A bioinformatic assay for pluripotency in human cells

Franz-Josef Müller[1,11], Bernhard M Schuldt[2,11], Roy Williams[3], Dylan Mason[4], Gulsah Altun[5], Eirini P Papapetrou[6], Sandra Danner[7], Johanna E Goldmann[5,8], Arne Herbst[1], Nils O Schmidt[9], Josef B Aldenhoff[1], Louise C Laurent[5,10] & Jeanne F Loring[5]

Pluripotent stem cells (PSCs) are defined by their potential to generate all cell types of an organism. The standard assay for pluripotency of mouse PSCs is cell transmission through the germline, but for human PSCs researchers depend on indirect methods such as differentiation into teratomas in immunodeficient mice. Here we report PluriTest, a robust open-access bioinformatic assay of pluripotency in human cells based on their gene expression profiles.

The current standard for demonstrating that human stem cells are pluripotent is based on their ability to generate a complex variety of tissues in tumors developed in immunodeficient mice. This teratoma assay is widely considered to be the most reliable and informative assay for pluripotency in human cells[1] and its use has increased substantially after the report of induction of pluripotency in somatic cells[2]. However, the generation of teratomas is technically challenging, resource-intensive, primarily qualitative and difficult to standardize, and as we have previously argued, may have limited value as a criterion for pluripotency[3]. With the rapid increase in generation of pluripotent human cells, especially induced pluripotent stem cell (iPSC) lines, there is a need for a cost-effective, animal-free alternative to the teratoma assay for assessing pluripotency in human cells[4]. The low cost and accessibility of microarray-based gene expression datasets makes transcription profiling an attractive alternative. We hypothesized that machine-learning methods that are capable of delineating stem cell phenotypes[5] based on microarray data could also predict the presence or absence of pluripotent features for unknown samples of cells.

We considerably expanded the gene expression database that we previously used for defining stem cell phenotypes[5] to a much larger dataset we term 'stem cell matrix 2' (SCM2). The SCM2 database contains ~450 genome-wide transcriptional profiles from diverse stem cell preparations from multiple laboratories, differentiated cell types, and developing and adult human tissues (**Supplementary Table 1**). The SCM2 dataset contains expression profiles from 223 human embryonic stem cell (hESC) and 41 iPSC lines. We analyzed the samples for the SCM2 dataset in a highly quality controlled pipeline, using Illumina microarrays. After appropriate transformation and normalization, we used non-negative matrix factorization (NMF) for dimension reduction and to identify unexpected patterns engrained in the datasets[6]. NMF is a systematic, unbiased approach to identify multigene features, frequently termed 'metagenes' in gene-expression studies[7], which can be used to characterize stem cell phenotypes[3].

We then assessed pluripotency of an unknown, potentially pluripotent sample by comparison of a 'query gene expression profile' from the sample to data models derived from the SCM2 dataset (**Fig. 1a** and **Supplementary Fig. 1**). Our goals were to not only develop a simple test for pluripotency but also obtain detailed information on features of the sample that deviate from typical PSC lines with a normal genome or epigenome. We based this approach on two related classifiers that use two differently constructed metagene models.

For the first classifier, termed the 'pluripotency score', we used all samples, pluripotent and non-pluripotent, to identify the metagenes that have the capability to separate pluripotent from non-pluripotent samples[5] in the SCM2 dataset (**Fig. 1b** and **Supplementary Figs. 2** and **3**). We selected the rank and number of metagenes by identifying those that provided the largest distance between margins of known pluripotent and non-pluripotent samples in the training set (**Fig. 1**, **Supplementary Fig. 4** and Online Methods). The pluripotency score is a logistic regression model that enables a probability-based choice between the two phenotypic classes.

The second classifier, termed the 'novelty score', measures the ability of an NMF model to approximate a given query gene expression profile (Online Methods)[8]. We compared the query sample to an NMF-reconstructed sample based on the well-characterized PSCs in the SCM2 dataset, determined model fit[8] and identified deviations from the expected gene expression patterns (**Fig. 1c–g**). The novelty score is a measure of technical as well as biological variations in the data; to de-emphasize the technical variation, we applied an exponential transformation to empirically weight biological over technical deviations from our model (Online Methods).

[1]Zentrum für Integrative Psychiatrie, Kiel, Germany. [2]Aachen Institute for Advanced Study in Computational Engineering Science, Rheinisch-Westfälische Technische Hochschule Aachen, Aachen, Germany. [3]Sanford Burnham Medical Research Institute, La Jolla, California, USA. [4]Independent consultant, Encinitas, California, USA. [5]Center for Regenerative Medicine, Department of Chemical Physiology, The Scripps Research Institute, La Jolla, California, USA. [6]Center for Cell Engineering and Molecular Pharmacology and Chemistry Program, Memorial Sloan-Kettering Cancer Center, New York, New York, USA. [7]Fraunhofer Research Institution for Marine Biotechnology, Lübeck, Germany. [8]Institut für Biochemie, Freie Universität Berlin, Berlin, Germany. [9]Department of Neurosurgery, University Medical Center Hamburg-Eppendorf, Hamburg, Germany. [10]University of California, San Diego, Department of Reproductive Medicine, La Jolla, California, USA. [11]These authors contributed equally to this work. Correspondence should be addressed to F.-J.M. (fj.mueller@zip-kiel.de).

RECEIVED 13 SEPTEMBER 2010; ACCEPTED 3 FEBRUARY 2011; PUBLISHED ONLINE 6 MARCH 2011; DOI:10.1038/NMETH.1580

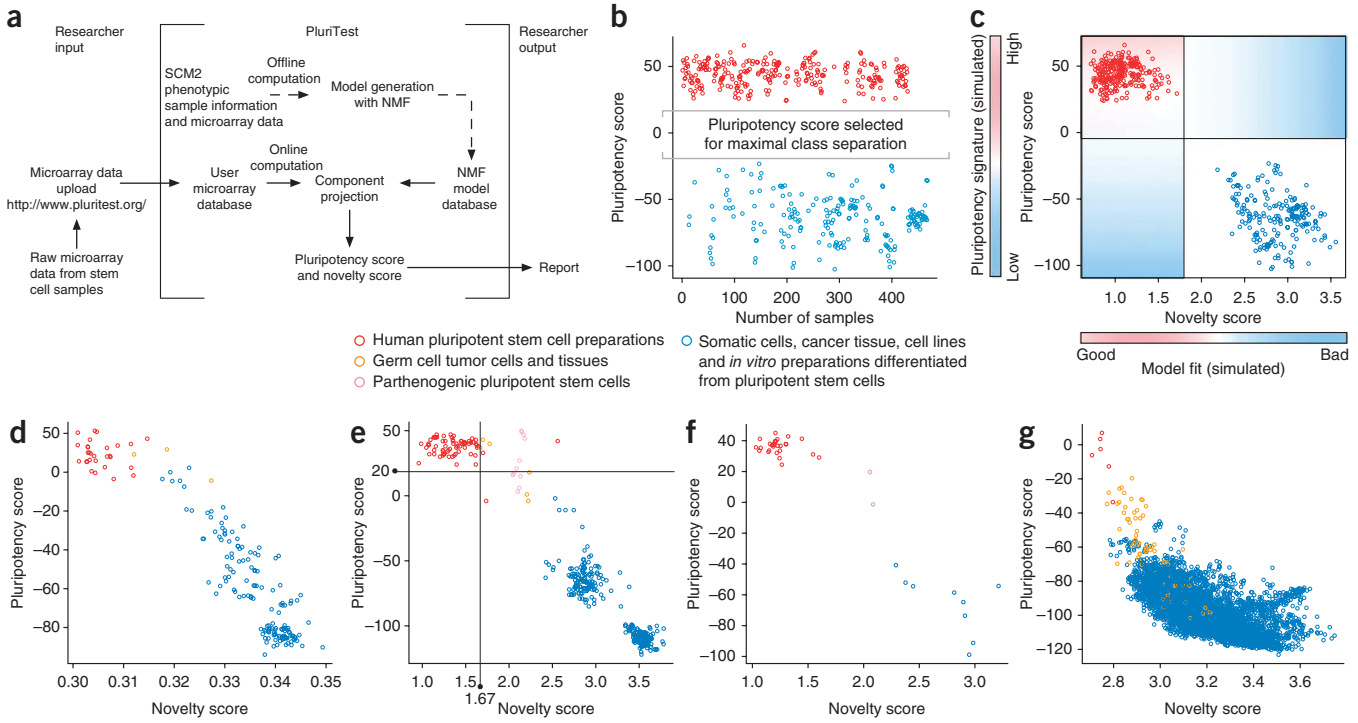

Figure 1 | A multidimensional data model for assessing PSCs. (**a**) Schematic for PluriTest. (**b**,**c**) Assessment of pluripotent and somatic cell samples in the training dataset with the pluripotency score only (**b**) and with both PluriTest scores (**c**). (**d**–**g**) PluriTest classifiers tested on datasets generated using four different microarray platforms: Illumina WG6v1 (**d**, 177 samples)[5], HT12v3 (**e**, 498 samples), HT12v4 (**f**, 38 samples) and Affymetrix U133A (**g**, 5,372 samples)[10]. Samples for these datasets were independently generated (**e**,**f**) and/or curated from published stu dies (**d**,**e**,**g**). In **e**, the lines in the plot indicate empirically determined thresholds for defining normal pluripotent lines.

The combination of the pluripotency score and the novelty score enables open-ended assessment of pluripotent features in a query sample when that sample is a new kind of PSC. The first classifier reports to what extent a query sample contains a pluripotent signature, and the second classifier reports how much of the signal measured in a query sample can be explained by the normal PSC lines contained in the SCM2 dataset (**Supplementary Note 1** and **Supplementary Fig. 1**). To test the two-classifier approach, we analyzed germ cell tumor cell lines. These cells are pluripotent and resemble normal PSCs but have genetic and epigenetic abnormalities[9]. These cells had high pluripotency scores, as expected, but the novelty score indicated that they deviate from the normal PSCs in the SCM2 dataset (**Fig. 1** and **Supplementary Fig. 2**).

We tested the combined classification approach and communication framework, which we termed PluriTest (http://www.pluritest.org/), using several independently generated test datasets containing pluripotent and non-pluripotent samples: Illumina WG6v1 (ref. 5), HT12v3 and HT12v4 datasets (**Fig. 1d**–**f**) generated in-house on our own microarray scanner and datasets that had been generated in six different core facilities (Online Methods and **Supplementary Table 1**). We also used PluriTest to examine data from a recently published human transcriptome atlas[10] based on Affymetrix U133A arrays (**Fig. 1g**).

PluriTest predicted pluripotency with excellent sensitivity and specificity. We set thresholds that separated pluripotent from non-pluripotent samples in HT12v3 test datasets with 98% sensitivity and 100% specificity, and also distinguished germ cell tumor cell lines and parthenogenetic stem cell lines from the bulk of PSCs (**Fig. 1d**–**f** and **Supplementary Fig. 2**). A few pluripotent samples had unusually high novelty scores (**Fig. 1e**), indicating that these test samples should be additionally evaluated for epigenetic or genetic abnormalities or unwanted differentiation (**Supplementary Fig. 1**). For the most informative analysis, the query sample should be analyzed on the same platform as the training dataset (Illumina HT12), but acceptable results can be obtained with data from other platforms (**Fig. 1f**, **Supplementary Fig. 3** and **Supplementary Note 2**).

We examined the performance of PluriTest on hESC lines (SIVF014, SIVF011, SIVF042, F4.2 and WA01) and human iPSC lines (HDF51IPS12 and HDF51IPS1), which were part of the training dataset; these lines grouped together and were separated from somatic samples (**Fig. 2a**). PluriTest also separated fully and partially reprogrammed iPSC lines (samples that were not in the training datatset; **Fig. 2b**); partially reprogrammed cell lines clustered with non-pluripotent cells. We then applied PluriTest to samples from a neural differentiation time-course experiment that also were not in the training dataset (**Fig. 2c**,**d**). We differentiated WA09 cells into neural precursors and collected three biological replicates on day 0, day 3, day 6 and day 14 after neural induction. The novelty score changed after 3 d of differentiation, but the pluripotency score was still high at this time point, whereas samples from later time points dropped out of the pluripotency score space and had increasingly higher novelty scores (**Fig. 2c**). In a mixing experiment in which we combined RNA samples from different time points (day 0 and day 14) at varying ratios, PluriTest could separate the differentially mixed samples (**Fig. 2d**).

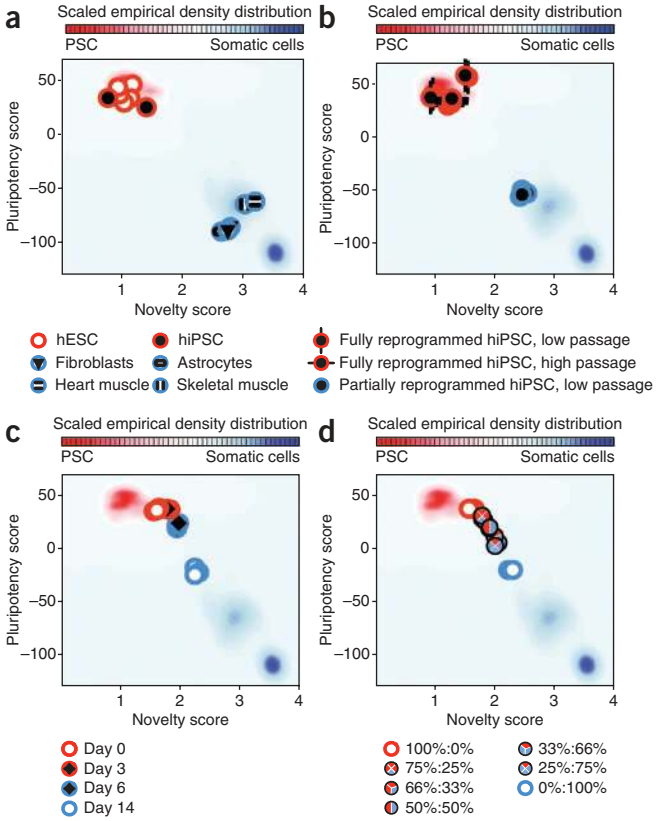

Figure 2 | Output of PluriTest. (**a**–**c**) PluriTest results for known pluripotent cells and somatic cells and tissues (**a**), for fully and partially reprogrammed iPSC lines (**b**) and for an hESC line (WA09) differentiated into neural precursors, at the indicated time points (**c**). (**d**) PluriTest results for mixed samples of hESC and hESC-derived neural precursor RNA (day 0 and day 14 from the data shown in **c**) at the indicated ratios. hiPSC, human iPSC. The background encodes an empirical density map indicating pluripotency and novelty as indicated by the color bar.

The PluriTest is contained in a single R/Bioconductor open-source, open-access workspace[11] (**Supplementary Data** and **Supplementary Note 3**) that also includes the SCM2 database–derived NMF models. To enable easy access to PluriTest, we programmed a rich internet application (RIA) using Microsoft Silverlight 4 and C# (http://www.pluritest.org/). The RIA automatically performs all data extraction and preprocessing steps after the upload of an unmodified microarray scanner output file. All data and results are stored securely in a Microsoft structured query language (MS-SQL) database. We used the binary microarray scanner output file (.idat file) as the most basic 'stem cell query term'. After upload, the results of our PSC-prediction algorithm are reported back to the user via a web interface (**Fig. 2** and **Supplementary Fig. 5**). PluriTest can run on every recent (Mac OS 10.5 and Windows XP or later) operating system, and requires internet access and a local installation of the Silverlight 4 plug-in. A typical online analysis with 12 samples takes less than 10 min including data upload (**Supplementary Note 2**).

Here we demonstrated the general feasibility of a web-based prediction of stem cell properties[12]. PluriTest breaks from the conventional marker-based approaches to assess pluripotency of human cells, which typically assay a few markers by methods such as quantitative real-time PCR. With the lowered cost of whole-genome analysis, reduction of a gene expression profile to a few markers is no longer necessary. Using all of the expression information available provides much higher discriminatory power and the ability to identify deviations from known patterns that may lead to additional insights into cellular phenotypes.

The PluriTest framework could be applied to any unbiased high-content dataset, such as global DNA methylation analysis or RNA sequencing data, provided that there is sufficient representation of a defined target phenotype in the training dataset. Our results suggest that it will be relatively straightforward to construct similar models of developmental pathways such as differentiation along the neural, endodermal or hematopoietic lineages. Such databases will inform subsequent experiments and may be applicable as a rapid method to quality control PSC-derived preparations for experimental and preclinical investigations.

METHODS

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

Note: Supplementary information is available on the Nature Methods website.

ACKNOWLEDGMENTS

F.-J.M. is supported by an Else-Kröner Fresenius Stiftung fellowship. J.F.L. is supported by grants from the California Institute for Regenerative Medicine (RT1-01108, TR1-01250 and CL1-00502), the US National Institutes of Health (R21 MH087925), the Bill and Melinda Gates Foundation, the Esther O'Keeffe Foundation and the Millipore Foundation. B.M.S. is supported by Bayer Technology Services GmbH and the Deutsche Forschungsgemeinschaft (GSC 111). L.C.L. is supported by a US National Institutes of Health National Institute of Child Health and Human Development K12 Career Development award. E.P.P. was supported by New York State Stem Cell Science grant N08T-060. We thank C. Lynch and H. Tran for preparing the samples and running the arrays; A. Schuppert, S. Peterson and K. Nazor for comments, criticisms and reading the manuscript for clarity; M. Sadelain (Memorial Sloan-Kettering Cancer Center) for providing samples and data; K. Haden and I. Mikoulitch (Illumina) for help with handling Illumina BeadArray data formats and letting us use the idat.reader.dll program module in PluriTest RIA; and C. Becker, D. Barker and A. Fritz for helpful discussions.

AUTHOR CONTRIBUTIONS

F.-J.M. conceived and designed the study. F.-J.M. and B.M.S. developed the PluriTest algorithm. F.-J.M., J.F.L., L.C.L. and J.B.A. oversaw the sample collection, microarray analysis and coordinated biological and bioinformatic experiments. R.W., D.M., B.M.S. and A.H. implemented the online bioinformatic platform. R.W., D.M., F.-J.M., B.M.S. and G.A. provided bioinformatic analyses. E.P.P., S.D., J.E.G. and N.O.S. prepared biological samples. F.-J.M., B.M.S. and J.F.L. wrote the manuscript with input from all authors.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/naturemethods/.
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/.

1. Daley, G.Q. *et al. Cell Stem Cell* **4**, 200–201 (2009).
2. Takahashi, K. & Yamanaka, S. *Cell* **126**, 663–676 (2006).
3. Müller, F.J. *et al. Cell Stem Cell* **6**, 412–414 (2010).
4. Russell, W.M.S. & Burch, R.L. *The Principles of Humane Experimental Technique* (Methuen, London, 1959).
5. Müller, F.J. *et al. Nature* **455**, 401–405 (2008).
6. Lee, D.D. & Seung, H.S. *Nature* **401**, 788–791 (1999).
7. Brunet, J.P. *et al. Proc. Natl. Acad. Sci. USA* **101**, 4164–4169 (2004).
8. Tax, D.M.J. & Muller, K.-R. *IEEE Proc. Pattern Recognition* **3**, 1051–4651/04 (2004).
9. Josephson, R. *et al. Stem Cells* **25**, 437–446 (2007).
10. Lukk, M. *et al. Nat. Biotechnol.* **28**, 322–324 (2010).
11. R Development Core Team. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, 2010).
12. Gray, J. in *Data-Intensive Scientific Discovery* (eds., Hey, T., Tansley, S. and Tolle, K.) XVII–XXXI (Microsoft Research, Redmond, Washington, USA, 2009).

ONLINE METHODS

Microarray analysis. Sample runs were analyzed in-house essentially as reported previously[5], except that Illumina HT12 arrays were used. We first filtered the probes that are present on both Illumina HT12v3 and HT12v4 arrays to ensure identical results when either of the two array versions were used with the PluriTest application. We filtered for probes that were detected with a *P* value of at least < 0.01 in at least ten samples of the SCM2 dataset. After filtering, 22,135 probes were retained and raw probe expression values were transformed and normalized with the variance stabilization transformation and robust spline normalization functions as implemented in the lumi R/Bioconductor package[13]. We normalized sample data to an in-house well-characterized pluripotent target sample (WA09).

Sample collection, test and training datasets. We analyzed 468 human samples for generating the PluriTest model. Of these, 204 were derived from somatic cells and tissues, 264 were pluripotent samples (223 hESC and 41 human iPSC; **Fig. 1b**,**c**). With these samples we trained both the multiclass and one-class classifiers. For our test datasets we analyzed samples in-house on Illumina HT12v3 (398 samples total; **Fig. 1e**) and v4 arrays (39 samples total; **Fig. 1f**) but also combined these samples with published datasets. J. Jeyakani (Genome Institute of Singapore), A. Tarca (Wayne State University), Toshima Parris (University of Gothenburg), M. Suarez-Farinas (Rockefeller University), S. Doulatov (Ontario Cancer Institute, Toronto), K. Gandhi and D. Booth (Westmead Millenium Institute, Sydney) shared the raw .idat files from their published studies (National Center for Biotechnology Information Gene Expression Omnibus (GEO) accession numbers GSE21973 (ref. 14), GSE204628 (ref. 15), GSE170489 (ref. 16), GSE2113510 (ref. 17) and GSE1868611 (ref. 18).

For the Illumina SCM1 dataset (GSE115081)[3] we focused on samples from our previous study that we analyzed on the WG6v1 platform (177 samples total; **Fig. 1d**).

For the Affymetrix U133A dataset (EM-Tab-6212; 5,372 samples total)[10] we translated the gene identifiers from the HT12v3 PLATFORM to the respective gene array annotation with a mapping table provided by Illumina (http://www.switchtoi.com/probemapping.ilmn, accessed 6 June 2010).

In the other cases (WG6v1 (GSE115081), Illumina WG6v3, HT12v4), most probes targeting specific transcripts were identical and matched based on their specific probe nucleotide universal identifiers (NuIDs)[13].

Details on all samples used for training and testing PluriTest are available in **Supplementary Table 1**.

Partially reprogrammed cell preparations. Human dermal fibroblasts (HDFs; Sciencell) were cultured in DMEM, 2 mM GlutaMax, 10% FBS and 0.1 mM non-essential amino acids (Life Technologies). HDFiPS cells were generated and maintained in standard hESC medium containing DMEM/F12 supplemented with 20% Knockout Serum Replacement (Life Technologies), 2 mM GlutaMax, 0.1 mM non-essential amino acids, 0.1 mM 2-mercaptoethanol and 12 ng ml^{-1} of bFGF (Stemgent). HDFiPS cells were cultured on irradiated mouse embryonic fibroblasts (MEFs) in hESC medium and mechanically passaged once a week. The hESC medium was changed daily.

PLAT-A packaging cells were plated onto six-well plates coated with poly(D-lysine) at a density of 1.5×10^6 cells per well without antibiotics and incubated overnight. Cells were transfected with 4 μg pMXs retroviral plasmids, which carry human *POU5F1* (also known as *OCT4*), *SOX2*, *KLF4* or *MYC* (Addgene 17217, 17218, 17219 and 17220, respectively), using Lipofectamine 2000 (Life Technologies) according to the manufacturer's instructions. Viral supernatants were collected at 48 h and 72 h after transfection, and filtered through a 0.45-μm pore size filter.

HDF cells were seeded onto a well of a six-well plate at a density of 1.5×10^6 cells per well 1 d before transduction. Cells were transduced (day 0) with equal volumes of fresh viral supernatants from all four transfections on day 2 and day 3, supplemented with 6 μg ml^{-1} of Polybrene (Sigma). On day 4, the transduced cells were split onto MEFs at a density of 10^4 cells per well of a six-well plate in hESC medium supplemented with 0.5 mM valproic acid (VPA; Stemgent). Cells were fed every other day with VPA-supplemented hESC medium for 14 d before VPA was withdrawn. The iPSC colonies were manually picked 3 weeks after transduction and transferred to MEF plates.

Twenty to thirty days after transduction the partially and fully reprogrammed cells were identified based on morphology and live staining with antibodies to TRA1-81 (1:200, R&D system, MAB1435) and SSEA4 (1:100, Stemgent 09-0011) as described previously[19] (**Supplementary Fig. 6**). Colonies that stained positive for TRA1-81 and SSEA4 and had hESC-like morphology (fully reprogrammed cells) were expanded on MEF feeders. Cells were collected for microarray analysis at passage 4 and passage 57. Colonies that showed no SSEA4 staining and very faint TRA1-81 staining (partially reprogrammed cells) were collected at passage 4. Before the cells were collected for whole-genome transcription microarrays they were again stained to confirm that the cells still expressed the correct surface cell marker.

Neural differentiation. We used a standard protocol for generating neural precursors from hESCs. hESCs were grown on Matrigel in StemPro medium (Life Technologies) until they were 30% confluent. We changed the medium then to DMEM/F12 (Life Technologies), 20% Knockout Serum Replacement with 5 mM dorsomorphin and 5 mM SB431542. Over the next 6 days, cells differentiated along a neuroectodermal lineage; on day 6 (**Fig. 2**), the population was ~95% PAX6$^+$, OTX2$^+$ and NES$^+$, and POU5F1$^-$ and Tra1-81$^-$ as assessed with flow cytometry (parallel cultures were analyzed by flow cytometry to estimate percentages). The cells were then passaged with Accutase (Life Technologies) onto a plate coated with Matrigel (BD Biosciences) and cultured in DMEM/F12 supplemented with N2/B27 media supplements (Life Technologies) and basic fibroblast growth factor (bFGF) for 8 d; during this time the primordial neural progenitor cells expanded and differentiated into more mature neural cells that were PAX6$^+$ and OTX2$^-$ (**Fig. 2**).

We profiled samples from this time course experiment in two ways: biological replicates (3 replicates) were collected on day 0 (undifferentiated hESCs), day 3 (differentiating hESCs) and before splitting the cells on day 6 (differentiating hESCs). Finally, three more biological replicates were collected after an additional 8 d in culture after the passage (day 14; neurally differentiated hESC).

In a second experiment we used the RNA obtained from the day 0 and day 14 cultures and mixed pooled RNA from those time points at seven ratios: 100% undifferentiated hESC RNA; 75% undifferentiated hESC RNA plus 25% neurally differentiated

doi:10.1038/nmeth.1580

RNA; 66% undifferentiated hESC RNA plus 33% neurally differentiated RNA; 50% undifferentiated hESC RNA plus 50% neurally differentiated RNA; 33% undifferentiated hESC RNA plus 66% neurally differentiated RNA; 25% undifferentiated hESC RNA plus 75% neurally differentiated RNA; and 100% neurally differentiated RNA.

For each of the experiments shown in **Figure 2** (different PSC lines, partially and fully reprogrammed iPSC samples, neural differentiation and RNA mixing experiments), we ran 12 samples on a single HT12v3 chip, which can be used to analyze 12 samples in parallel to minimize batch effects.

Model construction. We used a previously described dimension reduction algorithm[6] to compute NMFs.

Briefly, **V** is a data matrix from our microarray data. Each of the m columns contains gene expression values of one sample. Each of n rows in **V** contains the intensity values of a single gene probe across all samples.

$$\mathbf{V} \approx \mathbf{W} \times \mathrm{fH}$$

The NMF algorithm approximates a non-negative matrix **V** by the product of an $n \times r$ matrix **W** and an $r \times m$ matrix **H** with non-negative values with the variable n representing the number of rows, m the number of columns as above. r denotes the rank of the NMF decomposition. The column vectors of **W** can be seen as a basis that allows the approximation of **V** by linear combinations of the basis vectors. The **H** matrix contains the coordinates of the sample in the **W** basis[6].

The columns in **W** are standardized to sum to 1. We used the previously proposed procedure[6] to minimize the Euclidian distance between **V** and as implemented in the NMF R/Bioconductor package[20]. To compute the coordinates of a new sample in the basis **W** we implemented a multiplicative update algorithm[6] with a fixed matrix **W**.

$$\mathbf{H}_{ij} = \mathbf{H}_{ij} \frac{(\mathbf{W}^t \mathbf{V})_{ij}}{(\mathbf{W}^t \mathbf{W}\mathbf{H})_{ij}}$$

$\mathbf{W}^t$ denotes **W** transposed. The update process is iterated until either convergence or a maximum of 2,000 cycles.

We constructed two classifiers based on two different data subsets: a multiclass classifier based on all samples in the SCM2 dataset (tissues, somatic cells, PSCs and cells differentiated from PSCs, and a one-class classifier based on all PSC samples).

Selection of rank *k* and maximum number of features *l*. We used two different criteria to estimate the optimal number of factors determined by NMF for each of two classifiers. For the two-class classifier, we used NMF to find a low dimensional representation of all of our array data. Given a factorization of rank k we decided the optimal number of features $l < k$ to select for our classifier (rows in the **H** matrix). We calculated the area under the receiver operating characteristic (AUC) for each row of the H matrix using the sample information (pluripotency experimentally demonstrated or not) provided in the annotation file. The features h_i were ordered by the AUC and used to train a logistic regression model in R.

$$s = \mathrm{log\,it}(p) = c_0 + c_1 h_{(1)} \cdots c_l h_{(l)}$$

With the variable p representing the probability of pluripotency and the variable c denoting the logistic regression coefficients. Next, this information was used to compare the quality of different choices of k and l. We defined a quality measure r based on the margin between the pluripotent and non-pluripotent samples. As we were interested in a model that generalized well to new samples, logistic regression coefficients $c < 0$ were prohibited. This prevents the classifier from using the absence of specific non-pluripotent signatures, such as genes expressed specifically in fibroblasts, which may lead to inferior generalizability of our classifier and over-fitting to our training dataset. PSC is the set of samples defined as pluripotent:

$$r = \begin{pmatrix} \max(0, \min(s \in \mathrm{PSC}) - \max(s \notin \mathrm{PSC})) & , \min(c_{i \in \{1 \ldots l\}}) > 0 \\ 0 & , \min(c_{i \in \{1 \ldots l\}}) \leq 0 \end{pmatrix}$$

To allow comparison between different NMF factorizations we scaled r by the range of s:

$$\hat{r} = r / (\max(s) - \min(s))$$

In a more general setting a more robust quality measure may be required. We suggest using the other suitable quantiles instead the maximum minus minimum quantiles used in this case.

To select the optimal k and l values, we randomly split the training dataset (468 samples) in subtest and subtraining sets. NMF factorizations in the range from $k = 2$ to 25 were generated from the training set, with 8 random initializations for each k value. Classifiers with l values in the range 1:4 were trained. **Supplementary Figure 2** shows a plot of the mean r scaled by the range of s on the training set for the training (50% of samples chosen randomly from the 468 training samples) and test data (the remaining training samples). Classifiers with k-ranks lower than 10 achieved a good separation on the training set but did not generalize well to the test dataset. k-ranks of 13–17 resulted in classifiers that performed well on the training data. We therefore choose $k = 15$ and $l = 3$, and recalculated the classifier using the best out of 100 randomly initialized NMF approximations on the whole training dataset (468 samples). We tested the classifier on several independently generated datasets (**Fig. 1** and **Supplementary Figs. 3** and **4**).

We also derived a one-class novelty detection classifier on the samples in the training dataset based on a factorization of only the pluripotent samples in the SCM2 dataset, by using a previously described consistency approach[8] to limit the risk of over-fitting. We chose a rejection rate of 5% in a fivefold cross-validation setting. Well-characterized pluripotent samples in the SCM2 dataset were randomly assigned to one of five groups. Four of the randomly selected groups were used to train a NMF factorization, and the cutoff on the reconstruction error was set to reject the top 5% of samples with the biggest root mean squared error (RMSE).

The rejection of a sample can therefore be modeled as a binomial experiment. Given the number of test samples n we can compute the expectation and variance of the rejected samples based on the n repeated binomial experiments[8]. The samples in the test group were fitted to the $\mathbf{W}_{\mathrm{model}}$ matrix and the number of rejected samples was counted. This procedure was repeated for all five groups. A classifier was considered consistent if the mean rejection rate did not exceed the 2 s.d. (σ) bounds around the expected rejection rate. Rank $k = 12$ was the highest NMF decomposition that lead to a consistent classifier.

For the novelty classifier, we gauged the ability of the one-class NMF model to reconstruct a given query gene expression profile by the $\mathbf{W}_{\text{model}}$ basis. We first considered RMSE as suitable measure for estimating model fit. We noticed that the RMSE detected not only new biological features but also flagged some arrays analyzed in other core facilities as diverging from the one class classifier model; these particular samples were from the same PSC lines that we had analyzed in-house that did not diverge substantially from our PSC model. On the basis of such observations, we concluded that the RMSE as a novelty detection mechanism was more sensitive to technical variation than the pluripotency score. We observed in these cases that laboratory-specific variation changed most features on these arrays by a small distance, but biological variation (such as that observed in germ cell tumor cell lines) changed a restricted number of features in a sample by a large distance.

We therefore generalized the RMSE score to the *P*-weighted mean deviation (*P*-WMD) to empirically accommodate for technical variations across microarray core facilities. In the case $P = 2$ the *P*-WMD equals the RMSE and setting $N = 1$ the *P*-weighted mean deviation is reduced to a one-dimensional p-norm.

We defined the P-WMD as a distance between the reconstructed vector u and the measured vector v.

$$\text{P-WMD}(u,v) = \sqrt[p]{\frac{\Sigma_{i=1}^{N} |u_i - v_i|^p}{N}}$$

where $|u_i - v_i|$ denotes the component's absolute distances for N vector components and P the weighting exponent. As a result, for $P > 2$, components < 1 are reduced and those larger >1 gain more influence in *P*-WMD.

We determined that a *P* value in the range from 6 to 10 was optimal to increase the weight of biological variation over the technically induced deviations. Choosing $P = 8$ allowed us to reliably compare samples from several different core facilities without calibration.

To enable a probability-based assessment of the output score by PluriTest, we trained a logistical regression model for the novelty score as implemented in R/Bioconductor[11].

All model matrices and operations which are necessary to use PluriTest on novel query samples are contained in an R/Bioconductor workspace, which is available as **Supplementary Data**, and used on a local R/Bioconductor instance.

All offline computations were performed on a Cray CX1 16-core cluster with SUSE11 Enterprise and a custom compiled 64-bit R/Bioconductor implementation.

13. Du, P., Kibbe, W.A. & Lin, S.M. *Bioinformatics* **24**, 1547–1548 (2008).
14. Doulatov, S. *et al. Nat. Immunol.* **11**, 585–593 (2010).
15. Parris, T.Z. *et al. Clin. Cancer Res.* **16**, 3860–3874 (2010).
16. Gandhi, K.S. *et al. Hum. Mol. Genet.* **19**, 2134–2143 (2010).
17. Kunarso, G. *et al. Nat. Genet.* **42**, 631–634 (2010).
18. Fuentes-Duculan, J. *et al. J. Invest. Dermatol.* **130**, 2412–2422 (2010).
19. Chan, E.M. *et al. Nat. Biotechnol.* **27**, 1033–1037 (2009).
20. Gaujoux, R. & Seoighe, C. *BMC Bioinformatics* **11**, 367 (2010).

doi:10.1038/nmeth.1580