Interactome comparison of human embryonic stem cell lines with the inner cell mass and trophectoderm

Adam Stevens[1*], Helen Smith[1,2*], Terence Garner[1], Ben Minogue[2], Sharon Sneddon[2], Lisa Shaw[1], Rachel Oldershaw[2], Nicola Bates[2], Daniel R Brison[1,3], Susan J Kimber[2]

[1]Maternal and Fetal Health Research Centre, Division of Developmental Biology & Medicine, Faculty of Biology, Medicine and Health, University of Manchester

[2]Division of Cell Matrix Biology and Regenerative Medicine, Faculty of Biology, Medicine and Health, University of Manchester;

[3]Department of Reproductive Medicine, Saint Mary's Hospital, Manchester University NHS Foundation Trust; Oxford Road, Manchester M13 9WL.

And Manchester Academic Health Sciences Centre

*Authors contributed equally to this work

Address all correspondence to:

Susan J Kimber
Faculty of Biology Medicine and Health
Michael Smith Building
Oxford Road
Manchester M13 9PT, UK
Tel: +44 161 275 6773
E-mail: sue.kimber@manchester.ac.uk

Running title: Similarity of embryonic stem cell lines

Word count:

Figures:

Tables:

Abstract (150 words)

Networks of interacting co-regulated genes distinguish the inner cell mass (ICM) from the differentiated trophectoderm (TE) in the preimplantation blastocyst, in a species specific manner. In mouse the ground state pluripotency of the ICM appears to be maintained in murine embryonic stem cells (ESCs) derived from the ICM. This is not the case for human ESCs. In order to gain insight into this phenomenon, we have used quantitative network analysis to identify how similar human (h)ESCs are to the *human* ICM. Using the hESC lines MAN1, HUES3 and HUES7 we have shown that all have only a limited overlap with ICM specific gene expression, but that this overlap is enriched for network properties that correspond to key aspects of function including transcription factor activity and the hierarchy of network modules. These analyses provide an important framework which highlights the developmental origins of hESCs.

Highlights

- Similarity between clusters of co-expressed genes in the ICM and hESC differs between cell lines.
- hESC lines are enriched for highly connected genes in ICM and TE interactome network models
- MAN1 has a high proportion of genes with shared expression in ICM involved in transcriptional regulation.
- Hierarchy of network modules in ICM and TE can be used as a framework to compare hESC lines.

Glossary of Network Concepts

Modular Hierarchy – *Biological networks form regions of higher connectivity than would be expected by chance, known as modules. Modules represent functionally related elements of a network and their relative influence to a system can be estimated from their centrality.*

Metanode – *The most central ten connected genes within a module.*

Connectivity – *The number of links existing between a given node and its neighbours. An increased connectivity is indicative of a gene which is involved in numerous processes.*

Community Centrality – *A measure of the relative 'importance' of a node, characterised by high connectivity or connections between areas of high connectivity.*

Bridgeness – *A property of a node in a network which sits between two areas of high connectivity, such that if removed, it would cause the separation of a single module into two. These nodes act as 'bridges' between modules and an increased bridgeness identifies a node which connects multiple modules.*

Party hub – *A node with multiple connections which, in a biological system is thought to represent a gene with many active simultaneous interactions, such as protein complexes. It is characterised by a node which has a reduced bridgeness at a given centrality when compared to a date-hub*

Date hub – *A node with multiple connections which, in a biological system, has non-concurrent interactions with other nodes. These are thought to represent transcription factors. It is characterised by a node which has an increased bridgeness at a given centrality when compared to a party-hub*

Similarity Network Fusion – *A network approach which uses nearest neighbour relationships to combine datasets and identify regions of similarity within and between them. In the context of this manuscript, coherency between datasets represents genes whose expression patterns are conserved between cells derived from embryonic tissue and human embryonic stem cell lines.*

Introduction

Embryonic stem cell lines are generally derived from the inner cell mass of the preimplantation blastocyst. The proteins OCT4 (*POU5F1*), SOX2 and NANOG are core pluripotency-associated factors that define a network of interactions involved in self-renewal and maintenance of the pluripotent state for human and mouse embryonic stem cells[1]. Each of the core pluripotency factors has been detected in at least some early trophoblast cells, however, they have often not been detected in all cells of the inner cell mass (ICM)/epiblast, for a given embryo [2,3]. This heterogeneity has been confirmed by RNAseq analysis of single human preimplantation epiblast cells[4]. Recently the central role of OCT4 not only in maintenance of the inner cell mass stem cell population but also in the differentiation of the extra-embryonic trophectoderm (TE) has been established using CRISPR- Cas 9 gene editing in human preimplantation embryos and embryonic stem cells (ESCs)[5]. Data from the mouse and cynomolgus monkey indicate that the ICM generates a series of epiblast states before giving rise, after implantation, to progenitors of differentiated lineages[6-8]. Pluripotency–associated transcriptional networks continue to be expressed in the preimplantation human epiblast[4,9] and early post implantation cynomolgus epiblast[8]. Thus, the preimplantation epiblast has transcriptional heterogeneity which is likely to relate to initiation of differentiation events that take place in the early post implantation epiblast and will also impact the generation of ESCs.

Expression of a number of genes has been associated with the development of extraembryonic cell lineages including *Tead4*[10], *Tsfap2c*[11], *Gata3*[12] and *Cdx2*[13]. There is evidence suggesting divergence between species in the utilisation of some of these genes such as the Gata family[14-17] known to play a role in TE generation[8]. These observations imply that networks of interacting co-regulated proteins might distinguish the transiently pluripotent ICM/preimplantation epiblast from the early differentiated trophectoderm (TE) in a species specific manner.

In mouse the ground state pluripotency of the ICM appears to be maintained in murine ESCs derived from the ICM and cultured in the presence of LIF together with MEK and GSK3β inhibitors[7]. This is not the case for human ESCs derived from day 6-7 blastocysts and cultured in standard medium with TGFβ family molecules and FGF-2. It is established in the literature that human ESC lines have more similarities to the murine epiblast after implantation[18,19] than to the murine ICM and ESCs. In order to understand this difference, it is important to determine how similar hESCs are to the *human* ICM.

Transcriptional analysis of isolated ICM and TE samples from individual human embryos has also been performed, highlighting key metabolic and signalling pathways[20]. Recently a study of 1529 individual cells from 88 human preimplantation embryos has defined a transcriptional atlas of this stage of human development[4], however inter-individual heterogeneity has been shown to have a major effect on gene expression[21] (Smith *el al*, in review). Together these data show the relevance of transcriptome based analysis and highlight the need for approaches that account for inter-individual variation.

In the work presented here we have set out to examine how far the gene expression profile of ICM and TE have diverged from one another at the blastocyst stage when hESC derivation occurs and to compare these data to the transcriptome of hESCs. We have compared transcriptomic data between sets of matched TE and ICM pairs from the same human embryo and used these data to generate ICM and TE-specific interactome network models. This approach has allowed us to use quantitative network analysis to compare TE and ICM with hESCs and to evaluate the extent of similarity between ICM/TE and hESC cell lines. These analyses provide an important framework which highlights the development origins of hESCs.

Results

Similarities between the transcriptome of inner cell mass, trophectoderm and human embryonic stem cell lines.

Barcode Z scores for the entire transcriptome (n=54613 gene probe sets) were compared using PLSDA to assess the relationship between ICM, TE and the hESC lines MAN1, HUES3, HUES7 (**Figure 1**). The hESC sample groups were distinct from each other and from ICM and TE ($p<0.05$). All hESC cell lines were of equivalent distance from both ICM and TE along the X-axis (X-variate 1), however along the Y-axis (X-variate 2) MAN1 was closer to ICM than HUES3 or HUES7.

Gene expression unique to inner cell mass and trophectoderm and associated gene ontology

Gene barcode was used to isolate gene probe sets present in each embryonic cell line resulting in 2238 probe sets in ICM and 2484 probe sets in TE. This subset of transcriptome in the ICM and TE samples was used to determine the overlap and unique gene expression in each of these blastocyst tissues (**Figure 2A**). We found 881 and 1227 gene probesets uniquely expressed in the ICM and TE respectively, corresponding to 719 and 924 unique genes (**Supplemental Table S1**). The genes defined as having unique expression in ICM or TE significantly overlapped with single cell RNA-seq data from human epiblast and trophectoderm cells respectively (both $p<1.0\text{x}10^{-4}$), identified in previously published analysis[4].

The genes associated with ICM and TE were grouped by “biological process” ontology showing a similar proportion and ordering in both gene sets, the only difference being a reduction in the proportion of genes of the category “cell communication” in the TE compared to the ICM (**Figure 2B**). More detailed comparison of biological pathways identified “epithelial adherens junction signalling” (ICM $p=4.2 \times 10^{-5}$, TE $p=7.3 \times 10^{-4}$) as strongly associated with both TE and ICM, and EIF2 translation initiation activity (TE $p=4.4\text{x}10^{-6}$, ICM $p=0.39$) as significantly associated with TE, consistent with the TE being at an early stage of differentiation towards epithelium[22], with an active requirement for biosynthesis[23] (**Supplemental Table S2**).

It was noted that NANOG was strongly associated with the ICM ($p=5.9x10^{-6}$) but not the TE and that CDX2 was associated with TE ($p= 9.8x10^{-3}$) but not ICM, as would be anticipated[24]. Using causal network analysis we identified master regulators of gene expression associated with the transcriptomic data. This approach identified MYC ($p=7.6x10^{-8}$), a co-ordinator of OCT4 activity[25], and ONECUT1 (HNF6) ($p=4.0x10^{-8}$), a regulator of the development of epithelial cells[26], as the most significantly associated regulatory factors in ICM and TE respectively (**Supplemental Tables S3**).

Similarities between the inner cell mass and trophectoderm unique transcriptomes and the transcriptome of human embryonic stem cells

Similarity Network Fusion (SNF) was used to assess the similarity of gene expression patterns between cell lines. Genes form clusters within each cell line based on their expression patterns across each sample. We are able to identify regions where this pattern is *coherent* between MAN1, HUES3 or HUES7 and either ICM or TE. A region of coherency across a stem cell line and either TE or ICM represents a group of genes whose expression pattern is conserved between embryonic tissue and hESCs. The analysis highlighted a limited similarity of hESC lines with ICM (between 6% and 12% similarity) and TE (between 9% and 11%), consistent with the distance between the hESC lines and TE and ICM as observed by PLSDA analysis (**Figure 3A & Supplemental Figure S1**). Three primary clusters of similarity were identified in all comparisons between the hESC lines and ICM or TE (**Figure 3B**). These clusters were of equivalent similarity in TE with all hESC lines, as indicated by uniform yellow intensity indicating coherency with nearest co-expressed neighbours, implying highly co-ordinated expression. However, in ICM compared with hESCs, similar coherency was noted only with MAN1, and not with the other hESC lines. (**Figure 3B**).

An interactome network model of gene expression unique to ICM can be used as a framework to assess similarity with human embryonic stem cells.

An interactome network model can be used to consider the proteins derived from the differentially expressed genes and the proteins that they interact with. Using this approach allowed us to consider the wider context of biological influence generated by the gene expression unique to either the ICM or TE and to implement these models as a framework to assess similarity with the hESC lines. We first used the genes with unique expression in ICM and TE to generate interactome network models by inference to known protein-protein interactions (**Figure 4A & 4B**). As interactome networks account for inferred interactions these may be shared between models. Comparing the TE and ICM interactome network models there was an overlap of 5659 inferred genes that represent shared protein:protein interactions, accounting for 72% of the ICM interactome and 66% of the TE interactome .

Both networks were enriched for genes associated with pluripotency, for example NANOG with the ICM network and CDX2 with the TE network, as identified by gene ontology analysis. The ICM network contained 93/167 and 161/240 genes and the TE network contained 94/167 and 185/240 genes related to core pluripotency associated factors by RNAi[27-31] and protein interaction[31-35] screens respectively.

The shared transcriptome between ICM or TE and each human embryonic stem cell line was mapped onto the respective ICM or TE interactome network model. For ICM and MAN1 255 out of 517 shared genes (49%), for ICM and HUES3 405 out of 856 shared genes (47%) and for ICM and HUES7 463 out of 1010 shared genes (46%) mapped from the overlap of the transcriptomes to the network model. For TE and MAN1 335 out of 780 shared genes (43%), for TE and HUES3 512 out of 964 shared genes (53%) and for TE and HUES7 573 out of 1108 shared genes (52%) mapped from the overlap of the transcriptomes to the network model. Of the genes shared between the hESC lines and ICM there was no difference in the proportions shared with the network model (p=0.74), for the genes shared between the hESC lines and TE, MAN1 had a significantly smaller proportion of genes shared with the TE network model (p= 0.03).

Similarities and difference in topology between human embryonic stem cell lines in relation to inner cell mass and trophectoderm network models

As the ICM and TE interactome models shared a significant proportion of the same genes, we went on to assess the network topology of these models to determine further similarities and differences with the genes shared with the hESC lines. Analysis of the network topology of the ICM and TE interactome demonstrated that the genes shared with the hESC lines were enriched for highly connected genes (as measured by degree, the number of interactions made to other genes), with the enrichment seen not statistically different between the hESC lines (**Figure 6A & 6B**).

To further investigate the putative functional relevance of genes shared between the ICM or TE interactome models and the hESC lines we determined whether these genes had “party” or “date” like properties. In protein interaction networks party hubs co-ordinate local activity by protein complexes whereas date hubs regulate global effects and are assumed to represent the transient interactions that occur with transcription factors[36,37]. Date-like network hubs have been shown to possess a higher “bridgeness” property at any position within the interactome[38]. Bridgeness is a network property that measures overlap between network modules and this score can be compared at different positions within the network by plotting it against “centrality” a network property that measures the influence of a node in a network[38]. All three hESC lines were shown to be enriched for bridgeness score in relation to centrality when compared to the full ICM or TE networks (**Figure 6C & 6D**). This observation implies an enrichment for date-like network hubs in the genes shared between the hESC lines and the ICM or TE interactome network models, implying in turn an enrichment of transcription factor activity.

Previously we identified the overlap of genes expressed in the ICM or TE and the hESC cell lines (**Figure 5**). There were 590 and 652 genes shared between all the three hESC lines and ICM and TE respectively

(**Supplemental Figure S2A**). When we examined genes uniquely expressed in each of the hESC lines (**Supplemental Figure S2A**), the highly central genes in both networks (centrality score >100) were significantly enriched for bridgeness in ICM (p=0.016) but not TE (p=0.105), indicating more date-like properties in ICM (**Figure 6E & 6F**). In the ICM interactome network model MAN1 was significantly more date-like than HUES3 (p=0.048) and HUES7 (p=0.012). This observation implies that the MAN1 cell line shared significantly more transcription factor activity and that these are hierarchically more important within the ICM interactome, than either HUES3 or HUES7. Biological pathways associated with genes uniquely expressed in each of the hESC lines are shown in **Supplemental Figure S2B**. In MAN1 “PDGF signalling” and “cell cycle control of chromosome replication” were associated with the unique gene expression shared with ICM. PDGF signalling is required for primitive endoderm cell survival in the inner cell mass of the mouse blastocyst[39]. The cell cycle control of chromosome replication is influenced in the control of pluripotency by NANOG[40].

Modular hierarchy of the ICM and TE interactome network models reveal a greater proportion of modules enriched for MAN1 in ICM and HUES7 in TE

Network modules are sub-structures of a network that have a greater number of internal connections than expected by chance. Modules are known to represent functionally related elements of a network and can be ranked hierarchically by their centrality within a network, with the assumption that the more central modules are functionally dominant within the network. We defined modules within the TE and ICM interactome network modules allowing for overlap and arranged these into a hierarchy of influence by centrality score[38] (**Figure 7A**). The ICM and TE interactome network models had a hierarchy of 163 and 201 modules of different sizes respectively. There was no difference in the proportion of modules compared to network size between the ICM and TE interactome network models (p=0.2) (**Supplemental Figure S3 & Supplemental Tables S4 & S5**). The robustness of the definition of network modules in the ICM and TE interactome network models were confirmed by permutation analysis of the proportional random removal of genes (**Supplemental Figure S4**). This established that the majority of modules were robust to the removal of large proportions of the network, with only 2 of the top 47 ICM and 8 of the top 49 TE modules analysed experiencing a significant ($p<0.05$) reduction in connectivity within the module following the removal of a random 20% of the network iterated 100 times.

The genes with shared expression between ICM or TE and the hESC lines were mapped to each interactome module. In the ICM network 116/163 modules (71%) were enriched for gene expression shared between hESC lines and ICM. A greater proportion of hESC associated modules in the ICM interactome network model were enriched for MAN1 gene expression (0.46) compared to HUES3 (0.28) and HUES7 (0.25) ($p=9.0x10^{-4}$, chi squared test). In the TE interactome network model 132/201 modules (65%) were enriched for gene expression shared between hESC lines and TE. The smallest proportion of enriched hESC associated modules occurred in HUES7 (0.17) compared to MAN1 (0.39) and HUES3 (0.44) ($p=3.1x10^{-6}$, chi squared test) (**Figure 7B**).

The modules assessed as having enriched gene expression in specific hESC lines were mapped to the module hierarchy in the ICM or TE interactome network model (**Figure 7C**). These data show an enrichment of the modules that have the greatest proportion of shared gene expression with MAN1 in the upper part of the module hierarchy in both ICM and TE indicating that the MAN1 associated modules were likely to be more functionally active in both the ICM and TE interactomes.

Gene expression uniquely present in each of the hESC lines (**Supplemental Figure S2A**) was mapped to the central core (most central 10 genes) of each of the modules in the ICM and TE interactome network models (**Supplemental Figure S5**). This analysis highlighted only gene expression present uniquely in MAN1 or HUES7 in the upper part of the module hierarchy in the ICM and TE interactome network models indicating that HUES3 associated modules had a reduced role in the function of the ICM. The upper part of the TE network model module hierarchy was enriched for both HUES7 and MAN1 uniquely expressed genes, indicating a dominant effect of these hESC lines on TE function, compared to HUES3.

Finally, relating these analyses to the enrichment for pluripotency associated genes we defined in the ICM and TE interactome models, we examined this relationship to the modular hierarchy of the ICM and TE interactome network models. We assessed whether any of the pluripotent genes mapped to the central core of 10 genes in a network module (coloured black in **Figure 6C**). In the ICM modular hierarchy 16, 13 and 11 of the modules enriched for MAN1, HUES3 and HUES7 respectively were also mapped to by pluripotent genes. In the TE modular hierarchy 18, 11 and 15 of the modules enriched for MAN1, HUES3 and HUES7 respectively were also mapped to by pluripotent genes. It was noted that OCT4 (*POU5F1*), a primary marker of ICM[41], was present in the central core of the modules from the ICM but not the TE network models. *NANOG*, another marker of ICM[41], was present four times in the ICM and only once in the TE network models. Also *ESRRB*, a marker of TE[42,43], was present three times in the TE but not at all in the ICM network models. In the ICM network model, 2 of the 3 NANOG associated modules are enriched for MAN1 gene expression and the module associated with both NANOG and OCT4 had equivalent enrichment for MAN1, HUES3 and HUES7. In the TE network model the NANOG associated module was low in the hierarchy (76/201) and had equivalent enrichment for MAN1, HUES3 and HUES7. In the TE network model the three ESRRB associated modules were at the upper end of the module hierarchy with the highest ranked (8/201) being enriched for HUES3 and HUES7 and the other two being associated with MAN1 (**Figure 6C**). These data combined show that the key transcription factors (and partners) known to be associated with ICM and TE have the expected association with hESC lines within the modular hierarchies of the interactome network models.

Discussion

The analysis presented in this manuscript has defined a gene interactome network model of ICM and TE and used these to quantitatively assess the relationship to pluripotency of human embryonic stem cell lines derived from the ICM.

The MAN1 human embryonic stem cell line was furthest from both ICM and TE using distance metrics on the unsupervised transcriptome. Only ~10% of genes uniquely expressed by the ICM (compared to

TE) were shown to have similarity to expression patterns in MAN1, HUES3 and HUES7 using Similarity Network Fusion (SNF). However MAN1 was found to be most similar to ICM as it had both a greater enrichment of genes and a greater coherency with nearest neighbours in comparison to HUES3 and HUES7. Substantial enrichment of human embryonic stem cell line gene expression was also observed in relation to TE but, whilst this was shown to be coherent with nearest neighbours, it was at a reduced proportion of similarity compared to ICM in MAN1 and HUES7 and an increased proportion in HUES3.

We used interactome network models of ICM and TE as frameworks to map overlapping gene expression from MAN1, HUES3 and HUES7. Using network topology as a marker of functionality we demonstrated that all the human embryonic stem cell lines had increased connectivity in both the ICM and TE interactome network models generated from gene expression data. All human embryonic stem cell lines also showed an enrichment for network topology that was associated with date hubs more so than party hubs, in ICM and TE network models. Date hubs are network positions that are associated with non-concurrent signalling and are more likely to represent transcription factor activity related to the execution of a developmental programme[31,36-38]. A key finding of this study is that date hubs central to the network model and therefore likely to influence a greater proportion of network function were significantly enriched in the overlap of genes uniquely shared between MAN1 and the ICM compared to genes uniquely shared between HUES3 or HUES7 and ICM.

We defined a functional hierarchy of overlapping network modules in both the ICM and TE interactome network models and used this as a framework to study the relationship of MAN1, HUES3 and HUES7 with ICM and TE activity. MAN1 was shown to have the greatest proportion of shared expression with the ICM network modules and HUES7 had the greatest proportion of shared expression with the TE network modules. MAN1 had greater enrichment in the upper hierarchy for both ICM and TE network models both overall and for uniquely expressed genes.

Taken together these observations demonstrate the utility of network approaches to quantify underlying similarities based on the position of transcriptomic differences in an interactome network model. Quantitative comparison of the hierarchy of the ICM and TE interactome network modules in relation to the expressed genes in the human embryonic stem cell lines provided further insight into similarities and differences between the cell lines beyond those defined by traditional distance metrics.

An assessment of master regulators of transcription associated with the ICM and TE specific gene expression identified known tissue specific transcriptional regulators – NANOG in ICM[31,41] and CDX2 in TE[9,44]. Both the ICM and TE network models were enriched for genes associated with pluripotency[31,41]. The upper part of the hierarchy of network modules in both the ICM and the TE interactome network models was enriched for pluripotency associated genes. However MAN1 was more closely associated with gene modules including NANOG in the ICM interactome network model compared to HUES3 and HUES7 cell lines. In the TE interactome network model HUES3 and HUES7 were associated with the estrogen-related-receptor beta (*ESRRB*) related module at the highest position in the module hierarchy whilst MAN1 was also primarily associated with two further *ESSRB* related modules. ESSRβ, a direct target of Nanog[45], has been shown to be important in murine ES cells as a co-regulator of Oct 4 with Nanog[46] and a regulator of GATA6 though promoter binding[47]. ESSRβ works with p300 to maintain pluripotency networks generating a permissive chromatin state for binding of Oct4, Nanog and Sox2 and has been implicated in reprogramming Epistem cells to an iPSC state[48]. Thus the prevalence of ESRRβ in the hESC interactome could be interpreted as indicating hESC line position in the spectrum from the naïve to the Epistem like state, but further work would be needed to confirm this.

Overall these data identify that the MAN1 cell line had the greatest similarity to ICM compared to the other human embryonic stem cell lines despite being least related to ICM in the PLS-DA analysis. This observation is based on **I)** greater coherency in the SNF analysis with nearest neighbour genes, **II)** significantly increased proportion of genes with a date-like hub property in the ICM network, **III)** an increased proportion of genes mapping to ICM interactome network model modules and **IV)** an association with ICM network gene modules that map to NANOG activity. We propose that the network approach presented in this manuscript represents a significant advance on distance metrics in the comparison on hESC lines.

By using a barcode approach to define genes uniquely expressed we were able to define ICM- and TE-specific interactome network models, an important advance on more traditional comparative modelling using differential gene expression[49-51]. We also confirmed similarity of the underlying transcriptomic data with findings from single cell RNAseq data[4] adding confidence to our observations. These comparisons also confirmed the importance of network structure in the analysis we have undertaken[52]. We demonstrated the robustness of our network model by establishing module coherency over successive reductions of network model size (by gene removal) therefore establishing a high level of confidence in the analysis of related gene modules and network topology[53].

The differences between ICM and TE with all three hESC lines may partially reflect the genetic background of the infertile couples donating embryos for analysis and stem cell derivation. Our previous analysis of single cell ICM and TE RNAseq data showed a strong effect of inter-individual genetic variation (Smith *et al*, unpublished). To account for this we have restricted our analysis in this manuscript to only genetically matched pairs. The similarities we have established by comparison to other work[4] would suggest that the data presented in this manuscript is robust to inter-individual differences. The greater dissimilarity of MAN1 to HUES7 and HUES3, revealed in the overlap of the transcriptome to the ICM interactome network modules, may indicate genetic background has a greater effect than derivation regimen since HUES3 and HUES7 were derived in the same lab at a similar time[54,55]. However it should be noted that all hESC lines were enriched for connectivity, a marker of function, within the ICM interactome, an observation in alignment with a fundamental similarity between hESC lines, despite different genetic background and embryo generation and hESC derivation methods[54]. It was also noted that hESC lines are different in very many gene modules to ICM. Although the ICMs have totally different genetic back ground to the hESC lines assessed here, the fact that the genetically matched ICMs and TEs are inherently more similar to each other than to the hESCs does add further weight to this conclusion.

The use of network approaches to quantify similarities between hESCs and their tissue of origin is a developing field. Network summary approaches have been used with promising results (e.g. CellNet[56]). Correlation networks generated from gene expression have been used to generate quantitative comparison based on the analysis of discrete network modules[57]. Network driven approaches can also be used to deal with the large number of comparisons present in the analysis of 'omic data sets, e.g. topological data analysis (TDA)[52] and similarity network fusion (SNF)[58]. In the work presented here we have used an efficient method to generate hierarchies of overlapping gene modules[38,59], thus accounting for the underlying network topology, and supported this analysis using SNF[58] to generate quantitative comparison of hESC lines with ICM and TE. The approach we have developed accounts for both the hierarchy of modules within a network and the large number of comparisons performed in an unsupervised manner to generate robust conclusions. This has allowed us to apply quantitative approaches to determine the similarity of three hESC lines to each other in relation to ICM and TE. We have identified overall similarity of the transcriptomes and we have also defined how these similarities manifest at the level of the interactome. Our findings highlight the diversity inherent in the establishment of hESC lines and also present methods to quantitatively compare similarity and identify key differences using a network approach.

Methods

Embryos

Human oocytes and embryos were donated to research with fully informed patient consent and approval from Central Manchester Research Ethics Committee under Human Fertility and Embryology Authority research licences R0026 and R0171. Fresh oocytes and embryos surplus to IVF requirement were obtained from Saint Mary's Hospital Manchester, graded and prepared as described in Shaw et al 2013 [60].

Embryo sample preparation and microarray analysis of transcriptome

Donated embryos were cultured in ISM-1/2 sequential media (Medicult, Jyllinge, Denmark) until blastocyst formation. At embryonic day 6 the zona pellucida of the embryos were removed by brief treatment with acid Tyrode's solution pH 5.0 (Sigma-Aldrich, Gillingham, UK), and denuded blastocysts were washed in ISM2 (Medicult). Four blastocysts were lysed and reverse transcribed as previously described [61,62] and cDNA was prepared by polyAPCR amplification[63] which amplifies all polyadenylated RNA in a given sample, preserving the relative abundance` in the original sample[64,65]. A second round of amplification using EpiAmp™ (Epistem, Manchester, UK) and biotin-16-dUTP labelling using EpiLabel™ (Epistem) was performed in the Paterson Cancer Research Institute Microarray facility. For each sample, our minimum inclusion criterion was the expression of β-actin evaluated by gene-specific PCR. Labelled PolyAcRNA was hybridised to the Human Genome U133 Plus 2.0 Array (hgu133plus2.0, Affymetrix, SantaClara, CA, USA) and data initially visualised using MIAMIVICE software. Quality control of microarray data was performed using principal component analysis with cross-validation undertaken using Qlucore Omics Explorer 2.3 (Qlucore, Lund, Sweden).

The trophectoderm (TE) and inner cell mass (ICM) of 6 day human embryos were separated by immunosurgically lysing the whole TE (recovering RNA from both mural and polar TE), to leave a relatively pure ICM. Eight microarray datasets were obtained, corresponding to 4 genetically paired matched TE and ICM transcriptomes. Frozen robust multiarray averaging (fRMA)[66] was used to define absolute expression by comparison to publically available microarray datasets within R (3.1.2)[67]. An expression barcode and a z-score of gene expression in comparison to 63331 examples of hgu133plus2.0 was defined for each tissue [50,51] and used for unsupervised analysis. For analysis of gene expression specific to each tissue a z-score of 5 was used to call a gene present and a barcode was assigned scoring 1 for presence and 0 for absence of gene expression[49,50,66]. All Transcriptomic data will be made available on the Gene Expression Omnibus (GEO).

hESC lines

HuES7, HuES3 (kind gift of Kevin Eggan[55]) and MAN1[68] hESC lines were cultured as previously described[69]. Briefly, hESCs (p21-27) were cultured and expanded on Mitomycin C inactivated mouse embryonic fibroblasts (iMEFs) in hESC medium KO-DMEM (Invitrogen, Paisley, UK) with 20% knockout serum replacement (KO-SR, Invitrogen), 8 ng/ml basic fibroblast growth factor (bFGF, Invitrogen), 2 mM L-glutamine, 1% NEAA (both from Cambrex, Lonza Wokingham, UK), and 0.1 mM ß-mercaptoethanol (Sigma-Aldrich, Dorset, UK). For feeder-free culture, cells were lifted from the iMEF layers with TrypLE, and plated onto fibronectin-coated (Millipore) tissue culture flasks with Stem Pro feeder free medium. After 3 passages 100 hESC cells were isolated from each line (assessed separately as > 85% Oct4 positive), lysed and subjected to polyAPCR amplification, hybridisation and analysis as described above.

Analysis of differential gene expression

Principal component analysis was performed to provide further quality control using cross-validation (Qlucore Omics Explorer [QoE] 2.3). Partial least square discriminant analysis (PLSDA) was used to assess the Euclidean distance between the unsupervised transcriptomic samples using the MixOmics package for R[70].

We analysed published single-cell RNA-Seq data from human epiblast (inner cell mass) and trophectoderm tissue[4]. Transcripts per million (TPM) expression values were visualised in QoE and outliers were removed.

Similarity Network Fusion

Gene probe set similarity network fusion (SNF)[58] was performed on the fRMA derived data as an independent test for similarity, using the *SNFTool* R-package. Euclidean distances were calculated between gene probe sets for each hESC line as well as TE and ICM. Using a non-linear network method based on nearest neighbours, any two of the Euclidean distance matrices could be combined over 20 iterations to produce a final network which accurately describes the relationship between gene probe sets across both initial sets. This method was used to combine each hESC line with TE or ICM gene expression data. The fused data was subjected to spectral clustering to identify groups of gene probe sets with similar patterns of expression across the hESC and TE or ICM cell lines. This data was presented as a heatmap.

Network model construction and comparison

Lists of differentially expressed genes were used to generate interactome network models of protein interactions related to the transcriptomic data in Cytoscape[71] by inference using the BioGRID database[72].

The Cytoscape plugin Moduland[38,59] was applied to identify overlapping modules, an approach that models complex modular architecture within the human interactome[37] by accounting for non-discrete nature of network modules[38]. Modular hierarchy was determined using a centrality score and further assessed using hierarchical network layouts (summarising the underlying network topology). The overlap between the central module cores (metanode of the ten most central elements) was determined. Community centrality and bridgeness scores were assessed across network models using the Moduland package[59]. The bridgeness score was used in combination with centrality scores to categorise party and date hubs within the network i.e genes that interact simultaneously or sequentially respectively with neighbours[73,74].

The Network Analyser[75] Cytoscape plugin was used to calculate associated parameters of network topology. Hierarchical network layouts were used along with centrality scores to assess the hierarchy of network clusters. Significance of the overlap between network elements was calculated using the Fisher's exact test on the sum of each group compared to the expected sum.

The robustness of defined modules is an essential analytical step[53] and was assessed using permutation analysis in R (version 3.3.2)[76]. Robustness of network module and network topology properties was determined in the ICM interactome network model with 100 permutations of removal of 5, 10, 20, 30, 40 and 50% of the nodes, an approach that has been shown to assess the coherency of network modules[53]. These data were used to assess the stability of network observations.

Acknowledgements: Stuart Pepper Christie Genomics unit for Microarray hybridisation. We thank the Medical Research Council UK for funding (grants G0801057 MR/M017354/1) and the NIHR clinical research network for support. We would particularly like to thank the research nurses, IVF clinic staff and patients who donated embryos to this research.

References

1 Boyer, L. A. *et al.* Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell* **122**, 947-956, doi:10.1016/j.cell.2005.08.020 (2005).

2 Cauffman, G., De Rycke, M., Sermon, K., Liebaers, I. & Van de Velde, H. Markers that define stemness in ESC are unable to identify the totipotent cells in human preimplantation embryos. *Human reproduction (Oxford, England)* **24**, 63-70, doi:10.1093/humrep/den351 (2009).

3 Kimber, S. J. *et al.* Expression of genes involved in early cell fate decisions in human embryos and their regulation by growth factors. *Reproduction (Cambridge, England)* **135**, 635-647, doi:10.1530/REP-07-0359 (2008).

4 Petropoulos, S. *et al.* Single-Cell RNA-Seq Reveals Lineage and X Chromosome Dynamics in Human Preimplantation Embryos. *Cell* **165**, 1012-1026, doi:10.1016/j.cell.2016.03.023 (2016).

5 Fogarty, N. M. E. *et al.* Genome editing reveals a role for OCT4 in human embryogenesis. *Nature* **550**, 67-73, doi:10.1038/nature24033 (2017).

6 Han, D. W. *et al.* Epiblast stem cell subpopulations represent mouse embryos of distinct pregastrulation stages. *Cell* **143**, 617-627, doi:10.1016/j.cell.2010.10.015 (2010).

7 Weinberger, L., Ayyash, M., Novershtern, N. & Hanna, J. H. Dynamic stem cell states: naive to primed pluripotency in rodents and humans. *Nat Rev Mol Cell Biol* **17**, 155-169, doi:10.1038/nrm.2015.28 (2016).

8 Nakamura, T. *et al.* A developmental coordinate of pluripotency among mice, monkeys and humans. *Nature* **537**, 57-62, doi:10.1038/nature19096 (2016).

9 Niakan, K. K. & Eggan, K. Analysis of human embryos from zygote to blastocyst reveals distinct gene expression patterns relative to the mouse. *Dev Biol* **375**, 54-64, doi:10.1016/j.ydbio.2012.12.008 (2013).

10 Nishioka, N. *et al.* Tead4 is required for specification of trophectoderm in pre-implantation mouse embryos. *Mechanisms of Development* **125**, 270-283, doi:https://doi.org/10.1016/j.mod.2007.11.002 (2008).

11 Kuckenberg, P., Kubaczka, C. & Schorle, H. The role of transcription factor Tcfap2c/TFAP2C in trophectoderm development. *Reproductive biomedicine online* **25**, 12-20, doi:10.1016/j.rbmo.2012.02.015 (2012).

12 Home, P. *et al.* GATA3 is selectively expressed in the trophectoderm of peri-implantation embryo and directly regulates Cdx2 gene expression. *J Biol Chem* **284**, 28729-28737, doi:10.1074/jbc.M109.016840 (2009).

13 Strumpf, D. *et al.* Cdx2 is required for correct cell fate specification and differentiation of trophectoderm in the mouse blastocyst. *Development* **132**, 2093-2102, doi:10.1242/dev.01801 (2005).

14 Grabarek, J. B. *et al.* Differential plasticity of epiblast and primitive endoderm precursors within the ICM of the early mouse embryo. *Development* **139**, 129-139, doi:10.1242/dev.067702 (2012).

15 Rossant, J., Chazaud, C. & Yamanaka, Y. Lineage allocation and asymmetries in the early mouse embryo. *Philos Trans R Soc Lond B Biol Sci* **358**, 1341-1348; discussion 1349, doi:10.1098/rstb.2003.1329 (2003).

16 Stephenson, R. O., Rossant, J. & Tam, P. P. Intercellular interactions, position, and polarity in establishing blastocyst cell lineages and embryonic axes. *Cold Spring Harbor perspectives in biology* **4**, doi:10.1101/cshperspect.a008235 (2012).

17 Schrode, N. *et al.* Anatomy of a blastocyst: cell behaviors driving cell fate choice and morphogenesis in the early mouse embryo. *Genesis* **51**, 219-233, doi:10.1002/dvg.22368 (2013).

18 Faial, T. *et al.* Brachyury and SMAD signalling collaboratively orchestrate distinct mesoderm and endoderm gene regulatory networks in differentiating human embryonic stem cells. *Development* **142**, 2121-2135, doi:10.1242/dev.117838 (2015).

19 Tesar, P. J. *et al.* New cell lines from mouse epiblast share defining features with human embryonic stem cells. *Nature* **448**, 196-199, doi:10.1038/nature05972 (2007).

20 Adjaye, J. *et al.* Primary differentiation in the human blastocyst: comparative molecular portraits of inner cell mass and trophectoderm cells. *Stem cells (Dayton, Ohio)* **23**, 1514-1525, doi:10.1634/stemcells.2005-0113 (2005).

21 Stirparo, G. G. *et al.* Integrated analysis of single-cell embryo data yields a unified transcriptome signature for the human preimplantation epiblast. *Development*, doi:10.1242/dev.158501 (2018).

22 Marikawa, Y. & Alarcon, V. B. Creation of trophectoderm, the first epithelium, in mouse preimplantation development. *Results and problems in cell differentiation* **55**, 165-184, doi:10.1007/978-3-642-30406-4_9 (2012).

23 Hasegawa, Y. *et al.* Variability of Gene Expression Identifies Transcriptional Regulators of Early Human Embryonic Development. *PLoS Genet* **11**, e1005428, doi:10.1371/journal.pgen.1005428 (2015).

24 Niakan, K. K. & Eggan, K. Analysis of human embryos from zygote to blastocyst reveals distinct gene expression patterns relative to the mouse. *Developmental Biology* **375**, 54-64, doi:https://doi.org/10.1016/j.ydbio.2012.12.008 (2013).

25 Fang, L. *et al.* H3K4 Methyltransferase Set1a Is A Key Oct4 Coactivator Essential for Generation of Oct4 Positive Inner Cell Mass. *Stem cells (Dayton, Ohio)* **34**, 565-580, doi:10.1002/stem.2250 (2016).

26 Pierreux, C. E. *et al.* The transcription factor hepatocyte nuclear factor-6 controls the development of pancreatic ducts in the mouse. *Gastroenterology* **130**, 532-541, doi:10.1053/j.gastro.2005.12.005 (2006).

27 Hu, G. *et al.* A genome-wide RNAi screen identifies a new transcriptional module required for self-renewal. *Genes & development* **23**, 837-848, doi:10.1101/gad.1769609 (2009).

28 Ding, L. *et al.* A genome-scale RNAi screen for Oct4 modulators defines a role of the Paf1 complex for embryonic stem cell identity. *Cell stem cell* **4**, 403-415, doi:10.1016/j.stem.2009.03.009 (2009).

29 Zhang, J. Z. *et al.* Screening for genes essential for mouse embryonic stem cell self-renewal using a subtractive RNA interference library. *Stem cells (Dayton, Ohio)* **24**, 2661-2668, doi:10.1634/stemcells.2006-0017 (2006).

30 Ivanova, N. *et al.* Dissecting self-renewal in stem cells with RNA interference. *Nature* **442**, 533-538, doi:10.1038/nature04915 (2006).

31 Ng, P. M. & Lufkin, T. Embryonic stem cells: protein interaction networks. *Biomolecular concepts* **2**, 13-25, doi:10.1515/bmc.2011.008 (2011).

32 Liang, J. *et al.* Nanog and Oct4 associate with unique transcriptional repression complexes in embryonic stem cells. *Nature cell biology* **10**, 731-739, doi:10.1038/ncb1736 (2008).

33 Pardo, M. *et al.* An expanded Oct4 interaction network: implications for stem cell biology, development, and disease. *Cell stem cell* **6**, 382-395, doi:10.1016/j.stem.2010.03.004 (2010).

34 van den Berg, D. L. *et al.* An Oct4-centered protein interaction network in embryonic stem cells. *Cell stem cell* **6**, 369-381, doi:10.1016/j.stem.2010.02.014 (2010).

35 Wang, J. *et al.* A protein interaction network for pluripotency of embryonic stem cells. *Nature* **444**, 364-368, doi:10.1038/nature05284 (2006).

36 Agarwal, S., Deane, C. M., Porter, M. A. & Jones, N. S. Revisiting Date and Party Hubs: Novel Approaches to Role Assignment in Protein Interaction Networks. *PLOS Computational Biology* **6**, e1000817, doi:10.1371/journal.pcbi.1000817 (2010).

37 Chang, X., Xu, T., Li, Y. & Wang, K. Dynamic modular architecture of protein-protein interaction networks beyond the dichotomy of 'date' and 'party' hubs. *Sci Rep* **3**, 1691, doi:10.1038/srep01691 (2013).

38 Kovacs, I. A., Palotai, R., Szalay, M. S. & Csermely, P. Community landscapes: an integrative approach to determine overlapping network module hierarchy, identify key nodes and predict network dynamics. *PLoS One* **5**, doi:10.1371/journal.pone.0012528 (2010).

39 Artus, J., Kang, M., Cohen-Tannoudji, M. & Hadjantonakis, A. K. PDGF signaling is required for primitive endoderm cell survival in the inner cell mass of the mouse blastocyst. *Stem cells (Dayton, Ohio)* **31**, 1932-1941, doi:10.1002/stem.1442 (2013).

40 Apostolou, E. *et al.* Genome-wide chromatin interactions of the Nanog locus in pluripotency, differentiation, and reprogramming. *Cell stem cell* **12**, 699-712, doi:10.1016/j.stem.2013.04.013 (2013).

41 Hochedlinger, K. & Jaenisch, R. Induced Pluripotency and Epigenetic Reprogramming. *Cold Spring Harbor perspectives in biology* **7**, doi:10.1101/cshperspect.a019448 (2015).

42 Latos, P. A. *et al.* Fgf and Esrrb integrate epigenetic and transcriptional networks that regulate self-renewal of trophoblast stem cells. *Nature Communications* **6**, 7776, doi:10.1038/ncomms8776

https://www.nature.com/articles/ncomms8776#supplementary-information (2015).

43 Nicola, F., Nick, O. & Pablo, N. Esrrb, an estrogen-related receptor involved in early development, pluripotency, and reprogramming. *FEBS Letters* **592**, 852-877, doi:doi:10.1002/1873-3468.12826 (2018).

44 Niwa, H. *et al.* Interaction between Oct3/4 and Cdx2 determines trophectoderm differentiation. *Cell* **123**, 917-929, doi:10.1016/j.cell.2005.08.040 (2005).

45 Festuccia, N. *et al.* Esrrb is a direct Nanog target gene that can substitute for Nanog function in pluripotent cells. *Cell stem cell* **11**, 477-490, doi:10.1016/j.stem.2012.08.002 (2012).

46 Zhang, X., Zhang, J., Wang, T., Esteban, M. A. & Pei, D. Esrrb activates Oct4 transcription and sustains self-renewal and pluripotency in embryonic stem cells. *J Biol Chem* **283**, 35825-35833, doi:10.1074/jbc.M803481200 (2008).

47 Uranishi, K., Akagi, T., Koide, H. & Yokota, T. Esrrb directly binds to Gata6 promoter and regulates its expression with Dax1 and Ncoa3. *Biochemical and Biophysical Research Communications* **478**, 1720-1725, doi:https://doi.org/10.1016/j.bbrc.2016.09.011 (2016).

48 Adachi, K. *et al.* Esrrb Unlocks Silenced Enhancers for Reprogramming to Naive Pluripotency. *Cell stem cell* **23**, 266-275.e266, doi:10.1016/j.stem.2018.05.020 (2018).

49 McCall. Frozen Robust Multi-Array Analysis and the Gene Expression Barcode. (2015).

50 McCall, M. N. *et al.* The Gene Expression Barcode 3.0: improved data processing and mining tools. *Nucleic Acids Res* **42**, D938-943, doi:10.1093/nar/gkt1204 (2014).

51 Zilliox, M. J. & Irizarry, R. A. A gene expression bar code for microarray data. *Nat Methods* **4**, 911-913, doi:10.1038/nmeth1102 (2007).

52 Rizvi, A. H. *et al.* Single-cell topological RNA-seq analysis reveals insights into cellular differentiation and development. *Nat Biotechnol* **35**, 551-560, doi:10.1038/nbt.3854 (2017).

53 Reimand. Thread 2: Network models. *Nature Genetics* **45**, doi:10.1038/ng.2787 (2013).

54 De Sousa, P. A. *et al.* Clinically failed eggs as a source of normal human embryo stem cells. *Stem Cell Res* **2**, 188-197, doi:10.1016/j.scr.2009.01.002 (2009).

55 Cowan, C. A. *et al.* Derivation of embryonic stem-cell lines from human blastocysts. *N Engl J Med* **350**, 1353-1356, doi:10.1056/NEJMsr040330 (2004).

56 Cahan, P. *et al.* CellNet: network biology applied to stem cell engineering. *Cell* **158**, 903-915, doi:10.1016/j.cell.2014.07.020 (2014).

57 Huang, K., Maruyama, T. & Fan, G. The naive state of human pluripotent stem cells: a synthesis of stem cell and preimplantation embryo transcriptome analyses. *Cell stem cell* **15**, 410-415, doi:10.1016/j.stem.2014.09.014 (2014).

58 Wang, B. *et al.* Similarity network fusion for aggregating data types on a genomic scale. *Nat Methods* **11**, 333-337, doi:10.1038/nmeth.2810 (2014).

59 Szalay-Beko, M. *et al.* ModuLand plug-in for Cytoscape: determination of hierarchical layers of overlapping network modules and community centrality. *Bioinformatics (Oxford, England)* **28**, 2202-2204, doi:10.1093/bioinformatics/bts352 (2012).

60 Shaw, L., Sneddon, S. F., Zeef, L., Kimber, S. J. & Brison, D. R. Global gene expression profiling of individual human oocytes and embryos demonstrates heterogeneity in early development. *PLoS One* **8**, e64192, doi:10.1371/journal.pone.0064192 (2013).

61 Bloor, D. J., Metcalfe, A. D., Rutherford, A., Brison, D. R. & Kimber, S. J. Expression of cell adhesion molecules during human preimplantation embryo development. *Molecular human reproduction* **8**, 237-245 (2002).

62 Shaw, L., Sneddon, S. F., Brison, D. R. & Kimber, S. J. Comparison of gene expression in fresh and frozen-thawed human preimplantation embryos. *Reproduction (Cambridge, England)* **144**, 569-582, doi:10.1530/REP-12-0047 (2012).

63 Brady, G. & Iscove, N. N. Construction of cDNA libraries from single cells. *Methods in enzymology* **225**, 611-623 (1993).

64 Al-Taher, A., Bashein, A., Nolan, T., Hollingsworth, M. & Brady, G. Global cDNA amplification combined with real-time RT-PCR: accurate quantification of multiple human potassium channel genes at the single cell level. *Yeast (Chichester, England)* **17**, 201-210 (2000).

65 Iscove, N. N. *et al.* Representation is faithfully preserved in global cDNA amplified exponentially from sub-picogram quantities of mRNA. *Nat.Biotechnol.* **20**, 940-943 (2002).

66 McCall, M. N., Bolstad, B. M. & Irizarry, R. A. Frozen robust multiarray analysis (fRMA). *Biostatistics (Oxford, England)* **11**, 242-253, doi:10.1093/biostatistics/kxp059 (2010).

67 Team, R. C. (Foundation for Statistical Computing, Vienna, Austria, 2014).

68 Camarasa, M. V. *et al.* Derivation of Man-1 and Man-2 research grade human embryonic stem cell lines. *In Vitro Cell Dev Biol Anim* **46**, 386-394, doi:10.1007/s11626-010-9291-5 (2010).

69 Oldershaw, R. A. *et al.* Directed differentiation of human embryonic stem cells toward chondrocytes. *Nat Biotechnol* **28**, 1187-1194, doi:10.1038/nbt.1683 (2010).

70 Rohart, F., Gautier, B., Singh, A. & Le Cao, K.-A. mixOmics: an R package for 'omics feature selection and multiple data integration. *bioRxiv*, doi:10.1101/108597 (2017).

71 Su, G., Morris, J. H., Demchak, B. & Bader, G. D. Biological network exploration with Cytoscape 3. *Current protocols in bioinformatics* **47**, 8.13.11-24, doi:10.1002/0471250953.bi0813s47 (2014).

72 Chatr-Aryamontri, A. *et al.* The BioGRID interaction database: 2015 update. *Nucleic Acids Res* **43**, D470-478, doi:10.1093/nar/gku1204 (2015).

73 Yu, H., Kim, P. M., Sprecher, E., Trifonov, V. & Gerstein, M. The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. *PLoS Comput Biol* **3**, e59, doi:10.1371/journal.pcbi.0030059 (2007).

74 Komurov, K. & White, M. Revealing static and dynamic modular architecture of the eukaryotic protein interaction network. *Mol Syst Biol* **3**, 110, doi:10.1038/msb4100149 (2007).

75 Assenov, Y., Ramirez, F., Schelhorn, S. E., Lengauer, T. & Albrecht, M. Computing topological parameters of biological networks. *Bioinformatics (Oxford, England)* **24**, 282-284, doi:10.1093/bioinformatics/btm554 (2008).

76 RCoreTeam. R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria* **https://www.R-project.org/** (2016).

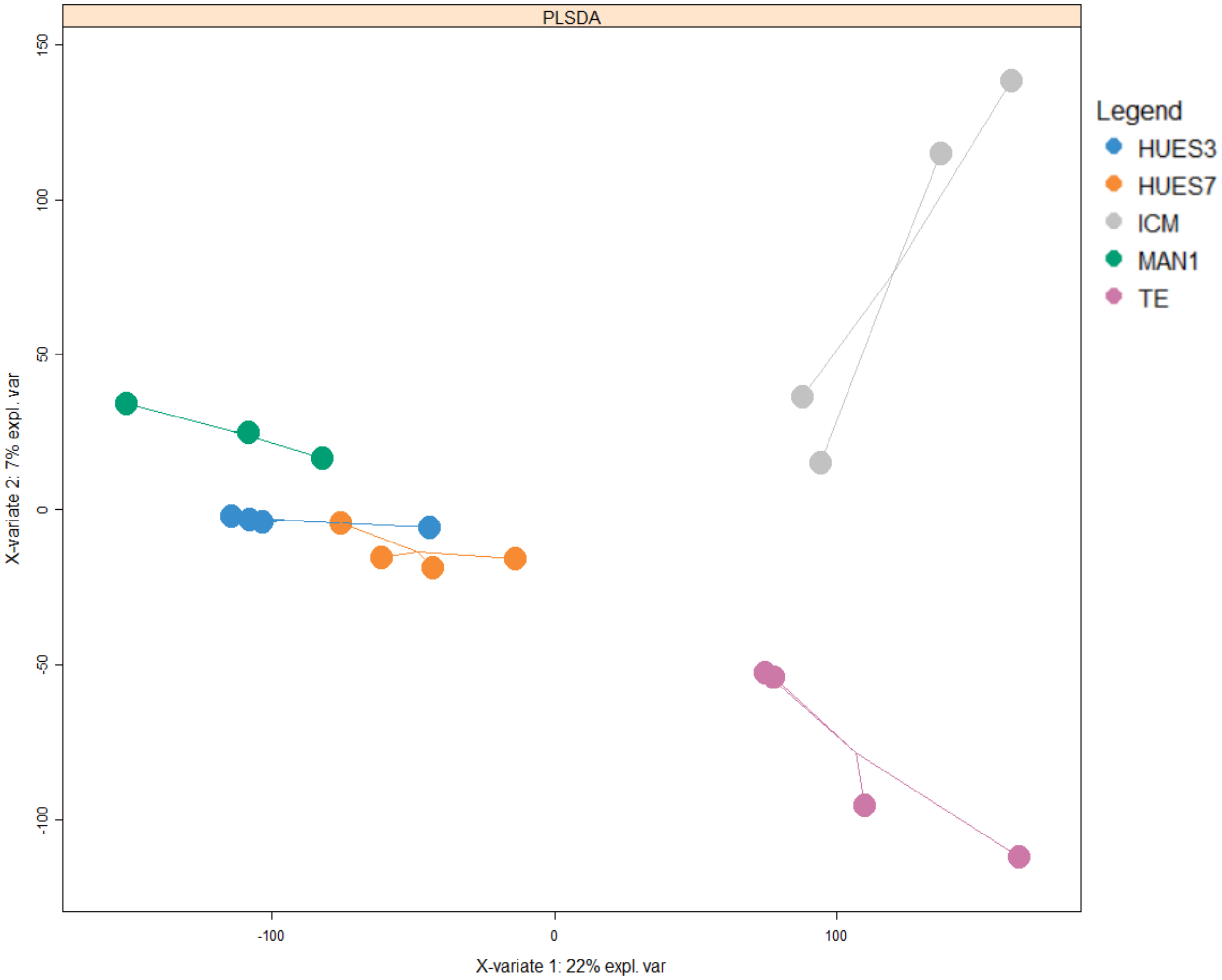


Figure 1. Distance between the transcriptomes of Inner cell mass, trophectoderm and human embryonic cell lines as a measure of similarity.

Gene expression over the entire transcriptome (54613 gene probesets) was defined using the gene barcode approach as a z-score in comparison to a database of 63331 examples of hgu133plus2.0. The Euclidean distances between samples were assessed using partial least square discriminant analysis (PLSDA).

Two components are used (X-variate 1 & 2) and the amount of explained variance is listed on the axis. The star plot shows sample distance from the centroid, the arithmetic mean position of all the points in each group.

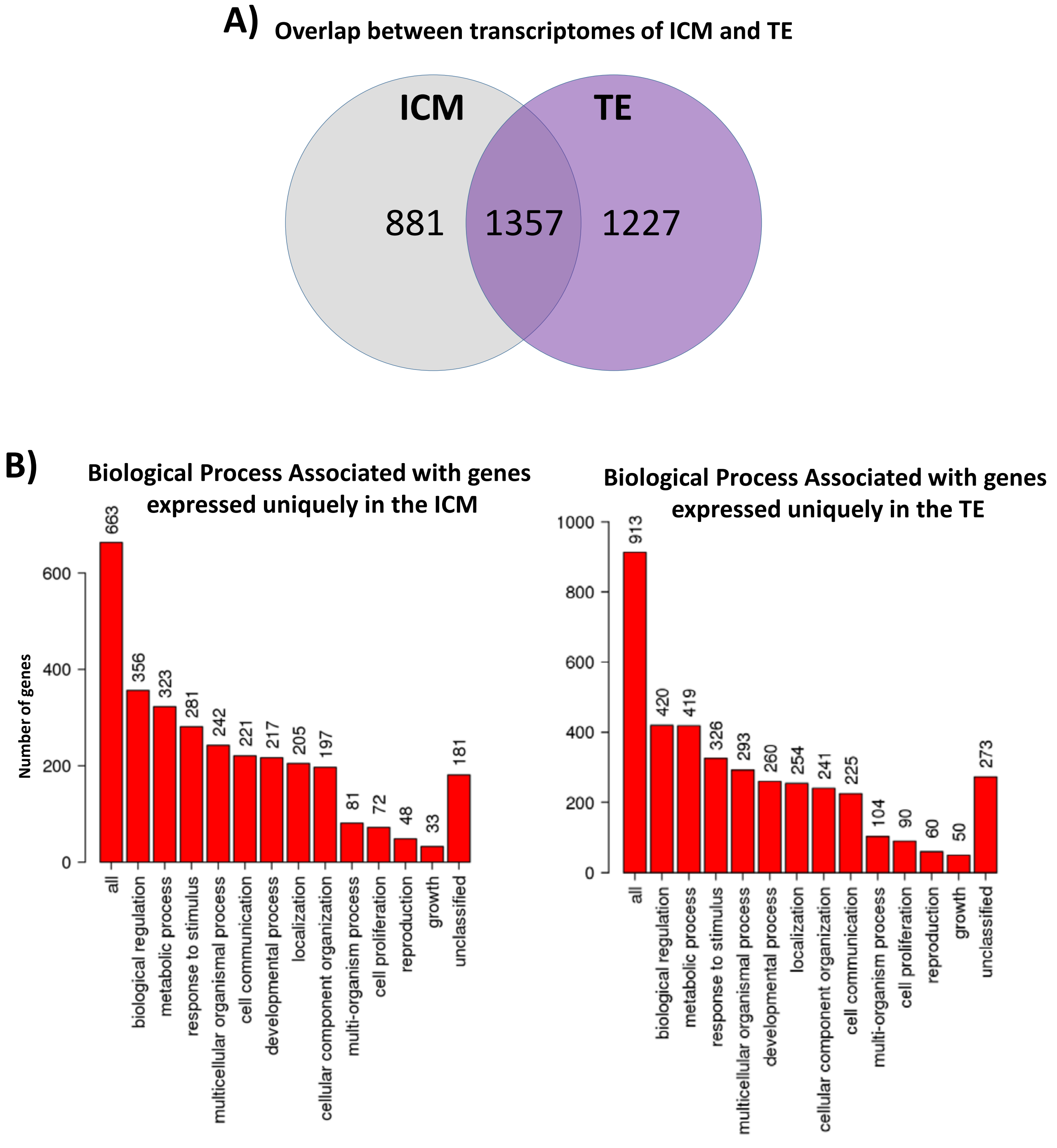


Figure 2. Inner cell mass and trophectoderm specific transcriptome and associated gene ontology

Gene expression over the entire transcriptome was assigned as present or absent using the gene barcode approach, present was defined as a z-score ≥ 5.0 for a gene probeset in comparison to a database of 63331 examples of hgu133plus2.0. This resulted in a set of 2238 gene probesets in ICM and 2484 gene probesets in TE. **A)** A Venn diagram showing the overlap and unique expression of gene probesets in the ICM and TE. **B)** Biological process gene ontology (GO Slim) for 663/719 genes used from 881 gene probesets uniquely mapped to the ICM and 913/924 genes used from 1227 gene probesets uniquely mapped to TE.

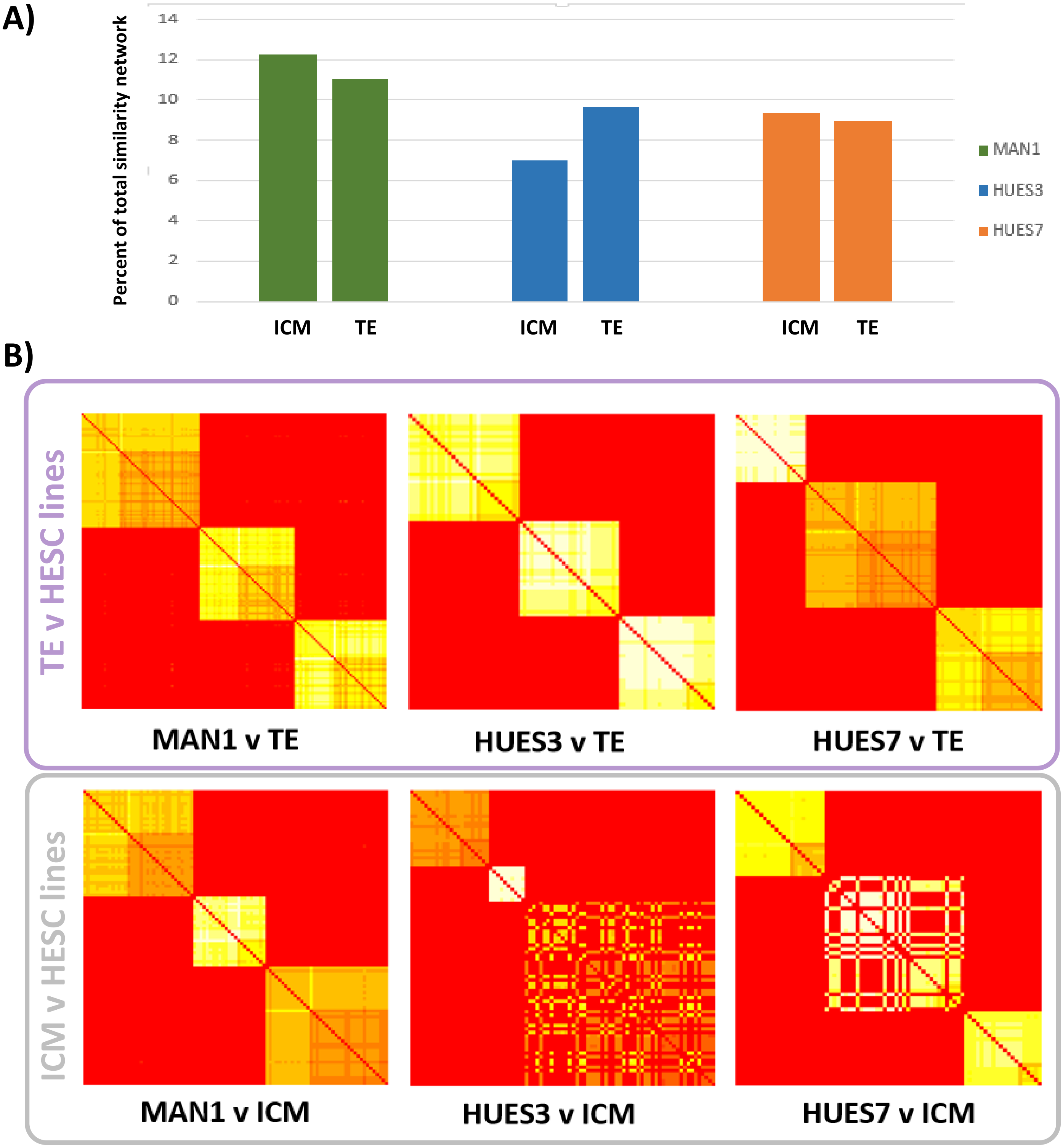


Figure 3. Similarity network fusion to compare homology between the transcriptome of inner cell mass and trophectoderm and human embryonic stem cell lines.

Similarity network fusion matrix showing similarity groups between the uniquely expressed ICM and TE gene probesets and the human embryonic stem cell line (square matrix of gene probesets with leading diagonal showing equivalence mapped to red). Similarity is coloured by intensity from white to yellow, red is dissimilar. Groups of genes with similar expression patterns across both comparisons appear as yellow, whilst those with dissimilar patterns of expression within or between cell lines appear red. Clusters therefore represent genes whose expression patterns are similar to one another both within and between input datasets. Similarity measures not only distance between ICM and the human embryonic stem cell lines but also coherency based on 15 nearest neighbours. **A)** Proportion of gene probesets in ICM or TE that are similar to human embryonic cell line transcriptome (**Supplementary Figure S1**). **B)** Similarity groups between ICM or TE and the human embryonic stem cell lines forming three clusters. Coherency in gene expression patterns with nearest neighbours is indicated by uniform yellow intensity.

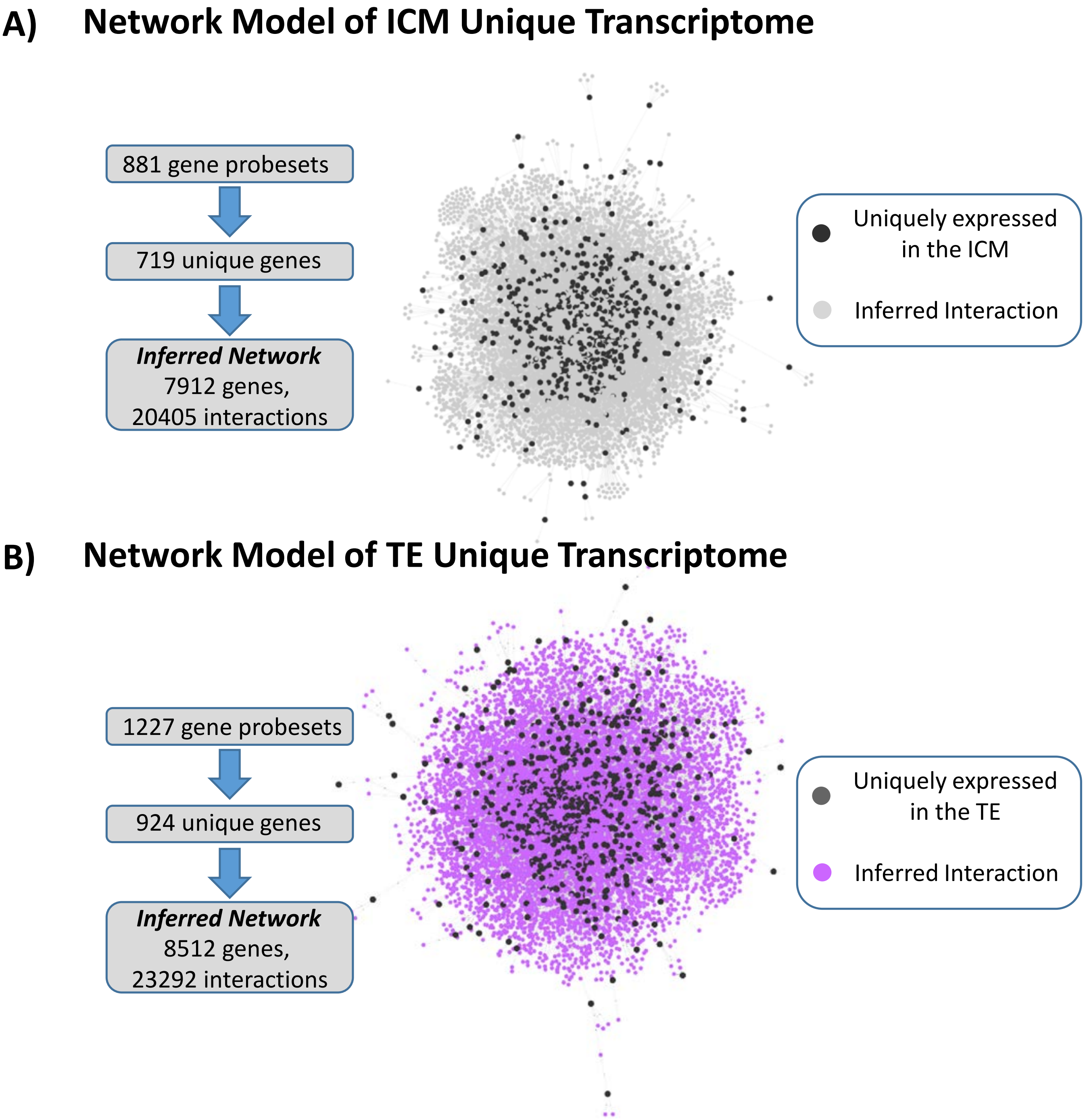


Figure 4. Interactome network models of gene expression unique to ICM or TE.

A) Interactome network model of the 719 genes (881 gene probesets) uniquely expressed in ICM.
B) Interactome network model of the 924 genes (1227 gene probesets) uniquely expressed in TE.
These were used to infer interactome network models using the BioGRID database version 3.4.158.

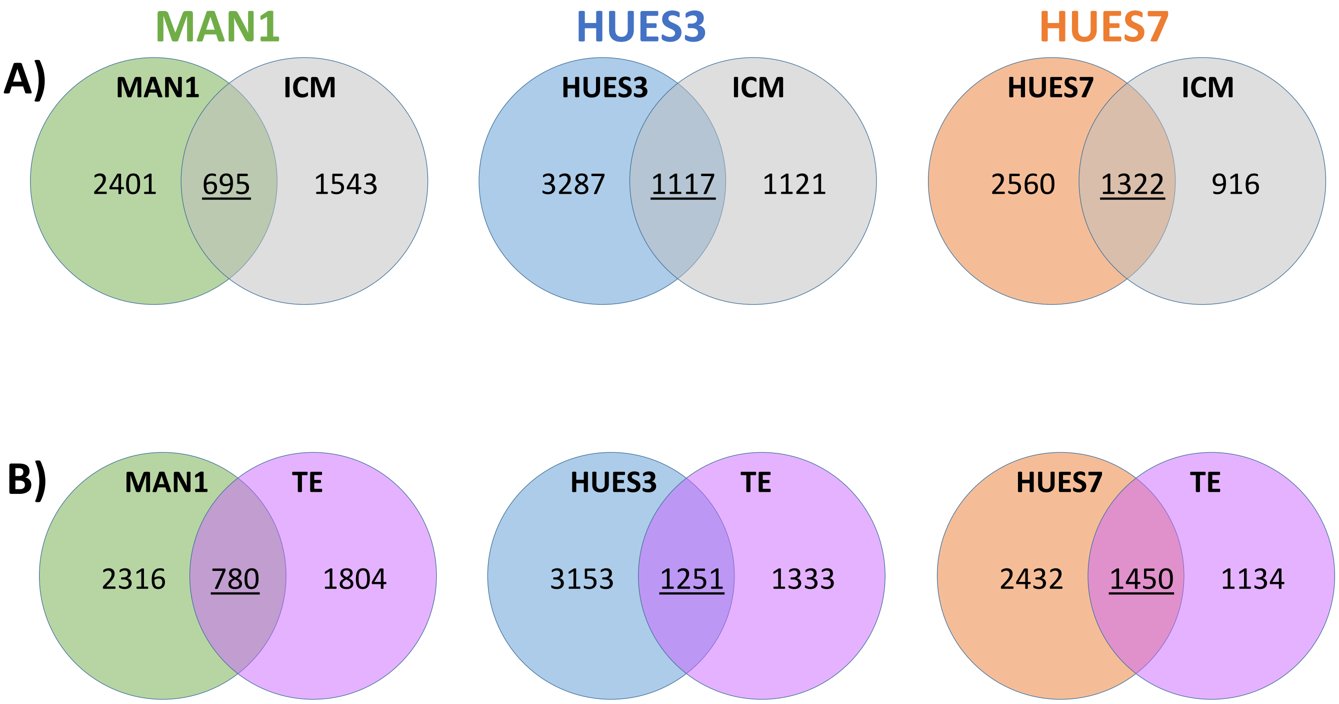


Figure 5. The interactome network models of gene expression unique to ICM or TE can be used as a framework to assess similarity with human embryonic stem cells.

Gene probesets were identified that had shared expression between human embryonic stem cell lines and the ICM (**A**) and in TE (**B**). The gene probe sets that were expressed in ICM or TE and the human embryonic stem cell lines were mapped to the ICM or TE interactome network models for the MAN1, HUES3 and HUES7 human embryonic stem cell lines. In ICM the overlaps were as follow: MAN1 cell line 695 (22%) gene probesets (517 genes), in HUES3 1117 (25%) gene probesets (856 genes) and in HUES7 1322 (34%) gene probesets (1010 genes). In TE the overlaps were as follows: MAN1 cell line 780 (30%) gene probesets (593 genes), in HUES3 1251 (48%) gene probesets (964 genes) and in HUES7 1450 (56%) gene probesets (1108 genes).

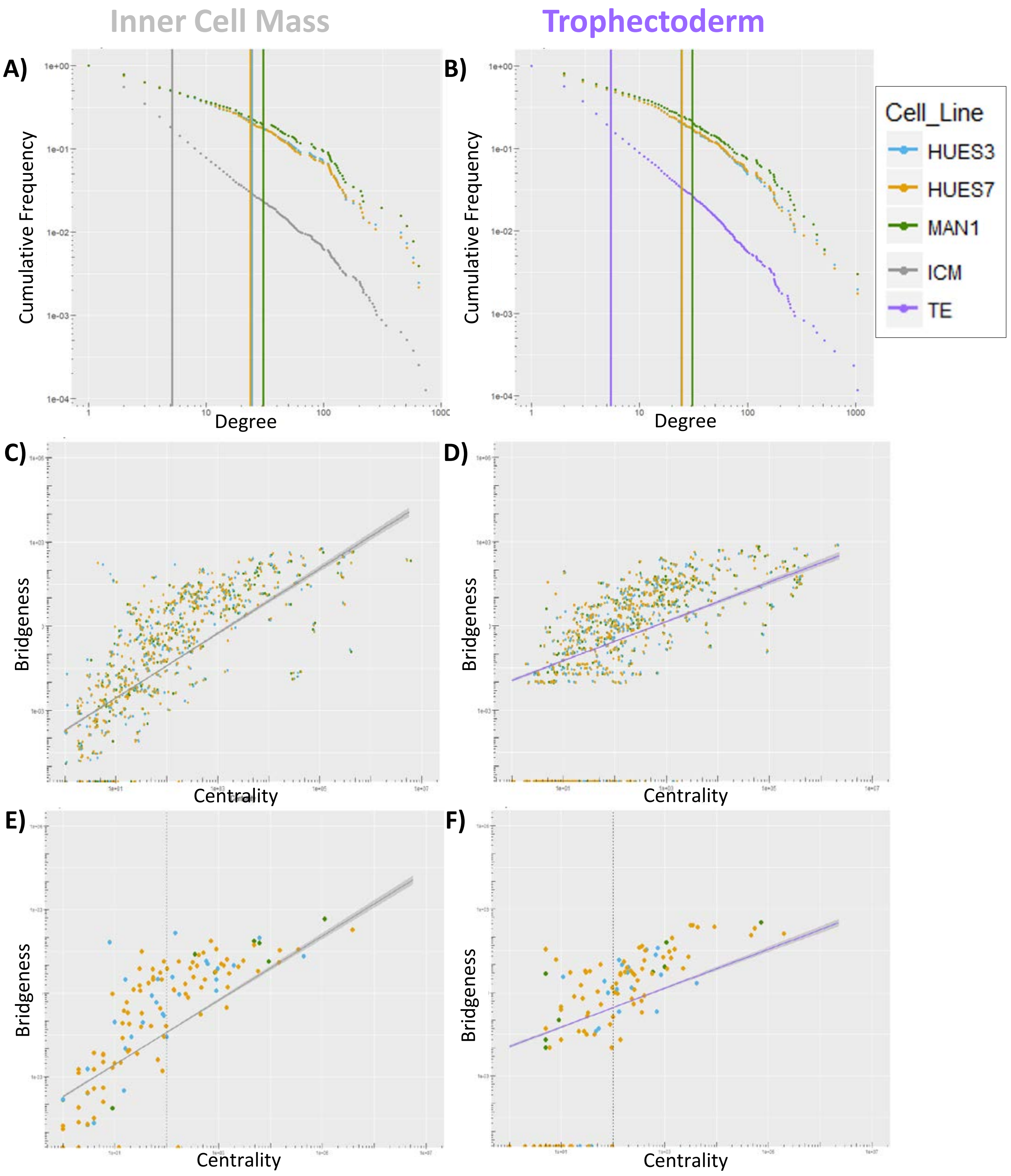


Figure 6. The network topology of the ICM and TE interactome is enriched in human embryonic stem cells.

A) ICM interactome connectivity and **B)** TE interactome network connectivity as measured by the degree (connectivity) of each gene within the network model (x-axis) plotted against the frequency of that connectivity within the network (y-axis). **C)** ICM interactome and **D)** TE interactome centrality score (x-axis), a network property that measures the influence of a node, plotted against bridgeness (y-axis), a network property measuring the bridge-like role of genes between network modules. The line with 95% confidence intervals shaded represents the centrality and bridgeness values over the entire network, genes shared with the human embryonic stem cells are marked. **E)** ICM interactome and **F)** TE interactome centrality versus bridgeness shown for genes uniquely expressed in each human embryonic stem cell line. Dotted vertical line placed at centrality value of 100 separates two perceived trajectories in the data.

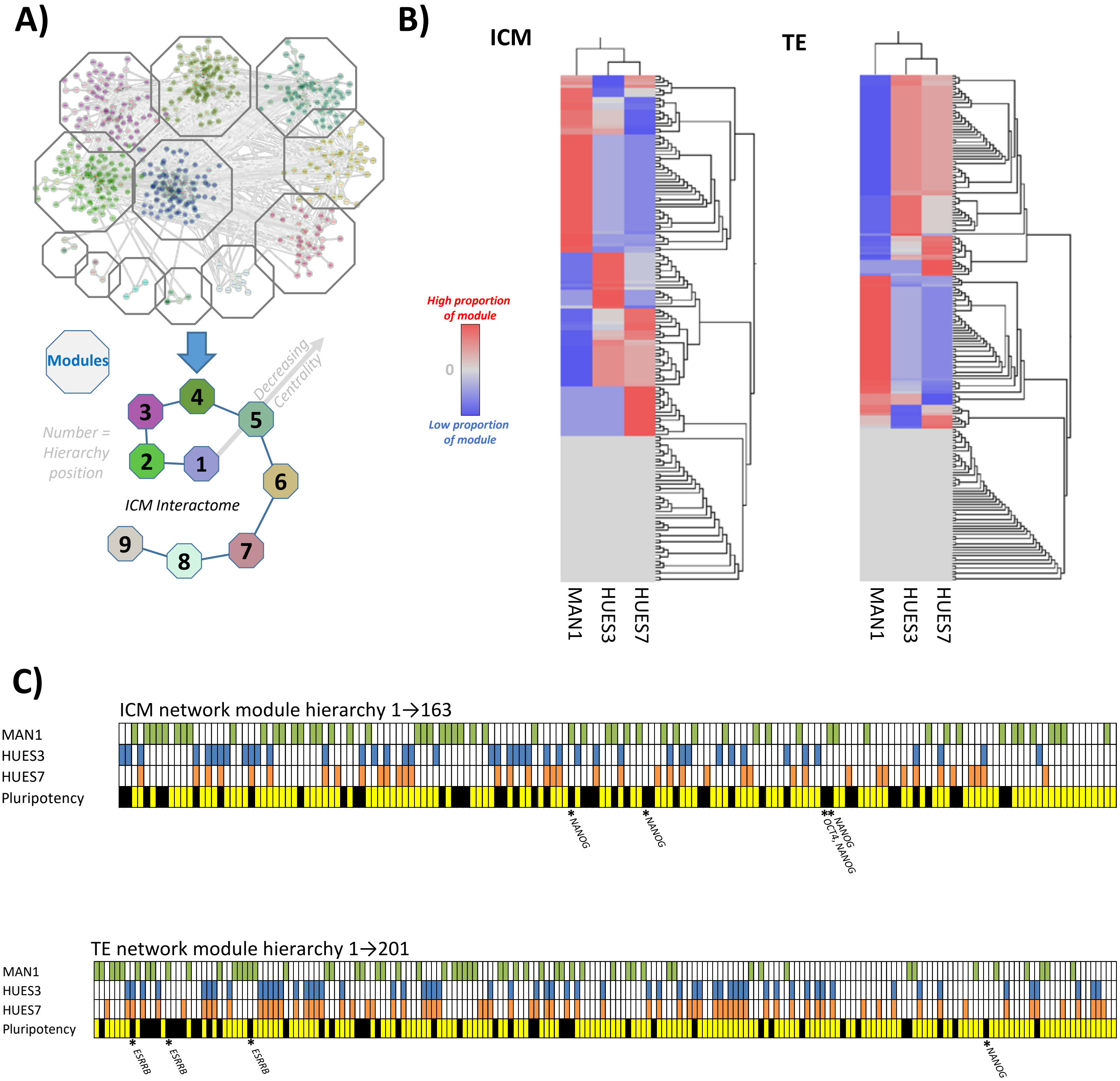
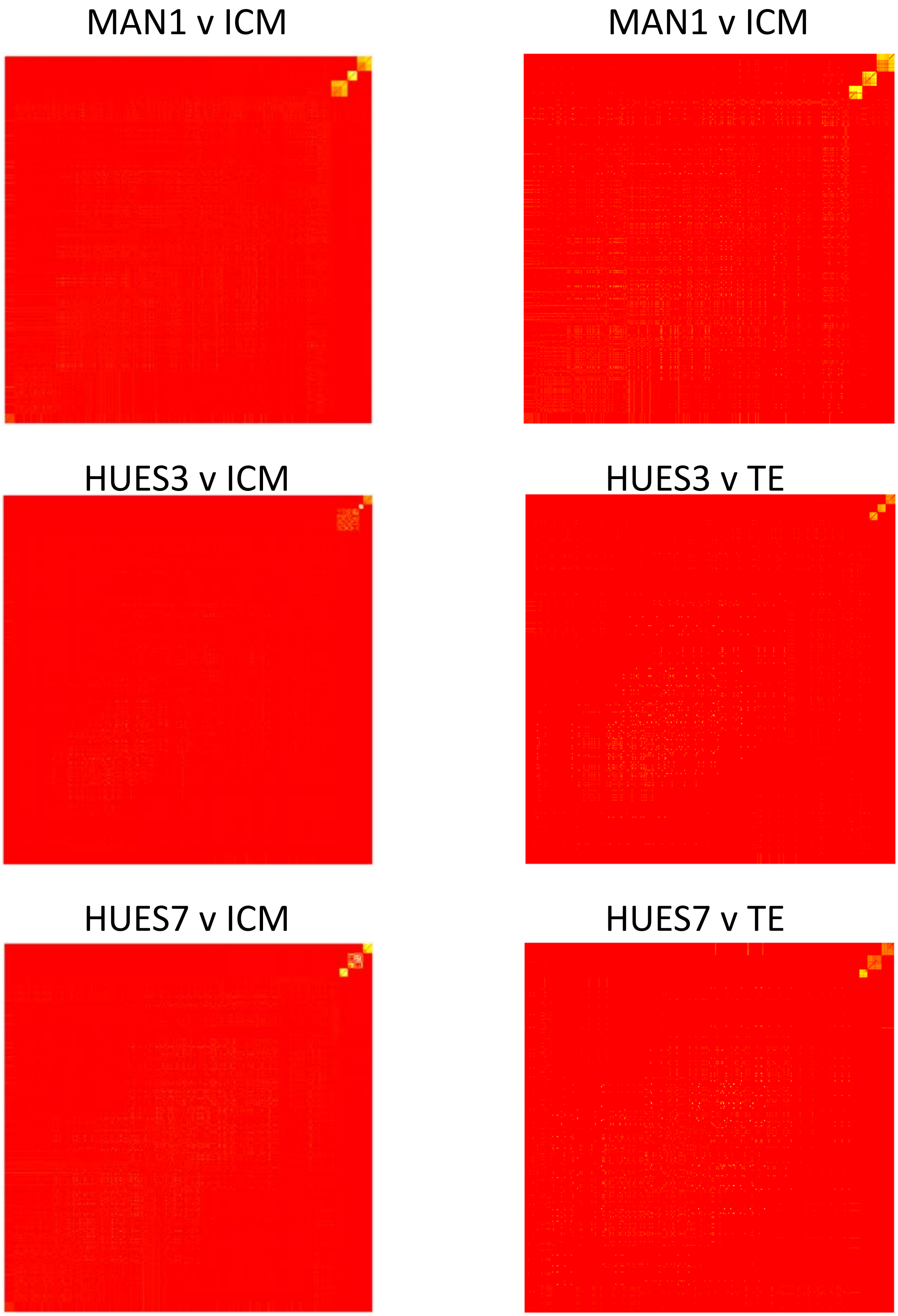


Figure 7. The modular structure of the interactome network model of gene expression unique to ICM and TE can be used as a framework to assess similarity with human embryonic stem cells.

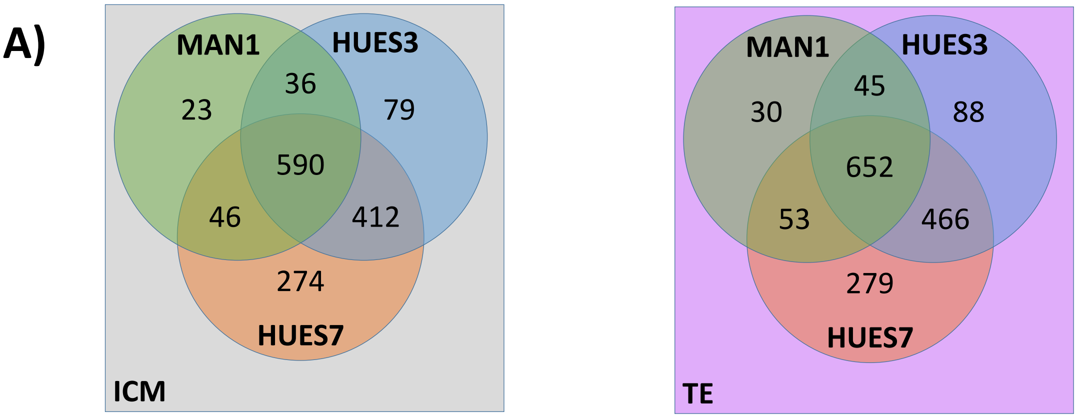
A) The modular structure of the ICM interactome was defined using the Moduland algorithm to assess the presence of highly connected gene modules. These were then formed into a hierarchy based on their centrality score, a measurement of network topology related to the influence of a network element on the rest of the network. **B)** The proportion of each module shared with the human embryonic stem cell lines was defined and clusters of modules with similar shared gene expression were assessed using a heatmap. **C)** The clusters of modules with similar proportions shared with specific human embryonic stem cell lines is represented in hierarchical order. Clusters are coloured to mark for which human embryonic cell line they are enriched. Pluripotency track represents which modules contain known pluripotency associated genes in black. An asterisk is used to mark where NANOG, OCT4 and ESRRB are situated in the modular hierarchy.

Supplementary Figures



Supplemental Figure S1. Full similarity network fusion to compare homology between the transcriptome of inner cell mass and trophectoderm and human embryonic stem cell lines.

Similarity network fusion matrix showing similarity groups between the uniquely expressed ICM gene probesets from both ICM and the human embryonic stem cell line (square matrix of gene probesets with leading diagonal showing equivalence mapped to red). Similarity is coloured by intensity from white to yellow, red is dissimilar. The proportion of genes which are similar between a hESC line and either ICM or TE can be determined by the proportion of either axis which contains yellow signal.

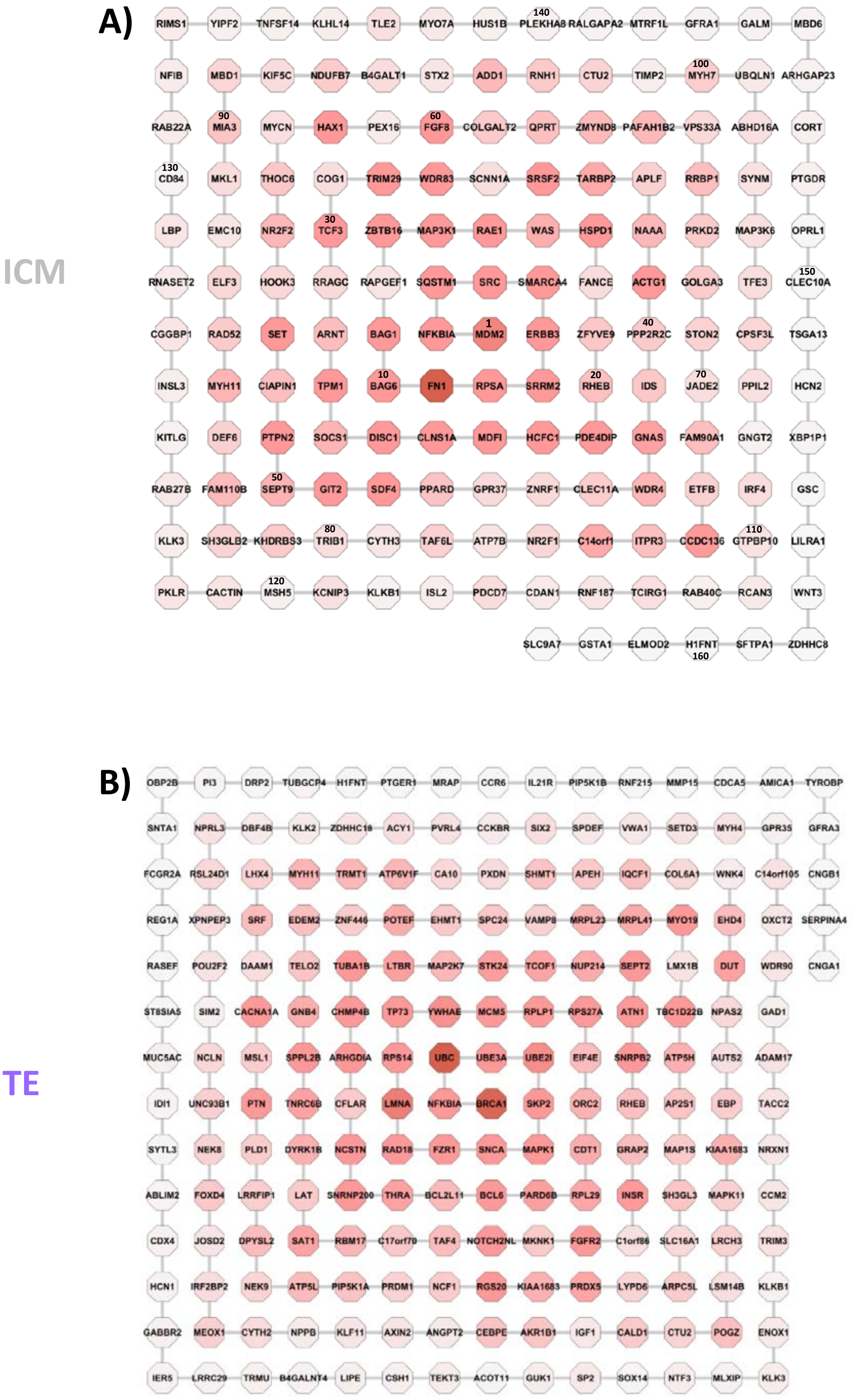


B)

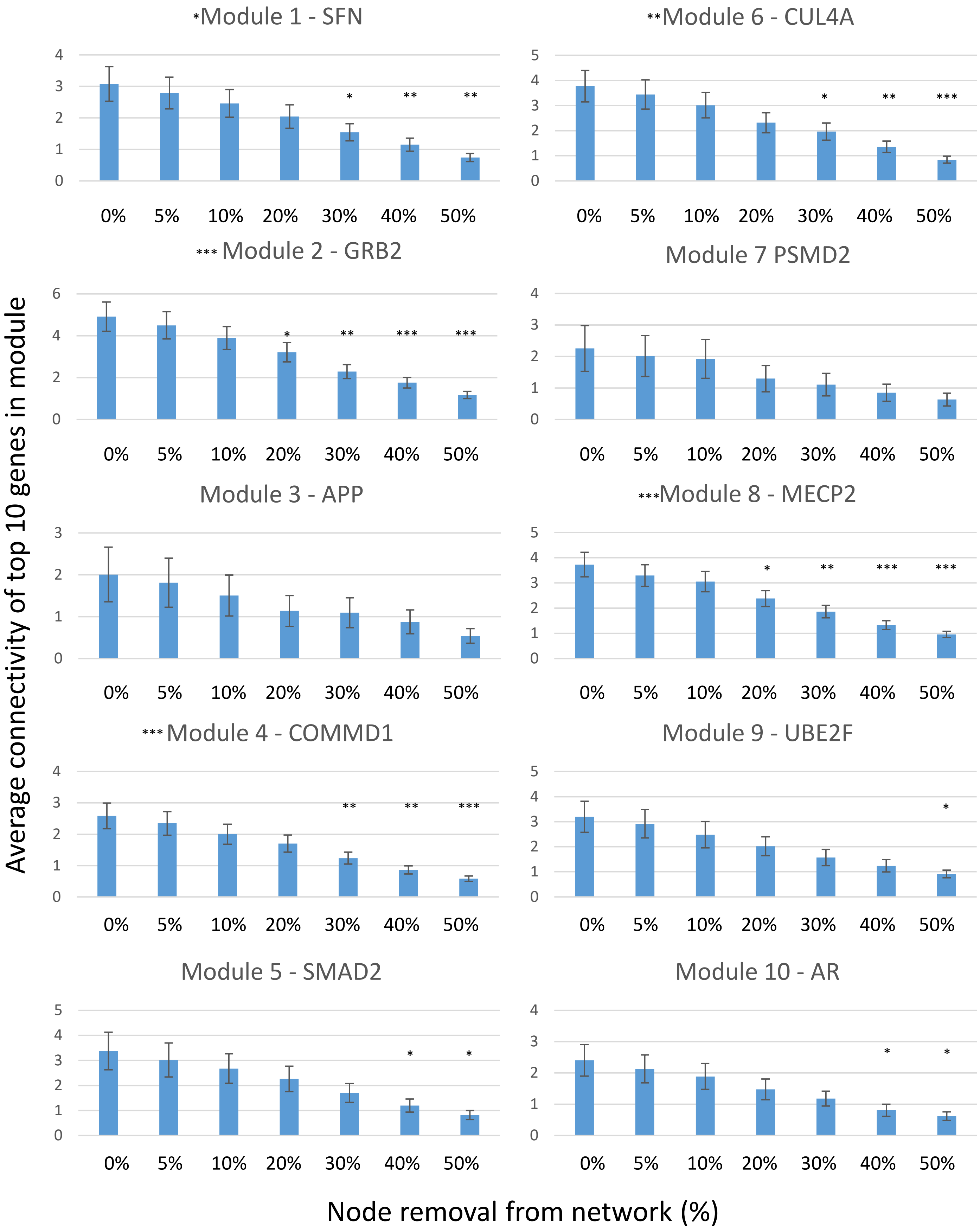
Canonical Pathway	HUES7 TE	HUES3 TE	MAN1 TE	HUES7 ICM	HUES3 ICM	MAN1 ICM
Intrinsic Prothrombin Activation Pathway	🟥	🟥			🟥	
Spermine and Spermidine Degradation I					🟥	
Role of Pattern Recognition Receptors in Recognition of Bacteria and Viruses		🟥			🟥	
Dolichyl-diphosphooligosaccharide Biosynthesis					🟥	
Differential Regulation of Cytokine Production by IL-17A and IL-17F					🟥	
Catecholamine Biosynthesis		🟥				
Dermatan Sulfate Degradation (Metazoa)		🟥				
Chondroitin Sulfate Degradation (Metazoa)		🟥				
PDGF Signaling						🟥
Cell Cycle Control of Chromosomal Replication						🟥
ERK5 Signaling			🟥			
Eicosanoid Signaling	🟥		🟥	🟥		
Myc Mediated Apoptosis Signaling	🟥		🟥			
Cell Cycle: G2/M DNA Damage Checkpoint Regulation	🟥		🟥			
Notch Signaling				🟥		
Gustation Pathway				🟥		
FXR/RXR Activation	🟥			🟥		
Parkinson's Signaling	🟥			🟥		
Glucocorticoid Receptor Signaling	🟥			🟥		
RhoGDI Signaling	🟥			🟥		
Glycerol-3-phosphate Shuttle	🟥			🟥		
Glutamate Receptor Signaling	🟥			🟥		
Gαs Signaling	🟥			🟥		
eNOS Signaling	🟥			🟥		
IL-17A Signaling in Gastric Cells	🟥			🟥		
Sperm Motility	🟥			🟥		
Signaling by Rho Family GTPases	🟥			🟥		
Heparan Sulfate Biosynthesis (Late Stages)	🟥			🟥		
Heparan Sulfate Biosynthesis	🟥			🟥		
Gα12/13 Signaling	🟥			🟥		
Dermatan Sulfate Biosynthesis (Late Stages)	🟥			🟥		
Chondroitin Sulfate Biosynthesis (Late Stages)	🟥			🟥		
Dermatan Sulfate Biosynthesis	🟥			🟥		
Chondroitin Sulfate Biosynthesis	🟥			🟥		

Supplemental Figure S2. Expressed genes uniquely shared between each human embryonic stem cell line and either the Inner Cell Mass (ICM) or the Trophectoderm (TE).

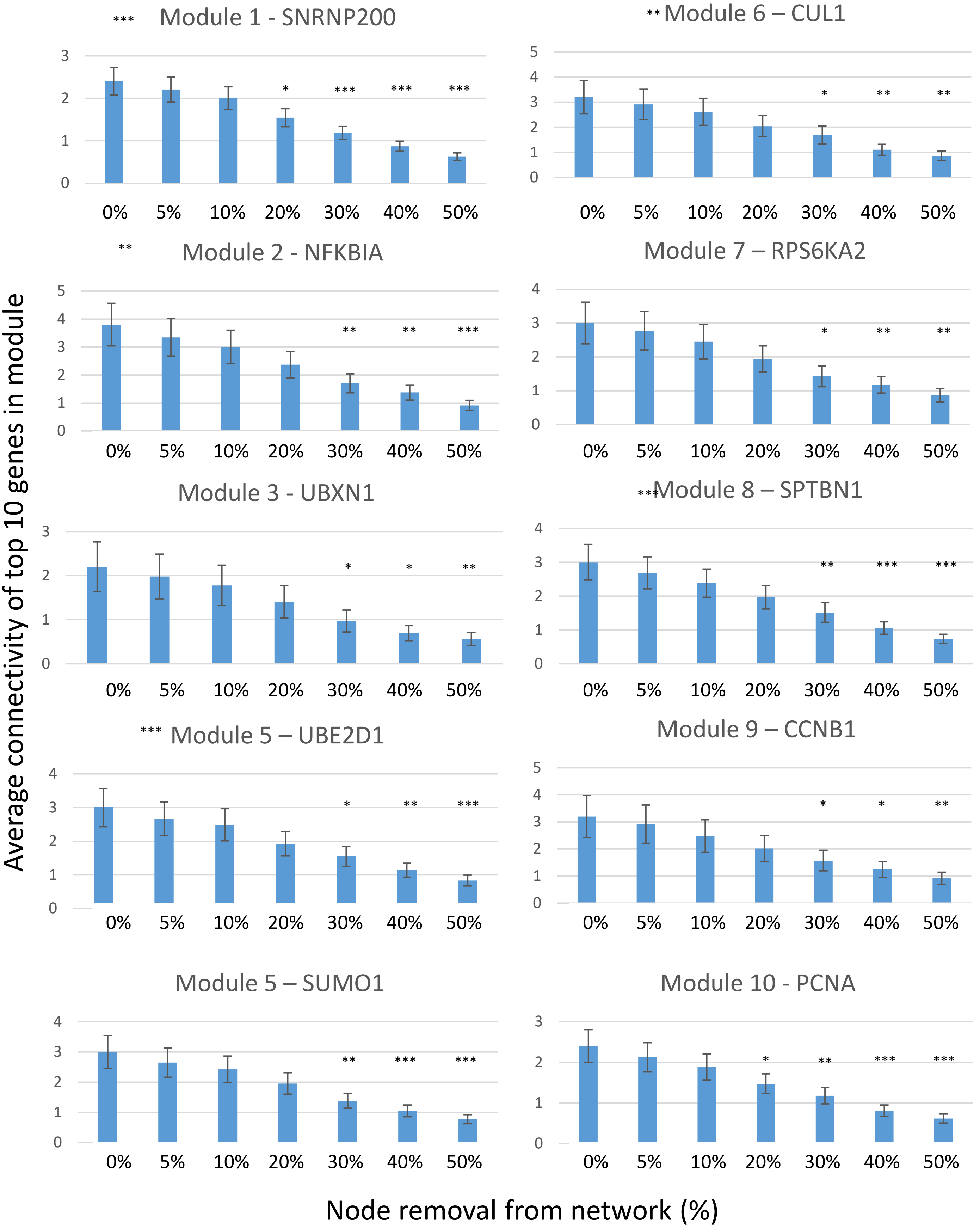
A) Overlap of the gene expression (gene probe sets) shared between the human embryonic stem cell lines and ICM or TE. **B)** Biological pathways associated with the gene expression uniquely shared between each human embryonic stem cell line and either ICM or TE. Intensity of red shade is proportional to p-value of right sided Fisher's Exact test.



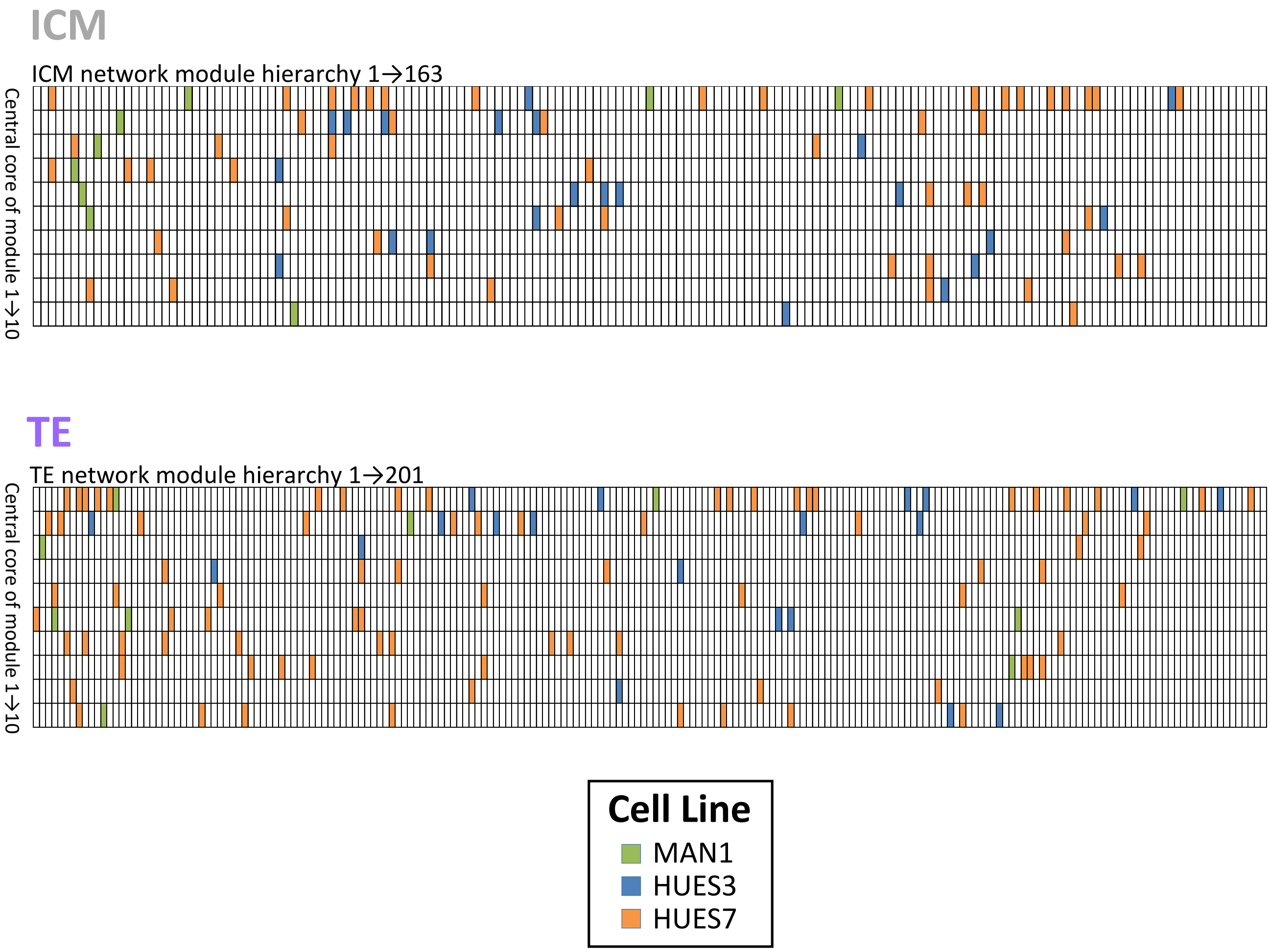
Supplemental Figure S3. Hierarchy of modules within the interactome network models of ICMN and TE.
A) The modules of the ICM and **B)** the TE interactome network represented as octagons named with the most central gene. Modules are arranged in a hierarchy represented as a spiral with numbers defining the position in the hierarchy. Modules are shaded red in relation to connectivity to highlight the relationship between network connectivity and centrality.



Supplemental Figure S4.a Robustness of 10 most central network modules of an ICM network. Robustness was determined by the mean change in connectivity between the 10 most connected nodes in each network module upon the removal of random nodes from the network. Up to 50% of nodes were removed before recalculating connectivity, iterated 100 times. Significance for each module was determined using ANOVAs whilst between samples t-tests determined significant differences from 0% node loss in each case. Modules whose mean connectivity was not significantly reduced at 20% node removal can be described as robust. [* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$].



Supplemental Figure S4.b Robustness of 10 most central network modules of a TE network. Robustness was determined by the mean change in connectivity between the 10 most connected nodes in each network module upon the removal of random nodes from the network. Up to 50% of nodes were removed before recalculating connectivity, iterated 100 times. Significance for each module was determined using ANOVAs whilst between samples t-tests determined significant differences from 0% node loss in each case. Modules whose mean connectivity was not significantly reduced at 20% node removal can be described as robust. [* p < 0.05; ** p < 0.01; *** p < 0.001].



Supplemental Figure S5. Gene expression uniquely present in each of the human embryonic stem cell lines mapped to the central core of each of the modules in the ICM and TE interactome network models The core of each module (listed horizontally) was defined as the most central ten genes (vertical columns). The overlap of these core genes with the unique gene expression shared with the human embryonic stem cell lines and either ICM or TE is shown coloured by cell line.