HHS Public Access

Author manuscript

*Stem Cells*. Author manuscript; available in PMC 2015 May 01.

Published in final edited form as:

*Stem Cells*. 2014 May ; 32(5): 1161–1172. doi:10.1002/stem.1612.

# Novel Insights into Embryonic Stem Cell Self-Renewal Revealed Through Comparative Human and Mouse Systems Biology Networks

**Karen G. Dowell**[1,2], **Allen K. Simons**[1], **Hao Bai**[3], **Braden Kell**[1], **Zack Z. Wang**[2,3], **Kyuson Yun**[1,2], and **Matthew A. Hibbs**[*,1,2,4]

[1]The Jackson Laboratory, Bar Harbor, Maine, United States of America

[2] Graduate School of Biomedical Sciences and Engineering, University of Maine, Orono, Maine, United States of America

[3]Johns Hopkins University School of Medicine, Baltimore, Maryland, United States of America

[4]Trinity University, Department of Computer Science, San Antonio, Texas, United States of America

## Abstract

Embryonic stem cells (ESCs), characterized by their ability to both self-renew and differentiate into multiple cell lineages, are a powerful model for biomedical research and developmental biology. Human and mouse ESCs share many features, yet have distinctive aspects, including fundamental differences in the signaling pathways and cell cycle controls that support self-renewal. Here, we explore the molecular basis of human ESC self-renewal using Bayesian network machine learning to integrate cell-type-specific, high-throughput data for gene function discovery. We integrated high-throughput ESC data from 83 human studies (~1.8 million data points collected under 1100 conditions) and 62 mouse studies (~2.4 million data points collected under 1085 conditions) into separate human and mouse predictive networks focused on ESC self-renewal to analyze shared and distinct functional relationships among protein-coding gene orthologs. Computational evaluations show that these networks are highly accurate, literature validation confirms their biological relevance, and RT-PCR validation supports our predictions. Our results reflect the importance of key regulatory genes known to be strongly associated with self-renewal and pluripotency in both species (e.g. *POU5F1*, *SOX2*, and *NANOG*), identify metabolic differences between species (e.g. threonine metabolism), clarify differences between human and mouse ESC developmental signaling pathways (e.g. LIF-activated JAK/STAT in mouse; NODAL/ACTIVIN-A-activated FGF in human), and reveal many novel genes and pathways predicted to be functionally associated with self-renewal in each species. These interactive networks are available online at www.StemSight.org for stem cell researchers to

[*] Corresponding author mhibbs@trinity.edu.

*Author contributions:* K.G.D.: conception and design, financial support, collection and/or assembly of data, data analysis and interpretation, manuscript writing; A.K.S.: collection and/or assembly of data, software development, technical support; H.B.: collection and/or assembly of data; B.K.: collection and/or assembly of data, software development; Z.Z.W.: financial support, data analysis and interpretation; K.Y.: data analysis and interpretation, manuscript editing; M.A.H.: conception and design, financial support, collection and/or assembly of data, data analysis and interpretation, manuscript writing, final approval of manuscript

develop new hypotheses, discover potential mechanisms involving sparsely annotated genes, and prioritize genes of interest for experimental validation.

## Introduction

Embryo-derived pluripotent stem cells are a powerful model system for biomedical research and the study of developmental biology. Among the most studied embryo-derived stem cells are human and mouse ESCs (hESCs and mESCs). Since first isolated from human embryos in 1998 [1], hESCs have been of particular interest to the research community as a tool for manipulating cell fate, analyzing characteristics of induced pluripotent stem (iPS) cells and cancer-stem-like cells, and testing potential medical and pharmaceutical applications [2].

All ESCs express common markers of pluripotency, such as *POU5F1* (also known as *OCT4*), *SOX2,* and *NANOG*, but there are many known differences between species, including morphological, molecular, and epigenetic characteristics that are reflected in ESCs grown in culture. hESCs are derived from the inner cell mass (ICM) of blastocysts 5-8 days post fertilization and form flat, two-dimensional round colonies with well defined boundaries, whereas mESCs are isolated from the ICM of blastocysts 3.5-4.5 days post fertilization and form tight, dome-shaped colonies [3]. hESCs respond to cooperative signaling between NODAL/ACTIVINA-activated FGF and TGF-β signaling pathways to sustain prolonged self-renewal *in vitro* [1, 3, 4]. In contrast, mESCs grown in culture require growth factors, LIF and BMP4, to activate JAK/STAT signaling [5-7]. Self-renewal in mESCs can be boosted by small molecule inhibitors that block differentiation cues from the FGF/ERK signaling cascade and mimic WNT/β-catenin signaling [8]. Interestingly, the same culture conditions that support self-renewal in mESCs can drive differentiation in hESCs. For example, hESCs exposed to LIF and BMP4 yield extraembryonic phenotypes, and FGF inhibition promotes neuroectoderm commitment [9]. In general, well-defined, standard protocols exist for growing mESCs in culture using cell lines of similar genetic backgrounds (predominantly derived from 129S/P/T substrains) [10, 11]. However, hESC cell lines have been derived from genetically distinct embryos in multiple laboratories, each using different media cocktails and protocols to promote self-renewal and the pluripotent state; consequently individual hESC cell lines may respond differently when grown under the same culture conditions [12].

mESCs share epigenetic traits of preimplantation blastocysts and are said to be at the “naïve” or primitive developmental ground state of pluripotency, prior to X-chromosome inactivation and genomic imprinting [11]. These naïve ESCs show no differentiation bias, can self-renew indefinitely *in vitro,* and can be expanded clonally without compromising the pluripotent state [13]. In addition, mESCs can be genetically modified, then reintroduced into preimplantation embryos to generate high-grade chimeric mice [11, 14]. In contrast, hESCs are said to be “primed” for differentiation and female cells have typically undergone random inactivation of one X chromosome [11, 15], however, there is variability in imprinted genes and other DNA methylation patterns depending on culture conditions and passage number [15, 16]. Intriguingly, hESCs share many molecular and epigenetic characteristics with mouse pluripotent stem cells isolated from the post-implantation epiblast

(mEpiSCs) [13, 14, 17], leading to the hypothesis that hESCs and mEpiSCs are both "primed." Although the chimerism assay may not be used with human cells for ethical reasons, studies have shown that lower primate ESCs cannot produce high-grade chimeras when injected into blastocysts, nor can they contribute to the germline [13, 18], indicating that lineage potential is limited *in vivo*. While these phenotypic contrasts are clear, the molecular foundations of naïve versus primed ESCs have not yet been fully characterized.

Understanding the differences between pluripotent cell types is of increasing interest and value as we develop new methods to "reprogram" adult cells to an ESC-like state for ongoing research or therapeutic applications. *In vitro* methods used to reprogram different types of human and mouse adult cells into induced pluripotent stem cells (iPSCs) vary widely, use different combinations of reprogramming factors (*e.g. POU5F1, SOX2, KLF4, KLF2, MYC, LIN28A*) and experimental conditions, and yield ESC-like cells with different degrees of developmental potential [19-21]. In mouse iPSCs, *Nanog* is critical for both blocking differentiation and achieving a naïve pluripotent state [9]. Human iPSCs have also been shown to require *NANOG* to block differentiation, but, to date, attempts to derive or reprogram hESCs to a naïve state have been unsuccessful [13].

Because naïve mESCs and primed hESCs respond to different signaling pathways to sustain and exit the self-renewing state [13], cross-species systems-level analyses of these cell types can reveal molecular details that will assist researchers in assessing the pluripotent state of embryo-derived cells and reprogrammed adult cells, and reveal novel functional homologs that support self-renewal and related early developmental processes. While many systems biology studies have been conducted to explore the molecular basis of self-renewal and pluripotency through gene expression profiles or regulatory networks of transcription factor binding, these efforts have been largely species-specific, restricted to data from a small number of cell lines, limited to a single experimental platform, and/or focused on a single regulatory characteristic influencing cell fate [3, 9, 22-24]. The goals of this study are to discover potential novel regulators of ESC stem cell renewal in hESCs and perform cross-species comparative analyses to further our understanding of shared and distinct molecular characteristics of hESCs and mESCs.

We applied a Bayesian network (Bayes net) machine learning approach based on species- and cell-type-specific data integration, which we previously applied to mESCs [25], to produce a consensus network that predicts gene function associations in the context of hESC self-renewal. Bayes nets, a type of supervised machine learning, are particularly useful for gene function discovery as they provide a statistically principled method to model relationships among genes given a solid foundation of biological knowledge [26-31]. To further our understanding of commonalities and differences between human and mouse ESCs, we generated comparative probabilistic networks for ESC self-renewal using a training set that included examples experimentally validated in hESCs, mESCs, or both species. The results of this study yield insights into novel genes involved in hESC self-renewal, and provide new comparisons of the molecular characteristics of the most widely-studied model ESC systems.

## Materials and Methods

### Master Gene List and Training Set Construction

Using a homology report from Mouse Genome Informatics (MGI) [32], we created a master gene list of 17,342 protein-coding mouse and human gene orthologs with one-to-one homology associations (Supplemental Table S1) [33]. Our positive training examples (Supplemental Table S2) consisted of 2077 manually curated pair-wise gene relationships involving 365 genes associated with mouse and/or human ESC self-renewal based on 108 recent journal articles (Supplemental Table S3), or annotated to early embryonic developmental signaling pathways by the Kyoto Encyclopedia of Genes and Genomes (KEGG) [34] and WikiPathways [35]. Negative/background training examples were generated in a 1:10 ratio by randomly selecting 20,770 gene pairs not involving any genes in positive training examples (Supplemental Table S4).

### Evidential Dataset Collection

We assembled a compendium of high-throughput hESC data, representing 83 independent research studies and consisting of ~1.8 million data points from 55 hESC cell lines (though predominantly H1 and H9), collected under 1100 conditions, using 6 different high-throughput data types, and encompassing more than 12 billion gene-pair measurements (Table 1; Supplemental Table S5). For mESCs, we updated our existing compendium of high-throughput mESC data [25] to include work from 62 independent research studies, consisting of ~2.3 million data points from 35 mESC cell lines (all derived from 129S/P/T substrains), spanning 1085 conditions, using 5 different high-throughput data types, and encompassing more than 7 billion gene-pair measurements (Table 1; Supplemental Table S6).

### Construction of Bayes Nets and Inference of Posterior Functional Relationship Scores

We trained separate naïve Bayes nets using the same gene list and training set, but with species-specific ESC data compendiums, to predict the probability of functional associations among 150,363,811 protein-coding gene pairs based on patterns observed in the integrated evidential data. Learning was achieved by computing the posterior probability of a functional relationship between training set gene pairs given all evidential data [36-39] as previously described (Equation 1) [25].

---

Supporting Information
Supplemental Tables S1 through S10 are available online at http://stemsight.org/stemdata.html#comp_info
**Table S1.** Protein Coding Gene Orthologs
**Table S2.** ESC Self-Renewal Positive Training Set Edges and Supporting Literature References
**Table S3.** Journal Articles Referenced to Curate Positive Training Set Edges
**Table S4.** ESC Self-Renewal Training Set Answer File (Positive and Negative Training Examples)
**Table S5.** hESC Data Compendium
**Table S6.** mESC Data Compendium
**Table S7.** hESC and mESC Network Gene Ontology Annotation Enrichment
**Table S8.** Comparative Network Topology Statistics
**Table S9.** Positive Training Set Posterior Edge Weight and Rank Order Novel Insights into Human Embryonic Stem Cell Self-Renewal
**Table S10.** Differential Network Analysis
**File S1.** Detailed Methods and Materials
In addition, the following data files are available online at www.StemSight.org: mESC Network graph file (~10 million edges; edge weight ≥0.25) and hESC Network ((~16 million edges; edge weight ≥0.25).

$$P(FR|E_1, E_2, \ldots E_n) = \frac{1}{Z} P(FR) \prod_{i=1}^{n} P(E_i|FR) \quad (1)$$

Where *FR* is a binary hidden variable representing whether a gene pair is functionally related, *P(FR=1)* is the predicted probability that a pair is functionally related, $E_i$ represents the evidence score of the gene pair for the $i^{th}$ dataset, and *Z* is a normalization factor.

Naïve Bayes nets impose a strict assumption of independence between evidence data that is likely violated by many of our input datasets, as such, we performed regularization to quantitatively correct for the amount of redundant information contained in each dataset as compared to all other datasets in the compendium (Equations 2 and 3) as previously described [25].

$$S_k = 1 + H(D_k)^{-1} \sum_{i \neq k} I(D_i; D_k) \quad (2)$$

$$P(FR_{i,j}|E_1, E_2, \ldots E_n) = \frac{1}{Z} \prod_{k=1}^{n} \frac{\alpha P[D_k = d_k(g_1 g_2)] + \alpha^{S_k} - 1}{\alpha + |D_k| \alpha^{S_k - 1}} \quad (3)$$

Where $S_k$ is a heuristic sum of shared information relative to the dataset's entropy used to weight the strength of prior belief in a uniform distribution for the dataset, *H* refers to Shannon entropy, and $I(D_i;D_k)$ refers to mutual information. This ultimately results in Equation 3 (a variation on Equation 1), such that $P(FR_{ij}|E_1, E_2,...E_n)$ is the predicted probability that there is a functional relationship between genes *i* and *j* given evidence in datasets *1* through *n*, *Z* is a normalization factor, $\alpha$ is a pseudocount regularization parameter used to modulate the strength of regularization (higher pseudocount values weaken influence of redundant datasets), and $D_k$ is the number of bins used to discretize continuous data values in dataset *K*.

A low $S_k$ indicates that the information contained in the dataset is highly unique, while a high score indicates that the datasets contained shared (redundant) information. The redundancy score for each dataset used to train the Bayesian classifier is listed in Supplemental Tables S5 (hESC) and S6 (mESC). We conducted performance tests to evaluate effects of regularization and selected a pseudocount value of 70 to regularize mESC evidential data, and a value of 100 for hESC evidential data to achieve similar posterior probability distributions.

## Computational Performance Assessment

To assess biological content and functional relevance of our networks, we used functional genomics tools [32, 40] to evaluate Gene Ontology (GO) term enrichment [41], validate that known gene pairs were strongly connected in the probabilistic network, and identify novel genes with strong functional linkages supported by evidential data. These computational validations, including standard machine learning metrics and cross validation, show that our networks are highly accurate and a powerful tool for comparative analysis (Supplemental Figure S1; additional details in Supplemental File S1, Supplemental Table S7).

### Network Topology and Correlation Analyses

To analyze the network topology we calculated the scaled degree ($K_i$) for each gene as previously described [25, 42]. To determine genes strongly associated with self-renewal (rather than the general connectivity measured by degree), we identified a set of "core" self-renewal genes from our positive training set for each species as those genes tightly connected to other training set genes as previously described (173 mESC genes; 111 hESC genes) [25]. Then for every gene in the genome, we calculated the average posterior probability of functional relationship to these "core" self-renewal genes, which we refer to as the self-renewal correlation (SRC) score (Supplemental Table S8). We used these SRC values to rank genes from most strongly correlated to self-renewal (ranks near 1) to least correlated to self-renewal (ranks near 17,342) (Supplemental Table S9). We produced a differential network by subtracting the posterior edge weights for each gene pair (mESC minus hESC probability) to identify pairs strongly associated in one network, but not the other. Spearman correlation coefficients were used to determine the conservation of edge weights between networks (Supplemental Table S10).

### Experimental Validation

H1 hESCs were obtained from WiCell Research Institute (Madison, WI). The hESCs (passages 30-50) were grown on mitotic-inactivated mouse embryonic fibroblast cells (MEFs) in ES Medium (ESM) containing DMEM/F-12 (Invitrogen, Carlsbad, CA), 20% knockout serum replacement (KSR, Invitrogen), 0.1 mM nonessential amino acids (Invitrogen), 2 mM L-glutamine (Mediatech, Inc, Herndon, VA), 0.1 mM beta-mercaptoethanol (Sigma, St. Louis, MO), and 4 ng/ml FGF2 (PeproTech, Rocky Hill, NJ) [43, 44]. Embryoid body (EB) differentiation was carried out as previously described [45]. EBs at day 4 and 6 were harvested for RNA isolation. Quantitative RT-PCR (qPCR) was performed and adjusted to yield equal amplification of glyceraldehyde-3-phosphate dehydrogenase (GAPDH) as an internal standard.

Further Materials and Methods details are provided in Supplemental File S1.

## Results and Discussion

We used a cell-type-specific naïve Bayes network methodology (Figure 1) [25] to create separate probabilistic biological networks for hESCs and mESCs focused on self-renewal and closely related biological processes (*e.g.* pluripotency and cell fate determination). As input data, our Bayes nets take: 1) a training set of prior knowledge (also known as a gold standard) comprised of protein-coding gene pairs known to be functionally related (positive training examples) and pairs believed to be unrelated (negative/background training examples), and 2) independent, whole-genome high-throughput datasets (observed evidential data). Based on these inputs, Bayes nets "learn" significant patterns in the evidence, assess data reliability, and then probabilistically predict novel relationships among protein-coding genes based on the most reliable data [46, 47]. (Additional details in Supplemental Figure S1 and Supplemental File S1.) The resulting networks consisted of 150,363,811 gene pairs that predicted the strength of functional associations among 17,342 protein-coding gene orthologs based on patterns observed in integrated genomic data.

Computational evaluations showed that these networks were highly accurate (Supplemental Figure S1). The topology of both the human and mouse networks show that a subset of genes (roughly 10%, or ~1700) are highly connected, as determined by scaled degree (a measure of global connectivity for each gene), while the majority of genes are less well connected. Further, the connectivity profiles are similar between both the human and mouse networks (Supplemental Figure S2, full results in Supplemental Table S8). We used GO term enrichment and network connectivity metrics to assess biological content and functional relevance of our these networks, validate that gene pairs known to significantly influence mESC self-renewal were strongly connected in these probabilistic networks, and identify novel genes with strong functional linkages supported by evidential data (Supplemental Tables S7-S8).

## hESC Network Predicts Novel Self-Renewal Genes

Functional associations to known self-renewal genes (SRC scores, see Methods and Materials) were calculated for all 17,342 genes in the hESC network and sorted to determine SRC ranks. As expected, many well-studied genes in our training examples are highly ranked by this metric, including *POU5F1*, *SOX2*, and *NODAL* (Supplemental Table S8). The top 12 novel genes (not included in the positive training set) identified by SRC ranks were TUBB, HSP90AB1, PELI1, SEMA6A, SLC7A5, RGMB, PCBP1, TUBB6, HSPA4, SFRP1, LRRN1, and *PRDM14* (Table 2). Literature validation of these genes confirmed that *PRMD14* has recently been shown to play a role in hESC self-renewal and pluripotency [48, 49], and several have been implicated in cancers, including *HSP90AB1, SFRP1, SEMA6A,* and *PRDM14* [50-53]. *TUBB* and *TUBB6* are β-tubulin genes, which could play a role in hESC colony morphology as β-tubulins help define cell shape [54] and could regulate self-renewal through nuclear positioning [55]. Several other highly ranked genes are known to be involved in embryonic developmental signaling pathways: *SFRP1* has been shown to play a key regulatory role in WNT receptor signaling and has been identified as a downstream target of Sonic Hedgehog signaling [56, 57]; *SEMA6* has been associated with non-canonical WNT Receptor Signaling and Planar Cell Polarity signaling [52]; and *RGMB* has been shown to play a regulatory role in BMP signaling [58]. In further support of these predictions, we measured the expression of our top novel genes (excluding β-tubulins) in undifferentiated H1 hESC cell lines, and during differentiation into embryoid bodies. In all cases, expression levels of our novel candidate genes were significantly reduced in the differentiated state compared to the self-renewing state (Supplemental Figure S3).

## Human and Mouse ESC Networks are Largely Similar and Significantly Correlated

To further evaluate our novel hESC network, we performed a comparative analysis with an updated version of our previous mESC network [25], generated using the same master list of protein-coding gene orthologs and training set. Differential network analysis showed that the mESC and hESC networks are highly correlated, and most genes share the same functional partners/interactors, as indicated by the posterior probability of edges (Spearman's rank correlation coefficient $\rho = 0.43$, p-value $< 1\times10^{-300}$; Supplemental Figure 4A). In particular, positive training set edges were significantly correlated between species (Spearman's rank correlation coefficient $\rho = 0.4435$, p-value $= 1.08^{-93}$). Our results confirm the conserved roles of genes known to be involved in early developmental transcriptional regulation and

stem cell maintenance in both species, including *POU5F1*|*Pou5f1* (hESC rank: 1|mESC rank: 1), *SOX2*|*Sox2* (2|5), and *NANOG*|*Nanog* (128|56). The top 1% of genes most conserved in both the mouse and human ESC networks significantly overlap (Hypergeometric p-value= $9.17^{-15}$) and were highly enriched for embryonic-stem-cell-related biological processes, including stem cell maintenance (GO:0010074), negative regulation of cell differentiation (GO:0045596), cell fate commitment (GO:0045165), regulation of transcription, DNA-dependent (GO:0042127), transcription factor activity (GO:0003712) and regulation of gene expression (GO:0010468) using *Mus musculus* (laboratory mouse) GO annotations (Table 3, Supplemental Tables S7 and S8). We observed some species-specific annotation biases [33] in that the same set of genes evaluated using *Homo sapien*s GO annotations indicated enrichment for transcription factor activity, regulation of transcription, and WNT receptor signaling pathway (GO:0016055).

As described in Methods, we identified 173 core self-renewal genes in the mESC network and 111 core self-renewal genes in the hESC network with a significant overlap of 73 shared genes (Supplemental Figure 4B; Supplemental Table S8; Hypergeometric p-value = $8.56^{-7}$). As expected, well-studied genes known to play a significant role in self-renewal and pluripotency across species were ranked highest in both networks. These key players included well-known transcriptional regulators: *POU5F1*|*Pou5f1* (1|1), *SOX2*|*Sox2* (2|5), and *NANOG*|*Nanog* (128|56) (Table 3, Supplemental Table S9). Highly conserved novel genes (not included in training set examples) included *PARP1*|*Parp1* (21|97), a regulator of chromatin structure, transcription, and DNA damage repair that has recently been shown to be required for reprogramming mouse embryonic fibroblasts (MEFs) to a pluripotent state [59, 60]; and *IFITM1*|*Ifitm1* (107|15), an interferon-induced transmembrane protein identified as a downstream target of WNT receptor signaling during gastrulation involved in somitogenesis and mesoderm formation [61].

In our training examples, we included components of KEGG signaling pathways and WikiPathways known to influence early development and ESC cell fate, such as FGF and JAK/STAT signaling [13, 35, 62, 63]. However, as these pathways include many homologous components, it is likely that only a subset of documented pathway participants are active in ESCs, and some interactions may be species specific (e.g. LIF-activated JAK/STAT signaling in mESCs, and NODAL/ACTIVIN-A-activated FGF signaling in hESCs). As such, the bulk of training set genes with low SRCs in both networks consist of signaling pathway participants that appear to be more likely involved in signaling in a cellular context other than ESCs, such as *FGF22*|*Fgf22* (15800|14199), *WNT10A*|*Wnt10a* (12140|10554), and *INHBA*|*Inhba* (12583|14301). Surprisingly, given its role in X-chromosome inactivation [15, 64], *XIST*|*Xist* (13765|11727) ranked poorly in both networks. This could be because, while it is measured by most human and mouse microarray platforms, it may not correlate with other specific gene expression regulatory and protein-DNA binding genes that are strongly associated with self-renewal. While our compendiums contain extensive high-throughput data from a broad range of experimental techniques (Table 1), the bulk of our data for both species is derived primarily from gene expression studies (using microarray or RNA-seq technologies) and from protein-DNA binding data (from ChIP-chip and ChIP-seq assays). We show that these data can be highly informative for elucidating protein function

in a specific cellular context; however, the scarcity of additional high-throughput data types, such as protein-protein-binding assays performed in ESCs, limited our approach to identifying genes related primarily through expression or direct transcriptional regulation.

## Predictive Networks Emphasize Important Signaling Pathway Differences

Although the mESC and hESC networks are highly correlated, suggesting that gene functional associations in the context of ESC self-renewal are largely conserved, there are striking differences (Table 3, Supplemental Table S10). Divergent functional linkages are particularly evident in developmental signaling pathways and our predictive networks recapitulate many known differences in signaling cues that prompt naïve mouse versus primed human ESCs to sustain or exit a self-renewing state. To visualize these differences, we created predictive mESC and hESC developmental signaling pathways using gene elements identified in subsections of KEGG pathways, overlaid with our predicted connection probabilities and functional correlations to self-renewal (SRCs) for those genes (Figures 2 and 3).

For example, FGF signaling has been shown to play an important role in regulating ESC self-renewal and differentiation (as well as myriad other processes) and may be involved in promoting the ESC transition from a naïve to primed pluripotent state [65]. Mice and humans have 22 FGF ligands and 5 FGFR receptors, each with specific expression patterns that change over time during early development, and different ligand-receptor pairs have varying functional roles depending on the cellular context. In mESCs, FGF4-activated ERK signaling promotes differentiation [65], whereas in hESCs, FGF2-activated ERK signaling sustains self-renewal [8, 65]. We compared functional associations among FGF ligand-receptor pairs using our hESC and mESC networks and found our predictions clearly recapitulate known differences between species (Figure 2). The *FGF2*|*Fgf2* (48|14300) ligand was strongly correlated to self-renewal in the hESC network only; while *FGF4*|*Fgf4* (3797|156) and *FGF5/Fgf5* (2965|308) ranked highest in the mESC network only. In both species, these active FGF ligands were strongly linked to the *FGFR1*|*Fgfr1* (126|302) and *FGFR2*|*Fgfr2* (198|379) receptors, while the *FGFR3*|*Fgfr3* (170|6063) and *FGFR4*|*Fgfr4* (1191|7560) receptors were strongly linked only in the human network. In addition, both our hESC and mESC networks showed high-ranking SRCs for the *FGF19*|*Fgf15* (144|622) ligand, though in humans it was associated with the *FGFR1*|*Fgfr1* (126|302) and *FGFR3*|*Fgfr3* (170|6063) receptors, while in mice this ligand was associated primarily with *FGFR2*|*Fgfr2* (198|379). In both species, *FGF19*|*Fgf15* is expressed in the fetal brain and mediates differentiation to the neuroectoderm [66, 67].

We also examined LIF-activated JAK/STAT signaling, as LIF promotes self-renewal by activating the JAK/STAT3 and PI3K/AKT signaling pathways in mESCs, but is not required by hESCs to sustain self-renewal [13, 62]. While *STAT3*|*Stat3* (202|485) is strongly associated with self-renewal in both species in our networks, it appears that the mode of regulation may be diverged. This difference is clearly supported in our probabilistic pathways focused on JAK/STAT signaling stimulated by the Interleukin-6 (IL-6) family of cytokines (Figure 3) [68]. In the mESC network, a strong connection exists between the *LIF*|*Lif* (3457|1317) cytokine and its well-characterized signal transducer *IL6ST*|*Il6st* (also

known as *gp130*; 9107|347), which in turn is tightly linked to tyrosine kinase *JAK3*|*Jak3* (10600|3126) and then on to *STAT3*|*Stat3* (202|485). These predicted linkages reflect the most well-documented LIF-activated IL-6 JAK/STAT signaling cascade in mESCs [69-71]; however these strong associations are not present in our hESC network. Our mESC results also showed functional linkages between *LIF*|*Lif* and the *IFNGR2*|*Ifngr2* (5801|946) receptor, which associates most strongly with *JAK1*|*Jak1* (2789|3808) and *JAK2*|*Jak2* (6625|1989), both of which strongly link to *STAT3*|*Stat3*.

In contrast, within the confines of this JAK/STAT pathway subset, the hESC network predicts that *LIF* and the many other IL-6 cytokines are not strongly associated with self-renewal (based on SRC scores), and shows only weak connections between *LIF* and the *Il6ST* signal transducer [72]. Rather, in hESCs, the cytokines *IL11*|*Il11* (2939|15541) and *LIF*|*Lif* were moderately associated with the *IFNGR1*|*Ifngr1* (2701|6655) and *OSMR*|*Osmr* (2060|8138) receptors, which were moderately linked to the *JAK1*|*Jak1* (2789|3801) and *TYK2*|*Tyk2* (11394|8747) tyrosine kinases, respectively (note that while *Tyk2* is annotated in the mouse JAK/STAT pathway by KEGG, it is not listed as a participant in the human KEGG pathway). Global expression profiling in mESCs has shown that the interferon gamma receptor *IFNGR1*|*Ifngr1* is regulated by *POU5F1*|*Pou5f1* [73], and although not documented as active in early embryonic development, *OSMR*|*Osmr* is predicted to have a strong functional association with *POU5F1*/*Pou5f1* supported by both gene expression and protein-DNA-interaction data in the human network. In contrast to these moderate associations with *STAT3,* our hESC network predicts strong *STAT3* associations with members of other signaling pathways, including *TCF7L1*, *FGF2*, *FGFR1*, and *BMPR2*, suggesting potentially significant cross-talk and/or alternate modes of *STAT3* regulation. However, these strong associations are not observed for *Stat3* in our mESC network, where *Stat3* is most strongly associated with known LIF and JAK/STAT signaling pathway members, such as *Fos*, *Pim1*, and *Spry4*.

## Metabolic Differences Between Species Highlighted in Predictive Networks

One of the most striking differences between our mouse and human ESC networks concerns threonine catabolism, which is required for mESC self-renewal, likely through the *TDH*|*Tdh* (L-threonine dehydrogenase) gene, which supports accelerated cell cycle kinetics by catabolizing threonine into glycine and acetyl-CoA, which is used by the TCA cycle to generate ATP [74, 75]. In our mouse network, *TDH*|*Tdh* (11393|26) has a strong correlation to mESC self-renewal and is tightly linked to many core self-renewal genes, including *POU5F1*|*Pou5f1* (1|1), *SOX2*|*Sox2* (2|5), *RIF1*|*Rif1* (14|33), and *NR0B1*|*Nr0b1* (5583|487). The functional relationship between *Tdh* and these genes is largely supported by ChIP-chip binding data from studies investigating the regulatory circuitry of mESCs and microarray data from a study analyzing mESC differentiation (Figure 4B). In contrast, *TDH* is not correlated to self-renewal in our hESC network and not tightly connected to any genes in our positive training set (Figure 4A). Differential and functional analyses of the mESC and hESC networks did not reveal a direct metabolic equivalent to threonine dehydrogenase in humans. Literature validation showed that the human *TDH* gene encoding threonine dehydrogenase has been rendered non-functional due to 3 mutations (2 AG-to-GG splice acceptor mutations in Exons 4 and 6, and a nonsense mutation in Exon 6) [74]. However,

hESCs grow at a slower rate than mESCs (with a doubling time of 35 hours as opposed to every 4-5 hours), and it is not yet known if the difference in growth rate might be due to the absence of *TDH* or if there may be some selective advantage for inactivating *TDH* in humans [74-76].

## Comparative Network Analyses Reveal Novel Species-Specific Differences

To discover novel species-specific differences, we focused on the top 0.001% (1503) of gene pairs with the greatest difference between the mESC versus hESC networks (Supplemental Table S10). There were 86 genes involved in edges strongly supported only in the mESC network, including genes annotated to chordate embryonic development (GO:0043009), stem cell maintenance (GO:0019827), negative regulation of cell differentiation (GO:0045596), and regulation of transcription, DNA-dependent (GO:0042127). In contrast, 179 genes were involved in edges supported only by the hESC network, including genes annotated to WNT Receptor Signaling (GO:0016055), MAPK Cascade (GO:0000165), FGF Signaling (GO:0008543), ATP-Binding (GO:0005524), and cell cycle regulation (GO:0051726), emphasizing the different set of developmental signaling cues observed in early cell fate decisions in hESCs.

These key differences were echoed when we compared the top 1% of novel genes ranked by SRC (the 173 highest ranking genes not included in our positive training set examples). Top-ranked novel mESC genes were predominantly associated with differentiation and transcriptional regulation processes, whereas the novel hESC genes were largely related to regulation of translation (GO:0006417), microtubule-based process (GO:0007017), regulation of cell cycle (GO:0051726), cell division (GO:0051301), and cell adhesion (GO:0007155). Although there are known differences in cell cycle controls that permit rapid cell cycling in ESCs, it is not yet known whether the cell cycle regulates pluripotency. Cell cycle length is not believed to determine pluripotency, because different pluripotent cell types divide at different rates [77]; however, the enrichment of cell cycle component genes in our novel hESC gene lists suggest that there may be functional differences in our mouse and human networks that can be mined to improve our understanding of how the core cell cycle machinery adapts to the timing requirements of primed and naïve ESCs.

Many of the genes most strongly correlated to self-renewal in the hESC network only (Table 3) are involved in translational regulation and protein synthesis, such as the ribosomal proteins *RPL10A*|*Rpl10a* (37|8606), *RPLP2*|*Rplp2* (38|7284), *RPS3A*|*Rps3a* (65|11303), and *RPL6*|*Rpl6* (101|11695), as well as the translation initiation factor *EIF3A*|*Eif3a* (115|12160). Intriguingly, genome-wide studies have shown that *LIN28*|*Lin28a* (62|224), a known regulator of self-renewal and an RNA-binding protein [78], targets RNAs in hESCs including ribosomal and translation-supporting genes, which are important for growth and survival [79]. In contrast, *Lin28a* targets in mESCs are enriched for translational repressors of membrane-bound proteins, secretory proteins, and proteins destined for the endoplasmic reticulum/Golgi lumen [80], suggesting alternative regulatory roles for *LIN28*|*Lin28a*, despite supporting self-renewal in both species.

### Comparative Network Visualization Promotes Novel Gene Discovery

To make our probabilistic comparative networks readily available to the stem cell research community, we provide interactive, online visualization resources at www.StemSight.org that can be used to view our hESC and mESC networks independently or comparatively. StemSight Scout enables users to explore subnetworks centered on user-provided genes of interest, and highlights potentially novel self-renewal genes by coloring nodes based on their SRC score and illustrates the weight of predicted interactions by coloring edges based on inferred posterior probabilities. Edges included in our training set examples are color-coded, making it easy to visually segregate novel from known. For positive training set edges, links are provided to the original articles documenting the relationship, while for novel edges, the evidential data supporting that functional relationship is displayed with links to the studies in PubMed. The comparative view enables users to search for interactions among one or more genes in both networks simultaneously, and displays results in a convenient side-by-side window for direct comparisons (Figure 4), which is especially helpful for contrasting connectivity of gene hubs and pathway participants in the hESC versus mESC networks.

## Conclusions

Determining the molecular underpinnings of stem cell self-renewal and cell fate determination is vital for understanding and refining the production of iPS cells, and has potential implications for targeting cancer stem cells and aging processes in adult stem cells. Our results confirm that multiple signaling pathways contribute to the balance between self-renewal and differentiation in ESCs, and further suggest that these complex pathways should not be considered monolithically, but recognized as general mechanisms whose details and interactions are only partially understood. For example, while our results indicate that *STAT3/Stat3* is strongly associated with self-renewal in both humans and mice, upstream regulation of signaling appears significantly diverged between species. Similar observations and conclusions can be drawn from our results for various members of WNT, NODAL/ACTIVIN-A, FGF, and TGF-β signaling, as well as many other signaling and metabolic pathways.

Our computational integration of hundreds of high-throughput datasets provides an entry point to the greater understanding of these pathways. For example, our evaluation of *STAT3* interactions suggests potentially significant crosstalk and alternate modes of human *STAT3* regulation. As *STAT3|Stat3* is a multi-functional gene implicated in phenotypes and disorders ranging from circadian rhythms to cancer, developing cancer therapies targeting the STAT3 pathway requires investigation of potential off-target effects, which in turn requires expanding our understanding of its context-specific regulation. Thus, unraveling the diversity of systems-level interactions and crosstalk is vital for the use of model organisms, and is likely key for dissecting the cell type-specific roles of multi-functional genes within humans. Similarly, the divergent role of *TDH|Tdh* between species in our networks likely reflects the distinct metabolic mechanisms of these cell types. While *Tdh* supports increased growth rates in mESCs, it is possible that the increased activity of translational regulatory genes in hESCs (such as the ribosomal and translation initiation genes identified only in our human networks) play a similar role in a human context.

By focusing here on human and mouse ESC self-renewal, we have developed methods for cross-species and cross-context comparisons to assist stem cell researchers in evaluating known and novel genes involved in ESC self-renewal and differentiation. In the future, comparative networks for additional stem cell types, including mEpiSCs, iPSCs, ESCs from other species, as well as adult stem cells, could be used to further elucidate the mechanisms of self-renewal in general. Primed mEpiSCs would be particularly useful for investigating differences between the naïve and primed pluripotent state within mouse (i.e. naïve mESCs versus primed mEpiSCs), as well as for comparing primed pluripotent cells in different species (i.e. primed mEpiSCs versus primed hESCs).

While computational systems biology efforts such as ours begin to scrutinize complex molecular interactions, additional laboratory efforts are required to confirm these hypotheses and to determine precise mechanisms. The data generated by these efforts can be re-incorporated into machine learning efforts to refine our training sets, provide additional input evidential data, and ultimately improve our understanding of context-specific, whole-genome interactomes. As such, the full results of this study are freely available to the stem cell community at www.StemSight.org, where users can explore our hESC and mESC self-renewal networks to place their results into a broader context, draw new hypotheses, or prioritize candidate genes.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgments

This project was partially funded by The Jackson Laboratory. KGD is supported by a Pharmaceutical Researchers and Manufacturers of America (PhRMA) Foundation Fellowship in Informatics, and has received funding from the University of Maine Graduate School of Biomedical Sciences, and the National Science Foundation Integrated Graduate Education and Research Traineeship (IGERT) program grant 0221625 to conduct this work. MAH is an Ellison Medical Foundation New Scholar in Aging; he is partially supported by National Institutes of Health / National Institute of Arthritis and Musculoskeletal and Skin (NIH/NIAMS) grant R21 AR069981-01 (Hibbs, Ackert-Bicknell), and National Institutes of Health (NIH) Grant P50 GM076468-06 (Churchill).

## REFERENCES

1. Thomson JA, Itskovitz-Eldor J, Shapiro SS, Waknitz MA, Swiergiel JJ, Marshall VS, Jones JM. Embryonic stem cell lines derived from human blastocysts. Science. 1998; 282(5391):1145–1147. [PubMed: 9804556]
2. Gokhale PJ, Andrews PW. The development of pluripotent stem cells. Current opinion in genetics & development. 2012; 22(5):403–408. [PubMed: 22868175]
3. Ginis I, Luo Y, Miura T, Thies S, Brandenberger R, Gerecht-Nir S, Amit M, Hoke A, Carpenter MK, Itskovitz-Eldor J, et al. Differences between human and mouse embryonic stem cells. Developmental biology. 2004; 269(2):360–380. [PubMed: 15110706]
4. Brown S, Teo A, Pauklin S, Hannan N, Cho CH, Lim B, Vardy L, Dunn NR, Trotter M, Pedersen R, et al. Activin/Nodal signaling controls divergent transcriptional networks in human embryonic stem cells and in endoderm progenitors. Stem Cells. 2011; 29(8):1176–1185. [PubMed: 21630377]
5. Welling M, Geijsen N. Uncovering the true identity of naive pluripotent stem cells. Trends in cell biology. 2013

6. Martin GR. Isolation of a pluripotent cell line from early mouse embryos cultured in medium conditioned by teratocarcinoma stem cells. Proceedings of the National Academy of Sciences of the United States of America. 1981; 78(12):7634–7638. [PubMed: 6950406]
7. Evans MJ, Kaufman MH. Establishment in culture of pluripotential cells from mouse embryos. Nature. 1981; 292(5819):154–156. [PubMed: 7242681]
8. Greber B, Wu G, Bernemann C, Joo JY, Han DW, Ko K, Tapia N, Sabour D, Sterneckert J, Tesar P, et al. Conserved and divergent roles of FGF signaling in mouse epiblast stem cells and human embryonic stem cells. Cell Stem Cell. 2010; 6(3):215–226. [PubMed: 20207225]
9. Saunders A, Faiola F, Wang J. Pursuing Self-Renewal and Pluripotency with the Stem Cell Factor Nanog. Stem Cells. 2013
10. Nichols J, Smith A. The origin and identity of embryonic stem cells. Development. 2011; 138(1): 3–8. [PubMed: 21138972]
11. Nichols J, Smith A. Naive and primed pluripotent states. Cell Stem Cell. 2009; 4(6):487–492. [PubMed: 19497275]
12. Lanza RP. Essentials of stem cell biology. 2006; xxxi:548.
13. De Los Angeles A, Loh YH, Tesar PJ, Daley GQ. Accessing naive human pluripotency. Current opinion in genetics & development. 2012; 22(3):272–282. [PubMed: 22463982]
14. Tesar PJ, Chenoweth JG, Brook FA, Davies TJ, Evans EP, Mack DL, Gardner RL, McKay RD. New cell lines from mouse epiblast share defining features with human embryonic stem cells. Nature. 2007; 448(7150):196–199. [PubMed: 17597760]
15. Tomoda K, Takahashi K, Leung K, Okada A, Narita M, Yamada NA, Eilertson KE, Tsang P, Baba S, White MP, et al. Derivation conditions impact X-inactivation status in female human induced pluripotent stem cells. Cell Stem Cell. 2012; 11(1):91–99. [PubMed: 22770243]
16. Nazor KL, Altun G, Lynch C, Tran H, Harness JV, Slavin I, Garitaonandia I, Muller FJ, Wang YC, Boscolo FS, et al. Recurrent variations in DNA methylation in human pluripotent stem cells and their differentiated derivatives. Cell Stem Cell. 2012; 10(5):620–634. [PubMed: 22560082]
17. Chenoweth JG, McKay RD, Tesar PJ. Epiblast stem cells contribute new insight into pluripotency and gastrulation. Development, growth & differentiation. 2010; 52(3):293–301.
18. Tachibana M, Sparman M, Ramsey C, Ma H, Lee HS, Penedo MC, Mitalipov S. Generation of chimeric rhesus monkeys. Cell. 2012; 148(1-2):285–295. [PubMed: 22225614]
19. Takahashi K, Yamanaka S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. Cell. 2006; 126(4):663–676. [PubMed: 16904174]
20. Ema M, Mori D, Niwa H, Hasegawa Y, Yamanaka Y, Hitoshi S, Mimura J, Kawabe Y, Hosoya T, Morita M, et al. Kruppel-like factor 5 is essential for blastocyst development and the normal self-renewal of mouse ESCs. Cell Stem Cell. 2008; 3(5):555–567. [PubMed: 18983969]
21. Buganim Y, Faddah DA, Jaenisch R. Mechanisms and models of somatic cell reprogramming. Nature reviews Genetics. 2013; 14(6):427–439.
22. Meissner A. Epigenetic modifications in pluripotent and differentiated cells. Nature biotechnology. 2010; 28(10):1079–1088.
23. Jaenisch R, Young R. Stem cells, the molecular circuitry of pluripotency and nuclear reprogramming. Cell. 2008; 132(4):567–582. [PubMed: 18295576]
24. Chambers I, Tomlinson SR. The transcriptional foundation of pluripotency. Development. 2009; 136(14):2311–2322. [PubMed: 19542351]
25. Dowell KG, Simons AK, Wang ZZ, Yun K, Hibbs MA. Cell-type-specific predictive network yields novel insights into mouse embryonic stem cell self-renewal and cell fate. PLoS One. 2013; 8(2):e56810. [PubMed: 23468881]
26. Huttenhower C, Hibbs MA, Myers CL, Caudy AA, Hess DC, Troyanskaya OG. The impact of incomplete knowledge on evaluation: an experimental benchmark for protein function prediction. Bioinformatics. 2009; 25(18):2404–2410. [PubMed: 19561015]
27. Hibbs MA. Advanced Bioinformatics Tools and Strategies. Principles and Practices of Plant Genomics. 2010; 3 Advanced Genomics.
28. Russell, S.; Norvig, P. Artifical Intelligence: A Modern Approach. 3rd edn.. Prentice Hall; Fort Collins, CO: 2009.

29. Charniak E. Bayesian Networks without Tears. AI Magazine. 1991; 91:50–63. Winter.
30. Darwiche, A. Modeling and reasoning with Bayesian networks. Cambridge University Press; Cambridge ; New York: 2009.
31. Hibbs MA, Myers CL, Huttenhower C, Hess DC, Li K, Caudy AA, Troyanskaya OG. Directing experimental biology: a case study in mitochondrial biogenesis. PLoS Comput Biol. 2009; 5(3):e1000322. [PubMed: 19300515]
32. Bult CJ, Kadin JA, Richardson JE, Blake JA, Eppig JT. The Mouse Genome Database: enhancements and updates. Nucleic Acids Res. 2010; 38(Database issue):D586–592. [PubMed: 19864252]
33. Dolan ME, Ni L, Camon E, Blake JA. A procedure for assessing GO annotation consistency. Bioinformatics. 2005; 21(Suppl 1):i136–143. [PubMed: 15961450]
34. Kanehisa M. The KEGG database. Novartis Found Symp. 2002; 247:91–101. discussion 101-103, 119-128, 244-152. [PubMed: 12539951]
35. Kelder T, van Iersel MP, Hanspers K, Kutmon M, Conklin BR, Evelo CT, Pico AR. WikiPathways: building research communities on biological pathways. Nucleic acids research. 2012; 40(Database issue):D1301–1307. [PubMed: 22096230]
36. Huttenhower C, Schroeder M, Chikina MD, Troyanskaya OG. The Sleipnir library for computational functional genomics. Bioinformatics. 2008; 24(13):1559–1561. [PubMed: 18499696]
37. Guan Y, Myers CL, Lu R, Lemischka IR, Bult CJ, Troyanskaya OG. A genomewide functional network for the laboratory mouse. PLoS Comput Biol. 2008; 4(9):e1000165. [PubMed: 18818725]
38. Myers CL, Barrett DR, Hibbs MA, Huttenhower C, Troyanskaya OG. Finding function: evaluation methods for functional genomic data. BMC Genomics. 2006; 7:187. [PubMed: 16869964]
39. Huttenhower C, Haley EM, Hibbs MA, Dumeaux V, Barrett DR, Coller HA, Troyanskaya OG. Exploring the human genome with functional maps. Genome Res. 2009; 19(6):1093–1106. [PubMed: 19246570]
40. Dennis G Jr. Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA. DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biol. 2003; 4(5):P3. [PubMed: 12734009]
41. The Gene Ontology in 2010: extensions and refinements. Nucleic Acids Res. 201038(Database issue):D331–335.
42. Horvath S, Dong J. Geometric interpretation of gene coexpression network analysis. PLoS Comput Biol. 2008; 4(8):e1000117. [PubMed: 18704157]
43. Bai H, Chen K, Gao YX, Arzigian M, Xie YL, Malcosky C, Yang YG, Wu WS, Wang ZZ. Bcl-xL enhances single-cell survival and expansion of human embryonic stem cells without affecting self-renewal. Stem Cell Res. 2012; 8(1):26–37. [PubMed: 22099018]
44. Bai H, Gao Y, Arzigian M, Wojchowski DM, Wu WS, Wang ZZ. BMP4 regulates vascular progenitor development in human embryonic stem cells through a Smad-dependent pathway. Journal of cellular biochemistry. 2010; 109(2):363–374. [PubMed: 19950207]
45. Bai H, Xie YL, Gao YX, Cheng T, Wang ZZ. The Balance of Positive and Negative Effects of TGF- beta Signaling Regulates the Development of Hematopoietic and Endothelial Progenitors in Human Pluripotent Stem Cells. Stem cells and development. 2013
46. Myers CL, Robson D, Wible A, Hibbs MA, Chiriac C, Theesfeld CL, Dolinski K, Troyanskaya OG. Discovery of biological networks from diverse functional genomic data. Genome Biol. 2005; 6(13):R114. [PubMed: 16420673]
47. Troyanskaya OG, Dolinski K, Owen AB, Altman RB, Botstein D. A Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae). Proc Natl Acad Sci U S A. 2003; 100(14):8348–8353. [PubMed: 12826619]
48. Tsuneyoshi N, Sumi T, Onda H, Nojima H, Nakatsuji N, Suemori H. PRDM14 suppresses expression of differentiation marker genes in human embryonic stem cells. Biochemical and biophysical research communications. 2008; 367(4):899–905. [PubMed: 18194669]
49. Chan YS, Goke J, Lu X, Venkatesan N, Feng B, Su IH, Ng HH. A PRC2-dependent repressive role of PRDM14 in human embryonic stem cells and induced pluripotent stem cell reprogramming. Stem Cells. 2012; 31(4):682–692. [PubMed: 23280602]

50. Cheng Q, Chang JT, Geradts J, Neckers LM, Haystead T, Spector NL, Lyerly HK. Amplification and high-level expression of heat shock protein 90 marks aggressive phenotypes of human epidermal growth factor receptor 2 negative breast cancer. Breast Cancer Res. 2012; 14(2):R62. [PubMed: 22510516]
51. Ren J, Wang R, Huang G, Song H, Chen Y, Chen L. sFRP1 Inhibits Epithelial-Mesenchymal Transition in A549 Human Lung Adenocarcinoma Cell Line. Cancer Biother Radiopharm. 2013
52. Katoh M. Comparative integromics on non-canonical WNT or planar cell polarity signaling molecules: transcriptional mechanism of PTK7 in colorectal cancer and that of SEMA6A in undifferentiated ES cells. Int J Mol Med. 2007; 20(3):405–409. [PubMed: 17671748]
53. Zhang T, Meng L, Dong W, Shen H, Zhang S, Liu Q, Du J. High expression of PRDM14 correlates with cell differentiation and is a novel prognostic marker in resected non-small cell lung cancer. Med Oncol. 2013; 30(3):605. [PubMed: 23690269]
54. Lodish HF. Molecular cell biology. 2008; 1 v. various pagings.
55. Gundersen GG, Worman HJ. Nuclear positioning. Cell. 2013; 152(6):1376–1389. [PubMed: 23498944]
56. Kele J, Andersson ER, Villaescusa JC, Cajanek L, Parish CL, Bonilla S, Toledo EM, Bryja V, Rubin JS, Shimono A, et al. SFRP1 and SFRP2 dose-dependently regulate midbrain dopamine neuron development in vivo and in embryonic stem cells. Stem Cells. 2012; 30(5):865–875. [PubMed: 22290867]
57. Shahi MH, Rey JA, Castresana JS. The sonic hedgehog-GLI1 signaling pathway in brain tumor development. Expert Opin Ther Targets. 2012; 16(12):1227–1238. [PubMed: 22992192]
58. Wu Q, Sun CC, Lin HY, Babitt JL. Repulsive guidance molecule (RGM) family proteins exhibit differential binding kinetics for bone morphogenetic proteins (BMPs). PLoS One. 2012; 7(9):e46307. [PubMed: 23029472]
59. Doege CA, Inoue K, Yamashita T, Rhee DB, Travis S, Fujita R, Guarnieri P, Bhagat G, Vanti WB, Shih A, et al. Early-stage epigenetic modification during somatic cell reprogramming by Parp1 and Tet2. Nature. 2012; 488(7413):652–655. [PubMed: 22902501]
60. Krishnakumar R, Kraus WL. The PARP side of the nucleus: molecular actions, physiological outcomes, and clinical targets. Molecular cell. 2010; 39(1):8–24. [PubMed: 20603072]
61. Lickert H, Cox B, Wehrle C, Taketo MM, Kemler R, Rossant J. Dissecting Wnt/beta-catenin signaling during gastrulation using RNA interference in mouse embryos. Development. 2005; 132(11):2599–2609. [PubMed: 15857914]
62. He S, Nakada D, Morrison SJ. Mechanisms of stem cell self-renewal. Annual review of cell and developmental biology. 2009; 25:377–406.
63. Vallier L, Touboul T, Brown S, Cho C, Bilican B, Alexander M, Cedervall J, Chandran S, Ahrlund-Richter L, Weber A, et al. Signaling pathways controlling pluripotency and early cell fate decisions of human induced pluripotent stem cells. Stem Cells. 2009; 27(11):2655–2666. [PubMed: 19688839]
64. Minkovsky A, Barakat TS, Sellami N, Chin MH, Gunhanlar N, Gribnau J, Plath K. The pluripotency factor-bound intron 1 of Xist is dispensable for X chromosome inactivation and reactivation in vitro and in vivo. Cell Rep. 2013; 3(3):905–918. [PubMed: 23523354]
65. Lanner F, Rossant J. The role of FGF/Erk signaling in pluripotent cells. Development. 2010; 137(20):3351–3360. [PubMed: 20876656]
66. Nishimura T, Utsunomiya Y, Hoshikawa M, Ohuchi H, Itoh N. Structure and expression of a novel human FGF, FGF-19, expressed in the fetal brain. Biochim Biophys Acta. 1999; 1444(1):148–151. [PubMed: 9931477]
67. Borello U, Cobos I, Long JE, McWhirter JR, Murre C, Rubenstein JL. FGF15 promotes neurogenesis and opposes FGF8 function during neocortical development. Neural Dev. 2008; 3:17. [PubMed: 18625063]
68. Kristensen DM, Kalisz M, Nielsen JH. Cytokine signalling in embryonic stem cells. Apmis. 2005; 113(11-12):756–772. [PubMed: 16480448]
69. Okita K, Yamanaka S. Intracellular signaling pathways regulating pluripotency of embryonic stem cells. Current stem cell research & therapy. 2006; 1(1):103–111. [PubMed: 18220859]

70. Ying QL, Wray J, Nichols J, Batlle-Morera L, Doble B, Woodgett J, Cohen P, Smith A. The ground state of embryonic stem cell self-renewal. Nature. 2008; 453(7194):519–523. [PubMed: 18497825]

71. Niwa H, Burdon T, Chambers I, Smith A. Self-renewal of pluripotent embryonic stem cells is mediated via activation of STAT3. Genes & development. 1998; 12(13):2048–2060. [PubMed: 9649508]

72. Liu N, Lu M, Tian X, Han Z. Molecular mechanisms involved in self-renewal and pluripotency of embryonic stem cells. J Cell Physiol. 2007; 211(2):279–286. [PubMed: 17195167]

73. Matoba R, Niwa H, Masui S, Ohtsuka S, Carter MG, Sharov AA, Ko MS. Dissecting Oct3/4-regulated gene networks in embryonic stem cells by expression profiling. PLoS One. 2006; 1:e26. [PubMed: 17183653]

74. Wang J, Alexander P, Wu L, Hammer R, Cleaver O, McKnight SL. Dependence of mouse embryonic stem cells on threonine catabolism. Science. 2009; 325(5939):435–439. [PubMed: 19589965]

75. Shyh-Chang N, Locasale JW, Lyssiotis CA, Zheng Y, Teo RY, Ratanasirintrawoot S, Zhang J, Onder T, Unternaehrer JJ, Zhu H, et al. Influence of threonine metabolism on S-adenosylmethionine and histone methylation. Science. 2013; 339(6116):222–226. [PubMed: 23118012]

76. Wang J, Alexander P, McKnight SL. Metabolic specialization of mouse embryonic stem cells. Cold Spring Harbor symposia on quantitative biology. 2011; 76:183–193.

77. Hindley C, Philpott A. The cell cycle and pluripotency. The Biochemical journal. 2013; 451(2): 135–143. [PubMed: 23535166]

78. Shyh-Chang N, Daley GQ. Lin28: primal regulator of growth and metabolism in stem cells. Cell Stem Cell. 2013; 12(4):395–406. [PubMed: 23561442]

79. Peng S, Chen LL, Lei XX, Yang L, Lin H, Carmichael GG, Huang Y. Genome-wide studies reveal that Lin28 enhances the translation of genes important for growth and survival of human embryonic stem cells. Stem Cells. 2011; 29(3):496–504. [PubMed: 21425412]

80. Cho J, Chang H, Kwon SC, Kim B, Kim Y, Choe J, Ha M, Kim YK, Kim VN. LIN28A is a suppressor of ER-associated translation in embryonic stem cells. Cell. 2012; 151(4):765–777. [PubMed: 23102813]

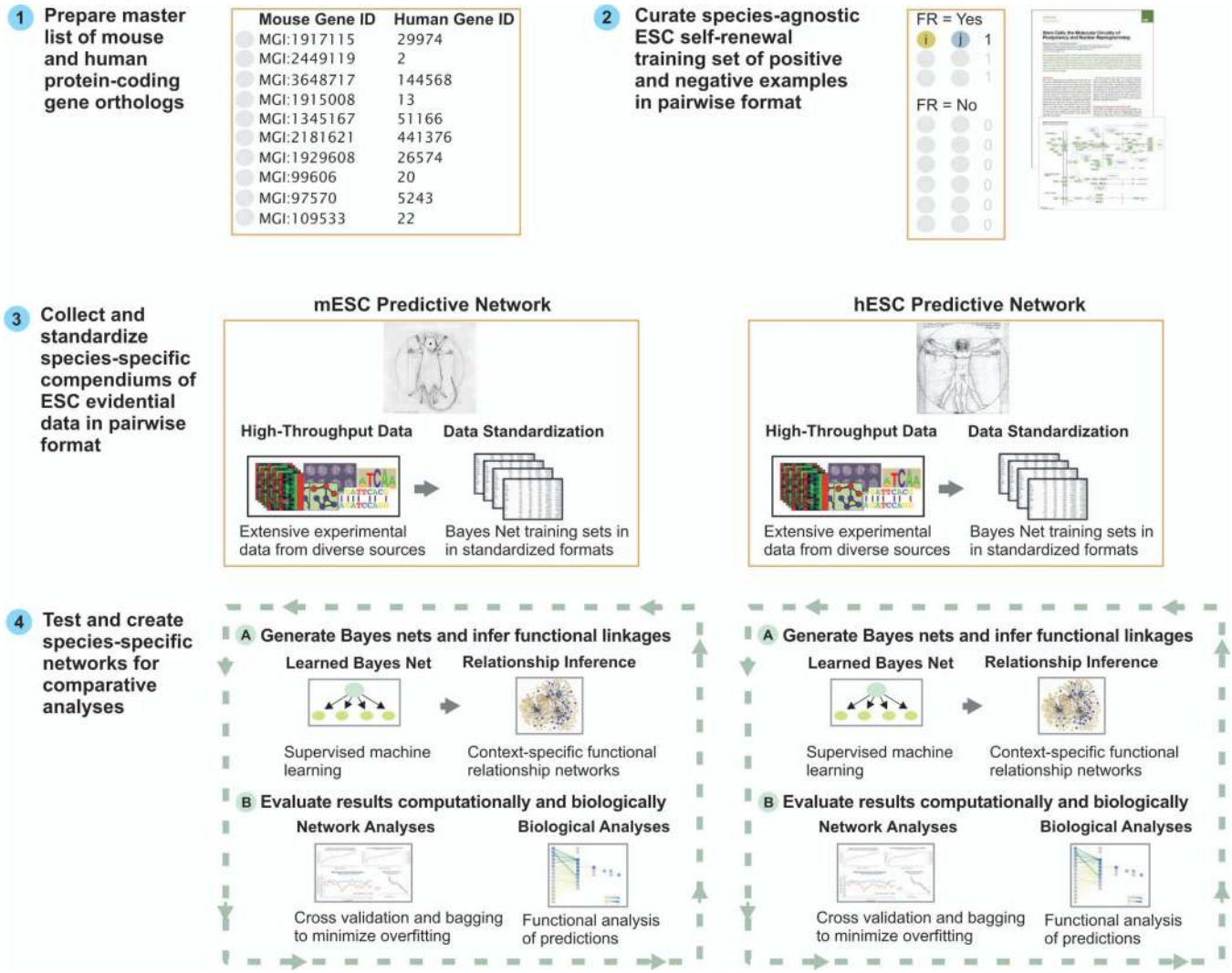


**Figure 1. Bayes Net Machine Learning Methodology for Cell-Type-Specific Comparative Networks**

To produce reliable, biologically relevant comparative predictive networks for mouse and human ESCs, we adapted our approach for cell-type-specific data integration and machine learning [25] as follows: **1.** Prepared a master list of protein-coding mouse and human gene orthologs. **2.** Curated a set of training examples focused on ESC self-renewal in a pairwise format. Positive (known) examples were extracted from recent literature and curated pathways; negative/background examples were randomly generated from the master gene list, excluding genes involved in positive edges. **3.** Collected and standardized high-throughput data from 83 human and 62 mouse ESC studies using diverse high-throughput data types (Table 1). Used distance/correlation metrics to distill data into a pairwise format. **4.** Iteratively tested and validated species-specific predictive networks for comparative analysis using the same training set, but with species-specific data compendiums as input.

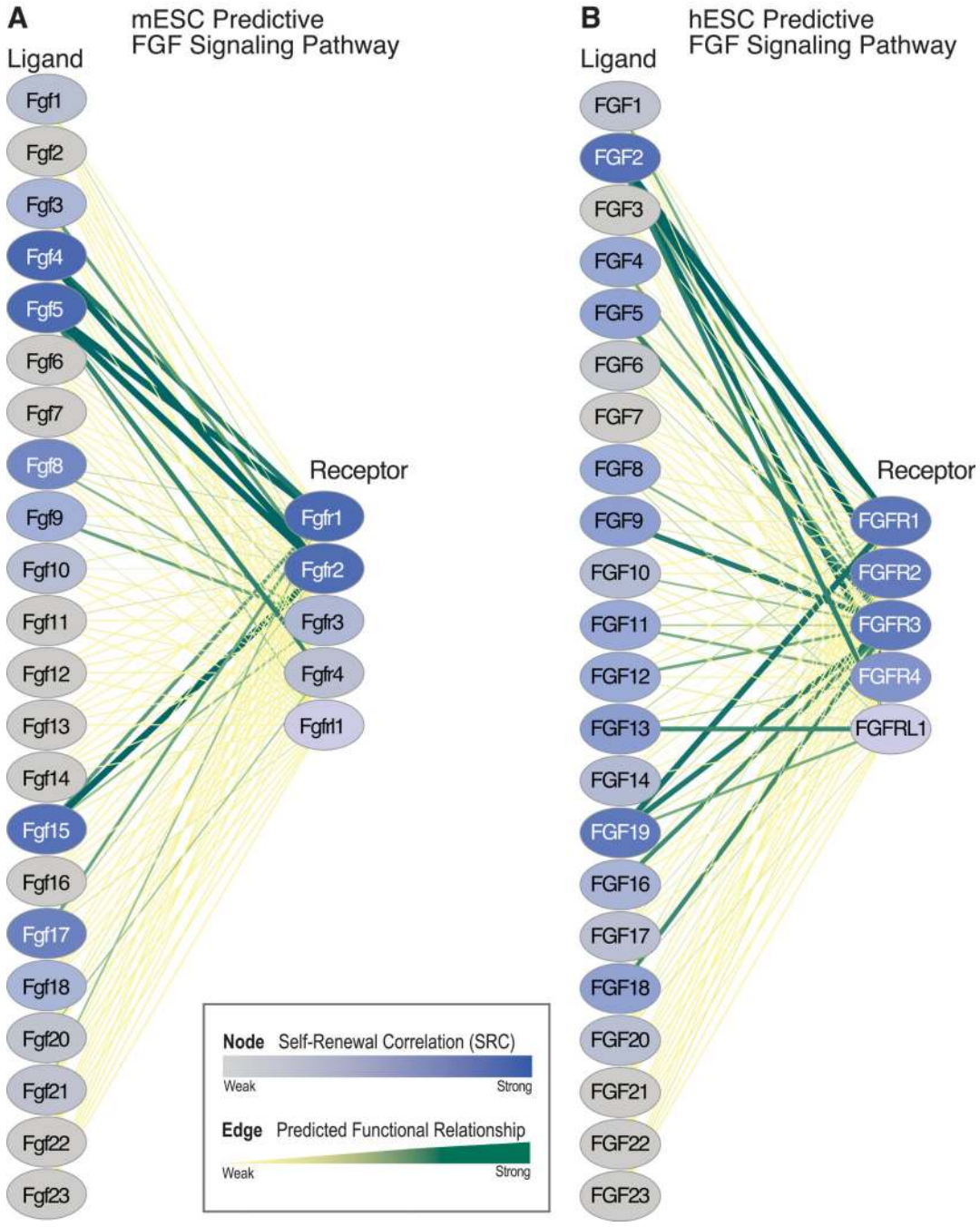


**Figure 2. Predicted FGF Signaling Pathway Relationships Across Species**

These network models show the predicted strength of relationships between all known FGF signaling ligand-receptor pairs [65]. Gene nodes are colored by SRC score (light gray = weak SRC; dark blue = strong SRC) and edge color/thickness indicates the strength of predicted functional association between ligand and receptor (yellow/thin = low probability; teal/thick = high probability). **A.** The hESC FGF signaling model shows that the most strongly connected ligand-receptor pairs with high-ranking SRCs are *FGF2* and *FGF19* (the human ortholog of *Fgf15*), which are most strongly associated with *FGFR1* and *FGFR3*.

*FGF2* is known to activate FGF signaling in hESCs, while *FGF19* has been associated with neuronal development. **B.** The mESC FGF signaling model showed that the most strongly connected ligand-receptor pairs with high ranking SRCs are *Fgf4* and *Fgf5*, which are both associated with *Fgfr1* and *Fgfr2*, and *Fgf15* which is most strongly connected to *Fgfr2*. In mESCs, *Fgf4* is known to activate FGF signaling, *Fgf5* is associated with FGF signaling in the late stage blastocyst and epiblast, and *Fgf15* has been shown to be involved in early neurodevelopment.

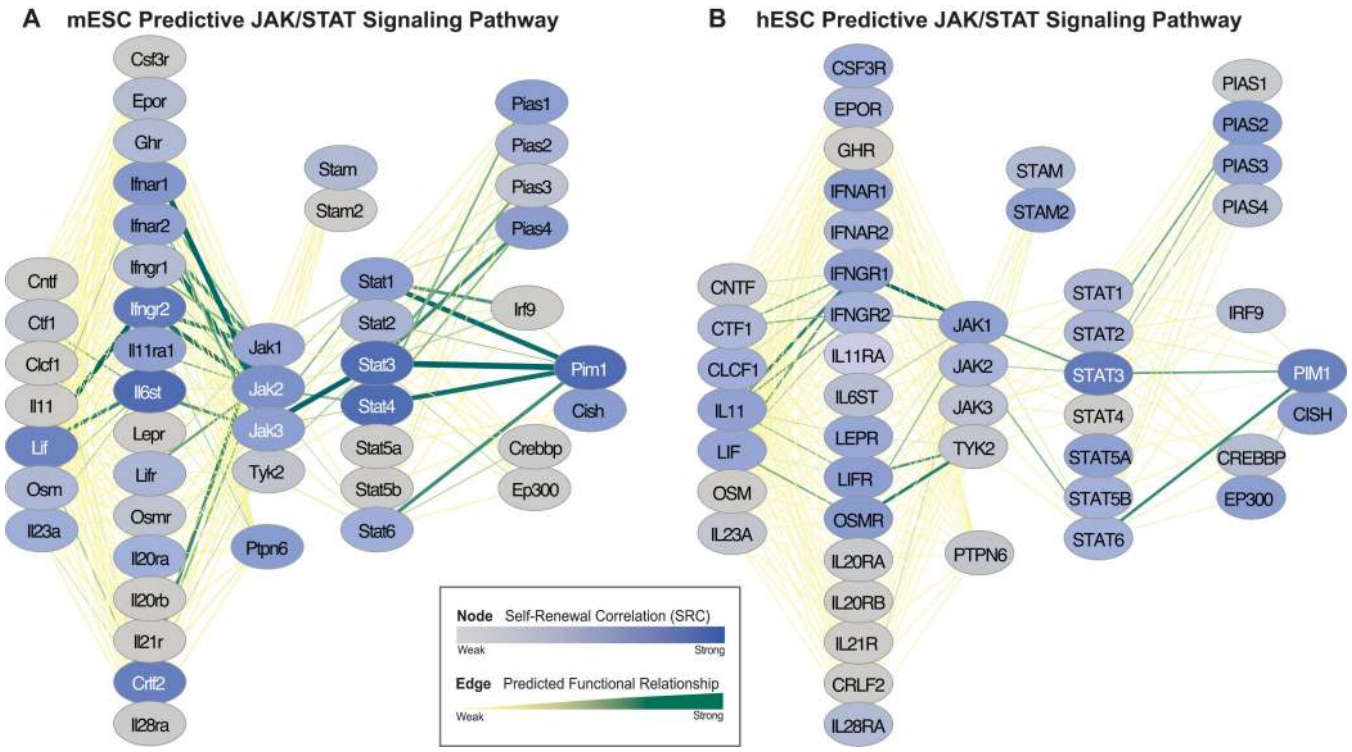


**Figure 3. Predicted JAK/STAT Signaling Pathway Relationships Across Species**

Analogously to Figure 2, species-specific predictions are shown for a portion of the KEGG pathway map for JAK/STAT signaling focused on the IL-6 family of cytokines. **A.** The hESC predicted JAK/STAT signaling pathway showed that *Stat3* is the gene most correlated to self-renewal, while upstream pathway participants exhibited lower SRC scores and weaker connection probabilities, suggesting that cytokines other than the IL-6 family, or signaling cross talk, may be required for *STAT3* activation in hESCs. **B.** The mESC predicted JAK/STAT signaling pathway showed a strong connection between *Lif* and its putative target, *Il6st*. Further, key known mouse self-renewal genes in this pathway, including *Lif, Il6st*, and *Stat3* are strongly correlated to other self-renewal genes as indicated by SRC score.

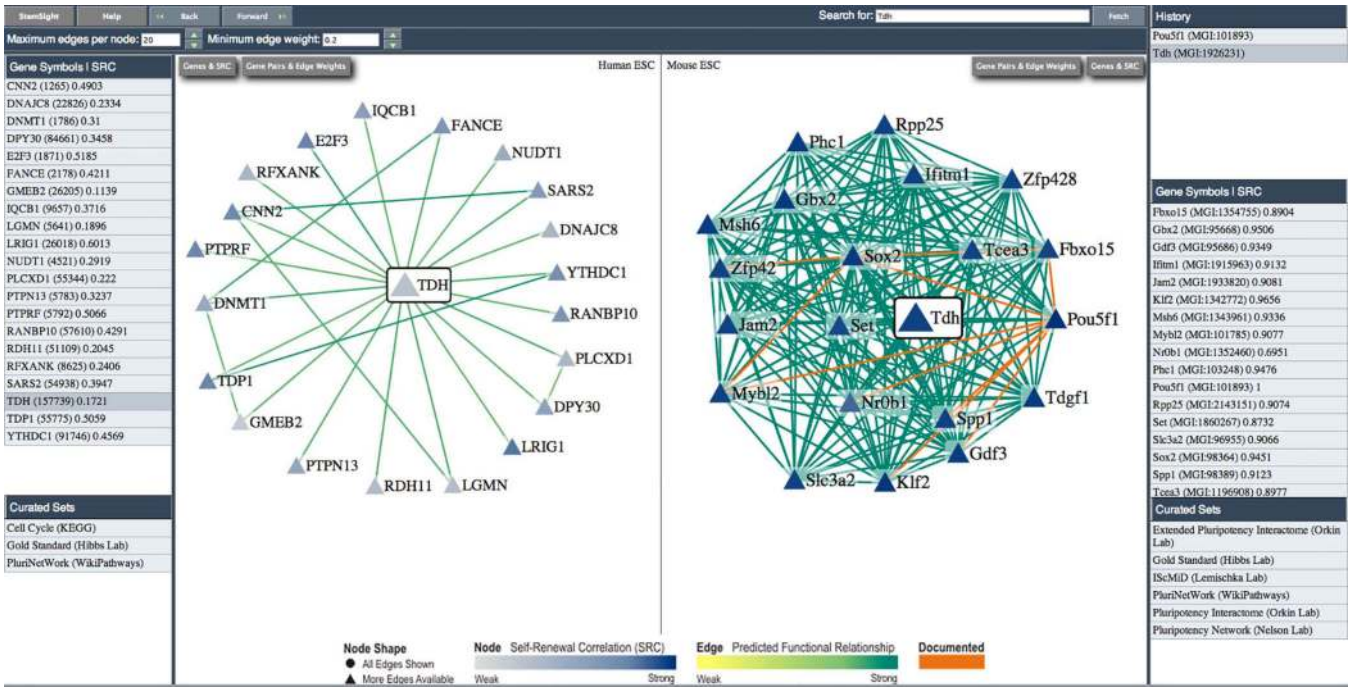


**Figure 4. Differences in mESC and hESC Threonine Metabolism**

We used our StemSight Scout data visualization tool (www.StemSight.org) to create network views centered around L-threonine dehydrogenase (*TDH*|*Tdh*), which supports accelerated cell cycle kinetics in mESCs, but is not functional in hESCs. Node and edge colors are as in Figures 2 and 3, except that edges contained in the positive training set were colored orange. **A.** The hESC *TDH*-centric network shows that *TDH* is weakly correlated to genes in our training set and has no strong functional associations to any known self-renewal genes. **B.** The mESC *Tdh*-centric network illustrated that *Tdh* is strongly correlated to self-renewal genes and had strong predicted functional associations with known self-renewal genes, including *Pou5f1, Sox2, Nr0b1, Klf2, Zfp42, Gdf3,* and *Fbx015*.

## Table 1

Summary of Integrated Data in Human and Mouse ESC Data Compendiums

| Data Type | Datasets\|Platforms | Genes Measured | Conditions | Gene Pairs | Redundancy |
|---|---|---|---|---|---|
| **hESC Data Compendium** | | | | | |
| Gene Expression | 72\|16 | 1,206,852 | 597 | 10,409,092,802 | 0.00872 |
| Protein-DNA Interactions | 28\|8 | 383,784 | 74 | 702,255,900 | 0.02718 |
| Protein-Protein Interactions | 3\|3 | 13,648 | 3 | 154,876 | 0.05047 |
| Epigenetic Markers | 9\|1 | 152,946 | 194 | 1,299,505,689 | 0.01053 |
| Phylogenetic Profiles | 1\|1 | 2391 | 229 | 3267 | 3.9E-6 |
| RNAi Screens | 3\|1 | 42,570 | 3 | 44,618,430 | 0.0195 |
| **mESC Data Compendium** | | | | | |
| Gene Expression | 61\|22 | 1,028,918 | 872 | 5,309,037,873 | 0.7110 |
| Protein-DNA Interactions | 102\|14 | 1,311,167 | 210 | 1,507,210,159 | 0.0218 |
| Protein-Protein Interactions | 1 \|1 | 207 | 1 | 207 | 2.2E-06 |
| Phylogenetic Profiles | 1\|1 | 15,703 | 1 | 123,284,253 | 0.1728 |
| RNAi Screens | 1\|1 | 16,891 | 2 | 131,795,730 | 0.1954 |

**Notes:** These data were collected from 83 hESC studies and 62 mESC studies, then standardized and integrated into a pair-wise format, and used as evidential data to generate predictive hESC- and mESC-specific networks focused on ESC self-renewal. Datasets were weighted based on the amount of shared mutual information (redundancy) contained in each as compared to all evidential datasets used by the Bayes net. A low mean redundancy indicates the dataset is highly unique. Dataset contributions to the top 0.001% (1503) of edges in the hESC and mESC networks are available in Supplemental Table S10.

## Table 2

Top Novel hESC Genes Most Strongly Correlated with Self-Renewal

| Gene Symbol | Gene Name | Rank | SRC (Scaled) |
|---|---|---|---|
| *TUBB* | Tubulin, beta class I | 1 | 0.8264 |
| *HSP90AB1* | heat shock protein 90kDa alpha (cytosolic), class B member 1 | 2 | 0.8053 |
| *PELI1* | pellino homolog 1 (Drosophila) | 3 | 0.7805 |
| *SEMA6A* | sema domain, transmembrane domain (TM), and cytoplasmic domain, (semaphorin) 6A | 4 | 0.7795 |
| *SLC7A5* | solute carrier family 7 (cationic amino acid transporter, y+ system), member 5 | 5 | 0.7693 |
| *RGMB* | RGM domain family, member B | 6 | 0.7684 |
| *PCBP1* | poly(rC) binding protein 1 | 7 | 0.7669 |
| *TUBB6* | Tubulin, beta 6 class V | 8 | 0.7184 |
| *HSPA4* | heat shock 70kDa protein 4 | 9 | 0.7653 |
| *SFRP1* | secreted frizzled-related protein 1 | 10 | 0.7587 |
| *LRRN1* | leucine rich repeat neuronal 1 | 11 | 0.7574 |
| *PRDM14* | PR domain containing 14 | 12 | 0.7458 |

**Notes:** Genes were identified by selecting the top novel genes (those not involved in positive training set edges) rank ordered by our self-renewal correlation (SRC) score as described in Methods and Materials and Supplemental File S1.

## Table 3

Functional Similarities and Differences Between Networks

| Shared | | Distinct | | | |
|---|---|---|---|---|---|
| ***Gene (h\|m)*** | ***Rank (h\|m)*** | ***hESC Gene*** | ***Rank (h\|m)*** | ***mESC Gene*** | ***Rank (h\|m)*** |
| *POU5F1\|Pou5f1* | 1\|1 | *PELI1* | 5\|8197 | *Utf1* | 7906\|6 |
| *SOX2\|Sox2* | 2\|5 | *LRRN1* | 15\|11119 | *Tcl1* | 14166\|17 |
| *NODAL\|Nodal* | 12\|114 | *TUBG1* | 32\|7004 | *Zfp428* | 13991\|22 |
| *RIF1\|Rif1* | 14\|33 | *RPL10A* | 37\|8606 | *Gjb3* | 13495\|23 |
| *PARP1\|Parp1* | 21\|97 | *RPLP2* | 38\|7284 | *Tdh* | 11393\|26 |
| *LEFTY2\|Lefty2* | 33\|85 | *TALDO1* | 63\|8745 | *Fbxo15* | 14921\|35 |
| *MYC\|Myc* | 40\|38 | *RPS3A* | 65\|11303 | *Zfp57* | 13784\|63 |
| *JARID2\|Jarid2* | 46\|11 | *RPL6* | 101\|11695 | *Eras* | 13432\|92 |
| *ZFP42\|Zfp42* | 78\|27 | *EIF3A* | 115\|12160 | *Dusp27* | 14812\|134 |
| *IFITM1\|Ifitm1* | 107\|15 | *USO1* | 151\|11350 | *Enpp3* | 16752\|169 |
| *NANOG\|Nanog* | 128\|56 | *GPM6B* | 419\|13693 | *Spats1* | 15033\|172 |

**Notes:** Shared genes strongly associated with ESC self-renewal were identified by finding the intersection of the highest-ranking 1% of genes (173 of 17,342) in each network ordered by SRC. Distinct genes were identified by taking the difference of SRC rank for all genes and selecting the top genes most strongly supported in one network, but not the other.