A peer-reviewed version of this preprint was published in PeerJ on 23 October 2014.

View the peer-reviewed version (peerj.com/articles/618), which is the preferred citable publication unless you specifically need to cite this preprint.

https://doi.org/10.7717/peerj.618
Network based meta-analysis prediction of microenvironmental relays involved in stemness of human embryonic stem cells

**Background.** Human embryonic stem cells (hESCs) are pluripotent cells derived from the inner cell mass of *in vitro* fertilised blastocysts, which can either be maintained in an undifferentiated state or committed into lineages under determined culture conditions. These cells offer great potential for regenerative medicine, but at present, little is known about the mechanisms that regulate hESC stemness; in particular, the role of cell-cell and cell-extracellular matrix interactions remain relatively unexplored. **Methods and results.** In this study we have performed an *in silico* analysis of cell-microenvironment interactions to identify novel proteins that may be responsible for the maintenance of hESC stemness. A hESC transcriptome of 8,934 mRNAs was assembled using a meta-analysis approach combining the analysis of microarrays and the use of databases for annotation. The STRING database was utilised to construct a protein-protein interaction network focused on extracellular and transcription factor components contained within the assembled transcriptome. This interactome was structurally studied and filtered to identify a short list of 92 candidate proteins, which may regulate hESC stemness. **Conclusion.** We hypothesise that this list of proteins, either connecting extracellular components with transcriptional networks, or with hub or bottleneck properties, may contain proteins likely to be involved in determining stemness.
Authors: Virginie Mournetas¹,², Quentin M. Nunes²,³, Patricia A. Murray¹, Christopher M. Sanderson¹, David G. Fernig²

Affiliations: ¹ Department of Cellular and Molecular Physiology, Institute of Translational Medicine, ² Department of Biochemistry, Institute of Integrative Biology, ³ NIHR Liverpool Pancreas Biomedical Research Unit, Institute of Translational Medicine, University of Liverpool, L697ZB Liverpool, United Kingdom
Introduction

Human embryonic stem cells (hESCs) are pluripotent cells present in the inner cell mass of the blastocyst (Pera et al. 2000). They give rise in vivo to the three germ layers (ectoderm, endoderm and mesoderm), and, therefore, have the ability to generate all tissues within the body. These cells can also be derived in vitro (Thomson et al. 1998), maintaining an ability to either self-renew or differentiate (Keller 2005). Human ESCs are a fundamental tool for understanding human embryonic development and constituent mechanisms of differentiation (Keller 2005).

Moreover, they represent a potentially powerful tool in drug screening (Jensen et al. 2009) and regenerative medicine (Aznar & Gomez 2012; Keller 2005; Wobus & Boheler 2005). However, in order to mobilise the potential of hESCs, it is necessary to understand the molecular determinants of self-renewal and differentiation.

The core transcriptional network regulating pluripotency (Babaie et al. 2007; Boyer et al. 2005; Chavez et al. 2009; Marson et al. 2008; Rodda et al. 2005), is composed of three transcription factors: octamer-binding protein 4 (OCT4) (Hay et al. 2004), sex determining region Y-box 2 (SOX2) (Fong et al. 2008) and NANOG (Hyslop et al. 2005; Zaehres et al. 2005). Interestingly, although these transcription factors clearly drive pluripotency (Li et al. 2009; Takahashi et al. 2007), their expression is not restricted to hESCs (Atlasi et al. 2008; Leis et al. 2012; Liedtke et al. 2007; Pierantozzi et al. 2011; Zangrossi et al. 2007). Thus, stemness must in part depend on other hESC specific characteristics, such as the context of expression of these three transcription factors. Protein-protein interaction networks may provide a valuable insight into this hESC specific context (Boyer et al. 2005; Muller et al. 2008). Proteins of the cell microenvironment may also be an important part of this network (Evseenko et al. 2009; Stelling et al. 2013; Sun et al. 2012), since this is the niche where cell-cell and cell-extracellular matrix (ECM) interactions...
occur, allowing selective cell communication. Indeed, it was through the addition of ECM proteins and growth factors that xeno-free culture conditions for hESCs were defined (Melkoumian et al. 2010; Rodin et al. 2010). These methods have facilitated investigation of the roles that extracellular molecules, such as heparan sulfate (HS) (Stelling et al. 2013), fibroblast growth factor (FGF)-2 (Eiselleova et al. 2009; Greber et al. 2010) and activin A (Xiao et al. 2006) play in hESC self-renewal and differentiation. However, such factors have not always been linked to specific transcriptional networks and many of the defined medium formulations do not completely sustain pluripotency (Baxter et al. 2009; Ludwig et al. 2006). Therefore, other factors involved in the maintenance of stemness must be missing. One key factor could be a wider link between ECM interactions and transcriptional networks, thereby establishing important relay mechanisms between endogenous and exogenous stemness regulators.

Data from large-scale transcriptomic and proteomic studies (Koh et al. 2012) facilitate the construction of large biological networks in which nodes and edges represent molecules and interactions respectively. Studying the topological properties of these networks may enable the elaboration of novel hypotheses. For instance, it has been shown that hubs, which are highly connected nodes within a network, are more likely to be important proteins in a protein-protein interaction network (Jeong et al. 2001), as well as bottlenecks, which are nodes with a high betweenness centrality, meaning many shortest paths within the network pass through them (Yu et al. 2007).

To gain a more global insight into the potential contribution of the cell-microenvironment to stemness, we employed an in silico systems-level approach where a meta-analysis of dozens of microarrays was performed to establish a stringent yet more representative hESC transcriptome. Transcripts of transcriptional and extracellular proteins were used to build a putative interactome.
or protein-protein interaction network. The organisation of this network was then analysed to identify extracellular proteins with hub or bottleneck properties, which may be involved in determining stemness, as well as proteins connecting the extracellular factors to transcription.
Materials and methods

Establishing hESC and hESC-derived transcriptomes

The microarray datasets used to establish a high coverage hESC transcriptome were raw data (.CEL image files) of single channel Human Genome U133 Plus 2.0 Affymetrix microarrays downloaded from the ArrayExpress public database (Parkinson et al. 2007). Probe intensity extraction and normalisation procedures were performed with BRB-ArrayTools 4.3.0 beta 1 (Simon et al. 2007) using default median array values (selected by BRB-ArrayTools 4.3.0 beta 1) as reference. The minimum required fold change was 1.5. If less than 20% of the expression values met this value, the gene was excluded. Each individual dataset was first analysed using the three available algorithms: Robust Multi-array Analysis (RMA) (Irizarry et al. 2003), GC-RMA (Wu et al. 2004) and Micro Array Suite 5.0 (MAS5.0) (Hubbell et al. 2002). The three lists of expressed genes were either combined to create a total list containing all expressed genes, or compared to create an intersection list containing only overlapping genes. For the hESC datasets, when the intersection list contained at least 50% of the genes of the total list, the dataset was used to perform a meta-analysis to establish the hESC transcriptome. Thus, all hESC datasets matching this criterion were grouped to be analysed together and generate the final intersection list used as the hESC transcriptome for further analysis (Fig. 1). For the hESC-derived cell datasets, if the intersection list contained at least 50% of the genes of the total list, the full transcriptome (fibroblasts and endothelial cells) was used for transcriptomic comparisons; otherwise the datasets were combined to build the final intersection list and form the hESC-derived cell transcriptome, which was used for transcriptomic comparisons (Fig. 1). The identifiers were EntrezGene IDs and Official Gene Symbol identifiers. The identifier conversion
was done with the database for annotation, visualization and integrated discovery (DAVID) 6.7


Selection of extracellular and transcription related sub-transcriptomes

The extracellular (EC) and the transcription factor related (TF) components of the transcriptomes were extracted using the Gene Ontology (GO) database (Ashburner et al. 2000). The terms used were: GO:0005576 (extracellular region) and GO:0009986 (cell surface) for the EC component; GO:0005667 (transcription factor complex), GO:0008134 (transcription factor binding), GO:0000988 (protein binding transcription factor activity) and GO:0001071 (nucleic acid binding transcription factor activity) for the TF component. Genes involved in biological processes (e.g. cell cycle (GO:0007049), cell adhesion (GO:0007155), cell communication (GO:0007154), cell junction (GO:0030054) and cytoskeleton organization (GO:0007010)) were also highlighted.

By using a published list of HS binding proteins (Ori et al. 2011), the EC component was divided into two distinct groups: genes coding for HS binding proteins and those coding for non-HS binding proteins.

The hESC transcriptome was compared with the three different hESC-derived cell transcriptomes to establish which mRNAs were only expressed in hESC (the specific part) and which ones were expressed in all analysed transcriptomes (the common part).

Construction and analysis of putative interactomes

Putative interactomes were built with the Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) 9.0 database (Szklarczyk et al. 2011) using interaction data from experimental/biochemical experiments and association in curated databases only, which excludes...
interaction predictions by neighbourhood in the genome, gene fusions, co-occurrence across
genomes, co-expression and text-mining (co-mentioned in PubMed abstracts). A stringent
interaction confidence of 0.7 was imposed, to ensure a higher probability that the predicted links
exist (Szklarczyk et al. 2011).

**Analysis of network structure**

Cytoscape 2.8.0 software (Shannon et al. 2003) and associated plug-ins were used to visualise
and analyse protein-protein interaction networks. Randomised networks were created by the
RandomNetworks v1.0 plug-in from the real protein-protein interaction networks. Therefore,
each random network had the same number of nodes N and edges L as its corresponding real
network. Network topological parameters, such as connected components, average degree <k>,
degree distribution P(k), average clustering coefficient <C>, clustering coefficient distribution
C(k) and characteristic path length <l>, were computed with the NetworkAnalyser plug-in.

**Enrichments analysis of interactome components**

Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa & Goto 2000) pathway and GO
Biological Processes term enrichments were processed using DAVID 6.7 (Huang da et al. 2009a;
Huang da et al. 2009b) for the analysis of transcriptome subsets. Terms were recorded when the
EASE score was ≤ 0.1 and considered significantly enriched when the false discovery rate was ≤
0.05. Enrichment was calculated through two different ways: the ratio of the ratio of proteins
belonging to the term in the analysed list and the ratio of proteins belonging to the term in Homo
sapiens, or hESCs.
Selection of candidate proteins

Proteins with a degree $k$ in the top 20% were considered as hubs, while proteins with a betweenness in the top 20% were considered as bottlenecks (Yu et al. 2007). The EC/TF and specific/common interfaces were established from the hESC sub-interactome, constructed with STRING data (edge confidence of 0.7) and containing EC and TF components only. To be part of the EC/TF interface network, an EC node had to be connected to a least one TF node and vice-versa. Similarly, to be part of the specific/common interface, a specific node had to be connected to a least one common node and vice-versa. In this complete (ALL_EC+TF) list of candidate proteins composed of the two interfaces, hubs and bottlenecks, only the EC nodes from the specific and common parts were kept to establish the final (C+S_EC) short list of candidate proteins (Fig. 2). The KEGG pathway and GO Biological Processes term enrichments were processed as previously. Statistical analysis was performed using IBM SPSS Statistics 21 software and presented as mean ± SEM.
Results

The hESC transcriptome

To discover new regulators of hESC pluripotency, 24 hESC microarrays were analysed from four different datasets (Table 1). A total of 8,934 genes were found to be expressed, which constitute the high coverage hESC transcriptome (Table S1). To establish hESC specific expression profiles, three different early hESC-derived cell transcriptomes were extracted from analogous fibroblast (5,086 mRNAs), endothelial cells (5,522 mRNAs) or mixed hESC-derived cells (10,730 mRNAs, Table 1 and Table S1).

The mRNAs specifically expressed by the hESCs (1,010 mRNAs) and those common to hESCs and hESC-derived cells (1,933 mRNAs) were identified by comparing the hESC transcriptome with the hESC-derived transcriptomes (Fig. 3A). Gene Ontology (GO) annotation database (Ashburner et al. 2000) was then used to identify the hESC transcription factor (TF) related (721 mRNAs) and extracellular (EC) transcripts. In this last set of mRNAs, a distinction between transcripts coding for HS binding proteins (191 mRNAs) and non-binding proteins (576 mRNAs, Fig. 3B and Table S1) was enabled by a published list of HS binding proteins (Ori et al. 2011).

Transcriptome analysis showed that genes known to be involved in stemness were represented in this hESC transcriptome, such as POU class 5 homeobox 1 (POU5F1, which encodes OCT4 protein) (Nichols et al. 1998) and SOX2 (Avilion et al. 2003). As expected, some of these were in the hESC specific sub-set, such as the telomerase reverse transcriptase (TERT) (Yang et al. 2008) and growth differentiation factor 3 (GDF3) (Levine & Brivanlou 2006) (Table 2A). Interestingly, NANOG (Chambers et al. 2003) was not present here. Some germ layer markers were also found
in the hESC transcriptome, but they were never specific (Table 2B). Lastly, many common
additions to cell culture medium, which have been observed to facilitate hESC growth in vitro,
such as FGF2 (Eiselleova et al. 2009; Vallier et al. 2005) and activin A (Xiao et al. 2006) were
also present (Table 2C).

Putative extracellular/transcriptional interactomes

As the aim of this study was to learn more about the potential importance of functional links
between cell/cell-matrix interactions and transcription, putative protein-protein interaction
networks containing only transcriptional and extracellular components (EC+TF) were established
by means of the STRING database (Szkłarczyk et al. 2011) using transcriptional expression data
as a proxy for protein expression profiles. Two interactomes were built: one (called ALL)
containing all identified EC+TF proteins, composed of 702 nodes and 3,201 edges (Data S1A),
and one (called C+S) containing only those transcripts/proteins that were either specific to hESCs
or common to hESCs and hESC-derived cells, comprising 209 nodes and 371 edges (Data S1B).

The average clustering coefficient \( <C> \) (indicating the network cohesiveness) was closer to zero
for all randomised networks compared to both ALL and C+S interactomes, implying a
significantly higher occurrence of clusters in these selected networks (Fig. 4A).

As observed in previous protein-protein interaction network studies (Albert et al. 2000; Jeong et
al. 2001), both selected networks (ALL and C+S) and randomised networks exhibit a scale-free
structure, where the degree distribution \( P(k) \) follows a power-law \( P(k) \propto k^{-\gamma} \), involving the
presence of hubs (Fig. 4B and Table S2), and the clustering coefficient distribution \( C(k) \) is
independent of \( k \) meaning there is no inherent presence of modules unlike hierarchical networks,
even if there was a tendency to be hierarchical ($C|k|^{-\beta}$) compared to the randomised versions (Fig. 4C).

These results demonstrate that the EC+TF putative protein-protein interaction networks were suitable for further analysis.

### Enrichment analysis

GO Biological Processes term and KEGG pathway enrichments were used to determine if the EC+TF putative interactomes contained significantly enriched sub-sets of proteins. As expected, terms related to EC (*extracellular matrix organization*), and TF (*transcription, DNA templated*) appeared. More interestingly, terms relating to development (*embryonic development*) and pathways already known to be involved in hESC stemness maintenance (transforming growth factor (TGF)-\(\beta\) (*James et al. 2005*) or wingless-type MMTV integration site family (Wnt) (*Sato et al. 2004*)) and differentiation (bone morphogenic protein (BMP) signalling (*Xu et al. 2005*)) were also identified. KEGG *Pathways in cancer* as well as GO terms of cell differentiation, cell adhesion, cell communication and cell proliferation were represented too (Fig. 5 and Table S3A).

Fewer terms were found to be significantly enriched when only the common and specific parts (from ALL to C+S) were analysed. However, when they were found significant, the vast majority was more enriched, except the terms related to TF (Table S3A). *Nuclear-transcribed mRNA catabolic process* (representing 56% of ALL and 48% of C+S) and *multicellular organismal development* (representing 47% of ALL and 54% of C+S) were the most represented non-related terms (Table S3A).
Interestingly, *regulation of cellular component movement* was well enriched with fold changes in ALL of 5.8 (*Homo sapiens* as background)/4.6 (hESC as background) and in C+S of 9.4 (*Homo sapiens* as background)/7.4 (hESC as background). 28%/51% of the proteins belonging to this term in *Homo sapiens*/hESC were represented in ALL and 17%/31% in C+S (Table S3A).

These data show that the EC+TF putative interactomes, both ALL and C+S, still contained the sub-sets of proteins involved in development, cell differentiation, cell adhesion and cell communication.

**Novel proteins potentially associated with stemness**

The final list of potential stemness proteins was established from ALL, the EC+TF putative protein-protein interaction network. This list (called ALL_EC+TF) was composed of nodes with hub or bottleneck features, as well as nodes within the specific/common and EC/TF interfaces (Fig. 2). Hubs are thought to be functionally important due to their high number of interactions, while bottlenecks form links between different processes. 58% of the bottlenecks in the ALL network were also hubs. The specific/common interface reflects the links between the more general cell functions and those specific to hESCs. The EC/TF interface represents points of communication between the genome and the cell's environment, including other cells. The ALL_EC+TF contained 387 candidates (49% EC and 55% TF) with 29 specific (8%) and 126 common (33%) nodes. The key transcription factors OCT4 and SOX2 were present as hubs and part of the EC/TF interface (Table S4A).

Considering GO set (TF, EC) enrichment with regards to proteins belonging to common or specific parts of the transcriptome, the specific sub-set was enriched in non-HS binding EC proteins (1.7-fold change), whereas the common sub-set was enriched in HS binding proteins.
In addition the common sub-set was found to be enriched in both hubs (1.3-fold change) and bottlenecks (1.6-fold change) (Fig. 6B). Finally, hubs were enriched in TF (1.4-fold change) and bottlenecks in HS binding proteins (1.3-fold change, Fig. 6B).

To assess the validity of the candidate prediction, a random ALL_EC+TF list was established the same way using a randomised version of the EC+TF putative interactome (Table S4B). The hub sub-set was identical in both real and random versions of the candidate list due to the way the randomised network was generated. However, the bottleneck sub-set in the real list had proteins with significantly higher betweenness centrality ($7,456 \pm 835$, paired sample test, $n=134$, $p$-value<0.001) than the one in the random list ($0.0108 \pm 0.0005$). Moreover, the random list with 581 proteins retained 83% of the original EC+TF putative interactome against 55% for the real list. The comparison between the real list of candidates and its random version showed that the filtering process was meaningful.

Three shortened lists were generated from ALL_EC+TF list to decrease the number of candidates by either keeping only EC proteins (ALL_EC, 188 proteins, Table S4C) or/and C+S proteins (C+S_EC+TF, 155 proteins, Table S4D and C+S_EC, 92 proteins, Table S4E) as described in Fig. 2. 59% of the common proteins in the longest ALL_EC+TF list and 62% of the specific ones were conserved in the shortest C+S_EC list. Similarly, 9% of the hubs and 20% of the bottlenecks were kept.

To determine if each list and each sub-set (hubs, bottlenecks and interfaces, as well as specific, common and other proteins from the complete (ALL_EC+TF) to the shortest (C+S_EC) list) still contained proteins potentially involved in stemness maintenance, we undertook further GO Biological Processes term and KEGG pathway enrichments (Table S3B-I). Only the sub-set
containing the hESC-specific proteins was found without any significant enrichment regarding
the analysed terms and pathways (Table S3C). However, the most represented term in both
specific sub-sets from the ALL_EC+TF and ALL_EC, as well as in the four full lists and in all
other sub-sets, was *multicellular organismal development* (Table S3B-I).

Again, terms and pathways related to TF appeared in ALL_EC+TF and C+S_EC+TF lists, as
well as in all the other sub-sets of these two lists (Table S3B,D-I). These TF terms and pathways
were logically lost in the ALL_EC and C+S_EC lists and sub-sets.

GO terms related to cell differentiation, cell adhesion, cell communication, cell movement or cell
proliferation, and KEGG *pathways of cancer* were still significantly enriched in the four lists and
in the vast majority of the analysed sub-sets (Table S3B,D-I).

These data demonstrate that the four lists of candidates, as well as each sub-set of proteins (hubs,
bottlenecks, specific/common and EC/TF interfaces) incorporated proteins involved in
development and cell communication. Focusing on the EC proteins that were either specific to
hESCs or common to hESCs and hESC-derived cells allowed us to reduce the number of
candidates to 92 proteins (Table 3 and Table S4E), while insure that proteins potentially
involved in stemness maintenance were retained. Among these proteins, some are already known
to be required for maintenance of hESC stemness, either directly, such as NODAL (*James et al.*
et al.* 2006), or indirectly through signalling pathways such as TGF-β (*James et al.* 2005) or Wnt
(*Sato et al.* 2004). Other proteins are also known to play a role in mouse ESC pluripotency, but
not yet in hESC, such as the transcription factor 3 (TCF3) (*Cole et al.* 2008). However, for the
majority of candidates, including titin, nothing is known yet about their functions in the context of hESCs.
Discussion

We provide a novel picture of the hESC transcriptome built from a meta-analysis and allowing the in silico analysis of a putative hESC protein-protein interaction network. This systems-level approach has been used to identify proteins potentially involved in the maintenance of stemness. Transcriptomic data provide the most comprehensive insight into variations in cell type or condition specific gene expression profiles. Therefore, data from multiple microarray studies were chosen to generate putative interactomes due to the lack of corresponding comprehensive proteomic profiles. Even if mRNA and protein levels have been suggested to correlate weakly, this correlation may be stronger than anticipated, though this depends on the techniques used to measure mRNA (Jingyi et al. 2014; Pascal et al. 2008; Schwanhausser et al. 2011; Schwanhäusser et al. 2013). Thus, the present study provides a predictive qualitative insight into sub-networks of proteins, which may mediate or maintain human stem cell pluripotency.

The decision to selectively include genes only found by three different algorithms allowed a reduction in the number of false positives in the whole transcriptome, but probably amplified the number of false negatives, which may explain the absence of NANOG. Regarding the specific/common distinction, this pipeline permitted confidence about the common mRNA sub-set, whereas it likely increased the false positive rate in the specific mRNA sub-set, which is still half the common one. However, the use of transcriptomic data from different hESC lines cultured under different conditions highlighted the core transcriptome of these cells.

Not all mRNAs were represented in the putative protein-protein interaction network, probably because coverage of human protein-protein interactions in all databases, including STRING, remains incomplete (De Las Rivas & Fontanillo 2010). High edge stringency limits imposed in
this study should minimise inclusion of false positive interactions (De Las Rivas & Fontanillo 2010), thereby increasing confidence in the relevance and utility of predicted networks.

The scale free nature of the EC+TF putative interactomes, mean that they should exhibit a high error tolerance thanks to redundancy and a high attack vulnerability, due to the presence of hubs (Albert et al. 2000).

Even incomplete interactomes are very complex structures. In order to focus on the likely most important proteins within this interactome, four selection criteria were applied, the first being the selection of hESC hubs. These proteins constitute a small, but often essential part of the interactome (Awan et al. 2007). For example, deletion of just one hub in yeast is often lethal (Jeong et al. 2001). The second was the selection of bottlenecks, which link processes and so permit cross-talk. The third and the fourth criteria involved were that proteins had to be in the specific/common or EC/TF interfaces. These interfaces are posited to be important, as they reflect communication links between the nucleus and the extracellular matrix, and between the specific and common proteins, which ultimately make hESCs different from other cell types.

Interestingly, the GO term related to cell motility regulation was strongly represented in the candidate lists. Cell movement is a key component of morphogenesis. It is usually accomplished by three steps (protrusion, adhesion and de-adhesion) where cytoskeleton and ECM are involved (Ananthakrishnan & Ehrlicher 2007). This may be significant as recent data indicates that cell motion may be an intrinsic feature of hESCs (Li et al. 2010).

Regulation of cell proliferation also appeared in our analysis of candidate lists. This may be significant, as cell proliferation is a key property of hESCs, since these cells are able to proliferate almost indefinitely in vitro (Miura et al. 2004). This capability is sustained by the EC part with growth factors (Activin A (Baxter et al. 2009) and FGF2 (Xu et al. 2005)) and ECM
molecules (fibronectin (Baxter et al. 2009) or laminin (Rodin et al. 2010)), as well as by the TF part through the Smad signalling pathway (James et al. 2005; Vallier et al. 2005). Cell proliferation can also be linked to the significant enrichment of cancer pathways in hESCs. Several links arise between cancer and hESCs, for example, the formation of teratomas as a test to assess pluripotency.

**Conclusion**

Mechanisms involved in stemness are complex, multi-level and determined by the intrinsic cell potential, cell/cell and cell/matrix interactions. The meta-analysis of transcriptomic data in this study has allowed the construction of a hESC putative protein-protein interaction network from which novel ECM proteins have been identified as potential stemness regulators. Networks are a snapshot of a dynamic model (Assmus et al. 2006; Peltier & Schaffer 2010). Notions of attractors (or cell stable stationary states), landscapes formed with valleys (attractors) and hills (barriers between attractors), and cell state transitions described by dynamic systems theory will complete this systems biology approach and bring new hypotheses on hESC behaviour (MacArthur et al. 2008; Macarthur et al. 2009; Peltier & Schaffer 2010; Roeder & Radtke 2009).
Acknowledgments

All publications based on BRB-ArrayTools analyses will contain the acknowledgment: "Analyses were performed using BRB-ArrayTools developed by Dr Richard Simon and BRB-ArrayTools Development Team."


Figure 1

Flow chart of the microarray dataset analysis

This flow chart describes the microarray meta-analysis process ending by the transcriptomes establishment of hESC, endothelial cells, fibroblasts and mixed hESC-derived cells.
Figure 2

Establishment of the list of candidates, a flow chart

This flow chart describes the candidate choice process, from the hESC transcriptome to the final list of 92 proteins.
Figure 3

Overlaps of transcriptomes and sub-transcriptomes

A) Main overlaps of hESC and hESC-derived cell transcriptomes. Grey: hESC transcriptome; Blue: endothelial cell transcriptome; Red: fibroblast transcriptome; Green: mixture of hESC-derived cell transcriptome. B) The overlaps of hESC sub-transcriptomes. The hESC transcriptome is composed of 8,934 mRNAs in total with a hESC-specific part (1,010 mRNAs, brown part), a common part (1,933 mRNAs, blue part) shared with the hESC-derived cells, and the rest of the mRNAs (grey). Sub-transcriptomes can be highlighted: the HS binding proteins part (191 mRNAs, specific in red and common in pink); the extracellular part (EC) without HS binding proteins (576 mRNAs, specific in orange and common in purple); the transcription factor related part (TF, 721 mRNAs, specific in yellow and common in light blue).
Figure 4

Figure 4 - General network parameters of EC+TF putative interactomes

A) The average clustering coefficient of real networks and their corresponding average randomised networks with SEM bars (One sample t-test, n=5, p-value<0.001). B) The node degree distribution $P(k)$ and C) the clustering coefficient distribution $C(k)$ (C: common part; S: specific part; R: random; EC: extracellular part; TF: transcription factor related part).
Figure 5

GO/KEGG analyses of EC+TF putative interactomes

A) GO Biological Processes term enrichment (against *Homo sapiens*), in fold change. B) Percentage of the total number of proteins in *Homo sapiens* related to GO Biological Processes that are present in ALL and C+S putative interactomes (C: common part; S: specific part).
Figure 6

Comparative enrichment trends within the candidate protein list

A) Enrichments in specific, common and other parts with HS, EC non-HS and TF. B) Enrichments in HS, EC non-HS and TF parts with EC/TF interface, C/S interface, hubs and bottlenecks (C: common part; S: specific part; EC: extracellular part; HS: heparan sulfate binding proteins; TF: transcription factor related part).
Table 1 (on next page)

Microarray datasets analysis

The access number (second column) gives access to the dataset in ArrayExpress database. Cell types and cell lines (first column), main cell culture conditions (third column), publications linked to the dataset (when available, fourth column) and the number of microarrays used per analysis (fifth column) are specified. Expressed gene lists for each algorithm (RMA, GC-RMA and MAS5.0) as the total and intersection lists are presented. Four hESC datasets (E-GEOD-6561, -15148, -18265 and -26672) have been used to build mix1. Six hESC-derived cell datasets (E-GEOD-9196, -9832, -9940, -14897, -19735 and -21668) have been used to build mix2. Datasets in bold represent the final transcriptomes used for further analysis: mix1 for hESCs, E-GEOD-9832 for the fibroblasts, E-GEOD-19735 for the endothelial cells and mix2 for the mixture of hESC-derived cells (MEFs mouse embryonic fibroblasts; HFFs human foreskin fibroblasts; SR serum replacer).
<table>
<thead>
<tr>
<th>Cell type (Cell line)</th>
<th>Access Number</th>
<th>Linked publication</th>
<th>Main cell culture feature</th>
<th>Number of microarrays</th>
<th>RMA</th>
<th>GC-RMA</th>
<th>MASS. 0</th>
<th>TOTA L</th>
<th>INTERSECTION</th>
</tr>
</thead>
<tbody>
<tr>
<td>hESCs (H14)</td>
<td>E-GEOD-6561</td>
<td>(Baker et al. 2007)</td>
<td>On feeder cells (irradiated MEFs) FGF2 (4 ng/mL) 20% KnockOut SR</td>
<td>4</td>
<td>9672</td>
<td>8680</td>
<td>9822</td>
<td>11930</td>
<td>7088</td>
</tr>
<tr>
<td>hESCs (H1, H7, H9, H13, H14)</td>
<td>E-GEOD-15148</td>
<td>(Yu et al. 2009)</td>
<td>On feeder cells (irradiated MEFs) FGF2 (100 ng/mL) 20% KnockOut SR OR On feeder-free matrigel conditioned medium</td>
<td>10</td>
<td>8798</td>
<td>10554</td>
<td>10293</td>
<td>12467</td>
<td>7395</td>
</tr>
<tr>
<td>hESCs</td>
<td>E-GEOD-18265</td>
<td></td>
<td>On feeder cells (inactivated HFFs) FGF2 (10 ng/mL) 20% KnockOut SR</td>
<td>5</td>
<td>9602</td>
<td>11325</td>
<td>10966</td>
<td>13499</td>
<td>7791</td>
</tr>
<tr>
<td>hESCs (H1)</td>
<td>E-GEOD-26672</td>
<td>(Hu et al. 2011)</td>
<td>On feeder cells (irradiated MEFs) FGF2 (4 ng/mL) 20% KnockOut SR</td>
<td>5</td>
<td>9565</td>
<td>11285</td>
<td>10843</td>
<td>12790</td>
<td>8235</td>
</tr>
<tr>
<td>hESCs</td>
<td>Mix1</td>
<td></td>
<td></td>
<td>24</td>
<td></td>
<td>11001</td>
<td>12173</td>
<td>11546</td>
<td>14043</td>
</tr>
<tr>
<td>Embryoid bodies</td>
<td>E-GEOD-9196</td>
<td>(Lu et al. 2007)</td>
<td></td>
<td>9</td>
<td>5142</td>
<td>6308</td>
<td>6667</td>
<td>8935</td>
<td>3576</td>
</tr>
<tr>
<td>Blast cells</td>
<td>E-GEOD-9832</td>
<td>(Park et al. 2008)</td>
<td></td>
<td>3</td>
<td>6471</td>
<td>7510</td>
<td>8072</td>
<td>9795</td>
<td>5086</td>
</tr>
<tr>
<td>Fibroblasts</td>
<td>E-GEOD-9940</td>
<td></td>
<td></td>
<td>12</td>
<td>6939</td>
<td>8432</td>
<td>7309</td>
<td>11517</td>
<td>3944</td>
</tr>
<tr>
<td>Neural progenitors</td>
<td>E-GEOD-14897</td>
<td></td>
<td></td>
<td>3</td>
<td>1356</td>
<td>2485</td>
<td>5009</td>
<td>5812</td>
<td>942</td>
</tr>
<tr>
<td>Embryoid bodies</td>
<td>E-GEOD-19735</td>
<td></td>
<td></td>
<td>4</td>
<td>7166</td>
<td>9237</td>
<td>8484</td>
<td>11085</td>
<td>5522</td>
</tr>
<tr>
<td>Hepatic cells</td>
<td>E-GEOD-21068</td>
<td>(Schley et al. 2010)</td>
<td></td>
<td>2</td>
<td>704</td>
<td>1832</td>
<td>1156</td>
<td>2079</td>
<td>583</td>
</tr>
<tr>
<td>Endothelial cells</td>
<td>Mix1</td>
<td></td>
<td></td>
<td>8</td>
<td>1617</td>
<td>1417</td>
<td>12134</td>
<td>17458</td>
<td>10730</td>
</tr>
<tr>
<td>Embryoid bodies</td>
<td>(Eyseenko et al. 2010)</td>
<td></td>
<td></td>
<td>3</td>
<td>861</td>
<td>1482</td>
<td>2543</td>
<td>3087</td>
<td>638</td>
</tr>
<tr>
<td>Differentiated cells</td>
<td>Mix2</td>
<td></td>
<td></td>
<td>48</td>
<td></td>
<td>16174</td>
<td>14172</td>
<td>12134</td>
<td>17458</td>
</tr>
</tbody>
</table>
Table 2 (on next page)

Transcriptomes and literature comparisons, a selection of markers

‘Transcriptome’ column: transcriptome(s) or sub-transcriptome containing the mRNAs (Endo: endothelial cell; F: fibroblast; Mix: mixture of hESC-derived cells). GO term column: GO terms found during the GO extraction (CA: cell adhesion; CC: cell cycle; CCo: cell communication; CS: cytoskeleton organisation; J: cell junction; EC: extracellular part; HS: heparan sulfate binding proteins; TF: transcription factor related part). A) Signalling molecules required for pluripotency/self-renewal; B) Germ layer markers and C) Molecules related to culture medium of hESCs.
<table>
<thead>
<tr>
<th>Marker/Family</th>
<th>Acronym</th>
<th>Name</th>
<th>Transcriptome</th>
<th>GO term</th>
</tr>
</thead>
<tbody>
<tr>
<td>Embryonic stem cell</td>
<td>PTEN</td>
<td>phosphatase and tensin homolog</td>
<td>hESC (COMMON)</td>
<td>CS/CA/CC/Cco</td>
</tr>
<tr>
<td></td>
<td>TERT</td>
<td>telomerase reverse transcriptase</td>
<td>hESC (SPECIFIC)</td>
<td>EC non-HS</td>
</tr>
<tr>
<td></td>
<td>GDF3</td>
<td>growth differentiation factor 3</td>
<td>hESC (SPECIFIC)</td>
<td>CCo/EC non-HS</td>
</tr>
<tr>
<td></td>
<td>NODAL</td>
<td>nodal homolog (mouse)</td>
<td>hESC, Mix</td>
<td>TF</td>
</tr>
<tr>
<td></td>
<td>ZIC3</td>
<td>Zic family member 3</td>
<td>hESC, Mix</td>
<td>CCo/TF</td>
</tr>
<tr>
<td></td>
<td>SOX2</td>
<td>SRY (sex determining region Y)-box 2</td>
<td>hESC, Mix</td>
<td></td>
</tr>
<tr>
<td></td>
<td>POU5F1</td>
<td>POU class 5 homeobox 1</td>
<td>hESC, Mix</td>
<td></td>
</tr>
<tr>
<td>Ectoderm</td>
<td>NEFH</td>
<td>neurofilament, heavy polypeptide</td>
<td>hESC (COMMON)</td>
<td>CS</td>
</tr>
<tr>
<td></td>
<td>TUBB3</td>
<td>tubulin, beta 3 class III</td>
<td>hESC, Mix</td>
<td>CCo</td>
</tr>
<tr>
<td>Endoderm</td>
<td>KRT19</td>
<td>keratin 19</td>
<td>hESC (COMMON)</td>
<td>CS</td>
</tr>
<tr>
<td></td>
<td>SOX7</td>
<td>SRY (sex determining region Y)-box 7</td>
<td>hESC, Endo, Mix</td>
<td>CCo/TF</td>
</tr>
<tr>
<td>Mesoderm</td>
<td>KDR</td>
<td>kinase insert domain receptor (a type III receptor tyrosine kinase) (VEGFR) platelet-derived growth factor receptor, alpha polypeptide</td>
<td>hESC, Mix</td>
<td>CA/CCo/HS</td>
</tr>
<tr>
<td></td>
<td>PDGFRA</td>
<td></td>
<td>hESC (COMMON)</td>
<td>CS/CA/Cco</td>
</tr>
<tr>
<td></td>
<td>VIM</td>
<td>Vimentin</td>
<td>hESC (COMMON)</td>
<td>CS</td>
</tr>
<tr>
<td>Fibronectin</td>
<td>FN1</td>
<td>fibronectin 1</td>
<td>hESC (COMMON)</td>
<td>CA/HS</td>
</tr>
<tr>
<td></td>
<td>ITGA5</td>
<td>integrin, alpha 5 (fibronectin receptor, alpha polypeptide)</td>
<td>hESC, Mix</td>
<td>CA/CCo/j/HS</td>
</tr>
<tr>
<td></td>
<td>ITGB1</td>
<td>integrin, beta 1 (fibronectin receptor, beta polypeptide, antigen CD29 includes MDF2, MSK12)</td>
<td>hESC, Endo, Mix</td>
<td>CS/CA/CCo/j/HS</td>
</tr>
<tr>
<td>Fibroblast Growth Factor</td>
<td>FGF2</td>
<td>fibroblast growth factor 2</td>
<td>hESC (COMMON)</td>
<td>CC/CCo/HS/TF</td>
</tr>
<tr>
<td></td>
<td>FGFR1</td>
<td>fibroblast growth factor receptor 1</td>
<td>hESC (COMMON)</td>
<td>CC/CCo/HS</td>
</tr>
<tr>
<td></td>
<td>FGFR2</td>
<td>fibroblast growth factor receptor 2</td>
<td>hESC, Endo, Mix</td>
<td>CC/CCo/HS</td>
</tr>
<tr>
<td></td>
<td>FGFR3</td>
<td>fibroblast growth factor receptor 3</td>
<td>hESC, Endo, Mix</td>
<td>CCo/j/HS</td>
</tr>
<tr>
<td></td>
<td>FGFR4</td>
<td>fibroblast growth factor receptor 4</td>
<td>hESC (SPECIFIC)</td>
<td>CCo/j/HS</td>
</tr>
<tr>
<td>Activin A</td>
<td>ACVR1B</td>
<td>activin A receptor, type IB (ALK4)</td>
<td>hESC (COMMON)</td>
<td>CC/CCo/EC non-HS</td>
</tr>
<tr>
<td></td>
<td>ACVR1C</td>
<td>activin A receptor, type IC</td>
<td>hESC (SPECIFIC)</td>
<td>CCo</td>
</tr>
<tr>
<td></td>
<td>ACVR2A</td>
<td>activin A receptor, type IIA</td>
<td>hESC, F, Mix</td>
<td>CCo</td>
</tr>
<tr>
<td></td>
<td>ACVR2B</td>
<td>activin A receptor, type IIB</td>
<td>hESC, Endo, Mix</td>
<td>CCo/EC non-HS</td>
</tr>
<tr>
<td></td>
<td>INHBA</td>
<td>inhibin, beta A / Activin A</td>
<td>hESC (COMMON)</td>
<td>CC/CCo/HS</td>
</tr>
</tbody>
</table>
Table 3 (on next page)

The list of candidates, an overview

The 'hubs' column gives the degree; the 'bottlenecks' column gives the betweenness; the 'S/C' column indicates if the protein is in the specific/common interface; the 'EC/TF' column indicates if the protein is in the EC/TF interface; the 'KEGG pathway' column indicates the number of pathway involving each protein (GO: Gene Ontology; KEGG: Kyoto Encyclopedia of Genes and Genomes; C: common part; S: specific part; EC: extracellular part; TF: transcription factor related part; J: cell junction). (see Table S4 for a complete list).
<table>
<thead>
<tr>
<th>Acronym</th>
<th>Name</th>
<th>GO term</th>
<th>Hu bs</th>
<th>Bottlenecks</th>
<th>S/C</th>
<th>EC/TF</th>
<th>KEGG pathway</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACTN4</td>
<td>actinin, alpha 4</td>
<td>EC non-HS/CS</td>
<td>17</td>
<td>X</td>
<td>X</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>ACVR1</td>
<td>activin receptor, type IB (ALK4)</td>
<td>EC non-HS/CCo</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ADM</td>
<td>adrenomedullin</td>
<td>EC non-HS/CCo</td>
<td>4050.8</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BMP2</td>
<td>bone morphogenetic protein 2</td>
<td>HS/CC/CCo</td>
<td>2693.9</td>
<td>X</td>
<td>X</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>DMD</td>
<td>dystrophin</td>
<td>EC non-HS/CCo</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FGFR1</td>
<td>Fibroblast growth factor receptor 1</td>
<td>HS/CC/CCo</td>
<td>2634.2</td>
<td>X</td>
<td>X</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>FN1</td>
<td>fibronectin 1</td>
<td>HS/CA</td>
<td>6988.8</td>
<td>X</td>
<td>X</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>IL6</td>
<td>Interleukin 6</td>
<td>HS/CCo</td>
<td></td>
<td>X</td>
<td></td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>INHBA</td>
<td>inhibin, beta A / Activin A</td>
<td>HS/CC/CCo</td>
<td></td>
<td>X</td>
<td></td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>ITGA6</td>
<td>integrin, alpha 6 (CD49f)</td>
<td>EC non-HS/CA/CCo</td>
<td>15218.1</td>
<td>X</td>
<td>X</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>ITGAV</td>
<td>integrin, alpha V (vitronectin receptor)</td>
<td>HS/CA/CCo</td>
<td>12409.4</td>
<td>X</td>
<td></td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>JAM3</td>
<td>junctional adhesion molecule 3</td>
<td>EC non-HS/CA/CCo</td>
<td></td>
<td>X</td>
<td>X</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>LAMA1</td>
<td>laminin, alpha 1</td>
<td>HS/CA/CCo</td>
<td>10501.7</td>
<td>X</td>
<td>X</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>MET</td>
<td>(hepatocyte growth factor receptor</td>
<td>HS/CS/CC/CCo</td>
<td>2552.8</td>
<td>X</td>
<td></td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>PLAT</td>
<td>plasminogen activator, tissue</td>
<td>HS/CCo</td>
<td></td>
<td>X</td>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>PLAU</td>
<td>plasminogen activator, urokinase</td>
<td>HS/CA/CCo</td>
<td>10022.8</td>
<td>X</td>
<td>X</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>SERPIN</td>
<td>serpin peptidase inhibitor, clade E,</td>
<td>HS/CC/CCo</td>
<td>10580.8</td>
<td>X</td>
<td>X</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td></td>
<td>member 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SERPIN</td>
<td>serpin peptidase inhibitor, clade I,</td>
<td>EC non-HS/CA/CCo</td>
<td>2066.0</td>
<td>X</td>
<td>X</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>member 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TGFB2</td>
<td>transforming growth factor, beta 2</td>
<td>HS/CA/CC/CCo</td>
<td>6817.1</td>
<td>X</td>
<td>X</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>THBS1</td>
<td>thrombospondin 1</td>
<td>HS/CA/CC/CCo</td>
<td>16</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VEGFA</td>
<td>vascular endothelial growth factor A</td>
<td>HS/CA/CCo</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CDH8</td>
<td>cadherin 8, type 2</td>
<td>HS/CA</td>
<td>X</td>
<td>X</td>
<td></td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>FGF4</td>
<td>fibroblast growth factor 4</td>
<td>HS/CA/CCo</td>
<td>X</td>
<td>X</td>
<td></td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>FGF4</td>
<td>fibroblast growth factor receptor 4</td>
<td>HS/CCo</td>
<td>X</td>
<td>X</td>
<td></td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>IDE</td>
<td>insulin-degrading enzyme</td>
<td>EC non-HS/CCo</td>
<td>X</td>
<td>X</td>
<td></td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>INHBE</td>
<td>inhibin, beta E</td>
<td>EC non-HS/CCo</td>
<td>X</td>
<td></td>
<td></td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>NODAL</td>
<td>nodal homolog (mouse)</td>
<td>EC non-HS/CCo</td>
<td>X</td>
<td>X</td>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>PLXNB1</td>
<td>plexin B1</td>
<td>EC non-HS/CCo</td>
<td>X</td>
<td></td>
<td></td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>TTN</td>
<td>titin</td>
<td>EC non-HS/CS/CC</td>
<td>X</td>
<td>X</td>
<td></td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>WIF1</td>
<td>WNT inhibitory factor 1</td>
<td>EC non-HS/CCo</td>
<td>X</td>
<td>X</td>
<td></td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>