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], Maria Keramari[2], 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: 6871

Figures: 6

Tables: 0

Abstract (193 words)

Human embryonic stem cells (hESCs) derived from the pluripotent Inner cell mass (ICM) of the blastocyst are fundamental tools for understanding human development, yet are not identical to their tissue of origin. To investigate this divergence we compared the transcriptomes of genetically paired ICM and trophectoderm (TE) samples with three hESC lines: MAN1, HUES3 and HUES7 at similar passage. We generated inferred interactome networks unique to the ICM or TE using transcriptomic data, and defined a hierarchy of modules (highly connected regions with shared function). We compared network properties and the modular hierarchy and show that the three hESC lines had limited overlap with the ICM specific transcriptome (6%-12%). However, this overlap was enriched for network properties related to transcriptional activity in ICM ($p=0.016$); greatest in MAN1 compared to HUES3 ($p=0.048$) or HUES7 ($p=0.012$). The hierarchy of modules in the ICM interactome contained a greater proportion of MAN1 specific gene expression (46%) compared to HUES3 (28%) and HUES7 (25%) ($p=9.0\times10^{-4}$).

These findings show that traditional methods based on transcriptome overlap are not sufficient to identify divergence of hESCs from ICM. Our approach also provides a valuable approach to the quantification of differences between 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 in a system can be estimated by 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 (Boyer, et al., 2005). 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 (Cauffman, et al., 2009; Kimber, et al., 2008). This heterogeneity has been confirmed by RNAseq analysis of single human preimplantation epiblast cells (Petropoulos, et al., 2016). 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)(Fogarty, et al., 2017). 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 (Han, et al., 2010; Nakamura, et al., 2016; Weinberger, et al., 2016). At the same time, pluripotency–associated transcriptional networks continue to be expressed in the preimplantation human epiblast (Niakan and Eggan, 2013; Petropoulos, et al., 2016) and early post-implantation cynomolgus epiblast (Nakamura, et al., 2016). 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 ESC lines.

Expression of a number of genes has been associated with the development of extraembryonic cell lineages including *Tead4* (Nishioka, et al., 2008), *Tsfap2c* (Kuckenberg, et al., 2012), *Gata3* (Home, et al., 2009) and *Cdx2* (Strumpf, et al., 2005). But there is evidence suggesting divergence between species in the utilisation of some of these genes such as the Gata family (Grabarek, et al., 2012; Rossant, et al., 2003; Schrode, et al., 2013; Stephenson, et al., 2012) known to play a role in TE

26 generation (Nakamura, et al., 2016). These observations imply that networks of interacting co-
27 regulated proteins might distinguish the transiently pluripotent ICM/preimplantation epiblast from
28 the early differentiated trophectoderm (TE) in a species-specific manner.

29 In mouse the ground state pluripotency of the ICM appears to be maintained in murine ESCs derived
30 from the ICM and cultured in the presence of LIF together with MEK and GSK3β inhibitors
31 (Weinberger, et al., 2016). This is not the case for human ESCs derived from day 6-7 blastocysts and
32 cultured in standard medium with TGFβ family molecules and FGF-2. It is established in the literature
33 that human ESC lines have more similarities to the murine epiblast after implantation (Faial, et al.,
34 2015; Tesar, et al., 2007) than to the murine ICM and ESCs. In order to understand this difference, it
35 is important to determine how similar hESCs are to the *human* ICM.

36 Transcriptional analysis of isolated ICM and TE samples from individual human embryos has also been
37 performed, highlighting key metabolic and signalling pathways (Adjaye, et al., 2005). A recent study
38 of 1529 individual cells from 88 human preimplantation embryos defined a transcriptional atlas of this
39 stage of human development (Petropoulos, et al., 2016), however cell lineage allocation can be
40 problematic and inter-individual heterogeneity has been shown to have a major effect on gene
41 expression (Smith, et al., 2019; Stirparo, et al., 2018). Together these data show the relevance of
42 transcriptome based analysis and highlight the need for approaches that account for inter-individual
43 variation.

44 Recently the heterogeneity present in available human blastocyst single cell RNAseq data has been
45 commented on and sample preparation methods have been questioned (Stirparo, et al., 2018). In the
46 work presented here we have set out to examine how far the gene expression profiles of ICM and TE
47 have diverged from one another at the blastocyst stage, when hESC derivation occurs, and to compare
48 these data to the transcriptome of hESCs using sets of transcriptomic data independent of preparation
49 method. We have defined paired transcriptomic data sets unique to the ICM and TE from the same
50 human embryo. Combined with accepted lists of genes that have differential expression between ICM

51 and TE defined by meta-analysis, we generated ICM-and TE-specific interactome network models. This
52 approach has allowed us to use quantitative network analysis to compare both TE and ICM with hESCs
53 and to evaluate the extent of similarity between ICM/TE and hESC cell lines as well as the hESC lines
54 with each other. These analyses provide an important framework which highlights the development
55 origins of hESCs.

56

57 ## Results

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

60 We used the significant transcriptomic differences between ICM and TE (512 genes) identified by
61 Stirparo *et al* in their meta-analysis of human blastocyst single cell RNAseq data (Stirparo, et al., 2018)
62 to map the relationship of our stem cell transcriptomic data (**Figure 1A**). These data demonstrated
63 that the MAN1, HUES3 and HUES7 transcriptomes identified using frozen RMA (McCall, 2015) were
64 similar to those hESCs previously examined by Yan et al (Yan, et al., 2013) and were in the direction of
65 the NANOG eigenvector. We also observed heterogeneity in the blastocyst single cell RNAseq from
66 Petropoulos *et al* as previously indicated by Stirparo *et al* (Stirparo, et al., 2018).

67 Frozen RMA barcode Z scores (McCall, 2015; McCall, et al., 2010) for the entire transcriptome
68 (n=54613 gene probe sets) were compared using partial least squares discriminant analysis (PLSDA)
69 to assess the relationship between ICM, TE and the hESC lines MAN1, HUES3, HUES7 (**Figure 1B**). The
70 hESC sample groups were distinct from each other and from ICM and TE ($p<0.05$). All three hESC cell
71 lines were of equivalent distance from both ICM and TE along the X-axis (X-variate 1), however along
72 the Y-axis (X-variate 2) MAN1 was closer to ICM than HUES3 or HUES7. Similar results were shown
73 with PCA (data not shown).

74

75

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

77 Frozen RMA gene barcode was used to isolate gene probe sets present in each embryonic tissue
78 resulting in 2238 probe sets in ICM and 2484 probe sets in TE. These data were used to determine the
79 overlap and unique gene expression in each of these blastocyst tissues (**Figure 2A**). We found 881 and
80 1227 gene probe sets uniquely expressed in the ICM and TE respectively, corresponding to 719 and
81 924 unique genes (**Supplemental Table S1**). The genes defined as having unique expression in ICM or
82 TE significantly overlapped with single cell RNA-seq data from human epiblast and trophectoderm
83 cells respectively (both $p<1.0x10^{-4}$), identified in previously published analysis (Petropoulos, et al.,
84 2016). Recognising the potential heterogeneity of samples within the Petropoulos data highlighted by
85 Stirparo et al we used the genes identified by frozen RMA as unique to ICM and TE in combination to
86 categorise the available single cell RNAseq blastocyst data. This analysis resulted in almost perfect
87 classification of the single cell RNAseq datasets from Yan *et al* (Yan, et al., 2013) and Blakely *et al*
88 (Blakeley, et al., 2015) and, as previously shown by Stirparo *et al* (Stirparo, et al., 2018), highlighted
89 the heterogeneity within the Petropoulos data (Petropoulos, et al., 2016) (**Figure 2B**). The stem cell
90 transcriptomic data generated by frozen RMA was no longer proximal to the stem cell data generated
91 by Yan *et al* (Yan, et al., 2013) but had moved further along the eigenvector implying ICM classification
92 (**Figure 2B**).

93 The genes associated with ICM and TE were grouped by "biological process" ontology showing a
94 similar proportion and ordering in both gene sets, the only difference being a reduction in the
95 proportion of genes of the category "cell communication" in the TE compared to the ICM (**Figure 2C**).
96 More detailed comparison of biological pathways identified "epithelial adherens junction signalling"
97 (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
98 initiation activity (TE $p=4.4x10^{-6}$, ICM $p =0.39$) as significantly associated with TE, consistent with the
99 TE being at an early stage of diverging differentiation towards trophectoderm epithelium (Marikawa
100 and Alarcon, 2012), with an active requirement for new biosynthesis (Hasegawa, et al., 2015)
101 (**Supplemental Table S2**).

102 It was noted that NANOG regulation was strongly associated with the ICM ($p=5.9x10^{-6}$) but not the TE
103 and that CDX2 regulation was associated with TE ($p= 9.8x10^{-3}$) but not ICM, as would be anticipated
104 (Niakan and Eggan, 2013). Using causal network analysis we identified master regulators of gene
105 expression associated with the transcriptomic data. This approach identified MYC ($p=7.6x10^{-8}$), a co-
106 ordinator of OCT4 activity (Fang, et al., 2016), and ONECUT1 (HNF6) ($p=4.0x10^{-8}$), a regulator of the
107 development of epithelial cells (Pierreux, et al., 2006), as the most significantly associated regulatory
108 factors in ICM and TE respectively (**Supplemental Tables S3**).
109
110 **Similarities between the inner cell mass and trophectoderm unique transcriptomes and the**
111 **transcriptome of human embryonic stem cells**
112 Firstly we used the 512 genes, defined by meta-analysis (Stirparo, et al., 2018), as differentially
113 expressed between ICM and TE to quantify correlations with gene expression within the
114 transcriptomes of the hESC lines using hypernetwork analysis. These data highlighted MAN1 as having
115 quantifiably more correlations (1.8 fold $p< 1x10^{-5}$) compared to HUES3 or HUES7 with gene expression
116 that is associated with the differentiation of ICM and TE (**Figure 3A**). The rank order of the stem cell
117 lines was MAN1 >> HUES3 > HUES7 as indicated by the number of co-expressed genes in the
118 interactome (increased proportion of yellow in the heatmap - **Figure 3A**).
119 Similarity Network Fusion (SNF) was used to assess the similarity of gene expression patterns between
120 cell lines and ICM or TE. SNF uses nearest neighbour component to its algorithm to identify regions
121 where this pattern is *coherent*. A region of coherency across a stem cell line and either TE or ICM
122 represents a group of genes whose expression pattern is conserved between embryonic tissue and
123 hESCs. The analysis highlighted a limited similarity of hESC lines with ICM (between 6% and 12%
124 similarity) and TE (between 9% and 11%), consistent with the distance between the hESC lines and TE
125 and ICM as observed by PLSDA analysis (**Figure 3B & Supplemental Figure S1**). Three primary clusters
126 of similarity were identified in all comparisons between the hESC lines and ICM or TE (**Figure 3C**). These
127 clusters were of equivalent similarity in TE with all hESC lines, as indicated by uniform yellow intensity

128 indicating coherency with nearest co-expressed neighbours, implying highly co-ordinated expression.
129 However, when ICM was compared with hESCs, coherency was noted only with MAN1 and not with
130 the other hESC lines (**Figure 3C**).
131
132 **An interactome network model of gene expression unique to ICM can be used as a framework to**
133 **assess similarity with human embryonic stem cells.**
134 An interactome network model can be used to consider the proteins derived from differentially
135 expressed genes and the proteins that they interact with. Using this approach allowed us to consider
136 the wider biological interactions generated by the gene expression unique to either the ICM or TE and
137 to implement these models as a framework to assess similarity with the hESC lines.
138 We used the genes with differential expression between ICM and TE as defined by Stirparo *et al*
139 (Stirparo, et al., 2018) to generate ICM and TE specific network models by using the genes with positive
140 fold change in expression in each specific tissue (337 for ICM and 175 for TE) as a basis for network
141 inference. We also separately used the genes with unique expression in either ICM or TE (**Figure 2A**)
142 to generate interactome network models by inference to known protein-protein interactions
143 (**Supplemental Figure S2A & 2B**).
144 As interactome networks account for inferred interactions these may be shared between different
145 network models. Comparing the TE and ICM interactome network models an overlap of 2517 and 5659
146 inferred genes was present in the Stirparo models and the models based on our de novo data
147 respectively. These overlaps represent protein:protein interactions, accounting for 85% (Stirparo) and
148 72% (our model) of the ICM interactome along with 66% (Stirparo) and 30% (our model) of the TE
149 interactome.
150 We examined further the network models based on the uniquely expressed genes in ICM and TE
151 identified by frozen RMA. Both networks were enriched for genes associated with pluripotency, for
152 example NANOG with the ICM network and CDX2 within the TE network, as identified by gene
153 ontology analysis. The ICM network contained 93/167 and 161/240 genes and the TE network

154 contained 94/167 and 185/240 genes related to core pluripotency associated factors by RNAi (Ding,
155 et al., 2009; Hu, et al., 2009; Ivanova, et al., 2006; Ng and Lufkin, 2011; Zhang, et al., 2006) and protein
156 interaction (Liang, et al., 2008; Ng and Lufkin, 2011; Pardo, et al., 2010; van den Berg, et al., 2010;
157 Wang, et al., 2006) screens respectively. The similarity of TE with ICM networks for pluripotency
158 factors is likely to reflect the fact that this tissue has only very recently begun to diverge.
159 Using uniquely expressed genes derived from our de novo transcriptomic analysis we were able to
160 determine the shared transcriptome between ICM or TE and each human embryonic stem cell line
161 and map these onto the respective ICM or TE interactome network model. Of the genes shared
162 between the hESC lines and ICM there was no difference between the proportions each line shared
163 with the network model (p=0.74), for the genes shared between the hESC lines and TE, MAN1 had a
164 significantly smaller proportion of genes shared with the TE network model (p= 0.03).
165
166 **Similarities and differences in topology between human embryonic stem cell lines in relation to**
167 **inner cell mass and trophectoderm network models**
168 As the ICM and TE interactome models shared a significant proportion of the same genes, we went on
169 to assess the network topology of these models to determine further similarities and differences with
170 the genes shared with the hESC lines. Analysis of the network topology of the ICM and TE interactome
171 demonstrated that the genes shared with the hESC lines were enriched for highly connected genes (as
172 measured by degree, the number of interactions made to other genes). We found that HUES3 and
173 MAN1 were more connected than HUES7 in the ICM network (MAN1vsHUES7 p=0.04, HUES3vsHUES7
174 p=0.04, MAN1vsHUES3 p=0.94) but not in the TE network (MAN1vsHUES7 p=0.21, HUES3vsHUES7
175 p=0.28, MAN1vsHUES3 p=0.89) (**Figure 4A & 4B**).
176 To further investigate the putative functional relevance of genes shared between the ICM or TE
177 interactome models and the hESC lines we determined whether these genes had “party” or “date”
178 like properties. In protein interaction networks party hubs co-ordinate local activity by protein
179 complexes, whereas date hubs regulate global effects and are assumed to represent the transient

180 interactions that occur with transcription factors (Agarwal, et al., 2010; Chang, et al., 2013). Date-like
181 network hubs have been shown to possess a higher "bridgeness" property at any position within the
182 interactome (Kovacs, et al., 2010). Bridgeness is a network property that measures overlap between
183 network modules and this score can be compared at different positions within the network by plotting
184 it against "centrality", a network property that measures the influence of a node in a network (Kovacs,
185 et al., 2010). HUES3 and MAN1 were shown to have a greater proportion of date-like hubs than HUES7
186 in either the ICM or TE network models based on the Stirparo data, demonstrating an increased
187 number of genes with network properties of transcription factors. (**Figure 4C & 4D**). This observation
188 implies an enrichment for date-like network hubs in the genes shared between the hESC lines and the
189 ICM or TE interactome network models, implying in turn an enrichment of transcription factor activity.
190 Using the interactome models derived from de novo transcriptomic analysis the network topology of
191 the ICM and TE demonstrated that the genes shared with the hESC lines were enriched for highly
192 connected genes (as measured by degree, the number of interactions made to other genes) and the
193 enrichment seen was not statistically different between the hESC lines (**Figure 5A & 5B**).
194 All three hESC lines were shown to be enriched for bridgeness score in relation to centrality when
195 compared to the full ICM or TE networks based on de novo transcriptomic data (**Figure 5C & 5D**). We
196 identified the overlap of genes expressed in the ICM or TE and the hESC cell lines (**Supplemental Figure**
197 **S2**). There were 590 and 652 shared genes between all the three hESC lines and ICM or TE respectively
198 (**Supplemental Figure S3A & S3B**). When we examined genes uniquely expressed in each of the hESC
199 lines (**Supplemental Figure S3A & S3B**), the highly central genes in both networks (centrality score
200 >100) were significantly enriched for bridgeness in ICM (p=0.016) but not TE (p=0.105), indicating
201 more date-like properties in ICM (**Figure 5E & 5F**). In the ICM interactome network model MAN1 was
202 significantly more date-like than HUES3 (p=0.048) and HUES7 (p=0.012). This observation implies that
203 the MAN1 cell line shared significantly more transcription factor activity with ICM which is
204 hierarchically more important within the ICM interactome, than do either HUES3 or HUES7. Biological
205 pathways associated with genes uniquely expressed in each of the hESC lines are shown in

206 **Supplemental Figure S3B**. In MAN1 "PDGF signalling" and "cell cycle control of chromosome
207 replication" were associated with the unique gene expression shared with ICM. PDGF signalling is
208 required for primitive endoderm cell survival in the inner cell mass of the mouse blastocyst (Artus, et
209 al., 2013) and the pluripotency associated transcription factor NANOG (referred to above) has been
210 shown to influence replication timing in the cell cycle (Apostolou, et al., 2013; Hiratani, et al., 2010).
211

212 **Modular hierarchy of the ICM and TE interactome network models reveal an enrichment in MAN1**
213 **for ICM and an enrichment in HUES7 for TE**

214 Network modules are sub-structures of a network that have a greater number of internal connections
215 than expected by chance. Modules are known to represent functionally related elements of a network
216 and can be ranked hierarchically by their centrality within a network, with the assumption that the
217 more central modules are functionally dominant within the network. We defined modules within our
218 TE and ICM interactome network models allowing for overlap and arranged these into a hierarchy of
219 influence by centrality score (Kovacs, et al., 2010) (**Figure 6A**). The ICM and TE interactome network
220 models had a hierarchy of 109 and 163 along with 71 and 201 modules of different sizes in the models
221 based on the Stirparo data and the de novo data respectively. There was no difference in the
222 proportion of modules compared to network size between the ICM and TE interactome network
223 models ($p<0.2$) using either the Stirparo data or our own data (**Supplemental Figure S4 &**
224 **Supplemental Tables S4 & S5**). The robustness of the definition of network modules in the ICM and
225 TE interactome network models based on our de novo transcriptomic data was confirmed by
226 permutation analysis of the proportional random removal of genes (Reimand, 2013) (**Supplemental**
227 **Figure S5**). This established that the majority of modules were robust to the removal of large
228 proportions of the network, with only 2 of the top 47 ICM and 8 of the top 49 TE modules analysed
229 experiencing a significant ($p<0.05$) reduction in connectivity within the module following the removal
230 of a random 20% of the network iterated 100 times.

231 Network module hierarchy in inner cell mass and trophectoderm network models based on the
232 Stirparo data was assessed and the proportion of each module that also mapped to genes within the
233 transcriptome of the human embryonic stem cell lines. This analysis again showed that MAN1 had a
234 greater number of unique genes represented in both TE (MAN1vsHUES3 p=0.004, MAN1vsHUES7
235 $p=2.95x10^{-6}$) and ICM (MAN1vsHUES3 p=0.026, MAN1vsHUES7 $p=2.95x10^{-8}$) networks, particularly in
236 the most central modules and to a greater extent in ICM than TE (**Figure 6B**).
237 The genes with shared expression between ICM or TE networks based on the de novo transcriptomic
238 data and the hESC lines were mapped to each interactome module. In the ICM network 116/163
239 modules (71%) were still enriched for gene expression shared between hESC lines and ICM. A greater
240 proportion of hESC associated modules in the ICM interactome network model were enriched for
241 MAN1 gene expression (0.46) compared to HUES3 (0.28) and HUES7 (0.25) ($p=9.0x10^{-4}$, chi squared
242 test). In the TE interactome network model 132/201 modules (65%) were enriched for gene expression
243 shared between hESC lines and TE. The smallest proportion of enriched hESC associated modules
244 occurred in HUES7 (0.17) compared to MAN1 (0.39) and HUES3 (0.44) ($p=3.1x10^{-6}$, chi squared test)
245 (**Figure 6C**).
246 The modules assessed as having enriched gene expression in specific hESC lines were mapped to the
247 module hierarchy in the ICM or TE interactome network model based on the de novo transcriptomic
248 data (**Figure 6D**). These data show an enrichment of the modules that have the greatest proportion of
249 shared gene expression with MAN1 in the upper part of the module hierarchy in both ICM and TE
250 indicating that the MAN1 associated modules were likely to be more functionally active in both the
251 ICM and TE interactomes.
252 Gene expression uniquely present in each of the hESC lines (**Supplemental Figure S2**) was mapped to
253 the central core (most central 10 genes) of each of the modules in the ICM and TE interactome
254 network models (**Supplemental Figure S3**). This analysis highlighted only gene expression present
255 uniquely in MAN1 or HUES7 in the upper part of the module hierarchy in the ICM and TE interactome
256 network models indicating that HUES3 associated modules had a reduced presence in the function of

257 the ICM. The upper part of the TE network model module hierarchy was enriched for both HUES7 and
258 MAN1 uniquely expressed genes, indicating a dominant effect of these hESC lines on TE function,
259 compared to HUES3.
260 Finally, relating these analyses to the enrichment for pluripotency associated genes we defined in the
261 ICM and TE interactome models, we examined this relationship to the modular hierarchy of the ICM
262 and TE interactome network models. We assessed whether any of the pluripotent genes mapped to
263 the central core of 10 genes in a network module (coloured black in **Figure 6D**). In the ICM modular
264 hierarchy 16, 13 and 11 of the modules enriched in MAN1, HUES3 and HUES7 respectively also
265 mapped to pluripotency genes. In the TE modular hierarchy 18, 11 and 15 of the modules enriched in
266 MAN1, HUES3 and HUES7 respectively also mapped to pluripotency genes. It was noted that OCT4
267 (*POU5F1*), a primary marker of ICM (Hochedlinger and Jaenisch, 2015), was present in the central core
268 of the modules from the ICM but not the TE network models. *NANOG*, another marker of ICM
269 (Hochedlinger and Jaenisch, 2015), was present four times in the ICM and only once in the TE network
270 models. Also estrogen-related-receptor beta (*ESRRB)*, a marker of TE (Latos, et al., 2015; Nicola, et al.,
271 2018), was present three times in the TE but not at all in the ICM network models. In the ICM network
272 model, 2 of the 3 NANOG associated modules are enriched for MAN1 gene expression and the module
273 associated with both NANOG and OCT4 had equivalent enrichment in MAN1, HUES3 and HUES7. In
274 the TE network model the NANOG associated module was low in the hierarchy (76/201) and had
275 equivalent enrichment in MAN1, HUES3 and HUES7. In the TE network model the three ESRRB
276 associated modules were at the upper end of the module hierarchy with the highest ranked (8/201)
277 being enriched in HUES3 and HUES7 and the other two being associated with MAN1 (**Figure 5C**). These
278 data combined show that the key transcription factors (and partners) known to be associated with
279 ICM and TE have biologically logical but different associations with hESC lines within the modular
280 hierarchies of the interactome network models.

281

282 **Discussion**

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

286 The MAN1 human embryonic stem cell line was furthest from both ICM and TE using distance metrics
287 on the unsupervised transcriptome. Only ~10% of genes uniquely expressed by the ICM (compared to
288 TE) were shown to have similarity to expression patterns in MAN1, HUES3 and HUES7 using SNF.
289 However MAN1 was found to be most similar to ICM as it had both a greater enrichment of genes and
290 a greater coherency with nearest neighbours in comparison to HUES3 and HUES7. Substantial
291 enrichment of human embryonic stem cell line gene expression was also observed in relation to TE
292 but, whilst this was shown to be coherent with nearest neighbours, MAN1 and HUES7 showed a
293 reduced similarity compared to that for ICM while HUES3 had an increased similarity to TE.

294 We used interactome network models of ICM and TE as frameworks to map overlapping gene
295 expression from MAN1, HUES3 and HUES7. Using network topology as a marker of functionality we
296 demonstrated that all the human embryonic stem cell lines had gene interaction networks with
297 increased connectivity in both the ICM and TE interactome network models generated from gene
298 expression data. All human embryonic stem cell lines also showed an enrichment for network
299 topology that was associated more with date hubs than with party hubs, in ICM and TE network
300 models. Date hubs are network positions that are associated with non-concurrent signalling and are
301 more likely to represent transcription factor activity related to the execution of a developmental
302 programme (Agarwal, et al., 2010; Chang, et al., 2013; Kovacs, et al., 2010; Ng and Lufkin, 2011). A key
303 finding of this study is that date hubs central to the network model, and therefore likely to influence
304 a greater proportion of network function, were significantly enriched in the overlap of genes uniquely
305 shared between MAN1 and the ICM compared to genes uniquely shared between HUES3 or HUES7
306 and ICM.

307 We defined a functional hierarchy of overlapping network modules in both the ICM and TE
308 interactome network models and used this as a framework to study the relationship of MAN1, HUES3
309 and HUES7 with ICM and TE gene expression. MAN1 had greater enrichment in the upper hierarchy
310 for both ICM and TE network models both overall and for uniquely expressed genes.
311 Taken together these observations demonstrate the utility of network approaches to quantify
312 underlying similarities based on the position of transcriptomic differences in an interactome network
313 model. Quantitative comparison of the hierarchy of the ICM and TE interactome network modules in
314 relation to the expressed genes in the human embryonic stem cell lines provided further insight into
315 similarities and differences between the cell lines beyond those defined by traditional distance
316 metrics.
317 An assessment of master regulators of transcription associated with the ICM and TE specific gene
318 expression identified known tissue specific transcriptional regulators – NANOG in ICM (Hochedlinger
319 and Jaenisch, 2015; Ng and Lufkin, 2011) and CDX2 in TE (Niakan and Eggan, 2013; Niwa, et al., 2005).
320 Both the ICM and TE network models were enriched for genes associated with pluripotency
321 (Hochedlinger and Jaenisch, 2015; Ng and Lufkin, 2011) an observation in alignment with recent
322 diversification of these tissues. The upper part of the hierarchy of network modules in both the ICM
323 and the TE interactome network models was enriched for pluripotency associated genes. However
324 MAN1 was more closely associated with gene modules including NANOG in the ICM interactome
325 network model compared to HUES3 and HUES7 cell lines. In the TE interactome network model HUES3
326 and HUES7 were associated with the *ESRRB* related module at the highest position in the module
327 hierarchy whilst MAN1 was also primarily associated with two further *ESSRB* related modules. ESSRβ,
328 a direct target of Nanog (Festuccia, et al., 2012), has been shown to be important in murine ES cells
329 as a co-regulator of Oct4 with Nanog (Zhang, et al., 2008) and a regulator of Gata6 though promoter
330 binding (Uranishi, et al., 2016). Using chromosome conformation capture sequencing Nanog
331 interacting modules were found to be more enriched with target sites for Esrrb as well as KLf4, Sox2
332 and cMyc target sequences with less consistency in Nanog and Oct4 target sequences (Apostolou, et

333 al., 2013). ESSRβ works with p300 to maintain pluripotency networks, generating a permissive
334 chromatin state for binding of Oct4, Nanog and Sox2 and has been implicated in reprogramming
335 epistem cells to an iPSC state (Adachi, et al., 2018).
336 Overall these data reveal that MAN1 had the greatest similarity to ICM compared to the other hESC
337 lines despite being least related to ICM in the PLSDA analysis. This observation is based on **I)** greater
338 coexpression with other tissue specific gene expression in the hypernetwork analysis, **II)** coherency in
339 the SNF analysis with nearest neighbour genes, **III)** significantly increased proportion of genes with a
340 date-like hub property in the ICM network, **IV)** an increased proportion of genes mapping to ICM
341 interactome network modules and **V)** an association with ICM network gene modules that map to
342 NANOG activity. Concordance has been identified between transcriptomic regulation in human
343 induced pluripotent stem cells and the ICM (Kilens, et al., 2018) but this has not been fully mapped at
344 the level of the interactome. We propose that the network approach presented in this manuscript
345 represents a significant advance on distance metrics in the comparison on hESC lines.
346 By using a barcode approach to define genes uniquely expressed we were able to define ICM- and TE-
347 specific interactome network models, an important advance from more traditional comparative
348 modelling using differential gene expression (McCall, 2015; McCall, et al., 2014; Zilliox and Irizarry,
349 2007). We also confirmed similarity of the underlying transcriptomic data with findings from single
350 cell RNAseq data (Petropoulos, et al., 2016) and the independent meta-analysis of that data (Stirparo,
351 et al., 2018) corroborating our observations. These comparisons also confirmed the importance of
352 network structure in the analysis we have undertaken (Rizvi, et al., 2017). We demonstrated the
353 robustness of our network models by establishing module coherency over successive reductions of
354 network model size (by gene removal), therefore establishing a high level of confidence in the analysis
355 of related gene modules and network topology (Reimand, 2013).
356 The differences between ICM and TE with all three hESC lines may partially reflect the genetic
357 background of the infertile couples donating embryos for analysis and stem cell derivation. Previously
358 we have performed re-analysis of single cell ICM and TE RNAseq from Petropoulos *et al* 2016

359 (Petropoulos, et al., 2016) and shown a strong effect of inter-individual genetic variation (Smith, et al.,
360 2019). To account for this we have restricted our analysis in this manuscript to only genetically
361 matched pairs of ICM and TE. The similarities we have established by comparison to other work
362 (Petropoulos, et al., 2016) indicate that the data presented in this manuscript is robust to inter-
363 individual differences. The greater dissimilarity of MAN1 to HUES7 and HUES3, revealed in the overlap
364 of the transcriptome to the ICM interactome network modules, may reflect differences in genetic
365 background of individual lines, or derivation regime since HUES3 and HUES7 were derived in the same
366 lab at a similar time (Cowan, et al., 2004; De Sousa, et al., 2009). However it should be noted that all
367 hESC lines were enriched for connectivity, a marker of function, within the ICM interactome, an
368 observation in agreement with a fundamental similarity between hESC lines, despite different genetic
369 background and embryo generation or hESC derivation methods (De Sousa, et al., 2009). It was also
370 noted that hESC lines are different in very many gene modules to ICM. Although the ICMs have totally
371 different genetic background to the hESC lines assessed here, the fact that the hESCs are more
372 dissimilar than the ICMs are to each other does add further weight to this conclusion.
373 Concern has been raised (Stirparo, et al., 2018) about the heterogeneity of tissue classification in the
374 single cell RNAseq data from Petropoulos *et al* (Petropoulos, et al., 2016). Our work broadly supports
375 these observations but also highlights that tissue classification can be made despite concerns in the
376 sample preparation (Stirparo, et al., 2018) or in inter individual differences (Smith, et al., 2019). This
377 observation would suggest that rigid definitions of tissue specific expression are not necessarily
378 helpful as we expand into single cell analysis. Whilst there is an inherent heterogeneity in the
379 transcriptome of the early embryo that has been defined in the work presented here.
380 The use of network approaches to quantify similarities between hESCs and their tissue of origin is a
381 developing field. Network summary approaches have been used with promising results (e.g. CellNet
382 (Cahan, et al., 2014)). Correlation networks generated from gene expression have been used to
383 generate quantitative comparison based on the analysis of discrete network modules (Huang, et al.,
384 2014). Network driven approaches can also be used to deal with the large number of comparisons

385 present in the analysis of 'omic data sets, e.g. topological data analysis (TDA) (Rizvi, et al., 2017) and
386 SNF (Wang, et al., 2014). In the work presented here we have used an efficient method to generate
387 hierarchies of overlapping gene modules (Kovacs, et al., 2010; Szalay-Beko, et al., 2012), thus
388 accounting for the underlying network topology, and supported this analysis using SNF (Wang, et al.,
389 2014) to generate quantitative comparison of hESC lines with ICM and TE. The approach we have
390 developed accounts for both the hierarchy of modules within a network and the large number of
391 comparisons performed in an unsupervised manner to generate robust conclusions. This has allowed
392 us to apply quantitative approaches to determine the similarities of three hESC lines to each other in
393 relation to ICM and TE. We have identified overall similarity of the transcriptomes and we have also
394 defined how these similarities manifest at the level of the interactome. Our findings highlight the
395 diversity inherent in the establishment of hESC lines and also present methods to quantitatively
396 compare similarity and identify key differences using a network approach.
397
398

399 ## Methods

400 **Embryos**

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

406 **Embryo sample preparation and microarray analysis of transcriptome**

407 Donated embryos were cultured in ISM-1/2 sequential media (Medicult, Jyllinge, Denmark) until
408 blastocyst formation. At embryonic day 6 the zona pellucida of the embryos were removed by brief
409 treatment with Acid Tyrode's solution pH 5.0 (Sigma-Aldrich, Gillingham, UK), and denuded
410 blastocysts were washed in ISM2 (Medicult). Blastocysts were lysed and reverse transcribed as

411 previously described (Bloor, et al., 2002; Shaw, et al., 2012) and cDNA was prepared by polyA-PCR
412 (Brady and Iscove, 1993) which amplifies all poly-adenylated RNA in a given sample, preserving the
413 relative abundance in the original sample (Al-Taher, et al., 2000; Iscove, et al., 2002). A second round
414 of amplification using EpiAmp™ (Epistem, Manchester, UK) and Biotin-16-dUTP labelling using
415 EpiLabel™ (Epistem) was performed in the Paterson Cancer Research Institute Microarray Facility. For
416 each sample, our minimum inclusion criterion was the expression of β-actin as evaluated by gene-
417 specific PCR. Labelled PolyAcRNA was hybridised to the Human Genome U133 Plus 2.0 Array
418 (HGU133plus2.0, Affymetrix, SantaClara, CA, USA) and data was initially visualised using MIAMIVICE
419 software. Quality control of microarray data was performed using principal component analysis (PCA)
420 with cross-validation undertaken using Qlucore Omics Explorer 2.3 (Qlucore, Lund, Sweden).
421 The trophectoderm (TE) and inner cell mass (ICM) of day 6 human embryos were separated by
422 immunosurgically lysing the whole TE (recovering RNA from both mural and polar TE), to leave a
423 relatively pure intact ICM. Eight microarray datasets were obtained, corresponding to 4 genetically
424 paired matched TE and ICM transcriptomes. Frozen robust multiarray averaging (fRMA) (McCall, et
425 al., 2010) was used to define absolute expression by comparison to publically available microarray
426 datasets within R (3.1.2) (Team, 2014). An expression barcode and a z-score of gene expression in
427 comparison to 63331 examples of HGU133plus2.0 was defined for each tissue (McCall, et al., 2014;
428 Zilliox and Irizarry, 2007) and used for unsupervised analysis. For analysis of gene expression specific
429 to each tissue a z-score of 5 was used to call a gene present and a barcode was assigned scoring 1 for
430 presence and 0 for absence of gene expression (McCall, 2015; McCall, et al., 2010; McCall, et al., 2014).
431 All transcriptomic data are available on the Gene Expression Omnibus (GEO) [GSE121982].

432 **hESC lines**

433 HUES7, HUES3 (kind gift of Kevin Eggan (Cowan, et al., 2004)) and MAN1 (Camarasa, et al., 2010) hESC
434 lines were cultured as previously described (Oldershaw, et al., 2010). Briefly, hESCs (p21-27) were
435 cultured and expanded on Mitomycin C inactivated mouse embryonic fibroblasts (iMEFs) in hESC
436 medium KO-DMEM (Invitrogen, Paisley, UK) with 20% knockout serum replacement (KO-SR,

437 Invitrogen), 8 ng/ml basic fibroblast growth factor (bFGF, Invitrogen), 2 mM L-glutamine, 1% NEAA
438 (both from Cambrex, Lonza Wokingham, UK), and 0.1 mM ß-mercaptoethanol (Sigma-Aldrich, Dorset,
439 UK). For feeder-free culture, cells were lifted from the iMEF layers with TrypLE (Thermo Fisher,
440 Loughborough, UK), and plated onto fibronectin-coated (Millipore) tissue culture flasks with StemPro
441 (Thermo Fisher, Loughborough, UK) feeder-free medium. After 3 passages 100 hESC cells were
442 isolated from each line (assessed separately as > 85% Oct4 positive), lysed and subjected to polyA-PCR
443 amplification, hybridisation to the microarray chip and analysis as described above.
444
445 **Analysis of differential gene expression**
446 Principal component analysis was performed to provide further quality control using cross-validation
447 (Qlucore Omics Explorer [QoE] 2.3). Partial least square discriminant analysis (PLSDA) was used to
448 assess the Euclidean distance between the unsupervised transcriptomic samples using the MixOmics
449 package for R (Rohart, et al., 2017).
450 We analysed published single-cell RNA-Seq data from human epiblast (inner cell mass) and
451 trophectoderm tissue (Blakeley, et al., 2015; Petropoulos, et al., 2016; Yan, et al., 2013). Transcripts
452 per million (TPM) expression values were visualised in QoE and outliers were removed.
453
454 **Similarity Network Fusion**
455 Gene probe set similarity network fusion (SNF) (Wang, et al., 2014) was performed on the fRMA
456 derived data as an independent test for similarity, using the *SNFTool* R-package. Euclidean distances
457 were calculated between gene probe sets for each hESC line as well as TE and ICM. Using a non-linear
458 network method based on nearest neighbours, any two of the Euclidean distance matrices could be
459 combined over 20 iterations to produce a final network which accurately describes the relationship
460 between gene probe sets across both initial sets. This method was used to combine each hESC line
461 with TE or ICM gene expression data. The fused data was subjected to spectral clustering to identify

462 groups of gene probe sets with similar patterns of expression across the hESC and TE or ICM samples.
463 This data was presented as a heatmap.
464
465 **Hypernetwork assessment of transcriptomic associations**
466 Hypernetworks were generated to understand the relationships between genes which distinguish the
467 trophectoderm and inner cell mass (Stirparo, et al., 2018). Correlations between these transcripts and
468 the rest of the transcriptome were calculated in R (v3.4.4) and the number of shared correlations
469 between pairs of genes was determined (Johnson, 2011). Hierarchical clustering was used to separate
470 a central cluster of genes with high inter-correlation from this network of transcriptomic associations.
471
472 **Network model construction and comparison**
473 Lists of differentially expressed genes were used to generate interactome network models of protein
474 interactions related to the transcriptomic data in Cytoscape (Su, et al., 2014) by inference using the
475 BioGRID database (Chatr-Aryamontri, et al., 2015).
476 The Cytoscape plugin Moduland (Kovacs, et al., 2010; Szalay-Beko, et al., 2012) was applied to identify
477 overlapping modules, an approach that models complex modular architecture within the human
478 interactome (Chang, et al., 2013) by accounting for non-discrete nature of network modules (Kovacs,
479 et al., 2010). Modular hierarchy was determined using a centrality score and further assessed using
480 hierarchical network layouts (summarising the underlying network topology). The overlap between
481 the central module cores (metanode of the ten most central elements) was determined. Community
482 centrality and bridgeness scores were assessed across network models using the Moduland package
483 (Szalay-Beko, et al., 2012). The bridgeness score was used in combination with centrality scores to
484 categorise party and date hubs within the network i.e genes that interact simultaneously or
485 sequentially respectively with neighbours (Komurov and White, 2007; Yu, et al., 2007).
486 The Network Analyser (Assenov, et al., 2008) Cytoscape plugin was used to calculate associated
487 parameters of network topology. Hierarchical network layouts were used along with centrality scores

488 to assess the hierarchy of network clusters. Significance of the overlap between network elements
489 was calculated using Fisher's exact test on the sum of each group compared to the expected sum.
490 The robustness of defined modules is an essential analytical step (Reimand, 2013) and was assessed
491 using permutation analysis in R (version 3.3.2) (RCoreTeam, 2016). Robustness of network module
492 and network topology properties was determined in the ICM and TE interactome network models with
493 100 permutations of removal of 5, 10, 20, 30, 40 and 50% of the nodes, an approach that has been
494 shown to assess the coherency of network modules (Reimand, 2013). These data were used to assess
495 the stability of network observations.

496 **Acknowledgements:** We thank Stuart Pepper at the Christie Genomics unit for Microarray
497 hybridisation. We thank the Medical Research Council UK for funding (grants G0801057 MR/M01
498 7354/1), and the NIHR clinical research network for support. We would particularly like to thank the
499 research nurses, IVF clinic staff and patients who donated embryos to this research.

500 **References**

501

502 Adachi, K.*, et al.* Esrrb Unlocks Silenced Enhancers for Reprogramming to Naive Pluripotency. *Cell*

503 *stem cell* 2018;23(2):266-275 e266.

504 Adjaye, J.*, et al.* Primary differentiation in the human blastocyst: comparative molecular portraits of

505 inner cell mass and trophectoderm cells. *Stem cells (Dayton, Ohio)* 2005;23(10):1514-1525.

506 Agarwal, S.*, et al.* Revisiting Date and Party Hubs: Novel Approaches to Role Assignment in Protein

507 Interaction Networks. *PLOS Computational Biology* 2010;6(6):e1000817.

508 Al-Taher, A.*, et al.* Global cDNA amplification combined with real-time RT-PCR: accurate

509 quantification of multiple human potassium channel genes at the single cell level. *Yeast (Chichester,*

510 *England)* 2000;17(3):201-210.

511 Apostolou, E.*, et al.* Genome-wide chromatin interactions of the Nanog locus in pluripotency,

512 differentiation, and reprogramming. *Cell stem cell* 2013;12(6):699-712.

513 Artus, J.*, et al.* PDGF signaling is required for primitive endoderm cell survival in the inner cell mass

514 of the mouse blastocyst. *Stem cells (Dayton, Ohio)* 2013;31(9):1932-1941.

515 Assenov, Y.*, et al.* Computing topological parameters of biological networks. *Bioinformatics (Oxford,*

516 *England)* 2008;24(2):282-284.

517 Blakeley, P.*, et al.* Defining the three cell lineages of the human blastocyst by single-cell RNA-seq.

518 *Development* 2015;142(20):3613.

519 Bloor, D.J.*, et al.* Expression of cell adhesion molecules during human preimplantation embryo

520 development. *Molecular human reproduction* 2002;8(3):237-245.

521 Boyer, L.A.*, et al.* Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell*

522 2005;122(6):947-956.

523 Brady, G. and Iscove, N.N. Construction of cDNA libraries from single cells. *Methods in enzymology*

524 1993;225:611-623.

525 Cahan, P.*, et al.* CellNet: network biology applied to stem cell engineering. *Cell* 2014;158(4):903-915.

526 Camarasa, M.V.*, et al.* Derivation of Man-1 and Man-2 research grade human embryonic stem cell
527 lines. *In Vitro Cell Dev Biol Anim* 2010;46(3-4):386-394.
528 Cauffman, G.*, et al.* Markers that define stemness in ESC are unable to identify the totipotent cells in
529 human preimplantation embryos. *Human reproduction (Oxford, England)* 2009;24(1):63-70.
530 Chang, X.*, et al.* Dynamic modular architecture of protein-protein interaction networks beyond the
531 dichotomy of 'date' and 'party' hubs. *Sci Rep* 2013;3:1691.
532 Chatr-Aryamontri, A.*, et al.* The BioGRID interaction database: 2015 update. *Nucleic Acids Res*
533 2015;43(Database issue):D470-478.
534 Cowan, C.A.*, et al.* Derivation of embryonic stem-cell lines from human blastocysts. *N Engl J Med*
535 2004;350(13):1353-1356.
536 De Sousa, P.A.*, et al.* Clinically failed eggs as a source of normal human embryo stem cells. *Stem Cell*
537 *Res* 2009;2(3):188-197.
538 Ding, L.*, et al.* A genome-scale RNAi screen for Oct4 modulators defines a role of the Paf1 complex
539 for embryonic stem cell identity. *Cell stem cell* 2009;4(5):403-415.
540 Faial, T.*, et al.* Brachyury and SMAD signalling collaboratively orchestrate distinct mesoderm and
541 endoderm gene regulatory networks in differentiating human embryonic stem cells. *Development*
542 2015;142(12):2121-2135.
543 Fang, L.*, et al.* H3K4 Methyltransferase Set1a Is A Key Oct4 Coactivator Essential for Generation of
544 Oct4 Positive Inner Cell Mass. *Stem cells (Dayton, Ohio)* 2016;34(3):565-580.
545 Festuccia, N.*, et al.* Esrrb is a direct Nanog target gene that can substitute for Nanog function in
546 pluripotent cells. *Cell stem cell* 2012;11(4):477-490.
547 Fogarty, N.M.E.*, et al.* Genome editing reveals a role for OCT4 in human embryogenesis. *Nature*
548 2017;550(7674):67-73.
549 Grabarek, J.B.*, et al.* Differential plasticity of epiblast and primitive endoderm precursors within the
550 ICM of the early mouse embryo. *Development* 2012;139(1):129-139.

551 Han, D.W.*, et al.* Epiblast stem cell subpopulations represent mouse embryos of distinct
552 pregastrulation stages. *Cell* 2010;143(4):617-627.
553 Hasegawa, Y.*, et al.* Variability of Gene Expression Identifies Transcriptional Regulators of Early
554 Human Embryonic Development. *PLoS Genet* 2015;11(8):e1005428.
555 Hiratani, I.*, et al.* Genome-wide dynamics of replication timing revealed by in vitro models of mouse
556 embryogenesis. *Genome Res* 2010;20(2):155-169.
557 Hochedlinger, K. and Jaenisch, R. Induced Pluripotency and Epigenetic Reprogramming. *Cold Spring*
558 *Harbor perspectives in biology* 2015;7(12).
559 Home, P.*, et al.* GATA3 is selectively expressed in the trophectoderm of peri-implantation embryo
560 and directly regulates Cdx2 gene expression. *J Biol Chem* 2009;284(42):28729-28737.
561 Hu, G.*, et al.* A genome-wide RNAi screen identifies a new transcriptional module required for self-
562 renewal. *Genes & development* 2009;23(7):837-848.
563 Huang, K., Maruyama, T. and Fan, G. The naive state of human pluripotent stem cells: a synthesis of
564 stem cell and preimplantation embryo transcriptome analyses. *Cell stem cell* 2014;15(4):410-415.
565 Iscove, N.N.*, et al.* Representation is faithfully preserved in global cDNA amplified exponentially from
566 sub-picogram quantities of mRNA. *Nat.Biotechnol.* 2002;20(9):940-943.
567 Ivanova, N.*, et al.* Dissecting self-renewal in stem cells with RNA interference. *Nature*
568 2006;442(7102):533-538.
569 Johnson, J. HYPERNETWORKS IN THE SCIENCE OF COMPLEX SYSTEMS. Imperial College Press; 2011.
570 Kilens, S.*, et al.* Parallel derivation of isogenic human primed and naive induced pluripotent stem
571 cells. *Nat Commun* 2018;9(1):360.
572 Kimber, S.J.*, et al.* Expression of genes involved in early cell fate decisions in human embryos and
573 their regulation by growth factors. *Reproduction (Cambridge, England)* 2008;135(5):635-647.
574 Komurov, K. and White, M. Revealing static and dynamic modular architecture of the eukaryotic
575 protein interaction network. *Mol Syst Biol* 2007;3:110.

576 Kovacs, I.A.*, et al.* Community landscapes: an integrative approach to determine overlapping
577 network module hierarchy, identify key nodes and predict network dynamics. *PLoS One* 2010;5(9).
578 Kuckenberg, P., Kubaczka, C. and Schorle, H. The role of transcription factor Tcfap2c/TFAP2C in
579 trophectoderm development. *Reproductive biomedicine online* 2012;25(1):12-20.
580 Latos, P.A.*, et al.* Fgf and Esrrb integrate epigenetic and transcriptional networks that regulate self-
581 renewal of trophoblast stem cells. *Nature Communications* 2015;6:7776.
582 Liang, J.*, et al.* Nanog and Oct4 associate with unique transcriptional repression complexes in
583 embryonic stem cells. *Nature cell biology* 2008;10(6):731-739.
584 Marikawa, Y. and Alarcon, V.B. Creation of trophectoderm, the first epithelium, in mouse
585 preimplantation development. *Results and problems in cell differentiation* 2012;55:165-184.
586 McCall. Frozen Robust Multi-Array Analysis and the Gene Expression Barcode. 2015.
587 McCall, M.N., Bolstad, B.M. and Irizarry, R.A. Frozen robust multiarray analysis (fRMA). *Biostatistics*
588 *(Oxford, England)* 2010;11(2):242-253.
589 McCall, M.N.*, et al.* The Gene Expression Barcode 3.0: improved data processing and mining tools.
590 *Nucleic Acids Res* 2014;42(Database issue):D938-943.
591 Nakamura, T.*, et al.* A developmental coordinate of pluripotency among mice, monkeys and humans.
592 *Nature* 2016;537(7618):57-62.
593 Ng, P.M. and Lufkin, T. Embryonic stem cells: protein interaction networks. *Biomolecular concepts*
594 2011;2(1-2):13-25.
595 Niakan, K.K. and Eggan, K. Analysis of human embryos from zygote to blastocyst reveals distinct gene
596 expression patterns relative to the mouse. *Dev Biol* 2013;375(1):54-64.
597 Niakan, K.K. and Eggan, K. Analysis of human embryos from zygote to blastocyst reveals distinct gene
598 expression patterns relative to the mouse. *Developmental Biology* 2013;375(1):54-64.
599 Nicola, F., Nick, O. and Pablo, N. Esrrb, an estrogen-related receptor involved in early development,
600 pluripotency, and reprogramming. *FEBS Letters* 2018;592(6):852-877.

601 Nishioka, N.*, et al.* Tead4 is required for specification of trophectoderm in pre-implantation mouse
602 embryos. *Mechanisms of Development* 2008;125(3):270-283.
603 Niwa, H.*, et al.* Interaction between Oct3/4 and Cdx2 determines trophectoderm differentiation. *Cell*
604 2005;123(5):917-929.
605 Oldershaw, R.A.*, et al.* Directed differentiation of human embryonic stem cells toward chondrocytes.
606 *Nat Biotechnol* 2010;28(11):1187-1194.
607 Pardo, M.*, et al.* An expanded Oct4 interaction network: implications for stem cell biology,
608 development, and disease. *Cell stem cell* 2010;6(4):382-395.
609 Petropoulos, S.*, et al.* Single-Cell RNA-Seq Reveals Lineage and X Chromosome Dynamics in Human
610 Preimplantation Embryos. *Cell* 2016;165(4):1012-1026.
611 Petropoulos, S.*, et al.* Single-Cell RNA-Seq Reveals Lineage and X Chromosome Dynamics in Human
612 Preimplantation Embryos. *Cell* 2016;167(1):285.
613 Pierreux, C.E.*, et al.* The transcription factor hepatocyte nuclear factor-6 controls the development
614 of pancreatic ducts in the mouse. *Gastroenterology* 2006;130(2):532-541.
615 RCoreTeam. R: A language and environment for statistical computing. *R Foundation for Statistical*
616 *Computing, Vienna, Austria* 2016;https://www.R-project.org/.
617 Reimand. Thread 2: Network models. *Nature Genetics* 2013;45(10).
618 Rizvi, A.H.*, et al.* Single-cell topological RNA-seq analysis reveals insights into cellular differentiation
619 and development. *Nat Biotechnol* 2017;35(6):551-560.
620 Rohart, F.*, et al.* mixOmics: an R package for 'omics feature selection and multiple data integration.
621 *bioRxiv* 2017.
622 Rossant, J., Chazaud, C. and Yamanaka, Y. Lineage allocation and asymmetries in the early mouse
623 embryo. *Philos Trans R Soc Lond B Biol Sci* 2003;358(1436):1341-1348; discussion 1349.
624 Schrode, N.*, et al.* Anatomy of a blastocyst: cell behaviors driving cell fate choice and morphogenesis
625 in the early mouse embryo. *Genesis* 2013;51(4):219-233.

626 Shaw, L.*, et al.* Comparison of gene expression in fresh and frozen-thawed human preimplantation
627 embryos. *Reproduction (Cambridge, England)* 2012;144(5):569-582.
628 Shaw, L.*, et al.* Global gene expression profiling of individual human oocytes and embryos
629 demonstrates heterogeneity in early development. *PLoS One* 2013;8(5):e64192.
630 Smith, H.L.*, et al.* Systems based analysis of human embryos and gene networks involved in cell
631 lineage allocation. *BMC Genomics* 2019;20(1):171.
632 Stephenson, R.O., Rossant, J. and Tam, P.P. Intercellular interactions, position, and polarity in
633 establishing blastocyst cell lineages and embryonic axes. *Cold Spring Harbor perspectives in biology*
634 2012;4(11).
635 Stirparo, G.G.*, et al.* Integrated analysis of single-cell embryo data yields a unified transcriptome
636 signature for the human pre-implantation epiblast. *Development* 2018;145(3).
637 Strumpf, D.*, et al.* Cdx2 is required for correct cell fate specification and differentiation of
638 trophectoderm in the mouse blastocyst. *Development* 2005;132(9):2093-2102.
639 Su, G.*, et al.* Biological network exploration with Cytoscape 3. *Current protocols in bioinformatics*
640 2014;47:8 13 11-24.
641 Szalay-Beko, M.*, et al.* ModuLand plug-in for Cytoscape: determination of hierarchical layers of
642 overlapping network modules and community centrality. *Bioinformatics (Oxford, England)*
643 2012;28(16):2202-2204.
644 Team, R.C. R: A language and environment for statistical computing. In. Foundation for Statistical
645 Computing, Vienna, Austria; 2014.
646 Tesar, P.J.*, et al.* New cell lines from mouse epiblast share defining features with human embryonic
647 stem cells. *Nature* 2007;448(7150):196-199.
648 Uranishi, K.*, et al.* Esrrb directly binds to Gata6 promoter and regulates its expression with Dax1 and
649 Ncoa3. *Biochemical and Biophysical Research Communications* 2016;478(4):1720-1725.
650 van den Berg, D.L.*, et al.* An Oct4-centered protein interaction network in embryonic stem cells. *Cell*
651 *stem cell* 2010;6(4):369-381.

652 Wang, B.*, et al.* Similarity network fusion for aggregating data types on a genomic scale. *Nat*
653 *Methods* 2014;11(3):333-337.
654 Wang, J.*, et al.* A protein interaction network for pluripotency of embryonic stem cells. *Nature*
655 2006;444(7117):364-368.
656 Weinberger, L.*, et al.* Dynamic stem cell states: naive to primed pluripotency in rodents and humans.
657 *Nat Rev Mol Cell Biol* 2016;17(3):155-169.
658 Yan, L.*, et al.* Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem
659 cells. *Nature Structural &Amp; Molecular Biology* 2013;20:1131.
660 Yu, H.*, et al.* The importance of bottlenecks in protein networks: correlation with gene essentiality
661 and expression dynamics. *PLoS Comput Biol* 2007;3(4):e59.
662 Zhang, J.Z.*, et al.* Screening for genes essential for mouse embryonic stem cell self-renewal using a
663 subtractive RNA interference library. *Stem cells (Dayton, Ohio)* 2006;24(12):2661-2668.
664 Zhang, X.*, et al.* Esrrb activates Oct4 transcription and sustains self-renewal and pluripotency in
665 embryonic stem cells. *J Biol Chem* 2008;283(51):35825-35833.
666 Zilliox, M.J. and Irizarry, R.A. A gene expression bar code for microarray data. *Nat Methods*
667 2007;4(11):911-913.

668

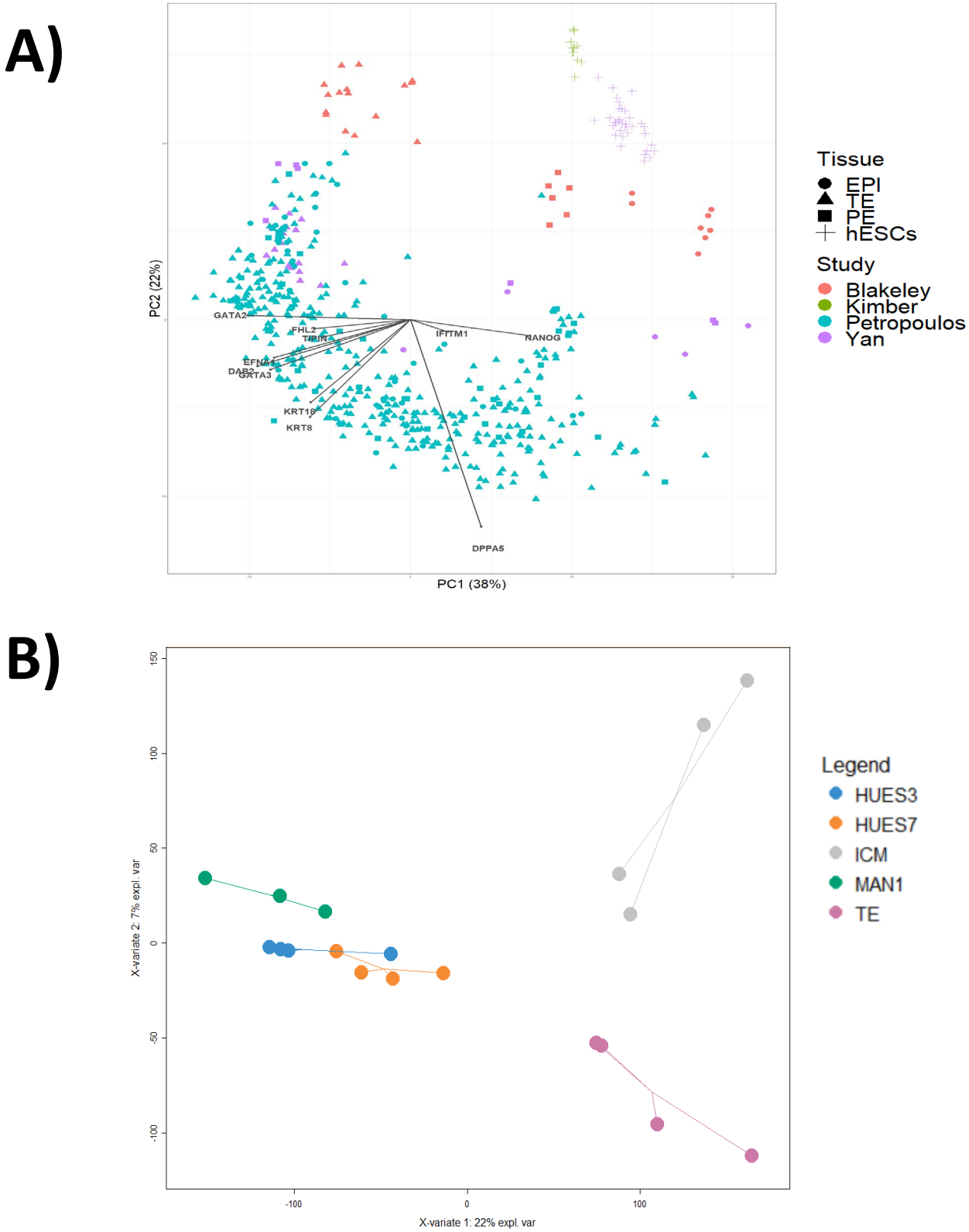


Figure 1. Distance between the transcriptomes of inner cell mass, trophectoderm and human embryonic cell lines as a measure of similarity.

A) Principal component analysis (PCA) using genes defined by Stirparo *et al* as distinguishing trophectoderm from epiblast. Performed using matching genes across 4 studies, including embryonic tissue and stem cells. Arrows show eigen vectors demonstrating the contributions of key genes to principal components. N-samples=520, N-genes=452.
B) 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.

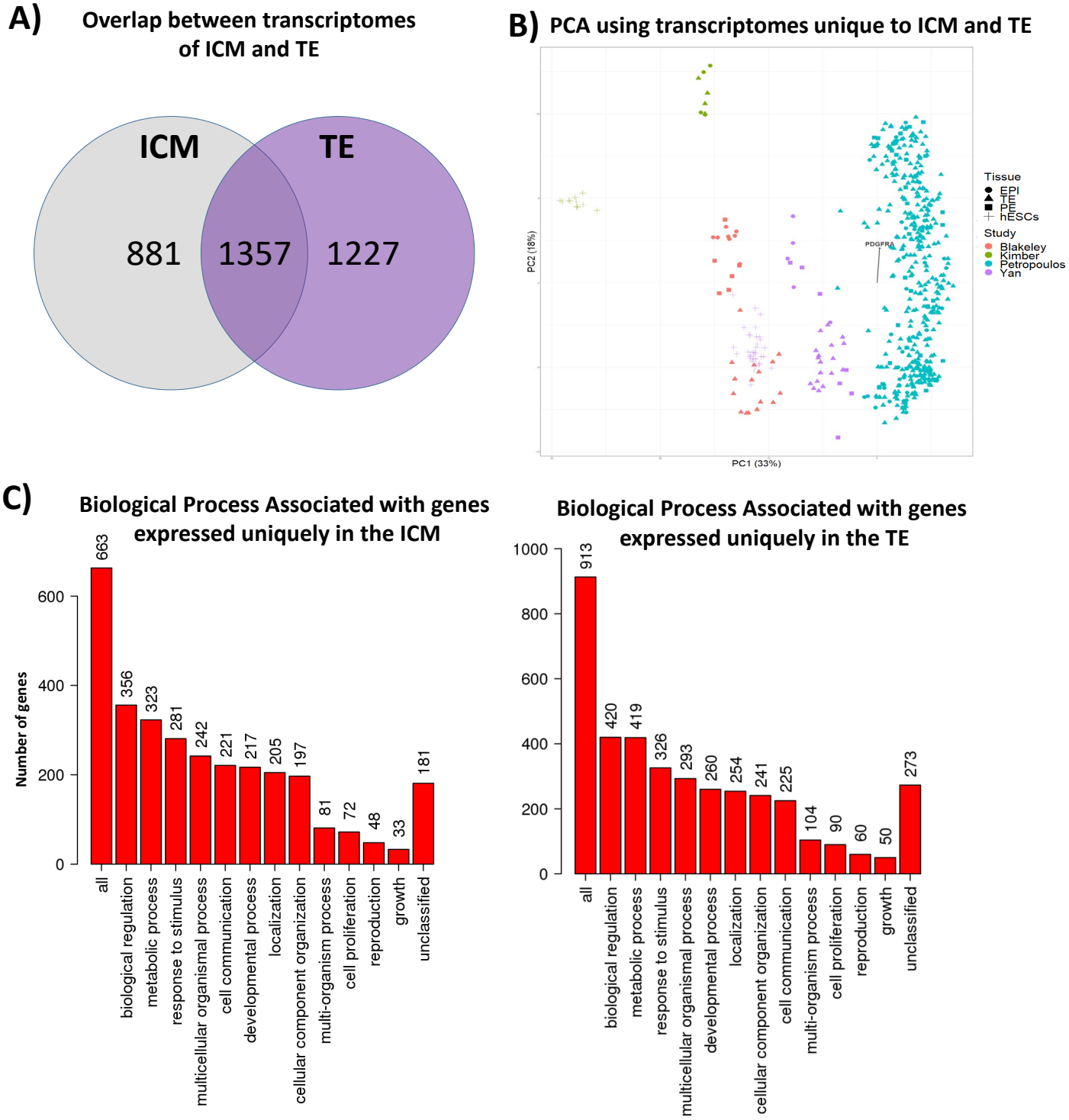


Figure 2. Inner cell mass and trophectoderm specific transcriptome and associated gene ontology

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)** PCA using genes defined to distinguish trophectoderm and epiblast by gene barcode on embryonic cells. N-samples=528, N-genes=452. **C)** 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.

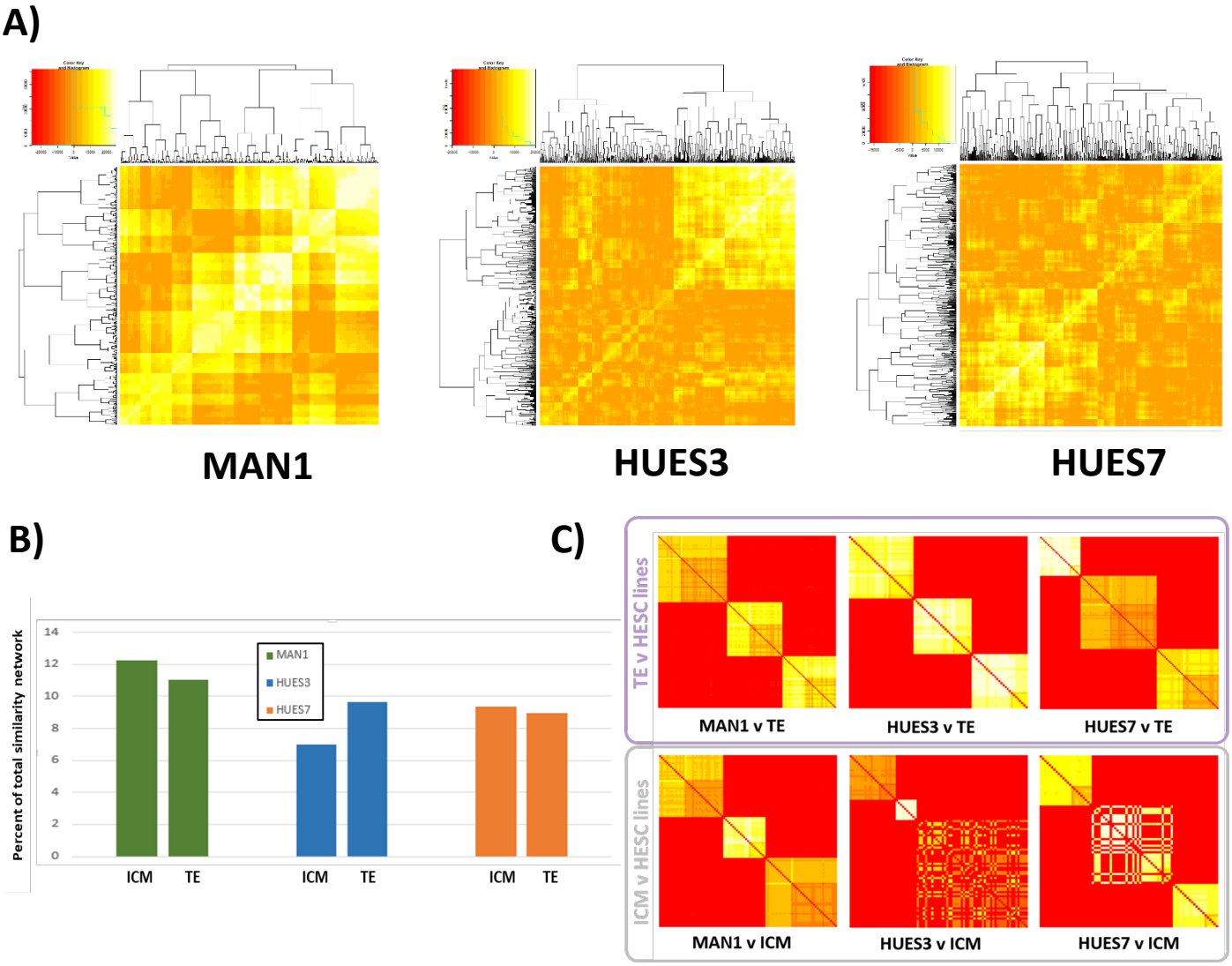


Figure 3. Similarity network fusion to compare homology between the transcriptome of inner cell mass and trophectoderm and human embryonic stem cell lines.

Similarity network fusion matrix showing similarity groups between the uniquely expressed ICM and TE gene probesets and the human embryonic stem cell lines (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. Hypernetworks of genes distinguishing trophectoderm from inner cell mass correlated against all other genes in three stem cell lines (HUES3, HUES7, MAN1). Yellow denotes high connectivity between epiblast-distinguishing genes and others.

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.

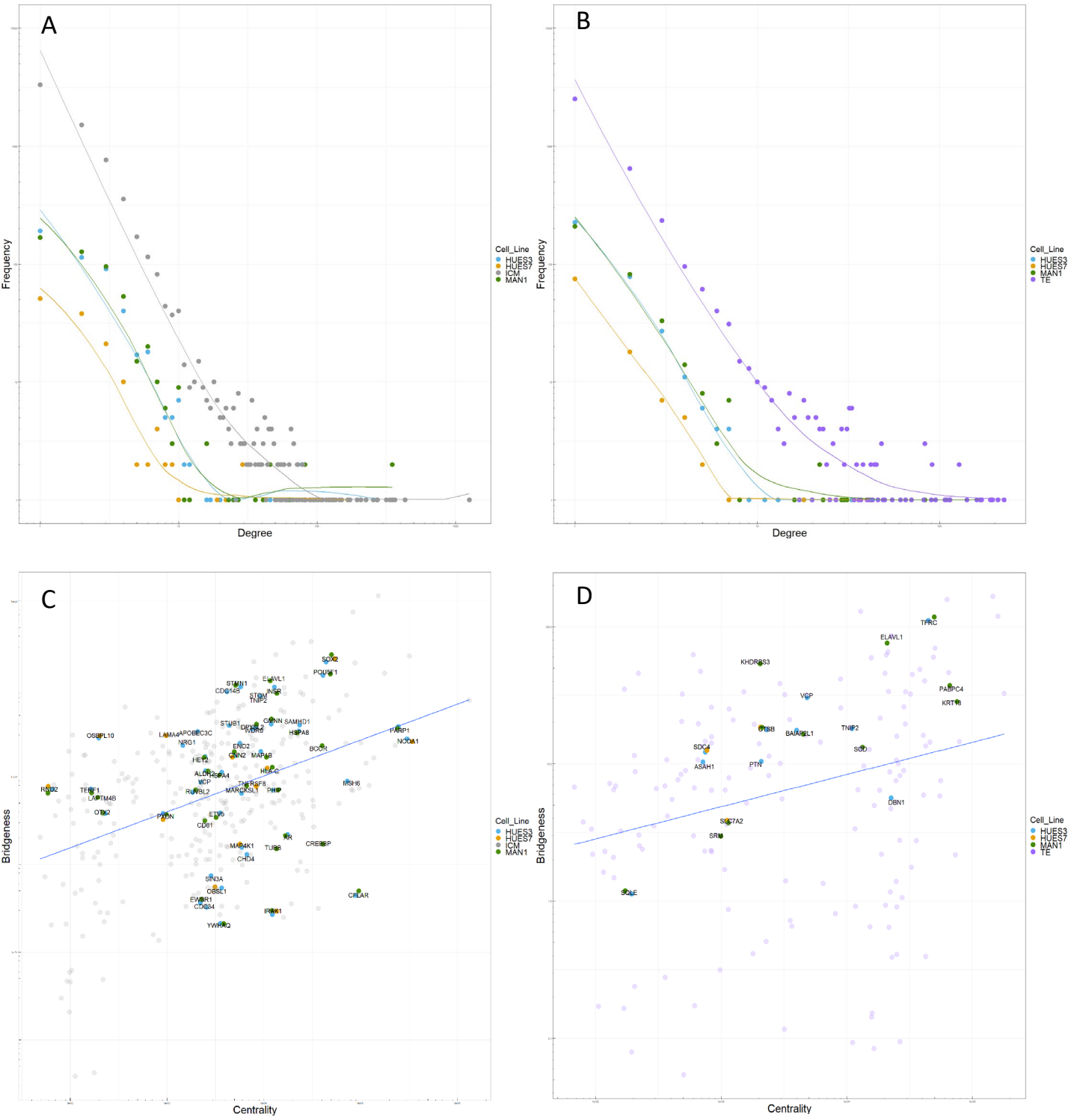


Figure 4. The network topology of the ICM interactome is enriched in human embryonic stem cells.
Degree distribution of unique genes in an inner cell mass (**A**) and trophectoderm (**B**) network model. HUES3, HUES7 and MAN1 are subsets of TE or ICM. These plots demonstrate that HUES3 and MAN1 are more connected than HUES7 in both networks. Plots are of log frequency and log degree.
Bridgeness vs centrality measures in an inner cell mass (**C**) and trophectoderm (**D**) network model. HUES3,HUES7 and MAN1 are subsets of TE or ICM. HUES3 and MAN1 have a greater proportion of date-like hubs than HUES7 or either ICM or TE, demonstrating an increased number of genes with network properties of transcription factors.

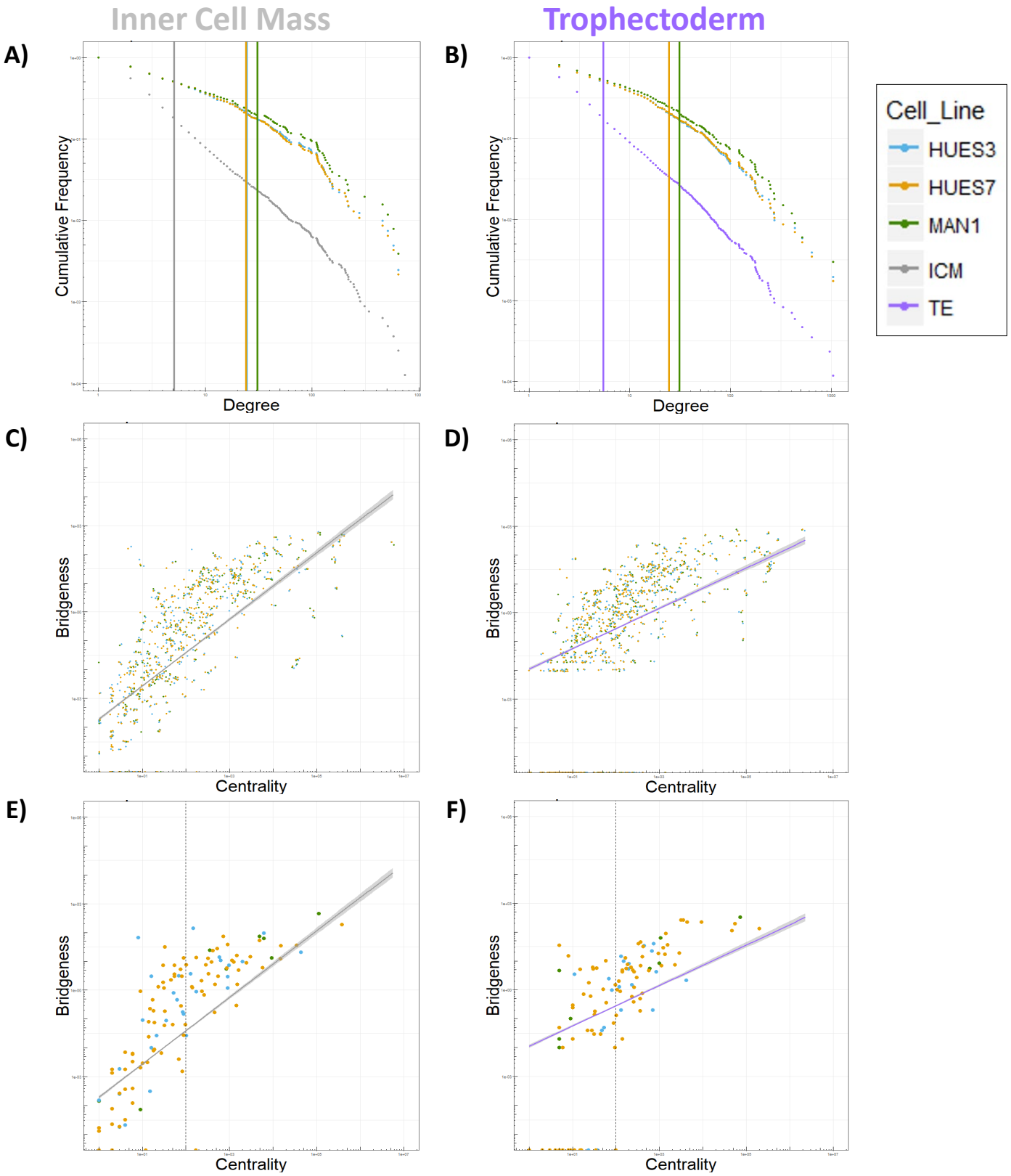


Figure 5. The network topology of the ICM interactome based on the unique expression of genes compared to TE is enriched in human embryonic stem cells.

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). Plots are of log frequency and log degree. **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.

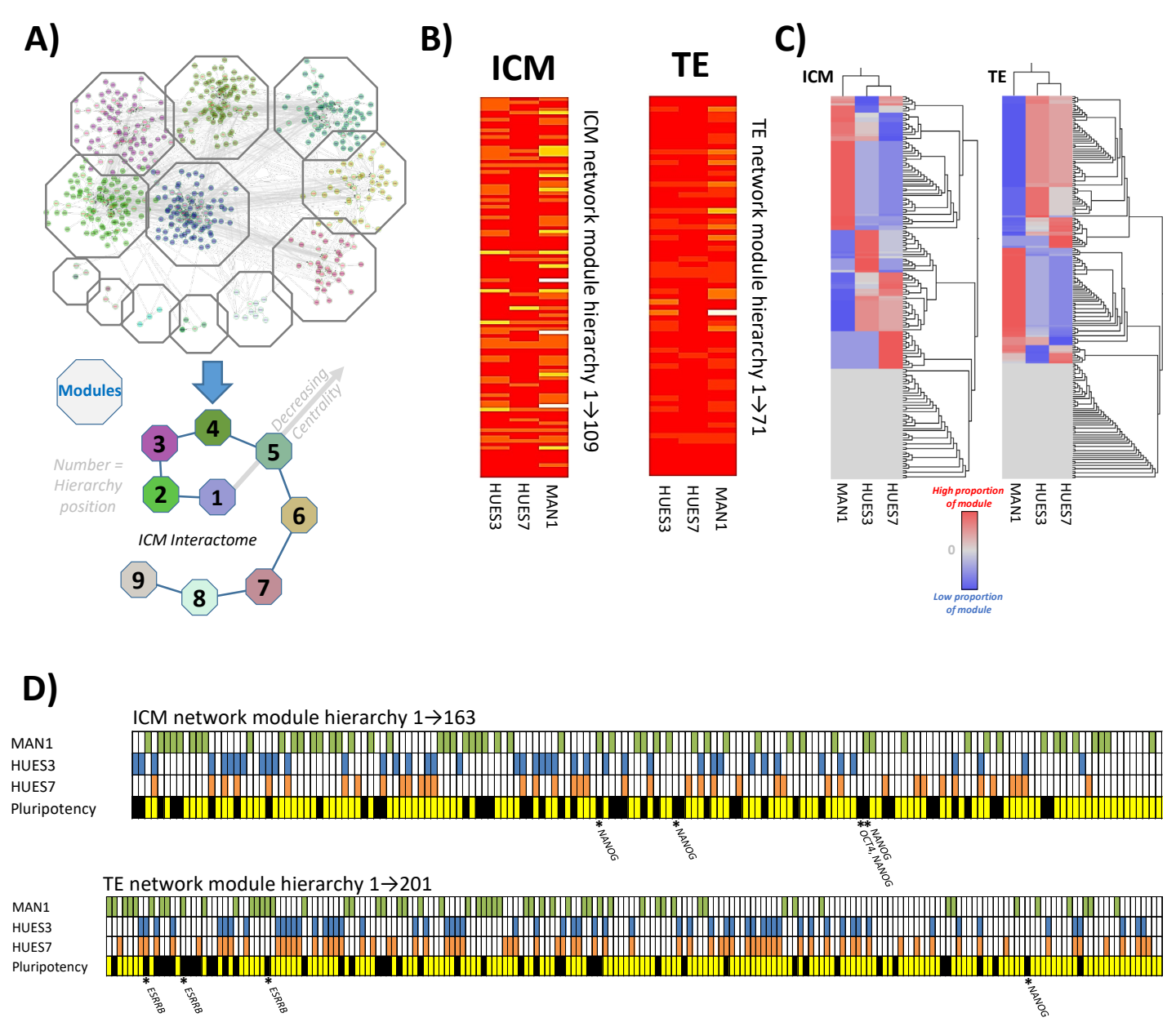
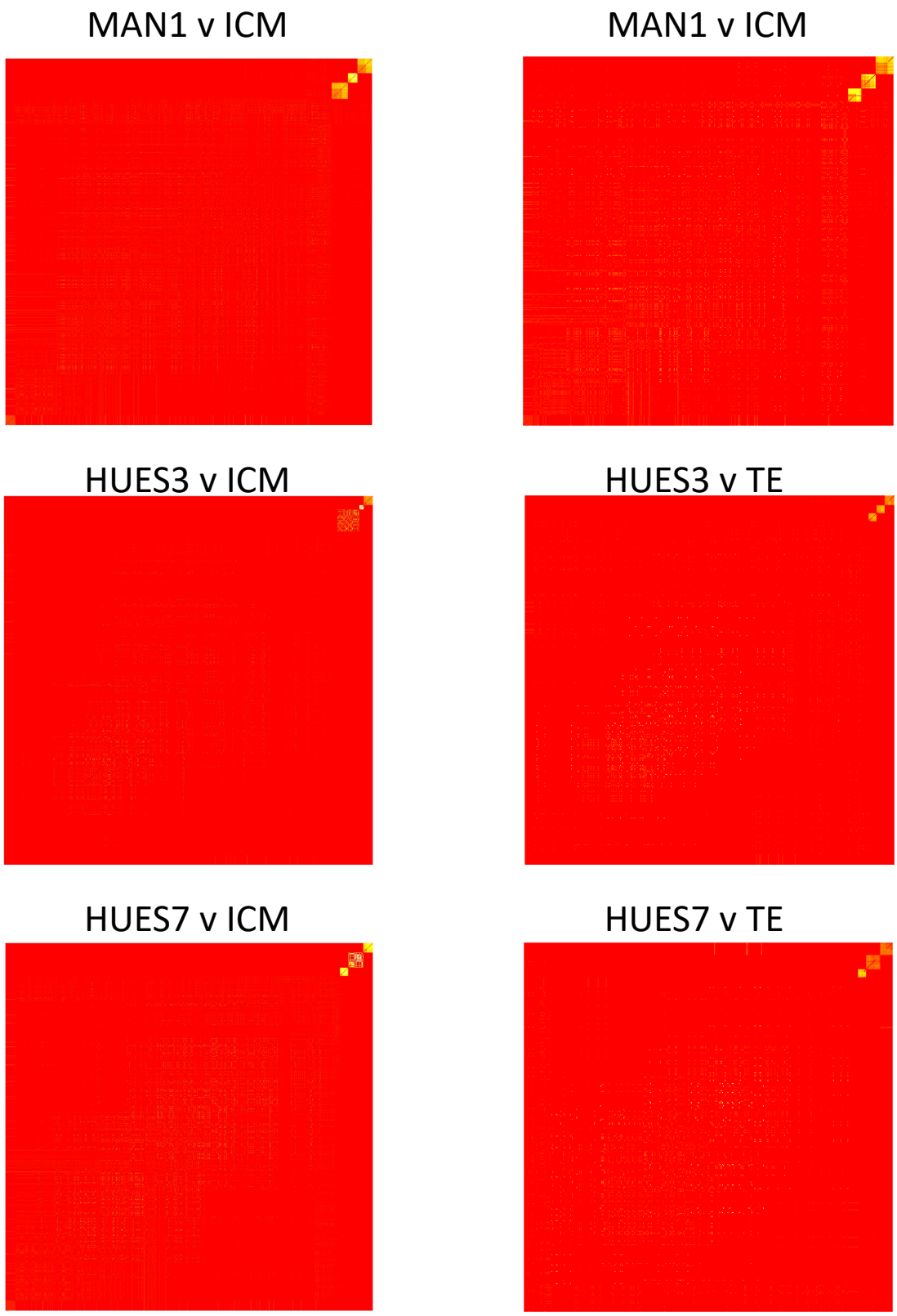


Figure 6. The modular structure of the interactome network model of gene expression unique to ICM and TE can be used as a framework to assess similarity with human embryonic stem cells.

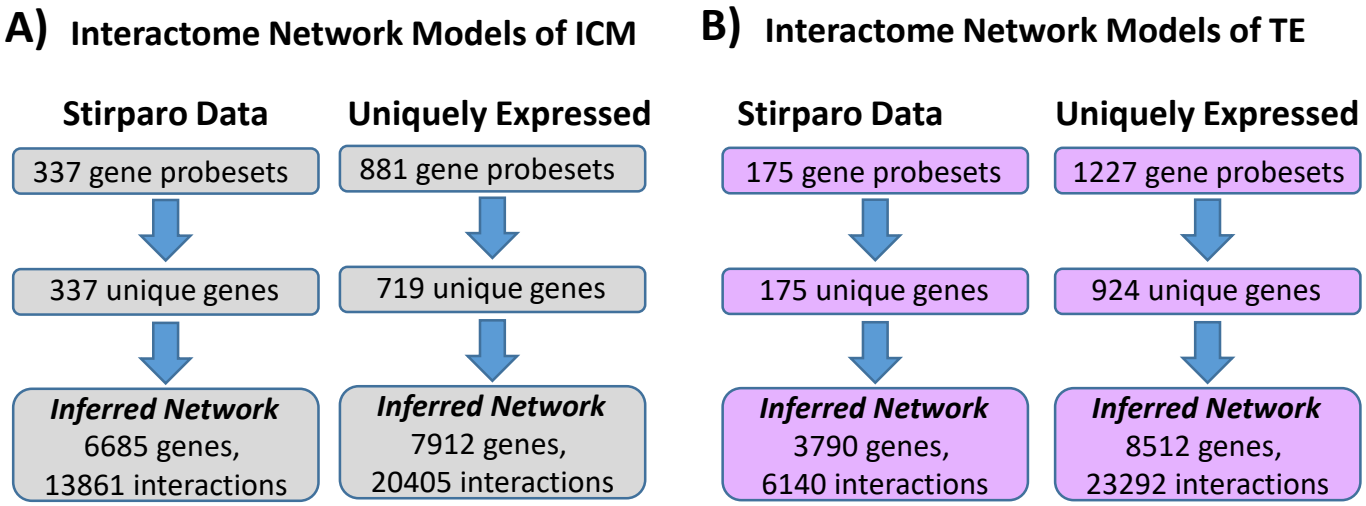
A) The modular structure of the ICM interactome was defined using the Moduland algorithm to assess the presence of highly connected gene modules. These were then formed into a hierarchy based on their centrality score, a measurement of network topology related to the influence of a network element on the rest of the network. **B)** Network module hierarchy in an inner cell mass and trophectoderm network models based on the Stirparo data. Yellow and orange bands demonstrate the presence of genes determined to be unique to each cell line using a gene barcode approach. The greater the intensity of the yellowness, the larger the number of unique genes represented in the meta-node (10 most central nodes in each module). This shows that MAN1 has a greater number of unique genes represented in both TE and ICM networks, particularly in the most central modules and to a greater extent in ICM than TE. **C)** 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. **D)** 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 lines (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.



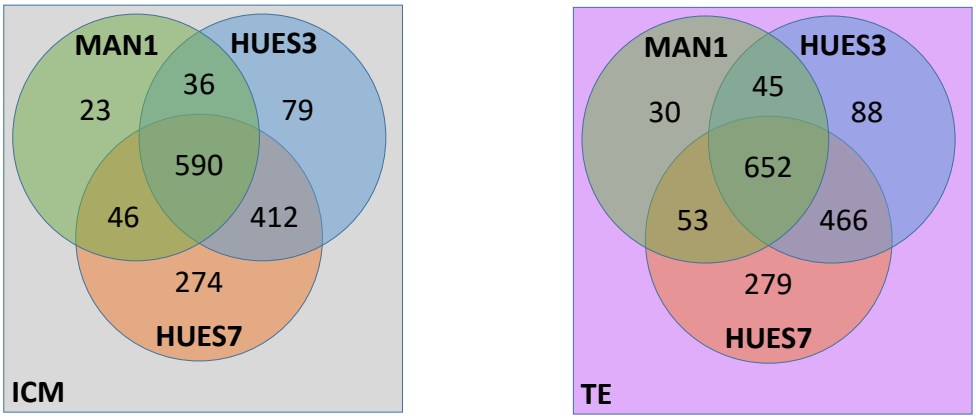
Supplemental Figure 2. Interactome network models of gene expression in ICM or TE.

A) In Inner Cell Mass (ICM). Using differentially expressed genes between ICM and TE, unique patterns of transcriptomic expression were defined. Genes with positive expression in ICM (337 from Stirparo meta-analysis) were used to generate an interactome network model for ICM. A second ICM interactome model was generated using the 719 genes (881 gene probesets) uniquely expressed in ICM from our de novo analysis.

B) In Trophectoderm (TE) The genes differentially expressed between ICM and TE were used (Stirparo datasets from meta-analysis) and genes with positive expression in TE (175) were used to generate an interactome network model. A second TE interactome model was generated using the 924 genes (1227 gene probesets) uniquely expressed in TE from our de novo analysis.

These were used to infer interactome network models using the BioGRID database version 3.4.158.

A)

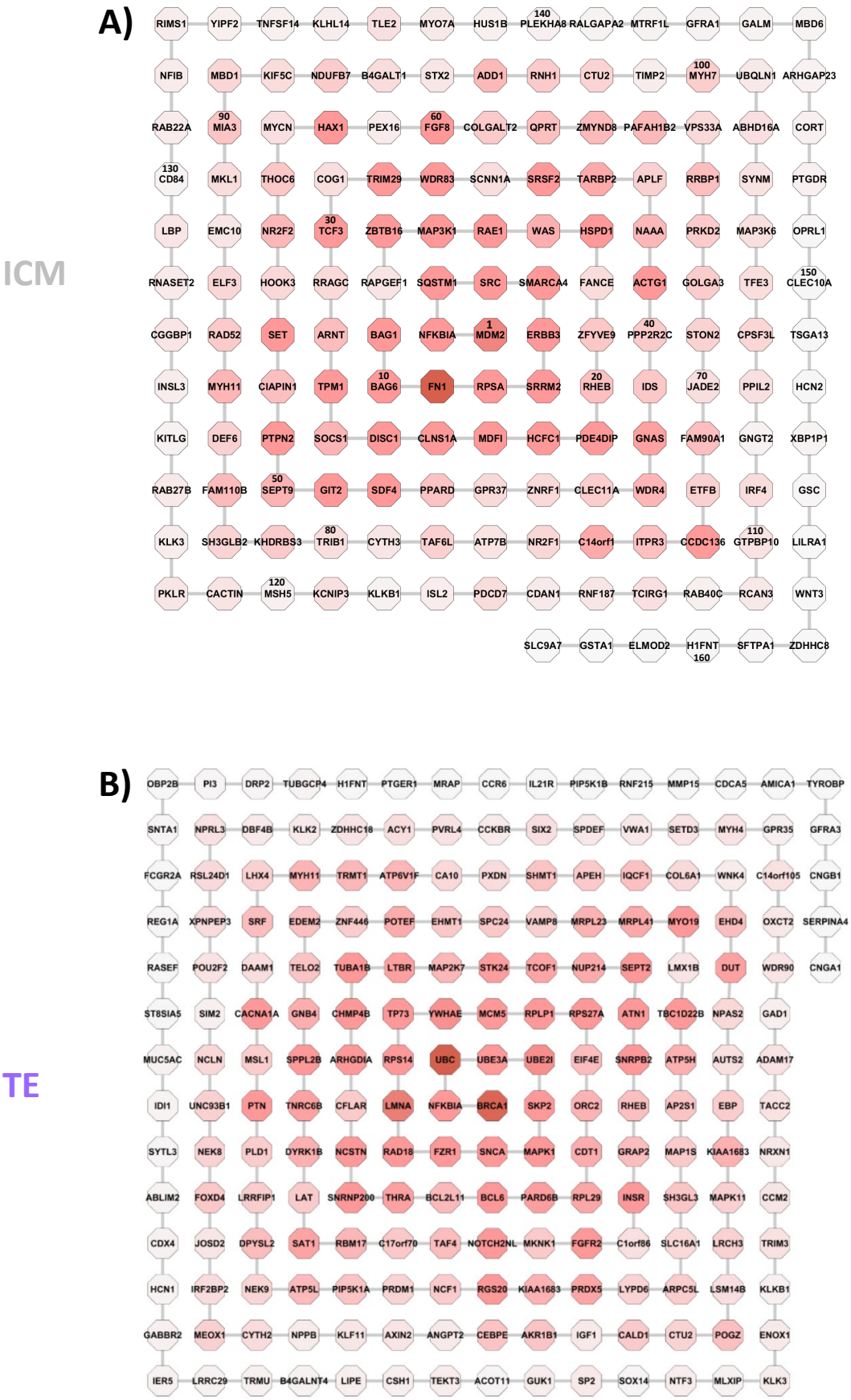


B)

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						

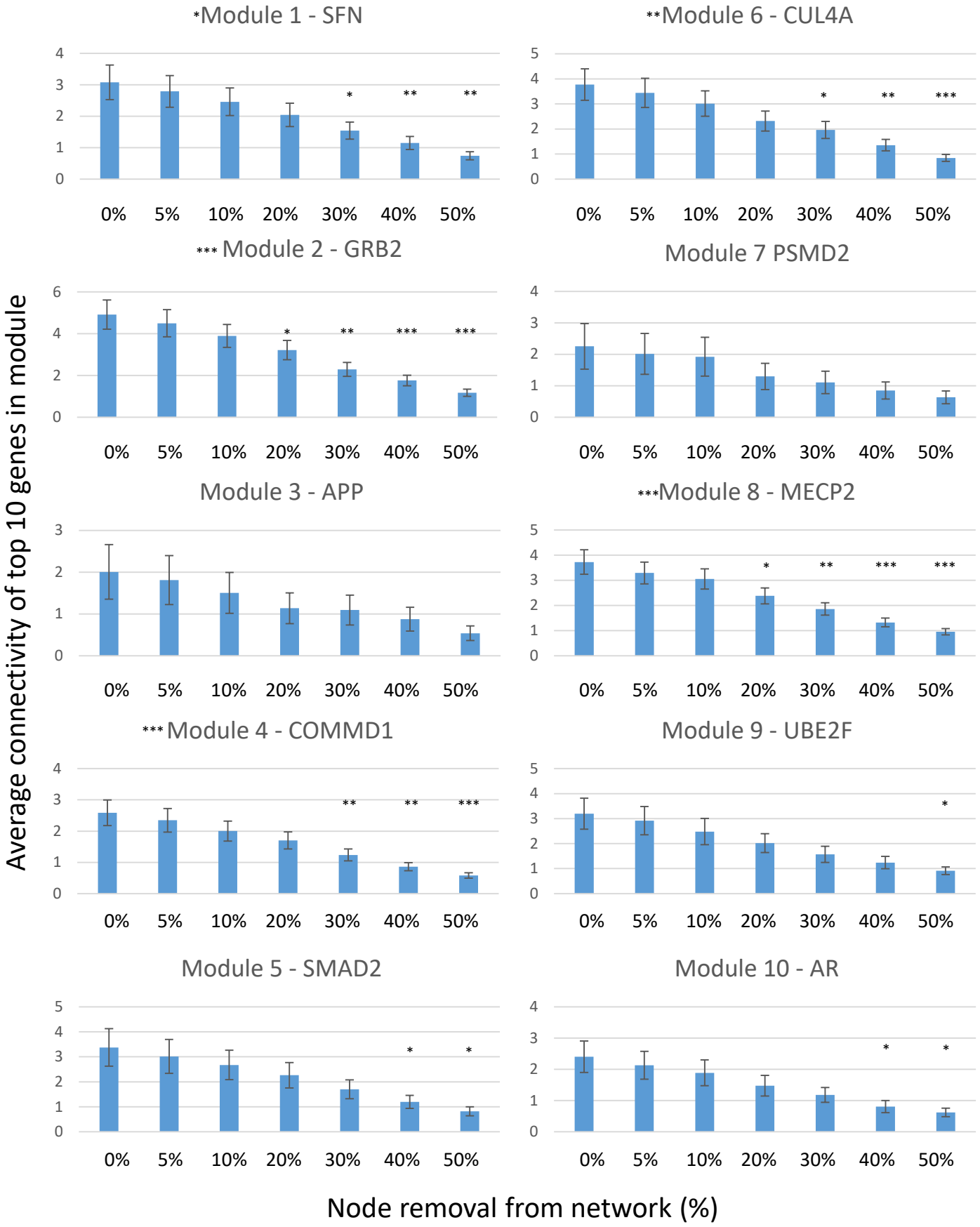
Supplemental Figure S3. Expressed genes uniquely shared between each human embryonic stem cell line and either the Inner Cell Mass (ICM) or the Trophectoderm (TE).

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 S4. Hierarchy of modules within the interactome network models of ICM and TE based on uniquely expressed genes.

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 5. 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. [* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$].