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.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.415v2 | CC-BY 4.0 Open Access | rec: 18 Sep 2014, publ: 19 Sep 2014

1 Authors: Virginie Mournetas[1,2], Quentin M. Nunes[2,3], Patricia A. Murray[1], Christopher M.

2 Sanderson[1], David G. Fernig[2]

3 Affiliations: [1] Department of Cellular and Molecular Physiology, Institute of Translational

4 Medicine, [2] Department of Biochemistry, Institute of Integrative Biology, [3]NIHR Liverpool

5 Pancreas Biomedical Research Unit, Institute of Translational Medicine, University of Liverpool,

6 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 (Huang da et al. 2009a; Huang da et al. 2009b).

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. Statistical analysis was performed using IBM SPSS Statistics 21 software.

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 trancriptome with the hESC-derived trancriptomes (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 (Szklarczyk 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 $P(k) \quad k^{-\gamma}$, 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,

176 even if there was a tendency to be hierarchical ($C(k) \ k^{-\beta}$) compared to the randomised

177 versions (Fig. 4C).

178 These results demonstrate that the EC+TF putative protein-protein interaction networks were

179 suitable for further analysis.

180 **Enrichment analysis**

181 GO Biological Processes term and KEGG pathway enrichments were used to determine if the

182 EC+TF putative interactomes contained significantly enriched sub-sets of proteins. As expected,

183 terms related to EC (*extracellular matrix organization*), and TF (*transcription, DNA templated*)

184 appeared. More interestingly, terms relating to development (*embryonic development*) and

185 pathways already known to be involved in hESC stemness maintenance (transforming growth

186 factor (TGF)- β (James et al. 2005) or wingless-type MMTV integration site family (Wnt) (Sato

187 et al. 2004)) and differentiation (bone morphogenic protein (BMP) signalling (Xu et al. 2005))

188 were also identified. KEGG *Pathways in cancer* as well as GO terms of cell differentiation, cell

189 adhesion, cell communication and cell proliferation were represented too (Fig. 5 and Table S3A).

190 Fewer terms were found to be significantly enriched when only the common and specific parts

191 (from ALL to C+S) were analysed. However, when they were found significant, the vast majority

192 was more enriched, except the terms related to TF (Table S3A). *Nuclear-transcribed mRNA*

193 *catabolic process* (representing 56% of ALL and 48% of C+S) and *multicellular organismal*

194 *development* (representing 47% of ALL and 54% of C+S) were the most represented non-related

195 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

(1.6-fold change) (Fig. 6A). 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 ± 835, paired sample test, n=134, p-value<0.001) than the one in the random list (0.0108 ± 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 insuring 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. 2005; Vallier et al. 2005), FGF2 (Eiselleova et al. 2009; Vallier et al. 2005) and activin A (Xiao 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."

References

Albert R, Jeong H, and Barabasi AL. 2000. Error and attack tolerance of complex networks. *Nature* 406:378-382.

Ananthakrishnan R, and Ehrlicher A. 2007. The forces behind cell movement. *Int J Biol Sci* 3:303-317.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, and Sherlock G. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 25:25-29.

Assmus HE, Herwig R, Cho KH, and Wolkenhauer O. 2006. Dynamics of biological systems: role of systems biology in medical research. *Expert Rev Mol Diagn* 6:891-902.

Atlasi Y, Mowla SJ, Ziaee SA, Gokhale PJ, and Andrews PW. 2008. OCT4 spliced variants are differentially expressed in human pluripotent and nonpluripotent cells. *Stem Cells* 26:3068-3074.

Avilion AA, Nicolis SK, Pevny LH, Perez L, Vivian N, and Lovell-Badge R. 2003. Multipotent cell lineages in early mouse development depend on SOX2 function. *Genes Dev* 17:126-140.

Awan A, Bari H, Yan F, Moksong S, and Yang… S. 2007. Regulatory network motifs and hotspots of cancer genes in a mammalian cellular signalling network. *Regulatory network motifs and hotspots of cancer genes in a mammalian cellular signalling network*.

Aznar J, and Gomez I. 2012. Possible clinical usefulness of embryonic stem cells. *Rev Clin Esp* 212:403-406.

Babaie Y, Herwig R, Greber B, Brink TC, Wruck W, Groth D, Lehrach H, Burdon T, and Adjaye J. 2007. Analysis of Oct4-dependent transcriptional networks regulating self-renewal and pluripotency in human embryonic stem cells. *Stem Cells* 25:500-510.

Baker DE, Harrison NJ, Maltby E, Smith K, Moore HD, Shaw PJ, Heath PR, Holden H, and Andrews PW. 2007. Adaptation to culture of human embryonic stem cells and oncogenesis in vivo. *Nat Biotechnol* 25:207-215.

Baxter MA, Camarasa MV, Bates N, Small F, Murray P, Edgar D, and Kimber SJ. 2009. Analysis of the distinct functions of growth factors and tissue culture substrates necessary for the long-term self-renewal of human embryonic stem cell lines. *Stem Cell Res*.

Boyer LA, Lee TI, Cole MF, Johnstone SE, Levine SS, Zucker JP, Guenther MG, Kumar RM, Murray HL, Jenner RG, Gifford DK, Melton DA, Jaenisch R, and Young RA. 2005. Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell* 122:947-956.

Chambers I, Colby D, Robertson M, Nichols J, Lee S, Tweedie S, and Smith A. 2003. Functional expression cloning of Nanog, a pluripotency sustaining factor in embryonic stem cells. *Cell* 113:643-655.

Chavez L, Bais AS, Vingron M, Lehrach H, Adjaye J, and Herwig R. 2009. In silico identification of a core regulatory network of OCT4 in human embryonic stem cells using an integrated approach. *BMC Genomics* 10:314.

Cole MF, Johnstone SE, Newman JJ, Kagey MH, and Young RA. 2008. Tcf3 is an integral component of the core regulatory circuitry of embryonic stem cells. *Genes Dev* 22:746-755.

De Las Rivas J, and Fontanillo C. 2010. Protein-protein interactions essentials: key concepts to building and analyzing interactome networks. *PLoS Comput Biol* 6:e1000807.

Eiselleova L, Matulka K, Kriz V, Kunova M, Schmidtova Z, Neradil J, Tichy B, Dvorakova D, Pospisilova S, Hampl A, and Dvorak P. 2009. A complex role for FGF-2 in self-renewal, survival, and adhesion of human embryonic stem cells. *Stem Cells* 27:1847-1857.

Evseenko D, Schenke-Layland K, Dravid G, Zhu Y, Hao QL, Scholes J, Wang XC, Maclellan WR, and Crooks GM. 2009. Identification of the critical extracellular matrix proteins that promote human embryonic stem cell assembly. *Stem Cells Dev* 18:919-928.

Evseenko D, Zhu Y, Schenke-Layland K, Kuo J, Latour B, Ge S, Scholes J, Dravid G, Li X, MacLellan WR, and Crooks GM. 2010. Mapping the first stages of mesoderm commitment during differentiation of human embryonic stem cells. *Proceedings of the National Academy of Sciences of the United States of America* 107:13742-13747.

Fong H, Hohenstein KA, and Donovan PJ. 2008. Regulation of self-renewal and pluripotency by Sox2 in human embryonic stem cells. *Stem Cells* 26:1931-1938.

Greber B, Wu G, Bernemann C, Joo JY, Han DW, Ko K, Tapia N, Sabour D, Sterneckert J, Tesar P, and Scholer HR. 2010. Conserved and divergent roles of FGF signaling in mouse epiblast stem cells and human embryonic stem cells. *Cell Stem Cell* 6:215-226.

Hay DC, Sutherland L, Clark J, and Burdon T. 2004. Oct-4 knockdown induces similar patterns of endoderm and trophoblast differentiation markers in human and mouse embryonic stem cells. *Stem Cells* 22:225-235.

Hu K, Yu J, Suknuntha K, Tian S, Montgomery K, Choi KD, Stewart R, Thomson JA, and Slukvin, II. 2011. Efficient generation of transgene-free induced pluripotent stem cells from normal and neoplastic bone marrow and cord blood mononuclear cells. *Blood* 117:e109-119.

Huang da W, Sherman BT, and Lempicki RA. 2009a. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res* 37:1-13.

Huang da W, Sherman BT, and Lempicki RA. 2009b. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* 4:44-57.

Hubbell E, Liu WM, and Mei R. 2002. Robust estimators for expression analysis. *Bioinformatics* 18:1585-1592.

Hyslop L, Stojkovic M, Armstrong L, Walter T, Stojkovic P, Przyborski S, Herbert M, Murdoch A, Strachan T, and Lako M. 2005. Downregulation of NANOG induces differentiation of human embryonic stem cells to extraembryonic lineages. *Stem Cells* 23:1035-1043.

Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, and Speed TP. 2003. Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res* 31:e15.

James D, Levine AJ, Besser D, and Hemmati-Brivanlou A. 2005. TGFbeta/activin/nodal signaling is necessary for the maintenance of pluripotency in human embryonic stem cells. *Development* 132:1273-1282.

Jensen J, Hyllner J, and Bjorquist P. 2009. Human embryonic stem cell technologies and drug discovery. *J Cell Physiol* 219:513-519.

Jeong H, Mason SP, Barabasi AL, and Oltvai ZN. 2001. Lethality and centrality in protein networks. *Nature* 411:41-42.

Jingyi JL, Peter JB, and Mark DB. 2014. System wide analyses have underestimated protein abundances and the importance of transcription in mammals. *PeerJ* 2.

Kanehisa M, and Goto S. 2000. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 28:27-30.

Keller G. 2005. Embryonic stem cell differentiation: emergence of a new era in biology and medicine. *Genes Dev* 19:1129-1155.

Koh GC, Porras P, Aranda B, Hermjakob H, and Orchard SE. 2012. Analyzing protein-protein interaction networks. *J Proteome Res* 11:2014-2031.

Leis O, Eguiara A, Lopez-Arribillaga E, Alberdi MJ, Hernandez-Garcia S, Elorriaga K, Pandiella A, Rezola R, and Martin AG. 2012. Sox2 expression in breast tumours and activation in breast cancer stem cells. *Oncogene* 31:1354-1365.

Levine AJ, and Brivanlou AH. 2006. GDF3, a BMP inhibitor, regulates cell fate in stem cells and early embryos. *Development* 133:209-216.

Li L, Wang BH, Wang S, Moalim-Nour L, Mohib K, Lohnes D, and Wang L. 2010. Individual cell movement, asymmetric colony expansion, rho-associated kinase, and E-cadherin impact the clonogenicity of human embryonic stem cells. *Biophys J* 98:2442-2451.

Li W, Wei W, Zhu S, Zhu J, Shi Y, Lin T, Hao E, Hayek A, Deng H, and Ding S. 2009. Generation of rat and human induced pluripotent stem cells by combining genetic reprogramming and chemical inhibitors. *Cell Stem Cell* 4:16-19.

Liedtke S, Enczmann J, Waclawczyk S, Wernet P, and Kogler G. 2007. Oct4 and its pseudogenes confuse stem cell research. *Cell Stem Cell* 1:364-366.

Lu SJ, Hipp JA, Feng Q, Hipp JD, Lanza R, and Atala A. 2007. GeneChip analysis of human embryonic stem cell differentiation into hemangioblasts: an in silico dissection of mixed phenotypes. *Genome biology* 8:R240.

Ludwig TE, Levenstein ME, Jones JM, Berggren WT, Mitchen ER, Frane JL, Crandall LJ, Daigh CA, Conard KR, Piekarczyk MS, Llanas RA, and Thomson JA. 2006. Derivation of human embryonic stem cells in defined conditions. *Nat Biotechnol* 24:185-187.

MacArthur BD, Ma'ayan A, and Lemischka IR. 2008. Toward stem cell systems biology: from molecules to networks and landscapes. *Cold Spring Harb Symp Quant Biol* 73:211-215.

Macarthur BD, Ma'ayan A, and Lemischka IR. 2009. Systems biology of stem cell fate and cellular reprogramming. *Nat Rev Mol Cell Biol* 10:672-681.

Marson A, Levine SS, Cole MF, Frampton GM, Brambrink T, Johnstone S, Guenther MG, Johnston WK, Wernig M, Newman J, Calabrese JM, Dennis LM, Volkert TL, Gupta S, Love J, Hannett N, Sharp PA, Bartel DP, Jaenisch R, and Young RA. 2008. Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells. *Cell* 134:521-533.

Melkoumian Z, Weber JL, Weber DM, Fadeev AG, Zhou Y, Dolley-Sonneville P, Yang J, Qiu L, Priest CA, Shogbon C, Martin AW, Nelson J, West P, Beltzer JP, Pal S, and Brandenberger R. 2010. Synthetic peptide-acrylate surfaces for long-term self-renewal and cardiomyocyte differentiation of human embryonic stem cells. *Nat Biotechnol* 28:606-610.

Miura T, Mattson MP, and Rao MS. 2004. Cellular lifespan and senescence signaling in embryonic stem cells. *Aging Cell* 3:333-343.

Muller FJ, Laurent LC, Kostka D, Ulitsky I, Williams R, Lu C, Park IH, Rao MS, Shamir R, Schwartz PH, Schmidt NO, and Loring JF. 2008. Regulatory networks define phenotypic classes of human stem cell lines. *Nature* 455:401-405.

Nichols J, Zevnik B, Anastassiadis K, Niwa H, Klewe-Nebenius D, Chambers I, Scholer H, and Smith A. 1998. Formation of pluripotent stem cells in the mammalian embryo depends on the POU transcription factor Oct4. *Cell* 95:379-391.

Ori A, Wilkinson MC, and Fernig DG. 2011. A systems biology approach for the investigation of the heparin/heparan sulfate interactome. *J Biol Chem*.

Park IH, Zhao R, West JA, Yabuuchi A, Huo H, Ince TA, Lerou PH, Lensch MW, and Daley GQ. 2008. Reprogramming of human somatic cells to pluripotency with defined factors. *Nature* 451:141-146.

Parkinson H, Kapushesky M, Shojatalab M, Abeygunawardena N, Coulson R, Farne A, Holloway E, Kolesnykov N, Lilja P, Lukk M, Mani R, Rayner T, Sharma A, William E, Sarkans U, and Brazma A. 2007. ArrayExpress--a public database of microarray experiments and gene expression profiles. *Nucleic Acids Res* 35:D747-750.

Pascal LE, True LD, Campbell DS, Deutsch EW, Risk M, Coleman IM, Eichner LJ, Nelson PS, and Liu AY. 2008. Correlation of mRNA and protein levels: cell type-specific gene expression of cluster designation antigens in the prostate. *BMC Genomics* 9:246.

Peltier J, and Schaffer DV. 2010. Systems biology approaches to understanding stem cell fate choice. *IET Syst Biol* 4:1-11.

Pera MF, Reubinoff B, and Trounson A. 2000. Human embryonic stem cells. *J Cell Sci* 113 (Pt 1):5-10.

Pierantozzi E, Gava B, Manini I, Roviello F, Marotta G, Chiavarelli M, and Sorrentino V. 2011. Pluripotency regulators in human mesenchymal stem cells: expression of NANOG but not of OCT-4 and SOX-2. *Stem Cells Dev* 20:915-923.

Rodda DJ, Chew JL, Lim LH, Loh YH, Wang B, Ng HH, and Robson P. 2005. Transcriptional regulation of nanog by OCT4 and SOX2. *J Biol Chem* 280:24731-24737.

Rodin S, Domogatskaya A, Strom S, Hansson EM, Chien KR, Inzunza J, Hovatta O, and Tryggvason K. 2010. Long-term self-renewal of human pluripotent stem cells on human recombinant laminin-511. *Nat Biotechnol* 28:611-615.

Roeder I, and Radtke F. 2009. Stem cell biology meets systems biology. *Development* 136:3525-3530.

Sato N, Meijer L, Skaltsounis L, Greengard P, and Brivanlou AH. 2004. Maintenance of pluripotency in human and mouse embryonic stem cells through activation of Wnt signaling by a pharmacological GSK-3-specific inhibitor. *Nat Med* 10:55-63.

Schwanhausser B, Busse D, Li N, Dittmar G, Schuchhardt J, Wolf J, Chen W, and Selbach M. 2011. Global quantification of mammalian gene expression control. *Nature* 473:337-342.

Schwanhäusser B, Busse D, Li N, Dittmar G, Schuchhardt J, Wolf J, Chen W, and Selbach M. 2013. Corrigendum: Global quantification of mammalian gene expression control. *Nature* 495:126-127.

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, and Ideker T. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 13:2498-2504.

Si-Tayeb K, Noto FK, Nagaoka M, Li J, Battle MA, Duris C, North PE, Dalton S, and Duncan SA. 2010. Highly efficient generation of human hepatocyte-like cells from induced pluripotent stem cells. *Hepatology* 51:297-305.

Simon R, Lam A, Li MC, Ngan M, Menenzes S, and Zhao Y. 2007. Analysis of gene expression data using BRB-ArrayTools. *Cancer Inform* 3:11-17.

Stelling MP, Lages YM, Tovar AM, Mourao PA, and Rehen SK. 2013. Matrix-bound heparan sulfate is essential for the growth and pluripotency of human embryonic stem cells. *Glycobiology* 23:337-345.

Sun Y, Villa-Diaz LG, Lam RH, Chen W, Krebsbach PH, and Fu J. 2012. Mechanics regulates fate decisions of human embryonic stem cells. *PLoS One* 7:e37178.

Szklarczyk D, Franceschini A, Kuhn M, Simonovic M, Roth A, Minguez P, Doerks T, Stark M, Muller J, Bork P, Jensen LJ, and von Mering C. 2011. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res* 39:D561-568.

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 JA, Itskovitz-Eldor J, Shapiro SS, Waknitz MA, Swiergiel JJ, Marshall VS, and Jones JM. 1998. Embryonic stem cell lines derived from human blastocysts. *Science* 282:1145-1147.

Vallier L, Alexander M, and Pedersen RA. 2005. Activin/Nodal and FGF pathways cooperate to maintain pluripotency of human embryonic stem cells. *J Cell Sci* 118:4495-4509.

Wobus AM, and Boheler KR. 2005. Embryonic stem cells: prospects for developmental biology and cell therapy. *Physiol Rev* 85:635-678.

Wu Z, Irizarry R, Gentleman R, Murillo F, and Spencer F. 2004. A Model Based Background Adjustment for Oligonucleotide Expression Arrays. *bepress*.

Xiao L, Yuan X, and Sharkis SJ. 2006. Activin A maintains self-renewal and regulates fibroblast growth factor, Wnt, and bone morphogenic protein pathways in human embryonic stem cells. *Stem Cells* 24:1476-1486.

Xu RH, Peck RM, Li DS, Feng X, Ludwig T, and Thomson JA. 2005. Basic FGF and suppression of BMP signaling sustain undifferentiated proliferation of human ES cells. *Nat Methods* 2:185-190.

Yang C, Przyborski S, Cooke MJ, Zhang X, Stewart R, Anyfantis G, Atkinson SP, Saretzki G, Armstrong L, and Lako M. 2008. A key role for telomerase reverse transcriptase unit in modulating human embryonic stem cell proliferation, cell cycle dynamics, and in vitro differentiation. *Stem Cells* 26:850-863.

Yu H, Kim PM, Sprecher E, Trifonov V, and Gerstein M. 2007. The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. *PLoS Comput Biol* 3:e59.

Yu J, Hu K, Smuga-Otto K, Tian S, Stewart R, Slukvin, II, and Thomson JA. 2009. Human induced pluripotent stem cells free of vector and transgene sequences. *Science* 324:797-801.

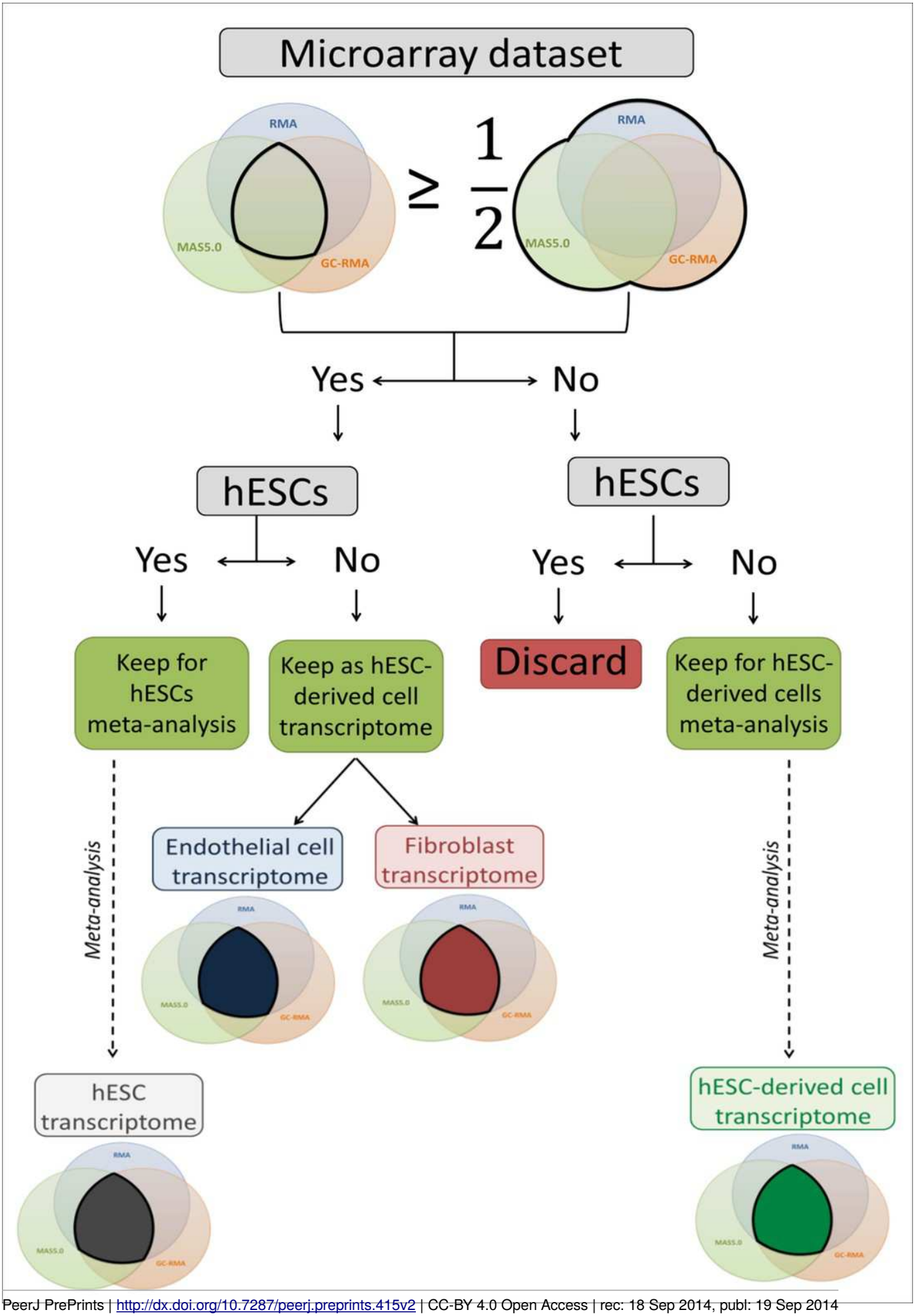
Zaehres H, Lensch MW, Daheron L, Stewart SA, Itskovitz-Eldor J, and Daley GQ. 2005. High-efficiency RNA interference in human embryonic stem cells. *Stem Cells* 23:299-305.

Zangrossi S, Marabese M, Broggini M, Giordano R, D'Erasmo M, Montelatici E, Intini D, Neri A, Pesce M, Rebulla P, and Lazzari L. 2007. Oct-4 expression in adult human differentiated cells challenges its role as a pure stem cell marker. *Stem Cells* 25:1675-1680.

Figure 1

Flow chart of the microarray dataset analysis

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



PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.415v2 | CC-BY 4.0 Open Access | rec: 18 Sep 2014, publ: 19 Sep 2014

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.

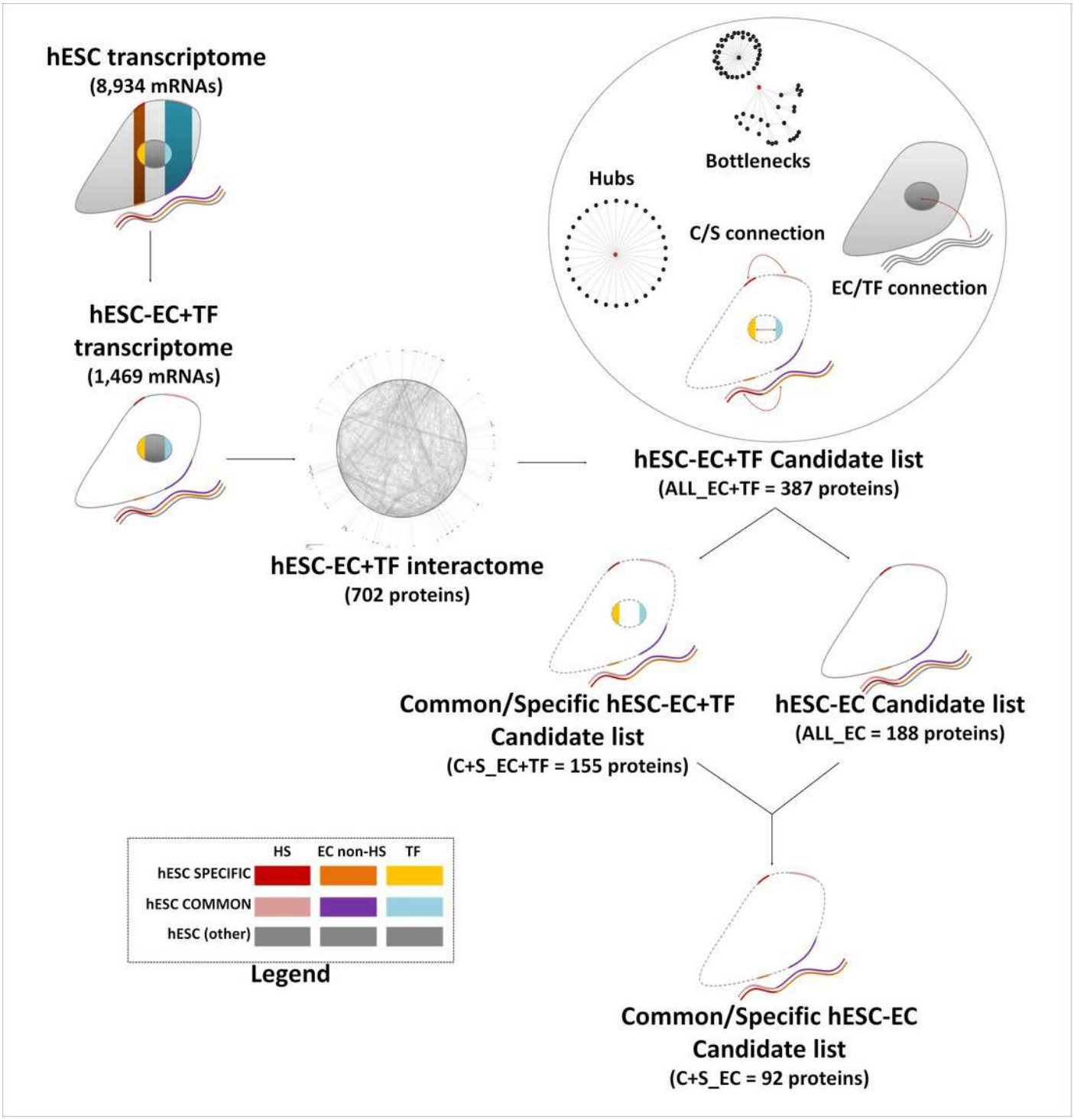


Figure 3

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).

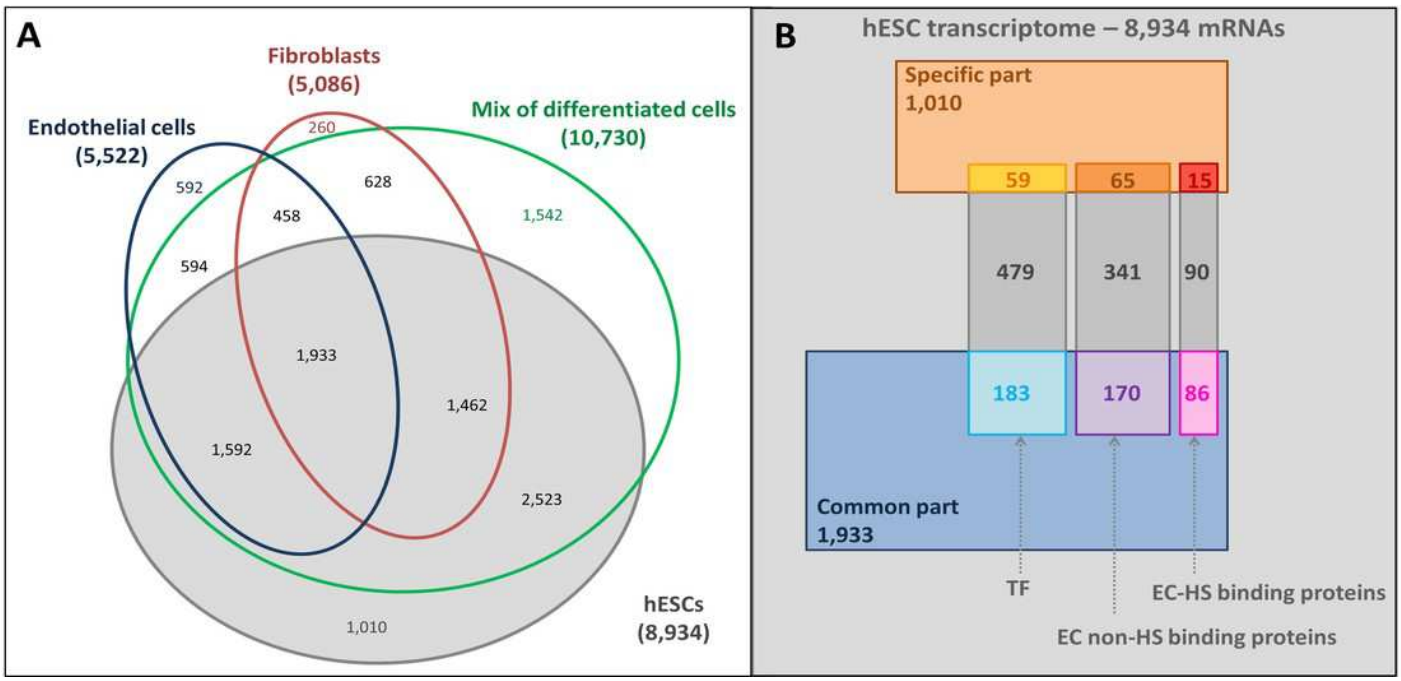


Figure 4

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).

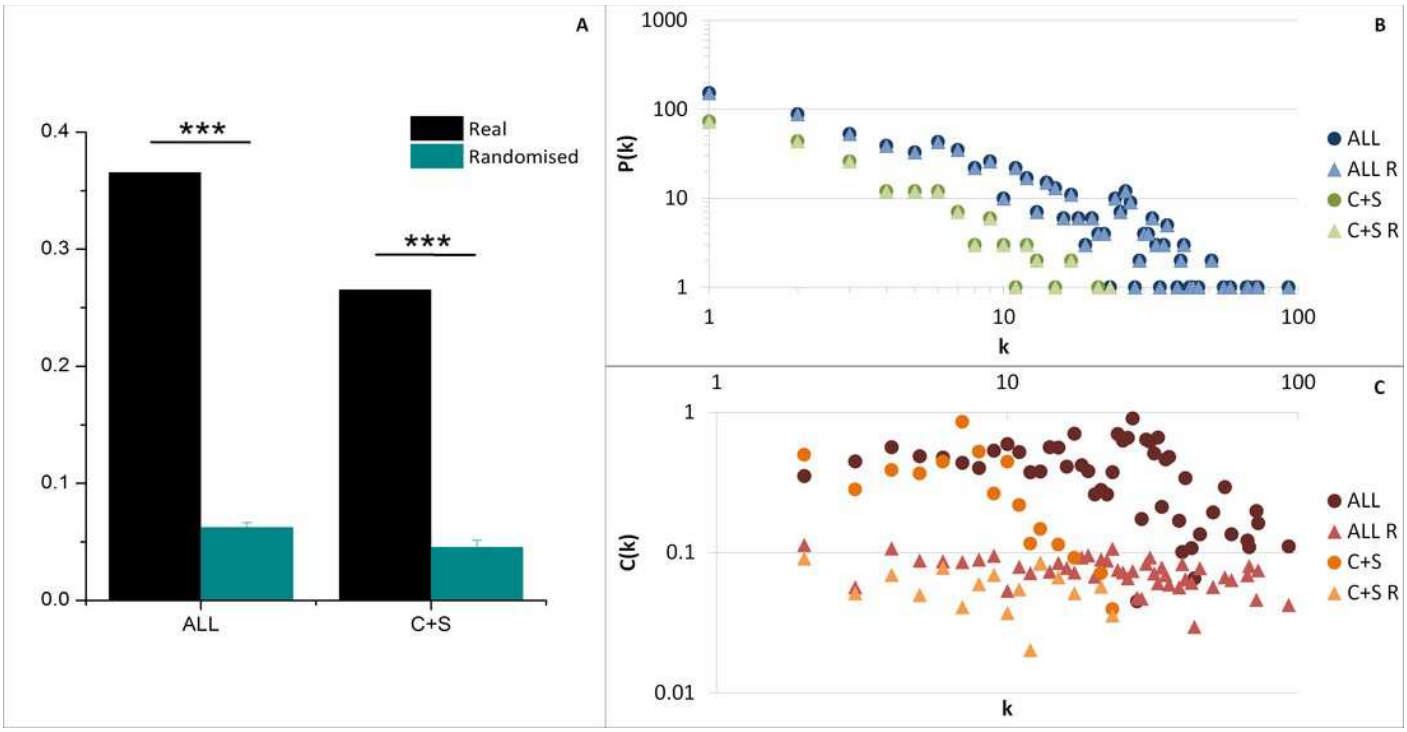
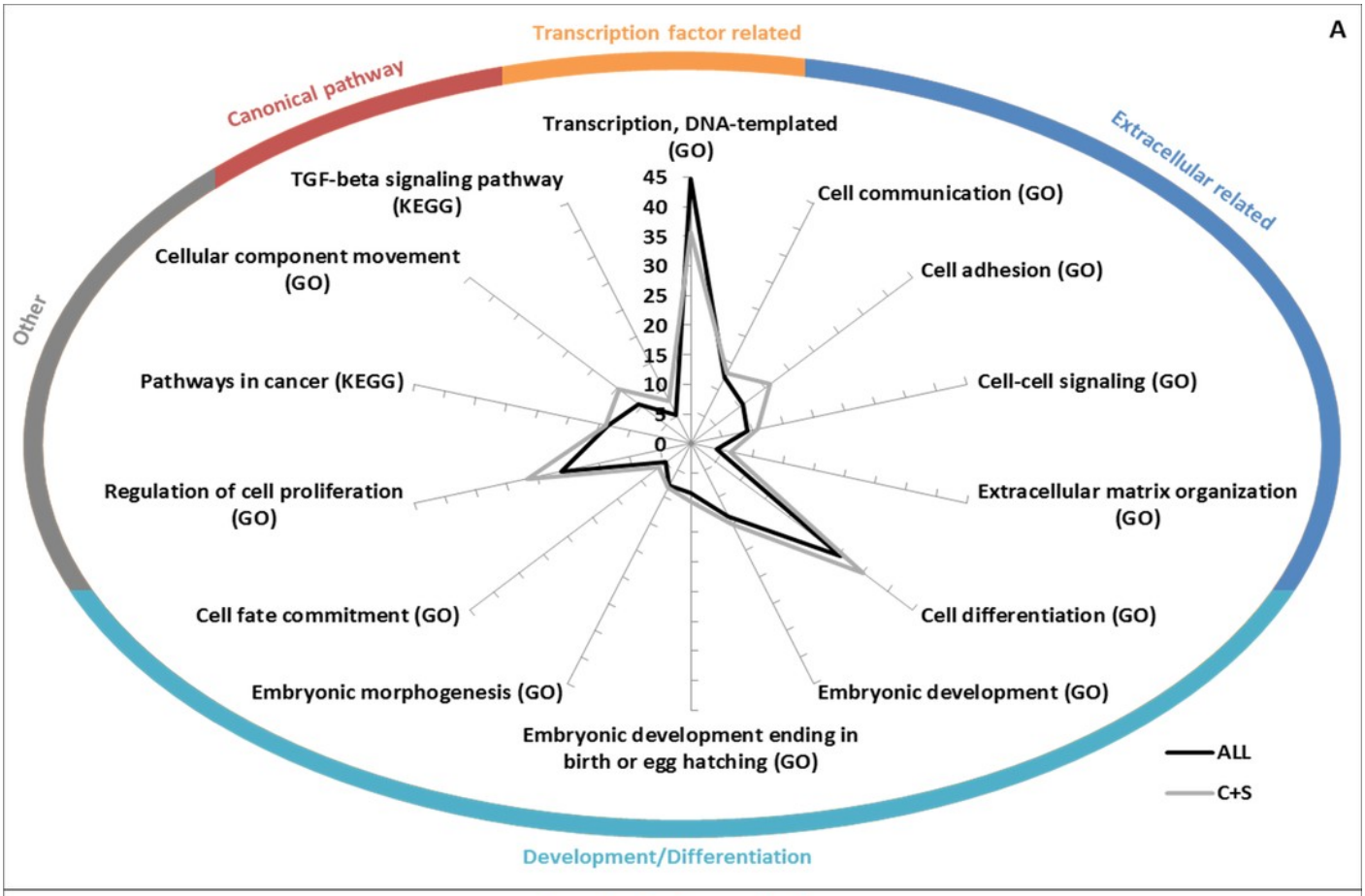


Figure 5

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).



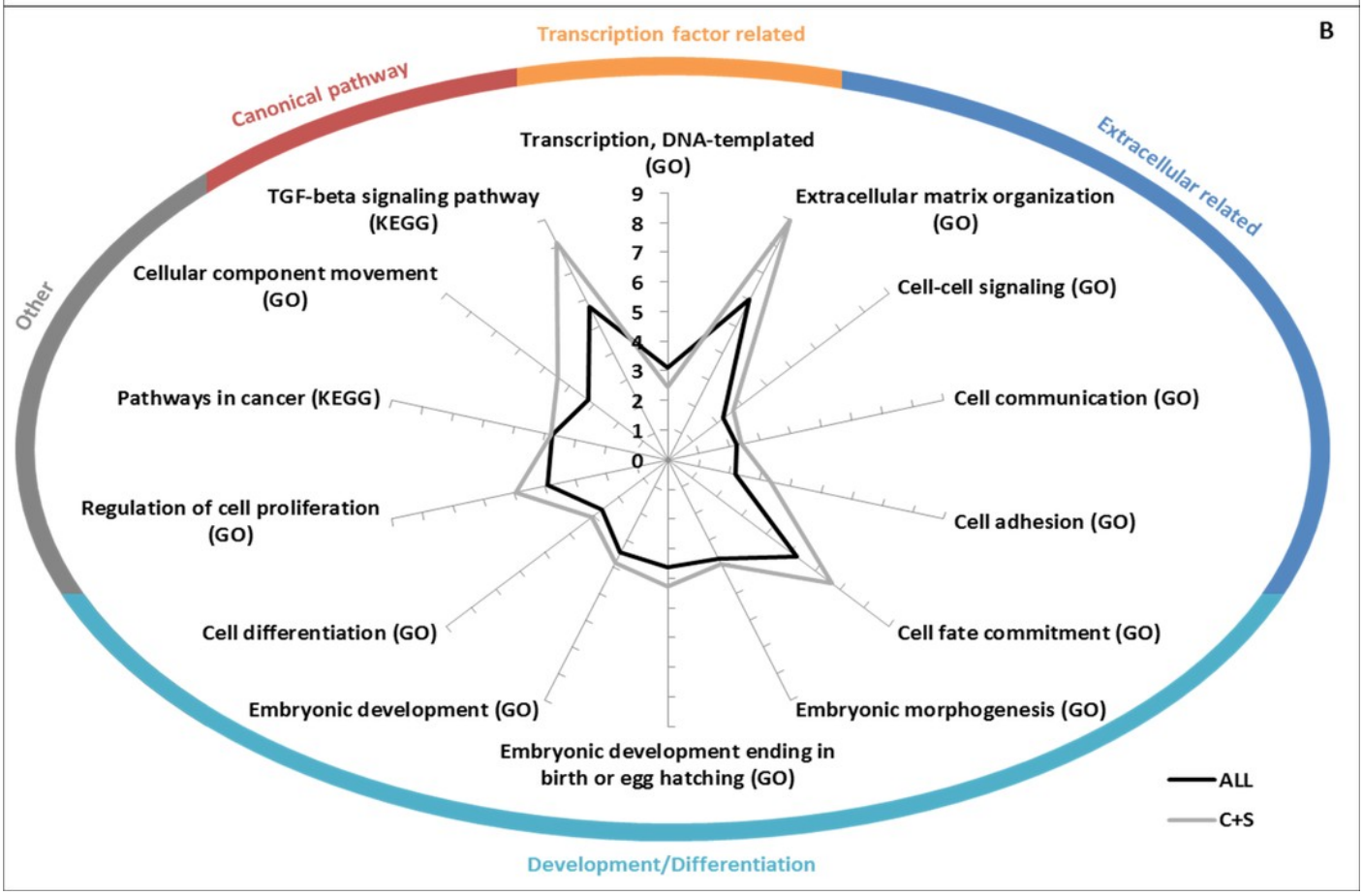


Figure 6

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).

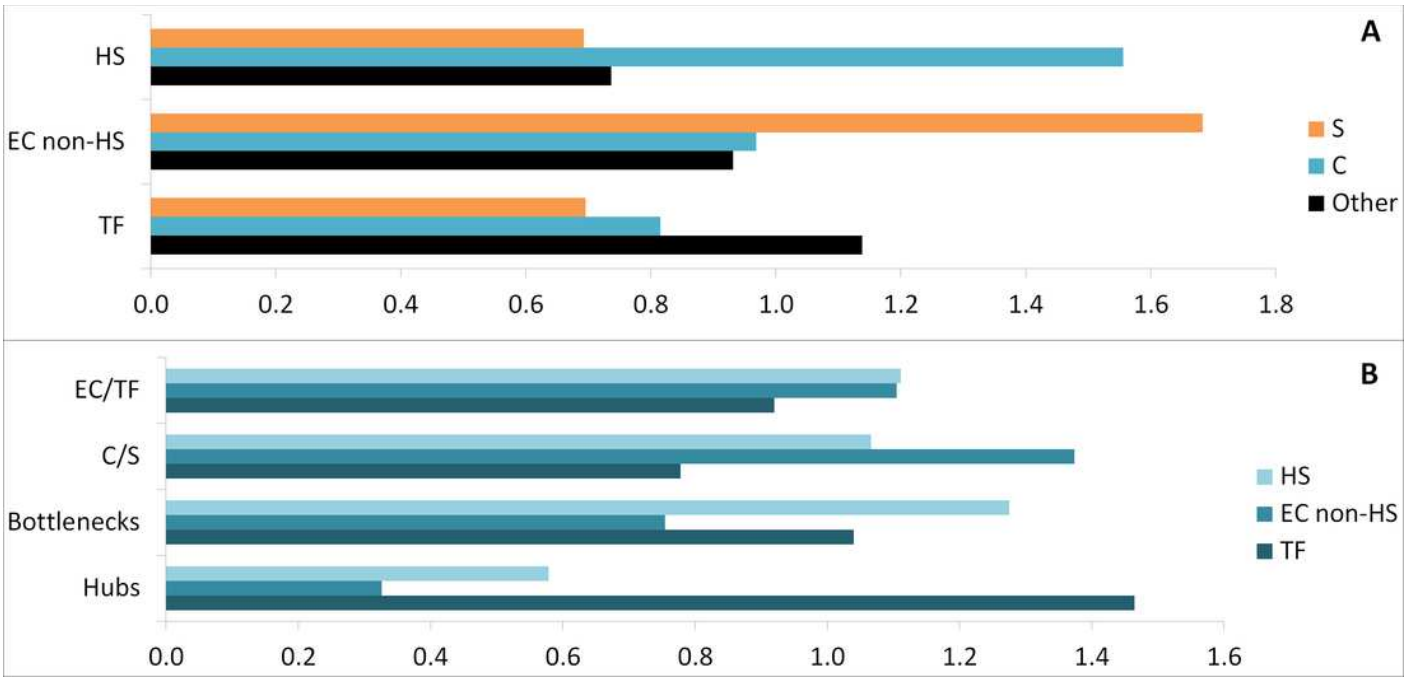


Table 1 (on next page)

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).

Cell type (Cell line)	Access Number	Linked publication	Main cell culture feature	Number of microarrays	RMA	GC-RMA	MAS5.0	TOTAL	INTERSECTION
hESCs (H14)	E-GEOD-6561	(Baker et al. 2007)	On feeder cells (irradiated MEFs) FGF2 (4 ng/mL) 20% KnockOut SR	4	9672	8680	9822	11930	7088
hESCs (H1, H7, H9, H13, H14)	E-GEOD-15148	(Yu et al. 2009)	On feeder cells (irradiated MEFs) FGF2 (100 ng/mL) 20% KnockOut SR OR On feeder-free matrigel Conditioned medium	10	8798	10554	10293	12467	7395
hESCs	E-GEOD-18265	/	On feeder cells (inactivated HFFs) FGF2 (10 ng/mL) 20% KnockOut SR	5	9602	11325	10966	13499	7791
hESCs (H1)	E-GEOD-26672	(Hu et al. 2011)	On feeder cells (irradiated MEFs) FGF2 (4 ng/mL) 20% KnockOut SR	5	9656	11285	10843	12790	8235
hESCs	**Mix1**	**/**		**24**	**11001**	**12173**	**11546**	**14043**	**8934**
Embryoid bodies	E-GEOD-9196	(Lu et al. 2007)		9	5142	6308	6667	8935	3576
Blast cells				9	4142	5866	6731	8617	3175
Fibroblasts	**E-GEOD-9832**	**(Park et al. 2008)**		**3**	**6471**	**7510**	**8072**	**9795**	**5086**
Neural progenitors	E-GEOD-9940	/		12	6939	8432	7309	11517	3944
Embryoid bodies				3	1356	2485	5009	5812	942
Hepatic cells	E-GEOD-14897	(Si-Tayeb et al. 2010)		3	1669	2659	4017	4864	1299
Endothelial cells	**E-GEOD-19735**	/		**4**	**7166**	**9237**	**8448**	**11085**	**5522**
Embryoid bodies				2	704	1832	1156	2079	583
Mesenchemal progenitors	E-GEOD-21668	(Evseenko et al. 2010)		3	861	1482	2543	3087	638
Differentiated cells	**Mix2**	**/**		**48**	**16174**	**14172**	**12134**	**17458**	**10730**

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.

	Marker/Family	Acronym	Name	Transcriptome	GO term
A	**Embryonic stem cell**	PTEN	phosphatase and tensin homolog	hESC (COMMON)	CS/CA/CC/Cco
		TERT	telomerase reverse transcriptase	hESC (SPECIFIC)	
		GDF3	growth differentiation factor 3	hESC (SPECIFIC)	EC non-HS
		NODAL	nodal homolog (mouse)	hESC (SPECIFIC)	CCo/EC non-HS
		ZIC3	Zic family member 3	hESC, Mix	TF
		SOX2	SRY (sex determining region Y)-box 2	hESC, Mix	CC/CCo/TF
		POU5F1	POU class 5 homeobox 1	hESC, Mix	CCo/TF
B	**Ectoderm**	NEFH	neurofilament, heavy polypeptide	hESC (COMMON)	CS
		TUBB3	tubulin, beta 3 class III	hESC, Mix	CCo
	Endoderm	KRT19	keratin 19	hESC (COMMON)	CS
		SOX7	SRY (sex determining region Y)-box 7	hESC, Endo, Mix	CCo/TF
	Mesoderm	KDR	kinase insert domain receptor (a type III receptor tyrosine kinase) (VEGFR)	hESC, Mix	CA/CCo/HS
		PDGFRA	platelet-derived growth factor receptor, alpha polypeptide	hESC (COMMON)	CS/CA/CCo
		VIM	Vimentin	hESC (COMMON)	CS
C	**Fibronectin**	FN1	fibronectin 1	hESC (COMMON)	CA/HS
		ITGA5	integrin, alpha 5 (fibronectin receptor, alpha polypeptide)	hESC, Mix	CA/CCo/J/HS
		ITGB1	integrin, beta 1 (fibronectin receptor, beta polypeptide, antigen CD29 includes MDF2, MSK12)	hESC, Endo, Mix	CS/CA/CC/CCo/J/HS
	Fibroblast Growth Factor	FGF2	Fibroblast growth factor 2	hESC (COMMON)	CC/CCo/HS/TF
		FGFR1	Fibroblast growth factor receptor 1	hESC (COMMON)	CC/CCo/HS
		FGFR2	Fibroblast growth factor receptor 2	hESC, Endo, Mix	CC/CCo/HS
		FGFR3	fibroblast growth factor receptor 3	hESC, Endo, Mix	CCo/J/HS
		FGFR4	Fibroblast growth factor receptor 4	hESC (SPECIFIC)	CCo/J/HS
	Activin A	ACVR1B	activin A receptor, type IB (ALK4)	hESC (COMMON)	CC/CCo/EC non-HS
		ACVR1C	activin A receptor, type IC	hESC (SPECIFIC)	CCo
		ACVR2A	activin A receptor, type IIA	hESC, F, Mix	CCo
		ACVR2B	activin A receptor, type IIB	hESC, Endo, Mix	CCo/EC non-HS
		INHBA	inhibin, beta A / Activin A	hESC (COMMON)	CC/CCo/HS

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 Onlotogy; 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).

	Acronym	Name	GO term	Hubs	Bottlenecks	S/C	EC/TF	KEGG pathway
C	ACTN4	actinin, alpha 4	EC non-HS/CS	17				4
	ACVR1B	activin A receptor, type IB (ALK4)	EC non-HS/CC/CCo			X	X	1
	ADM	adrenomedullin	EC non-HS/CCo		4050.8	X	X	0
	BMP2	bone morphogenetic protein 2	HS/CC/CCo		2693.9	X	X	3
	DMD	dystrophin	EC non-HS/CS/CCo			X		0
	FGFR1	Fibroblast growth factor receptor 1	HS/CC/CCo		2634.2	X	X	5
	FN1	fibronectin 1	HS/CA	26	6988.8		X	5
	IL6	Interleukin 6	HS/CCo				X	4
	INHBA	inhibin, beta A / Activin A	HS/CC/CCo			X		2
	ITGA6	integrin, alpha 6 (CD49f)	EC non-HS/CA/CCo/J	40	15218.1		X	6
	ITGAV	integrin, alpha V (vitronectin receptor)	HS/CA/CCo	40	12409.4		X	6
	JAM3	junctional adhesion molecule 3	EC non-HS/CA/CCo/J			X	X	2
	LAMA1	laminin, alpha 1	HS/CA/CCo	15			X	4
	MET	(hepatocyte growth factor receptor	HS/CS/CC/CCo	24	10501.7	X	X	6
	PLAT	plasminogen activator, tissue	HS/CCo		2552.8		X	0
	PLAU	plasminogen activator, urokinase	HS/CA/CCo				X	1
	SERPINE1	serpin peptidase inhibitor, clade E, member 1	HS/CA/CCo	24	10022.8		X	1
	SERPINI1	serpin peptidase inhibitor, clade I, member 1	EC non-HS/CA				X	0
	TGFB2	transforming growth factor, beta 2	HS/CA/CC/CCo	25	10580.8	X	X	6
	THBS1	thrombospondin 1	HS/CA/CC/CCo		2066.0		X	5
	VEGFA	vascular endothelial growth factor A	HS/CA/CCo	16	6817.1		X	6
S	CDH8	cadherin 8, type 2	HS/CA			X	X	0
	FGF4	fibroblast growth factor 4	HS/CA/CCo			X	X	3
	FGFR4	Fibroblast growth factor receptor 4	HS/CCo			X	X	4
	IDE	insulin-degrading enzyme	EC non-HS/CCo			X		0
	INHBE	inhibin, beta E	EC non-HS			X		2
	NODAL	nodal homolog (mouse)	EC non-HS/CCo			X		1
	PLXNB1	plexin B1	EC non-HS/CCo			X		0
	TTN	titin	EC non-HS/CS/CC			X	X	0
	WIF1	WNT inhibitory factor 1	EC non-HS/CCo			X	X	1