Developmental Cell
Resource

Transcriptome Analysis of Early Organogenesis in Human Embryos

Hai Fang,[1,2,5] Ying Yang,[1,3,5] Chunliang Li,[1,3] Shijun Fu,[1] Zuqing Yang,[4] Gang Jin,[1] Kankan Wang,[1,2] Ji Zhang,[1,2,*] and Ying Jin[1,3,*]

[1]Key Laboratory of Stem Cell Biology, Institute of Health Sciences, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences and Shanghai Jiao Tong University School of Medicine (SJTU-SM), 225 South Chongqing Road, Shanghai 200025, China
[2]Shanghai Institute of Hematology and Sino-French Research Center for Life Sciences and Genomics, Ruijin Hospital affiliated to SJTU-SM, Ruijin Road II, Shanghai 200025, China
[3]Shanghai Stem Cell Institute, SJTU-SM, 225 South Chongqing Road, Shanghai 200025, China
[4]Department of Obstetrics and Gynecology, Xinhua Hospital affiliated to SJTU-SM, 1665 Kongjiang Road, Shanghai 200025, China
[5]These authors contributed equally to this work
*Correspondence: jizhang@sibs.ac.cn (J.Z.), yjin@sibs.ac.cn (Y.J.)
DOI 10.1016/j.devcel.2010.06.014

SUMMARY

Genome-wide expression analysis of embryonic development provides information that is useful in a variety of contexts. Here, we report transcriptome profiles of human early embryos covering development during the first third of organogenesis. We identified two major categories of genes, displaying gradually reduced or gradually increased expression patterns across this developmental window. The decreasing group appeared to include stemness-specific and differentiation-specific genes important for the initiation of organogenesis, whereas the increasing group appeared to be largely differentiation related and indicative of diverse organ formation. Based on these findings, we devised a putative molecular network that may provide a framework for the regulation of early human organogenesis. Our results represent a significant step in characterization of early human embryogenesis and provide a resource for understanding human development and for stem cell engineering.

INTRODUCTION

The mammalian embryo develops from a single-cell zygote to a blastocyst (preimplantation stage), followed by gastrulation and then organogenesis (postimplantation stage). Organogenesis begins when ectodermal cells form the neural tube and mesodermal tissues become segmented into somites. Subsequently, organ primoridia begin to appear as organogenesis continues. Human organogenesis begins at Carnegie stage 9 (embryonic day 20, E20) and ends at Carnegie stage 23 (E56) (Carlson, 2004). This 36 day organogenesis period in the human is considerably longer than the 7 day period seen in the mouse. Nevertheless, the major changes in body form occur during the first 10 days. At E20, the embryo has the shape of the sole of a shoe, and the neural groove is evident dorsally. Soon after that, the neural folds form and primordial development of heart, optic system, thyroid, liver, and respiratory system takes place. At the end of Carnegie stage 14 (E32), the primary structure of many organs is distinguishable. Therefore, E20–E32 is a critical time period for human development, laying the foundation for subsequent developmental events.

During embryogenesis, early embryonic cells progressively differentiate into distinct cell types with a concomitant gradual loss of developmental potential, starting from the totipotent state, passing through the pluripotent state, and then to lineage commitment (Waddington, 1957; Yamanaka, 2009). Pluripotency, a characteristic of cells in the inner cell mass of the blastocyst, is defined as the potential of a cell to generate all cell types in an organism. Recently, it has been proposed that pluripotency could be more dynamic than previously thought (Smith et al., 2009). The pluripotency of human embryonic stem cells (hESCs), derived from the in vitro culture of human preimplantation embryos (Thomson et al., 1998), is difficult to evaluate developmentally for ethical reasons. Moreover, the recent advent in human induced pluripotent stem cells (iPSCs) renders the evaluation of pluripotency even more challenging (Takahashi et al., 2007; Yu et al., 2007), as reprogramming of somatic cells by defined factors is believed to be a continuous stochastic process, generating a heterogeneous population of cells at different states (e.g., fully reprogrammed pluripotent cells, partially reprogrammed cells, and nonreprogrammed differentiated cells) (Hanna et al., 2009). Transcriptome approaches have shown promise for gaining insights into the biology of undifferentiated hESCs (Abeyta et al., 2004) and human preimplantation embryos (Sudheer and Adjaye, 2007). For instance, the extent to which hESCs cultured in vitro reflect human embryos at the blastocyst stage in vivo can be assessed using such approaches (Sudheer and Adjaye, 2007). Similar efforts have also been made for looking at differentiated derivatives of hESCs (e.g., embryoid bodies, EBs), which are often used as in vitro models to study early human development (Dvash et al., 2004; Liu et al., 2006). However, it is still unknown whether the EB differentiation model recapitulates human early embryogenesis in vivo. Accordingly, information obtained from transcriptome analysis of human early embryos may help to

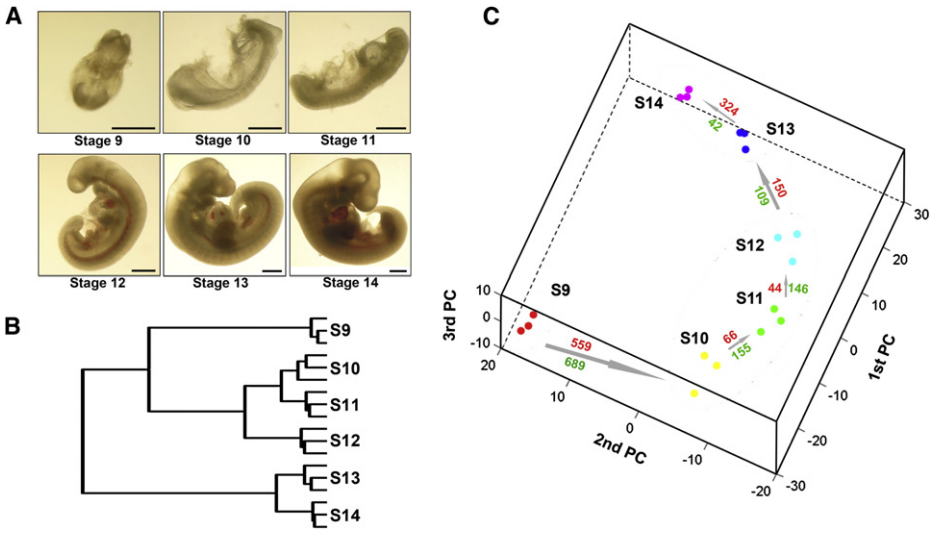

Figure 1. Global Analyses of the Transcriptome of Human Embryos during Early Organogenesis

(A) Morphological features of human embryos. According to the Carnegie criteria, the collected human embryos are grouped into six successive embryonic developmental stages (S9–S14). Scale bar in each photograph represents 1 mm.

(B) Transcriptome profiles of human embryos. Hierarchical classification analysis shows the reproducibility of transcriptome profiles of each staged embryos (sampled in triplicates) and divides the six developmental stages into three major branches: S9, S10–S12, and S13–S14.

(C) Stage-transitive transcriptome changes of human embryos. Human embryo samples are projected onto the three-dimensional space captured by principle component analysis (PCA). Each of the staged embryo samples are colored as indicated. Any two-successive-stage transition is illustrated using an arrow, adjacent to which shows the number of the transcripts significantly increasing (in red) or decreasing (in green) during the transition, as identified using Linear Models for Microarray Data (LIMMA). See also Figure S1 and Table S1.

define EB stages of hESC differentiation and evaluate cell populations reprogrammed to different extents.

Understanding of human early embryogenesis is also useful in regenerative medicine approaches aimed at differentiation of hESCs or iPSCs into functional cells (Gearhart, 2004). Recent evidence indicates that the most successful differentiation protocols for human pluripotent cells are those that most closely mimic the in vivo embryonic development of the particular cell lineage (Mayhew and Wells, 2010). A second potential application is in determining whether knowledge obtained from model organisms is truly representative of human development. The inaccessibility of the human embryo, especially at postimplantation stages, has long been a major limitation for the study of human early embryogenesis, and many studies have focused on mouse embryos instead. A recent study reported the transcriptome analysis of the mouse postimplantation embryo from gastrulation through early organogenesis and showed that morphological changes within the whole embryo are driven by molecular changes (Mitiku and Baker, 2007). However, it remains unknown to what extent the changes seen in mouse reflect developmental events in human embryos.

Overall, the limited nature of information about human early embryos has hampered many aspects of developmental biology and stem cell engineering. We describe here whole-genome expression array profiling of human postimplantation embryos at six successive time periods: Carnegie stages 9–14 (E20–E32), covering the first third of organogenesis. Using a range of data mining and information annotation approaches, we were able to identify a number of transcriptome features that may be significant for early human embryonic development.

RESULTS

Transcriptome Profiling of Human Embryos and Selection of Genes Informative to the Characterization of These Embryos

Human embryos from E20 to E32 were the most readily available from our clinic, spanning six successive Carnegie stages (i.e., S9–S14) that cover the first third of organogenesis. Embryos at these stages were collected, carefully staged based upon morphological criteria (Figure 1A; see Figures S1A and S1B available online) and then subjected to Affymetrix expression assays with three replicates (see Experimental Procedures). Three biological replicates were conducted for embryos at S10–S13; owing to practical limitations and material availability, embryos at S9 and S14 were pooled together and subjected to three technical replicate analyses (Figure S1A). After data normalization, an extraction of differential gene expression (EDGE)-based methodology (Storey et al., 2005) was applied to identify genes with expression that is consistent between replicates but differentially regulated across the various developmental stages. The resulting matrix contained expression measurements for 5441 transcripts across 18 samples, denoted as the human organogenesis (hORG) expression matrix (Table S1), and was used for the subsequent analyses. Independently, quantitative RT-PCR (qRT-PCR) was employed to validate genes with orthologs that are known to be developmental markers in mice, showing that these genes behaved as expected (Figure S1C). When hierarchical sample classification was applied to the hORG expression matrix, the 18 samples were clearly categorized into six sequential groups, corresponding to the six

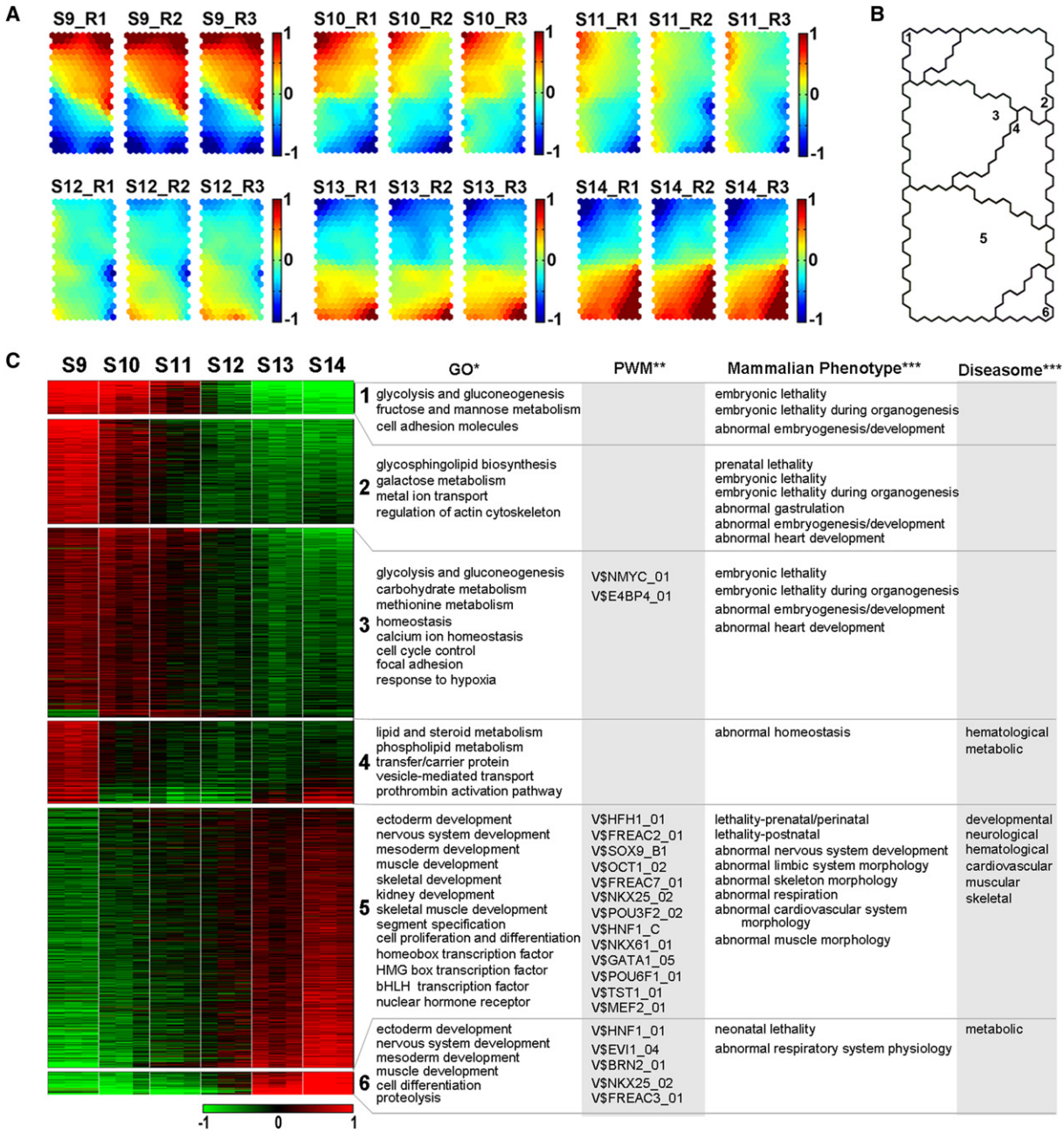

Figure 2. In-Depth Analyses of Transcriptome Features Characterizing Early Human Organogenesis

(A) Component plane presentation integrated self-organizing map (CPP-SOM) of human embryos, depicting dynamic transcriptome changes during early human organogenesis. Each presentation illustrates a sample-specific transcriptome map, in which all the up-regulated (in red), down-regulated (in blue), and moderately regulated (in yellow and green) genes are well delineated. Notably, the same position in all presentations contains the same group of coexpressed genes.

(B) Ideogram illustration of six gene clusters on a SOM grid map. These gene clusters are obtained through SOM-based two-phase gene clustering. The index of each cluster is marked in the seed neuron as indicated.

(C) Illustration of gene expression patterns and corresponding biological theme enrichments for each of the six gene clusters. Various biological annotations are mined to determine the enrichments of biological relevance, as highlighted by Gene Ontology (GO) and pathway for functional enrichments (*FDR <0.05), positional weighted matrix (PWM) of UCSC conserved transcription factor binding sites for regulatory enrichments (**FDR <0.01), Mouse Genome Informatics (MGI) phenotype ontology for mammalian phenotypic enrichments (***FDR <0.005), and Online Mendelian Inheritance in Man (OMIM) disorder-gene association information for disorder enrichments (***FDR <0.005). See also Figure S2 and Table S2.

successive Carnegie stages (Figure 1B; Figure S1D). Moreover, a dendrogram of the hierarchical sample classification (Figure 1B) and principal component analysis (PCA) (Figure 1C) showed that the S9 samples (E20–E21) were more distinct from those of S10 (E22–E23) as compared with other pairs of adjacent stages, suggesting that a major transition occurred

A

	Temporal patterns during early post-implantation mouse development		
	Organogenesis Clusters II, III, IV (591 #, 569 ※)	Gastrulation Clusters I, VI, VII (855 #, 809 ※)	Cell Cycle Cluster V (254 #, 246 ※)
Decreasing Clusters 1-3 (792 §)	51 (P=1.00)	76 ($P=5.28\times10^{-5}$)	18 (P=0.61)
Increasing Clusters 5-6 (690 §)	118 ($P=4.07\times10^{-12}$)	20 (P=1.00)	18 (P=0.30)

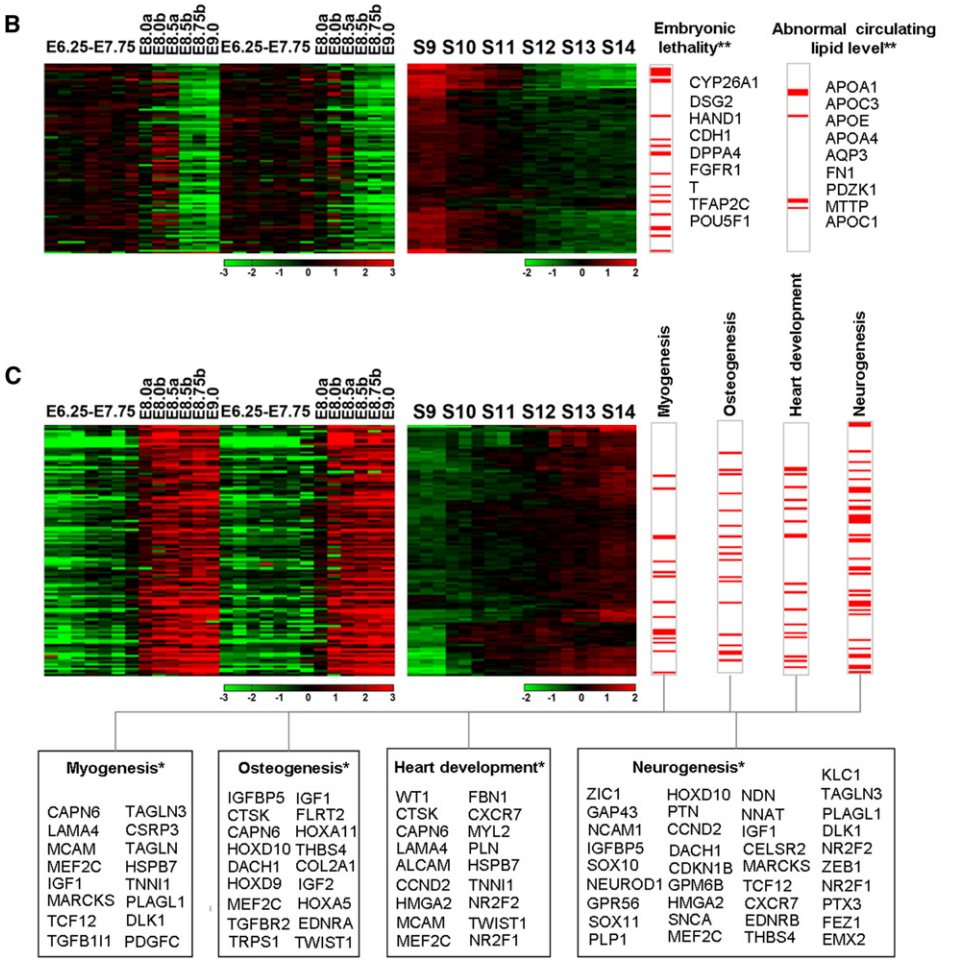

Figure 3. Illustration of Conserved Expression Patterns and Biological Characteristics of Human-Mouse Homologs during Early Organogenesis

(A) Comparison of temporal expression patterns during early organogenesis of human and mouse. The significance of overlaps was evaluated using Fisher's exact test. §Unique NCBI EntrezGenes in human. #Unique MGI IDs in mouse. ※The corresponding human homologs, transferred by INPARANOID homology data.

(B) Significant overlaps (termed as HM_HD-MD) between gastrulation clusters (I, VI, and VII) in mouse and decreasing clusters 1–3 in human. Mouse embryos spanning gastrulation and early organogenesis (E6.25–E9.0, in duplicates) are displayed in the left panel, followed by the stages of early human organogenesis. Enrichments of mammalian phenotypes are integrated on the right of the display (**FDR <0.01). Homologs harboring the specific enriched annotation are also marked in red in the bar.

(C) Significant overlaps (termed as HM_HI-MI) between organogenesis clusters (II, III, and IV) in mouse and increasing clusters 5 and 6 in human. Also, human-mouse homologs of HM_HI-MI annotated in myogeneisis, osteogeneisis, heart development, and neurogeneisis are listed (*FDR <0.05), serving as specific developmental markers to assess early mammalian organogenesis. See also Table S3.

between S9 and S10 at the transcriptome level. Consistent with this observation, only neural tissue and a few somites are morphologically evident at S9, whereas the emergence of diverse organ primordia occurs from S10–S12 (E26–E27) (Figure S1A).

Identification of Genes with Characteristic Expression Patterns in Early Organogenesis of Human Embryos

In an attempt to identify the transcriptome features inherent in the hORG expression matrix, we first applied an approach of combining the self-organizing map with singular value

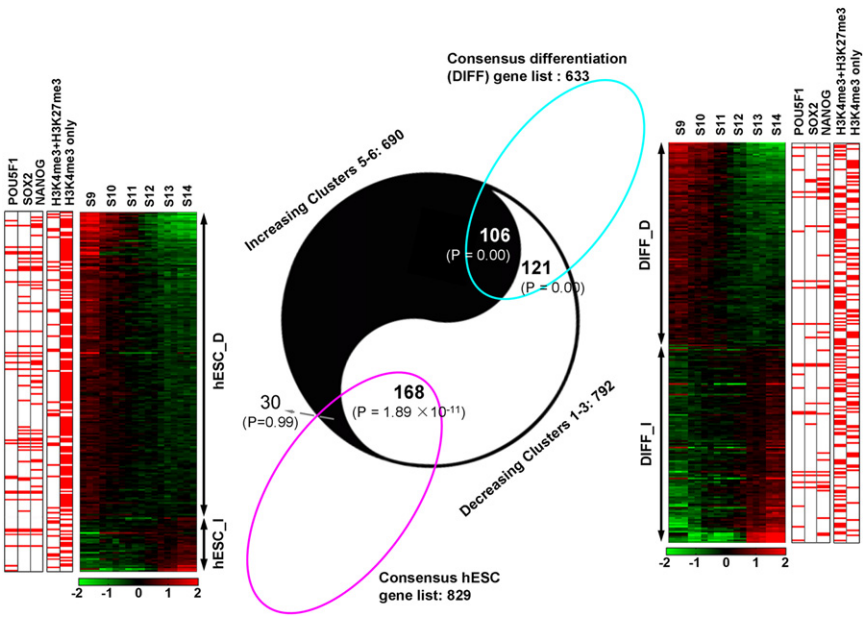

Figure 4. Coexistence of Stemness and Differentiation Potentials during Early Human Organogenesis

Yin-Yang Diagram includes the white part (representing the decreasing clusters 1–3) and the black part (representing increasing clusters 5 and 6). Those genes consistently overexpressed in hESCs (termed as "consensus hESC gene list") are circled in pink, while those underexpressed (termed as "consensus differentiation gene list") in cyan. Venn diagram in the middle illustrates their overlaps and the corresponding significance (Fisher's exact test). Overlaps of Consensus hESC genes with decreasing clusters 1–3 and increasing clusters 5 and 6 are labeled as hESC_D and hESC_I, respectively. The expression patterns of hESC_D and hESC_I are displayed in the left panel. The corresponding bars indicate genome-wide binding information of pluripotency-associated transcription factors POU5F1, SOX2, and NANOG and genome-wide histone modification sites of H3K4me3/H3K27me3 bivalent domains and H3K4me3 only domain. Likewise, the expression patterns for overlaps of consensus DIFF genes with decreasing clusters 1–3 (labeled as DIFF_D) and increasing clusters 5 and 6 (labeled as DIFF_I) are displayed in the right panel, with the regulatory and epigenetic information marked in red. See also Table S4.

decomposition (SOM-SVD) (Fang et al., 2008; Wang et al., 2009) for feature selection and artifact elimination (Figure S2). The resulting matrix containing expression measurements of 2148 transcripts across the 18 samples was then subjected to a gene clustering procedure (Vesanto and Sulkava, 2002), which revealed six clusters based on topological relationships (termed clusters 1–6) (see Table S2). To facilitate direct comparisons within/between developmental stages, component plane presentations (CPPs) (Xiao et al., 2003) were used to display sample-specific transcriptome changes (Figure 2A). The topological relationships of the six clusters are shown in Figure 2B. Genes within each of the six clusters displayed highly similar expression patterns (left panel of Figure 2C), suggesting that they may share common features. In addition, the expression patterns of genes in clusters 1, 2, and 3 showed overall similarity in terms of the gene expression level tending to be gradually repressed as development progressed, whereas the expression level of genes in clusters 5 and 6 gradually increased. Interestingly, the expression patterns of genes in these two major groups appeared to correlate well with the development potential of early embryonic cells, i.e., a gradual decrease in "stemness" and a concomitant increase in the diversity of cell types present.

Next, we performed enrichment analysis of the genes in these clusters to examine whether they shared functional or regulatory features using Gene Ontology (GO) (Mi et al., 2007) and the UCSC conserved transcription factor binding sites (TFBSs) (Miller et al., 2007). As shown in the middle panel of Figure 2C, significant GO terms in clusters 1–3 included those associated with cellular metabolism and homeostasis, but only a few significant TFBSs (or position weight matrixes [PWMs]) were observed, including survival-related transcription factors such as NMYC (Laurenti et al., 2008) and E4BP4 (Junghans et al., 2004). In contrast, the most significant GO terms in clusters 5 and 6 were more diverse, representing a wide spectrum of functions involved in the establishment of organ morphogenesis. In keeping with this functional spectrum, the most significant TFBSs included those of multiple organogenesis-related regulators such as the nervous system-specific OCT1 (Jin et al., 2009) and BRN2 (Castro et al., 2006), the muscle-specific MEF2 (Olson et al., 1995), the heart-specific NKX2-5 (Pashmforoush et al., 2004), and the skeletal-specific SOX5 (Smits et al., 2001).

Next, we used mammalian phenotype ontology from the Mouse Genome Informatics (MGI) (Bult et al., 2008) and the disorder-gene association information from Online Mendelian Inheritance in Man (OMIM) (Goh et al., 2007) to conduct enrichment analysis of the genes in the clusters. As illustrated in the right panel of Figure 2C, significant phenotypes related to genes clusters 1–3 were mostly associated with embryonic lethality and abnormal embryogenesis, whereas the genes in clusters 5 and 6 were primarily linked to postnatal lethality and diverse organ/system defects. Likewise, hardly any genetic disorders were linked to genes in clusters 1–3, but, in sharp contrast, many such disorders were linked to genes in clusters 5 and 6, including neurological, hematological, and cardiovascular disorders.

Taken together, above observations suggest that the genes in cluster 1–3 may play crucial roles in the initiation of organogenesis, whereas those in clusters 5 and 6 may primarily take part in the establishment of organogenesis. As genes in cluster 4 topologically share boundaries with both clusters 1–3 and clusters 5 and 6 (Figure 2B), with more diverse expression patterns (Figure 2C) and only representing a small percentage (11.8%), these genes were excluded from further consideration in this study.

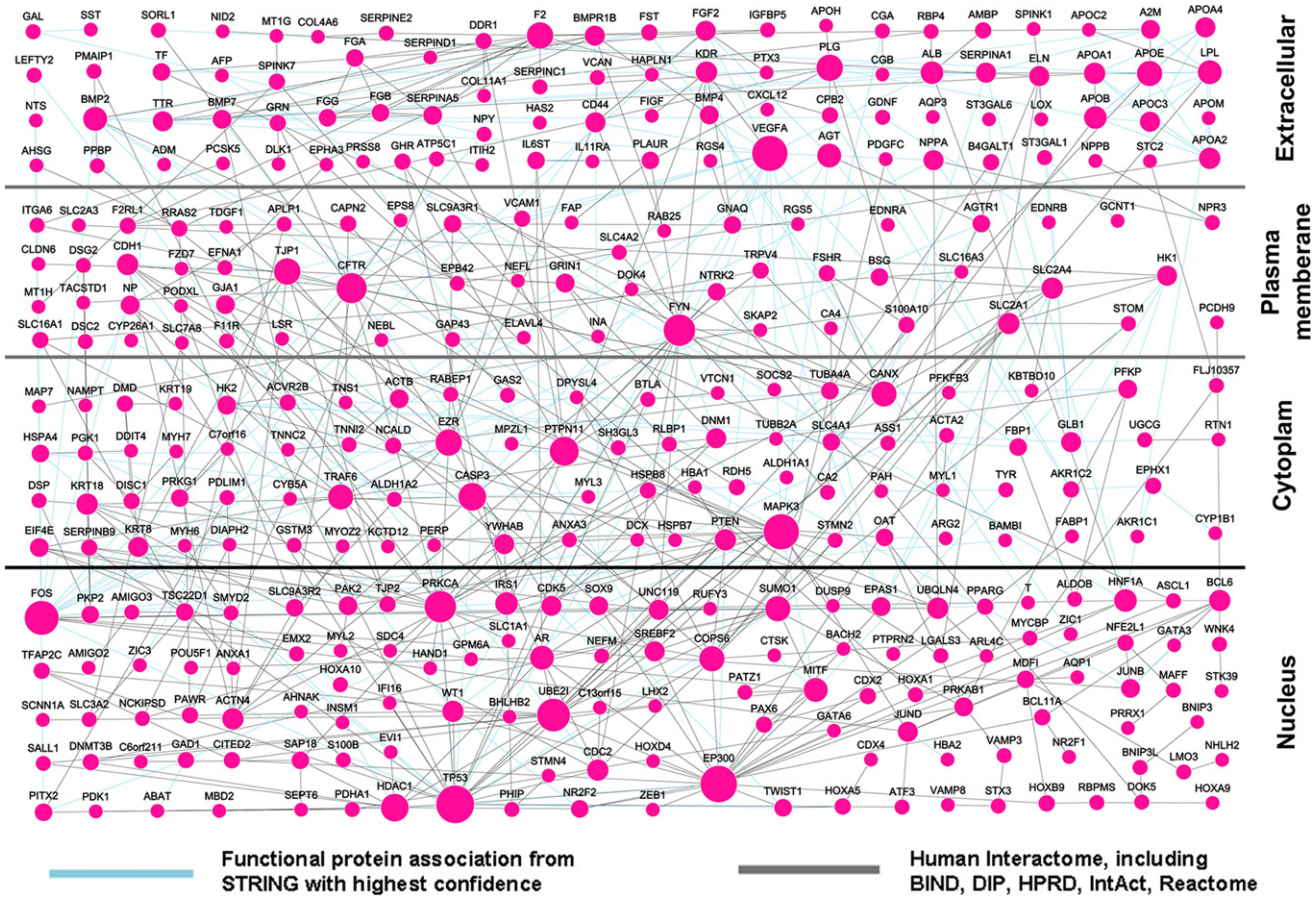

Figure 5. A Putative Molecular Interaction Network during Early Human Organogenesis

A connected network (hORGNet) shows significant changes in expression during early human organogenesis. The area of the gene node (circle) is proportional to node degree (i.e., the number of edges directly linked to the node). Highlighted in gray are the edges derived from the human interactome, while the edges in sky-blue are functional protein associations from the STRING database. Layouts of the hORGNet are based on subcellular localization information annotated in NCBI GO. The hORGNet is inherited with a stemness-relevant module (hStemModule), which dominates a hESC-context network (hESCNet). See also Figure S3 and Table S5.

Human-Mouse Comparative Analysis of Genes Important during Early Organogenesis

To further validate the above findings, we conducted a comparative analysis relative to previously published transcriptome data from mouse embryos covering gastrulation to early organogenesis (E6.25–E9) (Mitiku and Baker, 2007). In this study, gene clusters specific for gastrulation (i.e., clusters I, VI, and VII) and organogenesis (i.e., clusters II, III, and IV) were defined, in addition to a defined cell-cycle specific cluster (i.e., cluster V). Using human-mouse homologs in INPARANOID homology data (Berglund et al., 2008), we found that genes in clusters 1–3 were significantly ($p = 5.28 \times 10^{-5}$, Fisher's exact test) represented in the mouse gastrulation-specific clusters, whereas genes in clusters 5 and 6 were significantly ($p = 4.07 \times 10^{-12}$) enriched in the mouse organogenesis-specific clusters (Figure 3A). For the genes in the mouse cell cycle cluster V, we found no significant overlap with either clusters 1–3 ($p = 0.61$) or clusters 5 and 6 ($p = 0.30$).

When human-mouse homologs from clusters 1–3 were selected (labeled as HM_HD-MD) (see Table S3) and subjected to enrichment analysis using MGI mammalian phenotypes, embryonic lethality appeared to be the most prominent feature, consistent with the result shown in Figure 2C and associated with pluripotency-associated or germ-layer master genes within this group (e.g., *POU5F1*, *DPPA4*, *T*, *DSG2*, and *HAND1*) (Figure 3B). Another phenotype associated with these homologs was abnormal circulation of lipids, which could be attributed to the apolipoproteins-encoded genes (e.g., *APOA1*, *APOC1*, *APOC3*, *APOE*, *APOA4*, and *MTTP*), when disrupted in mice. In contrast, the human-mouse homologs from clusters 5 and 6 (labeled as HM_HI-MI) (see Table S3) were largely involved in myogenesis, osteogenesis, heart development, and neurogenesis (Figure 3C). These results further suggest that the genes in clusters 1–3 are primarily important for the initiation of organogenesis, whereas those in clusters 5 and 6 are important for the progression of organogenesis, and are probably involved in various types of cell differentiation or organ formation.

Integration of Multiple Layers of Information in hESCs to Identify Stemness- and Differentiation-Relevant Genes during Early Human Organogenesis

We next performed an enrichment analysis using a published data set from hESCs, from which a consensus hESC gene list

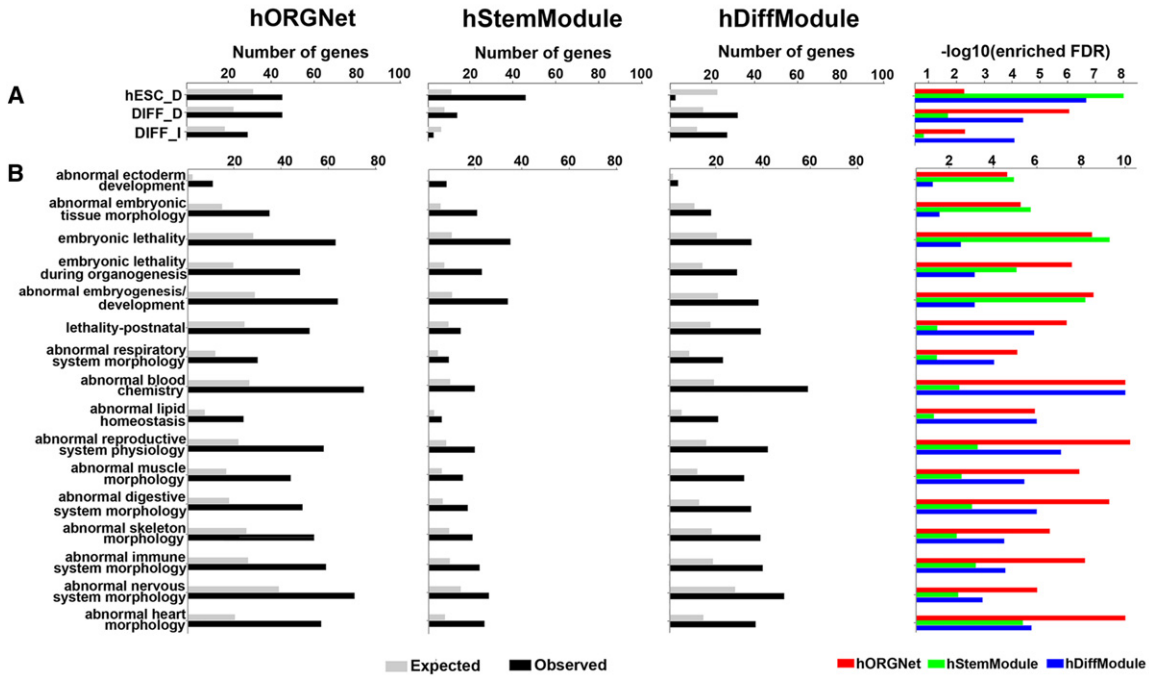

Figure 6. The Characteristics during Early Organogenesis of Human and Mouse Captured By the hORGNet and Its Two Modules (hStemModule and hDiffModule)

Enrichment analyses of the hORGNet, the hStemModule, and the hDiffModule (A) in the context of stemness-relevant genes (hESC_D) and differentiation-relevant genes (DIFF_D and DIFF_I), as identified in Figure 4 (B) in the context of the MGI mammalian phenotype ontology. The first column shows the observed number of genes overlapped between the hORGNet and specific gene lists (in black) verses the expected number of genes under null distribution (in gray). Similarly, the second and third columns are displayed for the hStemModule and the hDiffModule, respectively. The significance of the corresponding enrichments in the hORGNet, the hStemModule and the hDiffModule is illustrated in the right-most column.

and a consensus differentiation (DIFF) gene list were previously defined (Assou et al., 2007). Fisher's exact test was applied to compare these two gene lists with clusters 1–3 and clusters 5 and 6. As shown in the bottom-left of the middle panel in Figure 4, the defined consensus hESC genes were significantly overrepresented in clusters 1–3 ($p = 1.89 \times 10^{-11}$; the overlapping genes were denoted as hESC_D) (see Table S4), whereas no overrepresentation was observed in clusters 5 and 6 ($p = 0.99$). For the defined consensus DIFF genes, however, highly significant overrepresentation was observed in both clusters 1–3 ($p = 0.00$; overlaps were denoted as DIFF_D) and clusters 5 and 6 ($p = 0.00$; overlaps were denoted as DIFF_I) (the upper right of the middle panel of Figure 4).

The pluripotency-associated transcription factors POU5F1 (OCT4), NANOG, and SOX2 constitute a core transcriptional regulatory circuitry controlling stem cell identity (Boyer et al., 2005). Also, the histone modification H3K4me3 is associated with active promoters of genes that maintain the fundamental properties of hESCs, whereas H3K4me3/H3K27me3 bivalent modifications are linked to silenced promoters of genes that are poised for expression upon differentiation (Pan et al., 2007). We integrated above layers of information with our data set to provide an additional perspective on the overrepresented consensus hESC genes in clusters 1–3 (i.e., hESC_D), and the overrepresented consensus DIFF genes in clusters 1–3 (i.e., DIFF_D) and in clusters 5 and 6 (i.e., DIFF_I). As shown in Table S5, the potential binding sites of POU5F1, NANOG, and SOX2 were significantly enriched in hESC_D relative to DIFF_D or DIFF_I. Similarly, H3K4me3 modification was significantly enriched in the hESC_D group, whereas H3K4me3/H3K27me3 bivalent modifications were significantly enriched in both DIFF_D and DIFF_I (Table S5). The left panel of Figure 4 also illustrates a high frequency of H3K4me3 modification and triple occupancy by the pluripotency factors in genes of the hESC_D, which suggests that these genes may be important in controlling stem cell identity. In contrast, genes in the DIFF_D or DIFF_I groups showed features characteristic of cellular differentiation at both genetic and epigenetic levels (the right panel of Figure 4).

Considering these results together, it is tempting to assume that clusters 1–3 may contain at least two groups of genes essential for the initiation of organogenesis, one of which maintains the fundamental properties of hESCs (i.e., stemness) and the other of which is crucial for cellular differentiation. On the other hand, it appears that the genes in clusters 5 and 6 are involved in various types of cellular differentiation. These findings also provide further support for the idea that these clusters of genes are important for the initiation and progression of human organogenesis.

Integrative Mining Defines a Putative Molecular Network Depicting Early Human Organogenesis

Genes do not function in isolation and instead are interconnected into molecular networks that control biological processes

such as embryogenesis. It is appealing to hypothesize that the stemness- and differentiation-relevant genes identified in this setting are probably connected in an overall functional network. To investigate this idea, we employed the Cytoscape plug-in jActiveModules (Ideker et al., 2002) which integrates expression and interaction information to identify the expression-related subnetworks. Based on the intrinsic features of the hORG expression matrix in the context of the compiled human interaction/association network (for details see Supplementary Experimental Procedures), we detected a subnetwork (hORGNet) (see Table S5). Figure 5 shows the layout of the hORGNet configured by subcellular location using Cerebral (Barsky et al., 2007). In parallel, we employed the same approach to investigate stemness-relevant networks by limiting the input of the interaction information to those only involving stemness-relevant genes (i.e., hESC_D and hESC_I in the left panel of Figure 4). As a result, we obtained a subnetwork specific for stemness (hESCNet) (see Table S5). By comparing the obtained hESCNet with the hORGNet, we found that as many as 76% (105/129) of the genes in the hESCNet were also part of the hORGNet, implying that the hORGNet contains a stemness-relevant module (termed as hStemModule), represented in the leftmost region of Figure 5. Then, we tested the modularity of the hStemModule by examining the distribution of the pairwise shortest distance between genes in the hStemModule as well as in the hORGNet. Compared with the hORGNet, genes in the hStemModule showed closer topological distances ($p < 10^{-40}$, Kolmogorov-Smirnov test) (Figure S3A), suggesting that the hStemModule may function as a relatively distinct module in the hORGNet. As most of genes in the hORGNet were differentiation-associated (Figures S3B and S3C), the remainder of the hORGNet was defined as a differentiation-associated module (hDiffModule). As most genes in the hStemModule were gradually suppressed during early human organogenesis, whereas genes in the hDiffModule experienced more dynamic regulation (Figure S3D), we concluded that the stemness-associated hStemModule and the differentiation-associated hDiffModule, might coordinately regulate early organogenesis in human embryos within an overall molecular network.

In a converse manner, we then statistically tested for enrichment of the genes illustrated in Figure 4 (i.e., hESC_D, DIFF_D, and DIFF_I) in the hStemModule and in the hDiffModule. As shown in the second column of Figure 6A, genes of the hESC_D group but not the DIFF_D or DIFF_I groups were overrepresented in the hStemModule. Likewise, genes of the DIFF_D and DIFF_I groups but not the hESC_D group were overrepresented in the hDiffModule (third column in Figure 6A). These results provide further support for the idea that there exist stemness and differentiation modules within the overall framework of molecular networks, which may coordinately orchestrate early organogenesis in human embryos. Moreover, enrichment analyses of the hStemModule and the hDiffModule using MGI phenotype ontology suggested that the hStemModule is related to early embryonic morphological abnormalities (the second column of Figure 6B), whereas the hDiffModule is related to diverse organ/system defects (the third column of Figure 6B). More importantly, genes in the hStemModule were largely linked to embryonic lethality (the second column of Figure 6B), whereas genes in the hDiffModule were mostly linked to postnatal lethality (the third column of Figure 6B). Based on such comparative analysis of human-mouse homologs, it is deducible that genes in these two modules are probably crucial for ensuring the survival and normality of early embryos.

DISCUSSION

Transcriptome profiling of human embryos provides a useful tool for advancing our understanding of human development and for stem cell engineering. Using whole-genome expression arrays, we have profiled human embryos from Carnegie stages 9 to 14, covering the first third of organogenesis (Figure 1 ; Figure S1). Through in-depth data mining, we identified two major groups of genes whose expression patterns are consistent with the dynamic nature of early embryonic cells, i.e., gradually reduced stemness potential and concomitantly increased diversity of cell types as development progresses (Figure 2). Integration of multilayered information from mouse embryos and hESCs (Figures 3 and 4) allowed us to further divide the group of genes whose express levels were gradually reduced (clusters 1–3) into a stemness specific subgroup and a differentiation associated subgroup. Likewise, we were able to identify the group of genes whose expression levels were gradually increased (clusters 5 and 6) as differentiation-associated genes. Using advanced molecular network techniques, we were able to propose a putative molecular network within which a stemness-specific module and a differentiation-associated module could be defined (Figures 5 and 6). Although this putative molecular network remains to be validated through functional analyses, we can conclude that stemness-specific and differentiation-associated genes, as identified in this setting, are fundamentally important in orchestrating early organogenesis of human embryos.

Large-scale transcriptome analysis of this critical developmental window provides a wealth of information for studying mammalian developmental biology. For instance, mouse is one of the most widely used model organisms, and its application has facilitated many aspects of investigation in developmental biology. However, the extent to which mouse embryos are similar to human is still unclear. As shown in Figure 3, using comparative analysis we have found that genes of clusters 1–3 were significantly represented in the defined mouse gastrulation-specific clusters, whereas genes of clusters 5 and 6 were significantly enriched in the defined mouse organogenesis-specific clusters. These shared expression patterns and core functional roles of human-mouse homologs enhance our understanding of early mammalian organogenesis from an evolutionary perspective. Using functional annotation database, we also performed a computational survey looking into the functions of genes that are in human embryos but not previously implicated in mouse embryos (Table S3). This preliminary analysis reveals that most of these genes are likely to be of functional relevance to organogenesis. Still, it would be very interesting to undertake the follow-up experimental investigation of genes unique to human embryos so as to clarify their exact functional roles in early human organogenesis.

Another important area in which our data may be useful is stem cell biology. As demonstrated in Figure 4, by comparative analysis using published consensus pluripotency and differentiation genes from hESCs, followed by the integration of

information of relevant TFBSs and histone modifications, we were able to define stemness-specific and differentiation-associated genes in our clusters. As hESCs are propagated in vitro, inevitably with some artifacts, our comparative analysis may provide cross-validation and thus permit the recognition of bona fide stemness-specific/differentiation-specific genes in vivo. As many aspects of stem cell studies to date are based on in vitro culture, using our data and results as a reference for these aspects may be particularly valuable. Although we have analyzed our data intensively and identified a number of important features, there is still a lot of information within the data set that remains to be elucidated, which may provide additional insights for development biology and stem cell engineering. In the future, lineage- or organ-specific information about human embryos at comparable developmental stages could further refine the results we have obtained from whole-embryo analysis.

EXPERIMENTAL PROCEDURES

Human Embryo Collection

The protocol for collecting human embryos was reviewed and approved by the Ethical Review Board of the Xinhua Hospital, Shanghai, China. The experimental procedures involving human embryos conformed to the National Ethical Guideline on Human Embryo Research issued by the Committee on Bioethics, Chinese National Human Genome Center (Southern Headquarter). Human embryos were obtained from the Department of Obstetrics and Gynecology at the Xinhua Hospital during clinical drug abortion. All donors signed informed consent forms (see Supplemental Experimental Procedures). The age of the embryo was carefully determined according to the standard protocol (Carlson, 2004) (Figures S1A and S1B).

Gene Expression Profiling and Data Preprocessing

For embryos staged based on the Carnegie criteria (Figures S1A and S1B), the gene expression profiling was performed using the Affymetrix HG-U133A Genechip microarrays (Affymetrix, Santa Clara, CA) according to the standard protocol. Raw expression data were normalized using robust multiarray averaging (RMA) with quantile normalization. The pairwise Pearson's correlation coefficient was calculated to show a high degree of reproducibility of the embryo collection and transcriptome profiles (Figure S1D). The expression data were imported into the EDGE software (Storey et al., 2005) for detection of probesets/transcripts exhibiting consistent changes within the triplicates as well as differential expression across six developmental stages. Under Q-value thresholds of 0.001, the resulting 5441 transcripts across 18 samples (hORG expression matrix) were remained as the representative information for the characterization of human embryos. LIMMA bioconductor library (Gentleman et al., 2004) was used to identify stage-transitive transcriptome changes. The criteria for identifying the top significant probesets/transcripts were based on Benjamini and Hochberg-derived FDR (<0.01).

The Topology-Preserving Identification of Temporal Expression Patterns

The hORG expression matrix was subjected to the topology-preserving feature selection through SOM-SVD (Fang et al., 2008; Wang et al., 2009). The resulting data were then subjected to SOM-based two-phase gene clustering (Vesanto and Sulkava, 2002; Xiao et al., 2003). Subsequently, six clusters were identified based on topological relationships. See details in the Supplemental Experimental Procedures.

Comparisons of Gene Expression Patterns during Early Organogenesis between Human and Mouse

The transcriptome data of the mouse embryos covering from gastrulation to organogenesis (E6.25–E9.0, in duplicates) (Mitiku and Baker, 2007) were obtained from NCBI GEO (GSE9046). Gastrulation clusters (I, VI, and VII), organogenesis clusters (II, III, and IV) and cell cycle cluster V were then used for the comparison with clusters 1–3 and clusters 5 and 6 during early human organogenesis. Mouse genes were transferred by orthology to human homologs using INPARANOID homology data (Berglund et al., 2008). The significance of overlaps for each comparison was evaluated using the Fisher's exact test.

Multiple-Layer Genomic Data Sources Relevant to hESCs

Genes consistently overexpressed in hESCs (termed as "consensus hESC gene list") and genes underexpressed ("consensus differentiation gene list") were obtained according to a recent meta-analysis of transcriptomes in hESCs (Assou et al., 2007). Genome-wide binding information concerning pluripotency-associated transcription factors POU5F1 (OCT4), NANOG, and SOX2 was derived from the published study (Boyer et al., 2005), and genome-wide histone modification sites of H3K4me3 and H3K27me3 from the published report (Pan et al., 2007). Their comparisons with transcriptome data of human embryos were carried out using the Fisher's exact test.

Detection of the Expression-Active Subnetworks

The Cytoscape plug-in jActiveModules (Ideker et al., 2002) was modified to identify connected subnetworks from a human interaction network. This overall network includes a compiled human physical interactome (Bader et al., 2003; Salwinski et al., 2004; Kerrien et al., 2007; Mishra et al., 2006; Vastrik et al., 2007) and human protein-protein associatome from STRING (highest confidence; ≥ 0.9) (von Mering et al., 2007). The identified subnetworks contain groups of highly linked genes, most of which show dominant expression patterns during early human organogenesis (see Supplemental Experimental Procedures for details). Cerebral (Barsky et al., 2007) was used to visualize the subnetwork, which is configured based on subcellular location information of genes. These location data were obtained from the NCBI GO Cellular Component categories including the "nucleus" (GO:0005634), "cytoplasm" (GO:0005737), "plasma membrane" (GO:0005886), and "extracellular region" (GO:0005576).

Enrichment Analyses Using Various Biological Annotations

Hypergeometric distribution-based enrichment analyses (Wang et al., 2009) were performed to interpret gene groups of interest, using diverse external annotated databases including Gene Ontology (Mi et al., 2007), the UCSC conserved transcription factor binding sites (Miller et al., 2007), the MGI mammalian phenotype ontology (Bult et al., 2008), and OMIM disorder-gene association information (Goh et al., 2007). Besides these annotations, gene groups of interest identified in this study (i.e., hESC_D, DIFF_D, DIFF_I) were also used for enrichment analyses to examine their relevance to the hORGNet, the hStemModule, and the hDiffModule. Benjamini and Hochberg-derived FDR were applied to assess the significance of the enrichments. See details in the Supplemental Experimental Procedures.

For the enrichment analysis of gene groups (e.g., clusters 1–3 and clusters 5 and 6 in Figure 2C) using MGI mammalian phenotype ontology, we first retrieved mouse knockout phenotypes together with human-mouse homologs from the Mouse Genome Informatics (MGI) database (Bult et al., 2008), and then assessed their associations to mouse knockout phenotypes based on the hypergeometric distribution followed by multiple hypothesis tests (Wang et al., 2009). Only those phenotypes statistically enriched within the gene group were selected. In other words, these genes, once genetically disrupted, were likely to cause these phenotypes.

ACCESSION NUMBERS

The transcriptome profilings of human embryos during early organogenesis are deposited in NCBI GEO under accession number GSE18887.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, three figures, and five tables and can be found with this article online at doi:10.1016/j.devcel.2010.06.014.

ACKNOWLEDGMENTS

We thank Drs. Dangsheng Li, Tingxi Liu, and Zeguang Han for critical reading and helpful advice during preparation of the manuscript. We also thank Weicheng Wang and Lili Zhu for technical assistance. The study was supported by grants from the National Natural Science Foundation of China (30871257, 30730051, 90919059, 30730033, and 30971623), the National High Technology Research and Development Program of China (2006AA02Z197, 2006CB943901, 2006CB91040, 2007CB947904, 2007AA02Z335, 2007CB948004, 2009CB941103, and 2010CB945201), the Shanghai Leading Academic Discipline Project (S30201), Shanghai Postdoctoral Scientific Program (09R21414900), China Postdoctoral Science Foundation (20090450573), Shanghai Science & Technology Developmental Foundation (06DJ14001), and the Knowledge Innovation Program of Chinese Academy of Sciences (KSCX2-YW-R-46, KSCX2-YW-R-19 and KSCX1-YW-22-01). J.Z. is a member of the TB-VIR network (European Community Grants of FP7, 200973).

Received: May 29, 2008
Revised: January 29, 2010
Accepted: June 8, 2010
Published: July 19, 2010

REFERENCES

Abeyta, M.J., Clark, A.T., Rodriguez, R.T., Bodnar, M.S., Pera, R.A., and Firpo, M.T. (2004). Unique gene expression signatures of independently-derived human embryonic stem cell lines. Hum. Mol. Genet. *13*, 601–608.

Assou, S., Le Carrour, T., Tondeur, S., Strom, S., Gabelle, A., Marty, S., Nadal, L., Pantesco, V., Reme, T., Hugnot, J.P., et al. (2007). A meta-analysis of human embryonic stem cells transcriptome integrated into a web-based expression atlas. Stem Cells *25*, 961–973.

Bader, G.D., Betel, D., and Hogue, C.W. (2003). BIND: the Biomolecular Interaction Network Database. Nucleic Acids Res. *31*, 248–250.

Barsky, A., Gardy, J.L., Hancock, R.E., and Munzner, T. (2007). Cerebral: a Cytoscape plugin for layout of and interaction with biological networks using subcellular localization annotation. Bioinformatics *23*, 1040–1042.

Berglund, A.C., Sjolund, E., Ostlund, G., and Sonnhammer, E.L. (2008). InParanoid 6: eukaryotic ortholog clusters with inparalogs. Nucleic Acids Res. *36*, D263–D266.

Boyer, L.A., Lee, T.I., Cole, M.F., Johnstone, S.E., Levine, S.S., Zucker, J.P., Guenther, M.G., Kumar, R.M., Murray, H.L., Jenner, R.G., et al. (2005). Core transcriptional regulatory circuitry in human embryonic stem cells. Cell *122*, 947–956.

Bult, C.J., Eppig, J.T., Kadin, J.A., Richardson, J.E., and Blake, J.A. (2008). The Mouse Genome Database (MGD): mouse biology and model systems. Nucleic Acids Res. *36*, D724–D728.

Carlson, B.M. (2004). Human Embryology and Developmental Biology (St. Louis: Mosby).

Castro, D.S., Skowronska-Krawczyk, D., Armant, O., Donaldson, I.J., Parras, C., Hunt, C., Critchley, J.A., Nguyen, L., Gossler, A., Gottgens, B., et al. (2006). Proneural bHLH and Brn proteins coregulate a neurogenic program through cooperative binding to a conserved DNA motif. Dev. Cell *11*, 831–844.

Dvash, T., Mayshar, Y., Darr, H., McElhaney, M., Barker, D., Yanuka, O., Kotkow, K.J., Rubin, L.L., Benvenisty, N., and Eiges, R. (2004). Temporal gene expression during differentiation of human embryonic stem cells and embryoid bodies. Hum. Reprod. *19*, 2875–2883.

Fang, H., Wang, K., and Zhang, J. (2008). Transcriptome and proteome analyses of drug interactions with natural products. Curr. Drug Metab. *9*, 1038–1048.

Gearhart, J. (2004). New human embryonic stem-cell lines–more is better. N. Engl. J. Med. *350*, 1275–1276.

Gentleman, R.C., Carey, V.J., Bates, D.M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., et al. (2004). Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. *5*, R80.

Goh, K.I., Cusick, M.E., Valle, D., Childs, B., Vidal, M., and Barabasi, A.L. (2007). The human disease network. Proc. Natl. Acad. Sci. USA *104*, 8685–8690.

Hanna, J., Saha, K., Pando, B., van Zon, J., Lengner, C.J., Creyghton, M.P., van Oudenaarden, A., and Jaenisch, R. (2009). Direct cell reprogramming is a stochastic process amenable to acceleration. Nature *462*, 595–601.

Ideker, T., Ozier, O., Schwikowski, B., and Siegel, A.F. (2002). Discovering regulatory and signalling circuits in molecular interaction networks. Bioinformatics *18* (*Suppl 1*), S233–S240.

Jin, Z., Liu, L., Bian, W., Chen, Y., Xu, G., Cheng, L., and Jing, N. (2009). Different transcription factors regulate nestin gene expression during P19 cell neural differentiation and central nervous system development. J. Biol. Chem. *284*, 8160–8173.

Junghans, D., Chauvet, S., Buhler, E., Dudley, K., Sykes, T., and Henderson, C.E. (2004). The CES-2-related transcription factor E4BP4 is an intrinsic regulator of motoneuron growth and survival. Development *131*, 4425–4434.

Kerrien, S., Alam-Faruque, Y., Aranda, B., Bancarz, I., Bridge, A., Derow, C., Dimmer, E., Feuermann, M., Friedrichsen, A., Huntley, R., et al. (2007). IntAct–open source resource for molecular interaction data. Nucleic Acids Res. *35*, D561–D565.

Laurenti, E., Varnum-Finney, B., Wilson, A., Ferrero, I., Blanco-Bose, W.E., Ehninger, A., Knoepfler, P.S., Cheng, P.F., MacDonald, H.R., Eisenman, R.N., et al. (2008). Hematopoietic stem cell function and survival depend on c-Myc and N-Myc activity. Cell Stem Cell *3*, 611–624.

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

Mayhew, C.N., and Wells, J.M. (2010). Converting human pluripotent stem cells into beta-cells: recent advances and future challenges. Curr. Opin. Organ Transplant. *15*, 54–60.

Mi, H., Guo, N., Kejariwal, A., and Thomas, P.D. (2007). PANTHER version 6: protein sequence and function evolution data with expanded representation of biological pathways. Nucleic Acids Res. *35*, D247–D252.

Miller, W., Rosenbloom, K., Hardison, R.C., Hou, M., Taylor, J., Raney, B., Burhans, R., King, D.C., Baertsch, R., Blankenberg, D., et al. (2007). 28-way vertebrate alignment and conservation track in the UCSC Genome Browser. Genome Res. *17*, 1797–1808.

Mishra, G.R., Suresh, M., Kumaran, K., Kannabiran, N., Suresh, S., Bala, P., Shivakumar, K., Anuradha, N., Reddy, R., Raghavan, T.M., et al. (2006). Human protein reference database–2006 update. Nucleic Acids Res. *34*, D411–D414.

Mitiku, N., and Baker, J.C. (2007). Genomic analysis of gastrulation and organogenesis in the mouse. Dev. Cell *13*, 897–907.

Olson, E.N., Perry, M., and Schulz, R.A. (1995). Regulation of muscle differentiation by the MEF2 family of MADS box transcription factors. Dev. Biol. *172*, 2–14.

Pan, G., Tian, S., Nie, J., Yang, C., Ruotti, V., Wei, H., Jonsdottir, G.A., Stewart, R., and Thomson, J.A. (2007). Whole-genome analysis of histone H3 lysine 4 and lysine 27 methylation in human embryonic stem cells. Cell Stem Cell *1*, 299–312.

Pashmforoush, M., Lu, J.T., Chen, H., Amand, T.S., Kondo, R., Pradervand, S., Evans, S.M., Clark, B., Feramisco, J.R., Giles, W., et al. (2004). Nkx2-5 pathways and congenital heart disease; loss of ventricular myocyte lineage specification leads to progressive cardiomyopathy and complete heart block. Cell *117*, 373–386.

Salwinski, L., Miller, C.S., Smith, A.J., Pettit, F.K., Bowie, J.U., and Eisenberg, D. (2004). The Database of Interacting Proteins: 2004 update. Nucleic Acids Res. *32*, D449–D451.

Smith, K.P., Luong, M.X., and Stein, G.S. (2009). Pluripotency: toward a gold standard for human ES and iPS cells. J. Cell. Physiol. *220*, 21–29.

Smits, P., Li, P., Mandel, J., Zhang, Z., Deng, J.M., Behringer, R.R., de Crombrugghe, B., and Lefebvre, V. (2001). The transcription factors L-Sox5 and Sox6 are essential for cartilage formation. Dev. Cell *1*, 277–290.

Storey, J.D., Xiao, W., Leek, J.T., Tompkins, R.G., and Davis, R.W. (2005). Significance analysis of time course microarray experiments. Proc. Natl. Acad. Sci. USA *102*, 12837–12842.

Sudheer, S., and Adjaye, J. (2007). Functional genomics of human pre-implantation development. Brief. Funct. Genomics Proteomic. *6*, 120–132.

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

Thomson, J.A., Itskovitz-Eldor, J., Shapiro, S.S., Waknitz, M.A., Swiergiel, J.J., Marshall, V.S., and Jones, J.M. (1998). Embryonic stem cell lines derived from human blastocysts. Science *282*, 1145–1147.

Vastrik, I., D'Eustachio, P., Schmidt, E., Joshi-Tope, G., Gopinath, G., Croft, D., de Bono, B., Gillespie, M., Jassal, B., Lewis, S., et al. (2007). Reactome: a knowledge base of biologic pathways and processes. Genome Biol. *8*, R39.

Vesanto, J., and Sulkava, M. (2002). Distance matrix based clustering of the self-organizing map. In Artificial Neural Networks - Icann 2002 (London: Springer-Verlag), pp. 951–956.

von Mering, C., Jensen, L.J., Kuhn, M., Chaffron, S., Doerks, T., Kruger, B., Snel, B., and Bork, P. (2007). STRING 7–recent developments in the integration and prediction of protein interactions. Nucleic Acids Res. *35*, D358–D362.

Waddington, C.H. (1957). The strategy of the genes: A discussion of some aspects of theoretical biology (New York: Macmillan).

Wang, K., Fang, H., Xiao, D., Zhu, X., He, M., Pan, X., Shi, J., Zhang, H., Jia, X., Du, Y., et al. (2009). Converting redox signaling to apoptotic activities by stress-responsive regulators HSF1 and NRF2 in fenretinide treated cancer cells. PLoS ONE *4*, e7538.

Xiao, L., Wang, K., Teng, Y., and Zhang, J. (2003). Component plane presentation integrated self-organizing map for microarray data analysis. FEBS Lett. *538*, 117–124.

Yamanaka, S. (2009). Elite and stochastic models for induced pluripotent stem cell generation. Nature *460*, 49–52.

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