RESEARCH ARTICLE | Physiological Genomics of Cell States and Their Regulation and Single Cell Genomics

Comparative transcriptomics of hepatic differentiation of human pluripotent stem cells and adult human liver tissue

© Nidal Ghosheh,1,2 Barbara Küppers-Munther,3 Annika Asplund,3 Josefina Edsbagge,3 Benjamin Ulfenborg,1 Tommy B. Andersson,4,5 Petter Björquist,6 Christian X. Andersson,3 Helena Carén,7 Stina Simonsson,2 Peter Sartipy,1,8 and © Jane Synnergren1

1School of Bioscience, Systems Biology Research Center, University of Skövde, Skövde, Sweden; 2Institute of Biomedicine, Department of Clinical Chemistry and Transfusion Medicine, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden; 3Takara Bio Europe Aktiebolaget, Gothenburg, Sweden; 4Cardiovascular and Metabolic Diseases, Innovative Medicines and Early Development Biotech Unit, AstraZeneca, Mölndal, Sweden; 5Department of Physiology and Pharmacology, Section of Pharmacogenetics, Karolinska Institutet, Stockholm, Sweden; 6NovoHep Aaktiebolaget, Gothenburg, Sweden; 7Sahlgrenska Cancer Center, Department of Pathology, Institute of Biomedicine, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden; and 8AstraZeneca Research and Development, Global Medicines Development Cardiovascular and Metabolic Diseases Global Medicines Development Unit, Mölndal, Sweden

Submitted 23 January 2017; accepted in final form 28 June 2017

Ghosheh N, Küppers-Munther B, Asplund A, Edsbagge J, Ulfenborg B, Andersson TB, Björquist P, Andersson CX, Carén H, Simonsson S, Sartipy P, Synnergren J. Comparative transcriptomics of hepatic differentiation of human pluripotent stem cells and adult human liver tissue. Physiol Genomics 49: 430 – 446, 2017. — Hepatocytes derived from human pluripotent stem cells (hPSC-HEP) have the potential to replace presently used hepatocyte sources applied in liver disease treatment and models of drug discovery and development. Established hepatocyte differentiation protocols are effective and generate hepatocytes, which recapitulate some key features of their in vivo counterparts. However, generating mature hPSC-HEP remains a challenge. In this study, we applied transcriptomics to investigate the progress of in vitro hepatic differentiation of hPSCs at the developmental stages, definitive endoderm, hepatoblasts, early hPSC-HEP, and mature hPSC-HEP, to identify functional targets that enhance efficient hepatocyte differentiation. Using functional annotation, pathway and protein interaction network analyses, we observed the grouping of differentially expressed genes in specific clusters representing typical developmental stages of hepatic differentiation. In addition, we identified hub proteins and modules that were involved in the cell cycle process at early differentiation stages. We also identified hub proteins that differed in expression levels between hPSC-HEP and the liver tissue controls. Moreover, we identified a module of genes that were expressed at higher levels in the liver tissue samples than in the hPSC-HEP. Considering that hub proteins and modules generally are essential and have important roles in the protein-protein interactions, further investigation of these genes and their regulators may contribute to a better understanding of the differentiation process. This may suggest novel target pathways and molecules for improvement of hPSC-HEP functionality, having the potential to finally bring this technology to a wider use.

LIVER DISEASE ACCOUNTS for the death of ~1.7 million patients worldwide yearly (50), hereby becoming one of the leading cause of mortality in the world (34, 43, 49). End-stage liver disease is the condition when liver diseases lead to liver failure. This is the final stage of different liver diseases such as hepatitis infections, drug-induced liver injury, and hepatic carcinomas. The optimal treatment for end-stage liver disease is orthotopic liver transplantation (19, 59). However, the shortage of donor organs requires the search for alternative treatments such as hepatocyte transplantation and extra corporeal liver application (16, 23, 49). Freshly isolated human hepatocytes are currently the gold standard cell type for transplantation as well as for cell modeling for drug discovery and development, and various toxicity tests (15, 16, 21, 25, 34). However, limited availability, short life span, loss of functionality upon isolation and culturing in vitro, dedifferentiation, and failing to represent the polymorphic population urge the development of alternative cell sources that do not have these drawbacks (15, 25, 31, 34).

Human pluripotent stem cells (hPSCs) with their properties of infinite self-renewal and the ability to differentiate to all cell types in the body are an attractive cell source for regenerative medicine, cell therapy, drug discovery and development (23, 25). Moreover, human induced pluripotent stem cells (hiPSCs) generated from the transfection of somatic cells with pluripotency factors, may be employed in drug discovery as disease models (1, 4, 20, 31). In addition, these cells can also be utilized for patient specific treatment (1, 20). HPSCs have successfully been differentiated into hepatocytes (hPSC-HEP), sharing many features with in vivo hepatocytes, such as cytochrome P450 enzyme activity and glycogen storage, in addition to expression of several drug transporters (1, 17, 18, 52, 58). Furthermore, they are applicable for long-term toxicity studies (21) and can accurately classify toxic agents (55). HPSC-HEP have also been successfully and functionally engrafted in actual liver injured immune-compromised mice (37).
Improvements of the differentiation process of stem cells into hPSC-HEP have been achieved gradually in the last few years; however, methodology to generate fully functional hPSC-HEP is still lacking (9, 15, 23, 25, 42). The breakthrough of differentiating stem cells into hPSC-HEP with some activity of drug metabolic enzymes was accomplished by mimicking hepatogenesis in vitro (1, 7, 18, 58).

In previous work, we have applied robust and standardized definitive endoderm (DE) and hepatocyte differentiation protocols resulting in at least 90% pure cultures of DE and hPSC-HEP (1, 13), to differentiate hPSCs [including both human embryonic stem cells (hESCs) and hiPSCs] into hPSC-HEP. We investigated the differentiation process by monitoring the expression of 16 key developmental markers for hPSC, DE, hepatoblasts, fetal hepatocytes, and adult hepatocytes. The results revealed high synchronicity of the differentiation process across all cell lines tested. In addition, hPSC-HEP were shown to recapitulate important functionality of their in vivo counterparts, such as cytochrome P450 enzyme activity comparable to freshly isolated hepatocytes, expression of several drug transporters, and glycogen storage. Furthermore, no differences in differentiation efficiency between the hESCs and hiPSCs were observed (13), indicating a stringent and reproducible differentiation protocol.

In the present study, we applied genome-wide expression microarrays to investigate the progress of in vitro hepatic differentiation of hPSCs at DE, hepatoblast, early hPSC-HEP, and mature hPSC-HEP developmental stages at the transcriptome level, to identify putative functional improvement targets. Four time points during the differentiation procedure were identified as representative of DE, hepatoblast, early hepatocyte and mature hepatocyte stages, based on the results of our previous work (13). The cells were harvested at these time points, from the six different biological replicates (three hESC and three hiPSC lines) from the previous study (13). As reference samples, hPSCs and liver tissues containing ~80% hepatocytes (26, 34) were used. The 2,000 genes with the highest differential expression across the investigated time points and controls were selected for further analyses. Clustering analysis, functional annotation enrichment analysis, pathway analysis, and protein interaction network analysis were applied on these genes. The results revealed sets of genes with induced expression during different developmental stages, indicating essential roles for these genes in hepatic differentiation. As expected, some differences in the transcription profiles between the hPSC-HEP and liver tissue samples were observed. Moreover, hub proteins [proteins with high level of interactivity that usually are essential and play central role in protein interaction networks (53)], which appear to be involved in the different hepatocyte developmental stages, were identified.

MATERIALS AND METHODS

Cell Cultures and Hepatic Differentiation

All hPSC lines: three hESC lines Cellartis AS034, SA121, and SA181 (Takara Bio) referred to as A034, S121, and S181, respectively) and three hiPSC lines (human iPSC cell line ChiPSCb, P11012, and P11025, referred to as C6b, P12, and P25, respectively), Cellartis Definitive Endoderm Differentiation Kit with DEF-CS Culture System (cat. number Y30035), and a prototype of Cellartis Hepatocyte Differentiation Kit were provided by Takara Bio Europe (www.clontech.com). The culturing and differentiation of the hPSCs into hPSC-HEP were described previously (13). All hPSC lines are of XY karyotype. Liver tissues from two human male liver donors were purchased from tebu-bio (www.tebu-bio.com; H0529, Caucasian, age 28 yr, cause of death head trauma, and H0796, Caucasian, age 29 yr, cause of death anoxia).

RNA Extraction and Microarray Experiment

RNA processing and extraction were performed using MagMAX-96 Total RNA Isolation Kit and quantified by using GeneQuantpro spectrophotometer as described previously (13). RNA samples from two separate differentiation experiments from each of the six biological replicates day 5 (DE), day 14 (hepatoblast), day 25 (early hPSC-HEP), and day 30 (mature hPSC-HEP) during the hepatic differentiation were isolated. In addition, RNA was extracted from the reference samples; hPSCs (two samples from each of the six cell lines), and two liver tissue samples from different donors. All RNA samples were quality controlled using a Bioanalyzer. cDNA was synthesized from the RNA samples applying GeneChip WT PLUS Reagent Kit (Affymetrix) and subsequently applied to the GeneChip Human Transcriptome Array 2.0 (Affymetrix). In total 62 expression microarrays were run (Fig. 1A). Raw and processed data were submitted to ArrayExpress accession number: E-MTAB-5367.

Data Analysis

Filtering and preprocessing. To reduce nonbiological variation and make the transcript signal measures comparable across the microarray data set, the robust multiple average (RMA) normalization method was applied (22). To remove background signals from the data, probe sets with values below the median value of the spiked in negative controls on the array in 80% of all the samples included in the experiment, and given that these samples do not belong to only one sample group, were excluded. The output data from RMA are log2 transformed. Transcripts that lack official gene symbol were filtered from the data set. In total, 33,720 transcripts passed the above filtering criteria. The arithmetic mean of the replicated expression values of each of the cell lines was calculated and used for subsequent data analysis, resulting in 32 samples distributed on six groups (Fig. 1A). This preprocessed data are referred to as the data set throughout this paper (Supplemental File 1S). (The online version of this article contains supplemental material.)

Hierarchical clustering of samples included in the data set. To confirm the reproducibility of the differentiation and the microarray experiments, the samples in the data set were clustered using hierarchical clustering with Pearson correlation as distance measurement and the average linkage method by applying the genefilter package in R. The quality of the microarray data was also verified using spiked-in controls in standard quality control procedures.

Verification of the microarray results. Expression values of lineage specific genes that were investigated in our previous study [POU5F1 (OCT4), NANOG, SOX17, CXCR4, CER1, HHEX, TBX3, HNF6, PROX1, AFP, HNF4A, KRT18, ALB, SERPINA1 (AA7), CYP3A4] (13) were extracted from the microarray data and compared with the results of the real-time quantitative (q)PCR performed on the same RNA samples to verify the measures from the microarray experiment.

Identification of differentially expressed genes. Identification of the genes with the highest differential expression in the data set was performed applying the Significance Analysis of Microarray data (SAM) using the siggenes package and R. These genes were sorted on their d value, and the top 2,000 genes were selected for further in-depth analysis [false discovery rate (FDR) <3E-6]. The d value represents the score for relative difference in gene expression, which is the ratio of the change in gene expression to the standard deviation. Thus, the higher the d value is, the larger is the difference in gene expression among the investigated groups. SAM applies a modified t-test to detect significantly differentially expressed genes. Permutations are used to adjust the P values for multiple t-tests, and the q values represent the adjusted P values and determine the significance of the results (51).
Kmeans clustering of differentially expressed genes. The 2,000 genes with the highest differential expression were further analyzed and clustered using Kmeans clustering with the Pearson correlation as distance measure and by using R and the amap package. To determine an appropriate number of clusters for the Kmeans clustering the within-clusters sum of squares and the between-clusters sum of squares was calculated. Considering a trade-off between these two measurements, the number of clusters was set to 10, and the seed was set arbitrary to 1,987. Two of the clusters, numbers 4 and 6, showed very similar expression profiles with distinct peaks in gene expression at day 5. Only the magnitude of upregulation differed between these two clusters, and therefore these were merged into one larger cluster (cluster 4–6) in all subsequent analysis.

Functional annotation of the clusters. Functional annotation was performed by applying the CytoMine software (http://cytoscape.org/) with ClueGO plug-in (5). Each of the nine clusters was analyzed separately applying the following ClueGO settings: BiologicalProcess-GOA, CellularComponent-GOA, ImmuneSystemProcess-GOA, and MolecularFunction-GOA. For the “Evidence” parameter, only “All_Experimental (EXP, IDA, IPI, IGI, IEP)” was selected. In addition, the parameters “Use GO Term Fusion” and “Show only pathways with pV <0.05” were selected.

Pathway analysis. Each cluster was analyzed for pathway enrichment using the Enrichr software available at web site (http://amp.pharm.mssm.edu/Enrichr/) (8, 27). Results from the Wikipathways database with adjusted P value <0.05 were considered.

Protein interaction network analysis. Each cluster was further analyzed using the STRING database (https://string-db.org) to reveal interaction patterns among gene products within the same cluster. Only “Experiments,” “Gene Fusion,” “Databases,” “Co-occurrence,” and “Co-expression” were selected as active interaction sources in the “Data Settings.” Medium confidence (0.400) was selected for minimum required interaction score.

IDENTIFICATION OF PROTEIN HUBS. The identified protein interactions obtained from the STRING database were further analyzed with NetworkAnalyzer (2, 11) in Cytoscape to identify hub proteins. Proteins with node degree <10 were discarded. Nodes with clustering coefficient >0.7 were identified as major hubs (44). In addition, nodes with at least two combinations of the top 5% of the following network centralities (30), closeness centrality, betweenness centrality, and stress centrality, regardless of the clustering coefficient measurement, were considered as hubs. Hub proteins identified by applying any of these criteria were analyzed further by ClueGO to identify their biological function properties.

IDENTIFICATION OF PROTEIN MODULES. In addition, the interaction network derived by STRING were also analyzed by the MCODE application in Cytoscape, which applies clustering to identify modules, i.e., protein clusters in protein interaction networks (3). The
default settings with MCODE ≥7 were selected since they were demonstrated to give 100% accuracy of the predicted modules (3). The identified modules were analyzed further by ClueGO to determine their biological function properties. The enriched Gene Ontology (GO) terms for each modules were visualized using the REVIGO software available at (http://revigo.irb.hr/) (47).

RESULTS

Hierarchical Clustering

Hierarchical clustering showed consistent results from the array experiment and the differentiation procedure with all samples from the same development stages tightly grouped together (Fig. 1B). The dendrogram also confirms that the later differentiation stages (hepatoblast, early hPSC-HEP, and mature hPSC-HEP) show higher similarity to each other than to the earlier stages such as DE cells and hPSC. Early hPSC-HEP and late hPSC-HEP were clustered together, although there is a clear distinction between early and late hPSC-HEP illustrated by the dashed line in Fig. 1B. In addition, higher in the dendrogram, the human liver samples cluster with the later differentiation stages, as expected. Importantly, no distinction between hESC-lines and hiPSC-line was implied from the clustering results.

Verification of the Microarray Results

The expression of 15 lineage-specific genes that we monitored in our previous study (13) was extracted from the microarray data. The expression mean of all biological replicates was calculated and plotted against the time points during the differentiation. Figure 2 illustrates the log2 expression profiles of these genes. The microarray results show high consistency with the real-time qPCR results from our previous study (13), lending support to the reliability of the microarray experiment. The pluripotent genes POU5F1 (OCT4) and NANOG are expressed at day 0 and downregulated upon differentiation, where POU5F1 shows a more rapid downregulation than NANOG (Fig. 2A). The DE markers CER1, CXCR4, SOX17, and HHEX are highly expressed at day 5 (Fig. 2B). AFP, TBX3, and AAT are already expressed at the hepatoblast stage (Fig. 2C), and ALB and CYP3A4 are expressed in early and mature hPSC-HEP (Fig. 2D).

Identification of Differentially Expressed Genes

The SAM algorithm for multiple group comparison was applied to the six groups of samples, and the 2,000 genes that showed highest differential expression among these groups were identified (Supplemental File 2S). The highest FDR for genes in this list was < 3E-6. Table 1 shows a selection of the top 30 genes in this list. Notably, the lineage-specific genes AFP (fetal hepatocyte marker), CXCR4, and CER1 (DE markers) are present among these top 30 genes. However, the majority of the genes in the list are expressed at the later stages of the hepatic differentiation.
The top five differentially expressed genes from cluster 7 are F2RL1, LPQAT1, MPZL3, CYB5A, and HGSNAT. Cluster 8 includes only genes expressed in undifferentiated hPSCs. The top five differentially expressed genes from cluster 8 are CDC47L, JAKMIP2-AS1, PKR2B, ZNF589, and TERF1. Cluster 9 includes genes that are expressed in the later stages of hepatic differentiation, but these genes are slightly overexpressed compared with liver tissue controls. The top five differentially expressed genes from cluster 9 are AFP, TTR, FGA, FGB, and APOB. Cluster 10 includes genes that are expressed at lower levels in mature hPSC-HEP compared with the liver tissue controls. The top five differentially expressed genes from cluster 10 are ADH1B, APCS, AKR1C4, HAO1, and SLCO1B1. A complete list of all the genes in each of these clusters is available in Supplemental File 3S.

**Functional Annotation of the Gene Clusters**

The gene clusters were analyzed for enrichment of GO annotations using the ClueGO in Cytoscape to reveal biological functions for the genes in the different clusters. Significantly enriched GO annotations are shown in Figs. 4, 5, and 6, with only parent terms included in the pie charts. Significantly enriched annotations for genes in cluster 1 (Fig. 4A) are related to cell cycle activities. The genes in cluster 5 (Fig. 4C) are enriched for cell cycle as well. Enriched annotations for the genes in cluster 7 (Fig. 4B) include transcriptional regulation and chromatin organization. The genes in cluster 4 (Fig. 4C) show significant enrichment for terms that are involved in the determination of the DE fate, such as regulation of the activin receptor signaling pathway and negative regulation of the canonical Wnt signaling pathway, as well as cardiac markers arising from the common developmental stage mesendoderm. The genes in cluster 7 are enriched for GO terms for inter alia epithelial differentiation, tight junction, regulation of ion transport, vesicle, and vacuole (Fig. 5B). Figure 5C shows enrichment for positive regulation of endothelial cell proliferation and epithelial cell proliferation GO terms in cluster 8. Genes in cluster 9 (Fig. 6A) show enrichment for lipid metabolism processes, lipoprotein particle organization and remodeling, regulation of homeostasis and lipids synthesis, and protein activation cascade. Genes in cluster 10 (Fig. 6B) are enriched for various metabolic processes. No significantly enriched functional annotations were identified for the list of genes in cluster 2.

**Pathway Analysis**

To explore putative pathway activity for genes in the different clusters, pathway enrichment analysis was performed on each of the clusters separately using the Enrichr software. For this analysis, the Wikipathways database was referenced, and only pathways with adjusted P values < 0.05 were reported. In total 64 enriched pathways for genes in the clusters were identified, and some of these overlap between clusters. Among these are, for example, signaling of hepatocyte growth factor receptor, complement and coagulation cascades, drug induction of bile acid pathway, oxidation by cytochrome P450, and metabolic biomembrane transformation. A complete list of enriched pathways for each of the different clusters are shown in Supplemental Table 1S.

---

**Table 1. Top 30 differentially expressed genes during hepatic differentiation**

<table>
<thead>
<tr>
<th>Gene Symbol</th>
<th>d Value</th>
<th>Cluster</th>
</tr>
</thead>
<tbody>
<tr>
<td>LITD1</td>
<td>2,051.49</td>
<td>1</td>
</tr>
<tr>
<td>MIR302A</td>
<td>1,652.50</td>
<td>1</td>
</tr>
<tr>
<td>ZIC3</td>
<td>1,544.86</td>
<td>1</td>
</tr>
<tr>
<td>LINC00458</td>
<td>1,447.24</td>
<td>1</td>
</tr>
<tr>
<td>MCM10</td>
<td>1,350.34</td>
<td>1</td>
</tr>
<tr>
<td>TME458B</td>
<td>1,493.87</td>
<td>2</td>
</tr>
<tr>
<td>LGR5</td>
<td>3,012.81</td>
<td>4_6</td>
</tr>
<tr>
<td>EOMES</td>
<td>2,913.83</td>
<td>4_6</td>
</tr>
<tr>
<td>MIXL1</td>
<td>2,698.13</td>
<td>4_6</td>
</tr>
<tr>
<td>RP4-559A3.6</td>
<td>2,451.09</td>
<td>4_6</td>
</tr>
<tr>
<td>CXCR4</td>
<td>2,353.56</td>
<td>4_6</td>
</tr>
<tr>
<td>CER1</td>
<td>1,691.97</td>
<td>4_6</td>
</tr>
<tr>
<td>NCAF6</td>
<td>1,429.46</td>
<td>5</td>
</tr>
<tr>
<td>APOB</td>
<td>6,092.85</td>
<td>9</td>
</tr>
<tr>
<td>FGB</td>
<td>5,805.99</td>
<td>9</td>
</tr>
<tr>
<td>FGA</td>
<td>3,507.99</td>
<td>9</td>
</tr>
<tr>
<td>TTR</td>
<td>3,353.60</td>
<td>9</td>
</tr>
<tr>
<td>AFP</td>
<td>2,688.78</td>
<td>9</td>
</tr>
<tr>
<td>APOA4</td>
<td>2,410.96</td>
<td>9</td>
</tr>
<tr>
<td>FGG</td>
<td>1,675.20</td>
<td>9</td>
</tr>
<tr>
<td>SLC16A4</td>
<td>1,475.83</td>
<td>9</td>
</tr>
<tr>
<td>HAVCR1</td>
<td>1,395.49</td>
<td>9</td>
</tr>
<tr>
<td>MIA2</td>
<td>1,371.97</td>
<td>9</td>
</tr>
<tr>
<td>CRP</td>
<td>2,900.12</td>
<td>10</td>
</tr>
<tr>
<td>SLCO1B1</td>
<td>2,822.20</td>
<td>10</td>
</tr>
<tr>
<td>HA01</td>
<td>1,932.19</td>
<td>10</td>
</tr>
<tr>
<td>AKR1C4</td>
<td>1,779.06</td>
<td>10</td>
</tr>
<tr>
<td>APCS</td>
<td>1,663.55</td>
<td>10</td>
</tr>
<tr>
<td>ADH1B</td>
<td>1,399.21</td>
<td>10</td>
</tr>
</tbody>
</table>

---

Kmeans Clustering of the Differentially Expressed Genes

The Kmeans clustering of the differentially expressed genes identified by SAM recognized 10 distinct clusters that mirrored the expression dynamics of different groups of genes during the hepatic differentiation (Fig. 3). The cluster sizes range between 35 and 472 genes. Cluster 1 includes genes that are highly expressed in the early stages of hepatic differentiation (hPSC and DE). The top five differentially expressed genes from cluster 1 are LITD1, MIR302A, ZIC3, LINC00458, and MCM10. Cluster 2 includes genes that are upregulated in the early and mature hPSC-HEP. However, most of these genes are expressed at higher levels in the hPSC-HEP compared with the liver tissue controls. The top five differentially expressed genes from cluster 2 are LRRC19, ISX, AREG, SLC51B, and CYP1A1. Cluster 3 includes genes that are downregulated gradually during the hepatic differentiation. The top five differentially expressed genes from cluster 3 are TIAM1, BIRC5, GTF3C4, DNMT1, and PSMC3IP. Cluster 4 and cluster 6 have very similar expression profiles and include the DE genes. The top five differentially expressed genes from cluster 4_6 are CXCR4, RP4-559A3.6, MIXL1, EOMES, and LGR5. Cluster 5 includes genes that show high expression levels at early stages of the differentiation, and their expression decreases as the cells continue the differentiation process. The top five differentially expressed genes from cluster 5 are NCAPG2, GPM6B, TRIP13, CIT, and NCAF6. Cluster 7 includes genes that show higher expression levels at later stages of the differentiation. However, these genes are slightly higher expressed in early and mature hPSC-HEP compared with the liver tissue samples. The top five differentially expressed genes from cluster 7 are F2RL1, LPQAT1, MPZL3, CYB5A, and HGSNAT. Cluster 8 includes only genes expressed in undifferentiated hPSCs. The top five differentially expressed genes from cluster 8 are CDC47L, JAKMIP2-AS1, PKR2B, ZNF589, and TERF1. Cluster 9 includes genes that are expressed in the later stages of hepatic differentiation, but these genes are slightly overexpressed compared with liver tissue controls. The top five differentially expressed genes from cluster 9 are AFP, TTR, FGA, FGB, and APOB. Cluster 10 includes genes that are expressed at lower levels in mature hPSC-HEP compared with the liver tissue controls. The top five differentially expressed genes from cluster 10 are ADH1B, APCS, AKR1C4, HAO1, and SLCO1B1. A complete list of all the genes in each of these clusters is available in Supplemental File 3S.
Fig. 3. Kmeans clustering of top 2,000 differentially expressed genes during hepatic differentiation applying Pearson correlation as distance measure. The x-axis shows the biological replicates at day 0 (hPSCs), day 5 (DE), day 14 (hepatoblast), day 25 (early hPSC-HEP), day 30 (mature hPSC-HEP), and hLT_control (human liver tissue controls). The black lines represent the mean expression profiles for each of the clusters.
**Protein Interaction Network Analysis**

The results from the protein interaction network analysis showed interesting interactions between gene products for some of the clusters. Each cluster was also analyzed using the STRING database to identify any protein interaction between cluster members. The results from the STRING database were further explored in more details with the NetworkAnalyzer in Cytoscape. This analysis identified hub proteins only in cluster 1, 3, and 5, and these are listed in Supplemental Table 2S. As only two, one, and three proteins with node degree ≥ 10 were identified for clusters 7, 9, and 10, respectively, the top 5% of the network centralities criteria could not be fulfilled. However, due to the involvement of these clusters in the maturation stage of hepatocytes and since they have relatively high degree of betweenness and closeness centrality, these nodes were considered as hub proteins as well (Supplemental Table 2S).

Functional annotation analysis of hub proteins revealed that hubs from cluster 5 were enriched for inter alia “protein localization to chromosome,” “spindle midzone,” “regulation of chromosome segregation,” and “chromosome localization” GO terms. Hub proteins from cluster 1 were enriched for the following GO terms “kinetochore,” “chromosome, centromeric region,” “spindle microtubule,” “chromosome localization,” “establishment of chromosome localization,” and “metaphase plate congression.” Hubs from cluster 3 were not enriched for any GO term, and for the hubs identified in cluster 10 the drug metabolism term was enriched.

To identify modules (protein clusters with high level of interactivity), the MCODE application in Cytoscape was applied. One module was identified for cluster 1. Figure 7 illustrates the localization of the module in the protein interaction network generated by STRING (Fig. 7A) and the module (Fig. 7B). The protein members of the modules were analyzed with ClueGO to identify significantly enriched GO terms for biological processes (BP). Figure 7C shows enriched BP terms visualized using the REVIGO software. The results indicate that the module is involved in sister chromatid cohesion and chromosome localization. For cluster 3, two modules were identified (Fig. 8, B and C). Results from the GO enrichment analysis for module 1 and 2 show that proteins in module 1 is involved in RNA splicing (Fig. 8D), and proteins in module 2 are involved in chromosome segregation and...
sister chromatid cohesion (Fig. 8E). Three modules were identified in cluster 5 (Fig. 9, B–D). The results from GO enrichments for these modules indicate involvement of proteins in module 1 in DNA replication and phase G1/S transition in the mitotic cell cycle (Fig. 9F). For module 2 no significantly enriched terms for BPs were identified. GO results for module 3 show that this module is involved in spliceosomal snRNP assembly (Fig. 9E).

Interestingly, an additional module from cluster 10 with the MCODE score just below the threshold (Fig. 10B) was analyzed as well for GO enrichment for BP. The results show that this module is highly involved in drug metabolism, exogenous...
drug catabolism, exogenous P450 pathway, and fatty acid derivative metabolism (Fig. 10C).

**DISCUSSION**

There is a great need to find a renewable source of human hepatocytes with maintained functionality during extended culture in vitro, which in addition reflects the polymorphism in the human population. hPSC-HEP have the potential to be such a source (1, 7, 15, 18, 55, 58) and possibly replace current hepatic models lacking adequate properties. However, for some applications the functionality of hPSC-HEP still needs to be further improved (15, 25, 37, 42, 48, 56).

In this study, we applied whole genome transcriptomics to investigate the dynamics of gene expression during the differentiation of three hESC and three hiPSC lines to hPSC-HEP, through the investigation of four defined developmental stages.
For reference samples, we used undifferentiated hPSCs and human liver tissues. The choice of human liver tissue as a control instead of primary cells was made to ensure full functionality of hepatocytes in their natural microenvironment, since the dissociation and the culturing of hepatocytes in vitro reduce their functionality within hours after isolation (16). Moreover, liver tissues contain at least 80% hepatocytes (26, 34), which makes them suitable as hepatocyte controls that would convey the putative gene expression deviations of hPSC-HEP from in vivo hepatocytes regarding hepatic genes.

The selection of relevant time points representing DE, hepatoblast, early hPSC-HEP, and mature hPSC-HEP was based on our previous study (13). Day 5 was selected to represent the DE stage since it showed expression peaks for the DE markers CXCR4 and CER1. The expression of the early hepatic markers HNF6, HNF4a, TBX3, and AFP is initiated at the hepatoblast stage. These genes were all expressed at day 14; therefore, day 14 was selected to represent the hepatoblast stage. AFP is a fetal hepatocyte marker, and the expression of AFP was shown to decrease after day 25. Therefore, day 25 was selected to represent early hPSC-HEP. Finally, day 30 was selected to represent mature hPSC-HEP, as it showed relatively high expression of the mature hepatocyte markers ALB, AAT, and CYP3A4 and lower AFP expression (13).

Notably, our results from the hierarchical clustering clearly demonstrate that hESC lines and hiPSC lines are similar on the global transcriptomic level during the differentiation to hPSC-HEP. All the samples clustered with the corresponding samples from the same time point. This underscores the robustness of the applied differentiation protocol for both hESCs and hiPSCs.

We investigated the top 2,000 differentially expressed genes in the data set using Kmeans clustering algorithm, and the genes were grouped into nine representative gene clusters. Functional annotation and pathway analyses performed on genes in these clusters showed that the early stages during the
differentiation, i.e., the hPSC, the DE, and the hepatoblast, were enriched in genes involved in the cell cycle process, DNA synthesis and replication, and chromosome organization. However, the expression levels of these genes were much lower in hepatoblasts than in the hPSC and DE stages and slightly higher than in early and mature hPSC-HEP. These results are in agreement with other reports on liver organogenesis indicating the proliferative property of the early fetal liver (28, 57).

Furthermore, DE cells were enriched for pathways and GO terms involved in regulation of differentiation and proliferation of cells (regulation of ERK1 and ERK2 cascade) and regulation of the Wnt and Activin signaling pathways. These pathways are well-known regulators of DE differentiation (60). The later differentiation stages (i.e., early and mature hPSC-HEP) were enriched for pathways and GO terms involved in typical hepatocyte functions such as lipid metabolism, complement and cholesterol regulators (statin pathway), which is well aligned with liver development (6, 28). The late-stage clusters, starting from the hepatoblast stage and forward, were enriched for GO terms such as epithelial differentiation (implying the differentiation process of hepatoblasts, which are hepatic epithelial cells), bicellular tight junctions, vesicles, vacuoles, and regulation of ion transport involved in the clearance and transport of metabolites (40, 45, 46), in addition to pathways connected to important hepatocyte functions such as blood coagulation, complement system, lipid particle organization/remodeling, and lipid metabolism (60). Moreover, at later stages of hepatocyte differentiation there is enrichment for metabolic pathways, including several cytochrome P450 enzymes, e.g., CYP1A1, CYP1B1, and CYP3A5, which were expressed at lower levels in the liver tissue controls than in the hPSC-HEP. Interestingly several of the genes in cluster 2 are indicative of a mature hepatocyte phenotype, e.g., the transporters ABCB1 (MDR1, P-glycoprotein), ABCG2 (BCRP), and SLC51B (OST beta). In contrast, other genes in cluster 2 are rather typical for an immature phenotype. CYP1A1 and CYP1B1, for example, are expressed in fetal liver and subsequently downregulated or silenced in adult liver (35). These results imply inducibility of some maturation genes, in addition to either incomplete differentiation of hPSC toward mature hepatocytes or failure in turning off transcription of fetal genes, or both. These deviations should be addressed to improve the functionality of hPSC-HEPs.

Notably, cluster 10 includes many genes that are expressed at higher levels in liver tissue than in the hPSC-HEP. These genes are involved in hepatocyte activity pathways such as complement activation, metapathway biotransformation, oxidation by cytochrome P450 and drug induction of bile acid pathways. Increasing the expression of these genes would further improve the functionality of hPSC-HEP.

Protein interaction network analysis was performed cluster-wise to reveal interesting topological characteristics in the interactions among the gene products within each cluster. Several hub proteins with high connectivity were identified. Hub proteins are often essential and play central roles in protein interaction networks. In the present study, we applied a combination of node degree and centralities criteria to increase the possibility of identifying essential hub proteins. In addition, modules, i.e., protein clusters, were also identified and explored.
Fig. 8. A: protein interaction network of cluster 3; B: module 1; C: module 2; D: GO enrichment results for biological processes for module 1; E: GO enrichment results for biological processes for module 1.
Fig. 9. A: protein interaction network of cluster 5; B: module 3; C: module 1; D: module 2; E: GO enrichment results for biological processes for module 3; F: GO enrichment results for biological processes for module 1.
The vast majority of hub proteins and modules that were identified in this study are involved in cell cycle activities, DNA replication, and RNA transcription, including, inter alia, kinetochore and chromosome localization, RNA splicing, rRNA metabolism, sister chromatid cohesion and chromosome segregation, DNA replication, G1/S transition of mitotic cell cycle and nuclear envelope disassembly, and spliceosomal snRNP assembly. In general, modules included mostly hub proteins. However, module members that were not identified as hub proteins were associated with cell cycle, including all hub proteins except CDK1; in addition to BUB1 (mitotic checkpoint serine/threonine kinase) and NDC80 (kinetochore complex component), H2AFZ (H2A histone family member Z) is a variant of histone H2A. GINS4 is important for the GINS complex assembly that is essential to the initiation of DNA replication and progression of DNA replication forks (32).

Unlike genes associated with cell cycle, DNA replication, and RNA transcription activities, genes activated in hPSC and DE stages did not include proteins with the high connectivity level that matched the criteria of hub proteins. Apparently, these results indicate the critical role of cell cycle, DNA replication, and RNA transcription processes, as they exist constantly across all species and cell types.

Although no proteins matched our stringent criteria for hub proteins for the later differentiation stages (i.e., hepatoblasts, early hPSC-HEP, late hPSC-HEP and adult liver), six genes retained close values. MAPK1 is a member of the MAP kinase family, which is involved in various cellular processes such as proliferation, transcription regulation, differentiation, and development (38). MAPK1 is slightly upregulated in hPSC-HEP compared with liver tissue controls. PIK3R1, is a regulatory subunit of PIK3, which regulates cellular metabolism and growth. In addition, it is involved in the metabolism of insulin and modulates the cellular response to resolve endoplasmic reticulum stress that may lead to insulin resistance (54). However, PIK3R1 is slightly downregulated in hPSC-HEP compared with liver tissues. APOA1 (apolipoprotein A1) is involved in cholesterol efflux in hepatocytes (41). UGT2B4 is a member of the UDP glucuronosyltransferase family, which consists of phase II drug metabolism enzymes (10, 33). AOX1 is an oxidase that plays a key role in the metabolism of xenobiotics and drugs containing aromatic azaheterocyclic substituents (12). Finally, CYP2E1, a member of the cytochrome P450, is a phase I drug-metabolizing enzyme (33), and differences in epigenetic regulation of CYP2E1 between hPSC-HEP and primary hepatocytes have been determined (36), which may account for the impaired activity of this enzyme. As expected, AOX1 and CYP2E1 were not expressed in hPSC-HEP.

We identified an interesting module consisting of CYP4A11, CYP2A6, CYP2B6, CYP2C9, CYP2E1, UGT2B4, and AOX1, which are involved in drug and fatty acid metabolism, and drug catabolism. These gene products likely share many regulatory factors (3). Considering that these genes are expressed in liver tissues samples, but downregulated or silenced in hPSC-HEP samples, further investigation of this module may reveal important limitations in its regulatory pathways and provide possible approaches to further improve the functionality of hPSC-HEP.

Our results are consistent with similar studies comparing human primary hepatocytes to hPSC-HEP generated using an earlier generation of the differentiation protocol used in the present study (14) and with a different differentiation protocol analyzing several time points during the hepatic differentiation (39), where functional annotation analysis revealed that early stages of hepatic differentiation are GO enriched for terms in cell cycle and differentiation, while later stages are enriched for hepatocytes functional terms (14, 39). This is in agreement with studies of in vivo liver tissues (6, 24, 28, 29). Furthermore, similar results of clusters of genes that are either not expressed or low in hPSC-HEP compared with primary hepato-
cytoxenes are shown in (14), indicating remaining immature features of hPSC-HEP.

To conclude, a comprehensive global transcriptomics analysis during the differentiation of hPSCs into mature hPSC-HEP was performed. Functional annotation, pathway, and protein interaction network analyses were applied to the top 2,000 differentially expressed genes in the transcriptomics data set. The results revealed clusters of genes involved in the regulation and differentiation of defined developmental stages during hepatogenesis. However, deviation in expression compared with adult liver tissue, where genes were either overexpressed or downregulated in hPSC-HEP, was also observed. The regulation mechanism of these genes might be putative targets for product improvement strategies.

Notably, more than 40% of the genes were involved in cell proliferation and differentiation in general. The vast majority of the identified hub proteins and modules are related to the cell cycle process. Functional annotation analysis results suggest that the differentiation of hPSCs to hPSC-HEP recapitulated the in vivo hepatogenesis, where cell proliferation occurs in early developmental stages and maturation at later stages. These results promote the application of hPSC-HEP differentiation to study human hepatogenesis. Notably, most hub proteins that were identified were also members of the different modules, indicating the importance of these proteins during the hepatic differentiation. Finally, we have identified a module of genes expressed at lower levels in hPSC-HEP comprising CYP4A11, CYP2A6, CYP2B6, CYP2C9, CYP2E1, UGT2B4, and AOX1. Several studies suggest that epigenetic factors may underlie the impaired activity of hepatocyte maturity genes (19, 36). Thus, investigating the dynamics of these factors during differentiation, in addition to the pathways and transcription factors that regulate this module, may reveal several leads. These leads could be either chemically manipulated by regulatory modulators or small molecules and/or genetically manipulated by overexpression or knockout effect, to rectify the expression of maturity genes to match in vivo expression levels and subsequently allow the generation of fully functional hPSC-HEP. Nevertheless, further investigation of the regulation of genes with deviating expression identified in this study, in addition to the identified hub proteins and set of genes significant for liver development, has high potential to reveal mechanisms underlying these deviations, hereby promoting the development of methods to rectify gene expression and further improvement of the functionality of hPSC-HEP. Moreover, combination of results from present transcriptomics and planed methylation studies may ultimately leverage the unlimited population of functional liver cells highly demanded by the fields of drug discovery and regenerative medicine.

ACKNOWLEDGMENTS

We appreciate the Swegene Centre for Integrative Biology at Lund University Genomics for the microarray service. We also appreciate advice from Zelmina Lubovac for the network analysis.

GRANTS

This work was supported by Takara Bio Europe AB (Gothenburg, Sweden), AstraZeneca (Möln达尔, Sweden), and the University of Skövde, Sweden, under grants from the Knowledge Foundation [2012/0310], [2013/89] and [2014/0301].

DISCLOSURES

Authors Barbara Küppers-Munther, Annika Asplund, Josefina Edsbagge, Christian X. Andersson are employees of Takara Bio Europe AB. Authors Tommy B. Andersson and Peter Sartipy are employees of AstraZeneca.

AUTHOR CONTRIBUTIONS


REFERENCES


21. Holmgren

29. Lee

20. Irizarry

Guguen-Guillouzo

16. Hay

26. Jensen

14. Goldring

23. Jensen


