table 1; Supplementary Figures 10 – 13).

In summary, our data indicate that an unbiased global molecular profiling approach combined with a transcriptional phenotype collection using suitable machine learning algorithms can be used to understand and codify the phenotypes of stem cells.[4, 5, 25] Although it is more extensive than any stem cell dataset reported to date, we consider our database and the PluriNet to be a work in progress. As more direct evidence for protein-protein interactions in human cells becomes available, it will be possible to refine the networks we've defined and make them more useful for testing hypotheses about the nature of stem cell pluri- and multipotence. Also, our sample collection is limited to pluri- and multipotent stem cell types that grow well in culture, and does not include some of the most well-studied lineages, such as hematopoietic stem cells. Resolution and reliability of a context-based unsupervised classification can be expected to grow with the breadth and depth of the database content.[26] Even with these limitations, we have shown that the dataset and PluriNet have already proved useful for categorizing cell types using unbiased criteria. As more stem cell populations become available, cultured by new methods, isolated from new sources, or induced by new methods, we will use the PluriNet and the Stem Cell Matrix as a reference system for phenotyping the cells and comparing them with existing cell lines.

## Methods Summary

For an overview of the general workflow, please also refer to Figure 1. A detailed list of the samples, culture methods and reference publications is provided in the Supplementary materials.[11]. Generally, RNA from each sample was prepared from approximately $1 \times 10^6$ cultured cells. Sample amplification, labeling and hybridization on Illumina WG8 and WG6 Sentrix BeadChips were performed for all arrays in this study according to the manufacturer's instructions (http://www.illumina.com) at a single Illumina BeadStation facility. We used the Consensus Clustering framework[9] to cluster transcription profiles and to assess stability of the results. As the algorithm, we used sparse non-negative matrix factorization.[5] For data perturbation, 30 sub-sampling runs were performed for each considered number of clusters (k). In each run, 80% of the data was subjected to ten random restarts. The R-script can be downloaded at the accompanying website www.stemcellmachinelearning.org. Details on the application of GSA,[20] PAM,[27] MATISSE[6] as well as publicly available datasets used in this study can be found in the Methods section. We modified the MATISSE[6] computational framework to fit the goals of this study. For the present analysis we used the human physical interaction network that we had previously assembled[6] and augmented it with additional interactions from recent publications.[21, 28 29] The 64 interactions in Wang et al. 2006[21] were mapped to the corresponding human orthologs using the NCBI Homologene database. The microarray data has been deposited at NCBI GEO (GEO series accession number: GSE11508). It can also be accessed, processed and downloaded at www.stemcellmesa.org.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgements

## References

1. Müller FJ, Snyder EY, Loring JF. Gene therapy: can neural stem cells deliver? Nat Rev Neurosci 2006;7:75–84. [PubMed: 16371952]
2. Murry CE, Keller G. Differentiation of embryonic stem cells to clinically relevant populations: lessons from embryonic development. Cell 2008;132:661–80. [PubMed: 18295582]
3. Adewumi O, et al. Characterization of human embryonic stem cell lines by the International Stem Cell Initiative. Nat Biotechnol 2007;25:803–16. [PubMed: 17572666]
4. Brunet JP, Tamayo P, Golub TR, Mesirov JP. Metagenes and molecular pattern discovery using matrix factorization. Proc Natl Acad Sci U S A 2004;101:4164–9. [PubMed: 15016911]
5. Gao Y, Church G. Improving molecular cancer class discovery through sparse non-negative matrix factorization. Bioinformatics 2005;21:3970–5. [PubMed: 16244221]
6. Ulitsky I, Shamir R. Identification of functional modules using network topology and high-throughput data. BMC Syst Biol 2007;1:8. [PubMed: 17408515]
7. Carpenter MK, Rosler E, Rao MS. Characterization and differentiation of human embryonic stem cells. Cloning Stem Cells 2003;5:79–88. [PubMed: 12713704]
8. Goldman B. Magic Marker Myths. Nature Reports Stem Cells 2008. 2008
9. Monti S, Tamayo P, Mesirov J, Golub T. Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data. Machine Learning 2003;52:91–118.
10. Palmer TD, et al. Cell culture. Progenitor cells from human brain after death. Nature 2001;411:42–3. [PubMed: 11333968]
11. Schwartz PH, et al. Isolation and characterization of neural progenitor cells from post-mortem human cortex. J Neurosci Res 2003;74:838–51. [PubMed: 14648588]
12. Kornblum HI, Geschwind DH. Molecular markers in CNS stem cell research: hitting a moving target. Nat Rev Neurosci 2001;2:843–6. [PubMed: 11715062]
13. Takahashi K, Yamanaka S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. Cell 2006;126:663–76. [PubMed: 16904174]
14. Takahashi K, et al. Induction of pluripotent stem cells from adult human fibroblasts by defined factors. Cell 2007;131:861–72. [PubMed: 18035408]
15. Yu J, et al. Induced Pluripotent Stem Cell Lines Derived from Human Somatic Cells. Science. 2007
16. Park IH, et al. Reprogramming of human somatic cells to pluripotency with defined factors. Nature 2008;451:141–6. [PubMed: 18157115]
17. Okita K, Ichisaka T, Yamanaka S. Generation of germline-competent induced pluripotent stem cells. Nature. 2007
18. Zeng X, et al. BG01V: a variant human embryonic stem cell line which exhibits rapid growth after passaging and reliable dopaminergic differentiation. Restor Neurol Neurosci 2004;22:421–8. [PubMed: 15798361]
19. Cowan CA, et al. Derivation of embryonic stem-cell lines from human blastocysts. N Engl J Med 2004;350:1353–6. [PubMed: 14999088]
20. Bradley Efron RT. On testing the significance of sets of genes. The Annals of Applied Statistics 2007;1:107–129.
21. Wang J, et al. A protein interaction network for pluripotency of embryonic stem cells. Nature 2006;444:364–8. [PubMed: 17093407]

22. Wang QT, et al. A genome-wide study of gene activity reveals developmental signaling pathways in the preimplantation mouse embryo. Dev Cell 2004;6:133–44. [PubMed: 14723853]

23. Chambers I, et al. Nanog safeguards pluripotency and mediates germline development. Nature 2007;450:1230–4. [PubMed: 18097409]

24. Tesar PJ, et al. New cell lines from mouse epiblast share defining features with human embryonic stem cells. Nature 2007;448:196–9. [PubMed: 17597760]

25. Golub TR, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 1999;286:531–7. [PubMed: 10521349]

26. Donoho D, Stodden V. When Does Non-Negative Matrix Factorization Give Correct Decomposition into Parts? Advances in Neural Information Processing Systems NIPS*2003 Online Papers. 2003

27. Lacayo NJ, et al. Gene expression profiles at diagnosis in de novo childhood AML patients identify FLT3 mutations with good clinical outcomes. Blood 2004;104:2646–54. [PubMed: 15251987]

28. Ewing RM, et al. Large-scale mapping of human protein-protein interactions by mass spectrometry. Mol Syst Biol 2007;3:89. [PubMed: 17353931]

29. Mishra GR, et al. Human protein reference database--2006 update. Nucleic Acids Res 2006;34:D411–4. [PubMed: 16381900]
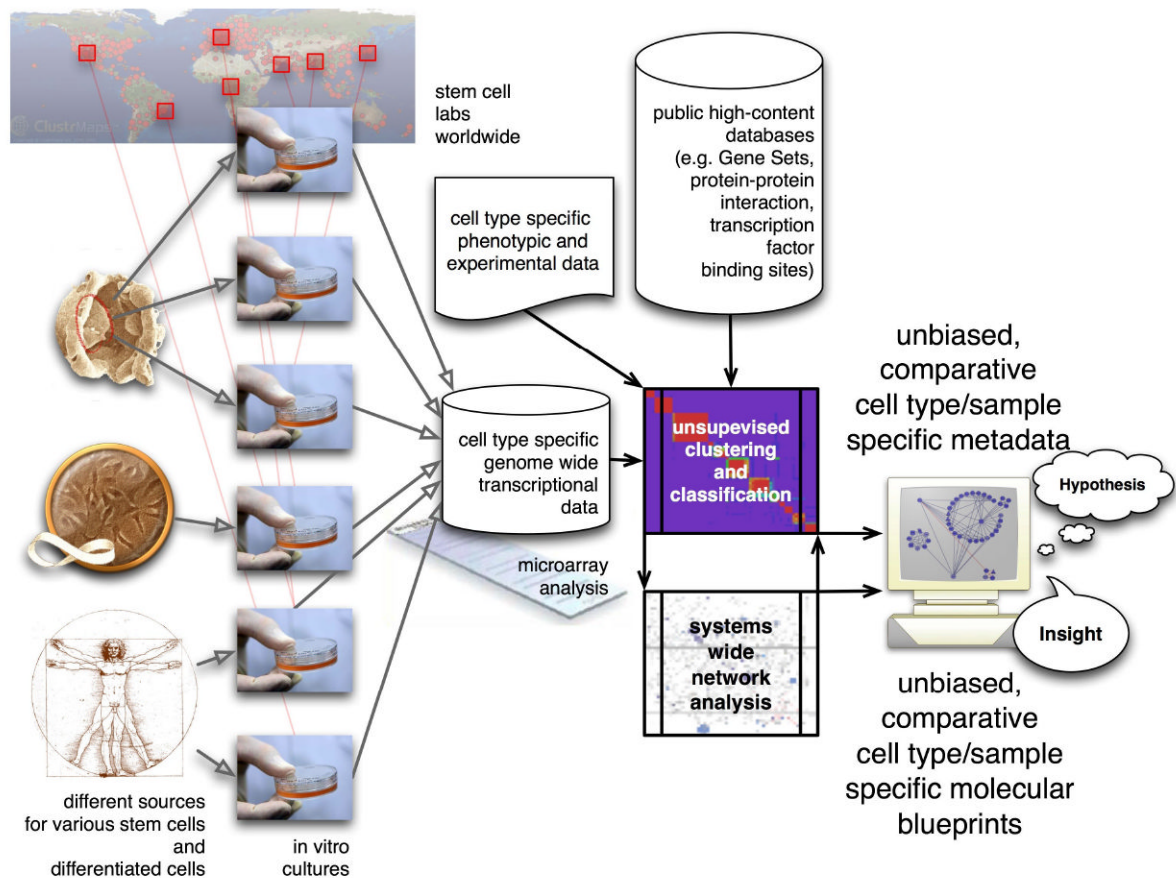
**Figure 1. Sample collection and analysis for the Stem Cell Matrix**
Cell preparations for the Stem Cell Matrix are cultured in the authors' laboratory or collected from other sources worldwide. Samples are assigned source codes that capture their biological origin and an relatively unbiased description of the cell type (such as BNLin for brain-derived neural lineage). Samples are collected and processed at a central lab for microarray analysis on a single Illumina BeadStation instrument.

The genomics data are processed by unsupervised algorithms that are capable of grouping the samples based on non-obvious expression patterns encoded in transcriptional phenotypes. For pathway discovery, existing high-content databases with experimental data (e.g. protein-protein-interaction data or gene sets) are combined with our transcriptional database, *a priori assumed* identity of cell types and bootstrapped sparse non-negative matrix factorization (sample clustering) to produce metadata that can be mined with Gene Set Analysis software and topology-based gene set discovery methods (systems wide network analysis). Web-based, computer-aided visualization methodologies can be used by researchers to formulate testable hypotheses and generate results and insights in stem cell biology.

Two exemplary results we report in this paper are the classification of novel stem cell types in the context of other better understood stem cell preparations, and a molecular map of interacting proteins which appear to function in concert in pluripotent stem cells.
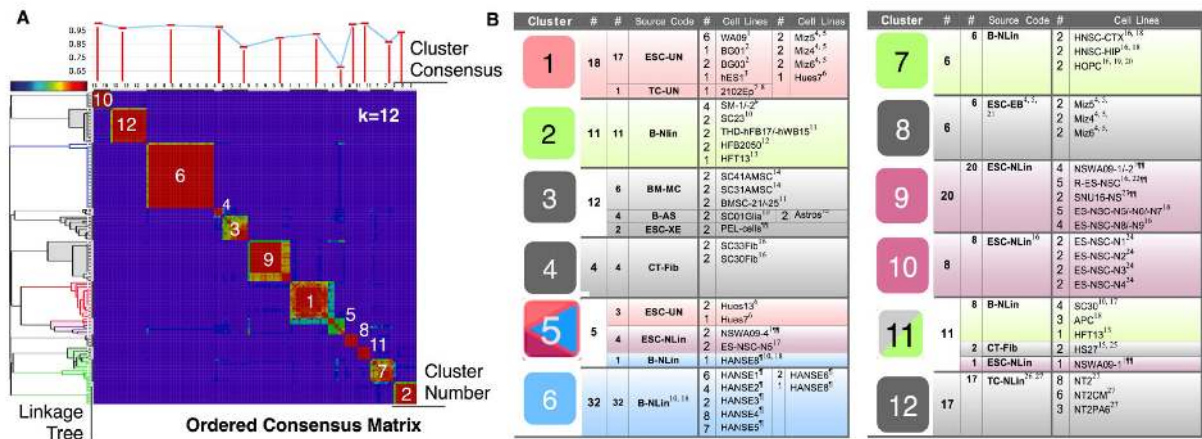
**Figure 2. Clusters of samples based on machine learning algorithm**
Samples were distributed on the basis of their transcriptional profiles into consensus clusters using sNMF.

**A.** Consensus matrix from consensus clustering results (center matrix plot). The consensus matrix is a visual representation of the clustering results and the separation of the sample clusters from each other. Blue indicates no consensus, and red very high consensus. The numbers (1-12) on the diagonal row of clusters indicate the number assigned to the cluster by sNMF. These numbers ("Cluster 1" …"Cluster 12") are used throughout the text to indicate the group of samples in that cluster. The bar graph above the consensus matrix plot shows the summary statistics assessing the overall quality of each cluster. The cluster consensus value (0-1) is plotted above the corresponding cluster in the matrix plot. Note that most clusters (Clusters 10, 12, 6, 4, 9, 1, 8, 11, 7, 2) have a high quality measurement. To the left of the consensus matrix is another view of the consensus data, visualized as a dendrogram. This is a representation of the hierarchical clustering tree of the consensus matrix

**B.** The content of the sample clusters resulting from the same sNMF run are displayed. Numbers are the same cluster numbers assigned by the consensus clustering algorithm that are used throughout the text and figures. For more information on samples and Source Code and references see Supplementary Tables 1 – 10.
# Number of samples,
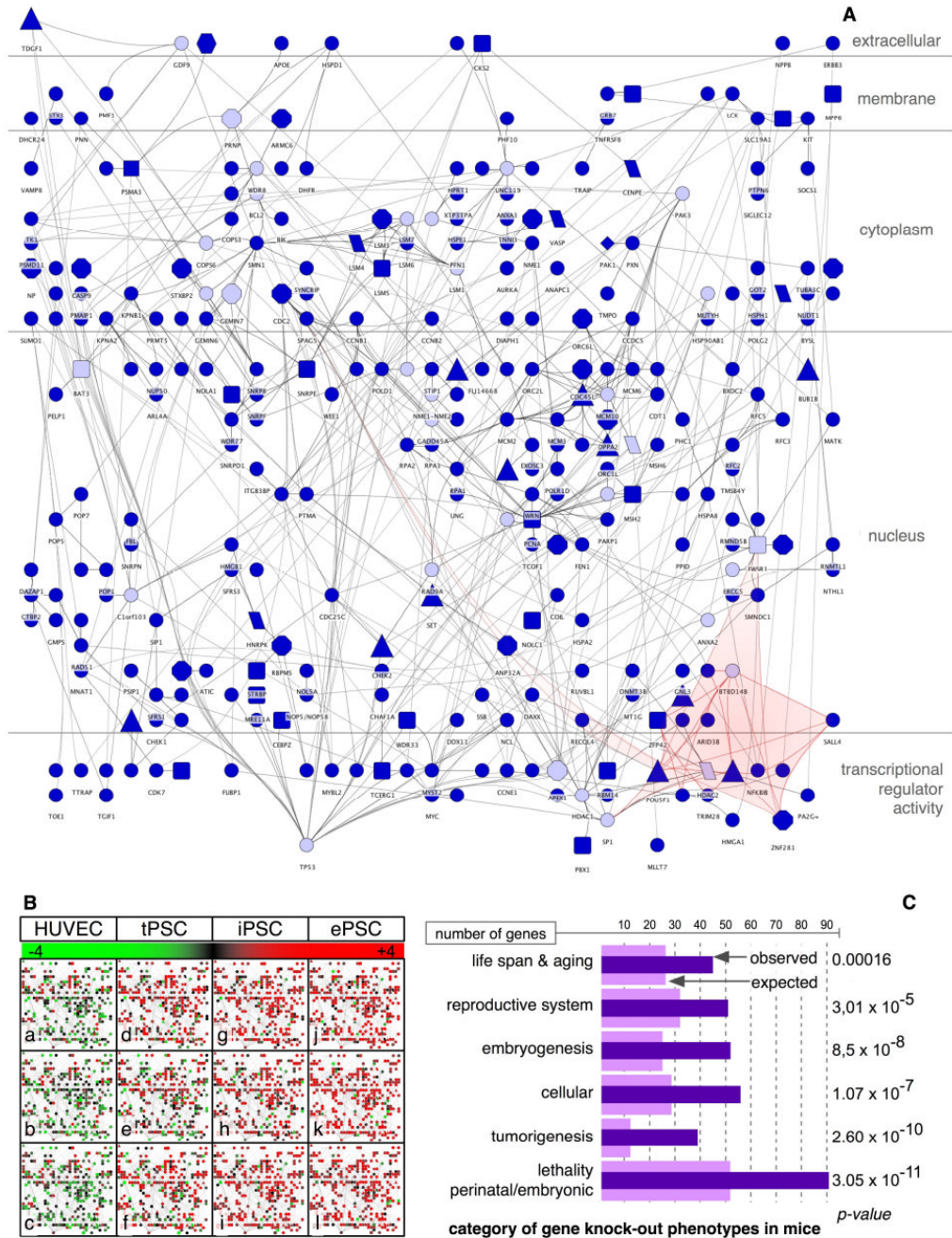¶ Samples were derived from adult brain specimens

**Figure 3. Pluripotent Stem Cell-specific protein-protein interaction network detected by MATISSE**
Clusters from the sNMF k=12 analysis were used in combination with the transcriptional database to identify protein-protein interaction networks enhanced in PSC.
**A.** A large differentially expressed connected subnetwork ("PluriNet") shows the dominance of cell cycle regulatory networks in PSC (see legend). All of the dark blue symbols are genes that are highly expressed in most PSCs compared to the other cell samples in the dataset. Front nodes as represented by Stem Cell Matrix expression data and back nodes as inferred by MATISSE are displayed with different colour shades.[6] Highlighted in red are the interactions of a group of proteins associated with pluripotency in murine ePSC[21]. Interestingly, this subnetwork shows a significant enrichment in genes that are targeted in the genome by the

transcription factors NANOG (p=$5.88 * 10^{-4}$), SOX2 (p=0.058) and E2F (p=$1.29 * 10^{-16}$, all p-values are Bonferroni corrected). For an interactive visualization of PluriNet, see www.stemcellmatrix.org.

**B.** Heat map-like visualization of PluriNet genes for samples from the test dataset: HUVEC (UC-EC, **a-b**, derived from three independent individuals), germ cell tumor derived pluripotent stem cells (tPSC-UN, **d-f**, lines GCT-C4, GCT-72, GCT-27X, derived from three independent individuals), induced pluripotent stem cells (iPSC-UN, g-i, BJ1-iPS12, MSC-iPS1, hFib2-iPS5 three independently derived lines from different somatic sources) and embryonic stem cells (ePSC-UN, j-l, lines Hues22, HSF6, ES2, derived from three independent blastocysts in three independent labs). Most PluriNet genes are markedly up-regulated in iPSC-UN and ePSC-UN. tPSC-UN do show a less consistent expression pattern. UC-EC show lower expression levels of most PluriNet genes. Please refer to Supplementary Figure 5 for a larger version of the same Net-Heatmaps

**C.** Analysis of genes from PluriNet in the context of phenotypes, which have been reported to result from specific genetic manipulations (e.g. gene knock-out) in mice in the MGI 3.6 phenotype ontology database (http://www.informatics.jax.org/). We find significant overrepresentation of phenotypes "lethality (perinatal/embryonic)", "tumorigenesis", "cellular", "embryogenesis", "reproductive system" and "life span and aging" among the genes in PluriNet. Although these broad categories might be rather unspecific surrogate markers for PSC function in mammals, this analysis might point towards PluriNet's role in vivo. For more details, see also Supplementary Figure 6 and Supplementary Table 12.

**Table 1**

PluriNet Expression patterns in various model system for pluripotecy

| A: Expression of PluriNet genes | |
|---|---|
| in murine model systems | |
| *MII Oocytes* | up-regulated[a] |
| *Zygote* | up-regulated[a] |
| *Embryo (2 cell–Blastocyst)* | up-regulated[a] |
| *ePSC* | up-regulated[b] |
| *EpiSC* | up-regulated[b] |
| *iPSC* | up-regulated[b] |
| *Fibroblasts (normal)* | down-regulated[b] |
| *Fibroblasts (transformed)* | down-regulated[b] |

| B: Successful PluriNet based, post-hoc classification for … | | |
|---|---|---|
| in murine model systems | pluripotency | germ-line transmission |
| *ePSC* | yes[c] | yes[c] |
| *EpiSC* | yes[c] | yes[c] |
| *iPSC* | yes[c] | yes[c] |
| *Fibroblasts (normal)* | yes[c] | yes[c] |
| *Fibroblasts (transformed)* | yes[c] | yes[c] |

| C: Expression of PluriNet genes | |
|---|---|
| in human model systems | |
| *MII Oocytes* | up-regulated[d] |
| *tPSC* | up-regulated[e] |
| *ePSC* | up-regulated[e/f] |
| *iPSC* | up-regulated[e/f] |
| *ePSC derived cell types* | down-regulated[f] |
| *somatic cell types* | down-regulated[e/f] |
| *somatic cancer line (HeLa)* | down-regulated[g] |

| D: Successful PluriNet based, post-hoc classification for … | |
|---|---|
| in human model systems | pluripotency |
| *tPSC* | yes[h] |
| *ePSC* | yes[h] |
| *iPSC* | yes[h] |
| *ePSC derived cell types* | yes[h] |
| *somatic cell types* | yes[h] |

upregulated

downregulated

*a*for more details see Supplementary Figure 8

*b*for more details see Supplementary Figures 9 and 10

*c*for more details see Supplementary Figure 10

*d*for more details see Supplementary Figure 7

*e*for more details see Figure 3B, Supplementary Figures 5 and 12

*f*for more details see Supplementary Figure 11

*g*for more details see Supplementary Discussion 2 PluriNet and Cell Cycle

*h*for more details see Supplementary Figure 12

PAM – Prediction Analysis of Microarray classifier with leave-one-out cross validation[27];

"yes" in Table 3B and 3D stands for: correct classification of pluripotent state (pluripotent or not pluripotent) in > 90% of samples.

This table summarizes the expression patterns of PluriNet in various model systems of pluripotence and differentiation. More details on the specific tests and explanations of the data sources for the results can be found at the respective Supplementary Figures and Materials as listed in the above.