

Unique gene expression signatures of independently-derived human embryonic stem cell lines

Michael J. Abeyta^{1,2,3}, Amander T. Clark^{1,2,3}, Ryan T. Rodriguez^{1,3}, Megan S. Bodnar^{1,3}, Renee A. Reijo Pera^{1,2,3,*} and Meri T. Firpo^{1,3}

¹Center for Reproductive Sciences, Department of Obstetrics, Gynecology and Reproductive Sciences, ²Departments of Physiology and Urology, Programs in Human Genetics, Cancer Genetics and ³Developmental and Stem Cell Biology, University of California at San Francisco, 513 Parnassus Avenue, San Francisco, CA 94143-0546, USA

Received October 15, 2003; Revised and Accepted January 19, 2004

Human embryonic stem cells (hESCs) have the potential to differentiate to diverse cell types. This ability endows hESCs with promise for the development of novel therapeutics, as well as promise for the development of a rigorous genetic system to probe human gene function. However, in spite of the impending utility of hESCs for clinical and basic applications, little is known about their fundamental properties. Recent reports have documented transcriptional profiles of mouse embryonic stem cells (mESCs), adult stem cells and a single hESC line, H9. To date, however, the transcriptional profiles of independently-derived hESC lines have not been compared. In order to examine the similarities and differences in multiple hESC lines, we compared gene expression profiles of the HSF-1, HSF-6 and H9 lines. We found that the majority of genes examined were expressed in all three cell lines. However, we also observed that each line possessed a unique expression signature; the expression of many genes was limited to just one or two hESC lines. We suggest that the observed differences in gene expression between independently-derived hESC lines may reflect inherent differences in the initial culture of each line and/or the underlying genetics of the embryos from which the lines were derived.

INTRODUCTION

All stem cells have the ability to self-renew and the ability to differentiate. Embryonic stem cells are unique in that they are derived from the inner cell mass of the developing blastocyst and, thus, can give rise to all the tissue types of the embryo. In recent years, mouse embryonic stem cells (mESCs) have been used widely as a model system to explore mammalian embryonic stem cell biology and to identify novel pathways such as those that function in maintaining pluripotency. While the mouse model provides a foundation for studying stem cell biology, distinct differences between mESCs and human embryonic stem cells (hESCs) have been observed. In particular, LIF (leukemia inhibiting factor) activity, modulated through gp130 and the JAK/STAT pathway, is sufficient for maintenance of undifferentiated mESCs (1). In contrast, addition of exogenous LIF is not sufficient to maintain

undifferentiated hESCs (2,3) and the role of gp130 and the JAK/STAT pathway has not been explored in detail. Differences such as this highlight the importance of studying stem cell biology in multiple systems, including humans.

A recent report compared the molecular signature of the hESC line, H9, to a pluripotent mESC line (4). A set of 918 genes was identified that was enriched in undifferentiated compared to differentiated H9 hESCs including several potentially relevant members of secretory pathways. In addition, the authors noted that a subset of hESC-enriched genes was also enriched in mESCs (5). This suggested that a comparison of the transcriptional profiles of different pluripotent cells may advance the identification of a core set of molecular components that define pluripotent cells.

Substantial data from model organisms has documented that gene expression profiles comprise a phenotype that differs between individuals of different genetic backgrounds (6–13).

*To whom correspondence should be addressed. Tel: +1 4154763178; Fax: +1 4154763121; Email: reijo@itsa.uscf.edu

Studies of the genetics of gene expression in *Saccharomyces cerevisiae*, for example, demonstrate that the expression of more than 1500 genes (of the total 6000 open reading frames in the yeast genome) differed between two closely-related strains and that the expression differences were modulated by complex genetic differences between the strains (6,7). Similar differences in gene expression were linked to genome variation in *Drosophila*, mice and primates (9–13). In light of studies such as these, it is notable that each hESC line that has been derived carries a unique version of the human genome that can be traced to the inner cell mass cells of a single embryo (2). Given this fact, and observations from model organisms, we expect that fundamental characteristics of each hESC line may be determined by the unique human genomic variants it carries. We sought to define inherent similarities and differences in independently-derived hESC lines by assessing differentiation to the three germ layers, endoderm, mesoderm and ectoderm, and then by comparing gene expression in multiple hESC lines in order to determine whether they possessed unique gene expression signatures that might reflect the genetic variants of each line. In addition, we directly assessed variation in expression of genes that may be critical for central properties of hESCs, such as maintenance of pluripotency, and cross-referenced our profiles with published data regarding both mESCs and adult stem cells.

RESULTS

The three hESC lines examined in this study were HSF-1, HSF-6 and H9 (NIH codes, UC01, UC06 and WA09, respectively). The karyotype of the HSF-6 and H9 cell lines is 46XX and that of HSF-1 is 46XY. Detailed information regarding these cell lines can be obtained at: <http://stemcells.nih.gov/stemcell/>. Of particular importance, however, is the observation that each of these lines is a pluripotent stem cell line, based on expression of markers of pluripotent cells such as SSEA-3 and SSEA-4 and the ability to differentiate to the three human germ layers (<http://stemcells.nih.gov/stemcell/>). Each cell line also has demonstrated the ability to form teratomas *in vitro* (2 and data not shown).

Differentiation of independently-derived hESC lines

To assess variation in independently-derived hESC lines, we examined their ability to contribute to ectoderm, endoderm and mesoderm, in more detail (Fig. 1). For this purpose, hESCs from three replicate cultures of each line, H9, HSF-1 and HSF-6, were removed from feeder layers and allowed to differentiate in suspension for up to 14 days. A sample of cells was removed at 0, 3, 7, 10 and 14 days of differentiation and gene expression of markers indicative of differentiation to endoderm (*AFP*), mesoderm [*VEGFR2* (*KDR*)] and ectoderm (*NCAMI*) were assayed. As expected, we observed that each replicate of each line differentiated to cells that expressed markers indicative of the three germ layers. Moreover, within each line, there was little variation in the kinetics of somatic differentiation (Fig. 1). In contrast, however, we observed that the expression patterns of the *AFP*, *VEGFR2* and *NCAMI* differentiation markers among the H9, HSF-1 and HSF-6 lines differed markedly. For example, expression of *AFP* increased significantly after 3 days

of differentiation of HSF-1, but did not increase until day 7 in HSF-6 or day 14 in H9 (Fig. 1). Similar differences were observed for the ectodermal marker, *NCAMI*, with significant expression observed after 3 days of differentiation in HSF-1 and HSF-6 but no significant expression until 7–10 days of differentiation with H9. Likewise, expression of the mesodermal marker, *VEGFR2*, increased after 7 days of differentiation of HSF-1 and HSF-6, but not until 10 days of differentiation of H9. Thus, although all three hESC lines have the potential to form all three germ layers, the time course of differentiation varied among the lines in this set of experiments. Whether this is a result of random fluctuation in the differentiation process or a reflection of variation inherent to each line merits further consideration.

Microarray analysis of multiple hESC lines

We next compared gene expression in the three independently-derived hESC lines, H9, HSF-1 and HSF-6. For this purpose, all three hESC lines were grown on three plates each of confluent CF1 feeder cells, using a single lot of feeders and medium. hESCs were harvested by incubation in culture medium and gentle pipetting to avoid collection of the highly adherent feeder cells with the hESC population. Cells from individual dishes were washed and frozen for RNA isolation. At the onset of cell culture for this study, the hESC line HSF-1 was at passage 36, HSF-6 was at passage 46 and H9 was at passage 51. For global gene expression analysis, we used Affymetrix oligonucleotide arrays that contained 44 794 oligonucleotide probes for known genes and expressed sequence tags (ESTs). RNA isolates were prepared from three independent cultures of each cell line, resulting in a total of nine microarray sets (hU133A and B GeneChips) analyzed in the experiment.

We analyzed our data from the microarray sets as follows: the robust multichip average (RMA) was used to calculate normalized expression values from the array dataset (14). We compared variability within each line to variability among the three independently-derived lines by calculating correlation coefficients from the normalized values. Consistent with array findings, the correlations between all samples were high, but in all instances correlation coefficients within a cell line were higher than those between lines. For example, the correlation coefficient for a comparison of HSF-6 sample 2 versus HSF-6 sample 3 was 0.9963, whereas the correlation coefficient for the comparison of HSF-6 sample 2 versus HSF-1 sample 2 was 0.9728.

Next, we compared arrays to identify significant differences in gene expression between pairs of different hESC lines as compared to within a particular line via analysis of variance (ANOVA) using an *F*-test with an adjusted *P*-value cut-off ≤ 0.01 . The *F* statistic (ANOVA) tests the null hypothesis that there is no difference between the different test groups; a high *F*-score (low adjusted *P*-value) is indicative of larger variation between groups than within a group. In contrast, a low *F*-score indicates high variability within a group. Genes that did not meet the adjusted *P*-value cut-off (low *F*-score) were eliminated from further analysis of relative expression. The list of genes that differed significantly among the three independent cell lines was further analyzed by *t*-test

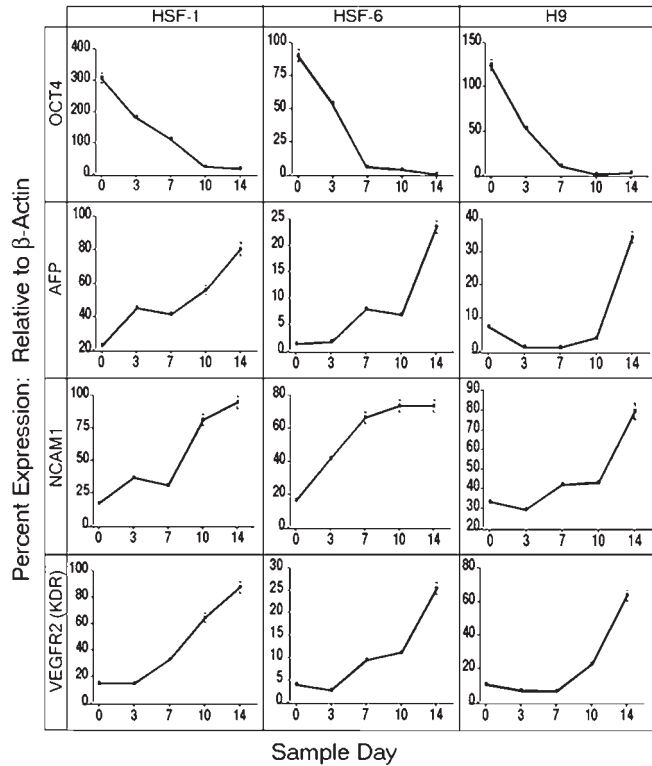


Figure 1. Semi-quantitative RT-PCR analysis of hESC-derived embryoid bodies (EBs). hESCs from each line were cultured under differentiation conditions, without feeders in serum-containing medium. *OCT4*, a marker of undifferentiated ES cells, *AFP* an endoderm marker, *VEGFR2* (*KDR*, *FLK1*), a mesoderm marker and *NCAM*, an ectoderm marker were assayed at each time point over a period of 2 weeks. Expression levels of the markers are shown relative to β -actin, a house-keeping gene that was assayed in the same reaction.

comparison (P -value ≤ 0.01) of each combination of ratio of medians (ROM; H9 versus HSF-6, H9 versus HSF-1 and HSF-6 versus HSF-1) to determine the comparisons in which the differences occurred. A cutoff of the ROM ≥ 1 and ≤ -1 (these \log_2 values correspond to ≥ 2 -fold change) was chosen to represent significant changes in gene expression. A stringent P -value was chosen to insure that we were conservative in assessing differences between cell lines. Further examination of several genes via quantitative PCR and subsequent gel electrophoresis demonstrates the correlation of gene expression of a subset of genes from the array (Fig. 2 and Table 1).

Unique expression signatures for each cell line

The MAS 5.0 software detection algorithm (Affymetrix) was used to determine the presence or absence of expression for each gene represented on the array. A gene with either the detection call of 'present' or 'marginal' was considered present for that array. We only considered a gene to be expressed in a cell line if that gene received a present call for all three arrays. A Venn diagram was constructed to illustrate the presence or absence of gene expression in the three hESC lines (Fig. 2). We found that 14136 (31.5%) of the total gene and EST probes on the oligonucleotide arrays (44 806) were expressed in one or more of the hESC lines. Surprisingly, however, only 7385 (52%) of those

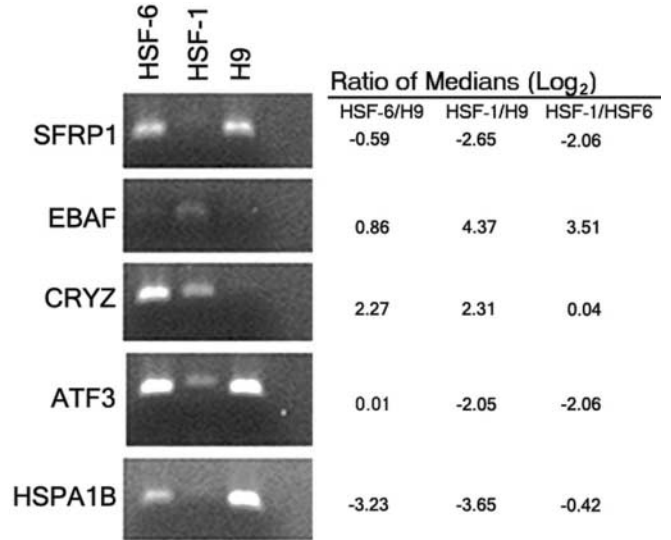


Figure 2. Verification of significant fold change differences among the three hESC lines. Gel electrophoresis from quantitative, real-time PCR performed on the following genes: (Affymetrix probe set ID in parenthesis): *SFRP1* (202037_s_at), *EBAF* (206012_at), *CRYZ* (202950_at), *ATF3* (202672_s_at; also shown in Table 1), and *HSPA1B* (202581_at). The numbers displayed to the right of the gel (expressed in \log_2) are ANOVA fold change data from genes showing large significant differences among the three cell lines. Gel bands on the left visually demonstrate these differences. Further explanation of ANOVA fold change statistics is given in Table 1.

were expressed in all three hESC lines (Supplementary Material, Table S3). An *in silico* analysis, done to identify the tissues in which these genes are expressed, showed that the vast majority of genes present in all three hESC lines are ubiquitously expressed (Supplementary Material, Table S4). As might be expected, no genes were found that were only expressed in embryonic stem cells, although several hundred genes and ESTs that have the potential to be important or specific to stem cells were lacking categorized expression information.

Of the 52% of probes expressed by all three lines, many were expressed at >2 -fold and significantly different levels among the three lines, according to the ANOVA data. Thus, in addition to differences in the presence or absence of expression of genes among the three lines, there was also substantial variability in the levels of gene expression among genes present in all three lines. Therefore, we conclude that the differential gene expression in the hESC lines imparts a unique expression signature on each cell line, as diagrammed in the non-overlapping portion of the circles (Fig. 3). As shown, 2279 probe sets were detected at significant levels in embryonic stem cell line H9 alone; 337 were unique to HSF-1 and 641 were unique to HSF-6. Genes expressed in all three hESC lines, as well as those expressed with unique signatures, are involved in diverse cellular processes ranging from RNA processing and transcriptional regulation to signaling and metabolism (Table 2 and Supplementary Material, Table S3). Given that each line can differentiate to the primary germ layers (Fig. 1), it is clear that the unique signature of each stem cell line does not prevent contribution of the hESC lines to multiple cell fates. However, particular gene expression signatures could be associated with

Table 1. Comparison of real-time data to array data

GENE	-Δ Ct from qPCR (log ₂)			ROM from arrays (log ₂)		
	HSF6/H9	HSF1/H9	HSF1/HSF6	HSF6/H9	HSF1/H9	HSF1/HSF6
P38IP	-0.07	-0.7	-0.6	-0.61 (A)	-0.82 (A)	-0.21 (A)
ATF3	-0.25	-4.55	-4.3	0.01 (C)	-2.05 (B)	-2.06 (B)
IVNS1ABP	-0.35	-1.15	-0.8	-0.41 (A)	-1.22 (B)	-0.81 (A)

These genes were chosen because they demonstrate three different statistical groups according to the ANOVA data. Expression data on the right side of the table labeled 'A' represents ratio of medians (ROM) data from the array that was statistically significant (*t*-test *P*-value ≤0.01). However, the actual differences in expression were <2-fold. Array data labeled 'B' indicates similar results except that the expression differences are >2-fold. Array data labeled 'C' indicates data that did not meet *P*-value cut-off criteria, and is therefore not statistically significant.

preferential differentiation along a particular course, or as suggested by Figure 1, variation in the kinetics or time course of differentiation.

Comparison of hESC data to known stem cell genes

In order to assess the biological relevance of the variable gene expression found in the three hESC lines, we focused more closely on the expression profiles of selected genes implicated in other stem cell systems and on the expression of a set of genes that might function in regulating self renewal and pluripotency in mouse embryonic and adult stem cells, according to published reports (5,15,16). We found that a number of genes important to other stem cells, including *POU5F1* (*OCT4*), *SOX2*, *ZFP42* (*REX1*), *TDGF1* (*CRIPTO*), *NODAL*, *EBAF* (*LEFTY*), *EPHA2*, *THY1*, *PROM1* (*AC133*), *CXCR4* and *TALI* (*SCL/TAL*), were expressed at detectable levels in all three of the undifferentiated hESC lines that we examined (Table 3). However, greater than 2-fold differences in expression were observed in a subset of these genes (*SOX2*, *LEFTY*, *AC133* and *CXCR4*). Moreover, the genes *GP130*, *STAT3*, *FOXD3*, *H19*, *FGF4*, *CD34*, *RUNX1*, *VEGFR2* (*KDR*, *FLK1*) and *KIT*, expressed in adult stem cell and mESC populations, were absent in one or more hESC lines. This data indicated that the variability in gene expression in hESC lines extended to several potentially important stem cell genes. Although these hESC lines are pluripotent, based on their SSEA-3 and SSEA-4 staining and ability to differentiate to the three human germ layers, these results suggested that there is likely to be functional significance associated with the unique gene expression signatures of independently-derived hESC lines.

Comparison of hESC data to enriched stem cell genes

Recently, two groups published reports of gene sets that might be common to multiple stem cell populations (5,15). We sought to determine whether there are common genes that are also expressed in multiple hESC lines, as well. We overlapped the two previously-reported sets of genes that were enriched in stem cells (adult and mouse embryonic) and identified a list of 11 genes with human homologs that were reported to be enriched in both datasets. A search of our dataset showed that each of the 11 genes was expressed in all three hESC lines (*PPIC*, *CCND1*, *CGI30*, *ITGA6*, *MSH2*, *BYSL*, *LAPTM4B*, *KIAA1018*, *ELOVL6*, *USP9X* and *ARCNI*) (5,15). Of these

genes, six were expressed in all three lines but with greater than 2-fold differences between the lines (*CGI30*, *ITGA6*, *MSH2*, *LAPTM4B*, *ELOVL6* and *USP9X*).

We further cross-referenced our data solely with the list of 'stemness' genes (5). Of the 216 mouse genes identified with 'stemness', we identified 125 with known human homologs that were represented on the Affymetrix hU133 oligonucleotide array set. Interestingly, 76 (60.8%) of these homologous genes were expressed in all three hESC lines, whereas the remaining 39.2% were either not detected or were expressed in two or fewer hESC lines (Fig. 4). Analysis of the 76 commonly-expressed, 'stemness' transcripts (Supplementary Material, Table S5) demonstrated that 17 of these genes displayed greater than a 2-fold difference in gene expression between the three lines. Given the fact that each hESC line has demonstrated self renewal capacity and pluripotency (Fig. 1), the observed variability suggests that hESCs may have different gene expression requirements than other stem cell populations.

DISCUSSION

Global gene expression analysis has identified expression profiles that are shared between, and specific to, distinct cell types (17). In most studies to date, ~10–20% of genes and ESTs were expressed in any one human cell or tissue type (18). In contrast, gene expression in stem cells appears to be elevated. For example, in one study, ~60% of the genes and ESTs probed were expressed in three mouse stem cell types (embryonic, neural and hematopoietic stem cells) (5). In a second study, ~40% of the sequences on oligonucleotide arrays were called present (expressed) in cultured human neural stem cells (19). Similarly, in our study, we found that 31.5% of the total gene and EST sequences on the arrays were expressed. Thus, in line with other studies, we found that indeed hESCs express a large proportion of the genes and ESTs probed, but certainly not the majority. However, of those gene and EST probes expressed, 52% were expressed in all three lines (albeit frequently at different levels) whereas 48% were restricted to one or two hESC lines.

Recent studies have elucidated a set of genes that are strictly housekeeping genes in that they are expressed in all human cell and tissue types examined to date (20). Notably, even within this fundamental set of commonly-expressed genes, variation in levels of expression are observed. This variation not only provides a reliable and unique signature to distinguish common tissues of the human body, but it may also contribute to

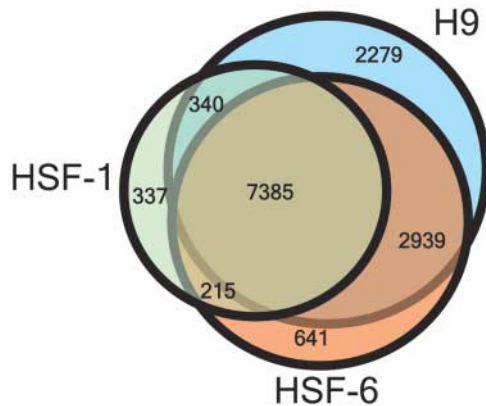


Figure 3. Venn diagram of shared and unique gene expression in independently-derived hESC lines. The region of overlap between all circles indicates the number of gene and EST probe sets expressed in all three hESC lines. Regions of overlap between two cell lines indicate gene expression in common between two of the three hESC lines. Regions that do not overlap between circles indicate unique gene expression signatures of particular hESC lines. Note that the numbers displayed correspond to the number of Affymetrix probe sets, not genes. Affymetrix estimates that the 44 794 probe sets present in the hU133 array set represent >39 000 transcripts from ~33 000 genes. One difficulty of microarray analysis is to determine the level of detection that distinguishes presence or absence of a gene product. User-defined parameters were set at the default settings listed in the Affymetrix manual to allow confidence in genes scored as present, and also for ease of standardization to future data sets. Consequently, some genes scored as absent might be detected by more sensitive assay methods. Table S3 (Supplementary Material) contains the complete list of genes present and absent determined via criteria described.

differences in basic biochemical processes in diverse cell types (20).

Given the considerations noted above, both the shared and unique gene expression signatures that we uncovered among the three independent hESC lines have important implications for further studies. First, we reason that if a gene is to be necessary for the fundamental functions of hESCs to self renew and to contribute to all three germ layers, then it should be expressed in all hESC lines. Thus, our results suggest that the list of 7385 commonly expressed hESC probe sets encompasses genes required for self-renewal and differentiation into multiple cell types. This list will be further refined when the expression profiles from other hESC lines from various laboratories are similarly analyzed and compared to these results. Second, the observation that each line possessed a unique expression signature suggests differences exist in either the culture history or the underlying genetics of each hESC line. We can discount several factors as the major determinants of expression differences. As indicated, all cell lines were at similar passage numbers. In addition, two of three cell lines have an XX karyotype and the third has an XY karyotype. Although Y chromosome genes were uniquely expressed in the XY cells, this cell line was not the most unique of the lines. Lastly, the growth conditions of all three lines were standardized such that they were grown with the same medium preparation, plating and passaging protocols and maintained on a common stock of feeder cells.

When we consider our results, along with the fact that each hESC line is derived from a single embryo that carries specific genetic variants of the human genome, we propose that

fundamental differences in the underlying genetics are the most likely determinant of variability observed among hESC lines. Variation in DNA sequences are common in the human genome; single nucleotide polymorphisms (SNPs) are present at an approximate frequency of 1 in every 1000 nucleotides (21). Given these observations, any given hESC line will carry thousands of SNPs that could potentially contribute to the unique kinetics of differentiation and gene expression that we observed. It is likely that a more complete understanding of the extent and consequences of this genetic variability is fundamental for maximizing the potential usefulness of hESC lines in both research and therapeutic applications. Currently, however, only a very limited number of hESC lines are available and they represent just a small fraction of the genetic diversity of the human population.

MATERIALS AND METHODS

Culture and maintenance of undifferentiated hESCs

hESC lines are routinely maintained in culture with replacement by frozen stocks every 20–30 passages. Cells are expanded on mouse primary feeder cells and routinely tested for sterility, hESC characteristics (expression of *OCT4*, *SSEA-3* and *SSEA-4*), and differentiation potential. Feeder cells were from CF-1 mice at day 13 or 14 of gestation (E13.5–14.5). Tissues were minced, washed several times, cultured in DMEM medium [plus 10% fetal bovine serum (FBS)]. Fibroblasts were passaged at least three times before use with hESCs to minimize residual non-fibroblast cells and were used as feeders between passage 3 and 7. Feeder cells were irradiated with gamma irradiation, frozen in liquid nitrogen in freezing solution (with 90% FBS, 10% DMSO). Feeders were thawed at 37°C, washed and plated onto gelatinized tissue culture plates. Confluent feeder plates were used for hESC cultures up to one week after thawing. Working stocks of hESCs were thawed at 37°C, washed and plated onto 10 cm plates with monolayers of feeders. Cultures were observed daily, first to confirm that clumps of hESCs were adhering to the feeders and spreading out into typical hESC colonies, and later to determine passaging. When colonies reached an average size of 300–400 cells, plates were passaged regardless of the number of colonies in the plate. Colonies were removed from the plate, disrupted with vigorous pipetting to clumps of 50–100 cells, and treated with growth medium containing 1 mg/ml type IV collagenase for 15–20 min. Cells were then washed to remove residual collagenase. Medium was then replaced and cells were replated.

hESC differentiation

hESCs from each cell line were cultured under differentiation conditions, without feeders in serum containing medium. Clumps of hESCs readily formed EBs similar to those formed by differentiating mESCs. EBs were harvested over a period of 2 weeks. *OCT4*, a marker of undifferentiated ES cells, *AFP* an endoderm marker, *VEGFR2* (*KDR*, *FLK1*), a mesoderm marker, and *NCAM*, an ectoderm marker were assayed at each time point. Expression levels of the markers are shown relative

Table 2. Abbreviated list of highly-expressed genes from the shared and unique gene expression signatures of the three hESC lines. Also refer to Supplementary Material, Table S3 for complete lists from each category

Category	Genes
Shared in all three hESC lines	<i>ACTG1, ATP5A1, CFL1, DJ-1, EEF1D, EIF3S6IP, EIF4A1, TPT1, ENO1, FAU, FTH1, FTL, GAPD, GLTSCR2, GNAS, GNB2L1, HMGB1, HNRPA1, HSPA8, K-ALPHA-1, LDHB, LENG7, MYL6, NACA, NGFRAP1, NPM1, NSEP1, OAZ1, PPIA, PRDX1,</i>
Unique to HSF-1	<i>A2M, ARHD, ATP2B4, BAALC, BART1, BMP7, C9orf25, CER1, CLDN5, CLDN7, CXCR4, FLJ22501, FOXG1B, GNG3, GPR56, GSTT2, H2AFM, HOXB2, HSPC002, IGFBP7, KPNB2, LM02, MFRP, MS4A6A, POU3F2, RPS4Y, SOST</i>
Unique to HSF-6	<i>ARL7, BCAN, CGI-36, CKIP-1, EGR1, EGR2, H2BFG, HIP-55, LGI1, NNAT, NOG, PCMI, PIGT, SERPINE1, SMOC1, SOX15, TRIM33, ZNF22, FLJ14299, KIAA0376, MGC2601, KIAA0644, KIAA0876, FLJ22955, HPI-BP74</i>
Unique to H9	<i>ABHD4, AKR1B10, AMOTL1, ANXA11, CA14, COX10, NDRG2, DDX11, DHRSX, DNAJB1, DOM3Z, DUSP16, FXYD5, GSTT1, HLA-E, MGST2, NCK2, NOT56L, PHC1, POLRMT, PRKCZ, PRO1853, SLC7A7, TNIP2, TCF7L2</i>

Table 3. Human embryonic stem cells exhibit variable expression of genes associated with other stem cell populations. Genes were selectively chosen from the literature (see references)

	Detected in all three hESC lines by array analysis	Fold changes ≥ 2 among genes detected in all three hESC lines	Absent in one or more hESC lines, by array analysis
Expressed in mESCs	<i>OCT4, SOX2, REXICRIPTO, NODAL, LEFTY, EPHA2, THY1 (2,23-33)</i>	<i>SOX2, LEFTY (24,25)</i>	<i>GPI30, STAT3, FOXD3, H19, FGF4 (25,34-36)</i>
Expressed in adult stem cells	<i>AC133, CXCR4, SCL/TAL (37-39)</i>	<i>AC133, CXCR4 (38,39)</i>	<i>CD34, RUNX1, VEGFR2 (KDR, FLK1), KIT (40-43)</i>

to β -actin, a housekeeping gene that was assayed in the same reaction. For PCR, total RNA was extracted with RNeasy (Qiagen); mRNA was prepared via Oligotex mRNA protocol (Qiagen). mRNA was reverse transcribed into cDNA (Invitrogen) and subsequently subjected to multiplex PCR for the semi-quantitative evaluation of the mRNA expression of *OCT4*, *AFP*, *VEGFR2* (*KDR*, *FLK1*) and *NCAM*. An aliquot of 50 ng of cDNA was amplified using 0.5 U/reaction Taq DNA Polymerase (Invitrogen), 2–4 ng *OCT4*, *AFP*, *VEGFR2* (*KDR*, *FLK1*) and *NCAM* primers. Primers were as follows: β -actin, 5'-CACCTGAAGTACCCCATCGAGCA-3', 5'-CAGGTCT-TTGGCGATGTCCACGTCAC-3'; *OCT4*, 5'-CTGCAGTG-TGGGTTTCGGGCA-3', 5'-CTTGCTGCAGAAAGTGGGTG-GAGGA-3'; *AFP*, 5'-ACTGCAATTGAGAAACCCACTG-GAGATG-3', 5'-CGATGCTGGAGTGGGCTTTTTGTGT-3'; *VEGFR2* (*KDR*, *FLK1*), 5'-AAGGTGACAGGAAAAGACG-AACT-3', 5'-TCCCCTCCATTGGCCCGCTTAAC-3'; *NCAM*, 5'-AGGAGACAGAAACGAAGCCA-3', 5'-GGTGTGGAA-ATGCTCTGGT-3'.

PCR conditions and electrophoresis were as previously described (35). Gel Expert software was used to quantify gene expression from gel bands. Expression levels are represented relative to β -actin expression.

Preparation of hESC RNA for Affymetrix arrays

Total RNA was extracted from cells using RNeasy (Qiagen). One microgram of total RNA was primed with 100 ng of Oligo dT-T7 primer and reverse transcribed with Superscript II (Invitrogen). A second strand was synthesized and the dscDNA was purified with DNA Clean and Concentrator (Zymo Research). The *in vitro* transcription reaction was incubated for 9 h with T7 RNA polymerase. First round aRNA was

purified with Qiagen RNeasy Mini Kit. Second round amplification was performed similar to the first round, but with 100 ng of aRNA and 500 ng random hexamers. Following the second round dscDNA, the ENZO BioArray HighYield RNA Transcript Labeling Kit (Affymetrix) was used to incorporate biotin-labeled nucleotides, then RNA was purified using RNeasy. Fragmentation was completed with the standard protocol (Affymetrix). Prior to hybridization on GeneChip arrays, a test3 array of housekeeping controls was analyzed to determine sample suitability for GeneChip arrays. Hybridized arrays were subsequently scanned for data analysis. A detailed RNA amplification protocol is available upon request.

Quantitative, real-time PCR

Total RNA was extracted from undifferentiated hESC lines, which were cultured under standardized conditions, using an RNeasy Mini Kit (Qiagen). For these experiments, equal amounts of total RNA (3 μ g) were then reverse transcribed from each cell line using SuperScript II (Invitrogen) according to the manufacturer's protocol. Each quantitative PCR reaction used 50 ng of this cDNA. The standardization of cDNA quantity allowed for direct comparison of Ct values in each cell line for a given gene. PCR reactions were optimized using 50 ng cDNA, 3 mM MgCl₂, 10 mM dATP, dGTP, dCTP and dTTP, 2 mM primers and 0.25 U platinum Taq (Invitrogen). PCR cycling: one cycle, 95°C, 5 min followed by 35 cycles 95°C, 30 s; 60°C, 30 s; 72°C, 30 s. Quantitative, real-time PCR was performed with the addition of 1 \times SYBR green (Molecular Probes), 1 \times fluorescein (Biorad) and 4% DMSO. Equal reaction efficiencies of all five genes were determined using 10-fold serial dilution of hESC cDNA over a 100-fold range. Calculations of relative levels were performed as

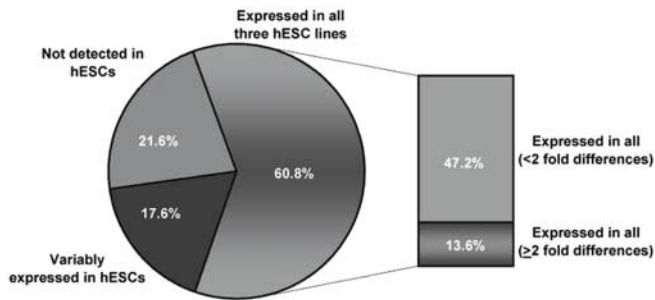


Figure 4. A comparison of the expression of 125 genes potentially associated with 'stemness' in mice and humans. Out of 216 potential 'stemness' genes, we identified 125 with human homologs that were present on the Affymetrix arrays. Of these 125 human homologs, 39.2% were detected in two or fewer hESC lines, whereas 60.8% were detected in all, albeit at variable levels (13.6% ≥ 2 -fold difference).

indicated (22). Primers were based on exemplar sequence of the Affymetrix probe set ID: *ATF3*, 5'-CTGAAGAAGATG-AAAGGAAAAAGA-3', 5'-ATGTATATCAAATGCTGCTTC-TCGT-3'; *IVNS1ABP*, 5'-TGAAAATAATACCTGGACTTTAATTGC-3', 5'-ATTTCTAGTTGGATCATAACATTTCCAC-3'; *P38IP*, 5'-CAACAAGCTTTAGAACTAGCTTTGG-3', 5'-TTCCTGGGTATAGATTGACCACTAA-3'; *SFRP1*, 5'-TAAGTGTGACAAGTTCCCGGA-3', 5'-TTCACCTCTTTTATTTTC-ATCCTCAGT-3'; *CRYZ*, 5'-AAGTTGATGAGAGCTGTTA-GAGTTT-3', 5'-ATAAGTACCAGAGCGAATGTATGTCTC-3'; *EBAF*, 5'-ACATGGAGAAGCTTGGTCATC-3', 5'-AAGC-TCTGGCTGAACCTCTTT-3'; *HSPA1B*, 5'-CTGTTTTCCTA-TGTCATTTCTGCTT-3', 5'-AAACACCCTTACAGTATCAA-CATTA-3'.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at HMG Online.

ACKNOWLEDGEMENTS

We thank C. Barker and K. Hanspers for valuable assistance in all aspects of Affymetrix array analysis, A. Dobson for helpful discussions, and F. Moore, A. Sehnert, R. Taylor, J. Tung, R. Weiner and E. Xu for critical reading of the manuscript. This work was carried out in part in the General Clinical Research Center (GCRC) at San Francisco General Hospital and supported by grant 5-MO1-RR00083 from the Division of Research Resources, National Institutes of Health. This work was also supported by National Institutes of Health grants from the National Center for Research Resources (grant R24-RRD17498) to M.T.F. and the National Institute of Child Health and Human Development (grant RO1-HD37095) and the Sandler Family Foundation to R.A.R.P.

REFERENCES

- Burdon, T., Smith, A. and Savatier, P. (2002) Signalling, cell cycle and pluripotency in embryonic stem cells. *Trends Cell Biol.*, **12**, 432–438.
- Thomson, J.A., Itskovitz-Eldor, J., Shapiro, S.S., Waknitz, M.A., Swiergiel, J.J., Marshall, V.S. and Jones, J.M. (1998) Embryonic stem cell lines derived from human blastocysts. *Science*, **282**, 1145–1147.
- Rose-John, S. (2002) GP130 stimulation and the maintenance of stem cells. *Trends Biotechnol.*, **20**, 417–419.
- Sato, N., Sanjuan, I.M., Heke, M., Uchida, M., Naef, F. and Brivanlou, A.H. (2003) Molecular signature of human embryonic stem cells and its comparison with the mouse. *Dev. Biol.*, **260**, 404–413.
- Ramalho-Santos, M., Yoon, S., Matsuzaki, Y., Mulligan, R.C. and Melton, D.A. (2002) 'Stemness': transcriptional profiling of embryonic and adult stem cells. *Science*, **298**, 597–600.
- Brem, R.B., Yvert, G., Clinton, R. and Kruglyak, L. (2002) Genetic dissection of transcriptional regulation in budding yeast. *Science*, **296**, 752–755.
- Cavaliere, D., Townsend, J.P. and Hartl, D.L. (2000) Manifold anomalies in gene expression in a vineyard isolate of *Saccharomyces cerevisiae* revealed by DNA microarray analysis. *Proc. Natl Acad. Sci. USA*, **97**, 12639–12374.
- Steinmetz, L.M., Sinha, H., Richards, D.R., Spiegelman, J.I., Oefner, P.J., McCusker, J.H. and Davis, R.W. (2002) Dissecting the architecture of a quantitative trait locus in yeast. *Nature*, **416**, 326–330.
- Jin, W., Riley, R.M., Wolfinger, R.D., White, K.P., Passador-Gurgel, G. and Gibson, G. (2001) The contributions of sex, genotype and age to transcriptional variance in *Drosophila melanogaster*. *Nature Genet.*, **29**, 389–395.
- Oleksiak, M.F., Churchill, G.A. and Crawford, D.L. (2002) Variation in gene expression within and among natural populations. *Nature Genet.*, **32**, 261–266.
- Sandberg, R., Yasuda, R., Pankratz, D.G., Carter, T.A., Rio, J.A.D., Wodicka, L., Mayford, M., Lockhart, D.J. and Barlow, C. (2000) Regional and strain-specific gene expression mapping in the adult mouse brain. *Proc. Natl Acad. Sci. USA*, **97**, 11038–11043.
- Enard, W., Khaitovich, P., Klose, J., Zollner, S., Heissig, F., Giavalisco, P., Niesel-Struwe, K., Muchmore, E., Varki, A., Ravid, R. et al. (2002) Intra- and interspecific variation in primate gene expression patterns. *Science*, **296**, 340–343.
- Cheung, V.G. and Spielman, R.S. (2002) The genetics of variation in gene expression. *Nature Genet. Suppl.*, **32**, S22–S25.
- Bolstad, B.M., Irizarry, R.A., Astrand, M. and Speed, T.P. (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, **19**, 185–193.
- Ivanova, N.B., Dimos, J.T., Schaniel, C., Hackney, J.A., Moore, K.A. and Lemischka, I.R. (2002) A stem cell molecular signature. *Science*, **298**, 601–604.
- Brivanlou, A.H., Gage, F.H., Jaenisch, R., Jessell, T., Melton, D. and Rossant, J. (2003) Setting standards for human embryonic stem cells. *Science*, **300**, 913–916.
- Smith, L. and Greenfield, A. (2003) DNA microarrays and development. *Hum. Mol. Genet.*, **12**, R1–R8.
- Warrington, J.A., Nair, A., Mahadevappa, M. and Tsyganskaya, M. (2000) Comparison of human adult and fetal expression and identification of 535 housekeeping/maintenance genes. *Physiol. Genomics*, **2**, 143–147.
- Wright, L.S., Li, J., Caldwell, M.A., Wallace, K., Johnson, J.A. and Svendsen, C.N. (2003) Gene expression in human neural stem cells: effects of leukemia inhibitory factor. *J. Neurochem.*, **86**, 179–195.
- Hsiao, L.L., Dangond, F., Yoshida, T., Hong, R., Jensen, R.V., Misra, J., Dillon, W., Lee, K.F., Clark, K.E., Haverty, P. et al. (2001) A compendium of gene expression in normal human tissues. *Physiol. Genomics*, **7**, 97–104.
- Kruglyak, L. and Nickerson, D. (2001) Variation is the spice of life. *Nature Genet.*, **27**, 234–236.
- Livak, K.J. and Schmittgen, T.D. (2001) Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) method. *Methods*, **25**, 402–408.
- Ben-Shushan, E., Thompson, J.R., Gudas, L.J. and Bergman, Y. (1998) Rex-1, a gene encoding a transcription factor expressed in the early embryo, is regulated via Oct-3/4 and Oct-6 binding to an octamer site and a novel protein, Rox-1, binding to an adjacent site. *Mol. Cell. Biol.*, **18**, 1866–1878.
- Nishimoto, M., Fukushima, A., Okuda, A. and Muramatsu, M. (1999) The gene for the embryonic stem cell coactivator UTF1 carries a regulatory element which selectively interacts with a complex composed of Oct-3/4 and Sox-2. *Mol. Cell. Biol.*, **19**, 5453–5465.
- Du, Z., Cong, H. and Yao, Z. (2001) Identification of Putative Downstream Genes of Oct-4 by Suppression-Subtractive Hybridization. *Biochem. Biophys. Res. Commun.*, **282**, 701–706.

26. Xu, C., Inokuma, M.S., Denham, J., Golds, K., Kundu, P., Gold, J.D. and Carpenter, M.K. (2001) Feeder-free growth of undifferentiated human embryonic stem cells. *Nat. Biotechnol.*, **19**, 971–974.
27. Henderson, J.K., Draper, J.S., Baillie, H.S., Fishel, S., Thomson, J.A., Moore, H. and Andrews, P.W. (2002) Preimplantation human embryos and embryonic stem cells show comparable expression of stage-specific embryonic antigens. *Stem Cells*, **20**, 329–337.
28. Reubinoff, B.E., Pera, M.F., Fong, C.Y., Trounson, A. and Bongso, A. (2000) Embryonic stem cell lines from human blastocysts: somatic differentiation *in vitro*. *Nat. Biotechnol.*, **18**, 399–404.
29. Avilion, A.A., Nicolis, S.K., Pevny, L.H., Perez, L., Vivian, N. and Lovell-Badge, R. (2003) Multipotent cell lineages in early mouse development depend on SOX2 function. *Genes Devel.*, **17**, 126–140.
30. Johnson, S.E., Rothstein, J.L. and Knowles, B.B. (1994) Expression of epidermal growth factor family gene members in early mouse development. *Dev. Dyn.*, **201**, 216–226.
31. Draper, J.S., Pigott, C., Thomson, J.A. and Andrews, P.W. (2002) Surface antigens of human embryonic stem cells: changes upon differentiation in culture. *J. Anat.*, **200**, 249–258.
32. Lickliter, J.D., Smith, F.M., Olsson, J.E., Mackwell, K.L. and Boyd, A.W. (1996) Embryonic stem cells express multiple Eph-subfamily receptor tyrosine kinases. *Proc. Natl Acad. Sci. USA*, **93**, 145–150.
33. Herrman, B. and Robertson, E.J. (1994) A primary requirement for nodal in the formation and maintenance of the primitive streak in the mouse. *Development*, **120**, 1919–1928.
34. Matsuda, T., Nakamura, T., Nakao, K., Arai, T., Katsuki, M., Heike, T. and Yokota, T. (1999) STAT3 activation is sufficient to maintain an undifferentiated state of mouse embryonic stem cells. *EMBO J.*, **18**, 4261–4269.
35. Hanna, L.A., Foreman, R.K., Tarasenko, I.A., Kessler, D.S. and Labosky, P.A. (2002) Requirement for *Foxd3* in maintaining pluripotent cells of the early mouse embryo. *Genes Devel.*, **16**, 2650–2661.
36. Poirier, F., Chan, C.T.J., Timmons, P.M., Robertson, E.J., Evans, M.J. and Rigby, P.W. (1991) The murine H19 gene is activated during embryonic stem cell differentiation *in vitro* and at the time of implantation in the developing embryo. *Development*, **113**, 1105–1114.
37. Sanchez, M.J., Bockamp, E.O., Miller, J., Gambardella, L. and Green, A.R. (2001) Selective rescue of early haematopoietic progenitors in *Scl(-/-)* mice by expressing *Scl* under the control of a stem cell enhancer. *Development*, **128**, 4815–4827.
38. Aiuti, A., Webb, I.J., Bleul, C., Springer, T. and Gutierrez-Ramos, J.C. (1997) The chemokine SDF-1 is a chemoattractant for human CD34+ hematopoietic progenitor cells and provides a new mechanism to explain the mobilization of CD34+ progenitors to peripheral blood. *J. Exp. Med.*, **185**, 111–120.
39. Uchida, N., Buck, D.W., He, D., Reitsma, M.J., Masek, M., Phan, T.V., Tsukamoto, A.S., Gage, F.H. and Weissman, I.L. (2000) Direct isolation of human central nervous system stem cells. *Proc. Natl Acad. Sci. USA*, **97**, 14720–14725.
40. Harder, F., Kirchhof, N., Petrovic, S., Schmittwolf, C., Durr, M. and Muller, A.M. (2002) Developmental potentials of hematopoietic and neural stem cells following injection into pre-implantation blastocysts. *Ann. Hematol.*, **81**, S20–S21.
41. North, T.E., Bruijn, M.F.d., Stacy, T., Talebian, L., Lind, E., Robin, C., Binder, M., Dzierzak, E. and Speck, N.A. (2002) Runx1 expression marks long-term repopulating hematopoietic stem cells in the midgestation mouse embryo. *Immunity*, **16**, 661–672.
42. Miyagi, T., Takeno, M., Nagafuchi, H. and Takahashi, M. (2002) Flk1+ cells derived from mouse embryonic stem cells reconstitute hematopoiesis *in vivo* in SCID mice. *Exp. Hematol.*, **30**, 1444–1453.
43. Orlic, D., Fischer, R., Nishikawa, S., Nienhuis, A.W. and Bodine, D.M. (1993) Purification and characterization of heterogeneous pluripotent hematopoietic stem cell populations expressing high levels of c-kit receptor. *Blood*, **82**, 762–770.