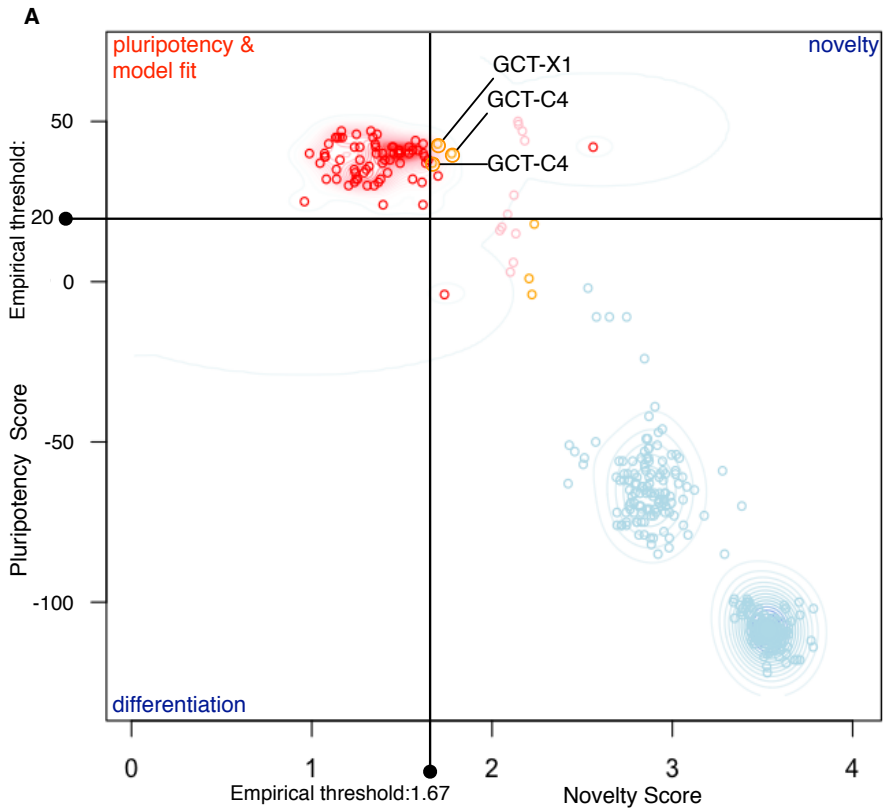


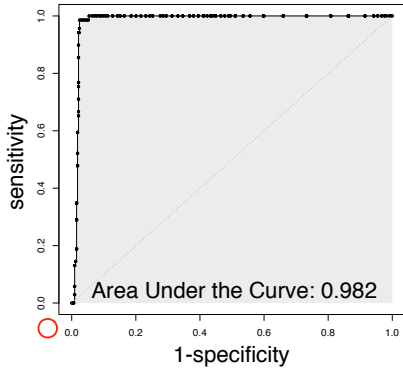
Supplementary Figure 1: Schematic for PluriTest interpretation.

We propose a structured approach toward the interpretation of the model-based summary statistics generated by PluriTest. After determination of the Pluripotency Score, the Novelty Score should be evaluated. If the Novelty Score for a pluripotent cell line is high, it could be useful to assess its genomic integrity and reprogramming or differentiation status.

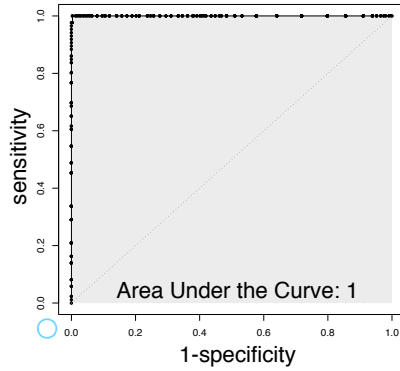
Supplementary Figure 2



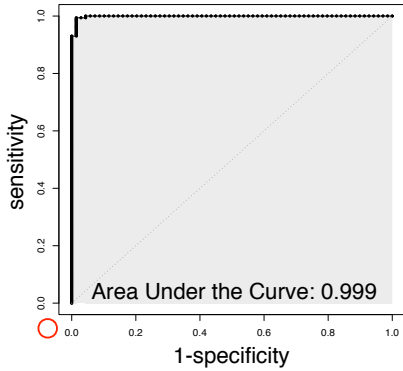
B Pluripotency Score Receiver Operating Characteristics



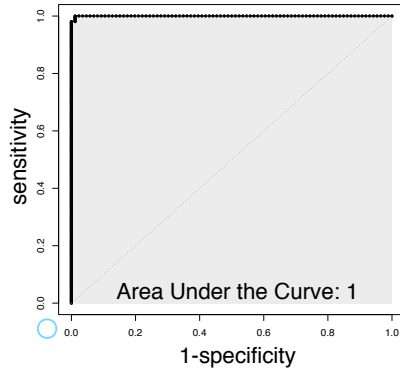
C Pluripotency Score Receiver Operating Characteristics



D Novelty Score Receiver Operating Characteristics



E Novelty Score Receiver Operating Characteristics



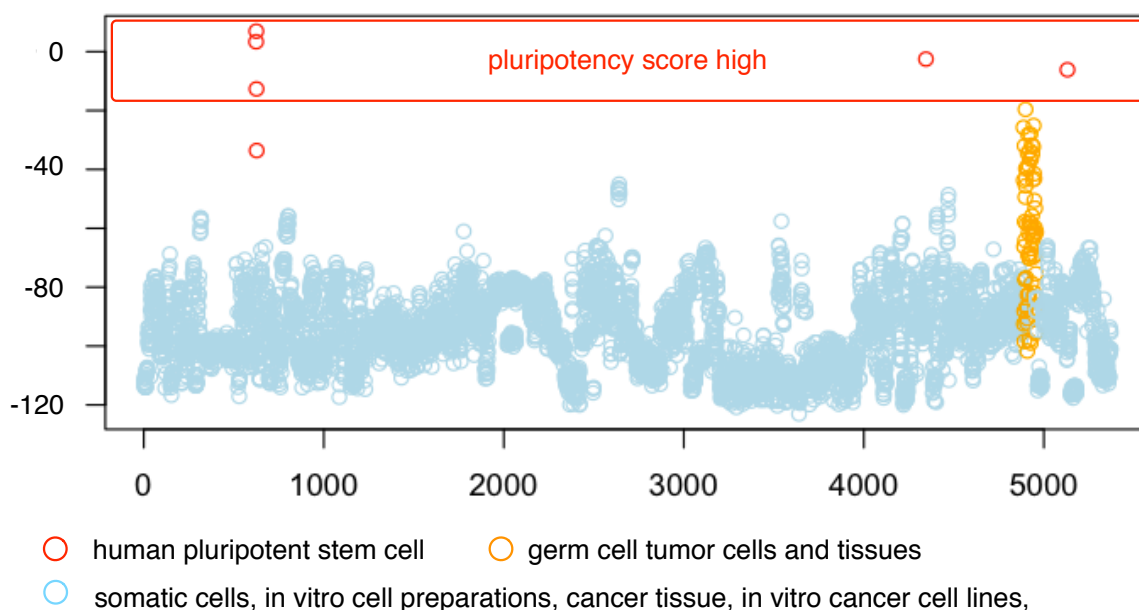
- Human pluripotent stem cell
- Germ cell tumor cells and tissues
- Somatic cells, in vitro cell preparations, cancer tissue, in vitro cancer cell lines,
- Parthenogenic pluripotent stem cells

Supplementary Figure 2: Threshold setting for PluriTest

We determined thresholds for acceptable Pluripotency and Novelty score empirically with an ROC/AUC approach based on our test data set for an sensitivity of 98.2% and specificity of 100% for the true positive identification of pluripotency in stem cell samples (see also **Online Methods**). **Supplementary Fig. 2a** depicts the same samples as in **Fig. 1e** with the empirical density distribution (as contour plot) of the same samples as background and lines indicating the empirically determined Pluripotency and Novelty Score thresholds. Future versions of PluriTest could incorporate user defined threshold settings to accommodate specific research questions. To exemplify the performance of our chosen thresholds we highlight the PluriTest performance with the Illumina HT12v3 test dataset (see also **Online Methods**). hPSC, germ cell tumor (teratocarcinoma) cells, partially reprogrammed somatic cells, differentiated hPSC and somatic cells and tissues in PluriTest. To illustrate the utility of a novelty detection classifier we tested 400 additional samples that were independently prepared and analyzed on HT12v3 microarrays on our own microarray scanner and in six other core facilities (Memorial-Sloan Kettering, NY, GIS Singapore, Singapore¹⁷, Wayne State University, MI, Rockefeller University, NY, University of Gothenburg, Sweden¹⁵, Westmead Millenium Research Institute, NSW, Australia¹⁶), Ontario Cancer Institute, Toronto, Canada¹⁴ and were not included in the model construction. Some germ cell tumor cell lines (GCT) have molecular features similar to hESC, and it would be desirable to be able to reliably detect and separate such cells from normal PSC. The *Novelty Score* was developed to detect patterns in global PSC profiles that were not present in the model data set. The GCT lines X1 and C4 show very similar levels on the *Pluripotency Score* as hiPSC and hESC lines. Other GCT lines and parthenogenic pluripotent stem cell lines (pink colored) can be easily separated from normal PSC. However, the Novelty Score is able to separate GCT-X1 and GCTC4 from normal PSC. It is straightforward to incorporate GCT lines into the pluripotency multiclass model and hence to improve the separation of these cells from karyotypically normal hESC and hiPSC on the level of the Pluripotency Score. Since we currently cannot predict what variation potentially underlies the PSC phenotype and future developments in stem cell culture and technology are unforeseeable, we have incorporated and tested an unbiased novelty detection mechanism into PluriTest with these preparations. The added value of the Novelty Score can be also gleaned from favorable Receiver Operating Characteristics (**b - e**)

While it is relatively easy to separate non-pluripotent samples from pluripotent samples with both the Pluripotency Score and Novelty Score (**c, e**), it is slightly more challenging to separate normal pluripotent samples from those with similarities to the normal pluripotent phenotype (e.g. GCT samples). Here the Novelty Score (**d**) appears to have a slight advantage over the Pluripotency Score (**b**).

Supplementary Figure 3

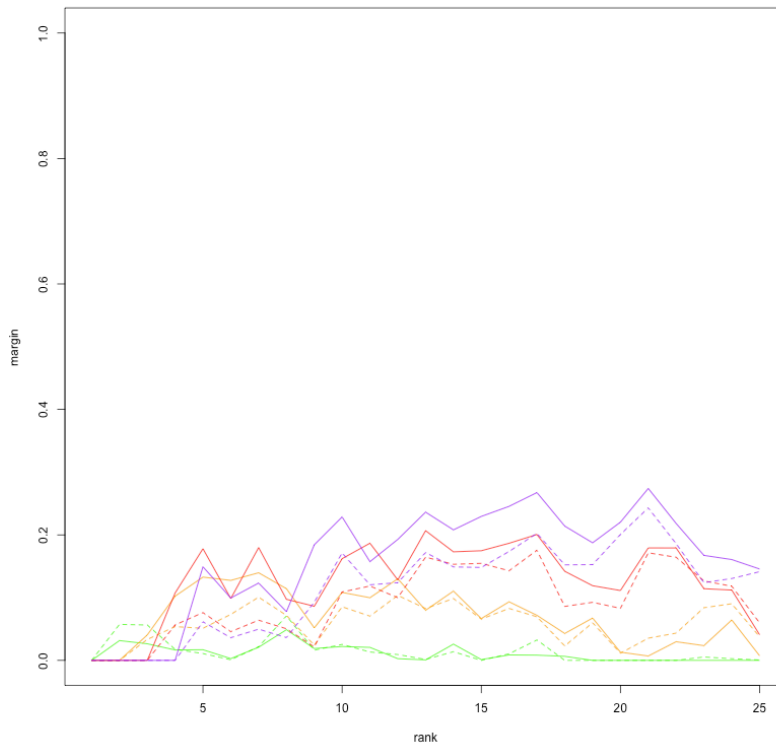


Supplementary Figure 3: Application of the Illumina-derived PluriTest data model on a large transcriptomic atlas based on Affymetrix U133A arrays

We mapped the Illumina probe identifiers onto the respective probes on the U133A platform and used our multiclass classifier derived from our in-house generated dataset on a large transcriptional atlas that was compiled from 5,372 samples representing 369 different cell types, tissues and disease states from 206 different studies conducted in 162 different laboratories.¹⁰

For this Supplementary Figure we have plotted the computed Pluripotency Score (y-axis) for each sample in the transcriptional atlas in the order of the expression matrix file (x-axis). We have plotted the Pluripotency Score vs. Novelty Score for the same samples in **Figure 1g**. Three of these studies contributed a total of six hESC samples, five of which we were able to reliably identify as pluripotent with our pre-computed data model. The one sample that we weren't able to confidently differentiate from germ cell tumor lines and tissues was an hESC sample (line HES2), which had been cultured at a high passage number (passage 128) at the time when sampled for the array analysis. Genomic analysis was not provided for this line, but aneuploidies are common in high-passage hESC lines, and this may account for this cell line's similarity to aneuploid germ cell tumors. These results support the idea that NMF-based models select stable features in PSC that can be generalized across different microarray platforms, labs and experiments.

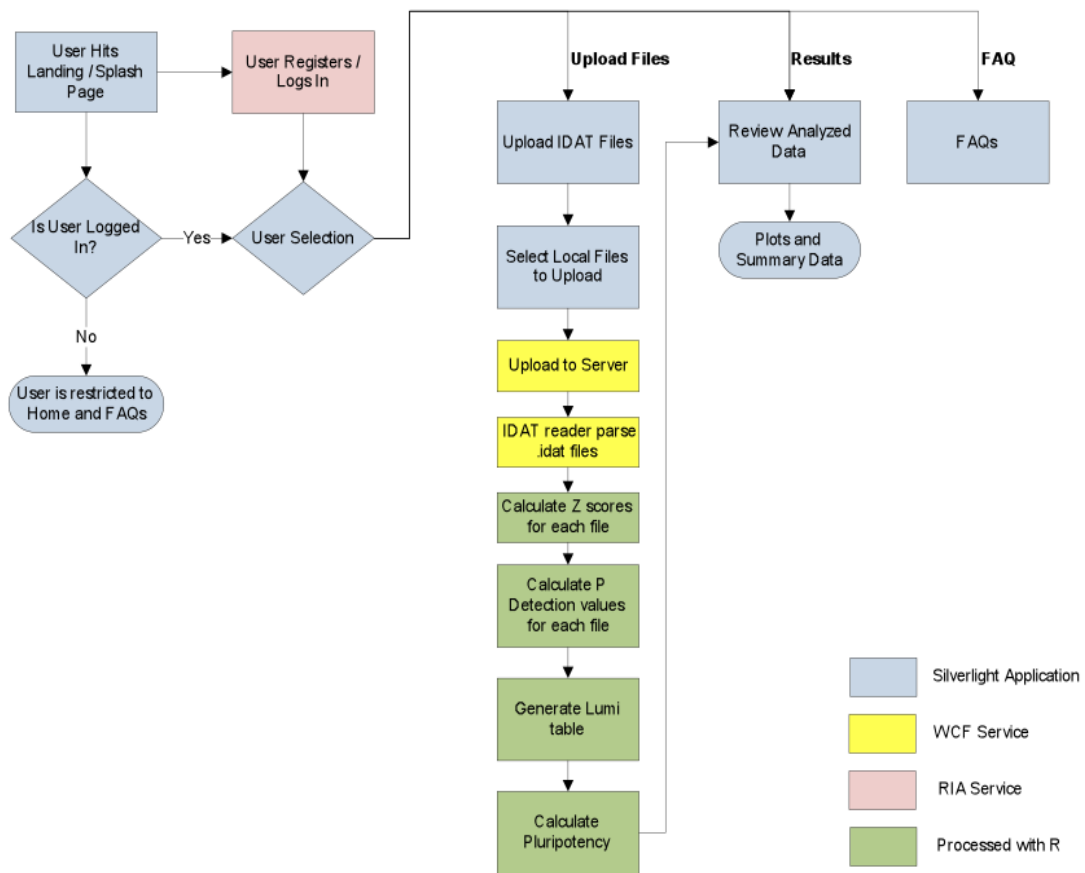
Supplementary Figure 4



Supplementary Figure 4: NMF model and feature selection for PluriTest

Quality achieved by the classifier on the test (dashed) and training set for l in 1:4 (green: $l=1$, orange: $l=2$, red: $l=3$, purple: $l=4$) We plotted the scaled margin(0-1 range) r' on the y-axis (i.e. the distance of the highest scoring, non-pluripotent sample form the lowest scoring pluripotent sample, see also **Fig. 1b** for a visualization of this concept) and the rank on the x-axis.

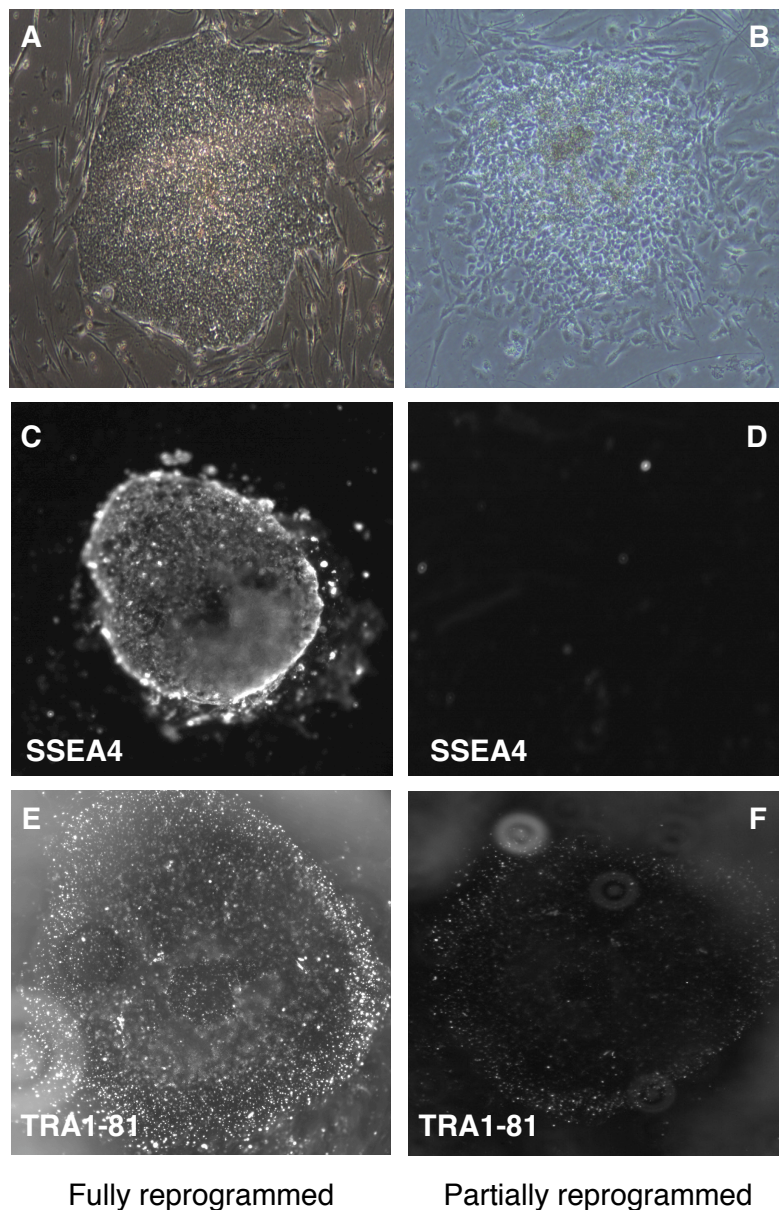
Supplementary Figure 5



Supplementary Figure 5: PluriTest Rich Internet Application (RIA) design data flow from the user in the form of uploaded .idat files

The advantage of beginning the analysis with the most “raw” file format is that all data extraction and preprocessing steps can be optimized for the specific task. The .idat files are uploaded to the server via a WCF HTTP service and are parsed into tab delimited text files, marked with a “.prs” extension. Each parsed file is used to calculate Z scores and P Detection values (using R/Bioconductor) which are written to another tab-delimited text file with a “.prc” extension. These text files are then used as inputs for calculating the Pluripotency Score and Novelty Score and generation of diagnostic plots in a specified modular R/Bioconductor¹¹ workspace. For the Z score and P detection value calculations, R is called using a COM-interop assembly, while the Lumi table generation and pluripotency calculation is done by calling R from the command-line as a shelled out process from .NET.

Supplementary Figure 6



Supplementary Figure 6: Characterization of fully and partially reprogrammed induced pluripotent stem cells

Recently, it was proposed that three kinds of colonies can be detected in reprogramming experiments, which were designated Type 1, 2, and 3.¹⁹ We were unable to maintain the Type 2 (intermediate) phenotype in our lab, but used the operational criteria proposed by Chan et al. 2009 for defining Type 1 (partially) and Type 3 (fully reprogrammed) cells.¹⁹ We also included Type 3 cells with a high passage number, since it has been suggested that iPSCs become more similar to hESCs over long term culture. We used the following criteria for identifying the two cell types: Colony-like growth on murine embryonic feeder cells (**a** and **b**), positive SSEA4 and TRA1-81 stains for fully reprogrammed iPSC (**c** and **e**) and no or faint staining for SSEA4 and TRA1-81 for partially reprogrammed cells (**d** and **f**).

Supplementary Note 1: Step-by-step explanation of NMF-model based prediction of pluripotent features in stem cell microarray data

I. Intended Audience

PluriTest uses an involved machine-learning algorithm and extensively validates the results with sometimes non-intuitive, but tried, tested and accepted statistical tools. The following section is intended to give non-experts an idea on what PluriTest involves and what it actually does. Such an explanatory section cannot replace or fully capture the procedures that we have outlined in the appropriate and correct technical terms in the Materials and Methods section. We have added links to related articles in Wikipedia for further reading.

II. Overview

PluriTest consists of two steps for two classifiers, a first step, which involves the (offline) creation of a data model and a second step, which is the actual (online) testing of new data (see also **Fig. 1a** and **Supplementary Fig. 1**). The text below explains the main procedures used for PluriTest.

III. NMF-model generation and use in PluriTest

1. Stem Cell Matrix 2 and samples used for establishment and validation of PluriTest

PluriTest is based on a large collection of gene expression microarrays that we term Stem Cell Matrix 2. The collection was initiated for our initial work on classifying cells based on their gene expression profiles ⁵, and for the current study, we collected many more hESC and iPSC lines from our own labs and from many collaborators. While we wanted diversity in the types and sources of samples, we also wanted strict quality control of the sample preparation and microarray analysis, so we prepared the RNA in house, hybridized the preparations to HT12v3 microarrays in our own lab, and scanned them on our own Illumina BeadArray scanner. A subset of this dataset (468 samples with pluripotent and non-pluripotent samples), was then used to construct the PluriTest classifiers (see the detailed explanations below).

After we had established the data model and classifier, we used additional samples run on the same array type with known identities (pluripotent vs. non-pluripotent) to determine empirical cutoffs for calling a sample pluripotent or non-pluripotent based on the combination of the *Pluripotency Score* and *Novelty Score*. This dataset did not only consist of samples that we had analyzed in house, but also included samples which we obtained from researchers who had scanned their samples with other Illumina array scanners and processed their samples in their respective labs following their own protocols.

We also tested samples of cells that we had demonstrated to be pluripotent but that had genomic

or epigenomic abnormalities (aneuploid, genomically abnormal by SNP assay, and parthenote-derived cells) to test the utility of the Novelty Score. These cells are the most stringent benchmark for validating any pluripotency assay; they appear to be normal pluripotent stem cells when examined by limited sets of markers, such as immunocytochemical assays and 96-gene RT-PCR assays, and express pluripotent gene signatures at the same or even higher level than normal pluripotent stem cells. We found that these cell types had high *Pluripotency Scores*, but also were clearly different from normal pluripotent cells, because they also showed high Novelty Scores.

We validated the PluriTest framework with data generated by other microarray versions and platforms. The Illumina microarray platform has undergone several evolutionary refinements and improvements over the last 6 years since we began collecting Stem Cell Matrix data. Those changes concerned mainly different (mostly improved) probe design (Illumina designs WG6v1 → WG6v2 → WG6v3, and HT12v3 → HT12v4) but also changes in the in bead technology (WG6v3 → HT12v3). We tested pluripotent and non-pluripotent samples that we had analyzed in house across the array generations WG6v1, WG6v3, HT12v3 and HT12v4. For the comparison of PluriTest across these versions we used probes that each of the other versions had in common with the HT12v3 array design, which we used for establishing the predictor. We also tested the predictor on data derived from another microarray platform (Affymetrix) to validate its generalizability across different technologies. Since the probe design differs fundamentally between these two microarray platforms, we could only map the genes that were represented in both Illumina and Affymetrix platforms.

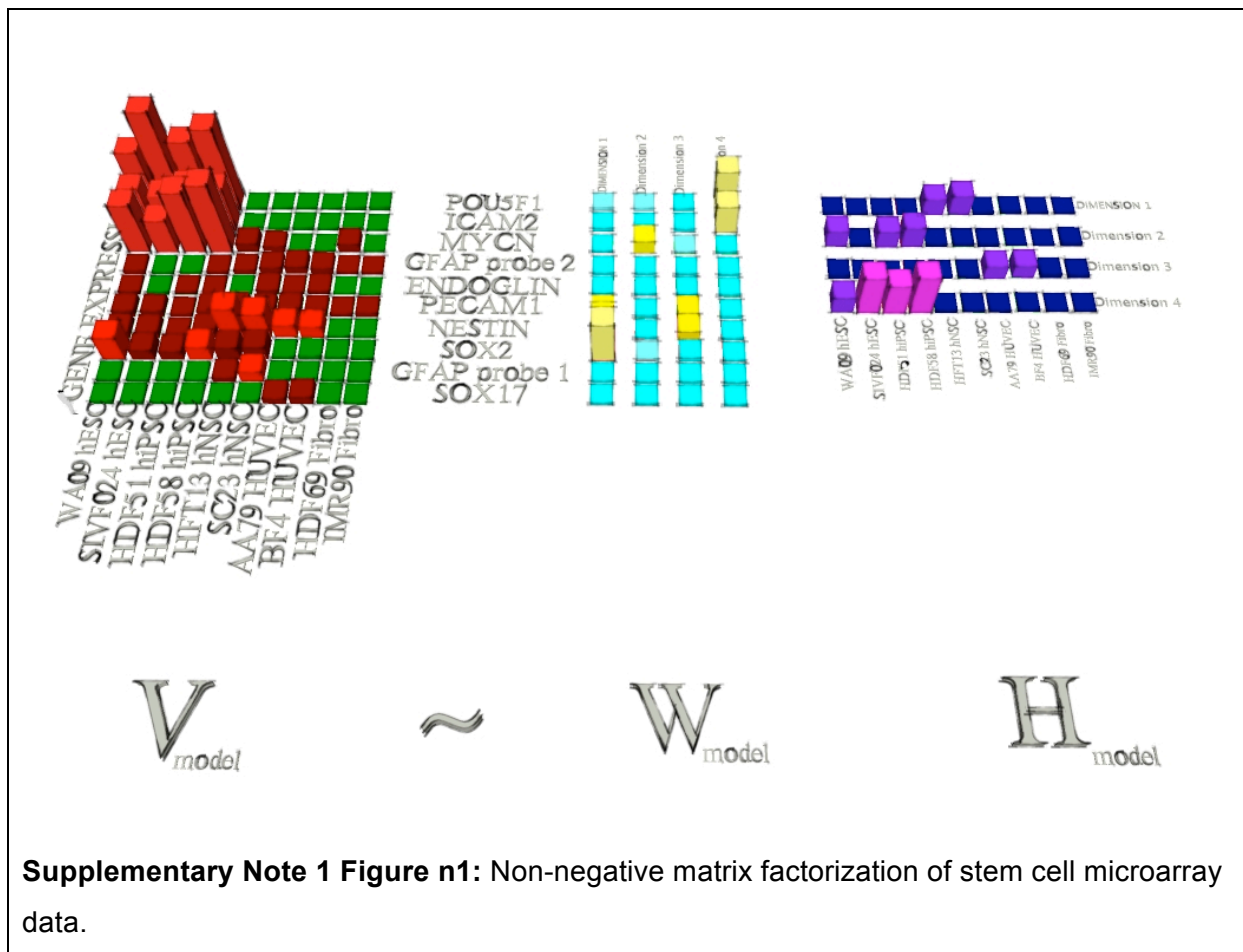
In **Fig. 1** of the main text, the raw *Pluripotency Scores* are higher and the raw Novelty Scores are lower in the pluripotent samples analyzed on HT12v3 chips when compared to all other platforms; this is because not all probes from the original HT12v3 predictor could be mapped to the other platforms and hence we see a “degeneration” of the signal through the mapping of probes from other array types. It is important to note that while the raw *Pluripotency Score* and Novelty Score values of samples are directly comparable only if they are run on the same platform, the critical information lies in the difference between the scores of pluripotent and non-pluripotent samples. As **Fig. 1b** illustrates, we have optimized the predictors for the largest distance between pluripotent and non-pluripotent signals and we would like to point out that this difference and thus the predictive capabilities of both scores remain intact in other data sets.

2. Model Generation Pluripotency Score

As first step, we construct a large gene-by-sample matrix V_{model} with microarray data from 468 samples we had analyzed in house on HT12v3 microarrays. Essentially we are searching for predictive parameters based solely on observational data. The observational data is the microarray collection of several hundreds of samples and non-negative matrix factorization (NMF [http://en.wikipedia.org/wiki/Non-negative_matrix_factorization]) is the algorithm that we use for identifying parameters that capture the information contained in the

data and which can be used for meaningful predictions.

For this explanatory section, we selected 10 genes and 10 samples from the Stem Cell Matrix 2 and represented these with columns on a grid. The heights of the columns represent the measured probe and sample-specific signal intensities. The color code stratifies measured probe intensities for greater clarity: high expression is identified with bright red, low expression intensities with dark red and expression values measured in the background-noise range with green (**Supplementary Note 1 Fig. n1**).



NMF essentially takes the observational data matrix and tries to find two smaller sub-matrices (W and H), which upon multiplication result in a “reconstructed” matrix that should resemble V . W and H capture patterns in the observed data matrix V and their multiplication “stacks” the different patterns on top of each other to reconstruct V .

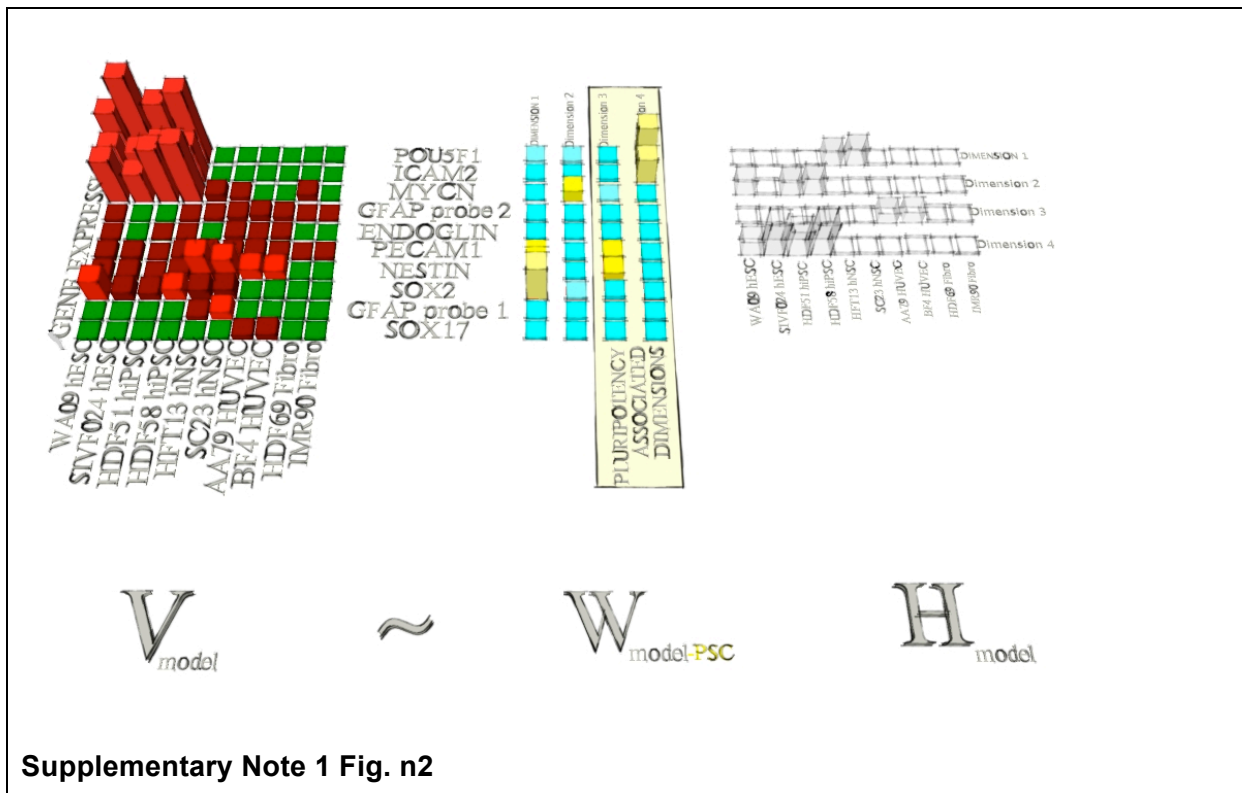
The algorithm optimizes W and H over and over again in such a way that the resulting reconstructed data becomes more and more similar to the actual data matrix. The sub-matrices have the same predefined number of columns (in the case of the W -matrix) and rows (in the case of the H -Matrix). This predefined number is called k . k is usually selected so that the columns and rows of the W - and H -matrices are much smaller than the original rows and columns in the observed data matrix. The Stem Cell Matrix we have used has

about 450 columns (samples) and about 23,000 rows (gene probes). We are using for the current PluriTest model $k=15$, which means we are reducing considerably the dimensionality (and size!) of the dataset with our model, while retaining the information necessary to reconstruct the complete database.

Finding the right k is sometimes referred to as a model selection. The choice of k influences the predictive qualities of any NMF derived signature, and hence we empirically optimized k for finding the best performing pluripotency predictor based on 1 to 4 W -dimensions ($i=1$ to 4) in our HT12v3 training and test data. We found that the best possible predictor that performs well in both datasets could be obtained with a $k=15$ and $i=3$ W -dimensions. The V -matrix is decomposed into two sub-matrices (W_{model} and H_{model}) and a selected k using the algorithm for non-negative factorization described by Lee and Seung.⁶

NMF optimizes W_{model} and H_{model} in such a way that only non-negative entries can exist in the sub-matrices. The quality of W_{model} and H_{model} determines how closely the newly generated matrix resembles V_{model} . The quality of this reconstruction can be measured by the reconstruction error. The reconstruction error is dependent on the specific NMF algorithm selected for a specific task, the data quality and number of samples in V -matrix and the number of iterative cycles NMF uses to optimize W_{model} and H_{model} .

For this explanatory section, we have chosen $k=4$ for the dimensionality in our example dataset. All data points visualized in the Supplementary Figures are similar to actually measured or computed values, but exact numbers have been omitted to emphasize the generality of this basic example (**Supplementary Note 1 Fig. n2**).



The *W*-matrix contains base vectors (columns), which can be seen as “meta-samples”. Such *W*-dimensions can be interpreted as a number of genes that are co-regulated in some sample classes. In our example, POU5F1 (OCT4) is co-regulated with ICAM2 in the reduced dataset. This exemplifies the power of such an unsupervised approach. NMF can detect significant and meaningful patterns in high-dimensional data, even if there is little prior knowledge on such a link exists (ICAM2 in this case is a hypothetical example to highlight the concept).

Importantly, each *W*-dimension associates each gene with a parameter that describes how much each gene contributes toward a specific gene expression pattern (‘load’). While a gene list with genes tested with e.g. RT-PCR or NanoString technologies has each gene weighted to the same extent, the *W*-matrix and microarray data allow for many more genes to contribute toward a specific gene signatures and weights each gene’s importance toward a single unique gene pattern. Technically, it is not possible to translate a *W*-matrix into a simple gene list because of the NMF-based, more nuanced signature parameterization and the (theoretical) possibility that all genes (with varying degrees) on an array can contribute to a specific signature. Kim and Tidor have proposed a method to extract the features/genes that pick up the most ‘loading’ of one dimension.²¹ When we apply their method to one of the three most pluripotency associated *W*-dimensions in our Pluripotency Score model, the following genes are highlighted:

ADA, AFP, ALDOC, ALPL, APLNR, APOA1, APOA2, APOA4, APOB, APOC1, APOE, ARHGAP28, ARID3B, ATP8B3, AURKB, AXIN2, BCL11A, BEX1, BEX2, BMP2, BMP4, BRSK1, BUB1, C1orf106, C20orf46, C20orf75, C21orf129, C6orf126, C9orf135, C9orf45, CA11, CA2, CACHD1, CAMKV, CBX2, CCDC81, CCKBR, CD24, CDC25A, CDCA7, CDH1, CDT1, CECR1, CELSR3, CENPF, CER1, CGN, CGNL1, CHD7, CHEK1, CHEK2, CHST4, CKB, CLDN10, CMTM7, CNTNAP2, COCH, COL4A6, COL9A2, CPVL, CRABP1, CRABP2, CRB3, CRMP1, CST1, CTSL2, CXADR, CYP26A1, CYP2S1, DBC1, DLK1, DMKN, DMKN, **DNMT3B**, DPPA4, DSC2, DSCR6, EOMES, EPHA4, EPSTI1, FAM184A, FAM46B, FAM64A, FAM69B, FAM89A, FBN3, FGA, FGB, FGFR3, FGFR4, FLJ25404, FLRT3, FOXA2, FOXA3, FRAS1, FRZB, FST, FST, FZD2, FZD7, GABRB3, GAD1, GAD1, GCA, GCHFR, GINS2, GLDC, GLIPR1L1, GPC3, GPC4, GPR19, GPR64, GPR98, GRP, GRPR, GSTA2, GYG2, HAND1, HAPLN1, HAS2, HELLS, HESX1, HLA-DOA, ID1, ID2, ID2, IGF2BP3, IGSF3, IRX2, IRX3, ISL1, KAL1, KCNG1, KCNG3, KCNK1, KCNK12, KCNN2, KIAA1553, KIF15, KIFC1, KRT19, L1TD1, LDB2, LEPREL1, **LIN28**, LIN28B, LOC133993, LOC388588, LOC401720, LOC440132, LOC57228, LOC642559, LOC644612, LOC644919, LOC645682, LOC646817, LOC91461, LPHN1, MAP4K1, MAP7, MATK, MBD2, MCM10, MCM2, MCM4, MEST, MGC39900, MGC39900, MGST1, MND1, MSX1, MT1F, MYCN, MYO5C, NEFM, NELL2, NKD2, NLGN4X, NPTX2, OIP5, OTX2, OVOL2, PACSIN1, PCDHB2, PCOLCE2, PDE9A, PDPN, PDZK1, PKMYT1, PLA2G3, PLCG2, PNCK, POLE2, **POU5F1**, POU5F1P1, PPP2R2B, PRDM14, PRKCB1, PRODH, PROM1, PUNC, RARRES2, RASL11B, RASL12, RASSF7, RBPMS2, RCOR2, RDM1, REEP6, RHBDL3, RHPN2, RIMS3, RND2, RPL29, RPL29, RPRM, RPRML, RPS15, RSPO3, SALL2, SALL4, SCNN1A, SEMA4D, SEMA6A, SEZ6L2, SFRP1, SFRP2, SH3GL3, SHISA2, SILV, SLC2A3, SLC7A3, SLC7A8, SMAD6, SMPDL3B, SORL1, SOX11, SOX17, **SOX2**, SP8, ST6GAL1, STC1, TACSTD1, TCEAL2, TCN2, TERF1, TJP3, TMEM27, TMEM88, TMPRSS2, TMSL8, TRIM71, TRO, TTR, TUBB2B, UGP2, UNQ2541, USP44, USP44, VANGL2, YBX2, ZIC2, ZIC3, ZNF423, ZNF589, ZNF702P, ZSCAN10.

For illustrative purposes, we have highlighted in red several well-studied genes that have been independently implicated in the molecular basis of pluripotency (such as POU5F1 [OCT4], SOX2 and LIN28). Also we note that many more genes with yet undetermined function in pluripotent stem cells (at least to our knowledge) are included in this list. We must emphasize that these genes differentiate PSC from other cellular phenotypes in the context of a large microarray dataset such as the Stem Cell Matrix 2 and that this gene list does not capture the more nuanced, individual parameters associated with each gene. Also, we would like to point out that gene lists as classifiers/signatures have to be cautiously interpreted. In a seminal paper, Ein-Dor used another signature detection method to find lists of genes correctly classifying breast cancer subtypes following tried and accepted methods.²² The first group of signature genes was removed from the data set, and another search for predictive gene signatures was performed; this search recovered a new gene list with equally predictive values. This procedure could be repeated several times. This important example is to caution the over-interpretation of predictive signatures in a biological context, since the data structures engrained in the whole dataset most likely contribute most to the predictive value of a signature, and the presence of a gene in a signature does not necessarily imply a biological function for that gene.

Finally, PluriTest was developed to optimize the predictive qualities of the W -dimension, not their biological interpretation. For this, we employed in our previous work⁵ additional algorithms (MATISSE) to allow for a biologically motivated interpretation of our exploratory findings.

The H_{model} -Matrix consists of “mixing values” for each of the four dimensions and each samples. Previously, we have shown that such H-matrices generated from the same V-matrix and the consensus clustering framework proposed by Monti et al. and Brunet et al. can be used to reliably classify stem cell samples into phenotypic and biologically distinct classes.⁷ Brunet and colleagues have introduced the term “meta-gene” for the rows in the H-matrix, and refer to them as “small number of gene combinations (‘metagenes’) that best capture the behavior of an expression data set”.²³

For establishing a generalizable, validated and reliable prediction algorithm for pluripotent features in human embryonic and induced stem cells, we identified and used subsets of the W -Matrix. Earlier work has shown that dimensions contained in the W -Matrix can be interpreted as meaningful parts out of which a complex dataset can be reconstructed. This was first demonstrated by Lee and Seung,⁶ and has been subsequently confirmed by Kim and Tidor²¹ and Gao and Church²⁴ for microarray analyses.

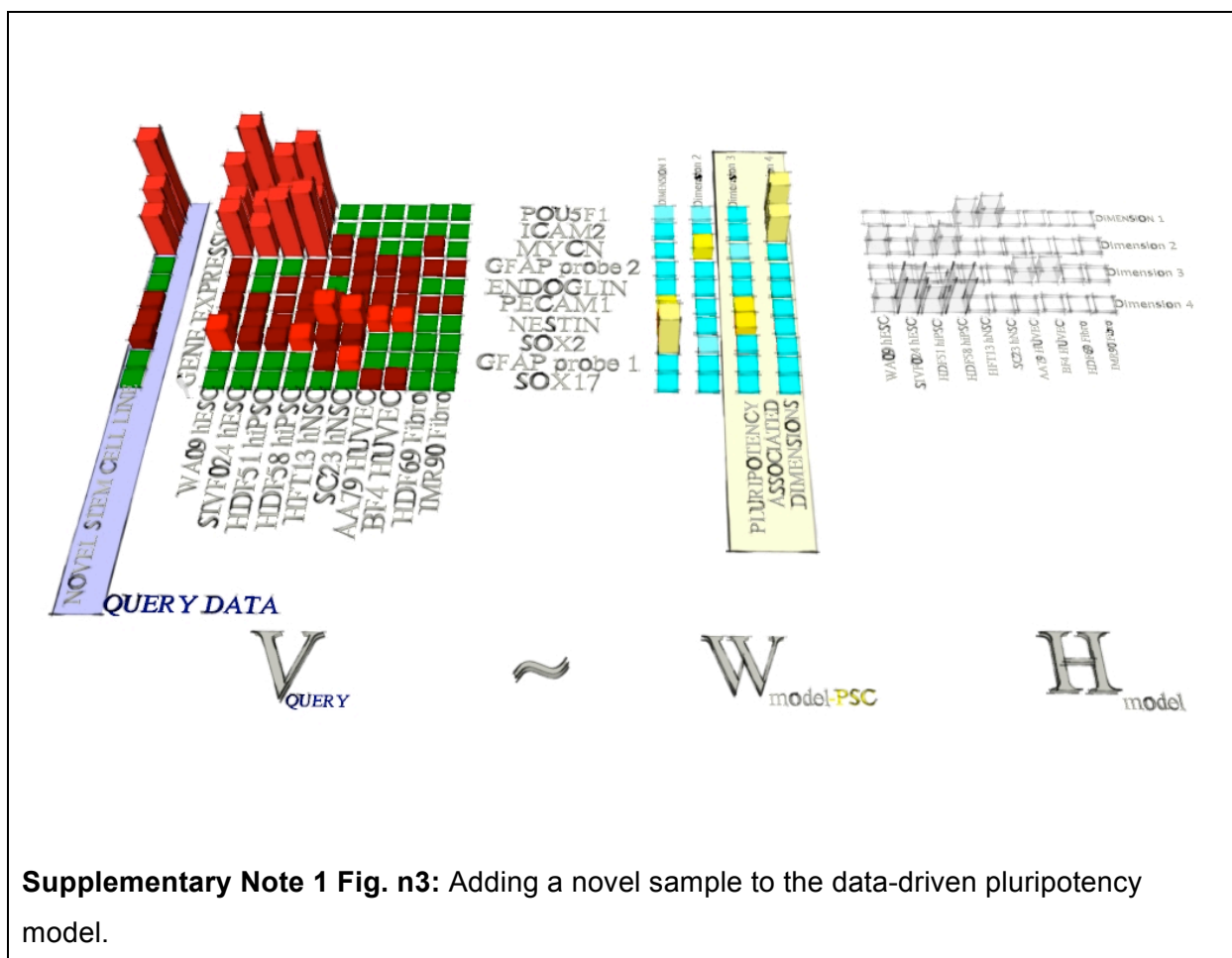
In our special case, we used our prior knowledge contained in the high-quality sample annotation of the Stem Cell Matrix 2 to identify the dimensions in the W_{model} -matrix that can explain most of the parts and variance observed in hPSC-cells in the SCM2. This is done by

selecting the three W -columns (dimensions) that best correlate with PSC. These dimensions are then set as W_{model} -PSC-matrix by using stringency criteria as described in the online Materials and Methods section.

These W -dimensions are then used in a similar manner to more conventional gene signatures, and we compute them with the H -Matrix to produce the Pluripotency Score. As expression levels of the PSC-associated gene lists go up, the Pluripotency Score will also increase since the associated metagenes also increase their values.

3. Testing an query sample with the Pluripotency Score

With these empirically and computationally generated data dimensions as our pluripotency model, we then conceptualized a novel stem cell sample as a 'query term'. This is similar to a query that is entered into a web search engine, which then uses this expression (e.g. 'pluripotent stem cell') to match it with relevant results from the search engine data model (e.g.: a web link to <http://stemcells.nih.gov/index.asp>). The main difference is that a stem cell microarray dataset is a more complex 'query term' than, for example, three words entered into a search engine search field.



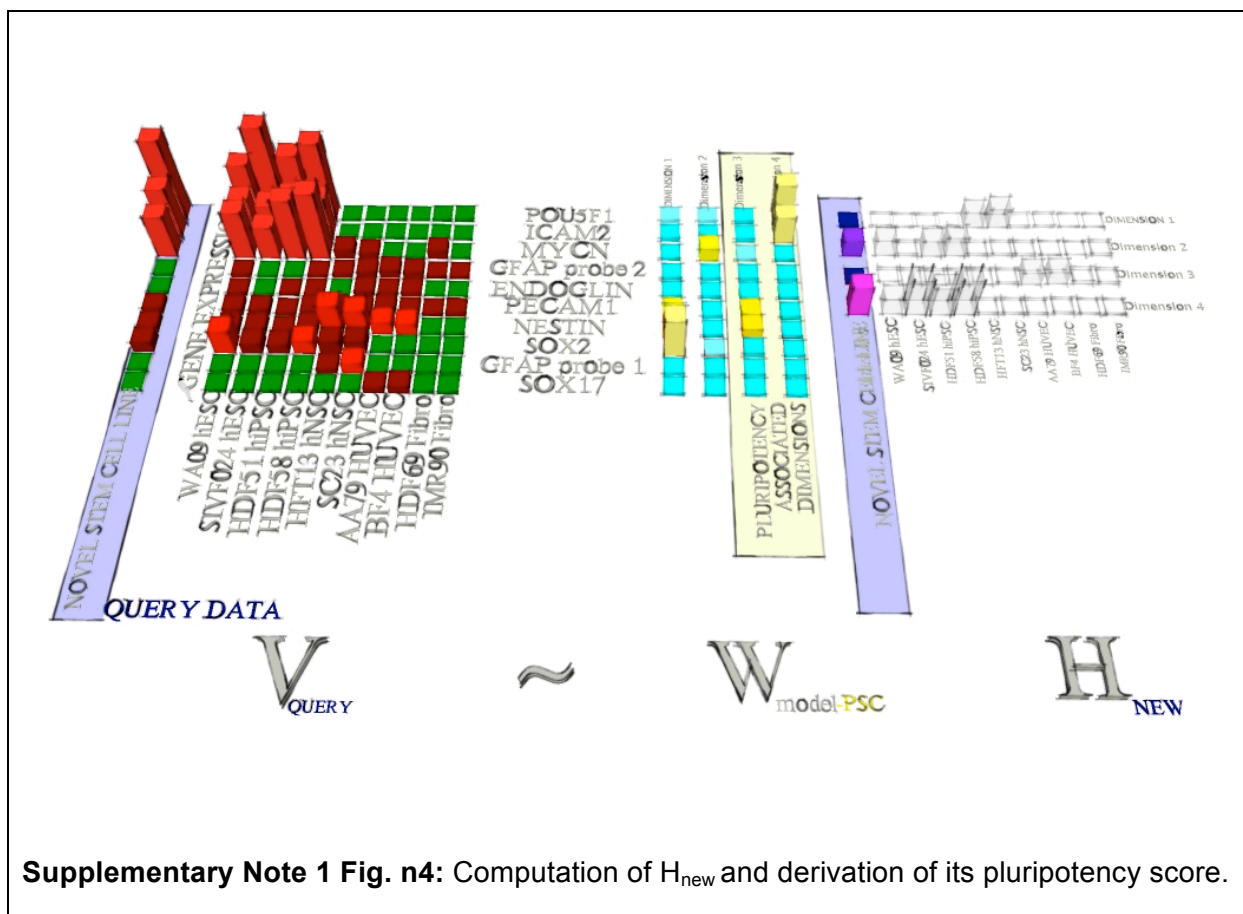
This necessitates an "interpretation" module in PluriTest that first performs data extraction, normalization and filtering steps following published microarray techniques to fit the query data

set into the PluriTest framework. Once preprocessing the query data set is accomplished, the query sample is added as novel sample entry to V_{model} (now termed V_{QUERY} , see also **Supplementary Note 1 Fig. n4**).

Based on V_{QUERY} and W_{MODEL} it is computationally relatively easy to generate H_{model} with the Lee and Seung algorithm.⁶ H_{model} from the query sample is used to compute the *Pluripotency Score*, which gives a relative measure of how much of a pluripotency signal (as defined by $W_{\text{MODEL-PSC}}$) can be detected in the query sample. The cut-off for the resulting *Pluripotency Score* is then calibrated by a test dataset [http://en.wikipedia.org/wiki/Test_set]. Data from known pluripotent cells are used to define what cut-off should be chosen to identify non-pluripotent cells.

In cancer diagnostics using microarrays, datasets are usually divided up into a training dataset (such as the Stem Cell Matrix) and a dataset from which we also know the true nature of the samples and can infer where to set thresholds to reliably differentiate one cell type from the other and determine the sensitivity or specificity of the assay.

[http://en.wikipedia.org/wiki/Receiver_operating_characteristic]. The procedure outlined so far generates a summary statistics term, which indicates whether or not a pluripotency signal is present in a cell preparation.



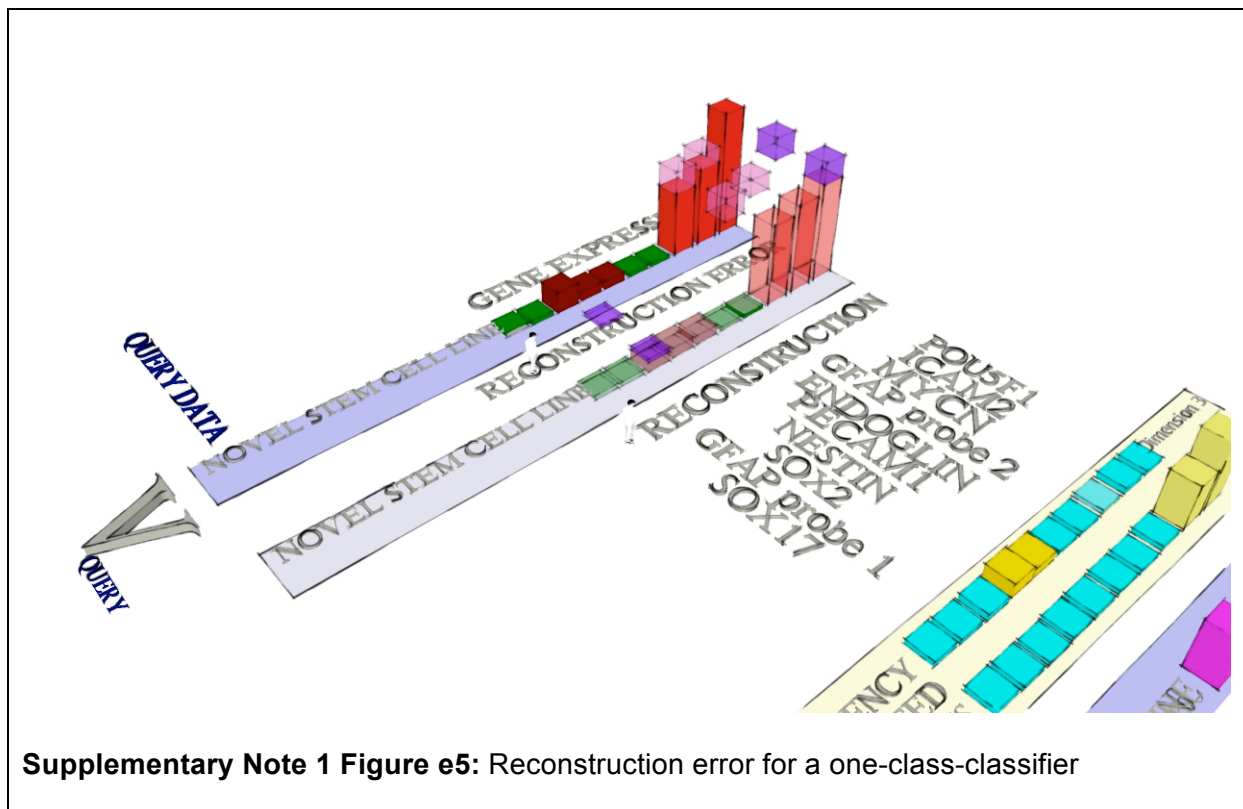
Supplementary Note 1 Fig. n4: Computation of H_{new} and derivation of its pluripotency score.

4. Model generation Novelty Score

It is important to know if there are other, yet unidentified, underlying patterns that should be detected and monitored. One scenario could be that a PSC preparation shows unwanted

differentiation into a specific lineage in a certain culture condition. Another example is a novel PSC type that may have a different transcriptional phenotype compared to known iPSC and hESC lines. Finally, we would like to distinguish genomically abnormal cells and malignant teratocarcinomas cells from normal cells. To detect deviations from a 'pure' and non-malignant PSC phenotype, we devised a second, one-class classifier that allows for an open-ended determination of aberrant transcriptional signals (**Supplementary Note 1 Fig. n5**). The challenge is to detect signals that cannot be anticipated or modeled with our current dataset or knowledge of pluripotency.

To solve this problem, we built a complete NMF model of all highly qualified PSC samples in the Stem Cell Matrix 2 and developed an algorithm that detects deviations from this model. Since all NMF decompositions are optimized to reconstruct an empirical data set, we can use the patterns engrained in the training dataset-derived submatrices to attempt to reconstruct a query sample. If the sample has qualities similar to the samples in the training dataset, we should be able to reconstruct it well with our NMF model. If the sample varies considerably



from the training set model, it should have “reconstruction error,” which gives rise to a difference in the *Novelty Score*.

Toward this end only the normal PSC samples in SCM2 were considered in generating a second non-negative matrix factorization to construct $W_{\text{MODEL-ONE}}$ and $H_{\text{MODEL-ONE}}$. When we reconstruct a query sample based on our model, we measure the reconstruction error and calibrate this summary statistics with the above-described training datasets.

5. Testing an query sample with the Novelty Score

As with the *Pluripotency Score*, the query sample is added to V_{MODEL} and $H_{\text{NEW-ONE}}$ was computed based on $W_{\text{MODEL-ONE}}$. From this knowledge – $H_{\text{NEW-ONE}}$ and $W_{\text{MODEL-ONE}}$ – the query sample is reconstructed and the reconstruction error is computed.

If a novel PSC sample is similar to the samples in the model, the reconstruction error will be small. Should a sample contain signatures that cannot be reconstructed by our model derived from highly qualified, normal PSC in the Stem Cell Matrix, the reconstruction error will be high. This reconstruction error is translated in PluriTest as the *Novelty Score*. In **Fig. 2** we show the Pluripotency and Novelty Scores generated from a time course experiment showing changes in PSC cultures at different stages of directed differentiation along a neuronal lineage. We also show the effects on the *Pluripotency* and *Novelty Scores* of mixing mRNA from undifferentiated and differentiated hESCs in various proportions.

Supplementary Note 2: Usage PluriTest online

I) Intended Audience

This is a description of the PluriTest online application to enable scientists to upload and analyze raw microarray data with the PluriTest assay.

II) Overview

This Note is a 'handbook' describing the use of PluriTest online (v1) as it is being released with the publication of the manuscript in Nature Methods.

The website has a two page structure. On the first page you upload your microarray data, the second page gives you access and displays your PluriTest results. This structure is combined with a secure login process and gives you access to all of your results, not only the ones from your most recent analysis. PluriTest online is designed to analyze up to twelve arrays at once. We recommend, to have all samples in one analysis run on a single HT12 chip (twelve samples in parallel) to minimize batch effects.

The PluriTest online website can be accessed under the address

www.pluritest.org.

Currently, the PluriTest online technology is based on the browser plugin Silverlight4 from Microsoft. If Silverlight for some reason should not be installed your computer, please follow the instructions online to install the plugin and also see Section X: Installing the Silverlight plugin in this Supplementary Note.

Technologies, especially computer, online and nucleotide-based analysis technologies undergo tremendous evolutions within short periods of time. PluriTest may adopt and expand other web technologies (such as HTML5) and other data formats (such as methylation microarrays, next generation sequencing technologies) in the mid-future. We will keep a link to the website in its current form (as of February 07th 2011) to PluriTest online as described in this note for the next 5 years.

We have tested PluriTest online with the Mozilla Fire Fox 3.5 and Internet Explorer 7/8 web browser. We recommend Firefox on all systems. Unfortunately, Apple Safari does not work with the current PluriTest online version.

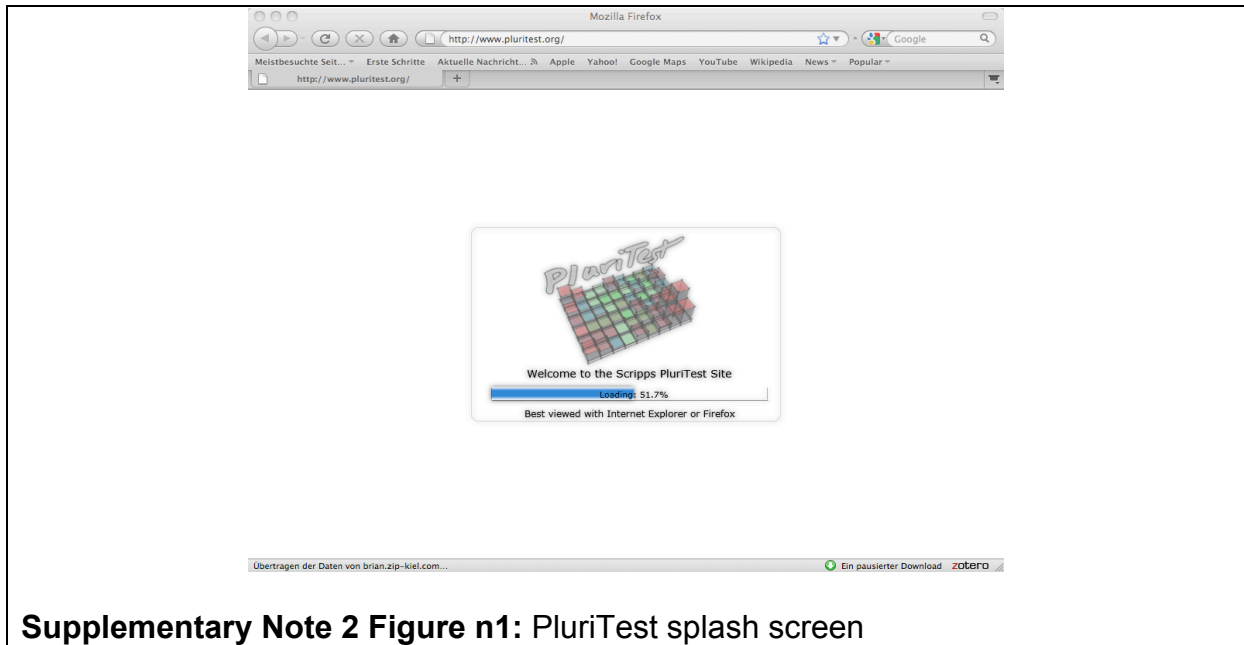
All microarray data that was used for generation of Fig. 2 is deposited on the PluriTest website. The data can be downloaded, independently analyzed and used with PluriTest.

Also, a brief video tutorial can be accessed after the user has logged into PluriTest and an example analysis can be viewed.

III. PluriTest online step-by-step description

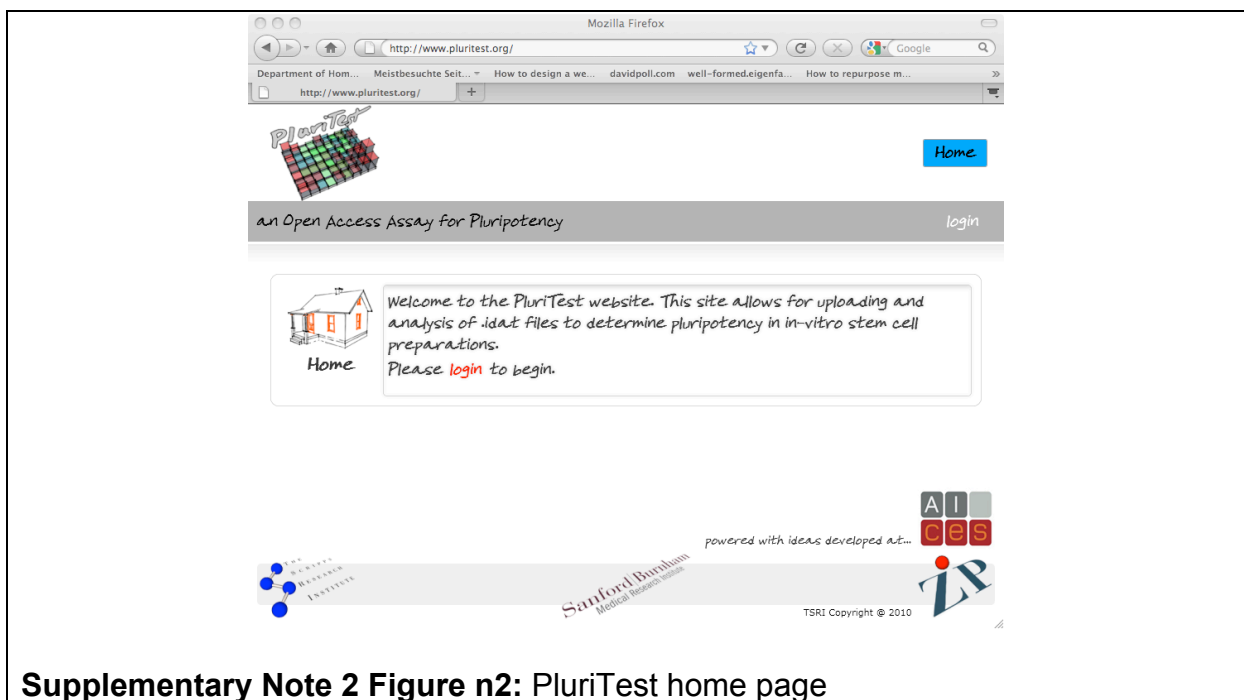
1. Login

After browsing to www.pluritest.org, a brief splash screen will indicate loading of the PluriTest Silverlight application.



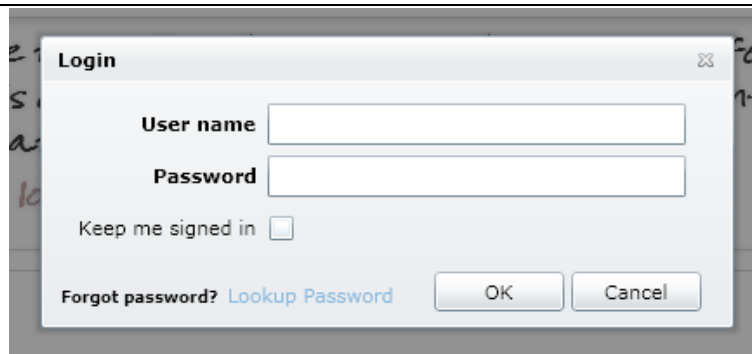
Supplementary Note 2 Figure n1: PluriTest splash screen

The duration of this process will depend on your internet connection. If you don't see this screen, but a message instructing you to install the Silverlight plugin, please follow the online instructions to install the plugin on your computer. After the login screen has loaded, please click on login.



Supplementary Note 2 Figure n2: PluriTest home page

A small window will appear, asking for your login name and your password. If you don't already have these, click on "Sign up for PluriTest" button to create an PluriTest online account.

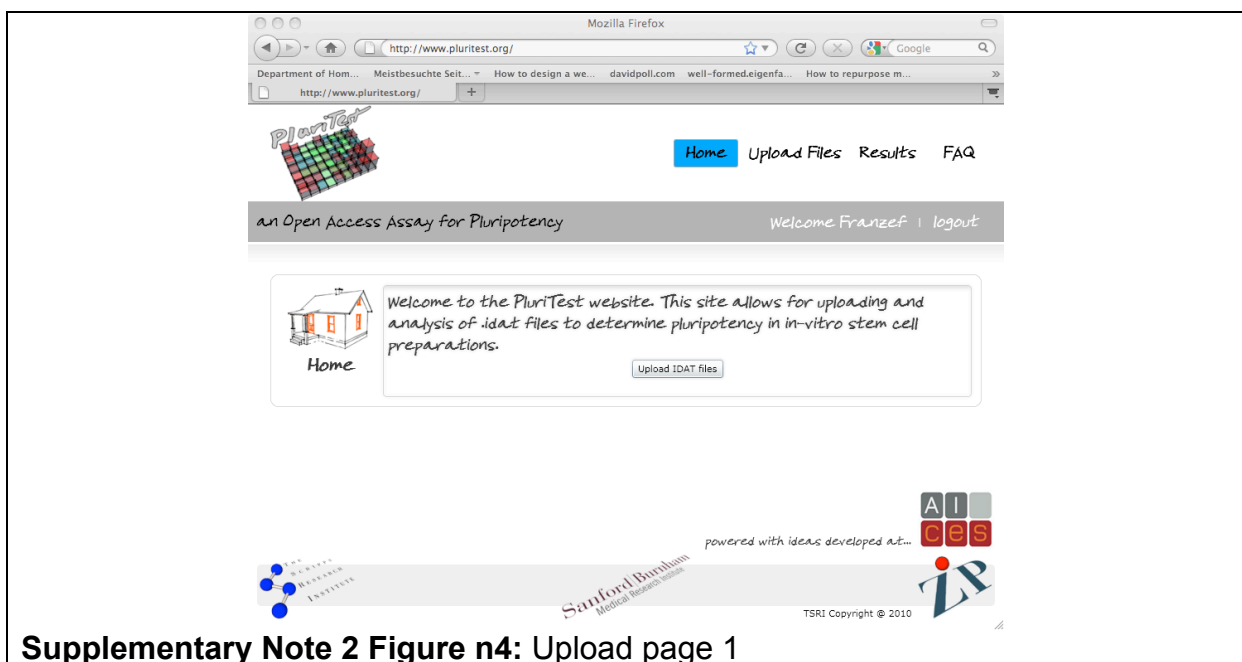


Supplementary Note 2 Figure n3: PluriTest logon

If you forgot your password, click on "Lookup Password" for retrieving your password with the email address that you have specified when you created your account.

2. Upload of .idat files

After you have successfully logged on, the upload screen will appear.

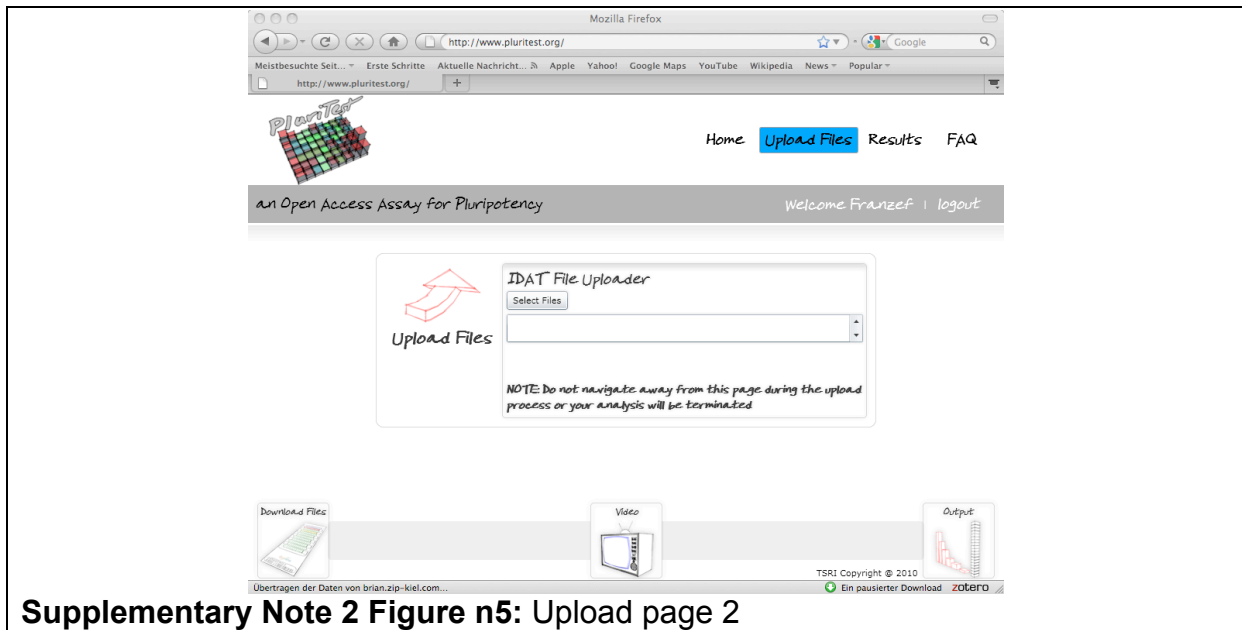


Supplementary Note 2 Figure n4: Upload page 1

Click on "Upload IDAT Files" for browsing for the files you want to analyze in PluriTest.

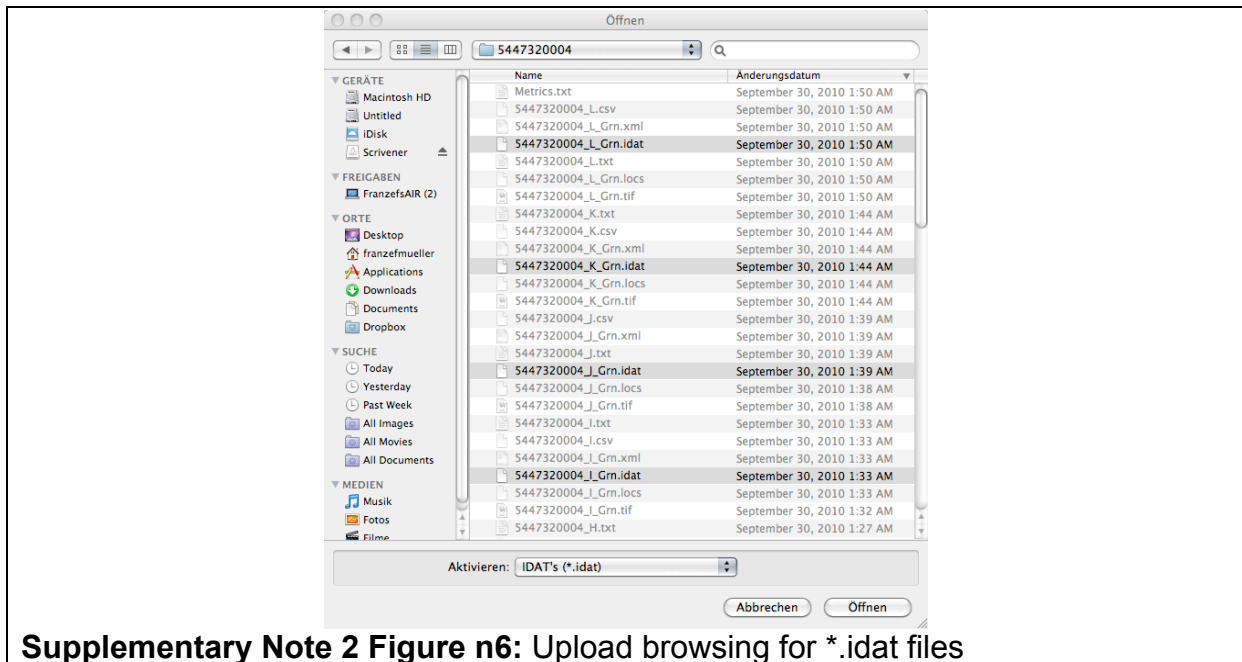
Note: Only the Illumina HT12v3 platform has so far been qualified for actual online testing of novel PSC preparations with a logistic regression model based on our own well-characterized PSC sample collection. While it is in principle possible to analyze Illumina WG6v3 IDAT files on PluriTest, the threshold values for pluripotency have so far not been validated with this array type. We will provide a validated model and threshold settings for the Illumina HT12v4 array

generation in the short term. For more details on the Illumina *.idat file format, see the paragraph 5. File formats.



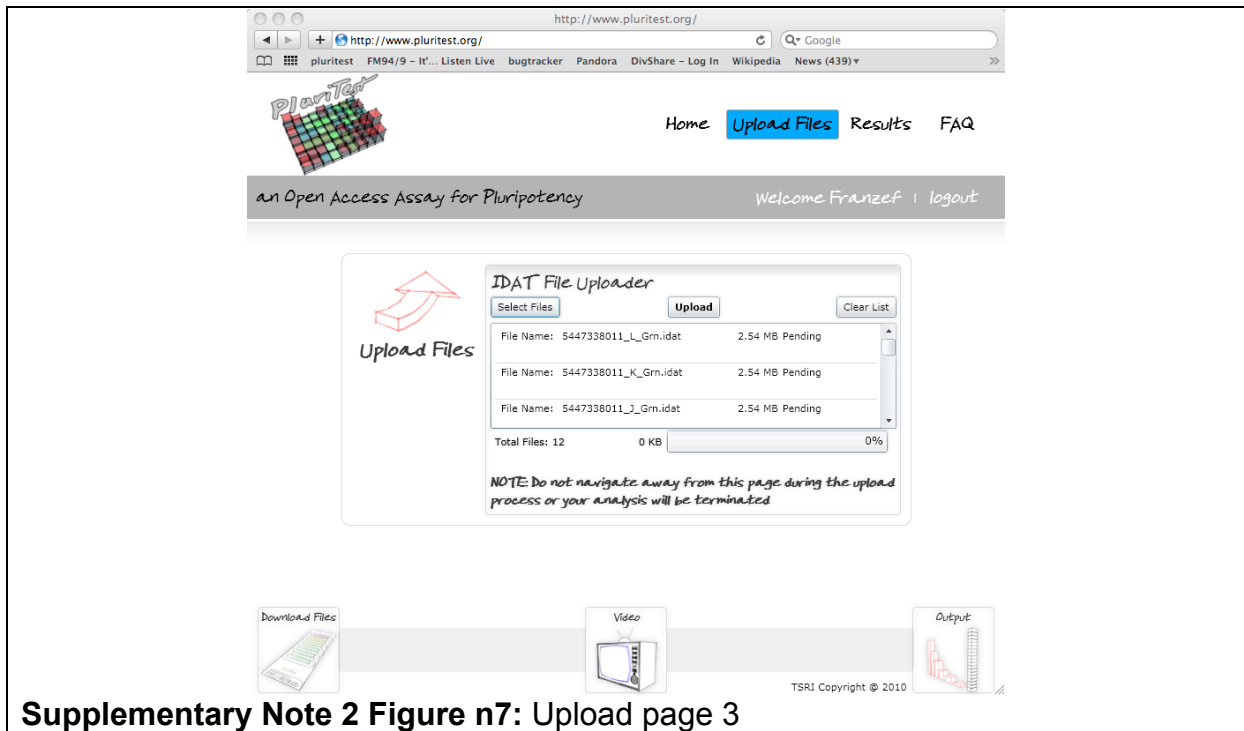
Supplementary Note 2 Figure n5: Upload page 2

On the IDAT file uploader screen, click on the “select files” button and an window will open that allows you to browse in the file system of your computer for IDAT files.



Supplementary Note 2 Figure n6: Upload browsing for *.idat files

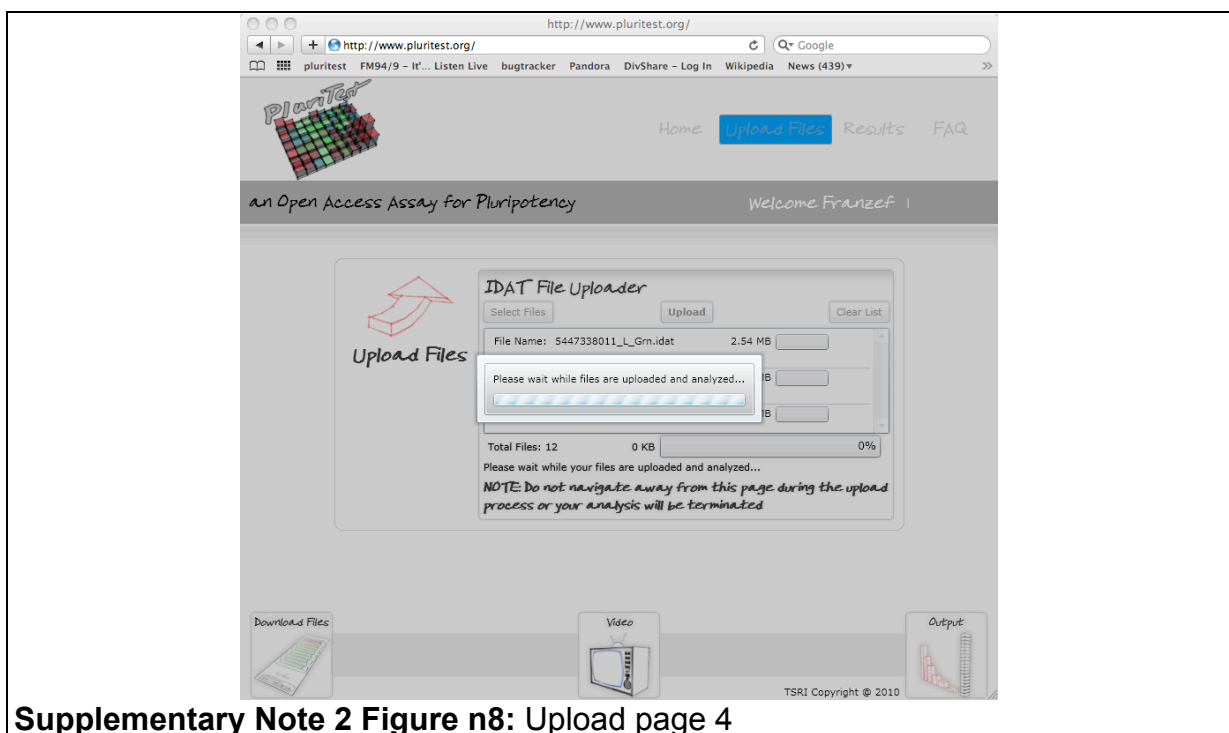
Please note that currently a minimum of three and a maximum of twelve IDATS can be uploaded in one analysis run. The raw data produced by an Illumina array scanner is contained in a folder that is identified by the barcode number of the specific Illumina HT12 array. When you open the folder, several different files and file types are contained in the file folder. Only the files with an .idat ending can be selected in the browser window. These files are the raw data IDATs from your samples. Select all IDAT that you wish to analyze and click on open. After a few seconds, the IDAT files will appear in the upload page.



Supplementary Note 2 Figure n7: Upload page 3

Note: Depending on your Internet connection and the work-load of the PluriTest server, this can take a few moments.

Then, click on the “Upload” Button and a progress bar will inform you on the upload and processing status of your files. The upload time will depend on your internet connection and work load of the PluriTest server and usually takes between two and fifteen minutes for twelve IDATS.



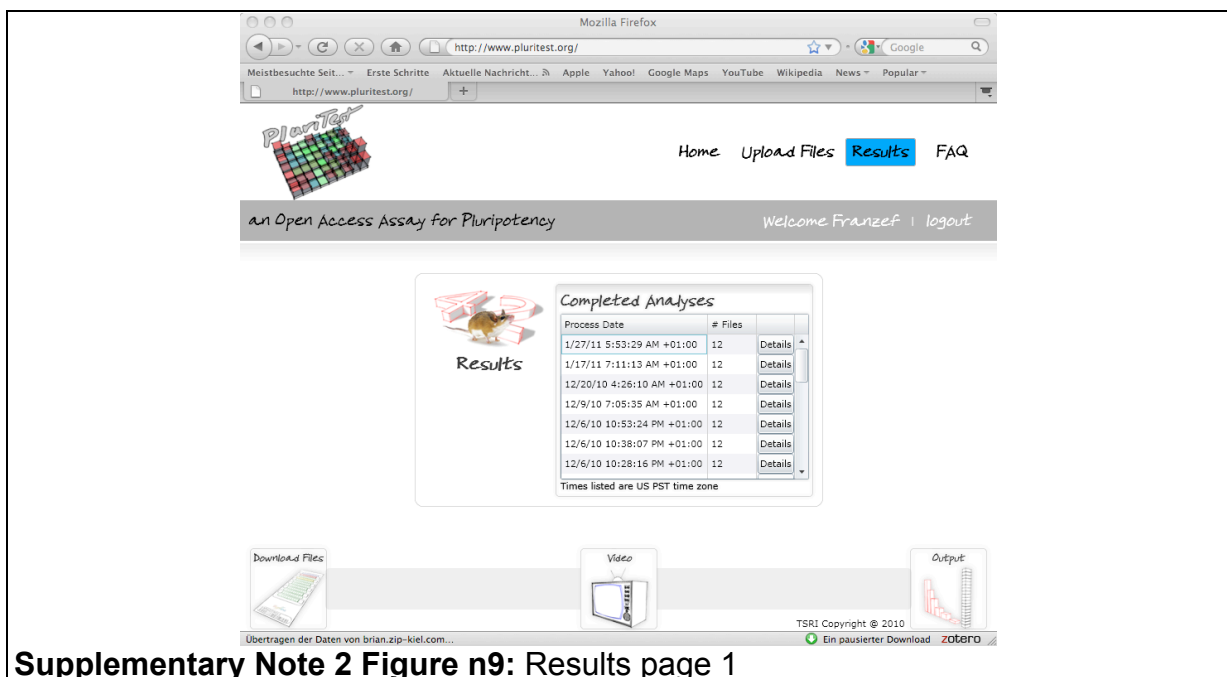
Supplementary Note 2 Figure n8: Upload page 4

Note: Do not navigate away from the site during the upload and processing of your raw files, as your analysis will be terminated.

Tip: Currently no sample tracking or annotation function is implemented in PluriTest. Usually, IDATS are labelled with a 10 digit number followed by an letter and look like 1234567890_K_Grn.idat and you will have to keep track on the identity of your samples. In your local computer file system you can rename the numbers to a clear name, like “hiPSC_retro_replicat1.idat”, which then will appear on the PluriTest readouts instead of the barcode number.

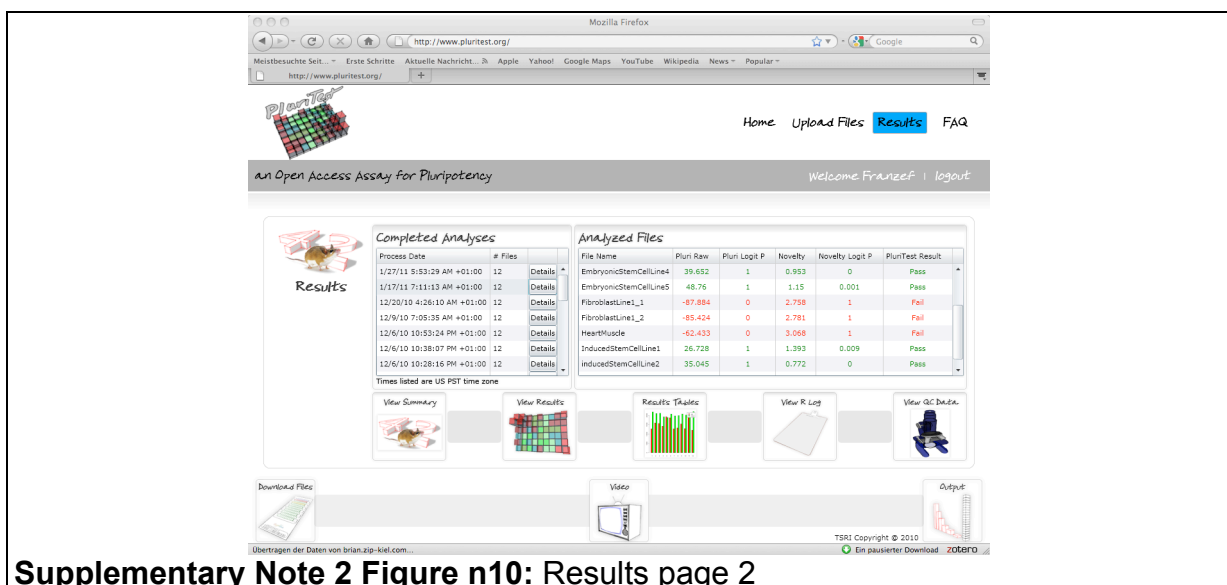
3. Results

In the results screen all of your previous analyses are stored and accessible through an scrollable list with the uppermost being your latest analysis.



Supplementary Note 2 Figure n9: Results page 1

Click on on the 'Details' button next to the analysis you want to inspect.



Supplementary Note 2 Figure n10: Results page 2

A second scrollable box will open that reports the Pluripotency and Novelty scores for each of your samples in your latest analysis. Based on the empirically set thresholds and the computed scores and likelihoods, we indicate the PluriTest results with the terms 'pass', 'further evaluate' and 'fail'.

We label samples with 'pass' if both thresholds, Pluripotency Score > 20 and Novelty Score < 1.67 are passed.

We label samples with 'further evaluate' if either one or the other thresholds, Pluripotency Score > 10 and Novelty Score < 1.67 are passed. These results indicate, that a cell preparation may show pluripotent features, but should be further evaluated for factors that could explain pervasive differences, such as technical problems with the array preparation, genomic abnormalities, contaminating cells, unwanted differentiation or epigenetic memory in the cells of interest.

We label samples with 'fail' if both of the thresholds Pluripotency Score > 10 and Novelty Score < 1.67 are not passed. The interpretation of this results depends on the prior knowledge on the sample.

If the samples was supposed to be an pluripotent sample, the preparation be further evaluated for factors that could explain pervasive differences, such as technical problems with the array preparation, genomic abnormalities, contaminating cells, unwanted differentiation or epigenetic memory in the cells of interest.

If the cell preparation in question is a differentiated sample, we would expect it to cluster with the other somatic samples (see view results section). Currently PluriTest is not designed to detect pluripotent contamination in differentiated samples, but it is planned to incorporate this function in future releases.

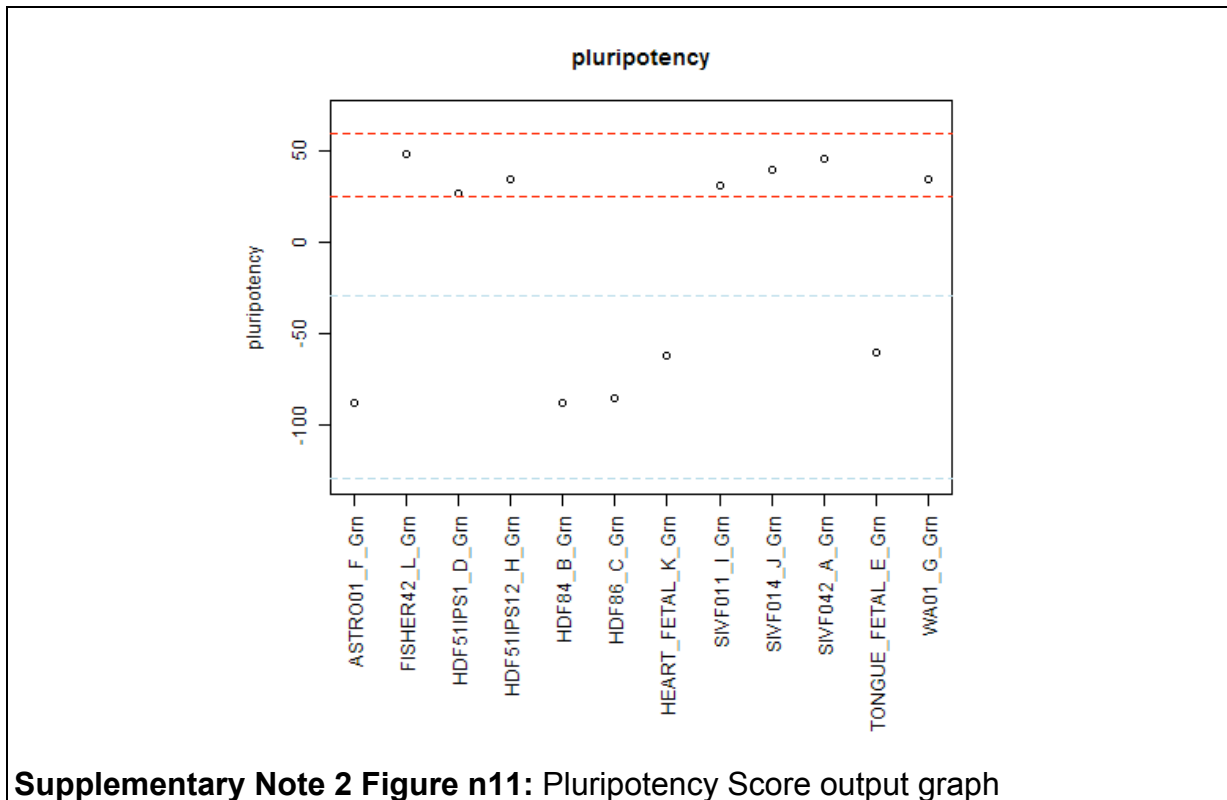
With the five icons below the both tables, more detailed information can be accessed. The most detailed PluriTest analysis is contained in the "View Results" section, while "Results Tables" offer .csv tables of Pluripotency and Novelty scores for all samples, "View R log" gives information on the processing of the sampels and "View QC Data" gives results from hierarchical clustering and box plots from all probes on each array.

The web page that opens is a plain html page and all text and graphs can be easily copy/pasted from this page into e.g. a word editor.

We have determined the thresholds for Pluripotency Score and Novelty Score following the Receiver Operator Curve/Area Under the Curve (ROC/AUC) approach based on our training dataset. These currently 'hard coded' cut offs may be handled more flexible in future versions of the assay, where users may be able to choose their cutoffs based on probabilities based on prior information.

In the following we demonstrate all online PluriTest outputs with example data that was also used for generation of **Fig. 2 a**

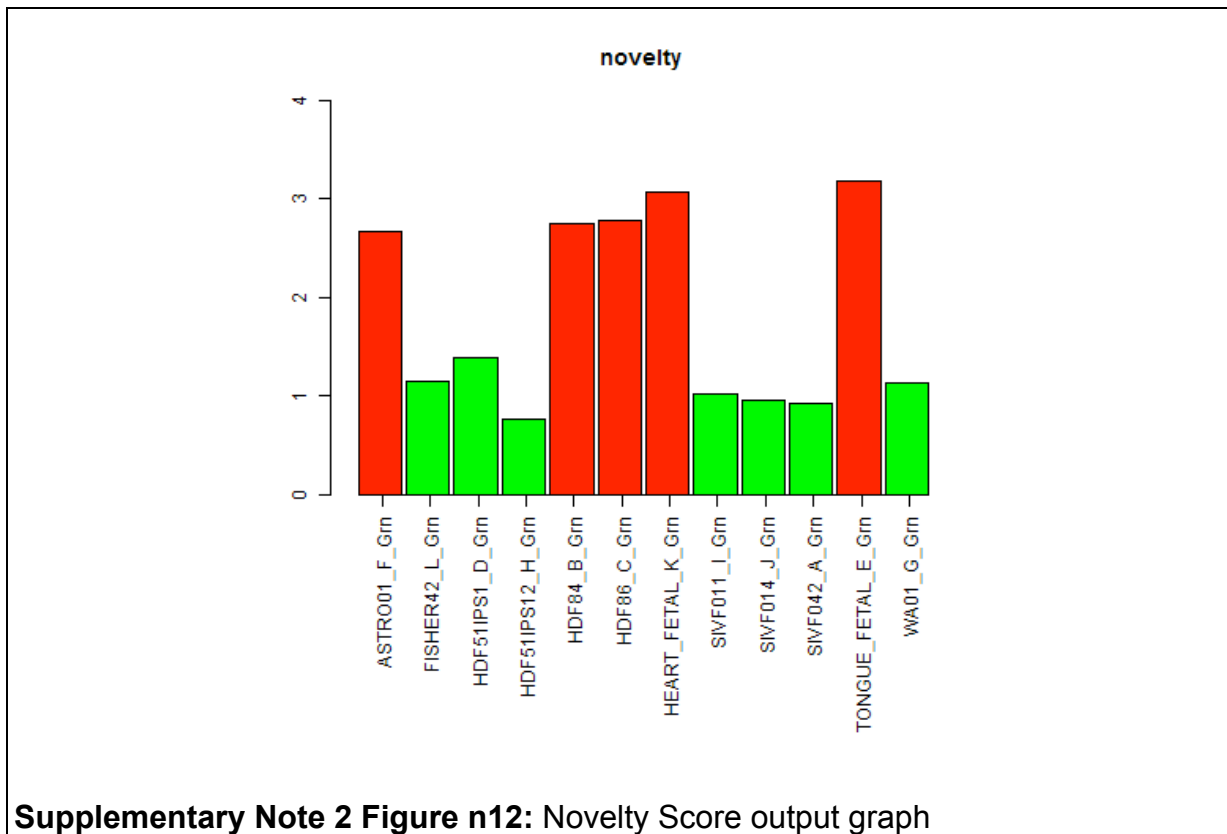
3.1 Model-based multi-class Pluripotency Score



Supplementary Note 2 Figure n11: Pluripotency Score output graph

Pluripotency Score: A score that is based on all samples (pluripotent cells, somatic cells and tissues) in the stem cell model matrix. Samples with positive values are more similar to the pluripotent samples in the model matrix than to all other classes of samples in the matrix. The area between the red lines indicates the range that contains approximately 95 percent of the pluripotent samples tested. The Pluripotency Score gives an indication if a sample contains a pluripotent signature, but not necessarily if the cell preparation is a normal, bona-fide hESC or iPSC. Partially differentiated pluripotent cells, teratocarcinoma cells or karyotypically abnormal embryonic stem cells may also have a high Pluripotency Score. The blue lines indicate those scores that we have observed in approximately 95 percent of the non-pluripotent samples.

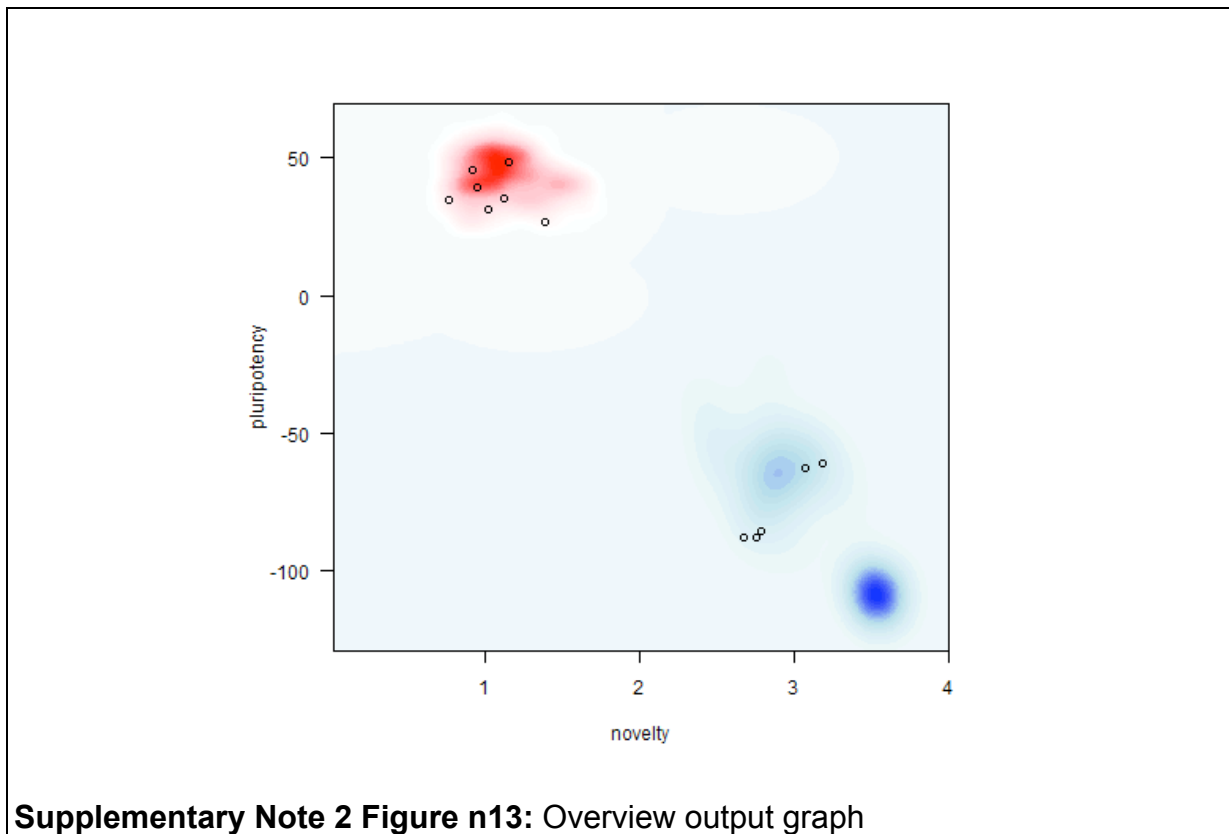
3.2 Novelty Score



Supplementary Note 2 Figure n12: Novelty Score output graph

Novelty Score: A score that is based on well-characterized pluripotent samples in the stem cell model matrix. Samples are color-coded green (pluripotent), orange, red (not-pluripotent) based on the probabilities given from the logistic regression model. Orange and red samples are more dissimilar to the pluripotent samples in the model matrix than the other pluripotent samples in the matrix. A low Novelty Score indicates that the test sample can be well reconstructed based on existing data from other well-characterized iPSC and ESC lines. A high Novelty Score indicates that there are patterns in the tested sample that cannot be explained by the currently existing data from well-characterized, karyotypic normal pluripotent stem cells. Partially differentiated pluripotent cells, teratocarcinoma cells or karyotypically abnormal embryonic stem cells may have a high pluripotency score but cannot be reconstructed well with data from well-characterized, normal pluripotent stem cells and thus are expected to have a high Novelty Score.

3.4 Overview plot



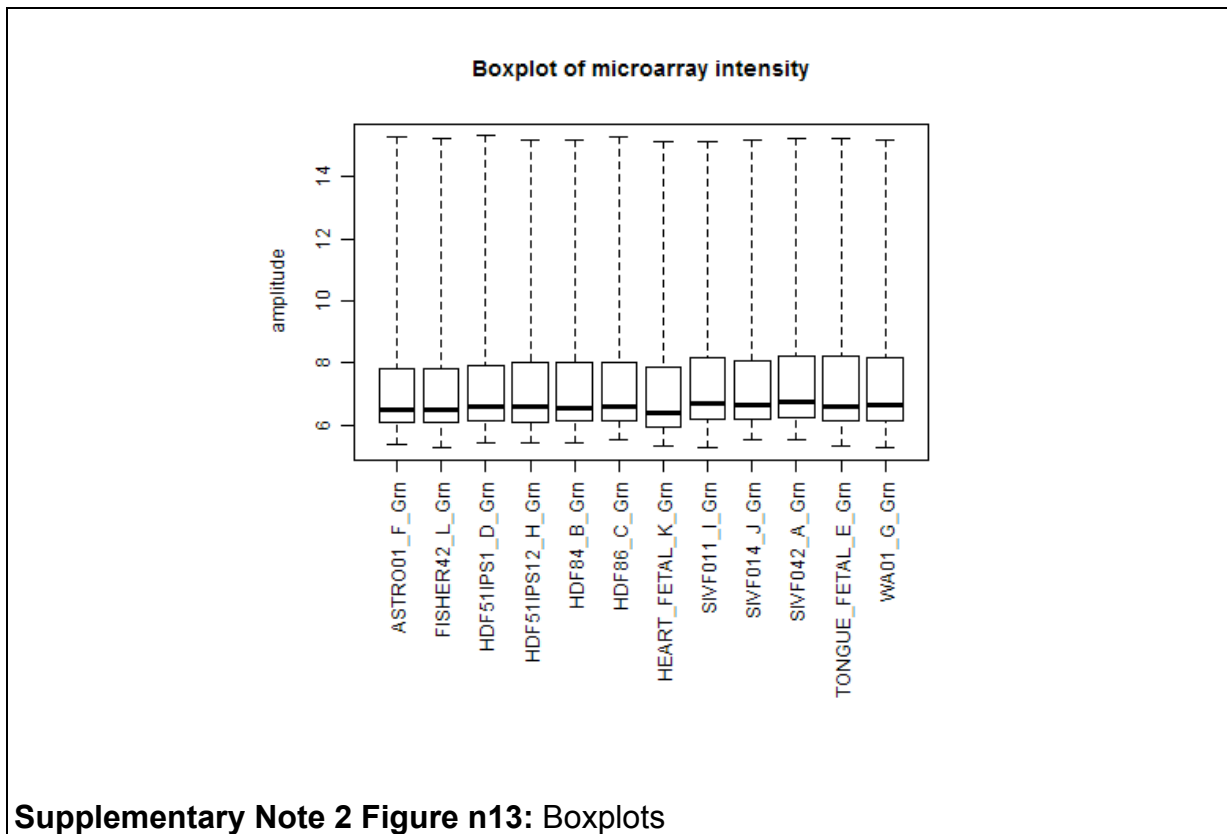
Combines the Pluripotency Score on the y-axis with the Novelty Score on the x-axis. The red and blue background hints to the empirical distribution of the pluripotent (red) and non-pluripotent samples (blue) in our own test data set.

3.5 Results table

	pluri-raw	pluri logit-p	nov elty	novelty logit-p	RM SD
ASTRO01_F_Grn	-87.97	0.00	2.68	1.00	0.76
FISHER42_L_Grn	48.76	1.00	1.15	0.00	0.31
HDF51IPS1_D_Grn	26.73	1.00	1.39	0.01	0.18
HDF51IPS12_H_Grn	35.05	1.00	0.77	0.00	0.16
HDF84_B_Grn	-87.88	0.00	2.76	1.00	0.81
HDF86_C_Grn	-85.42	0.00	2.78	1.00	0.78
HEART_FETAL_K_Grn	-62.43	0.00	3.07	1.00	0.90
SIVF011_I_Grn	31.20	1.00	1.02	0.00	0.21
SIVF014_J_Grn	39.65	1.00	0.95	0.00	0.19
SIVF042_A_Grn	45.66	1.00	0.92	0.00	0.17
TONGUE_FETAL_E_Grn	-60.64	0.00	3.18	1.00	0.84
WA01_G_Grn	35.21	1.00	1.13	0.00	0.19

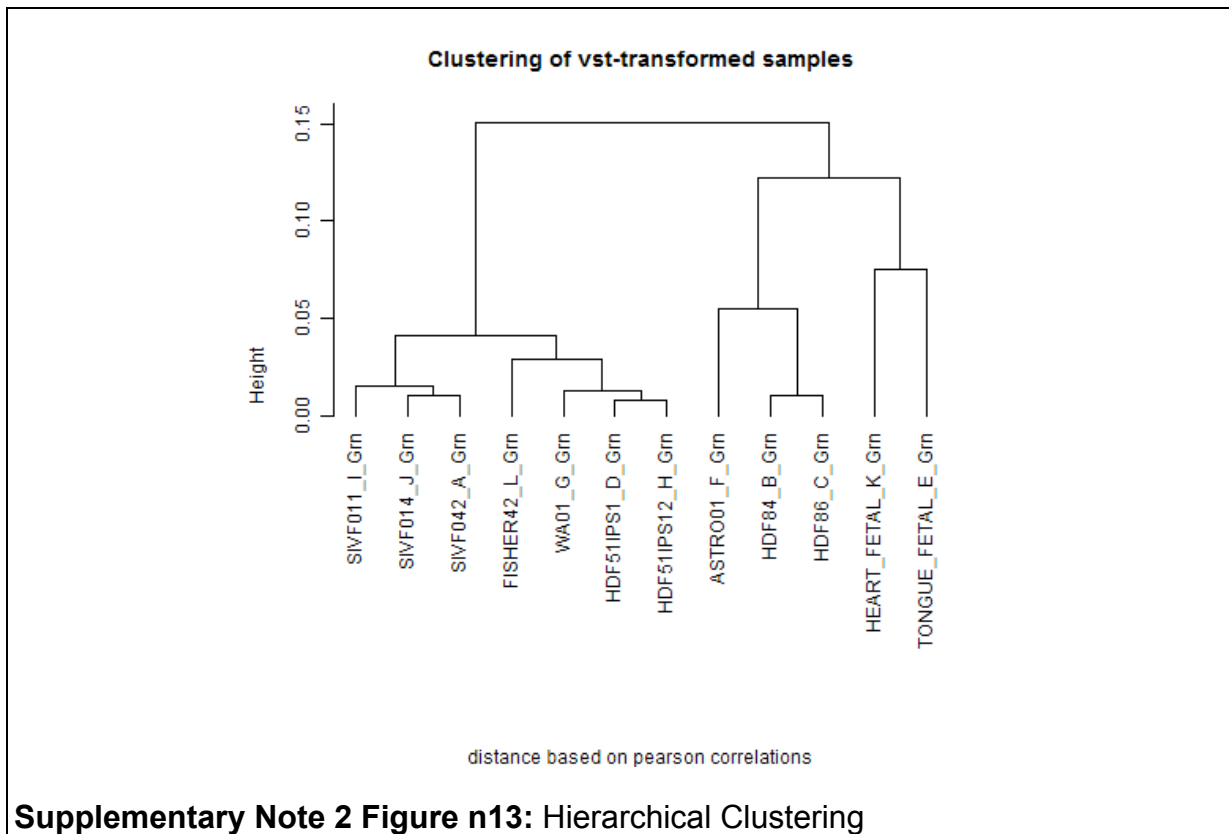
Supplementary Note 2 Table n1: PluriTest analysis values

3.6 Quality control: boxplots



A plot generated by the Lumi package after the samples were transformed with a variance stabilizing transformation (VST) and before robust spline normalization (RSN). Outlier arrays with too much technical variation might be spotted if they show a different probe intensity distribution pattern in the box-plots when compared to the other arrays on the same chip or when compared to arrays on other chips. For details please see Du, P., W.A. Kibbe, and S.M. Lin, lumi: a pipeline for processing Illumina microarray. *Bioinformatics*, 2008. 24(13): p. 1547-8.

3.7 Quality control: Hierarchical Clustering

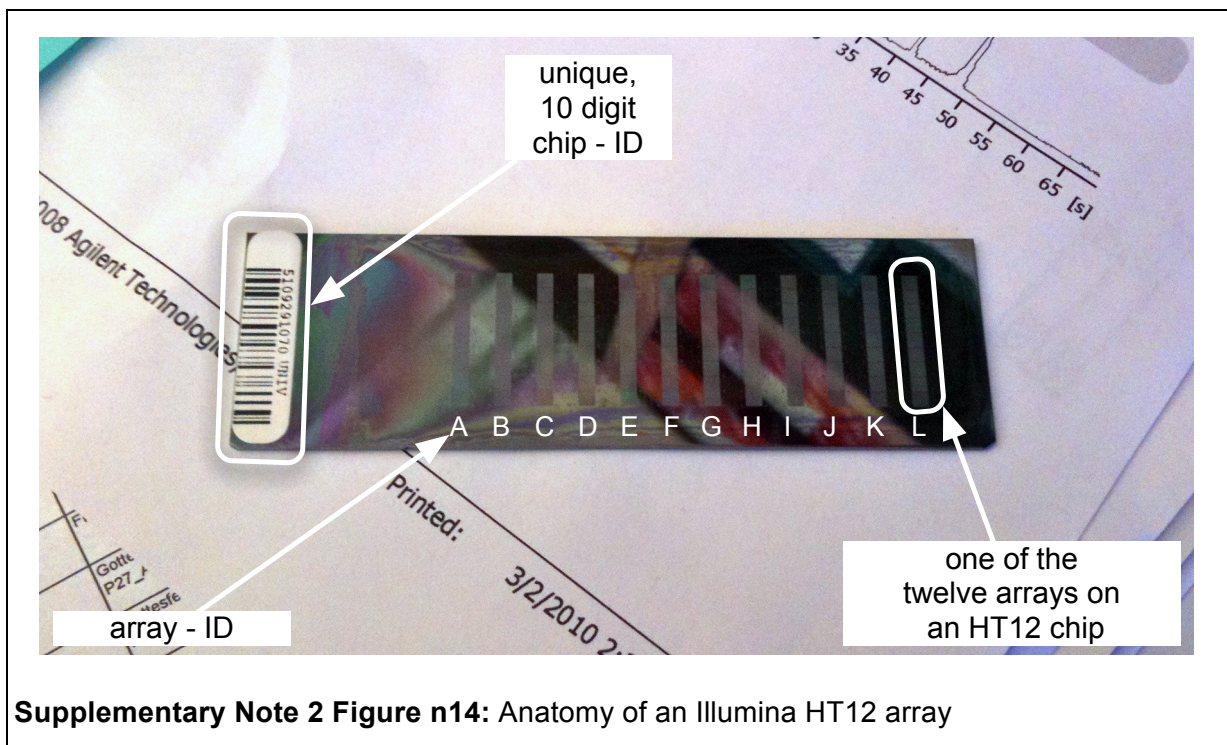


A plot generated by the Lumi package after the samples were transformed with a variance stabilizing transformation (VST) and before robust spline normalization (RSN). Outlier arrays with too much technical variation might be spotted if they do not cluster with their respective technical or biological replicates from the same sample or sample type. For example, if an array hybridized with the RNA from an pluripotent cell clusters with fibroblasts on the same chip, but not with other pluripotent samples, something might be wrong with technical aspects of your experiment. For details please see Du, P., W.A. Kibbe, and S.M. Lin, lumi: a pipeline for processing Illumina microarray. *Bioinformatics*, 2008. 24(13): p. 1547-8 and the vignette to the lumi package on the bioconductor website.

4. IDAT file formats

The *.idat file format is a proprietary file format created by Illumina, similar to the proprietary .CEL-file format from Affymetrix. These files contain raw, gene probe specific fluorescent intensity data read from a hybridized array chip in an Illumina array scanner. These files can only be read using Illumina's proprietary data extraction and analysis software, BeadStudio / GenomeStudio. This software translates the data into a ".txt" format that can be easily read and manipulated for compatibility with most available data analysis programs. In most cases, microarray core facilities use GenomeStudio to extract data from .idat files to provide their customers with .txt files that can be more broadly manipulated. We have incorporated the the .idat-reader module from Illumina into PluriTest. The .idat file is completely raw and unmodified data.

For each array on an HT12 chip, one unique .idat file is generated. An HT12 chip is a glass slide with similar dimensions as an microscope glass slide (see **Supplementary Note 2 Fig. n14**). Each chip does have a unique 10 digit ID number (e.g 449359422).



Supplementary Note 2 Figure n14: Anatomy of an Illumina HT12 array

It has 12 areas etched into the glass surface, each is a microarray. Each of the arrays can be hybridized with an different sample, hence the name H(igh) T(hroughput) 12(arrays). Each array is treated separately with biological material, yet all the hybridization steps, incubation etc. are done on all 12 arrays in parallel. This design is unlike the "classical" Affymetrix-system, where each GeneChip cartridge contains one array and more like the recently introduced 24- or 96- Affymetrix array plates.

Each slot/array on a chip is identified by an letter from A to L. Thus the unique ID for each

array consists of the 10 digit chip ID plus the letter denoting the array/slot on the HT12 array
(e.g. 449359422E)

Supplementary Note 3: Usage PluriTest R/Bioconductor workspace

I. Intended Audience

This is a description of the PluriTest R/Bioconductor workspace to enable scientists proficient in R/Bioconductor to run PluriTest locally on their machine. This document is not an introduction to R/Bioconductor. For this purpose, please refer to www.bioconductor.org and associated resources. If you are having trouble following this description, please contact us (schuldt@aices.rwth-aachen.de, fj.mueller@zipkiel.de) so we can update this description. All required packages will be loaded with the PluriTest script in the workspace. Required packages are: lumi, xtable, Go.db. This description assumes that a NMF- factorization

$$V \approx \tilde{V} = W \times H$$

has been calculated as described in the **Online Methods**. The matrices and coefficients needed to reproduce the reported prediction results are available in the R workspace.

II. Overview

The pluritest prediction workflow is divided in three parts

- (1) Array specific normalization and preprocessing
- (2) Calculation of the pluripotency and novelty scores
- (3) Generation of plots and output files for display

III. PluriTest R/Bioconductor dezailed description

1. Array specific normalization

1.1. Compatible arrays.

For compatible arrays-types running

```
pluritest(NewDataFileName,wd)
```

from within the workspace is the preferred method. This command writes all the results in the directory wd. If intermediated results are needed for further analysis a customized version pluritest can be created with the usual R commands, e.g.

```
pluritest.custom<-edit(pluritest)
```

This method can be used if the array output file is available in non-normalized file format readable by lumiR() and the file provides Probelds compatible with the Illumina Ids of the HT12v3 arrays. This should always be the case when starting with Beadstudio exports of HT12v3, HT12v4 and WG6v3 arrays.

The pluritest script uses the lumi package for normalization. Sample data ist vst-Transformed and rsn-normalized with a target array. For quality control, boxplots and hierarchical clustering plots are saved for later analysis (See **Supplementary Note Fig. n1** and **Supplementary Note Fig. n2**).

1.2. Other array-types.

Pluritest can be used with other array-types as well. The main step of the pluritest-workflow is the calculation of new H-matrices given a data-matrix V and the precomputed W matrices. Therefore the probe ids in the Data-matrix V have to be

matched to the probe ids used in the W matrices.

An array-specific normalization strategy (e.g. RMA for Affymetrix arrays) has to be chosen to make samples comparable.

Pluripotency and novelty scores will in general differ from the ones calculated for compatible arrays. As long as there is a high number of matching probes on the two different arrays this will only result in a shifting and scaling of the pluripotency scores that can be resolved by normalization or re-setting of the thresholds.

1.2.1. Affymetrix.

We have used the quintile pre-normalized data from Lukk et al. Nat Biotechnology 2010 for demonstrating the generalizeability of the Pluripotency/Novelty Score approach. A probe-mapping file is available from:

www.switchtoi.com/probemapping.ilmn

The following script takes a data-matrix `affytest` with `affyIds` in the rownames, replaces them with matching Illumina identifiers and removes duplicates.

```
illuminaoaffy<-read.table("ILMN_HumanWG6v3_AffyU133Plus2.0.txt", sep="\t")
sel=match(rownames(affytest),illuminaoaffy[,3])
sum(is.na(sel))
rownames(affytest)<-illuminaoaffy[sel,2]#
affytest<-affytest[!duplicated(rownames(test)),]
```

The resulting data-matrix `affytest` can then be used in the calculation of the pluripotency and novelty scores as described in section 2.

1.2.2. Other illumina array types.

The Illumina gene expression microarray platform has gone through much iteration in recent years and Probe annotations have changed considerably. For this reason older data sets are normalized internally without reference to a target sample.

Universal NuIDs were used as supplied by the `lumiHumanIDMapping` package. These IDs are based on Probe sequence and were used instead of the Illumina ProbeID as row identifiers for The V and W matrices.

The main step is the calculation of the H matrix given the data matrix V. This is done using the function `predictH()`, which implements the Lee and Seung scheme in the case of a fixed W matrix.

```
sel<-match(rownames(W15),fData(working.lumi)[,1])
# change the matching if your datamatrix is not stored in a lumi workspace
H15.new<-
predictH(exprs(working.lumi[sel,][!is.na(sel),]),W15[!is.na(sel),])
H12.new<-
predictH(exprs(working.lumi[sel,][!is.na(sel),]),W12[!is.na(sel),])
```

For the RMSE and Novelty score the Residuals $V - \tilde{V}$ are calculated and summarized in a single score per sample.

```
rss.new<-apply((exprs(working.lumi[sel,][!is.na(sel),])-
W12[!is.na(sel),]%*%H12.new)^2,2,sum)
RMSE.new<-sqrt(rss.new/sum(!is.na(sel))) novel.new<-
apply((exprs(working.lumi[sel,][!is.na(sel),])-
W12[!is.na(sel),]%*%H12.new)^8,2,sum) novel.new<-
(novel.new/sum(!is.na(sel)))^(1/8)
```

For the pluripotency score

```
#coef<-c(-1.267095e+02 ,4.567437e-03 , 4.377068e-03 ,1.043193e-03)
s.new<-drop(coef[1] +coef[2:4]*%*%H15.new[c(1,14,13),])
```

3. Output generation

The pluritest script generate plots and tables and save these files to the disc for further reference (**Supplementary Note 3 Figs. n3-n5**). The limits for the plots and the colors are set with respect to the Illumina ht12v3 ht12v4 platform and designed to be displayed on the web interface. For other uses it is best to first plot the pluripotency and novelty scores and to set the limits appropriately and to generate customized plots.

```
plot(s.new)
barplot(novel.new)
plot(s.new~novel.new)
```

Appendix 1: PluriTest script

```
pluritest <-function(NewDataFileName,wd) {

# Pluritest - skript for use with the open Access
# tool for classification of pluripotent stem cells
# Copyright (C) 2010 Bernhard Schuldt
# email: schuldt@AICES.rwth-aachen.de

# This program is free software: you can redistribute it and/or modify
# it under the terms of the GNU General Public License as published by
# the Free Software Foundation, either version 3 of the License, or
# (at your option) any later version.
#
# This program is distributed in the hope that it will be useful,
# but WITHOUT ANY WARRANTY; without even the implied warranty of
# MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. See the
# GNU General Public License for more details.
#
# You should have received a copy of the GNU General Public License
# along with this program. If not, see <http://www.gnu.org/licenses/>.
#Use NewDataFileName points to file that can be read by lumiR
# wd is the working directory where graphics and tables are stored
# NewData, workingdirectory for plots
# load packages

require(lumi)
require(xtable)
require(GO.db)

# setwd
setwd(wd)
sink(file="pluritest_log.txt")
##### IMPORT RAW DATA WITH LUMI
##### working.lumi<-lumiR(NewDataFileName,
convertNuID = FALSE, inputAnnotation=FALSE) # fData(working.lumi)[,1]<-
gsub("\",\"",fData(working.lumi)[,1]) ##### BOXPLOT OF
RAW VALUES #####

pdf(file="pluritest_image01.pdf")
plot(working.lumi, what='boxplot')
dev.off()
}
```

```

##### CLUSTERING OF VARIANCE STABILIZED SAMPLES
##### working.lumi<-lumiT(working.lumi)
hc<-hclust(as.dist(1-abs(cor(exprs(working.lumi[, ]))))))
pdf(file="pluritest_image02a.pdf")
plot(hc, hang=-1, main="Clustering of vst-transformed samples",
      sub="distance based on pearson correlations", xlab="")
dev.off()
##### RSN NORMALIZATION OF THE DATA
##### A <- fData(working.lumi)[,1] #matches on
ILMN_Ids for lumi/RSN
B <- fData(H9targetArray)[,1] #for matches on ILMN_Ids for lumi/RSN
sel.match <- match(B,A)
working.lumi <- working.lumi[sel.match,][!is.na(sel.match),]
#subsets the exprSet user.lumi down to subsetuser.lumi to match H9 ref
array rows working.lumi<-
lumiN(working.lumi, method="rsn", target=H9targetArray[is.na(sel.match)==FALSE,])
##### RSN NORMALIZATION OF THE DATA
#####

# assume that illumina Probe Id s are in fData[,1]
A <- fData(working.lumi)[,1]
sel.match<-match(colnames(W15),A)### needs to be changed.
##### calculations
##### try(
{
sel<-match(rownames(W15), fData(working.lumi)[,1])
coef<-c(-1.267095e+02 , 4.567437e-03 , 4.377068e-03 , 1.043193e-03)
H15.new<-
predictH(exprs(working.lumi[sel,][!is.na(sel),]), W15[!is.na(sel),])
H12.new<-
predictH(exprs(working.lumi[sel,][!is.na(sel),]), W12[!is.na(sel),])
rss.new<-apply((exprs(working.lumi[sel,][!is.na(sel),]) -
W12[!is.na(sel),] %*% H12.new)^2, 2, sum)
RMSE.new<-sqrt(rss.new/sum(!is.na(sel)))
novel.new<-apply((exprs(working.lumi[sel,][!is.na(sel),]) -
W12[!is.na(sel),] %*% H12.new)^8, 2, sum)
novel.new<-(novel.new/sum(!is.na(sel)))^(1/8)
s.new<-drop(coef[1] +coef[2:4] %*% H15.new[c(1,14,13),]) print(s.new)
}
)
##### plot MULTICLASS PLURITEST & overview
#####

table.results<-matrix(, nrow=ncol(exprs(working.lumi)), ncol=5)
rownames(table.results)<-colnames(exprs(working.lumi))
colnames(table.results)<-c("pluri-raw", "pluri logit-p", "novelty", "novelty
logit-p", "RMSD")
try(
{
print(s.new)
pdf(file="pluritest_image02.pdf")
par(mar=c(12, 4, 4, 2))
par(xaxt='n')
plot(s.new, main="pluripotency", xlab="", ylab="pluripotency", ylim=c(-130, 70))
abline(h=25.25, lty="dashed", col="red")
abline(h=59.95, lty="dashed", col="red")
abline(h=-28.92, lty="dashed", col="lightblue")
abline(h=-130, lty="dashed", col="lightblue")
par(xaxt='s')
axis(1, at=c(1:length(s.new)), labels=names(s.new), las=2)
dev.off()
}
)

```

```

)

table.results[,1]<-round(s.new,3)
table.results[,2]<-round(exp(s.new)/(1+exp(s.new)),3)
table.results[,3]<-round(novel.new,3)
table.results[,5]<-round(RMSE.new,3)

try(
{
pdf(file="pluritest_image03.pdf")
color.palette =
colorRampPalette(c("red","pink1","aliceblue","lightblue","blue"), bias=1)
filled.contour2(y=c(-129:70),x=c((1:200)/50)
,background129_70x1_4,col=colram(50),nlevels=35,xlab="novelty",ylab="
pluripotency")
points(s.new~novel.new,cex=.4,main="Overview")

dev.off()
}

)
try(
{

palette(colorRampPalette(c("green", "orange","orange","orange", "red"))(5))
df.novelty.new<-data.frame(novelty= novel.new)
pdf(file="pluritest_image03c.pdf")
par(mar=c(12,4,4,2))
par(xaxt='n')
barplot(novel.new,
col=pmin(5,10*predict(logit.novelty,type="response",newdata=
df.novelty.new)+1),
names.arg=c(1:length(novel.new)),xlab="",xlim=c(0,length(novel.new)),
width=.9,space=(1/9),ylim=c(0,4))
title(main="novelty")
par(xaxt='s')
axis(1,at=c(1:nrow(table.results))-4,labels=names(s.new),las=2)
table.results[,4]<-round(predict(logit.novelty,type="response",newdata=
df.novelty.new),3)
dev.off()

}
)

##### Save CSV FILE for TABLE
##### table.results[,5]<-round(RMSE.new,3)
write.csv(table.results,file="pluritest.csv")
sink()
}

```

Appendix 2: H matrix prediction script

The following function runs the Lee & Seung multiplicative update for a maximum of 2000 iterations, checking convergence every 20 steps.

```

predictH <-function(V,W){

H<-matrix(1,nrow=ncol(W),ncol=ncol(V))
tWW<-t(W)%*%W
tWVnew<-t(W)%*%V
for(j in 1:100){
for(i in 1:20){ Hold<-H

```

```

        H<- H*(tWVnew/(tWW%*%H))
    }
    if(sum((Hold-H)^2)<10e-4){
        return(H)
    }
}
return(H)
}
}

```

Appendix 2: Colored density distribution computation

```

density1<-
kde2d(p8wmd[colorcode=="red"],pls[colorcode=="red"],n=200,lims=c(0,4,-
130,70))

density2<-
kde2d(p8wmd[colorcode=="lightblue"],pls[colorcode=="lightblue"],n=200,lims=
c(0,4,-130,70))

colorRampPalette(c("red","pink","white","lightblue","blue"), bias=.9)-
>colram

filled.contour2(y=c(-129:70),x=c((1:200)/50),-
density1$z/sum(density1$z)+density2$z/sum(density2$z),col=colram(50),nlevel
s=35,xlab="novelty",ylab="pluripotency")

```

Additional References Supplementary Figures and Notes:

21. P. M. Kim and B. Tidor, *Genome Res* **13** (7), 1706 (2003).
22. L. Ein-Dor, I. Kela, G. Getz et al., *Bioinformatics* **21** (2), 171 (2005).
23. J. P. Daily, D. Scandfield, N. Pochet et al., *Nature* **450** (7172), 1091 (2007).
24. Y. Gao and G. Church, *Bioinformatics* **21** (21), 3970 (2005).