Identification of stable reference genes in differentiating human pluripotent stem cells

Gustav Holmgren,1,2 Nidal Ghosheh,1,2 Xiamin Zeng,3 Yalda Bogestal,1 Peter Sartipy,1,4 and Jane Synnergren1

1Systems Biology Research Center, University of Skövde, Skövde, Sweden; 2Department of Clinical Chemistry/Transfusion Medicine, Sahlgrenska University Hospital, Gothenburg, Sweden; 3Buck Institute for Research on Aging, Buck Institute, Novato, California; and 4AstraZeneca Research and Development, Global Medicines Development, Cardiovascular and Metabolic Diseases Global Medicines Development Unit, Mölndal, Sweden

Submitted 8 December 2014; accepted in final form 31 March 2015

Holmgren G, Ghosheh N, Zeng X, Bogestal Y, Sartipy P, Synnergren J. Identification of stable reference genes in differentiating human pluripotent stem cells. Physiol Genomics 47: 232–239, 2015. First published April 7, 2015; doi:10.1152/physiolgenomics.00130.2014.—Reference genes, often referred to as housekeeping genes (HKGs), are frequently used to normalize gene expression data based on the assumption that they are expressed at a constant level in the cells. However, several studies have shown that there may be a large variability in the gene expression levels of HKGs in various cell types. In a previous study, employing human embryonic stem cells (hESCs) subjected to spontaneous differentiation, we observed that the expression of commonly used HKG varied to a degree that rendered them inappropriate to use as reference genes under those experimental settings. Here we present a substantially extended study of the HKG signature in human pluripotent stem cells (hPSC), including nine global gene expression datasets from both hESC and human induced pluripotent stem cells, obtained during directed differentiation toward endoderm-, mesoderm-, and ectoderm derivatives. Sets of stably expressed genes were compiled, and a handful of genes (e.g., EID2, ZNF324B, CAPN10, and RABEP2) were identified as generally applicable reference genes in hPSCs across all cell lines and experimental conditions. The stability in gene expression profiles was confirmed by reverse transcription quantitative PCR analysis. Taken together, the current results suggest that differentiating hPSCs have a distinct HKG signature, which in some aspects is different from somatic cell types, and underscore the necessity to validate the stability of reference genes under the actual experimental setup used. In addition, the novel putative HKGs identified in this study can preferentially be used for normalization of gene expression data obtained from differentiating hPSCs.

housekeeping genes; reference genes; endogenous controls; human pluripotent stem cells; differentiation; gene expression; normalization

GENE EXPRESSION ANALYSIS is central in basic and applied biomedical research, with reverse transcription quantitative PCR (RT-qPCR) analysis as the far most common method of choice for accurate expression profiling of genes. Consistently increasing sensitivity, reproducibility, and dynamic range of this platform technology raises the requirements for proper internal reference genes that are appropriate to use for normalization purposes (52).

Housekeeping genes (HKGs) are typically constitutively expressed genes and are thus expected to display a stable expression pattern in different cell types (11, 16, 23). Ideally, the expression of HKGs should not vary under various experimental conditions or as a consequence of cell handling. Hence, HKGs are widely used as internal controls for normalization of gene expression data (9, 16). However, recent studies have shown that the expression levels of several of the commonly used HKGs vary in different cell types (29) as well as after different biological treatments (20), making the choice of appropriate reference genes for normalization challenging. Compared with somatic cells, HKGs in stem cells are poorly studied (16) and even less explored in human pluripotent stem cells (hPSCs).

hPSCs can differentiate into all derivatives of the three primary germ layers: ectoderm, endoderm, and mesoderm. Moreover, hPSCs have the capacity for unlimited proliferation (49). Based on these features, hPSCs serve as useful tools in basic research as well as provide unique opportunities for future regenerative medicine applications. Global gene expression studies on hPSCs are valuable for gaining deeper knowledge about the mechanisms directing cellular differentiation (1, 4, 44–47). For large-scale expression analyses there are sophisticated normalization algorithms available based on the expression of thousands of genes (6, 24).

On the other hand, for gene expression studies on a smaller scale the most commonly used strategy for normalization is the parallel amplification of a reference gene as an internal standard. For correct calculation of gene expression data, it is of critical importance that these reference genes are stably expressed during the whole experiment. The selection of appropriate reference genes for studies in hPSCs has been demonstrated to be particularly challenging since the differentiation may influence the gene expression stability (53). Some of the most frequently used HKGs in both somatic cells and in hPSCs include glyceraldehyde-3-phosphate dehydrogenase (GAPDH), β-tubulin (TUBB), and β-actin (ACTB) (12, 56). However, in a previous study, we showed that these commonly used HKGs are inappropriate to use as reference genes in hESCs due to their substantial variability during spontaneous cell differentiation (48). Recently, Vossaert et. al. (53) highlighted the problem with high variability of commonly used HKGs in an experimental setup based on retinoic acid induced differentiation and proposed a new normalization approach in hESCs taking advantage of Alu repeat expression. However, it is still unclear whether Alu repeats can be generally applied for any type of directed differentiation of hPSC.

This present study demonstrates a comprehensive global transcriptional analysis, including several large-scale datasets originating from studies of hPSCs differentiating toward all
three germ layers. The datasets were carefully mined for putative HKGs in hPSCs across the various differentiation protocols. A handful of HKGs displaying stable expression in all datasets were identified, and we also present selected sets of HKGs for each germ layer specific differentiation.

MATERIALS AND METHODS

Microarray Datasets

In total, nine different datasets including 144 microarrays were analyzed in this study. These datasets originate from directed differentiation experiments representing all three germ layers and include data from both hESCs and human induced pluripotent stem cells (hiPSCs) differentiated to ecto-, meso-, and endoderm (5, 31, 39, 42, 47, 50). All nine datasets were generated with whole transcript Gene ST 1.0 arrays (Affymetrix, http://www.affymetrix.com). More details related to each dataset are summarized in Table 1.

Data Analysis

Filtering of background signal and unknown transcripts. To avoid the risk of interpretation of background signal as stable expression, all transcripts with expression signals below or close to the detection level of the microarrays in all samples were removed from further analysis. The minimum expression value for each gene among all samples in one dataset was calculated, and the 25th quartile of the transcripts with the lowest minimum expression values was excluded from further analysis.

Furthermore, suitable reference gene candidates would preferably be previously known genes with detailed annotation. Therefore, all genes that lack annotation and an official gene symbol were excluded from further analysis.

Analysis of the stability of commonly used HKGs in hPSCs. The stability in gene expression of 30 commonly used HKGs (Supplementary Table S1) in studies on hPSCs identified from literature (40, 53) was investigated, and their suitability as reference genes was examined in the nine global expression datasets.

Identification of stably expressed transcripts. To identify putative HKGs the coefficient of variation (CV) was calculated and used for selection of stably expressed genes. To identify transcripts that show stable expression, we defined a CV < 10% as the threshold level. As a first step, the overlap between the stably expressed genes in the different datasets representing each germ layer-specific differentiation was investigated separately and illustrated in Venn diagrams. Next, the intersection of stably expressed genes across all data sets (endoderm, mesoderm, and ectoderm) was defined and illustrated in a Venn diagram.

Gene Ontology enrichment analysis. To further explore the functional properties of the stably expressed genes, a Gene Ontology (GO) (2) enrichment analysis and a Protein Information Resource (PIR)-keyword enrichment (http://pir.georgetown.edu) analysis were performed using the DAVID bioinformatics resource (15). Significantly overrepresented GO annotations for biological process (BP), molecular function (MF), and cellular component (CC) were identified as well as enriched PIR keywords. To identify enriched annotation terms, the modified Fisher’s exact test [EASE score (22)] was used. The genes that were stably expressed in any of the germ layer-specific differentiation protocols (in total 1,237 genes) were explored for overrepresentation of annotations from the three GO categories. For this analysis, the more specific GO FAT categories provided by DAVID were used in order to minimize the redundancy of the more general GO terms and to increase the specificity of the terms. Similarly as for the GO categories, the enrichment of PIR keywords was also investigated. All genes represented on the arrays were used as the reference list in both these enrichment analyses.

RT-qPCR Analysis

To evaluate the results from the stability analysis of the microarray datasets, four of the stably expressed genes were selected for additional expression studies by RT-qPCR analysis. Two pluripotency markers, one hepatic marker, and four commonly used HKGs were also assayed for further evaluation. Directed hepatic differentiation of the hESC line SA121 and the hiPSC line ChiPS-4, both from Takara Bio Europe, was performed. The cells were cultured under defined culture conditions over a period of 21 days to obtain a hepatic phenotype. In brief, the protocol as described in Ref. 50 recapitulates known stages of liver development in vivo. During the first 7 days the undifferentiated hPSCs were guided to differentiate into definitive endoderm (DE). Next, the DE was subsequently in a stepwise manner differentiated into foregut endoderm, hepatoblasts, and finally, matured hepatocytes. RNA was isolated at the undifferentiated state (day 0), at day 7, and at day 21 of differentiation using MagMAX Express from Life Technologies. cDNA was synthesized using iScript cDNA Synthesis Kit from Bio-Rad, and RT-qPCR was performed in triplicate using assay-on-demand probes from Life Technologies. The following 11 genes were assayed: AFP, ACTB, CAPN10, CREBBP, EID2, GAPDH, HPRT1, NANOG, POUSF1, RABEP2, ZNF324B.

RESULTS

This study describes the analysis of global transcriptional datasets for identification of stably expressed genes during directed differentiation of hPSCs. Initially, we surveyed the literature and identified 30 commonly used HKGs in various experimental settings and cells (40, 53). The stability in expression levels of these genes was investigated in all of the nine datasets, and the results are shown in Supplemental Table S1. Importantly, 12 of these (ACTB, B2M, CREBBP, GAPDH,
HMBS, HPRT1, PPIA, RPL13A, SDHA, TBP, UBC, and YWHAZ) have also been used as reference genes in studies of hPSC differentiation (40, 48, 53). Notably, many of these genes display high variability in several of the investigated datasets, illustrating the challenges in finding suitable reference genes for gene expression data normalization (Supplemental Table S1).

Three different datasets, which are representative of directed endoderm differentiation, (Table 1), were analyzed for stably expressed transcripts. We included 48 microarrays from endoderm/hepatic differentiation in this analysis. Approximately 1,200–2,000 transcripts show stable expression in each dataset, with a stability threshold of CV < 10% applied (Fig. 1A). In total, 278 of these stably expressed transcripts overlapped in all three endodermal datasets (Supplemental Table S2).

We further analyzed three other datasets representing directed mesoderm and cardiac differentiation (Table 1). Data from 57 microarrays from mesoderm/cardiac differentiation were included in this analysis. As shown in Fig. 1B, between 1,600 and 5,000 transcripts show stable expression in each individual datasets with use of a threshold of CV < 10%. In total, 258 of these stably expressed transcripts were overlapping in all three datasets (Supplemental Table S3).

In the same way as described above for endoderm and mesoderm differentiation we also analyzed three datasets representing directed ectoderm differentiation (Table 1). Data from 39 microarrays from ectoderm/neural differentiation were included in the analysis. Between 3,000 and 5,800 transcripts show stable expression levels in all cell lines and during all variants of directed differentiation evaluated in this study. Thus, these genes, listed in Table 2, may serve as suitable reference gene candidates for hPSC differentiation experiments.

Evaluation of Stably Expressed Genes by RT-qPCR Analysis

To further evaluate the stability of some of the genes we identified from the microarray datasets, we further analyzed the expression of these genes during endoderm/hepatic differenti-
ation and performed two separate differentiation experiments using both hES and hiPS cell lines. The expression profiles of four of the identified stably expressed genes, EID2, ZNF324B, CAPN10, and RABEP2 from Table 2, and four of the commonly used HKGs (GAPDH, B-ACTIN, HPRT1, and CREBBP) were analyzed with RT-qPCR (Fig. 2). The CVs for the Cq values for the novel candidate HKGs were in the range of 2.69 to 4.05, while for the commonly used HKGs the corresponding values were 5.22–7.90. We further analyzed the RT-qPCR data by applying the Bestkeeper algorithm (37) on the raw Cq values to rank these genes with respect to standard deviation. Expectedly, Bestkeeper ranked CAPN10 as the most stable gene (SD/H11021 0.88) followed by RABEP2, EID2, ZNF324B, B-ACTIN, GAPDH, TNFRSF13C, CREBBP, and HPRT1 in subsequent order. Thus, the results from the RT-qPCR analysis confirm the stability of the expression levels of the novel proposed reference genes in hPSCs and underscore the usefulness of these genes for normalization of hPSC differentiation experiments.

To illustrate the importance of using stably expressed genes for normalization, we calculated the relative gene expression levels of three genes POU5F1, NANOG (markers of pluripotency), and AFP (induced during hepatic differentiation) using eight different reference genes: four traditionally used genes (GAPDH, ACTB, CREBB, and HRTP) and four of our novel candidate reference genes for hPSC differentiation (EID2, ZNF324B, CAPN10, and RABEP2). As shown in Fig. 3, A–C, this comparison demonstrates the critical importance of selecting stable reference genes for normalization since the data obtained differ substantially depending on which of the reference genes were used. Notably, the decrease in expression levels of the pluripotency markers and the increase in expression level of AFP are more pronounced using the novel candidate reference genes identified in this study compared with the commonly used set of HKGs.

### Enrichment Analyses of Stably Expressed Genes

To gain additional insights about the functional properties of the genes that show stable expression during various forms of directed differentiation, a GO analysis was performed, and overrepresented functional annotations were identified. For the BP category, annotations such as “transcription,” “regulation of transcription (DNA dependent),” and “sensory organ development” were significantly enriched among the stably expressed genes. For the CC category, “keratin filament” and “intermediate filament (cytoskeleton)” were significantly enriched. Finally, for the MF category, “sequence-specific DNA binding,” “transcription regulatory activity,” “structural molecule activity,” and “transcription factor activity” were significantly overrepresented among the stably expressed genes. To understand more about the functionality of the proteins for which these stably expressed genes encode, we also investigated enrichment of PIR keywords, and results show that “cytokine,” “homeobox,” and “keratin” were significantly enriched among these proteins.

<table>
<thead>
<tr>
<th>UniGene ID</th>
<th>Gene Symbol</th>
<th>Gene Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hs.18949</td>
<td>EID2</td>
<td>EP300 interacting inhibitor of differentiation 2</td>
</tr>
<tr>
<td>Hs.344088</td>
<td>TNFRSF13C</td>
<td>tumor necrosis factor receptor superfamily, member 13C</td>
</tr>
<tr>
<td>Hs.186970</td>
<td>ZNF324B</td>
<td>zinc finger protein 324B</td>
</tr>
<tr>
<td>Hs.112218</td>
<td>CAPN10</td>
<td>calpain 10</td>
</tr>
<tr>
<td>Hs.555978</td>
<td>RABEP2</td>
<td>rabaptin, RAB GTPase binding effector protein 2</td>
</tr>
<tr>
<td>Hs.642693</td>
<td>LTB4R2</td>
<td>leukotriene B4 receptor 2</td>
</tr>
<tr>
<td>Hs.147762</td>
<td>CCDC108</td>
<td>coiled-coil domain containing 108</td>
</tr>
</tbody>
</table>

Table 2. Stably expressed genes during directed hPSC differentiation

The 1st column states the UniGene ID number, and the 2nd column provides the official gene symbol. The 3rd column states the full gene name for each of the 7 genes that were stably expressed with CV < 10% in all 9 datasets (Fig. 1D).

### Fig. 2. The gene expression profiles of 4 genes (EID2, ZNF324B, CAPN10, and RABEP2), selected from the stably expressed genes in Table 2, and 4 commonly used HKGs (GAPDH, B-ACTIN, HPRT1, and CREBBP) were evaluated in 2 additional endoderm/hepatocyte differentiation experiments including both hESC and hiPSC differentiation. The gene expression levels were assayed by RT-qPCR run in triplicate for each experiment. This figure shows combined data from these RT-qPCR experiments, and error bars represent SE based on all replicates. The results confirm the outcome from the analysis of the microarray data. All novel candidate reference genes identified in this study show a lower CV than the commonly used reference genes, d, Day.
DISCUSSION

hPSCs with their unlimited proliferation properties and their ability to differentiate into specialized cell types provide unique opportunities for usage in a wide range of research applications. The possibility to convert hPSCs into functional cells allows for novel opportunities in assay development and holds great potential for future regenerative medicine applications (33). However, to fully make use of the hPSCs and their derivatives, extensive functional and transcriptional characterization is needed as well as benchmarking to relevant in vivo counterparts.

One common way to study the differentiation state of hPSCs is to monitor the expression of marker genes by RT-qPCR. Importantly, to allow for comparison between different samples, accurate normalization is required to calculate the relative gene expression and correct for variation in the RT-qPCR reaction (55). This normalization can be done either against the total amount of cellular DNA or RNA content or to the expression of properly selected reference genes. Both methods are practiced although it has been stressed that normalization to reference genes is the most reliable method (7, 8).

Stably expressed HKGs are commonly used as reference genes for normalization of RT-qPCR data, but notably, several reports have shown that differentiating stem cells express a different set of HKGs than somatic cells (34, 40, 48, 53, 55). Investigators have underscored the importance of careful validation of reference genes in the experimental system under study, since high variation between different cell types and experimental conditions has been observed (10, 52). In the present study we investigate, for the first time in detail, the stability of HKGs under directed differentiation conditions including derivatives from both hESCs and hiPSCs. The results from our study confirm that commonly used HKGs vary substantially in some of the datasets investigated (Supplemental Table S1) and emphasize the importance of careful selection of reference genes in hPSCs and their differentiating progenies. For example, ACTB shows stable expression in five out of nine datasets but varied substantially during ectoderm differentiation. Remarkably, six HKGs (HPRT1, PPIA, SDHA, B2M, HMBS, CREBBP), which all have been used as reference genes in studies on stem cells (13, 18, 25–27, 30, 35, 36, 38, 51), varied above our threshold for stable expression (CV < 10%) in all nine investigated datasets. In particular, HPRT1 and B2M are, based on our results, particularly unsuitable for normalization of hPSCs experiments due to their high variability in these cell types (Supplemental Table S1). Two commonly used HKGs (GAPDH and UBC) show stable expression (CV < 10%) in eight out of nine investigated experiments. We

Fig. 3. Comparison of the effect of using different reference genes for normalization of RT-qPCR data. The Cq values for 2 pluripotency markers (POU5F1 and NANOG), and 1 hepatic marker (AFP) were normalized following the ΔΔCq method using 8 different reference genes, 4 traditional ones and 4 novel candidate reference genes. The resulting values represent the fold change in expression levels on log₁₀ scale. A: expression profile of POUF1 in ChiPS-4 cell line and in SA121 cell line, normalized to 8 different reference genes. B: expression profile of NANOG in ChiPS-4 cell line and in SA121 cell line, normalized to these 8 reference genes. C: expression profile of AFP in ChiPS-4 cell line and in SA121 cell line, normalized to these 8 reference genes.
also identified UBC as a stable HKG in our previous report using spontaneously differentiating hESCs (48). However, in one of the ectoderm differentiation experiments, these two HKGs showed a CV of 22 and 16%, respectively, which is clearly above our defined threshold for stability. This is an important observation and illustrates that it is critical to validate the HKGs in each experimental setting and not simply rely on the fact that the gene has proven stable in other cells and/or experiments.

Different approaches have been proposed for identification of proper reference genes based on microarray data (21, 41, 48), gel electrophoresis (19), and Northern blot (43). In the present study, nine different microarray datasets including both hESCs and hiPSCs, across various directed differentiation protocols, have been mined for stably expressed transcripts. Interestingly, this work has now led to the identification of a small set of genes (Table 2) that shows stable expression in all nine datasets. These genes are EID2, TNFRSF13C, ZNF324B, CAPN10, RABEP2, LTB4R2, and CCDC108. Interestingly, three of these genes (EID2, RABEP2, ZNF324B) have also been identified as having one single exon stably expressed across different experiments, and CAPN10 showed stable expression for three of the exons when analysis was performed at the individual exon level (17). However, to the best of our knowledge none of these genes have previously been used as reference genes for normalization of gene expression in differentiating hPSC. These genes are involved in various cellular functions. For example, EID2 is known to bind to SMAD and act as a repressor of MYOD-dependent transcription and muscle differentiation (28), and TNFRSF13C is a regulator of the peripheral B-cell population and involved in immune deficiencies (32). ZNF324B is suggested to be involved in transcriptional regulation, and CAPN10 is a calcium-regulated gene involved in cytoskeleton remodeling and signal transduction and has a role in development of diabetes mellitus type II (14). RABEP2 plays a role in membrane trafficking and homotypic early endosome fusion (3). LTB4R2 has an important role in inflammatory signaling pathways (54), and CCDC108 is a coiled-coil domain containing 108 that is poorly studied, but according to GO annotations it is related to structural molecule activities. These genes have a CV < 10% in all examined microarray datasets and thus appear suitable to use as reference gene candidates for studies of hPSCs and their differentiated progenies. A combination of more than one of these genes as reference would most likely be an advantage and increase the robustness further as demonstrated in work by Vandesompele et al. (52) (geNorm) and Pfaffl et al. (37) (BestKeeper).

To further explore the biology of the stably expressed genes in one or several of the investigated differentiation regimes, two enrichment analyses (GO terms and PIR keywords) were performed. Results from these analyses indicate that significant proportions of these genes are either involved in transcriptional activities or contribute to the cellular structure.

To verify our results and illustrate the importance of using stable reference genes for normalization of RT-qPCR data, we monitored the expression of three marker genes (POUSF1, NANOG, and AFP) during hepatic differentiation of hiPSCs. The raw Cq values were normalized with eight different reference genes, four commonly used HKGs and four of our novel candidate reference genes. As expected, the results show more homogenous differentiation profiles when the data were normalized with our novel candidate reference genes (Fig. 3, A–C). In addition, our results also show that the HKG signatures vary substantially between different hPSC experimental setups, which further emphasizes the need for validation of reference genes in the appropriate experimental setting before they are used for normalization purposes. It should be noted that, for one differentiation regime used in our analysis, the number of stably expressed genes in the different datasets varied more than in the others (Fig. 1, A–C). However, the available information for each dataset does not provide any explanation for these apparent differences.

To further understand if there are any common mechanisms underlying HKG expression in hPSCs, the seven stable expressed genes (Table 2) were investigated with respect to genomic localization, length of gene sequence, distribution of exons and introns, and contribution in different molecular pathways. In addition, we have also searched the promoter sequences for common regulatory motifs and binding transcriptional factors in attempts to find possible explanations for their common stable expression. However, none of these analyses have resulted in any conclusive mechanistic findings. This is, however, not unexpected, and with the exception of Eisenberg and Levanon (16), who reported higher compactness of HKGs, no common regulatory mechanism has been observed in other studies on HKGs.

Taken together, the results from this study underscore the problem with instability of common HKGs in hPSCs and their derivatives and emphasize the importance of validating reference genes in the experimental setup under study, before using them for normalization purposes of gene expression data. The use of reference genes with an unstable expression profile can have serious consequences and may lead to misinterpretation of gene expression data. A novel set of putative HKGs in differentiating hPSCs from various differentiation strategies has been identified in this study, and these may preferably be employed as novel candidate reference genes for gene expression studies on differentiating hPSCs.

ACKNOWLEDGMENTS

We appreciate the Swegene Center for Integrative Biology at Lund University Genomics facility for the microarray service.

Present address for Y. Bogestål: SP Chemistry Materials and Surfaces, Arvid Wallgrens Backe 20, SE-413 46, Gothenburg, Sweden (yalda.bogestal@sp.se).

GRANTS

This work was supported by Takara Bio Europe AB (Göteborg, Sweden) and the University of Skövde, Sweden, under Knowledge Foundation Grants 2011:0295, 2012:0310, and 2013:09.

DISCLOSURES

No conflicts of interest, financial or otherwise, are declared by the author(s).

AUTHOR CONTRIBUTIONS

Author contributions: G.H. and J.S. analyzed data; G.H. and J.S. prepared figures; G.H., N.G., X.Z., Y.B., P.S., and J.S. edited and revised manuscript; G.H., N.G., X.Z., Y.B., P.S., and J.S. corrected final version of manuscript; Y.B., P.S., and J.S. drafted manuscript; P.S. and J.S. conceived and designed experiments; P.S. and J.S. interpreted results of experiments; J.S. performed experiments.


REFERENCES


