OPEN ACCESS Freely available online

PLoS COMPUTATIONAL BIOLOGY

Meta-Analysis of Differentiating Mouse Embryonic Stem Cell Gene Expression Kinetics Reveals Early Change of a Small Gene Set

Clive H. Glover[1], Michael Marin[1,2], Connie J. Eaves[3,4], Cheryl D. Helgason[5,6], James M. Piret[1,7], Jennifer Bryan[1,2*]

1 Michael Smith Laboratories, University of British Columbia, Vancouver, British Columbia, Canada, **2** Department of Statistics, University of British Columbia, Vancouver, British Columbia, Canada, **3** Department of Medical Genetics, University of British Columbia, Vancouver, British Columbia, Canada, **4** Terry Fox Laboratory, British Columbia Cancer Agency, Vancouver, British Columbia, Canada, **5** Department of Surgery, University of British Columbia, Vancouver, British Columbia, Canada, **6** Department of Cancer Endocrinology, British Columbia Cancer Agency, Vancouver, British Columbia, Canada, **7** Department of Chemical and Biological Engineering, University of British Columbia, Vancouver, British Columbia, Canada

Stem cell differentiation involves critical changes in gene expression. Identification of these should provide endpoints useful for optimizing stem cell propagation as well as potential clues about mechanisms governing stem cell maintenance. Here we describe the results of a new meta-analysis methodology applied to multiple gene expression datasets from three mouse embryonic stem cell (ESC) lines obtained at specific time points during the course of their differentiation into various lineages. We developed methods to identify genes with expression changes that correlated with the altered frequency of functionally defined, undifferentiated ESC in culture. In each dataset, we computed a novel statistical confidence measure for every gene which captured the certainty that a particular gene exhibited an expression pattern of interest within that dataset. This permitted a joint analysis of the datasets, despite the different experimental designs. Using a ranking scheme that favored genes exhibiting patterns of interest, we focused on the top 88 genes whose expression was consistently changed when ESC were induced to differentiate. Seven of these (*103728_at, 8430410A17Rik, Klf2, Nr0b1, Sox2, Tcl1,* and *Zfp42*) showed a rapid decrease in expression concurrent with a decrease in frequency of undifferentiated cells and remained predictive when evaluated in additional maintenance and differentiating protocols. Through a novel meta-analysis, this study identifies a small set of genes whose expression is useful for identifying changes in stem cell frequencies in cultures of mouse ESC. The methods and findings have broader applicability to understanding the regulation of self-renewal of other stem cell types.

Citation: Glover CH, Marin M, Eaves CJ, Helgason CD, Piret JM, et al. (2006) Meta-analysis of differentiating mouse embryonic stem cell gene expression kinetics. PLoS Comput Biol 2(11): e158. doi:10.1371/journal.pcbi.0020158

Introduction

Various types of stem cells are now recognized as responsible both for the generation of tissues and organs during embryonic development and also for the subsequent maintenance and repair of these tissues and organs throughout adult life. This has led to considerable interest in the potential of these stem cell populations to be exploited as cellular therapies for medical conditions where tissue damage or malfunction is severe and irreversible. The clinical realization of stem cell–based therapies will, however, rely on the development of robust, scalable methods for the ex vivo expansion and controlled manipulation of these cell populations. Development of such protocols requires extensive testing of multiple factors and culture conditions [1], but this is currently inhibited by the lack of rapid endpoints of stem cell frequency that can be used in high-throughput assays.

The specific identification of most stem cell types currently relies on the use of functional assays to detect their developmental potential, either in vitro or in vivo [2]. Such assays are thus, by their very nature, retrospective, protracted, cumbersome, and labor-intensive. These features make such assays impractical for large-scale studies and rapid screening methodologies. Monitoring critical changes in gene expression using either microarray or high-throughput quantitative reverse transcription PCR (Q–RT–PCR) platforms offers a potentially attractive solution but depends on the identification of a set of genes whose expression changes predict decreased stem cell frequency with adequate precision and specificity.

Recently, several groups have described differences in the gene expression profiles of several types of stem cells and their differentiating progeny [3–8]. Most of these investigations have resulted in lists of genes that are too large for comprehensive assessment of their functional significance or specificity. Moreover, many have focused on the detection of altered patterns of gene expression that are more likely to be indicative of emerging differentiated lineages than of altered transcription of genes responsible for sustaining the stem cell state. In many cases, the actual stem cell content of the

Editor: Wei-Shou Hu, University of Minnesota, United States of America

Received March 9, 2006; **Accepted** October 6, 2006; **Published** November 24, 2006

Abbreviations: AA, ascorbic acid; CFC, colony forming cell; CV, confidence value; EB, embryoid body; ESC, embryonic stem cell(s); LIF, leukemia inhibitory factor; M–LR, multiple cell line LIF removal; PF, Pareto front; PFA, Pareto front analysis; Q–RT–PCR, quantitative reverse transcription PCR; R1–LR, R1 LIF removal; RA, retinoic acid

* To whom correspondence should be addressed. E-mail: jenny@stat.ubc.ca

Synopsis

Stem cells are able to develop into many specialized cell types and thus have the potential to be used to repair or replace damaged cells. One of the challenges that scientists face is learning how to multiply these cells in vitro without loss of their stem cell properties. The development of more rapid assays for stem cells in cultured populations would significantly aid the optimization of culture conditions for stem cells. The authors propose an assay for mouse embryonic stem cells based on the expression change of seven marker genes and show that it can detect both increases and decreases in the frequency of stem cells. The assay was developed by analyzing three independent microarray datasets that ask similar biological questions but use different experimental designs. Gene expression profiles were identified within each dataset that exhibited patterns consistent with loss of stem cell properties, and, using a novel statistical measure, these profiles were compared between datasets in an unbiased fashion. A similar experimental design could be used to develop other stem cell population assays, and the analytical methods are adaptable to unrelated biological questions where analysis of a diverse set of microarray experiments is useful.

population was insufficient to infer any changes in the stem cell subset. Mouse embryonic stem cells (ESC) are less problematic in this regard because of their ease of propagation in culture as a predominantly undifferentiated population [9,10] and the availability of well-defined protocols for inducing their rapid differentiation into particular lineages. A few genes that are important to the maintenance of the pluripotent status of mouse ESC, such as *Oct4* [11] and *Nanog* [12], have been identified. However, recent studies of the rate at which functional measures of stem cell frequency of mouse ESC are lost indicated that these occur well before changes in *Oct4* expression are initiated [13]. The goal of this study was to identify a robust set of early gene expression changes that would be reliable indicators of decreased pluripotent cell content in mouse ESC cultures, regardless of the differentiation stimulus applied. In the following, the ESC signature change is defined as a set of gene expression changes that are indicative of ESC loss from a population.

To identify candidates for inclusion in the ESC signature change, we sought genes that exhibited a pattern of expression consistent with functional assay output in three independently generated datasets from ESC-derived cell populations that had been treated for up to 96 h in several ways to induce differentiation. This strategy required innovation in statistical methodology since the ESC signature change is more complicated than simple differential expression. Here, we present a statistically rigorous approach where the probability that a gene exhibits a predetermined expression pattern is estimated using a semiparametric bootstrap. The definition of the ESC signature change was specific to each experimental context, and, therefore, we obtained from each dataset an objective summary of the evidence that a gene is part of the ESC signature change. Genes that exhibited the strongest evidence across all three datasets were then tested in other maintenance or differentiating conditions and shown to successfully predict functional assay readout, indicating their potential to be used as an assay in high-throughput screening experiments.

Results

Gene Expression Datasets

Gene expression data was obtained at several time points in three independent experiments in which various differentiation induction protocols were applied to three mouse ESC lines. Summaries of each of the experiments are shown in Table 1.

DMSO and retinoic acid-induced differentiation of R1 cells. Independent, duplicate cultures of R1 cells [14] were cultured for 96 h with leukemia inhibitory factor (LIF) ± 2 μM retinoic acid (RA) or without LIF + 1% DMSO. The DMSO/RA dataset samples were hybridized to Affymetrix MOE430 A and B Genechips (Table 1).

Induction of R1 cell differentiation by LIF removal. The data for R1 LIF removal (R1–LR dataset) have been described in detail previously [13]. Briefly, R1 cells were cultured in the absence of LIF for 0, 18, and 72 h. RNA for the 18-h sample was generated from cells cultured in suspension, while RNA for the 72-h sample was generated from cells cultured in methylcellulose-containing medium. Samples were generated independently in triplicate and hybridized to Affymetrix MGU74v2 A, B, and C Genechips (Table 1).

Induction of R1, J1, and V6.5 differentiation by LIF removal. The multiple cell line LIF removal (M–LR) dataset is available from StemBase (http://www.scgp.ca:8080/StemBase/) [15]. R1, J1 [16], and V6.5 [17] cells were transferred onto 0.1% gelatin-coated dishes for 48 h with LIF prior to inoculation in petri dishes in the absence of LIF and RNA extracts obtained from 0 to 96 h later. RNA samples were generated independently in triplicate and hybridized to MOE430 A and B Affymetrix Genechips. (Table 1).

Functional Assay Analysis

To define the time course of changes in the biological properties of ESC subjected to the differentiation protocols used for gene expression analyses, R1 ESC were cultured on 0.1% gelatin-coated tissue-culture dishes with LIF ± RA or without LIF ± DMSO, and aliquots were sequentially tested in two-colony assays for undifferentiated cell activity The colony-forming cell (CFC) assay performed in liquid cultures containing LIF and the embryoid body (EB)-forming cell assay performed in a semisolid medium without LIF (Figure

Table 1. Summary of All Microarray Experiments Used in This Study

Experiment Details	DMSO/RA	R1–LR	M–LR
Condition	1% DMSO/no LIF, 2 μM RA/10 ng/mL LIF, 10 ng/mL LIF (positive control)	LIF removal	LIF removal
Cell line	R1	R1	R1, J1, V6.5
Timepoint (h)	96	0, 18, 72	0, 6, 12, 18, 24, 36, 48, 96
Number of replicates per condition (total number of chips)	2 (6)	3 (9)	3 (72)
Affymetrix chip	MOE430	MG_U74v2	MOE430

Note that all replicates are biological.

doi:10.1371/journal.pcbi.0020158.t001

1A and 1B). The loss of these activities more closely parallels the loss of stem cell activity measured by contribution to chimeric mice than the loss of expression of SSEA-1 or Oct4 [13]. In the starting population, 20.9 ± 0.2% of the R1 cells were detectable as CFC and 11.2 ± 0.4% gave rise to EB. Exposure to RA had the fastest effect, causing a reduction of both these values ~20-fold within 24 h, whereas simply removing LIF (with or without DMSO addition) caused a corresponding reduction of these values ~6-fold and 12-fold in the same time frame. After 96 h, CFC and EB-forming cell frequencies were less than 1% in all treated cultures. In control cultures the frequencies of both CFCs and EB-forming cells were sustained at half of the starting value, as noted previously when R1 cells were transferred from cultures containing feeders to gelatin-coated dishes [13].

To verify that each treatment induced cells to differentiate towards different lineages, we used Q–RT–PCR to monitor changes in transcript levels for five differentiation markers in samples obtained after 96 h of treatment with the three differentiation protocols (Figure 1C). All changes were statistically significant relative to the +LIF condition unless otherwise stated. As expected, removal of LIF, with or without DMSO treatment, induced the increased expression of genes associated with ectoderm, neural, and mesoderm differentiation (*Fgf5* [18], *Nestin* [19], and *brachyury* [20], respectively), but had little effect on the expression of genes associated with endodermal differentiation (*Foxa2* and *Sox17* [21]). In contrast, treatment with RA strongly induced the markers of neural and endodermal differentiation, but decreased the expression of *brachyury* (mesodermal differentiation) and had no significant effect on *Fgf5* expression (ectodermal differentiation). Overall, all treatments generated mixed populations of cells.

Gene Expression Analysis

For the data from each experiment we applied a gene-specific linear model to separate the observed expression into a level for that gene under a reference condition (e.g., +LIF or time 0) plus effects due to treatment and random fluctuation due to biological variability and measurement error. For example, to analyze the DMSO/RA dataset, the following model was used:

expression = expression in " + LIF"

+change due to treatment (DMSO or RA)

+noise

In this case, the three model parameters of primary interest were: (1) the change attributable to the effect of DMSO, (2) the change attributable to the effect of RA, and (3) the typical magnitude of the noise. These changes can be visualized easily by plotting parameter estimates on a "transcriptome plot" where each gene is represented by a single point (Figure 2A–2C). For the data from the DMSO/RA experiments, most of the genes in such a plot were found to lie close to the origin (Figure 2A), indicative of their unaltered expression following either treatment. However, it is interesting to note that for those genes whose observed expression change was greater than two in either treatment (463 genes), both treatments appeared to have similar effects, as indicated by the fact that 98% of these were either increased (209 genes, top right

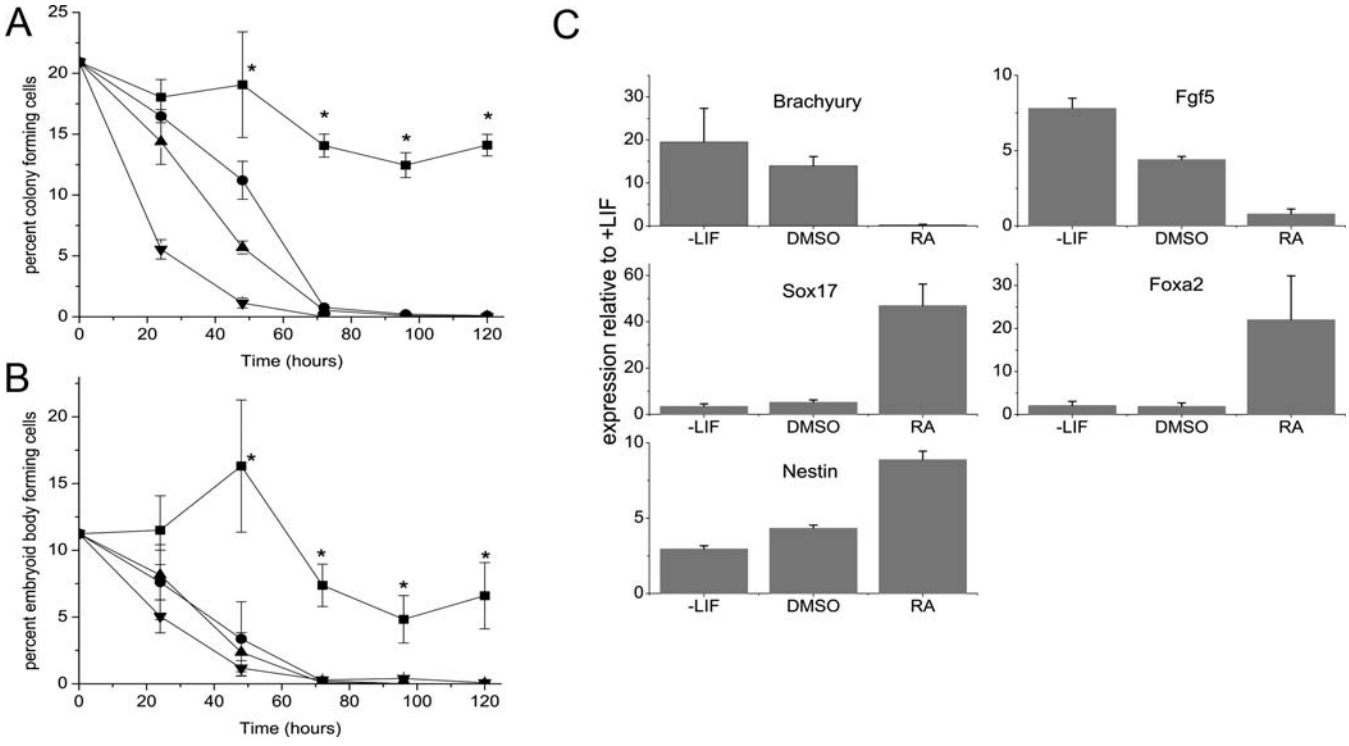

Figure 1. The Effect of LIF Removal with or without Addition of DMSO or RA on the Maintenance and Differentiation of ESC

(A) CFC frequency and (B) EB-forming ability of cells from the ESC cultures assessed at varying times after initiation of the treatment (+LIF controls = ■, –LIF = ●, –LIF + DMSO = ▲, +LIF + RA = ▼). * denotes the data for the +LIF sample are significantly different from all other treatments. (C) Gene expression of differentiation markers was monitored by Q–RT–PCR after 96 h of treatment. Results shown are relative to the +LIF control cells.

doi:10.1371/journal.pcbi.0020158.g001

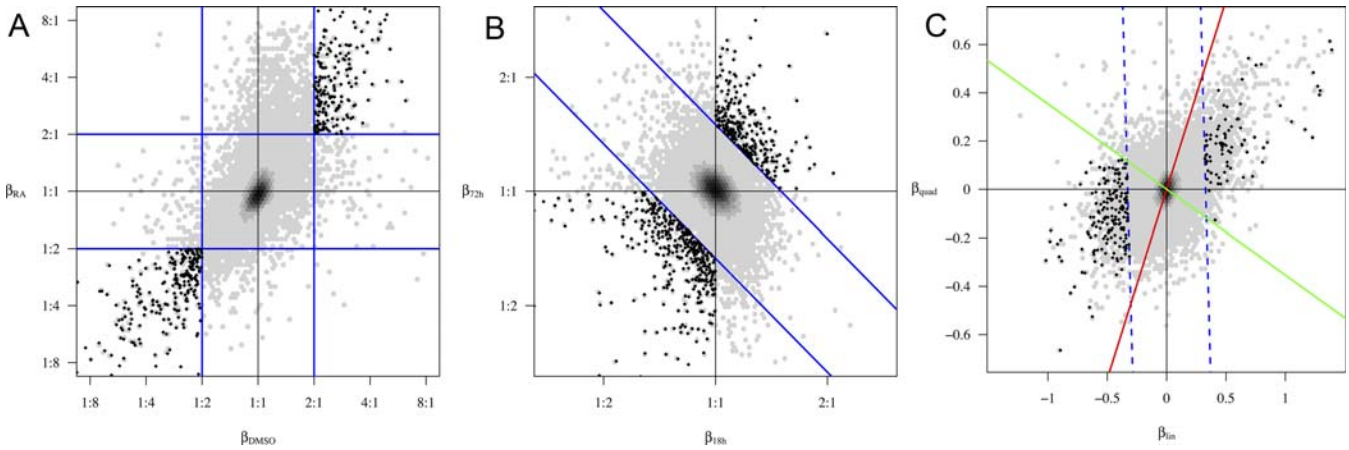

Figure 2. Transcriptome Plots of Estimated Expression Changes, Based on Fitting Models to Each Dataset

All plots have density shading to demonstrate the number of points (genes) in different regions. Lines illustrate examples of some of the requirements that make up the definition of the ESC signature change and observed gene expression patterns that fulfilled all requirements are marked as ♦. Experiment-specific implementations of requirements are explained below.

(A) DMSO/RA dataset. The requirement for large absolute changes is illustrated by the solid blue lines. Consistency across conditions implied that genes must exhibit a change in the same direction in both treatments (bottom left or top right quadrant).

(B) R1–LR dataset. Note that the *y*-axis is the change seen at 72 h relative to that seen at 18 h. The requirement for large absolute changes is illustrated by the solid blue lines. The criterion for consistency was applied by requiring that the change 18 h after LIF removal be in the same direction as that after 72 h (i.e., in the lower left or upper right quadrants), regardless of its magnitude

(C) M–LR dataset. The requirement for large absolute changes is illustrated by the dashed blue lines. To meet the consistency criterion, we required that a temporal gene expression trend either increase or decrease continuously over the duration of the experiment. This requirement was relaxed slightly to retain trends with a direction change occurring either very early (red line) or very late (green line).

doi:10.1371/journal.pcbi.0020158.g002

quadrant) or decreased (243 genes, bottom left quadrant) by both treatments. A similar model was fit to the R1–LR dataset to obtain estimates of expression changes 18 and 72 h after the removal of LIF, as well as a measure of the noise when this differentiation induction protocol was used (Figure 2B).

For the M–LR dataset, use of an ANOVA-type model, such as those described above, would have resulted in a large number of model parameters. Since interpretability of the parameters is so important in our context, we preferred a smaller, smoother quadratic model based on time. This model was able to capture the temporal trends of expression changes, and principal component analysis strongly suggested that a linear combination of constant, linear, and quadratic terms explained almost all of the data variability (Figure 2C).

Identification of a Robust Set of Early Gene Expression Changes That Indicate Decreased Frequency of Undifferentiated ESC

We defined the ESC signature change as a set of gene expression changes that were associated with decreased frequencies of ESC as indicated by functional assay readouts during ESC culture. To identify genes that exhibit patterns consistent with the ESC signature change, we imposed three requirements on each dataset. When customized to a specific experimental context, these requirements constitute an expression-based definition of the ESC signature change, a prerequisite for developing an appropriate statistical procedure.

The three requirements used to select expression changes for inclusion were: (1) large change in absolute magnitude, (2) consistent change for all treatments and cell lines, and (3) large change relative to gene-specific variability. The first two requirements can be visualized as retaining expression changes falling in certain regions of the transcriptome plots shown in Figure 2 (namely, those regions containing black points). The third requirement cannot be visualized directly in a transcriptome plot, but its effect is revealed by the fact that some expression changes in the highlighted regions are not retained, due to large gene-specific variance. Applications of these requirements are shown in Figure 2A–2C, with full explanation detailed in Materials and Methods. Specific values of the thresholds used for each dataset are shown in Table 2.

Confidence Values

One way to detect ESC signature change genes is to identify those whose observed expression patterns fall in regions of interest in transcriptome plots (Figure 2). However, this approach ignores the biological and technical noise contained in the observed data. Furthermore, it fails to distinguish between genes whose observed expression changes barely fulfill our requirements from those that substantially exceed the specified thresholds. For genes of the latter type, we have more confidence that their true, long-run expression patterns are compatible with our definition of the ESC signature change. We therefore decided to define and exploit a probabilistic quantity that measured our confidence, given the observed data, that a gene exhibits an expression pattern consistent with the ESC signature change [22,23]. Within each experiment, we defined a quantity p_g for each gene g: the probability that, in a hypothetical repeat of the experiment, the observed expression change of this gene would fulfill our requirements. Genes with true expression changes that substantially exceed all relevant thresholds have a p_g greater than those that barely fulfill the requirements. If two genes shared common expression changes but differed with respect to their background variability, the p_g of the gene with less variability would be greater. Also, as biological

Table 2. Thresholds Used in the Definition of the ESC Signature Change, in Terms of Gene Expression Changes

Parameter	DMSO/RA	R1–LR	M–LR
C_{abs}	2.0, 2.3, 3.0	1.5, 1.7, 2.0	1.4, 1.8, 2.2
C_{min}			0.4 (12 h)
C_{max}			0.9 (68 h)
C_{rel}	1.84 (DMSO), 2.72 (RA)	1.22	1.57
Number of Genes, CV > 0.5	265	269	240

doi:10.1371/journal.pcbi.0020158.t002

replication increases, the p_g of true ESC signature change genes approach 1 and those of all other genes approach 0. Just as p-values are used to rank genes with respect to differential expression, we used p_g to rank genes with respect to their consistency with the ESC signature change. Note that genes of primary interest have a p_g value near 1, not near 0, as is the case with p-values.

By definition, knowing p_g requires knowledge of the true change in expression following induction of differentiation, which is not available. Therefore, we estimated p_g by calculating the proportion of bootstrap datasets in which gene g exhibited data that fulfilled our requirements [23] and referred to this quantity as the confidence value (CV) [24]. All CVs are given in Tables S1 and S2. The methods used to generate the bootstrap data are described in the Materials and Methods and a more detailed explanation is contained in Protocol S1.

Meta-Analysis via Pareto Optimization

After calculating CVs for all genes in the three experiments, we conducted a meta-analysis to identify the gene expression changes across all datasets most compatible with the ESC signature change. Genes with expression changes most correlated with decreased frequency of ESC pluripotency would have CVs near 1 in all three experiments. If we were working with only one dataset, we could rank the genes by CV in decreasing order. However, with CVs arising from two or more datasets, the task of ranking becomes considerably more difficult. In fact, it is only possible to partially order the genes, and we accomplished this with Pareto Front Analysis (PFA) [25]. Briefly, in PFA, a comparison is made between all pairs of genes and gene g is said to dominate gene k if, in all experiments, the CVs of gene g are greater than or equal to those of gene k, with strict inequality in at least one experiment. The set of genes not dominated by any others is called the first Pareto front (PF) and contains the most promising candidates for the ESC signature change. This set is removed from the analysis and the same principle of nondomination is then used to derive successive PFs (Figure 3). A more detailed explanation of PFA can be found in [25]. The first five PFs identified changes in expression of 89 probesets representing 88 genes (10, 7, 17, 27, and 28 probesets on PFs 1 to 5, respectively). The genes on PFs 1 and 2 are shown in Table 3 and the additional three are listed in Table S3.

Q–RT–PCR of Array Results

Experiments were undertaken to test, by an independent strategy, the consistency of the candidates identified from the analysis with the definition of the ESC signature change. Accordingly, RNA extracts were prepared from R1 and J1 cells cultured for 0, 24, 72, and 96 h with LIF ± RA or without LIF ± DMSO. Q–RT–PCR was used to measure the changes in levels of 22 selected transcripts (relative to the cells cultured in the presence of LIF). Nine of these were for genes in the first PF (*103728_at, Esrrb, Nr0b1, Tcl1, Hck, Gbx2, Klf2, Fbxo1,5* and *Spp1*), four for genes in the second PF (*Tcfcp2l1, 8430410A17Rik, Zfp42, Klf4*), five for genes in the third PF (*Sox2, Jam2, Morc, Podxl, Sod2*), two in the fourth PF (*Nr1d2, Kit*) and two in the fifth PF (*Mtf2, Nmyc1*). These genes were purposefully chosen to have both high and moderate confidence in their ESC signature change membership (i.e., from the first to the fifth PFs). This tested the breadth of the correlation between the Q–RT–PCR and array results across the complete set of genes contained in the first five PFs.

Q–RT–PCR results were compared with their corresponding array data, except in the R1–LR dataset where the 24-hr Q–RT–PCR results were compared with the 18-h array data. All Q–RT–PCR data is shown in Table S4 and all comparisons of the datasets for matched treatments are plotted in Figure 4 (see Figure 4A and 4B). Results from the Q–RT–PCR measurements and the microarray analyses were strongly correlated in both cell lines (R1 cell line: Figure 4A, $r = 0.76$; J1 cell line: Figure 4B, $r = 0.82$), although the individual changes in gene expression measured by Q–RT–PCR were generally larger than those apparent from the microarray data. There was also a strong correlation between the data obtained for the two different cell lines tested ($r = 0.86$, Figure 4C). Overall, of the 22 genes tested, 18 demonstrated kinetics consistent with array results when assessed using Q–RT–PCR (Table S4).

In particular, seven genes evaluated (i.e., from the first PF: *103728_at, Klf2, Nr0b1,* and *Tcl1;* from the second PF: *8430410A17Rik* and *Zfp42;* and *Sox2* from the third PF) showed rapid (within 24 h) and sustained changes in expression in both ESC lines in all differentiation induction protocols (Figure 5). These seven genes were tested for their ability to predict the time course of functional changes in populations of ESC treated with another differentiation protocol, i.e., exposure to 50 μg/mL ascorbic acid (AA) in the absence of LIF, a treatment reported to promote the generation of cardiac myocytes from undifferentiated ESC [26]. Accordingly, R1 ESC were cultured for 5 d on 0.1% gelatin-coated tissue-culture dishes in standard maintenance conditions and with AA ± LIF and changes in gene expression compared with the loss of EB-forming potential.

As expected, EB potential decreased to 40% of its starting value in the first 24 h after transferring the cells to the control (+LIF) conditions (without feeders) and then stayed constant over the remaining 5 d of the experiment (Figure 6, see Figure 6A). Cells cultured with AA in the absence of LIF showed a rapid decrease in EB potential to almost undetectable levels by day 3. Interestingly, in the presence of AA and LIF, there was an enhanced yield of EB-forming cells (with a doubling of the proportion of EB-forming cells when compared with the control +LIF conditions).

Figure 6B shows the time course of changes in transcript levels for the seven genes that had previously been identified as showing rapid changes in expression in all tested differentiation conditions. It can be seen that all were reduced in the cells exposed to AA in the absence of LIF, consistent with the concordant rapid loss in EB potential. Moreover, when AA was added in the presence of LIF, the level of expression

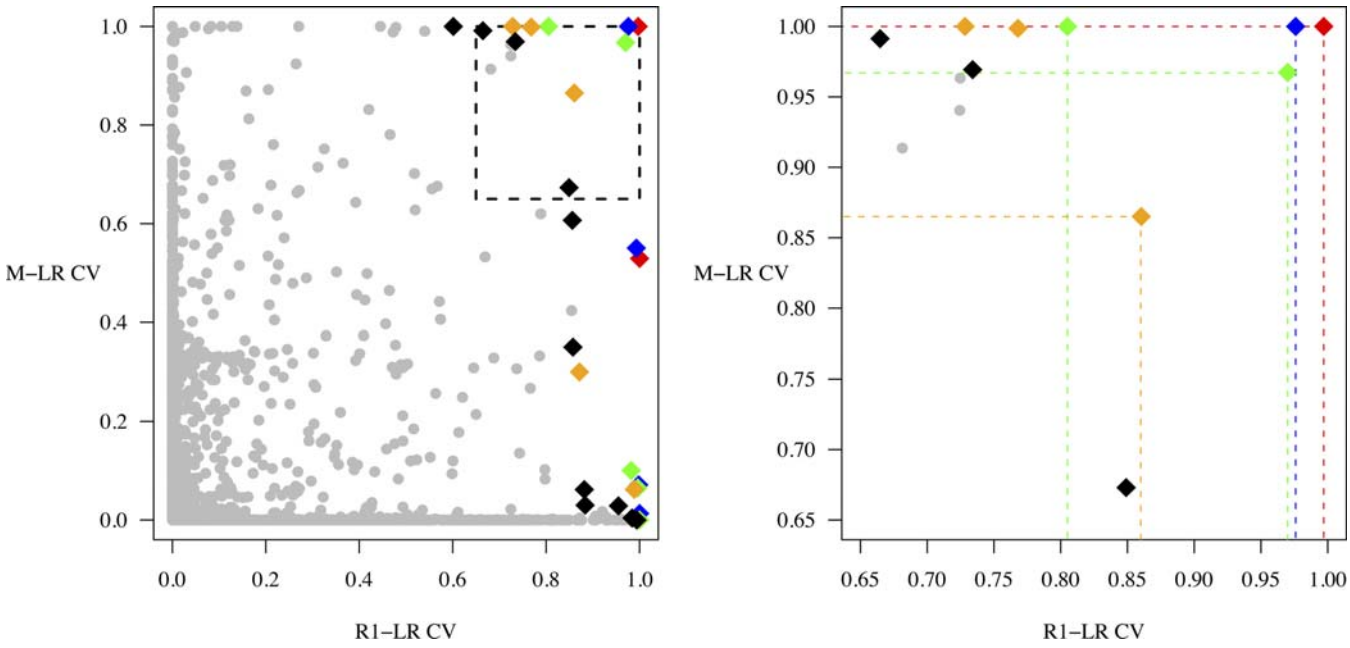

Figure 3. Two-way Pareto Front Analysis applied to CVs from the R1–LR Dataset and the M–LR Dataset

(Left) Shows the CVs for all comparable genes with the first five Pareto fronts highlighted (red, first PF; blue, second PF; green, third PF; gold, fourth PF; black, fifth PF).

(Right) Shows a magnification of the dashed box in the left panel. Here the red gene is said to dominate all other genes because, although it has an M–LR CV equal to that of several other genes, it has the highest R1–LR CV. Thus it lies on the first PF. In the same way, the blue gene dominates all genes except the red gene and thus lies on the second PF. It is not possible to choose between the two green genes because they each have larger CVs in one of the two experiments. They, therefore, lie on the same (third) PF. The highlighted yellow gene does have a larger CV in the R1–LR dataset than the green gene on the left but it falls on a lower PF because it is completely dominated by the green gene on the right.

doi:10.1371/journal.pcbi.0020158.g003

of these genes increased relative to the +LIF control cells, consistent with their opposite biological response to the AA treated cells in the absence of LIF. *Zfp42* showed the most rapid increase in expression in the AA+LIF-treated cells, and the increase in expression of *103728_at* and *Sox2* were the most delayed. Nevertheless, significantly increased expression of all seven genes relative to the +LIF controls was seen by 96 h. Together, these results indicate that changes in expression of these seven genes can be used to infer concordant functional changes in populations of ESC in culture.

Discussion

In this work, we have identified a small set of genes that exhibit the ESC signature change, i.e., whose altered expression is consistently and temporally correlated with an altered frequency of functionally defined, undifferentiated cells in ESC cultures. This result is important because we had previously found that significant changes of more established molecular markers of undifferentiated mouse ESC (Oct4 and SSEA-1) may not occur until well after the biological hallmarks of these cells have been lost [13]. By undertaking an integrated analysis of gene expression changes induced by exposure of ESC to multiple differentiation stimuli and the use of objective statistical methods, we identified seven genes whose altered expression correctly predicted concomitant functional changes induced by other treatments. Importantly, the expression changes of these genes reflected both decreased and increased frequency of ESC.

Four of these seven genes have been shown previously to be involved in the maintenance of mouse ESC or during early development. These include *Nr0b1* [27], *Sox2* [28], *Tcl1* [29], and *Zfp42* [30]. Among the remaining 81 genes on the first five PFs, an additional 11 have been reported to be involved in some aspect of development (see Table S5). Most notably *Hck* [31], *Fbx015* [32], *Dppa3/Stella* [33], and *Klf4* [34] have all been specifically implicated in maintenance or differentiation of ESC, while expression of *Eed* [35] is required for embryonic viability before implantation. Interestingly, *Oct4* [11] and *Nanog* [12] were not included in the first five PFs. While *Oct4* was differentially expressed in both the DMSO/RA and M–LR datasets, it was not changed in the R1–LR dataset as shown previously [13]. *Nanog* was differentially expressed in both the R1–LR and DMSO/RA datasets but did not show any change in the M–LR dataset.

Several previous microarray studies have been performed to uncover signalling pathways and regulatory factors required for the maintenance of human and mouse ESC [3–8,36,37]. These have each uncovered large numbers of genes whose expression was increased or absent in undifferentiated cells and, in some cases, little overlap has been found between the genes thus affected [38,39]. As a further step towards assessing the validity of our identified genes, we compared our data with two previously published datasets that sought to identify genes uniquely expressed in mouse ESC [4,5]. Of the 88 genes highlighted here, 75 were also changed in the other datasets (Table S6). This high degree of correspondence supported the validity of our very different approach to a similar biological question. Interestingly, a comparison of our results to genes whose expression has been reported to accompany the differentiation of human ESC [3,6,36,37],

Table 3. Genes Identified on the First Two Pareto Fronts

Pareto Front	MGU74v2	MOE430	CV			Fold Change at 96 h (DMSO/RA)		Fold Change Following LIF Removal (R1–LR)		Fold Change Following LIF Removal for 96 h (M–LR)			Gene Name	Gene Symbol
			DMSO/RA	R1–LR	M–LR	DMSO	RA	18 h	72 h	J1	R1	V6.5		
Pareto Front 1	103728_at	1456521_at	1.00	1.00	0.07	0.18	0.19	0.65	0.24	0.58	0.64	0.29	Transcribed locus	—
	168508_at	1436926_at	1.00	0.86	0.35	0.09	0.09	0.63	0.20	0.36	0.61	0.24	Estrogen-related receptor, beta	*Esrrb*
	93141_at	1417760_at	1.00	0.81	1.00	0.06	0.08	0.84	0.10	0.09	0.12	0.08	Nuclear receptor subfamily 0, group B, member 1	*Nr0b1*
	93296_at	1422458_at	1.00	0.98	0.10	0.19	0.19	0.78	0.22	0.48	0.73	0.42	T-cell lymphoma breakpoint 1	*Tcl1*
	93483_at	1449455_at	1.00	0.85	0.67	0.18	0.17	0.89	0.15	0.36	0.44	0.50	Hemopoietic cell kinase	*Hck*
	94200_at	1420337_at	0.92	0.98	1.00	0.23	0.30	0.69	0.22	0.26	0.29	0.25	Gastrulation brain homeobox 2	*Gbx2*
	96109_at	1448890_at	1.00	1.00	0.01	0.11	0.11	0.74	0.41	0.49	0.78	0.31	Kruppel-like factor 2 (lung)	*Klf2*
	96162_at	1427238_at	0.83	1.00	0.53	0.11	0.33	0.63	0.10	0.53	0.51	0.18	F-box protein 15	*Fbxo15*
	97519_at	1449254_at	0.02	0.99	0.55	0.08	0.77	0.54	0.19	0.54	0.44	0.10	Secreted phosphoprotein 1	*Spp1*
	99561_f_at	1448393_at	0.00	1.00	1.00	5.23	0.67	1.34	4.15	4.67	7.25	3.34	Claudin 7	*Cldn7*
Pareto Front 2	103761_at	1418091_at	1.00	1.00	0.07	0.19	0.26	0.64	0.13	0.60	0.66	0.22	Transcription factor CP2-like 1	*Tcfcp2l1*
	108097_at	1450626_at	0.93	0.80	0.10	0.15	0.29	0.74	0.27	0.69	0.71	0.25	Mannosidase, beta A, lysosomal	*Manba*
	108712_at	1434917_at	0.98	0.77	1.00	0.26	0.29	0.73	0.32	0.36	0.38	0.28	Cordon-bleu	*Cobl*
	160684_at	1423786_at	1.00	0.73	1.00	0.27	0.23	0.97	0.46	0.29	0.24	0.17	RIKEN cDNA 8430410A17 gene	*8430410A17Rik*
	95033_at	1426810_at	0.50	0.97	0.97	0.42	0.30	0.84	0.41	0.31	0.33	0.39	Jumonji domain containing 1A	*Jmjd1a*
	98414_at	1418362_at	1.00	0.73	0.97	0.09	0.10	0.88	0.14	0.28	0.40	0.11	Zinc finger protein 42	*Zfp42*
	99622_at	1417394_at	0.70	1.00	0.00	0.09	0.33	0.26	0.07	0.30	0.14	0.11	Kruppel-like factor 4 (gut)	*Klf4*

A complete table detailing all genes identified on the Pareto fronts is shown in Table S3.

doi:10.1371/journal.pcbi.0020158.t003

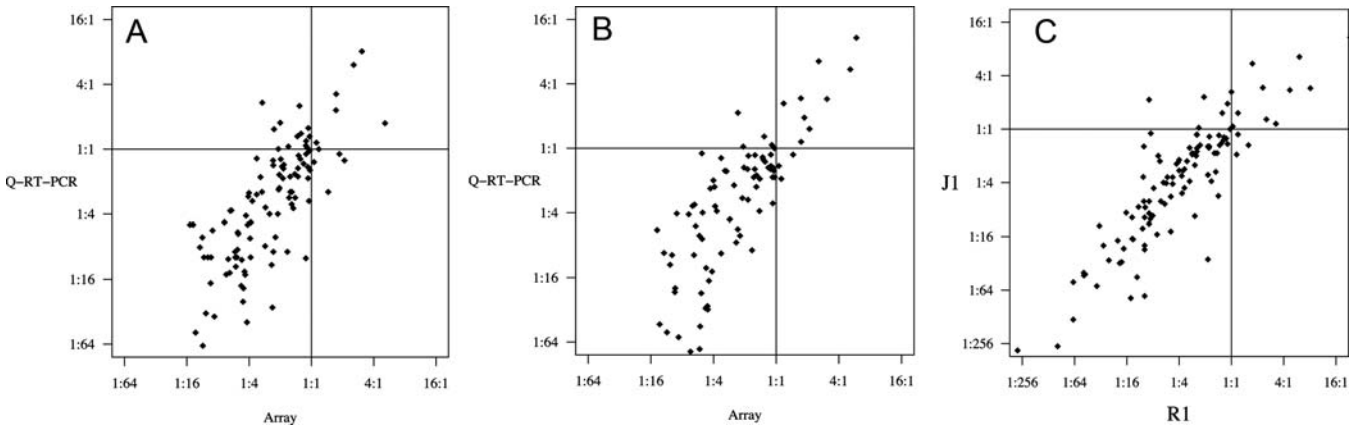

Figure 4. Comparison of Differently Measured Changes in Gene Expression within and between Two ESC Lines Grown for 0, 24, 72, and 96 h in +LIF ± RA or –LIF ± DMSO

(A) Comparison of Q–RT–PCR and microarray results for the R1 line ($r = 0.76$).
(B) Comparison of Q–RT–PCR and microarray results for the J1 line ($r = 0.82$).
(C) Comparison of results between the R1 and J1 lines ($r = 0.87$).
doi:10.1371/journal.pcbi.0020158.g004

revealed far fewer similarities, as reported by others [40]. We found that only 26 of the 88 genes exhibited similar expression changes in at least one of the four human ESC studies. Moreover, in some cases, the gene expression level changed in the opposite direction. For example, in this work, *Podxl* was found to be strongly increased as mouse ESC differentiate, whereas the human homolog was found to be decreased in three studies of human ESC differentiation [3,6,37]. Overall, only six of the 88 genes were not identified as being altered in any other published datasets of differentiating ESC.

The goal of this work was to identify a small number of genes suitable for the development of an expression-based assay to estimate the frequency of ESC in culture. To achieve this, we have taken care to seek gene expression changes that fulfill several criteria beyond simple differential expression, including large, rapid, and consistent changes in more than one cell line following the induction of differentiation using multiple methods. In terms of statistical analysis, we required (1) a quantitative index that reflected each gene's compliance with the predefined ESC signature change (for use within each dataset); and (2) a meta-analytic procedure for ranking genes based on their compliance with the ESC signature change definition in multiple datasets.

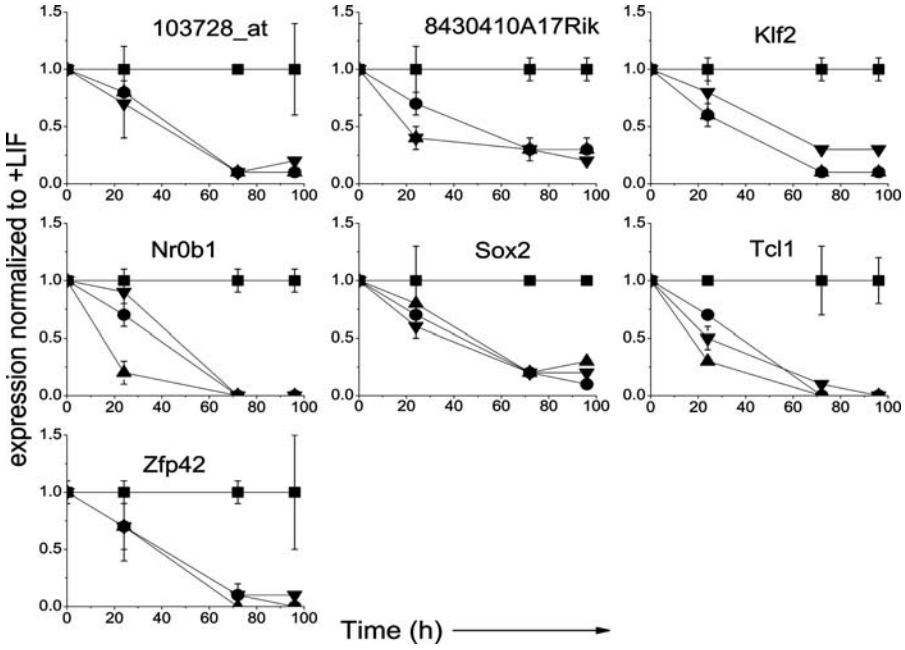

Figure 5. Q–RT–PCR Profiles of Transcript Levels for Seven Genes That Showed Rapid Decrease in ESC Populations When Subjected to a Variety of Differentiating Conditions

All levels of expression are relative to the +LIF control. +LIF control = ■, –LIF = ●, –LIF + DMSO = ▲, +LIF + RA = ▼.
doi:10.1371/journal.pcbi.0020158.g005

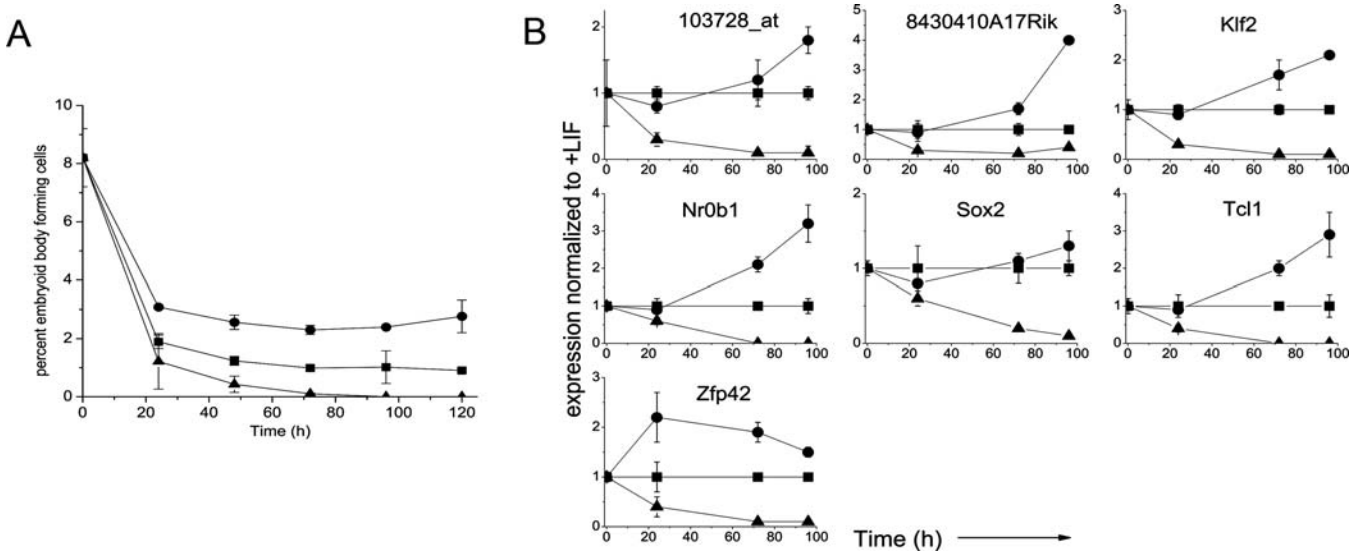

Figure 6. Comparison of Biological and Molecular Changes in ESC Stimulated to Differentiate by Exposure to AA

(A) Changes in EB frequency.

(B) Changes in the levels of transcripts for seven ESC signature change genes measured by Q–RT–PCR. +LIF (■), +LIF + AA (●), and –LIF + AA (▲).

doi:10.1371/journal.pcbi.0020158.g006

As our quantitative index, we chose the probability that a gene would exhibit the ESC signature change in a hypothetical repeat of the experiment, instead of the more conventional *p*-value, the probability that a gene would exhibit data as, or more "extreme" than, the observed change if its true expression were unchanged in the study. This choice was necessitated because in our application the ESC signature change is defined by more than a requirement for differential expression. In the majority of microarray applications, the genes of interest are characterized by differential expression, where the complementarity of the null hypothesis (no differential expression) and the biologically interesting state (differential expression) permits the *p*-value to serve as an index of biological interest. Here we employed an index that originates with an explicit definition of the biologically interesting target profile. This is an application of methods previously described for identifying genes with biologically specific expression patterns [22,23]. The task of identifying ESC signature change genes can be viewed also as an instance of the so-called "problem of regions" [24], in which the term "confidence value" is first established. In certain settings, CVs can be shown to be approximations of Bayesian a posteriori probabilities. Although not formally established here, the CV can be interpreted heuristically as the posterior probability that a gene truly exhibits the ESC signature change.

As our method of meta-analysis, we used PFA [25] to (partially) rank genes based on three independent measures of ESC signature change compliance, as opposed to the more prevalent practice of integrating experiment-specific fold-changes [41,42], *p*-values[43,44], test statistics [45,46], or effect size estimates [47,48]. In these works, the common goal is a unified list of differentially expressed genes that is accompanied by an estimated error rate, usually the false discovery rate [49,50]. The methodological choices and innovations of this work are motivated by departures from this common goal, and our techniques may prove useful in other studies where biological interest is not synonymous with differential expression. PFA was first applied to gene expression data by Fleury et al. [25]. In that work, PFA was used to optimize multiple indices within one study as opposed to our use, which is the optimization of a comparable index, the CV, across distinct but related studies. Yang et al. present another compelling technique for the synthesis of competing measures of differential expression within a single experiment [51], and it may be possible to extend their methodology for use in meta-analysis.

Meta-analysis of microarray data is an increasingly common technique to capitalize on the combined power of biologically related but distinct datasets [41–48]. In addition to the usual advantage of increasing the effective sample size, the primary benefit of meta-analysis in our application is to insulate our biological findings from confounding experimental and biological effects [44,46]. For example, in the R1–LR dataset, changes induced by differentiation could have been confounded with changes caused by the removal of feeders from the culture. However, this effect was not present in the other datasets; therefore, any common gene expression changes cannot be attributed to the removal of feeders. An example of profound differences in gene expression caused by small changes in culture conditions was reported by Skottman et al., who demonstrated effects on 1,417 genes in human ESC cultured in serum containing versus serum-free conditions, despite comparable levels of expression of other markers of their undifferentiated state [52].

In summary, meta-analysis of multiple gene expression datasets from populations of mouse ESC induced to differentiate has revealed multiple genes whose altered expression provides a robust and timely indication of changes in pluripotency. These findings suggest the importance of the products of these genes in the molecular regulation of the undifferentiated state of ESC and provide a useful basis for developing high-throughput approaches for the bio-monitoring of ESC cultures.

Materials and Methods

ESC maintenance cultures. J1 (passage 14) and R1 (passage 17) ESC lines were maintained on irradiated feeders at 37 °C in 5% CO_2 in air with daily exchange of ESC maintenance medium consisting of high glucose Dulbecco's Modified Eagles Media supplemented with 15% pre-screened fetal bovine serum (FBS), 0.1 mM nonessential amino acids, 2 mM glutamine, 100 U/mL penicillin, 100 μg/mL streptomycin, 10 ng/mL LIF (all reagents from StemCell Technologies, http://www.stemcell.com) and 100 μM monothioglycerol (MTG, http://www.sigmaaldrich.com). Cells were passaged every second day. Primary mouse embryo feeders (StemCell Technologies) were maintained at 37 °C in 5% CO_2 in air in DMEM supplemented with 10% FBS, 1 mM glutamine, 100 U/mL penicillin, 100 μg/mL streptomycin, and 100 μM MTG. Feeders were irradiated by exposure to 80 Gy 300 kvP X-rays.

ESC experimental cultures. Cells were thawed and maintained on irradiated feeders for two passages prior to initiation of differentiation cultures. To remove contaminating feeders, cells harvested from maintenance cultures were plated onto tissue culture dishes (Sarstedt, http://www.sarstedt.com/php/main.php) in maintenance medium for 1 h at 37 °C. All suspended and loosely adherent cells were harvested by gently pipetting medium onto the surface of the tissue-culture dish. Following this, feeder contamination was estimated at <1% based on cell size during counting.

Experimental cultures were performed on tissue culture dishes (Sarstedt), coated with 0.1% porcine gelatin (Sigma) with cells plated at a density of 80–1,500 cells per cm^2 depending on the day of harvest. Differentiation media were based on maintenance medium with the following differences; (a) LIF removal—maintenance medium minus LIF, (b) DMSO—maintenance medium minus LIF plus 1% DMSO (Sigma), (c) RA—maintenance medium plus 2 μM RA (Sigma), (d) AA + LIF—maintenance medium plus 50 μg/mL AA (Sigma), and (e) AA-LIF—maintenance medium without LIF plus 50 μg/mL AA. Concentrated RA was prepared at a concentration of 10 mM by dissolving 30 mg powder in 10 mL 100% ethanol and stored at 4 °C in the dark. Media was prepared by adding 10 μL of RA stock solution to 50 mL maintenance media. Cells were harvested daily for functional assay analysis (CFC and EB assays, see below) or for RNA extraction.

Colony forming cell assay. Test cells were plated at a density of 1,000–2,500 cells on gelatinized 60-mm tissue-culture dishes (with grid) in maintenance medium at 37 °C in 5% CO_2 in air. Five days later, colonies were stained for alkaline phosphatase (Kit 86-R; Sigma) and counted. Colonies were classified as differentiated (colourless), undifferentiated (pink), or mixed. Assay output was calculated as the percentage of undifferentiated colonies based on the assay seeding density.

Embryoid body assay. Test cells were plated at a density of 1,000–5,000 cells per 35-mm low-adherence petri dish in Iscove's Modified Dulbecco's Media supplemented with 15% prescreened FBS, 0.9% methylcellulose, 2 mM glutamine, and 150 μM MTG (all reagents from StemCell Technologies). EB were counted 5 d later and the frequency of EBs was calculated as the number of EBs generated per 100 ESC plated.

RNA extraction and array hybridization. Cytoplasmic RNA was extracted using the RNeasy mini kit (Qiagen, http://www1.qiagen.com). Standard Affymetrix protocols (Affymetrix, http://www.affymetrix.com/index.affx) were used to generate RNA probes from 5 μg of extracted RNA. Samples were hybridized to MOE430 A & B chips on a Genechip system (Affymetrix) at the Ottawa Genomics Innovation Centre (http://www.ottawagenomecenter.ca/) according to the manufacturer's instructions.

Gene expression analysis. In this analysis, as is common practice hybridization data for each probeset was considered independently, although we recognize that transcripts for many genes would be captured by multiple probesets. Furthermore, although the correspondence between probeset and gene is not, as a rule, one-to-one, we refer to the expression from each probeset as if it reflected the expression of one gene, unless otherwise stated.

All preprocessing, including background correction, normalization, probeset summarization, and $\log_2$ transformation, was carried out with the RMA algorithm [53] in the affy package [54] from Bioconductor [55] and processed data returned by RMA are referred to as expressions. The R code [56] for all the data analysis shown below is available from http://www.stat.ubc.ca/~jenny/webSupp/gloverSCmeta/index.html.

In the equations below we followed these conventions:

Observed intensities were denoted by $Y_{i,cond}$, where i indexed the biological replicate, i.e., $i \in \{1, \ldots, N\}$, and *cond* denoted the corresponding condition, i.e., treatment or time.

All models are gene-specific, although, for the sake of simplicity, an explicit index for gene was avoided.

Within the observations for one gene, the random errors ε were assumed to be independent and identically distributed and to have expectation zero and a finite, gene-specific variance σ^2_{exp}, where *exp* is DMSO/RA, R1–LR, or M–LR.

To summarize gene expression changes, we fit linear models to the RMA processed data. For the DMSO/RA data we used the following model:

$$Y_{i,cond} = \mu_{+LIF} + \beta_{cond} + \varepsilon_{i,cond}$$

where μ_{+LIF} was the expected intensity in the +LIF control condition, $\beta_{+LIF} = 0$ by definition, and β_{DMSO} (β_{RA}) was the effect of DMSO (RA) treatment, relative to +LIF.

For the R1–LR data we used the following model:

$$Y_{it} = \mu_{0h} + \beta_{18h} * I_{t \geq 18h} + \beta_{72h} * I_{t \geq 72h} + \varepsilon_{it}$$

where μ_{0h} was the expected intensity at time 0, β_{18h} (β_{72h}) was the effect of 18 h versus 0 h (72 h versus 18 h), and $I_{statement}$ was 1 if the statement was true and $I_{statement}$ was 0 otherwise.

For the M–LR data we used the following model:

$$Y_{it} = \mu_0 + \beta_{lin} t + \beta_{quad} t^2 + \varepsilon_{it} \quad (1)$$

where μ_0 was the expected intensity at time 0, t was $\log_2$ transformed and centered time where 0 h was changed to 3 h to avoid undefined values, and β_{lin} and β_{quad} gave the linear and quadratic effects of time, respectively.

Defining the ESC signature change in terms of gene expression. For the DMSO/RA data, a gene had to fulfill the following requirements to be included in the ESC signature change:

Absolute change: $\beta_{cond} > C_{abs}$, or $\beta_{cond} < -C_{abs}$ for $cond = DMSO$ and $cond = RA$

Change relative to variability: $\frac{|\beta_{cond}|}{\sigma} > C_{rel}$ for $cond = DMSO$ and $cond = RA$

Consistency: $\beta_{DMSO} > 0$ and $\beta_{RA} > 0$, or $\beta_{DMSO} < 0$ and $\beta_{RA} < 0$

For the R1–LR data, expression requirements were as follows:

Absolute change: $\beta_{18h} + \beta_{72h} > C_{abs}$, or $\beta_{18h} + \beta_{72h} < C_{abs}$

Change relative to variability: $\frac{|\beta_{18h} + \beta_{72h}|}{\sigma} > C_{rel}$

Consistency: $\beta_{18h} \geq 0$ and $\beta_{72h} \geq 0$, or $\beta_{18h} \leq 0$ and $\beta_{72h} \leq 0$

For the M–LR data, the definition of an interesting expression pattern was more complicated. We required a large absolute difference in the expected expression intensity between the start time, t_{min}, of the study and the end, t_{max}. Specifically, we required that

$$|E(Y_{t_{\max}}) - E(Y_{t_{\min}})| > C_{abs}$$

Given Equation 1, it can be shown that this is equivalent to the following requirement:

$$\beta_{quad} > \frac{C_{abs}}{TD_1} - \frac{TD_1}{TD_2}\beta_{lin} \text{ or } \beta_{quad} < \frac{-C_{abs}}{TD_1} - \frac{TD_1}{TD_2}\beta_{lin}$$

where $TD_k = t^k_{\max} - t^k_{\min}$. The relative expression change requirement that

$$\frac{|E(Y_{t_{\max}}) - E(Y_{t_{\min}})|}{\sigma} > C_{rel}$$

is equivalent to the following condition:

$$\beta_{quad} > \frac{C_{rel}}{TD_2}\sigma - \frac{TD_1}{TD_2}\beta_{lin} \text{ or } \beta_{quad} < \frac{-C_{rel}}{TD_2}\sigma - \frac{TD_1}{TD_2}\beta_{lin}$$

Consistency was built into the ESC signature change definition by noting the location in the time course of the vertex of the quadratic fit (Equation 1). The requirement for strictly increasing or decreasing expression patterns was relaxed by allowing genes whose vertex fell before C_{min} or after C_{max} to be retained, where C_{min} and C_{max} were specified relative to the standardized, transformed study time. This requirement is captured in the following condition:

$$\frac{\left(-\frac{\beta_{lin}}{2\beta_{quad}} - t_{\min}\right)}{TD_2} < C_{\min} \text{ or } \frac{\left(-\frac{\beta_{lin}}{2\beta_{quad}} - t_{\min}\right)}{TD_2} > C_{\max}$$

In all experiments, the specific values of the user-specified thresholds are given in Table 2. Note that the final results of this analysis are highly robust to modest differences in these thresholds.

Confidence values. One thousand simulated datasets were generated for each of the original datasets by adding the original fitted averages and a randomly sampled residual (with replacement) from the residuals associated with that gene within the dataset. Note that for the M–LR dataset, the bootstrap data was derived from time and cell line specific averages and residuals, not from quadratic fits. To

maintain the covariance between genes, the same random selection of residuals were used for all genes in a simulated dataset. The simulated data was then assessed relative to the ESC signature change definitions given above. The proportion of times that a gene fulfilled the definition, i.e., the CV, was calculated. As a practical measure, within each experiment, we used several values for C_{abs} and averaged the resulting CVs to obtain the CVs used for PFA. This was an expedient method for reducing the frequency of CVs equal to 0 or 1, which are highly undesirable for PFA. The use of multiple thresholds makes the final results of our analysis robust to modest changes. The most stringent values of C_{abs} were chosen such that the number of genes retained in each dataset was approximately equal, and a range of less stringent thresholds was applied to create further distinction among the CVs.

Comparison of MOE430 and MG_U74v2 chips. To compare gene samples hybridized to the MOE430 and MG_U74 chips, six different comparisons were used. Two comparisons were generated from the Affymetrix-defined "good" or "best" comparisons (for more information see http://www.affymetrix.com). Resourcerer [57] was used to generate lists of comparable probesets based on EGO, Unigene, Locuslink, and Refseq comparisons. A list of two-way comparable probesets was generated by ordering evidence for comparison as follows: Affymetrix Best > Affymetrix Good > EGO > Unigene = Locuslink = Refseq. In this way 21,271 one-to-one mappings between the two chips were made (Table S7).

Gene ontology. Functional information about differentially expressed genes was obtained by loading Affymetrix identifiers into the Database for Annotation, Visualization, and Integrated Discovery 2.0 (DAVID2) [58]. Gene ontologies were determined and compiled at several different levels of the ontology.

Quantitative reverse transcription-PCR. Q–RT–PCR was performed as previously described [13]. Relative expression changes were determined with the $2^{-\Delta\Delta CT}$ method [59] and the *Gapdh* transcript was used to normalize results between samples. Primers were manufactured by Invitrogen (http://www.invitrogen.com/) and sequences are shown in Table S8.

Supporting Information

Protocol S1. Example of Confidence Value Calculation

Found at doi:10.1371/journal.pcbi.0020158.sd001 (89 KB PDF).

Table S1. Confidence Values for All Genes in the R1–LR Dataset

Found at doi:10.1371/journal.pcbi.0020158.st001 (686 KB PDF).

Table S2. Confidence Values for All Genes in the DMSO/RA Dataset and M–LR Dataset

Found at doi:10.1371/journal.pcbi.0020158.st002 (859 KB PDF).

Table S3. The First Five Pareto Fronts

Found at doi:10.1371/journal.pcbi.0020158.st003 (18 Kb PDF).

Table S4. Raw Data from Q–RT–PCR

Found at doi:10.1371/journal.pcbi.0020158.st004 (8 KB PDF).

Table S5. Biological Process and Molecular Function Classification of Genes in Pareto Fronts

Found at doi:10.1371/journal.pcbi.0020158.st005 (8 KB PDF).

Table S6. Comparison of Table S3 with Other Publications

Found at doi:10.1371/journal.pcbi.0020158.st006 (61 KB PDF).

Table S7. Probeset Translations between the MG_U74v2 and MOE430 Genechips

Found at doi:10.1371/journal.pcbi.0020158.st007 (365 KB PDF).

Table S8. Primers Used for Q–RT–PCR

Found at doi:10.1371/journal.pcbi.0020158.st008 (18 KB PDF).

Accession Numbers

Accessions numbers from the Ensembl database (http://www.ensembl.org/index.html) for the genes mentioned in this paper are: *103728_at* (ENSMUST00000027649), *8430410A17Rik* (ENSMUST00000032141), *Klf2* (ENSMUST00000067912), *Nr0b1* (ENSMUST00000026036), *Sox2* (ENSMUST00000099151), *Tcl1* (ENSMUSG00000041359), *Zfp42* (ENSMUST00000082120), *Oct4* (ENSMUST00000025271), *Nanog* (ENSMUST00000012540), *Fgf5* (ENSMUST00000031280), *Nestin* (ENSMUST00000090973), *Brachyury* (ENSMUST00000074667), *Foxa2* (ENSMUST00000047315), *Sox17* (ENSMUST00000027035), *Esrrb* (ENSMUST00000021680), *Hck* (ENSMUST00000003370), *Gbx2* (ENSMUST00000036954), *Fbxo15* (ENSMUST00000037718), *Spp1* (ENSMUST00000031243), *Tcfcp2l1* (ENSMUST00000027629), *Jam2* (ENSMUST00000057513), *Morc* (ENSMUST00000023330), *Podxl* (ENSMUST00000026698), *Sod2* (ENSMUST00000007012), *Nr1d2* (ENSMUST00000090543), *Kit* (ENSMUST00000005815), *Mtf2* (ENSMUST00000081567), *Nmyc1* (ENSMUST00000043396), *Dppa3/Stella* (ENSMUST00000049644), *Klf4* (ENSMUST00000003116), *Eed* (ENSMUST00000032850), *Gdf3* (ENSMUST00000032211), *Krt1–18* (ENSMUST00000023803), *Krt1–19* (ENSMUST00000007317), *Cldn7* (ENSMUST00000018713), *Manba* (ENSMUST00000029814), *Cobl* (ENSMUSG00000020173), and *Jmjd1a* (ENSMUSG00000053470).

Accessions numbers from the ArrayExpress database (http://www.ebi.ac.uk/arrayexpress) for the microarrays mentioned in this paper are: for the DMSO/RA dataset (E-MEXP-412), and for the R1-LR dataset (E-MEXP-414).

Accession numbers from the Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo) for RNA samples mentioned in this paper are: R1 data (GSE2972), V6.5 data (GSE3231), and J1 data (GSE3749).

Acknowledgments

Author contributions. CHG, CJE, CDH, and JMP conceived and designed the experiments. CHG and CDH performed the experiments. CHG, MM, and JB analyzed the data. CDH contributed reagents/materials/analysis tools. CHG, CJE, JMP, and JB wrote the paper.

Funding. This work was supported by a Stem Cell Network grant to JMP and CJE, a Mathematics of Information Technology and Complex Systems grant to JMP, a Genome BC/Canada grant to CJE, and a Natural Sciences and Engineering Research Council grant to JB. CHG is a recipient of a Stem Cell Network trainee award and a Canadian Institutes of Health Research Doctoral Research Award. CDH and JB are Scholars of the Michael Smith Foundation for Health Research, and CDH is a Canadian Institutes of Health Research New Investigator.

Competing interests. The authors have declared that no competing interests exist.

References

1. Zandstra PW, Nagy A (2001) Stem cell bioengineering. Annu Rev Biomed Eng 3: 275–305.
2. van Os R, Kamminga LM, de Haan G (2004) Stem cell assays: Something old, something new, something borrowed. Stem Cells 22: 1181–1190.
3. Brandenberger R, Wei H, Zhang S, Lei S, Murage J, et al. (2004) Transcriptome characterization elucidates signaling networks that control human ES cell growth and differentiation. Nat Biotechnol 22: 707–716.
4. Ivanova NB, Dimos JT, Schaniel C, Hackney JA, Moore KA, et al. (2002) A stem cell molecular signature. Science 298: 601–604.
5. Ramalho-Santos M, Yoon S, Matsuzaki Y, Mulligan RC, Melton DA (2002) "Stemness": Transcriptional profiling of embryonic and adult stem cells. Science 298: 597–600.
6. Sato N, Sanjuan IM, Heke M, Uchida M, Naef F, et al. (2003) Molecular signature of human embryonic stem cells and its comparison with the mouse. Dev Biol 260: 404–413.
7. Sharov AA, Piao Y, Matoba R, Dudekula DB, Qian Y, et al. (2003) Transcriptome analysis of mouse stem cells and early embryos. PLoS Biol 1 (3): e74.
8. Tanaka TS, Kunath T, Kimber WL, Jaradat SA, Stagg CA, et al. (2002) Gene expression profiling of embryo-derived stem cells reveals candidate genes associated with pluripotency and lineage specificity. Genome Res 12: 1921–1928.
9. Martin GR (1981) Isolation of a pluripotent cell line from early mouse embryos cultured in medium conditioned by teratocarcinoma stem cells. Proc Natl Acad Sci U S A 78: 7634–7638.
10. Evans MJ, Kaufman MH (1981) Establishment in culture of pluripotential cells from mouse embryos. Nature 292: 154–156.
11. Niwa H, Miyazaki J, Smith AG (2000) Quantitative expression of Oct-3/4 defines differentiation, dedifferentiation or self-renewal of ES cells. Nat Genet 24: 372–376.
12. Chambers I, Colby D, Robertson M, Nichols J, Lee S, et al. (2003) Functional expression cloning of Nanog, a pluripotency sustaining factor in embryonic stem cells. Cell 113: 643–655.
13. Palmqvist L, Glover CH, Hsu L, Lu M, Bossen B, et al. (2005) Correlation of

murine embryonic stem cell gene expression profiles with functional measures of pluripotency. Stem Cells 23: 663–680.
14. Nagy A, Rossant J, Nagy R, Abramow-Newerly W, Roder JC (1993) Derivation of completely cell culture–derived mice from early-passage embryonic stem cells. Proc Natl Acad Sci U S A 90: 8424–8428.
15. Perez-Iratxeta C, Palidwor G, Porter CJ, Sanche NA, Huska MR, et al. (2005) Study of stem cell function using microarray experiments. FEBS Lett 579: 1795–1801.
16. Li E, Bestor TH, Jaenisch R (1992) Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. Cell 69: 915–926.
17. Eggan K, Akutsu H, Loring J, Jackson-Grusby L, Klemm M, et al. (2001) Hybrid vigor, fetal overgrowth, and viability of mice derived by nuclear cloning and tetraploid embryo complementation. Proc Natl Acad Sci U S A 98: 6209–6214.
18. Haub O, Goldfarb M (1991) Expression of the fibroblast growth factor-5 gene in the mouse embryo. Development 112: 397–406.
19. Dahlstrand J, Lardelli M, Lendahl U (1995) Nestin mRNA expression correlates with the central nervous system progenitor cell state in many, but not all, regions of developing central nervous system. Brain Res Dev Brain Res 84: 109–129.
20. Wilkinson DG, Bhatt S, Herrmann BG (1990) Expression pattern of the mouse T gene and its role in mesoderm formation. Nature 343: 657–659.
21. Kubo A, Shinozaki K, Shannon JM, Kouskoff V, Kennedy M, et al. (2004) Development of definitive endoderm from embryonic stem cells in culture. Development 131: 1651–1662.
22. Bryan J, Pollard KS, van der Laan MJ (2002) Paired and unpaired comparison and clustering with gene expression data. Stat Sinica 12: 87–110.
23. van der Laan MJ, Bryan J (2001) Gene expression analysis with the parametric bootstrap. Biostatistics 2: 445–461.
24. Efron B, Tibshirani R (1998) The problem of regions. Ann Stat 26: 1687–1718.
25. Fleury G, Hero AO, Yosida S, Carter T, Barlow C, et al. (2002) Pareto analysis for gene filtering in microarray experiments. Proceedings of the XI European Signal Processing Conference; 3–6 September 2002; Toulouse, France. pp. 165–168.
26. Takahashi T, Lord B, Schulze PC, Fryer RM, Sarang SS, et al. (2003) Ascorbic acid enhances differentiation of embryonic stem cells into cardiac myocytes. Circulation 107: 1912–1916.
27. Swain A, Narvaez V, Burgoyne P, Camerino G, Lovell–Badge R (1998) Dax1 antagonizes Sry action in mammalian sex determination. Nature 391: 761–767.
28. Avilion AA, Nicolis SK, Pevny LH, Perez L, Vivian N, et al. (2003) Multipotent cell lineages in early mouse development depend on SOX2 function. Genes Dev 17: 126–140.
29. Narducci MG, Fiorenza MT, Kang SM, Bevilacqua A, Di Giacomo M, et al. (2002) TCL1 participates in early embryonic development and is overexpressed in human seminomas. Proc Natl Acad Sci U S A 99: 11712–11717.
30. Rogers MB, Hosler BA, Gudas LJ (1991) Specific expression of a retinoic acid-regulated, zinc-finger gene, Rex-1, in preimplantation embryos, trophoblast and spermatocytes. Development 113: 815–824.
31. Ernst M, Gearing DP, Dunn AR (1994) Functional and biochemical association of Hck with the LIF/IL-6 receptor signal transducing subunit gp130 in embryonic stem cells. EMBO J 13: 1574–1584.
32. Tokuzawa Y, Kaiho E, Maruyama M, Takahashi K, Mitsui K, et al. (2003) Fbx15 is a novel target of Oct3/4 but is dispensable for embryonic stem cell self-renewal and mouse development. Mol Cell Biol 23: 2699–2708.
33. Bortvin A, Eggan K, Skaletsky H, Akutsu H, Berry DL, et al. (2003) Incomplete reactivation of Oct4-related genes in mouse embryos cloned from somatic nuclei. Development 130: 1673–1680.
34. Li Y, McClintick J, Zhong L, Edenberg HJ, Yoder MC, et al. (2005) Murine embryonic stem cell differentiation is promoted by SOCS-3 and inhibited by the zinc finger transcription factor Klf4. Blood 105: 635–637.
35. Morin-Kensicki EM, Faust C, LaMantia C, Magnuson T (2001) Cell and tissue requirements for the gene *eed* during mouse gastrulation and organogenesis. Genesis 31: 142–146.
36. Sperger JM, Chen X, Draper JS, Antosiewicz JE, Chon CH, et al. (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.
37. Bhattacharya B, Miura T, Brandenberger R, Mejido J, Luo Y, et al. (2004) Gene expression in human embryonic stem cell lines: Unique molecular signature. Blood 103: 2956–2964.
38. Evsikov AV, Solter D (2003) Comment on "Stemness": Transcriptional profiling of embryonic and adult stem cells and "a stem cell molecular signature. rdquo; Science 302: 393. And Author's Reply.
39. Fortunel NO, Otu HH, Ng HH, Chen J, Mu X, et al. (2003) Comment on "Stemness": Transcriptional profiling of embryonic and adult stem cells and "a stem cell molecular signature. " Science 302: 393. And Author's Reply.
40. Ginis I, Luo Y, Miura T, Thies S, Brandenberger R, et al. (2004) Differences between human and mouse embryonic stem cells. Dev Biol 269: 360–380.
41. Stevens JR, Doerge RW (2005) Combining Affymetrix microarray results. BMC Bioinformatics 6: 57.
42. Wang J, Coombes KR, Highsmith WE, Keating MJ, Abruzzo LV (2004) Differences in gene expression between B-cell chronic lymphocytic leukemia and normal B cells: A meta-analysis of three microarray studies. Bioinformatics 20: 3166–3178.
43. Rhodes DR, Barrette TR, Rubin MA, Ghosh D, Chinnaiyan AM (2002) Meta-analysis of microarrays: Interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer. Cancer Res 62: 4427–4433.
44. Levesque MP, Vernoux T, Busch W, Cui H, Wang JY, et al. (2006) Whole-genome analysis of the SHORT-ROOT developmental pathway in *Arabidopsis*. PLoS Biol 4 (5): e143. doi:10.1371/journal.pbio.0040143
45. Ghosh D, Barette TR, Rhodes D, Chinnaiyan AM (2003) Statistical issues and methods for meta-analysis of microarray data: A case study in prostate cancer. Funct Integr Genomics 3: 180–188.
46. Estrada B, Choe SE, Gisselbrecht SS, Michaud S, Raj L, et al. (2006) An integrated strategy for analyzing the unique developmental programs of different myoblast subtypes. PLoS Genet 2 (2): e16. doi:10.1371/journal.pgen.0020016
47. Choi JK, Yu U, Kim S, Yoo OJ (2003) Combining multiple microarray studies and modeling interstudy variation. Bioinformatics 19 (Supplement 1): i84–90.
48. Hu P, Greenwood CM, Beyene J (2005) Integrative analysis of multiple gene expression profiles with quality-adjusted effect size models. BMC Bioinformatics 6: 128.
49. Storey JD, Tibshirani R (2003) Statistical significance for genomewide studies. Proc Natl Acad Sci U S A 100: 9440–9445.
50. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. J Roy Stat Soc B 57: 289–300.
51. Yang YH, Xiao Y, Segal MR (2005) Identifying differentially expressed genes from microarray experiments via statistic synthesis. Bioinformatics 21: 1084–1093.
52. Skottman H, Stromberg AM, Matilainen E, Inzunza J, Hovatta O, et al. (2006) Unique gene expression signature by human embryonic stem cells cultured under serum-free conditions correlates with their enhanced and prolonged growth in an undifferentiated stage. Stem Cells 24: 151–167.
53. Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, et al. (2003) Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Res 31: e15.
54. Gautier L, Cope L, Bolstad BM, Irizarry RA (2004) affy—Analysis of Affymetrix GeneChip data at the probe level. Bioinformatics 20: 307–315.
55. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, et al. (2004) Bioconductor: Open software development for computational biology and bioinformatics. Genome Biol 5: R80.
56. Ihaka R, Gentleman R (1996) R: A language for data analysis and graphics. J Comput Graph Stat 5: 299–314.
57. Tsai J, Sultana R, Lee Y, Pertea G, Karamycheva K, et al. (2001) RESOURCERER: A database for annotating and linking microarray resources within and across species. Genome Biol 2: software0002.0001–0002.0004.
58. Dennis G Jr, Sherman BT, Hosack DA, Yang J, Gao W, et al. (2003) DAVID: Database for annotation, visualization, and integrated discovery. Genome Biol 4: P3.
59. Livak KJ, Schmittgen TD (2001) Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. Methods 25: 402–408.