Online Methods

Microarray analysis

Analysis of samples run in-house was essential the same as reported[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 are used with the PluriTest application. We filtered for probes that were detected with a p-detection value of at least < 0.01 in at least 10 samples of the Stem Cell Matrix 2. After filtering, 22,135 probes were retained and raw probe expression values were transformed and normalized with the variance stabilization transformation (VST) and robust spline normalization (RSN) functions as implemented in the lumi R/Bioconductor package.[13] We normalized sample data toward an in-house well-characterized pluripotent target sample (WA09).

Sample collection, Test and Training datasets

We collected 468 human samples for generating the PluriTest model. 204 were derived from somatic cells and tissues, 264 were pluripotent samples (223 hESC and 41 hiPSC, **Fig. 1b** and **c**). With these samples we trained both the multi-class and one-class classifiers. For our tests 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. Justin Jeyakani (Genome Institute of Singapore, Singapore), Adi Tarca (Wayne State University, Detroit, MI USA), Toshima Parris (University of Gothenburg,Gothenburg, Sweden), Mayte Suarez-Farinas (Rockefeller University, New York, NY, USA), Sergei Doulatov (Ontario Cancer Institute, Toronto, Canada), Kaushal Gandhi and David Booth (Westmead Millenium Institute, Sydney, Australia) shared the raw .idat files from their published studies (NCBI GEO accession numbers: GSE21973[14], GSE204628[15], GSE170489[16], GSE2113510 [17], GSE1868611[18]).

For the Illumina Stem Cell Matrix 1 dataset (GSE115081)[3] we focused on those samples in our previous study that were analyzed on the WG6v1 platform (177 samples total, **Fig 1d**)

For the Affymetrix U133A dataset (EM-Tab-6212, 5372 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 Inc (http://www.switchtoi.com/probemapping.ilmn, accessed 06/10/2010).

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

A Supplementary Table (.xls) provides details on all samples used for training and testing PluriTest.

Partially reprogrammed cell preparations:

Human dermal fibroblasts (HDFs; Sciencell) were cultured in DMEM, 2mM GlutaMax, 10% fetal bovine serum 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), 2mM GlutaMAX, 0.1 mM nonessential amino acids, 0.1mM 2-Mercaptoethanol, and 12 ng/ml 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 1.5x106 cells per well without antibiotics and incubated overnight. Cells were transfected with 4 ug pMXs retroviral plasmids, which carry human OCT4/POU5F1, SOX2, KLF4 or MYC (Addgene catalog number 17217, 17218, 17219, and 17220 respectively) by Lipofectamine 2000 (Life Technologies) according to the manufacturer's instructions. Viral supernatants were collected at 48 and 72 hours post- 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 1.50 x 106 cells per well one day before transduction. Cells were transduced with equal volumes of fresh 48 hr and 72

hr viral supernatants containing each of four retroviruses on day two and day three, supplemented with 6 µg/ml of Polybrene (Sigma). On day four, the transduced cells were split onto MEFs at a density of 104 cells per well of a six well plate in hESC medium supplemented with 0.5mM valproic acid (VPA. Stemgent). Cells were fed every other day with VPA-supplemented hESC medium for 14 days before VPA was withdrawn. The iPSC colonies were manually picked three weeks post transduction and transferred onto MEF plates.

20-30 days post-transduction the partially and fully reprogrammed cells were identified based on morphology and live staining with TRA1-81 (1:200, R&D system MAB 1435) and SSEA4 (1:100, Stemgent 09-0011) as described previously (see also **Supplementary Fig. 6**).[19] Colonies that were TRA1-81 and SSEA4-positive and had hESC-like morphology (fully reprogrammed cells) were expanded on MEF feeders. Cells were harvested 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 harvested at passage 4. Before the cells were harvested for whole genome transcription microarrays they were again stained to confirm that the cells still carried the correct surface cell marker expression.

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, 20% Knockout Serum Replacement (Life Technologies), with 5mM dorsomorphin and 5mM SB431542. Over the next six days, cells differentiated along a neuroectodermal lineage; on the 6th day (day 6 in Figure 2-1), the population was approximately 95% PAX6-, OTX2- and NES-positive, and OCT4/POU5F1- and Tra1-81-negative (parallel cultures were analyzed by flow cytometry to estimate percentages). The cells were then passaged with Accutase™ onto a Matrigel™ - coated plate and cultured in DMEM/F12 supplemented with N2B27 and bFGF for eight days, during which the primordial neural progenitor cells expanded and differentiated into more mature neural cells that were PAX6- and OTX2- negative (day 14 in **Fig. 2**).

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

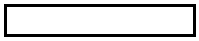
In a second experiment we used the RNA obtained from the day 0 and day 14 cultures and mixed pooled RNA from those time points at seven ratios:

100% undifferentiated hESC RNA;
75% undifferentiated hESC RNA plus 25% neurally differentiated 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;
100% neurally differentiated RNA

For each of the experiments shown in Fig. 2 (different PSC lines, partially and fully reprogrammed iPSC samples, neural differentiation and RNA mixing experiments), we run 12 samples on a single HT12v3 chip, which analyzes 12 samples in parallel to minimize batch effects.

Model construction

We used a dimension reduction algorithm first described by Lee and Seung to compute non-negative matrix factorizations (NMF).[6]
Briefly, V is a data matrix from our microarray data, the columns contain the gene expression values of each experiment.



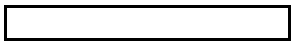
The NMF algorithm approximates a non-negative matrix V by the product of a n x r matrix W and an r x m matrix H with non-negative values. The column-vectors of W can be seen as a basis that allows the approximation of V by linear combinations of the basis vectors. The H-matrix contains the coordinates of the sample in the W-basis.[6]
The columns in W are standardized to sum to 1. We used the procedure proposed by Lee and Seung[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.

The update process is iterated until either convergence or a maximum of 2000 cycles.
We constructed two classifiers based on two different data subsets: a multi-class classifier based on all samples in the SCM2 (tissues, somatic cells, PSC and cells differentiated from PSC, and a one-class classifier based on all PSC samples.

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 (i.e. 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 were ordered by the AUC and used to train a logistic regression model in R.

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. Since we are interested in a model that generalizes well to new samples, logistic regression coefficients <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 generaliziability of our classifier and over-fitting to our training dataset. PSC is the set of samples defined as pluripotent:

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



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

Selection of rank k and maximum number of features l.

To select the optimal k and l we randomly split the training data set (468 samples, see also 'Sample collection, Test and Training datasets' paragraph above for details) in sub-test and sub-training sets. NMF factorizations in the range from k=2 to 25 were generated from the training set, with 8 random initializations for each k. Classifiers with l in the range 1:4 were trained. **Supplementary Fig. 2** shows a plot of the mean r scaled by the range of s on the training-set for the training (468 samples) and test data (398 samples, see also 'Sample collection, Test and Training datasets' paragraph above for details). Classifiers with k-ranks

lower than 10 achieved a good separation on the training set, but did not generalize well to the test data set. Ranks k in the range of 13 to 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 data set (468 samples). We tested the classifier on several independently generated data sets (see **Fig. 1**, **Supplementary Figs. 3** and **4 and** 'Sample collection, Test and Training datasets' paragraph above).

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, by using a previously described consistency approach[8] to limit the risk of over-fitting. We chose a rejection rate of 5% in a five-fold cross-validation setting. Well-characterized pluripotent samples in the SCM2 were randomly assigned to one of 5 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).

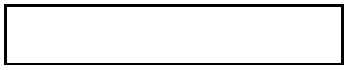
The rejection of a sample can therefore be modeled as a binomial experiment. Given the number of test samples n we can compute the expectation and variance of the rejected samples based on the n repeated binomial experiments.[8] The samples in the test group were fitted to the W_{model} -matrix and the number of rejected samples was counted. This procedure was repeated for all 5 groups. A classifier was considered to show consistency if the mean rejection rate is not exceeding the 2σ 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 gauge the ability of the one-class NMF model to reconstruct a given query gene expression profile by the W_{model}-basis. We first considered the root mean square error (RMSE) as suitable measure for estimating model fit. We noticed that the RMSE detected not only novel biological features but also flagged some arrays analyzed in other core facilities as diverging from the on class classifier model; these particular samples were from the same PSC lines that we had analyzed in-house that did not diverge significantly 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, while biological variation (such as that observed in germ cell tumor cell lines) changes a restricted number of features in a sample by a large distance.

We therefore generalized the RMSE score to the p-weighted mean deviation 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.

For p>2, values <1 are reduced and values larger >1 are amplified.



We determined that a p 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 can be downloaded 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 64bit, R/Bioconductor implementation.

References Main Text:

1. G. Q. Daley, M. W. Lensch, R. Jaenisch et al., *Cell Stem Cell* **4** (3), 200 (2009).
2. K. Takahashi and S. Yamanaka, *Cell* **126** (4), 663 (2006).
3. F. J. Müller, J. Goldmann, P. Loser et al., *Cell Stem Cell* **6** (5), 412 (2010).
4. W.M.S. Russell and R.L. Burch, *The Principles of Humane Experimental Technique*. (Methuen, London, 1959).
5. F. J. Müller, L. C. Laurent, D. Kostka et al., *Nature* **455** (7211), 401 (2008).
6. D. D. Lee and H. S. Seung, *Nature* **401** (6755), 788 (1999).
7. J. P. Brunet, P. Tamayo, T. R. Golub et al., *Proc Natl Acad Sci U S A* **101** (12), 4164 (2004).
8. David M. J. Tax and Klaus-Robert Muller, in *Proceedings of the Pattern Recognition, 17th International Conference on (ICPR'04) Volume 3 - Volume 03* (IEEE Computer Society, 2004).
9. R. Josephson, C. J. Ording, Y. Liu et al., *Stem Cells* **25** (2), 437 (2007).
10. M. Lukk, M. Kapushesky, J. Nikkila et al., *Nat Biotechnol* **28** (4), 322.
11. R_Development_Core_Team, R: A Language and Environment for Statistical Computing (R Foundation for Statistical Computing, Vienna, Austria, 2010).
12. Jim Gray, in *Data-Intensive Scientific Discovery*, edited by Toney Hey, Stewart Tansley, and Kristin Tolle (Microsoft Research, Redmond, WA, 2009).
13. P. Du, W. A. Kibbe, and S. M. Lin, *Bioinformatics* **24** (13), 1547 (2008).
14. S. Doulatov, F. Notta, K. Eppert et al., *Nat Immunol* **11** (7), 585 (2010).
15. T. Z. Parris, A. Danielsson, S. Nemes et al., *Clin Cancer Res* **16** (15), 3860 (2010).
16. K. S. Gandhi, F. C. McKay, M. Cox et al., *Hum Mol Genet* **19** (11), 2134 (2010).
17. G. Kunarso, N. Y. Chia, J. Jeyakani et al., *Nat Genet* **42** (7), 631 (2010).
18. J. Fuentes-Duculan, M. Suarez-Farinas, L. C. Zaba et al., *J Invest Dermatol* **130**

(10), 2412 (2010).
19. E. M. Chan, S. Ratanasirintrawoot, I. H. Park et al., *Nat Biotechnol* **27** (11), 1033 (2009).
20. R. Gaujoux and C. Seoighe, *BMC Bioinformatics* **11**, 367 (2010).

Supplementary Materials

Supplementary File	Title
Supplementary Figure 1	Schematic for PluriTest workflow and interpretation
Supplementary Figure 2	Threshold setting for PluriTest
Supplementary Figure 3	Application of the Illumina-derived PluriTest data model on a large transcriptomic atlas based on Affymetrix U133A arrays
Supplementary Figure 4	NMF model and feature selection for PluriTest
Supplementary Figure 5	PluriTest Rich Internet Application (RIA) design data flow from the user in the form of uploaded .idat files
Supplementary Figure 6	Characterization of fully and partially reprogrammed induced pluripotent stem cells
Supplementary Note 1	Step-by-step explanation of NMF-model based prediction of pluripotent features in stem cell microarray data
Supplementary Note 2	Usage PluriTest online
Supplementary Note 3	Usage PluriTest R/Bioconductor workspace
Supplementary Table	Samples used for PluriTest model generation and validation
Supplementary Data	PluriTest R/Bioconductor workspace