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] & Jeanne F. Loring[1,6]

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] has 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 (which we call the '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, whereas 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 seems to be a common characteristic of pluripotent cells, including mouse embryonic stem and induced pluripotent cells and human oocytes. Our results offer a new strategy for classifying stem cells and support the idea that pluripotency 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 multipotent 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 messenger RNA expression phenotypes (Fig. 1). The core reference data set, 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 designated the input cell types with terminology that carried as little preconception about

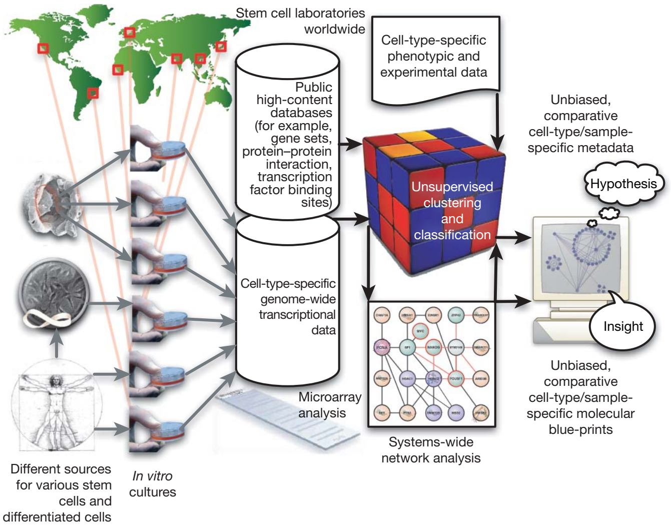

Stem cell laboratories worldwide
Cell-type-specific phenotypic and experimental data
Public high-content databases (for example, gene sets, protein–protein interaction, transcription factor binding sites)
Unbiased, comparative cell-type/sample-specific metadata
Hypothesis
Unsupervised clustering and classification
Cell-type-specific genome-wide transcriptional data
Insight
Microarray analysis
Unbiased, comparative cell-type/sample-specific molecular blue-prints
Systems-wide network analysis
Different sources for various stem cells and differentiated cells
In vitro cultures

Figure 1 | Sample collection and analysis for the stem cell matrix. Cell preparations for the stem cell matrix are cultured in the authors' laboratories or collected from other sources worldwide. Samples are assigned source codes that capture their biological origin and a relatively unbiased description of the cell type (such as BNLin for brain-derived neural lineage). Samples are collected and processed at a central laboratory 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 (for example, 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 GSA 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 that appear to function together in pluripotent stem cells.

[1]Center for Regenerative Medicine, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, California 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, California 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 North Torrey Pines Road, La Jolla, California 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, Maryland 21704, USA. [9]Center for Stem Cell Biology, Buck Institute on Aging, 8001 Redwood Boulevard, Novato, California 94945, USA. [10]Center for Neuroscience Research, Children's Hospital of Orange County Research Institute, 455 South Main Street, Orange, California 92868, USA. [11]Developmental Biology Center, University of California, Irvine, 4205 McGaugh Hall, Irvine, California 92697, USA. [12]Department for Neurosurgery University Medical Center Hamburg-Eppendorf, Martinistrasse 52, D-20246 Hamburg, Germany. †Present address: Genome and Biomedical Sciences Facility and Department of Statistics, University of California, Davis 451 Health Sciences Drive, Davis, California 95616, USA.

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.

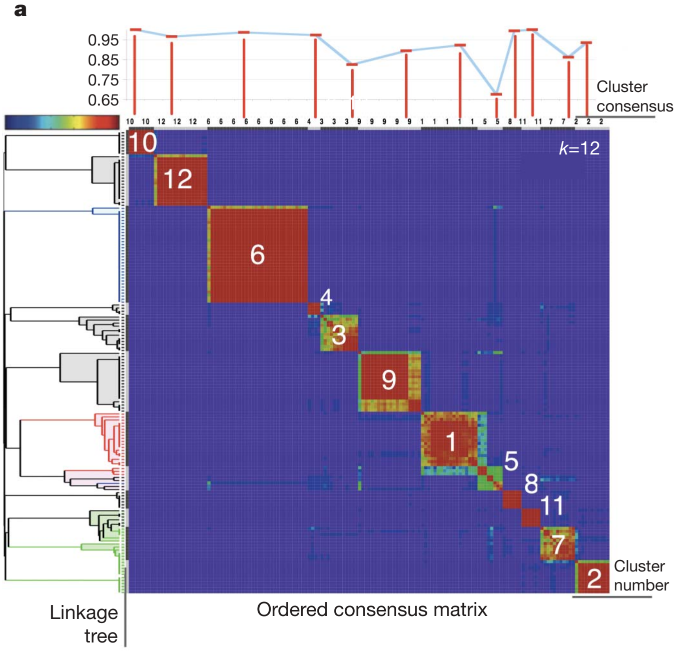
To sort the cell types we used an unsupervised machine learning approach to cluster transcriptional profiles of the cell preparations into stable distinct groups. Sparse non-negative matrix factorization (sNMF) was adjusted for this task by implementing a bootstrapping algorithm to find the most stable groupings (see also Supplementary Discussion 1)[4,5]. The stability of the clustering[9] indicated that the data set most likely contained about 12 different types of samples (Fig. 2a and Supplementary Methods 2). The composition of the stable clusters revealed both predictable and unpredicted groupings of a priori designations (Fig. 2b and Supplementary Fig. 1). The 20 samples identified as undifferentiated human pluripotent stem cell (PSC) preparations were grouped together in one dominant cluster (Fig. 2, cluster 1) and one secondary cluster (Fig. 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 (Fig. 2, cluster 6; see also Supplementary Fig. 3 and Supplementary Methods 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 Fig. 4), but the clustering clearly separated them from the other samples that had been derived from post-mortem brains of prematurely born infants (SC23 and SC30, see Fig. 2b)[10,11].

We tested the ability of our data set 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 (HUVECs)) and a recently developed new source of pluripotent cells called induced pluripotent stem cells[13–16] (iPSCs, 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 data set (Supplementary Table 10). All of the HUVEC samples clustered together and formed a distinct group. Most of the additional PSC lines (human embryonic stem cells (embryonic PSCs; ePSCs) and iPSCs) from several different laboratories were placed into a context that contained solely PSC lines. Three additional germ cell tumour lines clustered together with the tumour-derived pluripotent stem cell (tPSC) line 2102Ep and samples of three human embryonic stem (ES) cell lines: BG01v (ref. 18), Hues7 (ref. 19) and Hues13 (ref. 19). BG01v is an established aneuploid variant line and the two Hues lines are 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 data set. Gene Set Analysis[20] (GSA) is a means to identify the underlying themes in transcriptional data in terms of their biological relevance.

GSA uses lists of genes[20] 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 Fig. 2, Supplementary Table 11, Supplementary Methods and http://www.stemcellmatrix.org).

Although 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 (for example, 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 subnetworks that are co-expressed in many samples but also significantly upregulated or downregulated in a specific sample cluster.



b

Cluster	No.	No.	Source code	No.	Cell lines	No.	Cell lines
1	18	17	ePSC-UN	6 1 2 1	WA09 BG01 BG03 hES1	2 2 2 1	Miz5 Miz4 Miz6 Hues7
		1	tPSC-UN	1	2102Ep		
2	11	11	B-NLin	4 2 2 2 1	SM-1/-2 SC23 THD-hFB17/-hWB15 HFB2050 HFT13		
3	12	6	BM-MC	2 2 2	SC41AMSC SC31AMSC BMSC-21/-25		
		4	B-AS	2	SC01Glia	2	Astros
		2	ePSC-XE	2	PEL-cells		
4	4	4	CT-Fib	2 2	SC33Fib SC30Fib		
5	5	3	ePSC-UN	2 1	Hues13 Hues7		
		4	ePSC-NLin	2 2	NSWA09-4 ES-NSC-N5		
		1	B-NLin	1	HANSE8		
6	32	32	B-NLin	6 4 2 8 7	HANSE1¶ HANSE2¶ HANSE3¶ HANSE4¶ HANSE5¶	2 1	HANSE6¶ HANSE8¶

Cluster	No.	No.	Source code	No.	Cell lines
7	6	6	B-NLin	2 2 2	HNSC-CTX HNSC-HIP HOPC
8	6	6	ePSC-EB	2 2 2	Miz5 Miz4 Miz6
9	20	20	ePSC-NLin	4 5 2 5 4	NSWA09-1/-2 R-ES-NSC SNU16-NS ES-NSC-N5/-N6/-N7 ES-NSC-N8/-N9
10	8	8	ePSC-NLin	2 2 2 2	ES-NSC-N1 ES-NSC-N2 ES-NSC-N3 ES-NSC-N4
11	11	8	B-NLin	4 3 1	SC30 APC HFT13
		2	CT-Fib	2	HS27
		1	ePSC-NLin	1	NSWA09-1
12	17	17	tPSC-NLin	8 6 3	NT2 NT2CM NT2PA6

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 (centre 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; red indicates 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 to 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 and 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, source code and references see Supplementary Tables 1–10. No., number of samples. The symbol '¶' indicates that samples were derived from adult brain specimens.

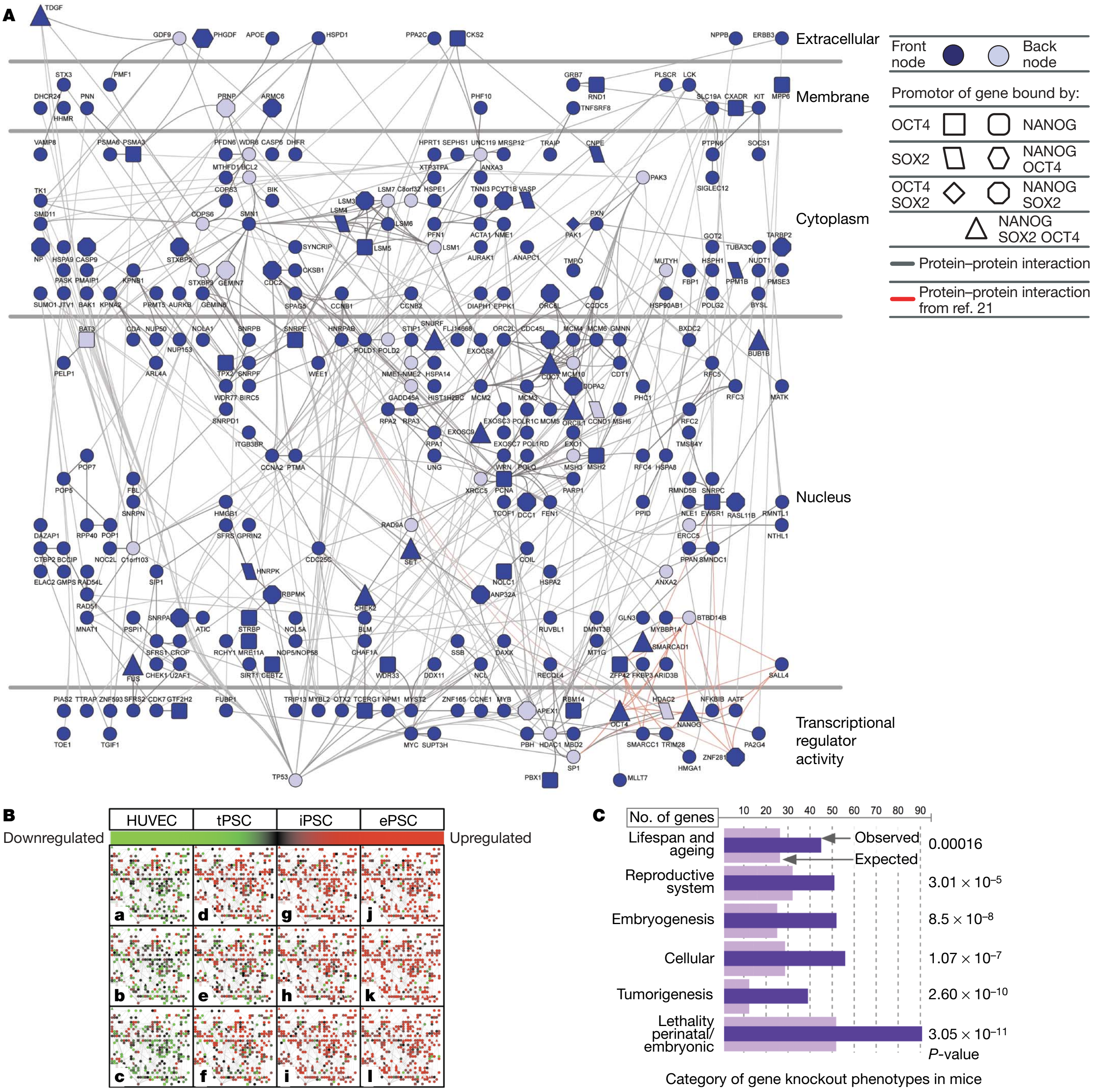


Figure 3 | Pluripotent stem-cell-specific protein–protein interaction network detected by MATISSE. Clusters from the sNMF $k = 12$ analysis were used in combination with the transcriptional database to identify protein–protein interaction networks enhanced in PSCs. **A**, A large differentially expressed connected subnetwork (PluriNet) shows the dominance of cell cycle regulatory networks in PSCs (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 data set. 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 ePSCs[21]. This subnetwork shows a significant enrichment in genes that are targeted in the genome by the transcription factors NANOG ($P = 5.88 \times 10^{-4}$), SOX2 ($P = 0.058$) and E2F ($P = 1.29 \times 10^{-16}$, all *P*-values are Bonferroni corrected). For an interactive visualization of PluriNet, see http://www.stemcellmatrix.org. **B**, Heat-map-like visualization of PluriNet genes for samples from the test data set: HUVECs (UC-EC, **a**–**c**, derived from three independent individuals), germ cell tumour-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 laboratories). Most PluriNet genes are markedly upregulated in iPSC-UN and ePSC-UN cells. tPSC-UN cells show a less consistent expression pattern. UC-EC cells show lower expression levels of most PluriNet genes. See Supplementary Fig. 5 for a larger version of the same heat maps. **C**, Analysis of genes from PluriNet in the context of phenotypes that have been reported to result from specific genetic manipulations (for example, gene knockout) in mice in the MGI 3.6 phenotype ontology database (http://www.informatics.jax.org/). We find significant over-representation of phenotypes 'lethality (perinatal/embryonic)', 'tumorigenesis', 'cellular', 'embryogenesis', 'reproductive system' and 'lifespan and ageing' 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 Fig. 6 and Supplementary Table 12.

Table 1 | PluriNet expression patterns in various model systems for pluripotency

a Expression of PluriNet genes in murine model systems

Cell type	Upregulated/downregulated
MII oocytes	Upregulated*
Zygote	Upregulated*
Embryo (two-cell blastocyst)	Upregulated*
ePSC	Upregulated†
EpiSC	Upregulated†
iPSC	Upregulated†
Fibroblasts (normal)	Downregulated†
Fibroblasts (transformed)	Downregulated†

b Successful PluriNet-based, post-hoc classification in murine model systems

Cell type	Upregulated/downregulated	Pluripotency (PAM)	Germline transmission (PAM)
ePSC	Upregulated	Yes‡	Yes‡
EpiSC	Upregulated	Yes‡	Yes‡
iPSC	Upregulated	Yes‡	Yes‡
Fibroblasts (normal)	Downregulated	Yes‡	Yes‡
Fibroblasts (transformed)	Downregulated	Yes‡	Yes‡

c Expression of PluriNet genes in human model systems

Cell type	Upregulated/downregulated
MII oocytes	Upregulated§
tPSC	Upregulated‖
ePSC	Upregulated‖¶
iPSC	Upregulated‖¶
ePSC-derived cell types	Downregulated‖
Somatic cell types	Downregulated‖¶
Somatic cancer cell line (HeLa)	Downregulated#

d Successful PluriNet-based, post-hoc classification in human model systems

Cell type	Upregulated/downregulated	Pluripotency (PAM)
tPSC	Upregulated	Yes**
ePSC	Upregulated	Yes**
iPSC	Upregulated	Yes**
ePSC-derived cell types	Downregulated	Yes**
Somatic cell types	Downregulated	Yes**

This table summarizes the expression patterns of PluriNet in various model systems of pluripotency and differentiation. More details on the specific tests and explanations of the data sources for the results can be found as indicated below. EpiSC, epiblast-derived stem cells[24]; PAM, prediction analysis of microarray, classifier with leave-one-out cross validation[27]. 'Yes' in parts **b** and **d** indicates correct classification of pluripotent state (pluripotent or not pluripotent) in >90% of samples.

* For more details see Supplementary Figs 8 and 9.

† For more details see Supplementary Fig. 10.

‡ For more details see Supplementary Fig. 10.

§ For more details see Supplementary Fig. 7.

‖ For more details see Fig. 3B and Supplementary Figs 5 and 12.

¶ For more details see Supplementary Fig. 11.

# For more details see Supplementary Discussion 2.

** For more details see Supplementary Fig. 12.

Because 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 pluripotency and self-renewal. A Nanog-associated regulatory network has been outlined in mouse embryonic PSCs[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 PSCs possess a similar NANOG-linked network (Fig. 3A; elements labelled in red). However, we also discovered that the human NANOG network seems to be integrated as a small component of a much larger protein–protein interaction network that is upregulated in human PSCs (Fig. 3). Notably, this PSC-specific network (termed pluripotency-associated network, 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 http://www.openstemcellwiki.org). Many of the genes in the PluriNet have been linked to embryogenesis, tumorigenesis and ageing (Fig. 3C and Supplementary Fig. 6). We further explored the hypothesis that pluripotency is closely linked to PluriNet expression by analysing published gene expression data sets 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 data set[22] that spans development from murine oocytes to the late blastocyst stage revealed that the PluriNet expression is dynamic and upregulated during early mammalian embryogenesis (Table 1 and Supplementary Figs 7–9)[23]. Also, our preliminary analyses indicate that the PluriNet is strongly upregulated 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 and Supplementary Figs 10–13).

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 data set reported so far, 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 have defined and make them more useful for testing hypotheses about the nature of stem cell pluripotency and multipotency. 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 haematopoietic 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 data set 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 Fig. 1. A detailed list of the samples, culture methods and reference publications is provided in Supplementary Information[11]. Generally, RNA from each sample was prepared from approximately 1×10^6 cultured cells. Sample amplification, labelling 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 subsampling 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 http://www.stemcellmatrix.org. Details on the application of GSA[20], PAM[27], MATISSE[6] as well as publicly available data sets 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 ref. 21 were mapped to the corresponding human orthologues using the NCBI HomoloGene database.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 15 December 2007; accepted 26 June 2008.
Published online 24 August 2008.

1. Müller, F. J., Snyder, E. Y. & Loring, J. F. Gene therapy: can neural stem cells deliver? *Nature Rev. Neurosci.* **7**, 75–84 (2006).
2. Murry, C. E. & Keller, G. Differentiation of embryonic stem cells to clinically relevant populations: lessons from embryonic development. *Cell* **132**, 661–680 (2008).
3. Adewumi, O. *et al.* Characterization of human embryonic stem cell lines by the International Stem Cell Initiative. *Nature Biotechnol.* **25**, 803–816 (2007).
4. Brunet, J. P., Tamayo, P., Golub, T. R. & Mesirov, J. P. Metagenes and molecular pattern discovery using matrix factorization. *Proc. Natl Acad. Sci. USA* **101**, 4164–4169 (2004).

5. Gao, Y. & Church, G. Improving molecular cancer class discovery through sparse non-negative matrix factorization. *Bioinformatics* **21**, 3970–3975 (2005).
6. Ulitsky, I. & Shamir, R. Identification of functional modules using network topology and high-throughput data. *BMC Syst. Biol.* **1**, 8 (2007).
7. Carpenter, M. K., Rosler, E. & Rao, M. S. Characterization and differentiation of human embryonic stem cells. *Cloning Stem Cells* **5**, 79–88 (2003).
8. Goldman, B. Magic marker myths. *Nature Reports Stem Cells*. doi:10.1038/stemcells.2008.26 (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. *Mach. Learn.* **52**, 91–118 (2003).
10. Palmer, T. D. *et al.* Cell culture. Progenitor cells from human brain after death. *Nature* **411**, 42–43 (2001).
11. Schwartz, P. H. *et al.* Isolation and characterization of neural progenitor cells from post-mortem human cortex. *J. Neurosci. Res.* **74**, 838–851 (2003).
12. Kornblum, H. I. & Geschwind, D. H. Molecular markers in CNS stem cell research: hitting a moving target. *Nature Rev. Neurosci.* **2**, 843–846 (2001).
13. Takahashi, K. & Yamanaka, S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* **126**, 663–676 (2006).
14. Takahashi, K. *et al.* Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell* **131**, 861–872 (2007).
15. Yu, J. *et al.* Induced pluripotent stem cell lines derived from human somatic cells. *Science* **318**, 1917–1920 (2007).
16. Park, I. H. *et al.* Reprogramming of human somatic cells to pluripotency with defined factors. *Nature* **451**, 141–146 (2008).
17. Okita, K., Ichisaka, T. & Yamanaka, S. Generation of germline-competent induced pluripotent stem cells. *Nature* **448**, 313–317 (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.* **22**, 421–428 (2004).
19. Cowan, C. A. *et al.* Derivation of embryonic stem-cell lines from human blastocysts. *N. Engl. J. Med.* **350**, 1353–1356 (2004).
20. Efron, B. & Tibshirani, R. On testing the significance of sets of genes. *Ann. Appl. Stat.* **1**, 107–129 (2007).
21. Wang, J. *et al.* A protein interaction network for pluripotency of embryonic stem cells. *Nature* **444**, 364–368 (2006).
22. Wang, Q. T. *et al.* A genome-wide study of gene activity reveals developmental signaling pathways in the preimplantation mouse embryo. *Dev. Cell* **6**, 133–144 (2004).
23. Chambers, I. *et al.* Nanog safeguards pluripotency and mediates germline development. *Nature* **450**, 1230–1234 (2007).
24. Tesar, P. J. *et al.* New cell lines from mouse epiblast share defining features with human embryonic stem cells. *Nature* **448**, 196–199 (2007).
25. Golub, T. R. *et al.* Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**, 531–537 (1999).
26. Donoho, D. & Stodden, V. When does non-negative matrix factorization give correct decomposition into parts? *Proc. NIPS* (2003) ⟨http://books.nips.cc/papers/files/nips16/NIPS2003_LT10.ps.gz⟩.
27. Lacayo, N. J. *et al.* Gene expression profiles at diagnosis in *de novo* childhood AML patients identify FLT3 mutations with good clinical outcomes. *Blood* **104**, 2646–2654 (2004).
28. Ewing, R. M. *et al.* Large-scale mapping of human protein-protein interactions by mass spectrometry. *Mol. Syst. Biol.* **3**, 89 (2007).
29. Mishra, G. R. *et al.* Human protein reference database–2006 update. *Nucleic Acids Res.* **34**, D411–D414 (2006).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank C. Stubban, H. Dittmer, S. Zapf and H. Meissner for their work with various cell cultures. We are grateful to D. Wakeman, R. Gonzalez, S. McKercher, J. P. Lee, H.-S. Park and S. Y. Moon for sharing their cell preparations for the type collection. We are also grateful to R. Wesselschmidt and M. Pera for their unique GCT lines and G. Daley for providing human iPSCs. A. M. Kocabas and J. Cibelli shared their human oocyte expression data with us. A. Barsky let us use the Cerebral 2.0 plug-in before its publication. M. Rosentraeger helped to compile the cell culture metadata. We thank J. Aldenhoff, D. Hinze-Selch, M. Westphal, K. Lamszus, U. Kehler, D. Barker and A. Fritz for their support and discussions of this project. This study has been supported by the following grants and awards: Christian-Abrechts University Young Investigator Award (F.-J.M.), SFB-654/C5 Sleep and Plasticity (F.-J.M. and D. Hinze-Selch), Hamburger Krebsgesellschaft Grant (N.O.S.), Edmond J. Safra Bioinformatics program fellowship at Tel-Aviv University (I.U.), Converging Technologies Program of The Israel Science Foundation Grant No 1767.07 (R.S.), Raymond and Beverly Sackler Chair in Bioinformatics (R.S.), Reproductive Scientist Development Program Scholar Award K12 5K12HD000849-20 (L.C.L.), California Institute for Regenerative Medicine Clinical Scholar Award (L.C.L.), NIH P20 GM075059-01 (J.F.L.), the Alzheimer's Association (J.F.L.), and anonymous donations in support of stem cell research.

Author Contributions J.F.L. and F.-J.M. designed the study and wrote the manuscript; I.U., R.W., D.K., R.S., L.C.L. and F.-J.M. designed and conducted the bioinformatics analysis; L.C.L., C.L., P.H.S., M.S.R., I.-H.P., F.-J.M. and N.O.S. conducted experiments and provided essential materials for this study.

Author Information The microarray data have been deposited at NCBI GEO (accession number GSE11508) and can also be accessed, processed and downloaded at http://www.stemcellmesa.org. Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to F.-J.M. (fj.mueller@zip-kiel.de) or J.F.L. (jloring@scripps.edu).

METHODS

Compilation of type collection. Samples were either grown in our own laboratory or provided by collaborators. Each sample was prepared from approximately 1×10^6 cultured cells, which were mechanically harvested, pelleted and snap frozen in liquid nitrogen. Biological replicates were produced for almost all samples. Details on the included cell lines and culture methods can be found in the Supplementary Tables 3–8.

Neural progenitor cultures (HANSE) from neurosurgical specimens. Brain tissue samples were obtained from patients who underwent surgery for intractable temporal lobe epilepsy at the Department of Neurosurgery, University Medical Center Hamburg-Eppendorf, Germany ($n = 6$; 4 males and 2 females; mean age 33). All procedures were performed with informed consent and in accordance with institutional human tissue handling guidelines. We used modifications of reported protocols for establishing neural progenitor cultures from fetal and postmortem brain tissue[10,30]. A more detailed description can be found in Supplementary Methods 1.

Whole-genome gene expression. All RNA was purified in our laboratory using standard methods. Sample amplification, labelling and hybridization on Illumina WG8 and WG6 Sentrix BeadChips were performed for all arrays in this study according to the manufacturer's instructions (Illumina) using an Illumina BeadStation (Burnham Institute Microarray Core).

Microarray data pre-processing. Raw data extraction was performed with BeadStudio v1.5 and probes with a detection score of less than 0.99 in all of the samples were discarded. The resulting probes were then quantile-normalized to correct for between-sample variation[31]. The sample data were quality controlled before normalization using the quality parameters provided by BeadStudio software. Before and after normalization the arrays were inspected with signal distribution box plots and by using the maCorrPlot package[32].

Parameters for unsupervised classification. The data sets and the sparseness factor λ were adjusted for the unsupervised clustering task following previous reports[4,5]. Parameters we have used for this study are: SCM core data set (153 samples), $\lambda = 0.01$; SCM test data set (219 samples), $\lambda = 0.1$. The pre-processed data sets used can be downloaded at http://www.stemcellmatrix.info.

Gene expression and gene set analysis. To screen for differentially expressed groups of genes between computationally defined sample clusters we used the Gene Set Analysis (GSA) methods proposed previously[33,34]. GSA was chosen because it uses a stringent max-mean algorithm to identify significantly differentially regulated gene sets. The cutoff *P*-value was adjusted to accommodate a false discovery rate (FDR) of 10%. A translation file was built to use GSA with Illumina expression data. We collected gene lists from recent publications and public repositories (MolSigDB2, Stanford repository). These files can be downloaded from http://www.stemcellmatrix.org. To screen for differentially expressed genes between computationally defined sample clusters we used standard *t*-test-based methods implemented in the R Bioconductor package[35]. The cutoff *P*-value was adjusted to accommodate a FDR of 5%.

Detection of cluster-specific subnetworks using MATISSE. MATISSE[6] (http://acgt.cs.tau.ac.il/matisse) was adjusted to detect differentially expressed connected subnetworks (DECSs), corresponding to connected subnetworks in a physical interaction network that show a significant co-expression pattern. The physical network used by MATISSE contains vertices corresponding to genes and edges corresponding to protein–protein and protein–DNA interactions. 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]. In total, the network contained 34,212 interactions among 9,355 proteins.

Originally, MATISSE used the Pearson correlation coefficient as a measure of similarity between the expression patterns of gene pairs. These similarity values serve as a starting point for the computation of pair-wise weights using a probabilistic model. The Pearson correlation between a pair of genes captures a global similarity trend between their expression patterns. In this work we were interested in extracting groups of genes that are not only similar across the experimental conditions, but also show significantly high or significantly low expression values in a specific subset of the samples, identified using the sNMF clustering scheme. To this end we devised a hybrid similarity score that captures two features: (1) both genes show differential expression; (2) the genes have similar expression patterns, regardless of their differential expression.

We denote the expression pattern of gene i by $x^i = (x^i_1, x^i_2, \ldots, x^i_m)$. Assume that we are interested in DECSs upregulated in a condition subset $A \subseteq \{1, \ldots, m\}$. To address goal (1), we use an 'ideal' expression profile $p = (p_1, p_2, \ldots, p_m)$ where $p_i = 1$ if $i \in A$ and $p_i = -1$ otherwise. The signs are reversed if we are interested in a DECS downregulated in A. r_{kp} is the Pearson correlation coefficient between x^k and p. Intuitively, r_{kp} is close to 1 if the corresponding transcript is strongly upregulated in A compared to the other conditions, and close to -1 if it is strongly downregulated in A. This measure has been suggested as an aparametric differential expression score[36]. Note that the Pearson correlation is invariant under normalization of the patterns to zero mean and standard deviation of 1. For every gene pair (i,j) we compute $S_{\text{diff}}(i,j) = (r_{ip} + r_{jp})/2$. To address goal (2) we use the partial correlation coefficient between the gene patterns conditioned on the ideal profile. Formally, $S_{\text{part}}(i,j) = \dfrac{r_{x^i,x^j} - r_{x^i,p}\, r_{x^j,p}}{\sqrt{(1 - r^2_{x^i,p})(1 - r^2_{x^j,p})}}$, where r_{yz} is the Pearson correlation coefficient between the profiles y and z. Intuitively, S_{part} conveys the information about how similar x^i and x^j are, regardless of their differential expression in A. Finally, we use the similarity score $S = \lambda S_{\text{diff}} + S_{\text{part}}$, where λ is a trade-off parameter setting the relative importance of the differential expression in the similarity score. We used $\lambda = 3$ for the analysis described in this paper. These S scores are then modelled using the probabilistic model described previously[6]. The advantage of using this pair-wise scoring scheme over the use of gene-specific differential expression scores, such as those proposed by others[37], is that it will prefer gene groups that are not only differentially expressed in the specified condition subset, but also have coherent expression profiles.

To diminish the effect of the size difference between the clusters, we reduced the number of conditions in clusters 1, 2, 3, 6, 9, 10 and 12, by including fewer replicates. Overall, 105 samples were used in the MATISSE analysis and can be downloaded at http://www.stemcellmatrix.org. We executed this MATISSE variant iteratively, each time setting A to contain all the samples of a single cluster or a cluster pair. The upper bound on module size was set to 300 and the rest of the parameters were as previously reported[6]. We then filtered the resulting networks by removing DECSs that overlapped more than 50% with other, higher scoring DECSs. The full set of the DECSs is available at http://www.stemcellmatrix.org.

Visualization. For visualization of the selected DECS we used Cytoscape 2.5 (ref. 38) and Cerebral 2.0 (ref. 39). Localization data from HRPD and the GO-Molecular function categories were also used[29]. NANOG, POU5F1/OCT4 and SOX2 promoter binding information was used to code the ESC-specific regulation of nodes[40]. Permutmatrix was used for heat maps[41]. Data for the analysis of human oocytes were accessed on the authors' or the journals' website[42]. For analysis of iPSCs induced with LIN28, OCT4, NANOG and SOX2, the data set was obtained from the Thomson laboratory[15].

Classification based on PluriNet. We used the 299 genes from DECS (Up(1,5)A) (PluriNet) with the PAM[20] software package. Class probabilities were re-computed 10,000 times; average scores are reported in Supplementary Figs 10 and 12. We translated the human genes into their murine orthologues from PluriNet using the NCBI HomoloGene database for re-analysing murine expression profiles. The expression array data from murine fibroblasts, induced pluripotent cells, epiblast-derived stem cells and murine embryonic stem cells were downloaded from NCBI GEO[21–24].

30. Imitola, J. *et al.* Directed migration of neural stem cells to sites of CNS injury by the stromal cell-derived factor 1α/CXC chemokine receptor 4 pathway. *Proc. Natl Acad. Sci. USA* **101,** 18117–18122 (2004).

31. Barnes, M., Freudenberg, J., Thompson, S., Aronow, B. & Pavlidis, P. Experimental comparison and cross-validation of the Affymetrix and Illumina gene expression analysis platforms. *Nucleic Acids Res.* **33,** 5914–5923 (2005).

32. Ploner, A., Miller, L. D., Hall, P., Bergh, J. & Pawitan, Y. Correlation test to assess low-level processing of high-density oligonucleotide microarray data. *BMC Bioinformatics* **6,** 80 (2005).

33. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102,** 15545–15550 (2005).

34. Efron, B. & Tibshirani, R. On testing the significance of sets of genes. *Ann. Appl. Stat.* **1,** 107–129 (2007).

35. R Development Core Team, R. A language and environment for statistical computing, help files. ⟨http://www.bioconductor.org/⟩ (2007).

36. Troyanskaya, O., Garber, M., Brown, P., Botstein, D. & Altman, R. Nonparametric methods for identifying differentially expressed genes in microarray data. *Bioinformatics* **18,** 1454–1461 (2002).

37. Ideker, T., Ozier, O., Schwikowski, B. & Siegel, A. F. Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics* **18** (suppl. 1), S233–S240 (2002).

38. Cline, M. S. *et al.* Integration of biological networks and gene expression data using Cytoscape. *Nature Protocols* **2,** 2366–2382 (2007).

39. Barsky, A., Gardy, J. L., Hancock, R. E. & Munzner, T. Cerebral: a Cytoscape plugin for layout of and interaction with biological networks using subcellular localization annotation. *Bioinformatics* **23,** 1040–1042 (2007).

40. Boyer, L. A. *et al.* Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell* **122,** 947–956 (2005).

41. Caraux, G. & Pinloche, S. PermutMatrix: a graphical environment to arrange gene expression profiles in optimal linear order. *Bioinformatics* **21,** 1280–1281 (2005).

42. Kocabas, A. *et al.* The transcriptome of human oocytes. *Proc. Natl Acad. Sci. USA* **103,** 14027–14032 (2006).