Stem Cell Research (2013) 10, 57–66

Available online at www.sciencedirect.com

SciVerse ScienceDirect

www.elsevier.com/locate/scr

SHORT REPORT

StemCellDB: The Human Pluripotent Stem Cell Database at the National Institutes of Health

Barbara S. Mallon [a,*,1], Josh G. Chenoweth [b,1,2], Kory R. Johnson [c], Rebecca S. Hamilton [a], Paul J. Tesar [b,3], Amarendra S. Yavatkar [c], Leonard J. Tyson [c], Kyeyoon Park [a], Kevin G. Chen [a], Yang C. Fann [c], Ronald D.G. McKay [a,b,2]

[a] *NIH Stem Cell Unit, Division of Intramural Research, National Institute of Neurological Disorders and Stroke, National Institutes of Health, Bethesda, MD 20892, USA*
[b] *Laboratory of Molecular Biology, Division of Intramural Research, National Institute of Neurological Disorders and Stroke, National Institutes of Health, Bethesda, MD 20892, USA*
[c] *Bioinformatics Section, Information Technology and Bioinformatics Program, Division of Intramural Research, National Institute of Neurological Disorders and Stroke, National Institutes of Health, Bethesda, MD 20892, USA*

Received 14 August 2012; received in revised form 12 September 2012; accepted 18 September 2012
Available online 26 September 2012

Abstract Much of the excitement generated by induced pluripotent stem cell technology is concerned with the possibility of disease modeling as well as the potential for personalized cell therapy. However, to pursue this it is important to understand the 'normal' pluripotent state including its inherent variability. We have performed various molecular profiling assays for 21 hESC lines and 8 hiPSC lines to generate a comprehensive snapshot of the undifferentiated state of pluripotent stem cells. Analysis of the gene expression data revealed no iPSC-specific gene expression pattern in accordance with previous reports. We further compared cells, differentiated as embryoid bodies in 2 media proposed to initiate differentiation towards separate cell fates, as well as 20 adult tissues. From this analysis we have generated a gene list which defines pluripotency and establishes a baseline for the pluripotent state. Finally, we provide lists of genes enriched under both differentiation conditions which show the proposed bias toward independent cell fates.
Published by Elsevier B.V.

* Corresponding author at: 9000 Rockville Pike, Bldg 37/Rm 1000, Bethesda, MD 20892, USA. Fax: +1 301 480 1022.
E-mail address: mallonb@mail.nih.gov (B.S. Mallon).
URL: http://stemcells.nih.gov/research/nihresearch/scunit/ (B.S. Mallon).
[1] Equal contribution.
[2] Current address: Lieber Institute for Brain Development, Johns Hopkins Medical Campus, 855 N. Wolfe Street, Suite 300, Baltimore, MD 21205.
[3] Current address: Department of Genetics, School of Medicine, Case Western Reserve University, Cleveland, OH 44106, USA.

1873-5061/$ - see front matter. Published by Elsevier B.V.
http://dx.doi.org/10.1016/j.scr.2012.09.002

Introduction

As alternatives to human embryonic stem cells (hESCs), such as induced pluripotent stem cells (hiPSCs) (Park et al., 2008; Takahashi et al., 2007; Takahashi and Yamanaka, 2006; Yu et al., 2007) are explored, an accurate definition of what constitutes pluripotency becomes important. Continued progress toward realizing the potential of human pluripotent stem cells will be facilitated by robust datasets and complementary resources that are easily accessed and interrogated by the stem cell community. Many genome-wide microarray expression studies have been performed on hESCs using a variety of different technologies (Bock et al., 2011; Chin et al., 2009; Liu et al., 2006; Muller et al., 2011; Rao et al., 2004; Skottman et al., 2005; Sperger et al., 2003 and reviewed in Bhattacharya et al., 2009). To complement the existing data, we report here the establishment of the Human Pluripotent Stem Cell Database at the National Institutes of Health (NIH), StemCellDB, where we provide an in-house dataset of pluripotent human stem cells. StemCellDB provides data on all twenty one hESC lines available on the pre-2008 NIH Human Pluripotent Stem Cell Registry and eight human induced pluripotent stem cells (iPSCs), derived in-house by retroviral transduction of human fibroblasts. To facilitate comparisons of gene expression data between human pluripotent stem cells for the casual user, in both the undifferentiated and differentiated states, we have created a user-friendly search engine. This may be accessed directly at http://stemcelldb.nih.gov or through the 'Searchable Databases' link on the NIH Stem Cell Unit homepage, http://stemcells.nih.gov/research/nihresearch/scunit/. Here, a single gene portal allows users to examine individual genes for expression under all culture conditions.

To demonstrate the value of the database, we have compared the microarray gene expression profiles from undifferentiated and differentiated hESCs, as well as from 20 adult tissues and provide a list of 169 gene probes which can be used to define pluripotency at the gene expression level. Although overall gene expression is similar in the hESC lines, reproducible differences in expression between certain genes are observed. In addition to gene expression microarray data, StemCellDB provides access to data for single nucleotide polymorphism (SNP) genotyping, array-based comparative genomic hybridization (aCGH), miRNA array and DNA methylation analysis from matched samples (http://stemcelldb.nih.gov). The data may also be accessed through the NCBI GEO public database (Superseries number GSE34200). This facilitates interrogation and comparison of transcriptional regulation to advance our understanding of the pluripotent state. Taken together, the data deposited in StemCellDB constitute a benchmark reference data set which should be of great interest to the scientific community.

Materials and methods

Human ES cell culture

All culture reagents were acquired from Invitrogen unless stated otherwise. Standard culture conditions of 37 °C, 5% CO_2 and 95% humidity were maintained for all cells. Cell lines used and their suppliers are listed in Table 1.

Human ES cells (hESCs) were cultured on a feeder-layer of irradiated CF1 mouse embryonic fibroblasts (MEFs) in DMEM:F12(Cat# 11330–032) containing 20% Knockout Serum Replacement (KSR) (Cat# 10828–028), 1 mM glutamine (Cat# 25030–081), 0.1 mM β-mercaptoethanol (β-ME; Sigma), 1× non-essential amino acids (NEAA; Cat# 11140–050) and 4 ng/ml bFGF (R&D Systems) (Cat# 233-FB). Fibroblasts were cultured in DMEM (Cat# 11965–092) containing 10% fetal bovine serum (FBS) (Gemini Bio-products), 2 mM glutamine and 1× NEAA. Fibroblasts were irradiated with ~6500 rads using a Faxitron RX650 X-irradiator. They were subsequently plated on Falcon 6-well tissue culture dishes, coated with 0.1% gelatin, at a density of 0.1875×10^6/well. hESCs were plated in small clumps the following day, medium was exchanged every day and colonies were passaged by collagenase treatment every 3–4 days. Briefly, cultures were treated with 1.5 mg/ml collagenase IV for 20–40 min and either tapped sharply or scraped to dislodge colonies. Colonies were allowed to sediment for 5 min, the supernatant was removed and fresh media added. This process was repeated for a total of 3 sediments. At this point cells were triturated to generate colonies of approximately 10–100 cells for passaging or 50–250 cells for embryoid body (EB) formation. Embryoid bodies were cultured in fibroblast medium (FBS; EB_mesend) or in hESC medium without bFGF (EB_ecto) in 60 mm Corning Low Attachment dishes for a total of 8 days. Media were changed by sedimentation every 2 days. An important point to note is that the same lot number of FBS was used for all studies.

Nucleic acid extractions

For Comparative Genomic Hybridization (CGH), genomic DNA was extracted using the Wizard Genomic DNA Purification Kit (Promega) according to the manufacturer's instructions.

For gene expression microarray analysis, RNA was extracted using a modification of the basic Trizol (Invitrogen) protocol. Briefly, 1 ml of Trizol was added to sedimented colonies or EBs and triturated to dissociate the cells. At this point the lysates were stored at −80 °C until all samples for that cell line were collected. Upon thaw, lysates were incubated at room temperature for 10 min, mixed with 200 μl chloroform and centrifuged in a Phase-Lock Gel (heavy) Eppendorf tube (Qiagen). RNA was precipitated from the aqueous phase by the addition of 250 μl of isopropanol and 250 μl of a high salt buffer (0.8 M sodium citrate and 1.2 M NaCl) followed by centrifugation. The RNA pellet was washed twice with 75% ethanol, dried and resuspended in nuclease-free water. RNA was DNase treated for 20 min and the DNase removed using Ambion's DNA-Free kit. Concentration was determined using a NanoDrop ND-1000 UV–VIS spectrophotometer.

Array technologies

Global gene expression analysis was performed using Agilent human One Color Gene Expression Oligo arrays, reagents

Table 1 Karyotype and FISH analysis for chromosomes 12 and 17 are provided where performed (ND — not done). Both NIH and supplier nomenclature are given for all hESCs. The Coriell reference number is given for the hiPSC lines generated from that source cell line.

Cell line	Supplier name	Supplier	Passage #	Karyotype	Chr 12&17 fish
BG01	hESBGN-01	BresaGen, Inc	79	Normal	Normal
BG02	hESBGN-02	BresaGen, Inc	54	Normal	Normal
BG03	hESBGN-03	BresaGen, Inc	ND	ND	ND
ES01	HES-1	ES Cell International	72	Normal	ND
ES02	HES-2	ES Cell International	49	Normal	Normal
ES03	HES-3	ES Cell International	88	Normal	Normal
ES04	HES-4	ES Cell International	76	Normal	Normal
ES05	HES-5	ES Cell International	59	Normal	Normal
ES06	HES-6	ES Cell International	62	Normal	Normal
SA01	Sahlgrenska-1	Cellartis AB	32	Normal	Normal
SA02	Sahlgrenska-2	Cellartis AB	39	Abnormal [a]	Normal
TE03	I3	Technion — Israel Institute of Technology	70	Normal	Normal
TE04	I4	Technion — Israel Institute of Technology	ND	ND	ND
TE06	I6	Technion — Israel Institute of Technology	64	Abnormal [b]	Normal
UC01	HSF-1	University of California, San Francisco	64	Normal	1/200 trisomy 12
UC06	HSF-6	University of California, San Francisco	59	Normal	Normal
UC06	HSF-6	University of California, San Francisco	114	Normal	Normal
WA01	H1	WiCell Research Institute	57	Normal	Normal
WA07	H7	WiCell Research Institute	54	Normal	2/200 trisomy 17
WA09	H9	WiCell Research Institute	45	Normal	Normal
WA13	H13	WiCell Research Institute	ND	ND	ND
WA14	H14	WiCell Research Institute	40	Normal	ND
NIH-i1	Neonatal HFF	NIH/Vogel Lab	16	Normal	Normal
NIH-i2	AG20443	Coriell	24	Normal	Normal
NIH-i4	AG20443	Coriell	21	Abnormal [c]	Normal
NIH-i5	AG20443	Coriell	21	Normal	Normal
NIH-i7	AG08395	Coriell	21	Normal	Normal
NIH-i11	AG20443	Coriell	25	Abnormal [c]	Normal
NIH-i12	AG08396	Coriell	21	Normal	Normal
NIH-i13	AG08396	Coriell	18	Normal	Normal

[a] Trisomy 13 characteristic of this line.
[b] Nonclonal aberrations in 2/20.
[c] Balanced translocation present in the parent fibroblasts.

and software as previously described (Tesar et al., 2007). Comparative genomic hybridization and analysis was performed using Agilent software, reagents and arrays according to the manufacturer's instructions using 3 μg genomic DNA. Control male and female DNA was obtained from Promega. SNP analysis and methylation profiling were conducted by AGTC, Fairfax, VA using the Illumina Human1M-Duov3 and Human Methylation 27 k platforms respectively. MicroRNA arrays were performed using Agilent Human miRNA microarray kits, reagents and software.

Microarray data statistical analysis

The statistical programming language R (http://cran.r-project.org/) was used. Details are also shown in Supplemental Fig. 1. Raw expression measurements for all gene probes for all samples were log (base=2) transformed then quantile normalized. Quality of data was assured via sample-level inspection by Tukey box plot, covariance-based PCA scatter plot and correlation-based Heat Map. Raw expression measurements for samples deemed outliers were discarded and quantile normalization repeated. Gene probes not having at least one expression measurement greater than system noise post normalization were deemed "noise-biased" and discarded. System noise was defined as the lowest observed expression measurement at which the LOWESS (locally weighted scatterplot smoothing) fit of the CV (coefficient of variation) by mean for each gene probe for each class of samples (i.e., "ES undiff", "ES EB_ecto", "ES EB_mesend") grossly deviates from linearity. For gene probes not discarded, expression measurements were floored to equal system noise if less than system noise and were then subject to the one-factor ANOVA (analysis of variance) under BH (Benjamini and Hochberg) FDR (false discovery rate) MCC (multiple comparison correction) condition. Gene probes with a

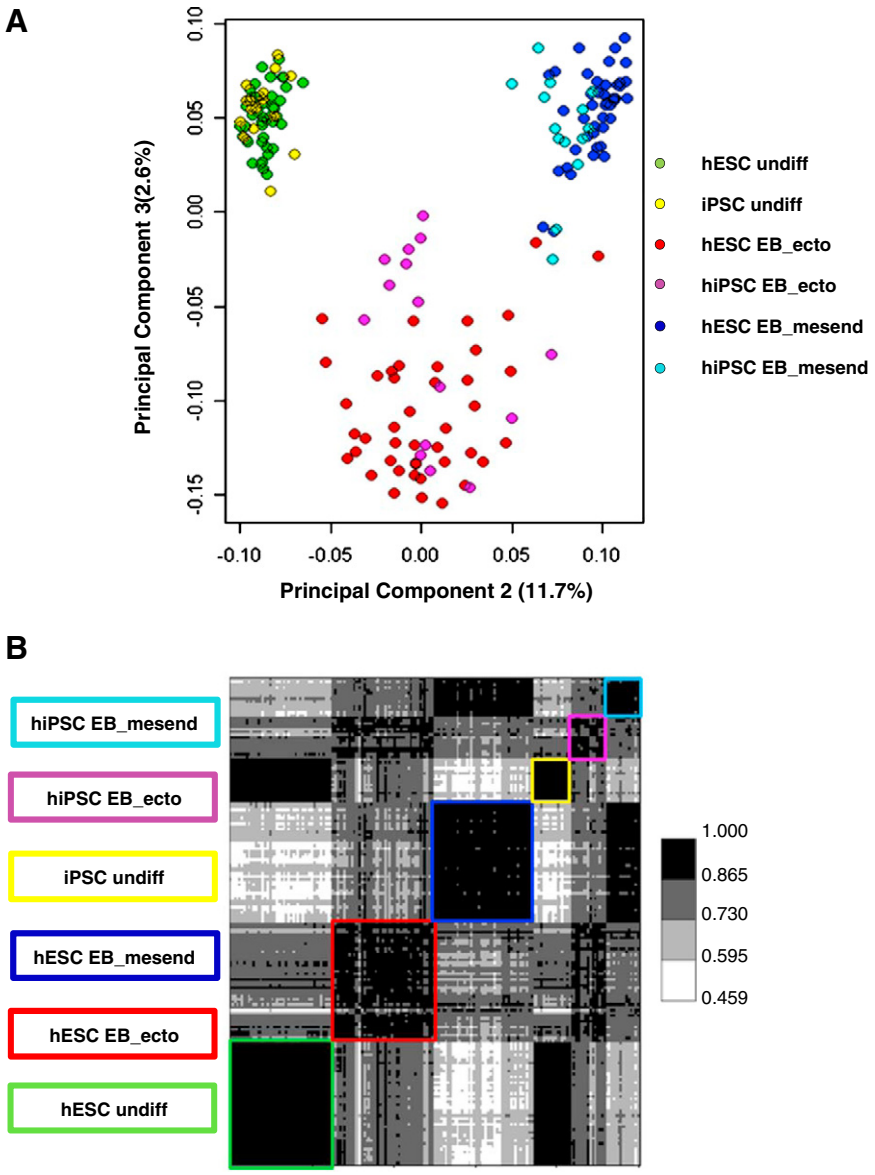

Figure 1 (A) Covariance PCA scatterplot and (B) Pearson correlation heat map depicting 175 samples using log (base=2) transformed quantile normalized expression for 4482 gene probes per sample. Gene probes selected for use represent those not noise-biased, not processing-biased, not gender-biased that pass ANOVA under multiple comparison correction condition ($P<0.05$) using class as the factor, pass post-hoc testing for at least one pair-wise class comparison (Tukey HSD $P<0.05$) and pass a mean–difference criteria (≥ 1.75) for the same pair-wise class comparison having a Tukey HSD $P<0.05$.

corrected P-value <0.05 were deemed "potentially informative" and subject to the TukeyHSD (honestly significant difference) post-hoc test. Gene probes having a post-hoc P-value <0.05 and a difference of class means ≥ 1.50 for a specific comparison of classes were deemed to have expression "significantly different" between the two classes. For these gene probes, measurements were subsequently interrogated for association with processing time and/or differences in gender using PolySerial correlation and ANOVA respectively under BH FDR MCC condition (alpha<0.05). Those gene probes having measurements significantly associated with processing time were deemed "processing-biased" while gene probes having measurements significantly associated with differences in gender were deemed "gender-biased". Annotations and associated functions for each gene probe were obtained using IPA (Ingenuity, Inc.).

Results and discussion

Comparison of hESCs and hiPSCs gene expression profiles

All twenty one hESC lines available on the pre-2008 NIH Human Pluripotent Stem Cell Registry and eight human iPSCs, derived in-house by retroviral transduction of human fibroblasts were adapted to one standard culture protocol. The cells were expanded to assess their identity and genomic integrity. Short

Table 2 Putative markers of pluripotency. Of the original 489 gene probes down-regulated in both differentiation conditions, 169 were found to be expressed in somatic tissues at a level less than the 5th percentile observed in hESCs and are designated markers of pluripotency.

GeneProbe	Gene	Gene_description
A_32_P74847	LARP7	La ribonucleoprotein domain family, member 7
A_24_P668974	LARP7	La ribonucleoprotein domain family, member 7
A_24_P383640	POU5F1P3	POU class 5 homeobox 1 pseudogene 3
A_32_P211752	LOC100506507	Hypothetical LOC100506507
A_32_P132563	POU5F1	POU class 5 homeobox 1
A_24_P144601	POU5F1	POU class 5 homeobox 1
A_24_P214841	POU5F1	POU class 5 homeobox 1
A_23_P327910	ZIC3	Zic family member 3
A_23_P140362	VRTN	Vertebrae development homolog (pig)
A_23_P204640	NANOG	Nanog homeobox
A_23_P25587	LECT1	Leukocyte cell derived chemotaxin 1
A_23_P329798	CER1	Cerberus 1, cysteine knot superfamily, homolog (*Xenopus laevis*)
A_23_P59138	POU5F1	POU class 5 homeobox 1
A_23_P72817	GDF3	Growth differentiation factor 3
A_23_P380526	DPPA4	Developmental pluripotency associated 4
A_32_P135985	TDGF1	Teratocarcinoma-derived growth factor 1
A_23_P127322	NODAL	Nodal homolog (mouse)
A_23_P137484	L1TD1	LINE-1 type transposase domain containing 1
A_23_P374844	GAL	Galanin prepropeptide
A_23_P366376	TDGF1	Teratocarcinoma-derived growth factor 1
A_23_P216149	TERF1	Telomeric repeat binding factor (NIMA-interacting) 1
A_24_P357266	GRPR	Gastrin-releasing peptide receptor
A_32_P220696	TERF1	Telomeric repeat binding factor (NIMA-interacting) 1
A_23_P137573	LEFTY2	Left–right determination factor 2
A_24_P90022	SEPHS1	Selenophosphate synthetase 1
A_24_P192434	TERF1	Telomeric repeat binding factor (NIMA-interacting) 1
A_23_P207999	PMAIP1	Phorbol-12-myristate-13-acetate-induced protein 1
A_23_P102471	MSH2	MutS homolog 2, colon cancer, nonpolyposis type 1 (*E. coli*)
A_24_P392475	BPTF	Bromodomain PHD finger transcription factor
A_23_P28153	SCLY	Selenocysteine lyase
A_23_P209337	METTL21A	Methyltransferase like 21A
A_24_P50458	TERF1	Telomeric repeat binding factor (NIMA-interacting) 1
A_23_P204246	PHC1	Polyhomeotic homolog 1 (Drosophila)
A_23_P156310	SKP2 (includes EG:27401)	S-phase kinase-associated protein 2 (p45)
A_32_P137926	MMS22L	MMS22-like, DNA repair protein
A_23_P14821	GABRB3	Gamma-aminobutyric acid (GABA) A receptor, beta 3
A_32_P87531	DNAH14	Dynein, axonemal, heavy chain 14
A_23_P256142	AKIRIN1	Akirin 1
A_24_P162929	METTL21A	Methyltransferase like 21A
A_32_P741851	GLB1L3	Galactosidase, beta 1-like 3
A_24_P118452	SEPHS1	Selenophosphate synthetase 1
A_23_P47058	CUZD1	CUB and zona pellucida-like domains 1
A_24_P655268	LOC729082	Hypothetical LOC729082
A_24_P916586	BICD1	Bicaudal D homolog 1 (Drosophila)
A_23_P156842	EEF1E1	Eukaryotic translation elongation factor 1 epsilon 1
A_23_P259127	ESRP1	Epithelial splicing regulatory protein 1
A_32_P76091	HSPD1	Heat shock 60 kDa protein 1 (chaperonin)
A_24_P134727	TFAM	Transcription factor A, mitochondrial
A_23_P160336	LEFTY1	Left–right determination factor 1
A_24_P244699	NUDT15	Nudix (nucleoside diphosphate linked moiety X)-type motif 15
A_24_P52921	BCAT1	Branched chain amino-acid transaminase 1, cytosolic
A_23_P214907	MTHFD1L	Methylenetetrahydrofolate dehydrogenase (NADP+dependent) 1-like
A_32_P213091	SHISA9	Shisa homolog 9 (*Xenopus laevis*)
A_23_P323094	PHC1	Polyhomeotic homolog 1 (Drosophila)

(continued on next page)

Table 2 (*continued*)

GeneProbe	Gene	Gene_description
A_23_P82823	PINX1	PIN2/TERF1 interacting, telomerase inhibitor 1
A_23_P162256	DENR	Density-regulated protein
A_23_P365060	MDN1	MDN1, midasin homolog (yeast)
A_23_P18818	CNOT6	CCR4-NOT transcription complex, subunit 6
A_23_P148484	RLIM	Ring finger protein, LIM domain interacting
A_23_P111373	MRS2 (includes EG:380836)	MRS2 magnesium homeostasis factor homolog (*S. cerevisiae*)
A_23_P203201	DDX6	DEAD (Asp-Glu-Ala-Asp) box polypeptide 6
A_23_P92410	CASP3	Caspase 3, apoptosis-related cysteine peptidase
A_23_P216118	UNC5D	Unc-5 homolog D (*C. elegans*)
A_23_P214111	KIF13A	Kinesin family member 13A
A_23_P138465	NOLC1	Nucleolar and coiled-body phosphoprotein 1
A_23_P121423	CDC25A	Cell division cycle 25 homolog A (*S. pombe*)
A_23_P136504	SLC25A21	Solute carrier family 25 (mitochondrial oxodicarboxylate carrier), member 21
A_23_P73220	FGD6	FYVE, RhoGEF and PH domain containing 6
A_23_P421436	ADD2	Adducin 2 (beta)
A_23_P23356	RRP15 (includes EG:327053)	Ribosomal RNA processing 15 homolog (*S. cerevisiae*)
A_32_P34826	C21orf88	Chromosome 21 open reading frame 88
A_24_P128977	G3BP2	GTPase activating protein (SH3 domain) binding protein 2
A_23_P405761	RRAS2	Related RAS viral (r-ras) oncogene homolog 2
A_23_P70168	TARS	Threonyl-tRNA synthetase
A_24_P415260	DDX21	DEAD (Asp-Glu-Ala-Asp) box polypeptide 21
A_24_P253215	EMG1	EMG1 nucleolar protein homolog (*S. cerevisiae*)
A_23_P54834	NIP7	Nuclear import 7 homolog (*S. cerevisiae*)
A_23_P155407	RTP1	Receptor (chemosensory) transporter protein 1
A_24_P297888	MTAP	Methylthioadenosine phosphorylase
A_23_P351215	SKIL	SKI-like oncogene
A_32_P1614	LOC100506054	Hypothetical LOC100506054
A_24_P213794	CCRN4L	CCR4 carbon catabolite repression 4-like (*S. cerevisiae*)
A_23_P10966	GABRB3	Gamma-aminobutyric acid (GABA) A receptor, beta 3
A_23_P160881	SMPDL3B	Sphingomyelin phosphodiesterase, acid-like 3B
A_23_P373119	HMG4L	High mobility group box 3 pseudogene 1
A_23_P27167	RNASEH1	Ribonuclease H1
A_24_P49747	HMGB3P24	High mobility group box 3 pseudogene 24
A_23_P213908	PHAX	Phosphorylated adaptor for RNA export
A_23_P358417	TIMM8A	Translocase of inner mitochondrial membrane 8 homolog A (yeast)
A_24_P902052	SNHG13	Small nucleolar RNA host gene 13 (non-protein coding)
A_24_P24685	HMGB3P22	High mobility group box 3 pseudogene 22
A_24_P13533	LRR1	Leucine rich repeat protein 1
A_23_P215484	CCL26	Chemokine (C–C motif) ligand 26
A_23_P252362	MRPS30	Mitochondrial ribosomal protein S30
A_24_P943922	CACHD1	Cache domain containing 1
A_32_P194264	CHAC2	ChaC, cation transport regulator homolog 2 (E. coli)
A_24_P922606	NUP160	Nucleoporin 160 kDa
A_23_P133216	NLN	Neurolysin (metallopeptidase M3 family)
A_23_P128991	SLIRP	SRA stem-loop interacting RNA binding protein
A_23_P56553	METTL8	Methyltransferase like 8
A_23_P355075	CENPN	Centromere protein N
A_23_P134008	USP45	Ubiquitin specific peptidase 45
A_23_P41255	G3BP2	GTPase activating protein (SH3 domain) binding protein 2
A_23_P145724	C7orf16	Chromosome 7 open reading frame 16
A_23_P87759	EMG1	EMG1 nucleolar protein homolog (*S. cerevisiae*)
A_23_P56865	DDX18	DEAD (Asp-Glu-Ala-Asp) box polypeptide 18
A_24_P134626	TXLNG	Taxilin gamma
A_24_P234196	RRM2	Ribonucleotide reductase M2
A_23_P214354	EXOC2	Exocyst complex component 2
A_23_P5370	RPRM	Reprimo, TP53 dependent G2 arrest mediator candidate

Table 2 (*continued*)

GeneProbe	Gene	Gene_description
A_24_P12573	CCL26	Chemokine (C–C motif) ligand 26
A_23_P72770	USP44	Ubiquitin specific peptidase 44
A_24_P272389	LOC285216	Methylenetetrahydrofolate dehydrogenase (NADP+dependent) 2, methenyltetrahydrofolate cyclohydrolase pseudogene
A_23_P54540	EIF2AK4	Eukaryotic translation initiation factor 2 alpha kinase 4
A_24_P347624	SNURF	SNRPN upstream reading frame
A_24_P128085	RC3H2	Ring finger and CCCH-type domains 2
A_23_P102183	PNO1	Partner of NOB1 homolog (*S. cerevisiae*)
A_32_P71788	FKBP4	FK506 binding protein 4, 59 kDa
A_23_P204170	TMPO	Thymopoietin
A_32_P44775	C9orf85	Chromosome 9 open reading frame 85
A_23_P143958	RPL22L1	Ribosomal protein L22-like 1
A_24_P914479	SNX5	Sorting nexin 5
A_23_P427217	JMJD1C	Jumonji domain containing 1C
A_23_P204380	GNPTAB	N-acetylglucosamine-1-phosphate transferase, alpha and beta subunits
A_24_P344307	PSME3	Proteasome (prosome, macropain) activator subunit 3 (PA28 gamma; Ki)
A_24_P100664	MKKS	McKusick-Kaufman syndrome
A_23_P218918	FGF2	Fibroblast growth factor 2 (basic)
A_24_P314477	TUBB2B	Tubulin, beta 2B
A_24_P15754	TOMM40	Translocase of outer mitochondrial membrane 40 homolog (yeast)
A_23_P37497	MYO1E	Myosin IE
A_24_P143843	LOC729566	Zinc finger and BTB domain containing 8 opposite strand pseudogene 1
A_24_P152404	C10orf76	Chromosome 10 open reading frame 76
A_23_P125001	RAC3	Ras-related C3 botulinum toxin substrate 3 (rho family, small GTP binding protein Rac3)
A_23_P135063		
A_24_P161773		
A_24_P178523		
A_24_P179646		
A_24_P195286		
A_24_P221285		
A_24_P340659		
A_24_P341106		
A_24_P341731		
A_24_P358302		
A_24_P367326		
A_24_P392505		
A_24_P410000		
A_24_P41189		
A_24_P455060		
A_24_P560332		
A_24_P58597		
A_24_P67063		
A_24_P67681		
A_24_P695223		
A_24_P707102		
A_24_P711050		
A_24_P752362		
A_24_P76142		
A_24_P901084		
A_24_P928765		
A_32_P104334		
A_32_P146320		
A_32_P152696		
A_32_P157504		
A_32_P207147		
A_32_P24068		
A_32_P63086		

(continued on next page)

Table 2 (*continued*)

GeneProbe	Gene	Gene_description
A_32_P65691		
A_32_P885123		
A_32_P89049		

Tandem Repeat (STR) and single nucleotide polymorphism (SNP) genotyping confirmed that each line was genetically unique. Cytogenetic and array comparative genomic hybridization (aCGH) analysis showed that most cell lines have a normal chromosome complement (Table 1). In addition, flow cytometry demonstrated that nearly all cells expressed the pluripotent markers POU5F1 (Oct-4) and Tra-1-81. Quality control reports are available on our website, http://stemcelldb.nih.gov.

Covariance principal component analysis (PCA) and Pearson correlation of the gene expression microarray data indicated that hESCs and hiPSCs are grossly similar (class means>0.865) in the undifferentiated and differentiated states (Figs. 1A and B). In no class was any gene found to be exclusively expressed by one population of pluripotent cell versus the other. Thus, in agreement with published reports (Guenther et al., 2010), we conclude by this measure that there is no absolutely unique gene expression profile that can be assigned to hESCs or hiPSCs.

Pluripotency-associated genes

We assessed the expression and regulation of pluripotency markers in hESCs only and generated a list of 489 gene probes which are down-regulated in both differentiation conditions (Supplemental Table 1). Of this list, 169 gene probes were found to be expressed in somatic tissues at a level less than the 5th percentile observed in hESCs and are designated markers of pluripotency (Table 2). Included in this "pluripotency" list are genes involved in maintenance of the pluripotent state such as *POU5F1* and *NANOG* (Figs. 2A and B) as well as many components/targets of the TGFβ-superfamily signaling network including *NODAL* and *TDGF1* (Figs. 2C and D). This is consistent with a requirement for Activin/Nodal signaling in the maintenance of hESCs as described previously (James et al., 2005; Vallier et al., 2004). Also in the "pluripotency" list are gene probes that have not been annotated at this time, raising the possibility of novel pluripotency-associated genes. The use of these 169 probes in a focused array could possibly be used as a fingerprint for pluripotent stem cells.

Differentiation pathways in two embryoid body culture media

The differentiation conditions selected for embryoid body (EB) differentiation were designated EB_ecto, for ectodermal lineage, and EB_mesend, for mesendodermal lineage. We examined which genes changed under each condition to see if the differentiation media truly affected fate bias. We found

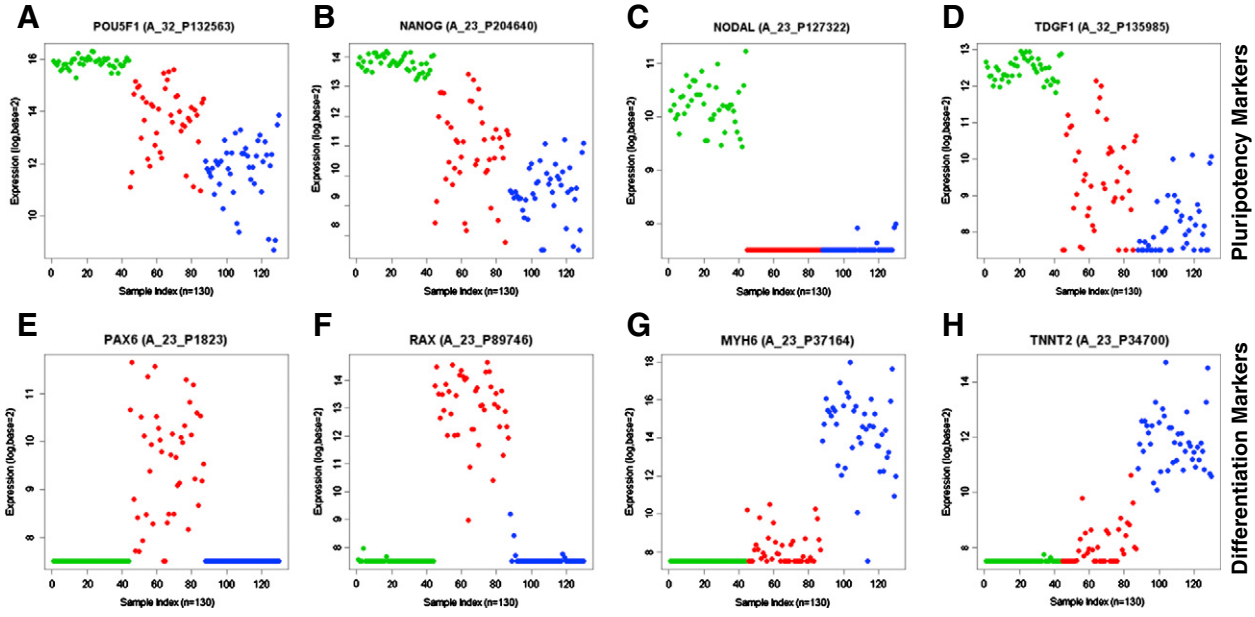

Figure 2 A–D) eNorthern of pluripotency genes — A) *POU5F1/Oct4*; B) *NANOG*; C) *NODAL*; D) *TDGF1*; E–F) eNorthern of differentiation associated gene expression — E) *PAX6*; F) *RAX*; G) *MYH6*; H) *TNNT2*. Green = undifferentiated samples, red = EB_ecto; blue = EB_mesend.

595 gene probes up-regulated in both conditions, 243 gene probes enriched in the EB_ecto condition, and 1086 gene probes enriched in the EB_mesend condition (Supplemental Table 2). Many genes encoding neurectodermal markers, such as *PAX6*, *RAX* (Figs. 2E and F), *LHX2*, and *LMO1*, are detected in the EB_ecto-enriched group, fitting gene ontology analysis indicating roles for this group in nervous system development and function. The 1086 genes enriched in the EB_mesend group include many genes encoding mesendodermal markers, such as *MYH6* and *TNNT2* (Figs. 2G and H), as well as many *HOX* and hemoglobin genes. Gene ontology analysis of this group demonstrates roles in cardiogenesis, vasculogenesis as well as muscular development. Taken together, the gene probes found to be up-regulated in EB_ecto or EB_mesend can discriminate lineage differentiation.

Using the StemCellDB gene expression search engine

The StemCellDB website is designed primarily to facilitate interrogation of the gene expression information by a casual user. Upon accessing the site, either directly at http://stemcelldb.nih.gov or through the 'Searchable Databases' link on the NIH Stem Cell Unit homepage http://stemcells.nih.gov/research/nihresearch/scunit/, an option to search the Agilent or Affymetrix datasets is presented. The Affymetrix dataset is a minor subset of the Agilent data which we have provided for comparison to other datasets commonly available. Selecting either dataset will not only present various options to search for a gene of interest for a casual user but also allows access to all datasets, for an advanced user, using the GEO submissions link on the sidebar (Fig. 3A). Upon searching the preferred cell type or tissue data, multiple probes may be returned as hits, not all of which may give useful information depending on probe location and other factors. With the Agilent dataset, a pop-up menu gives a snapshot of the data spread for each probe to allow the user to select the most informative probes for further evaluation (Fig. 3B). Once a probe is selected, the data is available for download or for charting. The Agilent dataset provides median normalized data as well as quantile raw-based, quantile median normalized-based and quantile $\log_2$ raw-based data, which are used for our analysis (Fig. 3C). Using the drop-down menus, the dataset may be downloaded as PDF, MS-Excel or text formats and charts plotted according to the desired data type. A more detailed tutorial for the use of the gene expression search engine may be found on the navigational sidebar (Fig. 3C).

We have also provided the quantile $\log_2$ gene expression data for the cell lines as a Microsoft Excel spreadsheet in Supplemental data (Supplemental Table 3).

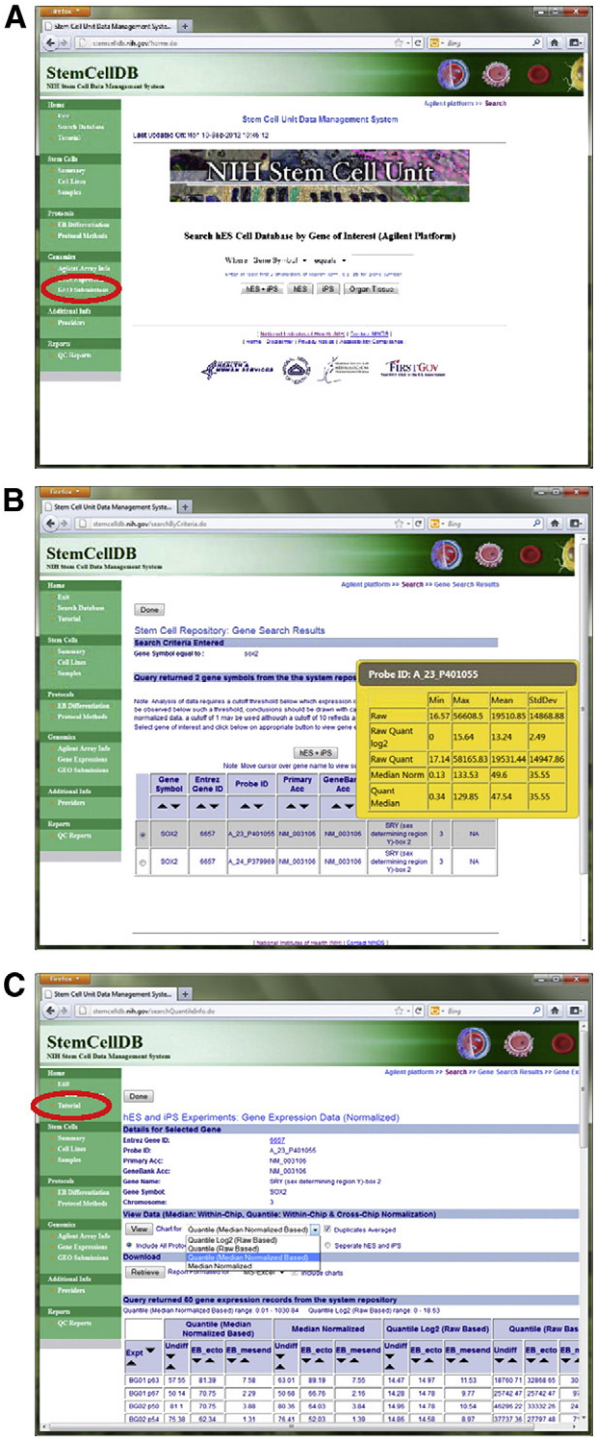

Figure 3 A) The StemCellDB gene expression search engine. A casual user can search their gene of interest directly or a more experienced user can access all raw data deposited with NCBI GEO by using the GEO Submissions link on the sidebar. B) Returned search results. Pop-up menus for each probe allow a user to select the most informative probes for further examination. C) Data is available in several formats for instant viewing as a chart or for download.

Conclusion

Here we report the launch of StemCellDB, a database of molecular profiles which together provide a comprehensive snapshot of human embryonic stem cells in their undifferentiated state including general differentiation potential. As described in other studies, we find no iPSC-specific gene expression pattern under any of the three culture conditions. We have analyzed the data to provide a list of 169 gene probes, which may be used as a fingerprint of pluripotency and show that 2 independent differentiation conditions can upregulate genes associated with different lineages. We have designed a user-friendly

search engine to facilitate casual interrogation of the gene expression data. Together, this provides a useful resource for the stem cell community.

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.scr.2012.09.002.

Acknowledgments

We would like to thank Dr. Pamela Gehron Robey for helpful discussions and Dr. Jeanette Beers for help with the fibroblast culture. This research was supported by the Intramural Research Program of the NIH.

References

Bhattacharya, B., Puri, S., Puri, R.K., 2009. A review of gene expression profiling of human embryonic stem cell lines and their differentiated progeny. Curr Stem Cell Res Ther 4, 98–106.

Bock, C., Kiskinis, E., Verstappen, G., Gu, H., Boulting, G., Smith, Z.D., Ziller, M., Croft, G.F., Amoroso, M.W., Oakley, D.H., et al., 2011. Reference maps of human ES and iPS cell variation enable high-throughput characterization of pluripotent cell lines. Cell 144, 439–452.

Chin, M.H., Mason, M.J., Xie, W., Volinia, S., Singer, M., Peterson, C., Ambartsumyan, G., Aimiuwu, O., Richter, L., Zhang, J., et al., 2009. Induced pluripotent stem cells and embryonic stem cells are distinguished by gene expression signatures. Cell Stem Cell 5, 111–123.

Guenther, M.G., Frampton, G.M., Soldner, F., Hockemeyer, D., Mitalipova, M., Jaenisch, R., Young, R.A., 2010. Chromatin structure and gene expression programs of human embryonic and induced pluripotent stem cells. Cell Stem Cell 7, 249–257.

James, D., Levine, A.J., Besser, D., Hemmati-Brivanlou, A., 2005. TGFbeta/activin/nodal signaling is necessary for the maintenance of pluripotency in human embryonic stem cells. Development 132, 1273–1282.

Liu, Y., Shin, S., Zeng, X., Zhan, M., Gonzalez, R., Mueller, F. J., Schwartz, C. M., Xue, H., Li, H., Baker, S. C., et al., 2006. Genome wide profiling of human embryonic stem cells (hESCs), their derivatives and embryonal carcinoma cells to develop base profiles of U.S. Federal government approved hESC lines. BMC Dev Biol 6, 20.

Muller, F.J., Schuldt, B.M., Williams, R., Mason, D., Altun, G., Papapetrou, E.P., Danner, S., Goldmann, J.E., Herbst, A., Schmidt, N.O., et al., 2011. A bioinformatic assay for pluripotency in human cells. Nat Methods 8, 315–317.

Park, I.H., Zhao, R., West, J.A., Yabuuchi, A., Huo, H., Ince, T.A., Lerou, P.H., Lensch, M.W., Daley, G.Q., 2008. Reprogramming of human somatic cells to pluripotency with defined factors. Nature 451, 141–146.

Rao, R.R., Calhoun, J.D., Qin, X., Rekaya, R., Clark, J.K., Stice, S.L., 2004. Comparative transcriptional profiling of two human embryonic stem cell lines. Biotechnol Bioeng 88, 273–286.

Skottman, H., Mikkola, M., Lundin, K., Olsson, C., Stromberg, A.M., Tuuri, T., Otonkoski, T., Hovatta, O., Lahesmaa, R., 2005. Gene expression signatures of seven individual human embryonic stem cell lines. Stem Cells 23, 1343–1356.

Sperger, J.M., Chen, X., Draper, J.S., Antosiewicz, J.E., Chon, C.H., Jones, S.B., Brooks, J.D., Andrews, P.W., Brown, P.O., Thomson, J.A., 2003. Gene expression patterns in human embryonic stem cells and human pluripotent germ cell tumors. Proc Natl Acad Sci U S A 100, 13350–13355.

Takahashi, K., Tanabe, K., Ohnuki, M., Narita, M., Ichisaka, T., Tomoda, K., Yamanaka, S., 2007. Induction of pluripotent stem cells from adult human fibroblasts by defined factors. Cell 131, 861–872.

Takahashi, K., Yamanaka, S., 2006. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. Cell 126, 663–676.

Tesar, P.J., Chenoweth, J.G., Brook, F.A., Davies, T.J., Evans, E.P., Mack, D.L., Gardner, R.L., McKay, R.D., 2007. New cell lines from mouse epiblast share defining features with human embryonic stem cells. Nature 448, 196–199.

Vallier, L., Reynolds, D., Pedersen, R.A., 2004. Nodal inhibits differentiation of human embryonic stem cells along the neuroectodermal default pathway. Dev Biol 275, 403–421.

Yu, J., Vodyanik, M.A., Smuga-Otto, K., Antosiewicz-Bourget, J., Frane, J.L., Tian, S., Nie, J., Jonsdottir, G.A., Ruotti, V., Stewart, R., et al., 2007. Induced pluripotent stem cell lines derived from human somatic cells. Science 318, 1917–1920.