Derivation of an interaction/regulation network describing pluripotency in human

Anup Som[† a], Mitja Luštrek[† a, b], Nitesh Kumar Singh[a], Georg Fuellen[a]*

[a] Institute for Biostatistics and Informatics in Medicine and Ageing Research, University of Rostock, Ernst-Heydemann-Str. 8, 18057, Rostock, Germany

[b] Jožef Stefan Institute, Department of Intelligent Systems, Jamova cesta 39, 1000 Ljubljana, Slovenia

[†]These authors contributed equally to this work

*To whom correspondence should be addressed.

Tel. +49-381-4947360. Fax. +49-381-4947203. Email: fuellen@uni-rostock.de (GF)

Abstract

Identification of the key genes/proteins of pluripotency and their interrelationships is an important step in understanding the induction and maintenance of pluripotency. Experimental approaches have accumulated large amounts of interaction/regulation data in mouse. We investigate how far such information can be transferred to human, the species of maximum interest, for which experimental data are much more limited. To address this issue, we mapped an existing mouse pluripotency network (the *PluriNetWork*) to human. We transferred interaction and regulation links between genes/proteins from mouse to human on the basis of orthologous relationship of the genes/proteins (called interolog mapping). To reduce the number of false positives, we used four different methods: phylogenetic profiling, Gene Ontology semantic similarity, gene co-expression, and RNA interference (RNAi) data. The methods and the resulting networks were evaluated by a novel approach using the information about the genes known to be involved in pluripotency from the literature. The RNAi method proved best for filtering out unlikely interactions, so it was used to construct the final human pluripotency network. The RNAi data are based on human embryonic stem cells (hESCs) that are generally considered to be in a (primed) epiblast stem cell state. Therefore, we assume that the final human network may reflect the (primed) epiblast stem cell state more closely, while the mouse network reflects the (unprimed/naïve) embryonic stem cell state more closely.

Keywords

Interolog mapping; Phylogenetic profiling; Co-expression; Gene Ontology; RNA interference; Stem cell

1. Introduction

In recent years the field of embryonic stem cell (ESC) and pluripotency research has gained importance because of its therapeutic potential in regenerative medicine (Boiani and Scholer, 2005; Rao and Orkin, 2006; Jaenisch and Young, 2008). Unraveling the mechanisms underlying pluripotency and reprogramming in human may open a new era in medicine (Cohen and Melton, 2011). These mechanisms involve protein-protein interactions, which participate in key biological processes in cells. They are supplemented by protein-DNA interactions describing gene regulation by the control of transcription. Describing and interpreting such a network of interaction and regulation (i.e., stimulation and inhibition) links is an essential task of computational biology.

Considering how beneficial the knowledge of the mechanisms underlying pluripotency and reprogramming in human would be, the description of the gene/protein interaction/regulation underlying pluripotency (called the pluripotency network) in human is remarkably limited. By inferring interactions in human based on mouse, which is the most closely related species with rich interaction/regulation data, we created a view of the human pluripotency network, an important first step in systems-level understanding of the underlying mechanisms.

To make full use of the currently available interaction (and regulation) data, computational methods have been developed to predict new interactions. These methods are based on diverse attributes, concepts, and data types, such as interologs (Matthews *et al*., 2001; Yu *et al*., 2004), gene expression profiles (Ideker at al. 2002), Gene Ontology (GO) annotations (Wu *et al*., 2006), phylogenetic profiling (Pellegrini *et al*., 1999), domain interactions (Ng *et al*. 2003), and co-evolution (Jothi *et al*., 2005). Some machine learning methods, such as support vector machines (SVMs), were also used to predict protein-protein interactions based on sequence data (Shen *et al*., 2007; Guo *et al*., 2008). Among these methods, the interolog approach has been widely implemented (Rhodes *et al*., 2005). The method assumes that protein-protein interactions are conserved between organisms and that pairs of proteins whose orthologs are known to interact in a model

organism probably interact in the organism of interest (target organism) as well (Walhout *et al*., 2000). Numerous studies were published in which mainly human protein-protein interactions were predicted based on interolog detection (Han *et al*., 2004; Lehner and Fraser, 2004; Rhodes *et al*., 2005; Tirosh and Barkai, 2005; Brown and Jurisica, 2005; Persico *et al*., 2005; Huang *et al*., 2007).

A potential problem in predicting protein-protein interactions using an interolog-based method is that it may generate false positive interactions, i.e., interactions that are falsely predicted to exist in the target organism. The false positive interactions appear due to two reasons. The first reason is false positives in original interactions obtained experimentally in the model organism (von Mering *et al*., 2002; Sprinzak *et al*., 2003; Yu *et al*., 2008). The second reason is the lack of evolutionary conservation of interactions, in particular when applied to phylogenetically distant organisms (Mika and Rost, 2006; Brown and Jurisica, 2007). In such cases an interaction does exist in the model organism but not in the target organism.

We minimized the appearance of false positive interactions in the interaction data of the model organism (i.e., the first reason) by considering the high-quality literature-curated mouse *PluriNetWork* as the model network (Som *et al*., 2010). To reduce the number of false positives due to the lack of evolutionary conservation of interactions (i.e., the second reason), we filtered the interactions using four methods: (1) phylogenetic profiling, (2) GO semantic similarity, (3) gene co-expression, and (4) considering RNAi data. A novel approach was adopted to evaluate the relative performance of these four methods. The best of them (that is, RNAi) was finally selected to filter out the unlikely interactions, resulting in the final predicted human pluripotency network.

2. Materials and methods

Fig. 1 shows the flowchart of our approach to mapping the mouse pluripotency network to human. Its steps are described in detail in the rest of the paper.

2.1. Model network: Mouse pluripotency network

We previously assembled a network of 547 molecular interactions, stimulations and inhibitions involved in mouse pluripotency called *PluriNetWork* (Som *et al*., 2010), which we consider the model network. It is shown in Fig. 2 (a high-resolution JPEG image and a Cytoscape version of the network are given in Supplementary Fig. S1). It is based on a collection of primary research data from 177 publications involving 264 mouse genes/proteins. It includes the core circuit of Oct4 (Pou5f1), Sox2, Nanog and Klf4, its periphery Esrrb, c-Myc, Nr5a2, Stat3, and Sall4 (red region), connections to upstream signaling pathways such as Activin, Wnt, FGF, BMP, Insulin, Notch, and LIF (green region), and epigenetic regulators such as Dnmt3a, Dnmt3b, Hdac1, Hdac2, and Kdm3a (blue region). A detailed description of the network assembly, its properties, the associated biological information, and its applications is found in the publication of Som *et al*., (2010).

2.2. Ortholog identification

Orthologs of mouse pluripotency genes/proteins in human were identified from three publicly available ortholog databases: (1) Ensembl (Release 62, April 2011) [http://www.ensembl.org], (2) InParanoid (Version 7.0, June 2009) (Berglund *et al*., 2008), and (3) HomoloGene (Release 64, February 2011) [http://www.ncbi.nlm.nih.gov/homologene/], as follows. We exported mouse-human ortholog pairs from Ensembl by the help of BioMart software (Haider *et al*. 2009). The mouse Ensembl gene ID in the PluriNetWork was used as an identifier to mine Ensembl with BioMart. From the InParanoid database, the complete dataset of mouse-human orthologs was downloaded. We then extracted the orthologs of mouse pluripotency genes in human. InParanoid is one of the best ortholog databases, especially as it identifies more correct co-orthologs (defined as two or more genes that were duplicated after the speciation and hence are orthologs to one or more genes in another species) than other such databases (Chen *et al*., 2007). Finally, the HomoloGene online web interface was used to establish mouse-human ortholog pairs. All three databases

contained orthologs for all the genes except two. None had an ortholog for Ins1 (*Insulin 1*) and only HomoloGene had an ortholog for Ctbp2 (*C-terminal binding protein 2*), which we nevertheless decided to omit from our set of human orthologs. We also looked for co-orthologs of mouse pluripotency proteins in human. The Ensemble ortholog dataset showed that two mouse proteins Lefty1 (*Left right determination factor 1*) and Zfx (*Zinc finger protein X-linked*) have co-orthologs in human, whereas Inparanoid and HomoloGene did not support these co-orthologs. Therefore, we assumed that only one-to-one relationships exist in the ortholog data. In summary, we obtained a clean set of human orthologs of mouse pluripotency players that contains 262 genes/proteins (Supplementary Table S1).

2.3. Mouse-to-human interolog mapping

In the *PluriNetWork*, protein-protein interactions and regulatory protein-DNA interactions (i.e., stimulations and inhibitions) are collectively called *links*. We do not distinguish a gene and its protein product – they are both referred to by the gene name. For all links between mouse genes, the mouse-human ortholog pairs were investigated. If both genes comprising a mouse link have human orthologs, then these human orthologs were predicted to be linked by an interolog. The interolog detection strategy was initially developed to transfer information on protein-protein interactions (from yeast to higher organisms) (Walhout *et al*., 2000; Matthews *et al*., 2001), but it can also be employed for regulatory links (Yu *et al*., 2004; Yellaboina *et al*., 2007). This method assumes that links are conserved between organisms: pairs of proteins whose orthologs are known to interact in other species probably interact in the species of interest as well. Based on human orthologs of mouse pluripotency players (genes/proteins), we transferred the links from mouse to human, resulting in the initial human version of the mouse PluriNetWork. The predicted network consists of 262 nodes (genes/proteins) linked by 545 links (Supplementary Fig. S2). Thus, with the exception of two links to genes that had no ortholog (to Ins 1 and to Ctbp2), the predicted network is identical to the original one.

2.4. Filtering out false positive Interologs

The links transferred from mouse to human may include false positives due to false positives in the original interactions obtained experimentally in mouse (von Mering *et al*., 2002; Sprinzak *et al*., 2003; Yu *et al*., 2008) or the lack of evolutionary conservation of interactions (Mika and Rost, 2006; Brown and Jurisica, 2007). We minimized the first reason (i.e., the appearance of false positive interactions in the model organism) by considering the high-quality literature-curated *PluriNetWork* as the model network. To evaluate whether the transferred links truly belong to the human pluripotency network and thus to reduce the number of false positives due to the second reason, we used four different link evaluation methods: (1) phylogenetic profiling, (2) GO semantic similarity, (3) gene co-expression, and (4) considering RNAi data. The first three methods are well known for their potential to improve the quality of predicted interactions, while the fourth one is new.

For each link, each of the four link evaluation methods provided a value (we called it link value) corresponding to the probability that the link is involved in pluripotency. However, in order to actually filter out false positive interologs, we needed a threshold for the values provided by each of the four methods to separate the links to be included in the network from those to be excluded. Furthermore, in order to select the best of the four filtering methods, we need to evaluate the networks constructed using each of them. The four methods are described in the following four subsections. The selection of the thresholds and the evaluation of the networks is described in the final subsection.

2.4.1. Phylogenetic profiling method

Phylogenetic profiling assumes that two proteins displaying a similar phylogenetic profile (i.e., a similar presence/absence pattern in a set of reference organisms) are functionally linked (Pellegrini *et al*., 1999). In other words, if both proteins are either conserved or deleted in several organisms, this is an indication of a link between them. A binary phylogenetic profile of a gene is represented by a vector of 0 and

1, depending on the presence or absence of the gene's homolog in the set of reference organisms. We constructed binary profiles for all mouse pluripotency genes, using 14 reference vertebrate genomes (incl. human) listed in Supplementary Table S2. A Blastp phylogenetic profile (Enault *et al*., 2003) of a mouse gene is represented by a vector of normalized bit scores obtained from BLAST when searching for homologs of the protein encoded by the gene in the 14 other genomes. The normalized Blastp (Altschul *et al*., 1997) bit scores were taken from the InParanoid ortholog dataset. We then calculated an "evolutionary dissimilarity score" (EDS) for each link using binary and Blastp profiles. The EDS of the linked genes *i* and *j* is defined as the sum of the absolute differences of binary or Blastp scores across the profile, i.e.

$$EDS_{ij} = \sum_{k=1}^{N} \left| P_{ik} - P_{jk} \right|,$$

where P_{ik} and P_{jk} denote the presence or absence of the homologs of the genes *i* and *j* in the genome *k*, or the Blastp bit scores of the proteins encoded by the genes *i* and *j* when searching for them in the genome *k*, and *N* is the number of genomes.

We defined the EDS based on the hypothesis that a pair of interacting genes should feature similar evolutionary changes among species, elaborating on the fundamental assumption of phylogenetic profiling that co-evolving genes are functionally linked. According to our definition of EDS, two proteins with similar phylogenetic profiles should have a low EDS, and two proteins with dissimilar profiles should have a high EDS. A link with a low EDS indicates that the link should be included in the human pluripotency network, whereas a link with a high EDS is likely a false positive and should be excluded from the network.

2.4.2. GO semantic similarity method

This method assumes that interacting proteins share the same subcellular localization (Shin *et al*., 2009) and are involved in similar biological processes (Ewing, 2007). We assessed these two properties by the similarity of the genes according to their Cellular Component (CC) and Biological Process (BP) GO terms (Schlicker *et*

al., 2006). We used three variants of the GO semantic similarity method: (1) BP similarity, (2) CC similarity, and (3) BP + CC similarity. The BP and CC similarities were computed as the similarities of the GO BP and CC terms of the genes that are linked. The BP + CC was computed as the average of the BP and CC similarities if the link consisted of either two transcription factors (TFs) or two non-TFs, or as the BP similarity otherwise. The reason for not considering the CC similarity of links between a TF and non-TF is as follows. The *PluriNetWork* contains several transcriptional links (i.e., a link between a TF and a non-TF, such as a signaling protein), e.g. Sox2-Fgf4 and Pou5f1-Fgf4. Naturally, the TFs, Sox2 and Pou5f1 are located in the nucleus, whereas the location of Fgf4 is the extracellular space. In such cases the CC similarity cannot reflect the probability of stimulation/inhibition.

We calculated the GO BP and CC semantic similarities between interacting genes using Resnik's term similarity method as implemented in the GOSim package (Frohlich *et al*., 2007). Resnik's method is based on the information content of the lowest common ancestor (LCA) of two terms (Resnik, 1999). The more frequently a term occurs, the lower is its information content. If the LCA of two terms describes a generic concept, these terms are not very similar and this is reflected in the low information content of their LCA. The corresponding genes probably do not interact and are thus assigned a low link value. Resnik's method is considered the best among the existing methods for measuring the semantic similarity (Guo *et al*., 2006; Wang *et al*., 2007).

2.4.3. Gene co-expression method

This method assumes that interacting pairs of proteins tend to be co-expressed. Human pluripotency-specific gene expression data were used to measure the co-expression values (twelve samples: GSM530601-3, 6, 9, 11, and 13-18 from the Gene Expression Omnibus series GSE21222). Of the twelve samples, six samples (GSM530601-3, 6, 9, 11) are of the (primed) epiblast type (hESC), and six samples (GSM530613-18) are of

the (unprimed) naïve type (see Fig. 4C of Hanna *et al*., 2010 from the paper that is associated with the GSE21222 dataset). Expression values were preprocessed using MAS5 algorithm (Hubbell *et al*., 2002).

Co-expression of a pair of genes was first assessed by calculating the Pearson correlation coefficient of their expression values. We calculated the correlation coefficient for each link in three different manners: (1) using the hESC type data only (six samples), which we called hESC correlation score of the link, (2) using the naïve type data only (also six samples), which we called naïve correlation score, and (3) using both hESC and naïve type data (i.e., using all the twelve samples), which we called hESC + naïve correlation score. The correlation coefficients range from –1 to 1. Values close to 1 or –1 indicate likely interaction (negative values mean one gene/protein probably inhibits the other) and values close to zero indicate no interaction. We thus used the absolute values of correlation coefficients of pairs of genes as the link values.

Additionally, we used a method to measure the change of frequency of interaction (that is, its startup or shutdown) known as LinkScore (Warsow *et al*., 2010). The LinkScore method calculates the amount of change in interaction between two genes/proteins, usually measured by two gene expression experiments. We computed the LinkScores of each pair of linked genes from pluripotent states (we again considered hESC, naïve, and hESC+naïve samples separately) to non-pluripotent state, and used them as the link values. For the pluripotent states we used the already described twelve samples. For the non-pluripotent state, we used the non-pluripotent counterparts of our samples (GSE7178 and GSE 17772, samples GSM172865-73 and GSM443832-34). A high LinkScore means that both genes interact more in the non-pluripotent state, which indicates that the link should not be included in the human pluripotency network.

2.4.4. RNAi method

Recently, a genome-wide RNAi screen was conducted to identify genes which regulate self-renewal and pluripotency properties in hESC (Chia *et al*., 2010). In this study, the authors screened a small interfering RNA (siRNA) library targeting 21,121 human genes and reported the importance of their role in hESC. Each

of the 21,121 human genes was assigned a numerical score called F_{av}. Higher values of F_{av} indicated a more important role in hESC. For example, in their score sheet Pou5f1 was placed on top with the highest score, indicating Pou5f1's critical importance in hESC, a fact established by several other studies.

For the purpose of filtering out false positive interologs, we needed to assign values to links instead of genes. We considered three approaches (i.e., three variants of processing the RNAi data) to transform two gene scores to a link score: (1) the minimum of the F_{av} scores of the linked genes, (2) the arithmetic mean of the scores, and (3) the geometric mean of the scores. The rationale for (1) is that if one of the linked genes is not involved in human pluripotency, the link between it and another gene does not belong to the human pluripotency network, either. The rationale for (2) is that since the relation between the F_{av} score and the involvement in pluripotency is somewhat uncertain, averaging the scores of both genes cancels out some of the uncertainty. The last method, (3) geometric mean, is striving for a balance between (1) and (2): it averages the F_{av} scores, but if one of the scores if very low and the other very high, their geometric mean is still very low, as in (1).

2.5. Evaluation of the methods

Each of the four link evaluation methods provided us with link values for a subset of the 545 links in the initial predicted human pluripotency network. This is because some of the data required to use the methods was not available for all the links – for example, a number of genes have no GO annotations, so the links associated with these genes have no GO semantic similarity scores. The sizes of these subsets are given in Table 1. The size of the intersection of the four subsets, which contains the links for which all four methods provided link values, is 406. The link evaluation methods were compared on these 406 links.

In order to compare the link evaluation methods and to evaluate the resulting pluripotency networks, we needed some reference data – reliable and independent information on which links belong to the human pluripotency network. Such data for links between genes is hard to come by, but there is some literature on

genes/proteins experimentally shown to be or not to be involved in human pluripotency. We found 15 genes known to be involved in human pluripotency and 12 genes known not to be involved (Table 2). We translated this information on genes/proteins into information on links as follows. If both genes in a link are known to be involved in human pluripotency, we considered the link to belong to the human network. If at least one of the genes is known not to be involved, we considered the link not to belong to the human network. If one of the genes is known to be involved and we had no information on the other, we assigned a probability to that link belonging to the human network. Since one gene in such a link is already known to be involved, the probability for the whole link is equal to the probability of the other gene being involved. Therefore, the probability of such a link belonging to the human network was computed as the number of link endpoints known to be involved in human pluripotency (314, that is the number of times that the 15 genes known to be involved are the endpoint of a link), divided by the number of link endpoints on which we had some literature-based information (436 endpoints, 314 involved and 122 not involved), which equals 0.72. (Note that a "link endpoint" is a gene, but we use this term to convey that we count it once for every occurance in a link.) We used the counts of link endpoints instead of simple gene counts because we wished to count each gene once for each occurrence in a link, since this gives a greater weight to the genes involved in more links. We ignored the links for which we had no information on either gene. The correctness scores for inclusion in or exclusion from the human pluripotency network are given in Table 3. A link is assigned the same score if the involvements of the first and the second gene are reversed. For example, if the third case were (yes, no) instead of (no, yes), the correctness scores for that case would be the same.

Based on the link values assigned by the link evaluation methods, the links were ranked in the order of the probability that they should be included in the human pluripotency network (i.e., the higher rank of a link by a given method, the more likely the link is involved in pluripotency according to that method). The link values and the corresponding ranks for the four link evaluation methods are shown in Supplementary Table S3. To construct the human network, we needed to decide for each link whether to include it or not, which

means that we needed thresholds for the values provided by the link evaluation methods. Ideally, the links above such a threshold would all be involved in human pluripotency, and the links below it would not be involved. As the threshold we used the number of links to be included in the human network, which we denoted *n*. This made it possible to compare the four link evaluation methods, since each method uses a different scale. Each method and each value *n* split the links into those to be included in the human network (the top-ranked *n* links) and those not to be included (the bottom-ranked 406 – *n* links). For example, if we chose the RNAi method and the threshold 100, the 100 links with the highest link values according to the RNAi method would be included in the human network, and the remaining 306 links would be excluded. For the links to be included, the correctness scores from the third column (inclusion) of Table 3 were added up, and for the links to be excluded, the scores from the fourth column (exclusion) were added up. The total sum of these gave the overall quality of the human pluripotency network for a given link evaluation method and a given number of links *n*. The value *n* for which this score was the largest, was considered the optimal threshold for a given method.

3. Results

We first constructed the initial human pluripotency network by transferring the links between the genes from the mouse pluripotency network. Afterwards we compared the four link evaluation methods used to filter out the false positive links, as described in the Materials and method section, and constructed the final human network using the best of them. The following two sections present the results of the comparison of the link evaluation methods and the final human network.

3.1. Comparison of the link evaluation methods

Each of the four link evaluation methods has multiple variants, which we compared to choose the best variant for each method. For phylogenetic profiling, we compared binary profiles with profiles consisting of Blastp

bit scores. For GO semantic similarity, we compared similarities of BP, CC and BP + CC GO terms. For gene co-expression, we compared the degree of co-expression computed with the Pearson correlation coefficient and LinkScore using hESC, naïve and hESC + naïve samples. For the RNAi method, we compared the link values computed as the minimum, arithmetic mean and geometric mean of gene values. The results of the comparisons revealed that for the phylogenetic profiling method binary profile is the best variant (see Supplementary Figure S3), for the GO semantic similarity method CC similarity is the best variant, for the gene co-expression method LinkScore using hESC samples is the best variant, and for the RNAi method arithmetic mean of gene values is the best variant (the comparison of the variants of each of the four link evaluation methods, as described in the next paragraph, are presented in Supplementary Fig. S3). We selected as the best variant the one with the highest average and peak correctness score.

After selecting the best variant of each link evaluation method, we compared the four methods using these variants. The results are presented in Fig. 3. The horizontal axis of the graph represents the threshold (n), which ranges from 0 (no links included) to 406 (all links included). The vertical axis shows the number of links whose inclusion in the human network or exclusion from it is correct based on the literature. The correctness curves show the correctness of each method at a given threshold. Of particular interest is the threshold at which a method reaches the highest correctness, since that is the threshold for inclusion in the pluripotency network that would be chosen for that method. The comparison of the link evaluation methods shows that the RNAi method is the best performer followed by the gene co-expression method. The most commonly used phylogenetic profiling and GO semantic similarity methods performed poorly.

3.2. Derivation of human pluripotency network

Since the RNAi method proved to be the best link evaluation method, we used it to filter out false positive interactions from the human pluripotency network. Fig. 4 shows the correctness score with respect to the threshold for the RNAi method for all the links for which we had RNAi information, which are 540 out of

545 links. RNAi data (F_{av} score) was missing for four genes (Kdm1a, Kdm4c, Kdm5c, and mTOR), which appeared in five links. The threshold thus ranges from 0 (no links included) to 540 (all links included).

The maximum possible correctness that could be achieved by the method is 307.52, which is calculated as follows. We have literature-based information on 373 links. For 139 of them we know that both genes in the link are involved in pluripotency (so the link is involved as well) or that one of the genes is not involved (so the link is not involved, either). If all of these links are included in / excluded from the network correctly, they contribute 139 to the correctness score. For 234 links we know that one gene is involved, so if all of them are included in the network, they contribute $234 \times 0.72 = 168.52$ to the correctness score (as per Table 3). The maximum number achieved by the RNAi method is 250.29 when the threshold is 215, which indicates that the top-ranking 215 links are to be retained in the human network. Thereafter, we used the gene co-expression method to evaluate the remaining five links (on which we had no RNAi data). All the five links were filtered out by the co-expression method, so we deleted them from the human network.

After filtering out possible false positive interactions by the RNAi method, we observed that for a number of genes the majority of their links were filtered out. This observation led us to the conjecture that they are not involved in human pluripotency. We decided to delete all genes from the network for which at least 80% of the links were filtered out. The threshold of 80% was set to the highest value such that the remaining genes not involved in human pluripotency based on the literature were deleted. By applying this criterion, ten more genes (and 18 links) shown in Table 4 were deleted from the filtered network. Out of the ten deleted genes, the literature reported that seven genes (Ctnnb1, Esrrb, Klf2, Klf5, Nr5a2, Smad1, and Stat3) are not involved in human pluripotency (Table 2) and the role of the other genes (Mbd3, Myc and Sall4) in hESCs is unknown. Finally, we deleted the remaining gene that the literature reported not to be involved in human pluripotency – LIF – and its links. As a result, LIFR was disconnected from the network, so we deleted it as well. The final human pluripotency network retained 196 links and 136 genes (see Supplementary Table S4). This means that the majority of links (approximately 64%) were filtered out from

the predicted network. Fig. 5 thus shows the final putative interaction/regulation network of human pluripotency (a high-resolution JPEG image and a Cytoscape version of the network are given in Supplementary Fig. S4).

4. Discussion

In this study, we mapped a mouse pluripotency network to human by interolog detection and then used the RNAi method as a filter to increase the quality of the transferred network. We first established orthologous relationships of mouse-human pluripotency genes/proteins by the combination of the three most popular publicly available databases, namely Ensembl, InParanoid and HomoloGene. Even given an adequate set of orthologs, the predicted network is expected to contain several false positive interactions. To filter out such false positives, we used four methods to evaluate the interactions. To find out the best method, we evaluated their relative performance. Interestingly, we found that the most widely used methods (i.e., phylogenetic profiling, GO semantic similarity, and gene co-expression) performed worse than the RNAi method. The high performance of the RNAi method is not entirely surprising, because the RNAi experiment was conducted precisely to identify the genes which regulate self-renewal and pluripotency in hESCs (i.e., it directly measured the involvement and/or criticality of the genes in hESCs). However, it is worth examining why the other methods performed relatively poorly.

A likely reason for the poor performance of phylogenetic profiling is that it is too general. It predicts interactions based on evolutionary conservation of genes only and does not take into account the function of the genes. For example, Esrrb, LIF, and Il6st (also called Gp130) are well conserved across the genomes, so phylogenetic profiling is unable to filter them out, even though they are known not to be involved in human pluripotency (i.e., the phylogenetic profiling method is unable to measure the degree of involvement in pluripotency; rather it measures the degree of interaction between genes). However, it is possible that these

genes are involved in other cellular phenomena (i.e., they are false positives not because they do not interact, but because they are not involved in pluripotency).

The underlying hypothesis of the GO semantic similarity method is that interacting proteins share the same sub-cellular localization and are involved in similar biological processes. As described in the GO semantic similarity method, because of transcriptional links the CC score cannot reflect the probability of interaction. The BP score is not reliable, either, because the observations that pair of proteins shares the same biological process do not guarantee that they in fact interact. Finally, GO annotations are known to be incomplete and erroneous (Done *et al*., 2010). Particularly for the species human, GO annotation is fairly problematic. A large number of GO annotations of human genes come from mouse genes (the annotation that was made for the mouse gene was transferred to the human gene) (http://www.geneontology.org/).

Even though the gene co-expression methods performed better than phylogenetic profiling and GO semantic similarity methods, it was still unable to filter many genes properly. The method is based on the notion that interacting proteins are co-expressed. However, expression data are usually derived from a heterogeneous mixture of cells and cellular compartments. Genes may have very specific expression patterns based on a variety of cellular activities. Thus, overlapping local expression patterns may not be identifiable in a global co-expression measure. Moreover, the human pluripotency network is composed of TFs, signaling proteins and epigenetic factors. It is, therefore, possible that interacting signaling proteins or interacting TFs are strongly co-expressed, but this may not be true for an interacting pair of a signaling protein and a TF (i.e., both signaling proteins and TFs may have their specific expression patterns, including time delays if a TF is activated by a signal, or if a TF activates a signal). The Pearson correlation coefficient variant of the gene co-expression method utilizes co-expression pattern only, while the LinkScore variant looks for a difference between pluripotent and non-pluripotent samples. The latter performed better which confirms our observation that false positives are caused by the lack of involvement in pluripotency, not the lack of interaction.

After filtering out false positives, the final human pluripotency network retained 196 links and 136 genes, which means that the majority of links (approximately 64%) were filtered out from the predicted network. This result implies that the underlying mechanisms of pluripotency significantly differ between mouse and human. This is somewhat surprising considering that 99% of mouse pluripotency genes have human orthologs. Furthermore, mouse and human genomes are highly conserved in general, with about 85% of all the mouse genes having human orthologs. However, the difference in the mechanism of pluripotency between the species was explained by Tesar *et al*., (2007), who found fundamental differences between the mouse and human state of pluripotency usually investigated that is an embryonic mouse stem cell (naïve/unprimed) versus an epiblast-like human stem cell (primed).

We inspected the pathways in the final human network, namely TGF-beta/Activin/Nodal, Wnt signalling and LIF signalling, and found that the TGF-beta/Activin/Nodal pathway exists (Fig. 5; yellow region), the Wnt pathway was disconnected from the core network (brown region) and the LIF signaling pathway was deleted. It was reported that the activation of the TGF-beta/Activin/Nodal branch through SMAD2/3 is associated with pluripotency in human and is required for the maintenance of the undifferentiated state in hESCs (Vallier *et al*., 2005; James *et al*., 2005). Our filtered human network thus agrees with the current state of experimental knowledge of the TGF-beta/Activin/Nodal pathway. However, Wnt proteins are also believed to play an important role in controlling hESC maintenance (Sato *et al*., 2004), but the Wnt pathway was disconnected. A possible reason is that different Wnt genes are required for the maintenance of the undifferentiated state of the ESCs in human and in mouse. The Wnt family has 19 members (genes). In the mouse *PluriNetWork*, two Wnt genes, namely Wnt3a and Wnt5a, were included. In human, it was reported in the literature that Wnt3, Wnt5a and Wnt10B are involved in pluripotency mediated by the Wnt pathway (Katoh, 2008). Furthermore, the RNAi screening result confirmed that Wnt3 and Wnt10B (together with Wnt2B and Wnt9B) play a more important role in hESCs than Wnt3a and Wnt5a (Chia *et al*., 2010). Several experimental studies reported that unlike in mouse, the LIF signaling pathway is

not required to maintain hESC (Okita and Yamanaka, 2006; Sun *et al*., 2006). Our filtered human network matches these experimental results.

The main limitation of the human pluripotency network derived from the mouse network is the missing links. Links are missing because they are absent from the mouse network or are human-specific. We believe that the mouse network reflected the current knowledge of mouse pluripotency fairly well, and as new links are discovered, they can be added to the mouse network and transferred to human. Some of the missing links, however, are human-specific. Considering that the size of the derived human network is less than half of the mouse network and that we have no reason to believe that the true human network is smaller than the mouse network, the human-specific links may well be in the majority. Some of these links may be transferred from other species or inferred from gene expression and RNAi data, although it is doubtful that these approaches would yield much reliable information. We may consider them in the future, but the only sure way to fill in the missing links is experimental identification. It is also future work to find out how useful networks from other murine cell types may be for estimating the human pluripotency network. We expect that the main determinant of usefulness will be the proximity of the murine cell type to the specific kind of pluripotency featured by hESC. In particular, a network from mouse Epiblast stem cells (EpiSC) may be very close, whereas networks of proliferating (cancer) cells of mice are expected to be less related, though they may still share features related to proliferation / renewal. Networks from very different cell types are expected to be only remotely related; they should miss the core pluripotency network around Pou5f1, Sox2 and Nanog as well as much of its periphery.

5. Conclusions

In conclusion, we derived a putative human pluripotency network for mouse, for which experimental data are much more plentiful than for human. The quality of the predicted network was improved by using genome-wide RNAi screening data, which directly measured the involvement and/or criticality of the genes in

hESCs. The predicted network will be useful to understand the biology underlying pluripotency, and scientists are expected to benefit from the access to a human network of pluripotency players and mechanisms, which will help them make sense of high-throughput data. Most importantly, given recent investigations, we assume that the human network may reflect the "primed" epiblast stem cell state more closely, while the mouse network reflects the "unprimed", or, "naïve", ESC state more closely. It is future work, requiring more experimental data, to disentangle the difference in the developmental state and the species difference.

In the future, we are interested in mapping the pluripotency network to more organisms of interest where the experimental data are also limited (e.g., Axolotl, Chicken, Rat, etc.) and to investigate their evolution. Multi-species pluripotency networks should be useful to identify species-specific pathways evolution, and afford a deeper understanding of the evolution of pluripotency.

Acknowledgements

We thank the two anonymous referees for their helpful comments. Funding by the DFG SPP 1356, *Pluripotency and Cellular Reprogramming* (FU583/2-1), is also gratefully acknowledged.

Appendix A. Supplementary data

Supplementary data to this article can be found online.

References

Adewumi, O., Aflatoonian, B., Ahrlund-Richter, L., *et al*., 2007. Characterization of human embryonic stem cell lines by the International Stem Cell Initiative. Nature Biotechnology 25, 803-816.

Altschul, SF., Madden, T.L., Schaffer, A.A., *et al*., 1997. Gapped BLAST and PSIBLAST: a new generation of protein database search programs. Nucleic Acids Res 25, 3389-3402.

Assou, S., Le Carrour, T., Tondeur, S., *et al*., 2007. A meta-analysis of human embryonic stem cells transcriptome integrated into a web-based expression atlas. Stem Cells 25, 961-973.

Berglund, A.C., Sjolund, E., Ostlund, G., Sonnhammer, E.L.L., 2008. InParanoid 6: eukaryotic ortholog clusters with inparalogs. Nucleic Acids Res 36, D263-266.

Boiani, M., Scholer, H.R., 2005. Regulatory networks in embryo-derived pluripotent stem cells. Nat Rev Mol Cell Biol 6(11), 872-84.

Brown, K.R., Jurisica, I., 2005. Online predicted human interaction database. Bioinformatics 21(9), 2076-82.

Brown, K.R., Jurisica, I., 2007. Unequal evolutionary conservation of human protein interactions in interologous networks. Genome Biol 8(5), R95.

Chen, F., Mackey, A.J., Vermunt, J.K., Roos, D.S., 2007. Assessing performance of orthology detection strategies applied to eukaryotic genomes. PLoS ONE 2, e383.

Chia, N.Y., Chan, Y.S., Feng, B., *et al*., 2010. A genome-wide RNAi screen reveals determinants of human embryonic stem cell identity. Nature 468, 316-320.

Cohen, D.E., Melton, D., 2011. Turning straw into gold, directing cell fate for regenerative medicine. Nature Reviews Genetics 12, 243-252.

Done, B., Khatri, P., Done, A., Draghici, S., 2010. Predicting novel human Gene Ontology annotations using semantic analysis. IEEE/ACM Transactions on Computational Biology and Bioinformatics 7(1), 91-99.

Enault, F., Suhre, K., Abergel, C., Poirot, O., Claverie, J.M., 2003. Annotation of bacterial genomes using improved phylogenomic profiles. Bioinformatics 19(Suppl 1), i105-107.

Ewing, R.M., Chu, P., Elisma, F., Li, H., *et al*., 2007. Large-scale mapping of human protein–protein interactions by mass spectrometry. Mol Syst Biol 3, 89.

Frohlich, H., Speer, N., Poustka, A., BeiSZbarth, T., 2007. GOSim – an R-package for computation of information theoretic GO similarities between terms and gene products. BMC Bioinformatics 8, 166.

Greber, B., Lehrach, H., Adjaye, J., 2007. Silencing of core transcription factors in human EC cells highlights the importance of autocrine FGF signaling for self-renewal. BMC Dev Biol 7, 46.

Guo, X., Liu, R., Shriver, C.D., Hu, H., Liebman, M.N., 2006. Assessing semantic similarity measures for the characterization of human regulatory pathways. Bioinformatics 22, 967-973.

Guo, Y., Yu, L., Wen, Z., Li, M., *et al*., 2008. Using support vector machine combined with auto covariance to predict protein-protein interactions from protein sequences. Nucleic Acids Res 36, 3025-3030.

Haider, S., Ballester, B., Smedley, D., Zhang, J., Rice, P., Kasprzyk, A., 2009. BioMart central portal–unified access to biological data. Nucleic Acids Res 37, 23-27.

Han, K., Park, B., Kim, H., Hong, J., Park, J., 2004. HPID: the human protein interaction database. Bioinformatics 20(15), 2466-70.

Hanna, J., Cheng, A.W., Saha, K., *et al*., 2010. Human embryonic stem cells with biological and epigenetic characteristics similar to those of mouse ESCs. Proc Natl Acad Sci USA 107, 9222-9227.

Huang, T.W., Lin, C.Y., Kao, C.Y., 2007. Reconstruction of human protein interolog network using evolutionary conserved network. BMC Bioinformatics 8, 152.

Hubbell, E., Liu, W.M., Mei, R., 2002. Robust estimators for expression analysis. Bioinformatics 18, 1585-1592.

Ideker, T., Ozier, O., Schwikowski, B., Siegel, A.F., 2002. Discovering regulatory and signalling circuits in molecular interaction networks. Bioinformatics 18(Suppl 1), S233-240.

Jaenisch, R., Young, R.A., 2008. Stem Cells, the Molecular Circuitry of Pluripotency and Nuclear Reprogramming. Cell 132, 567-582.

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

Jothi, R., Kann, M.G., Przytycka, T.M., 2005. Predicting protein-protein interaction by searching evolutionary tree automorphism space. Bioinformatics 21(Suppl 1), i241-250.

Katoh, M., 2008. WNT signaling in stem cell biology and regenerative medicine. Curr Drug Targets 9, 565-570.

Lam, H., Patel, S., Wong, J., Chu, J., Lau, A., Li, S., 2008. Localized decrease of ß-catenin contributes to the differentiation of human embryonic stem cells. Biochem Biophys Res Commun 372, 601-606.

Lehner, B., Fraser, A.G., 2004. A first-draft human protein-interaction map. Genome Biol 5(9), R63.

Matthews, L.R., Vaglio, P., Reboul, J., *et al*., 2001. Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "interologs". Genome research 11, 2120-2126.

Mika, S., Rost, B., 2006. Protein-protein interactions more conserved within species than across species. PLoS Comput Biol 2(7), e79.

Ng, S.K., Zhang, Z., Tan, S.H., 2003. Integrative approach for computationally inferring protein domain interactions. Bioinformatics 19(8), 923-929.

Okita, K., Yamanaka, S., 2006. Intracellular signaling pathways regulating pluripotency of embryonic stem cells. Curr Stem Cell Res Ther 1(1), 103-11.

Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D., Yeates, T.O., 1999. Assigning protein functions by comparative genome analysis, protein phylogenetic profiles. Proc Natl Acad Sci USA 96(8), 4285-8.

Pera, M.F., Tam, P.P., 2010. Extrinsic regulation of pluripotent stem cells. Nature 465(7299), 713-720.

Persico, M., Ceol, A., Gavrila, C., *et al*., 2005. HomoMINT: an inferred human network based on orthology mapping of protein interactions discovered in model organisms. BMC Bioinformatics 6(Suppl 4), S21.

Rao, M., 2004. Conserved and divergent paths that regulate self-renewal in mouse and human embryonic stem cells. Dev Biol 275, 269-286.

Rao, S., Orkin, S.H., 2006. Unraveling the transcriptional network controlling ES cell pluripotency. Genome Biol 7, 230.

Resnik, P., 1999. Semantic similarity in a taxonomy, an information-based measure and its application to problems of ambiguity in natural language. J Artif Intell Res 11, 95-130.

Rhodes, D.R., Tomlins, S.A., Varambally, S., *et al*., 2005. Probabilistic model of the human protein-protein interaction network. Nat Biotechnol 23(8), 951-959.

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

Schlicker, A., Domingues, F.S., Rahnenfuhrer, J., Lengauer, T., 2006. A new measure for functional similarity of gene products based on Gene Ontology. BMC Bioinformatics 7, 302.

Schnerch, A., Cerdan, C., Bhatia, M., 2010. Distinguishing between mouse and human pluripotent stem cell regulation, the best laid plans of mice and man. Stem Cells 28, 419-30.

Shen, J., Zhang, J., Luo, X., *et al*., 2007. Predicting protein-protein interactions based only on sequences information Proc Natl Acad Sci USA 104(11), 4337-4341.

Shin, C.J., Wong, S., Davis, M.J., Ragan, M.A., 2009. Protein–protein interaction as a predictor of subcellular location. BMC Syst Biol 3, 28.

Som, A., Harder, C., Greber, B., *et al*., 2010. The PluriNetWork, an in-silico representation of the network underlying pluripotency in mouse, and its applications. PLoS ONE 5(12), e15165.

Sprinzak, E., Sattath, S., Margalit, H., 2003. How reliable are experimental protein-protein interaction data? J Mol Biol 327(5), 919-23.

Sun, Y., Li, H., Yang, H., Rao, M.S., Zhan, M., 2006. Mechanisms controlling embryonic stem cell self-renewal and differentiation. Crit Rev Eukaryot Gene Expr 16(3), 211-31

Tesar, P.J., Chenoweth, J.G., Brook, F.A., *et al*., 2007. New cell lines from mouse epiblast share defining features with human embryonic stem cells. Nature 448, 196-199.

Tirosh, I., Barkai, N., 2005. Computational verification of protein-protein interactions by orthologous co-expression. BMC Bioinformatics 6, 40.

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

von Mering, C., Krause, R., Snel, B., *et al*., 2002. Comparative assessment of large-scale data sets of protein-protein interactions. Nature 417, 399-403.

Walhout, A.J., Sordella, R., Lu, X., *et al*., 2000. Protein interaction mapping in C elegans using proteins involved in vulval development. Science 287, 116-122.

Wang, J.Z., Du, Z., Payattakool, R., Yu, P.S., Chen, C.F., 2007. A new method to measure the semantic similarity of go terms. Bioinformatics 23, 1274-1281.

Warsow, G., Greber, B., Falk, S., *et al*., 2010. ExprEssence - Revealing the essence of differential experimental data in the context of an interaction/regulation network. BMC Systems Biology 4, 164.

Wu, X., Zhu, L., Guo, J., Zhang, D.Y., Lin, K., 2006. Prediction of yeast protein-protein interaction network, insights from the Gene Ontology and annotations. Nucleic Acids Res 34(7), 2137-2150.

Xie, C.Q., Jeong, Y., Fu, M., *et al*., 2009. Expression profiling of nuclear receptors in human and mouse embryonic stem cells. Mol Endocrinol 23, 724-733.

Yellaboina, S., Goyal, K., Mande, S.C., 2007. Inferring genome-wide functional linkages in E coli by combining improved genome context methods, comparison with high-throughput experimental data. Genome research 17, 527-535.

Yu, H., Braun, P., Yildirim, M.A., *et al*., 2008. High-quality binary protein interaction map of the yeast interactome network. Science 322(5898), 104-110.

Yu, H., Luscombe, N.M., Lu, H.X., *et al*., 2004. Annotation transfer between genomes, protein-protein interologs and protein-DNA regulogs. Genome research 14, 1107-1118.

Table 1: The number of links (genes/proteins) in the pluripotency network at various steps of our approach to transfer the mouse pluripotency network to human.

Step	Number
Interactions identified experimentally in mouse	547 (264)
Interologs transferred to human based on orthologous relationship	545 (262)
Interactions for which phylogenetic profiling data was available	545
Interactions for which GO semantic similarity data was available	480
Interactions for which gene co-expression data was available	453
Interactions for which RNAi data was available	540
Interactions for which data for all four link evaluation methods was available	406
Interactions remaining after filtering by the RNAi method	215 (148)
Interactions remain in the final network	196 (136)

Table 2: Genes experimentally shown to be required, or shown not to be required for the induction and/or maintenance of pluripotency in human.

Genes required	References	Genes not required	References
ACVR1	Schnerch *et al*. (2010)	CTNNB1	Lam *et al*. (2008)
DNMT3B	Adewumi *et al*. (2007)	ESRRB	Xie *et al*. (2009)
DPPA4	Assou *et al*. (2007)	FBX15	Rao (2004)
FGFR1	Schnerch *et al*. (2010)	IL6ST	Schnerch *et al*. (2010)
HELLS	Assou *et al*. (2007)	JAK1	Schnerch *et al*. (2010)
KLF4	Pera and Tam (2010)	KLF2	Greber *et al*. (2007)
LEFTY1	Schnerch *et al*. (2010)	KLF5	Greber *et al*. (2007)
NANOG	Pera and Tam (2010)	LIF	Pera and Tam (2010)
NODAL	Pera and Tam (2010)	NR5A2	Xie *et al*. (2009)
PHF17	Assou *et al*. (2007)	SMAD1	Schnerch *et al*. (2010)
POU5F1	Pera and Tam (2010)	STAT3	Schnerch *et al*. (2010)
SMAD2	Schnerch *et al*. (2010)	TBX3	Greber *et al*. (2007)
SOX2	Pera and Tam (2010)		
TGFB1	Schnerch *et al*. (2010)		
ZIC3	Assou *et al*. (2007)		

Table 3: Correctness scores for the inclusion of a link between genes in the human pluripotency network or the exclusion from the network, depending on whether the genes that are linked are known to be involved in human pluripotency based on the literature.

Involvement in human pluripotency based on the literature		**Inclusion in the human network**	**Exclusion from the human network**
First gene	**Second gene**		
yes	yes	1	0
no	no	0	1
yes	no	0	1
yes	unknown	0.72	0.28
no	unknown	0	1
unknown	unknown	0	0

Table 4: The nodes deleted by the 80% criterion (i.e., a node was deleted when at least 80% of the links attached to it were filtered out).

Gene name	**Total no. of links attached to the gene**	**No. of links deleted**	**Fraction of links deleted**
ESRRB	16	14	87%
KLF2	5	4	80%
KLF5	11	10	91%
MBD3	6	5	83%
NR5A2	23	19	82%
SALL4	12	10	83%
SMAD1	11	9	82%
MYC	15	13	86%
STAT3	32	30	94%
CTNNB1	7	6	86%

Figures

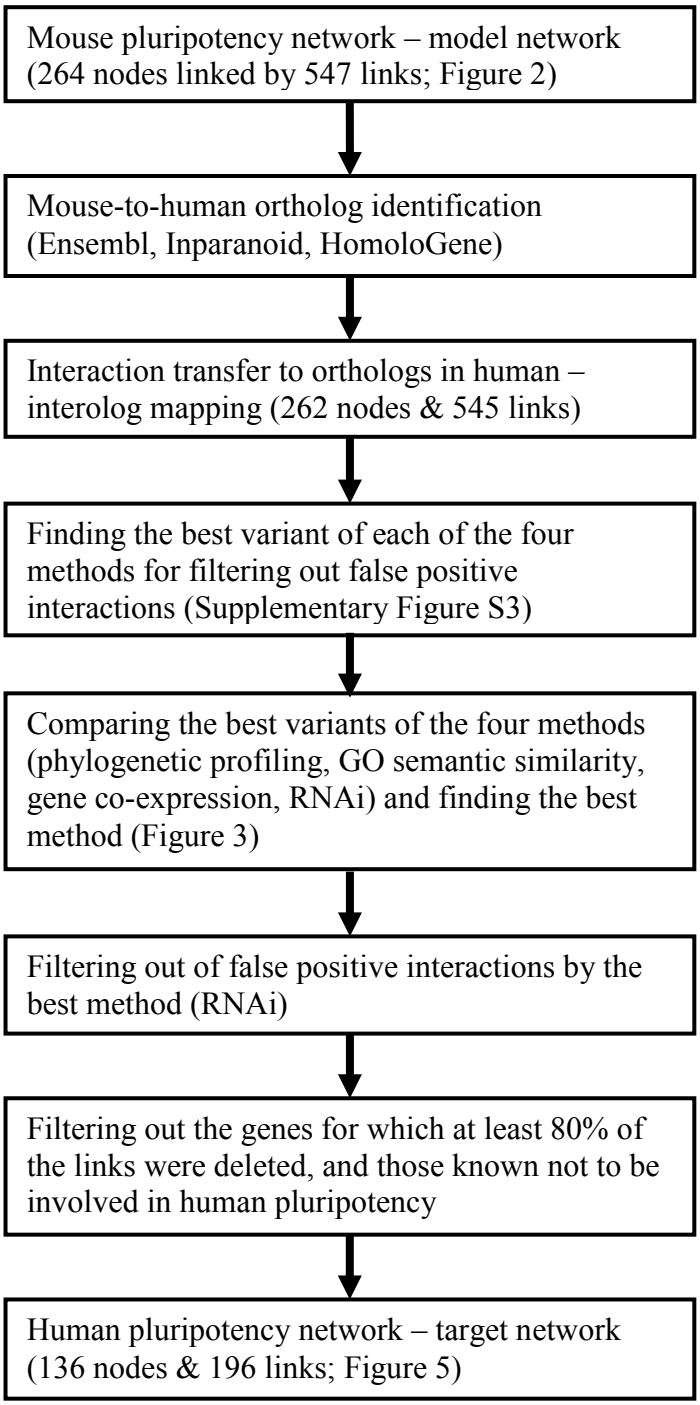

Fig. 1: Flowchart of the approach used for the derivation of human pluripotency network.

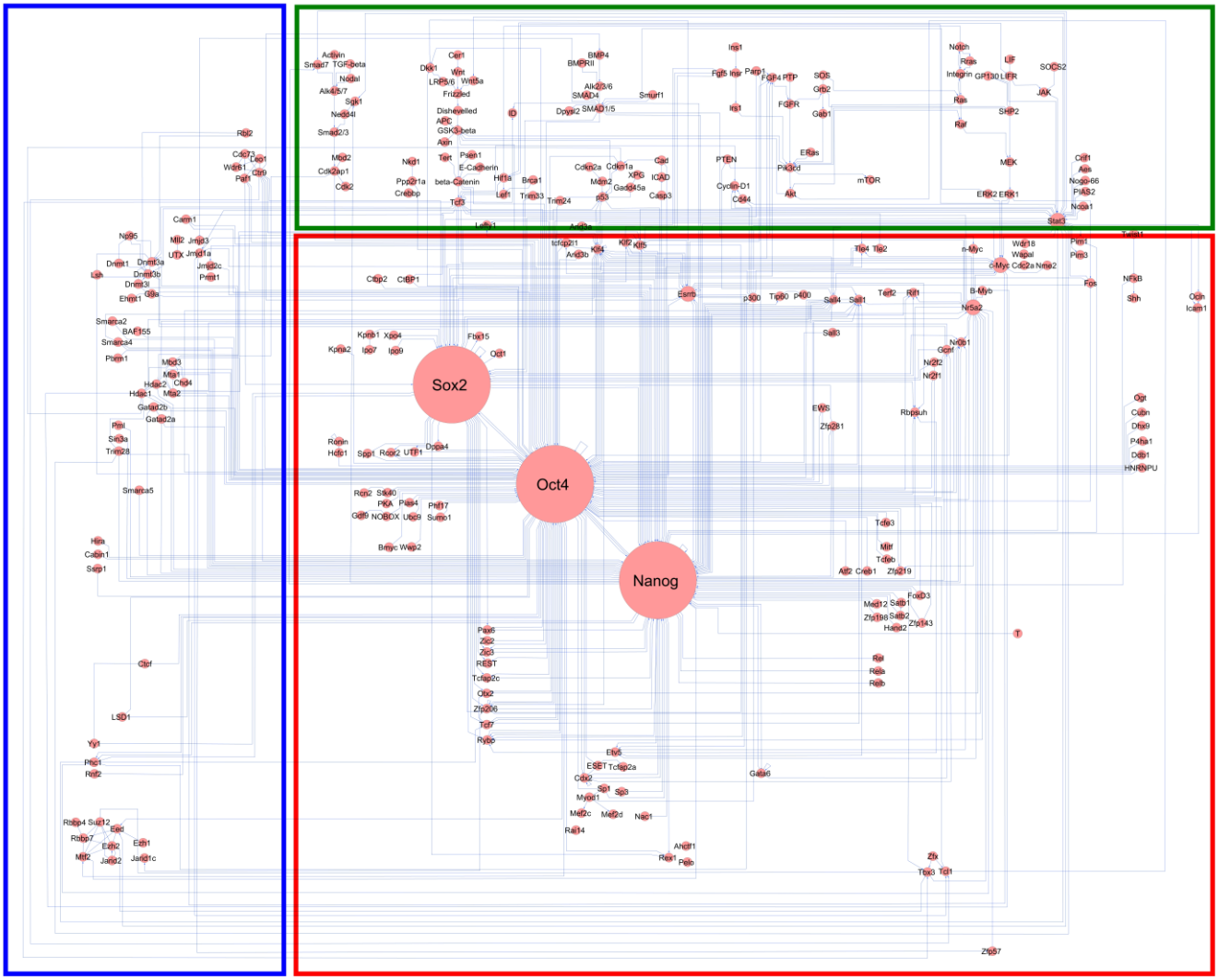

Fig. 2: Manual layout of the *PluriNetWork* in Cytoscape. Nodes (264) are genes/proteins, edges (547) are stimulations (arrows), inhibitions (T-bar arrows) and interactions (lines). The top third of the network includes upstream signaling pathways (green region), the middle is composed of the core circuitry of pluripotency (Pou5f1 – also known as Oct4, Sox2 and Nanog) and its periphery (red region), and the left part includes epigenetic factors and related mechanisms (blue region). A high-resolution JEPG image and a Cytoscape version of the network are presented in Supplementary Fig. S1.

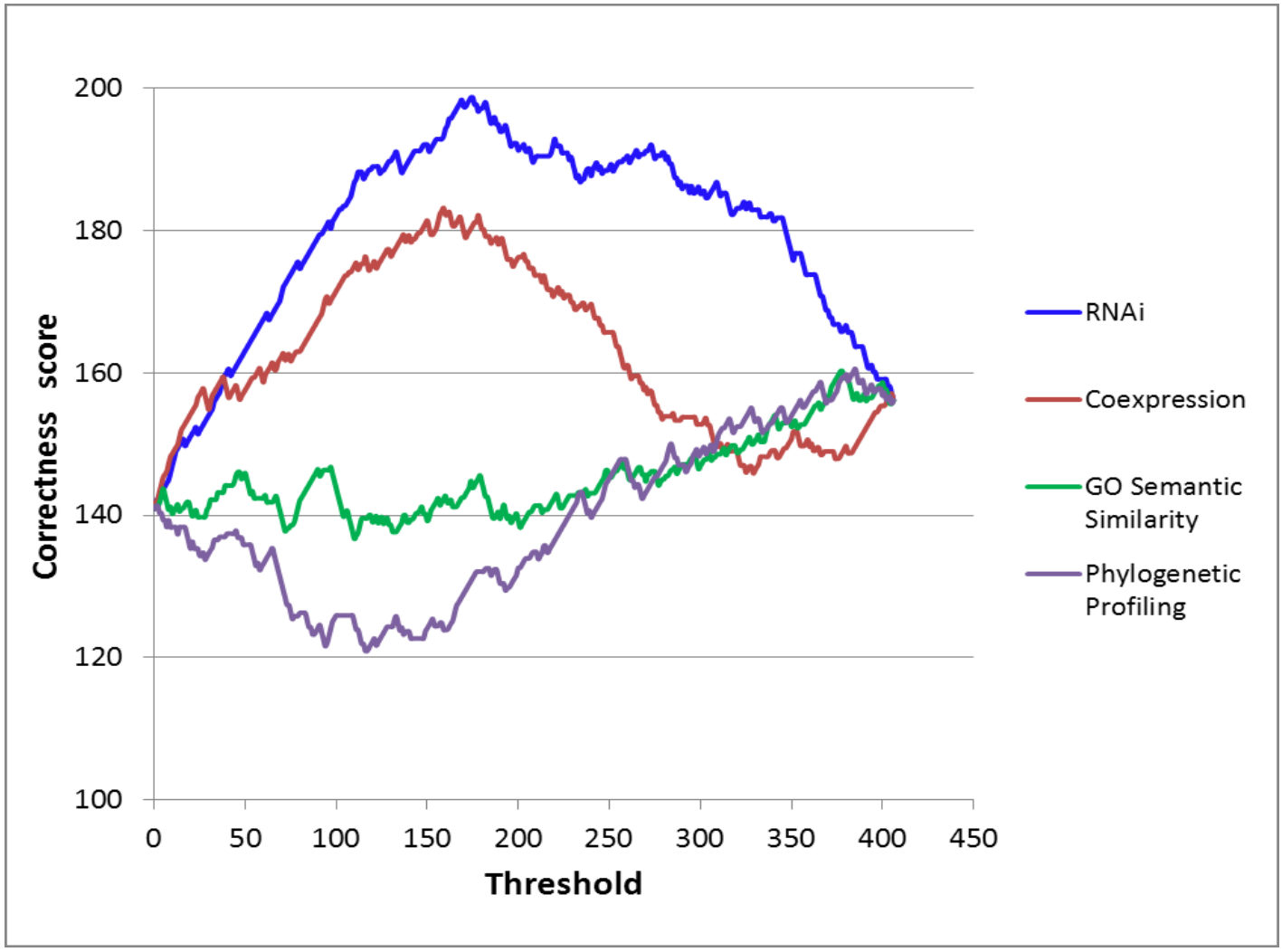

Fig. 3: Comparison of the interolog filtering methods. Threshold versus correctness score is plotted for the four link evaluation methods. A higher correctness score indicates that the method better filters out possible false positive interactions. The plot shows that the RNAi method is the best method followed by the co-expression method.

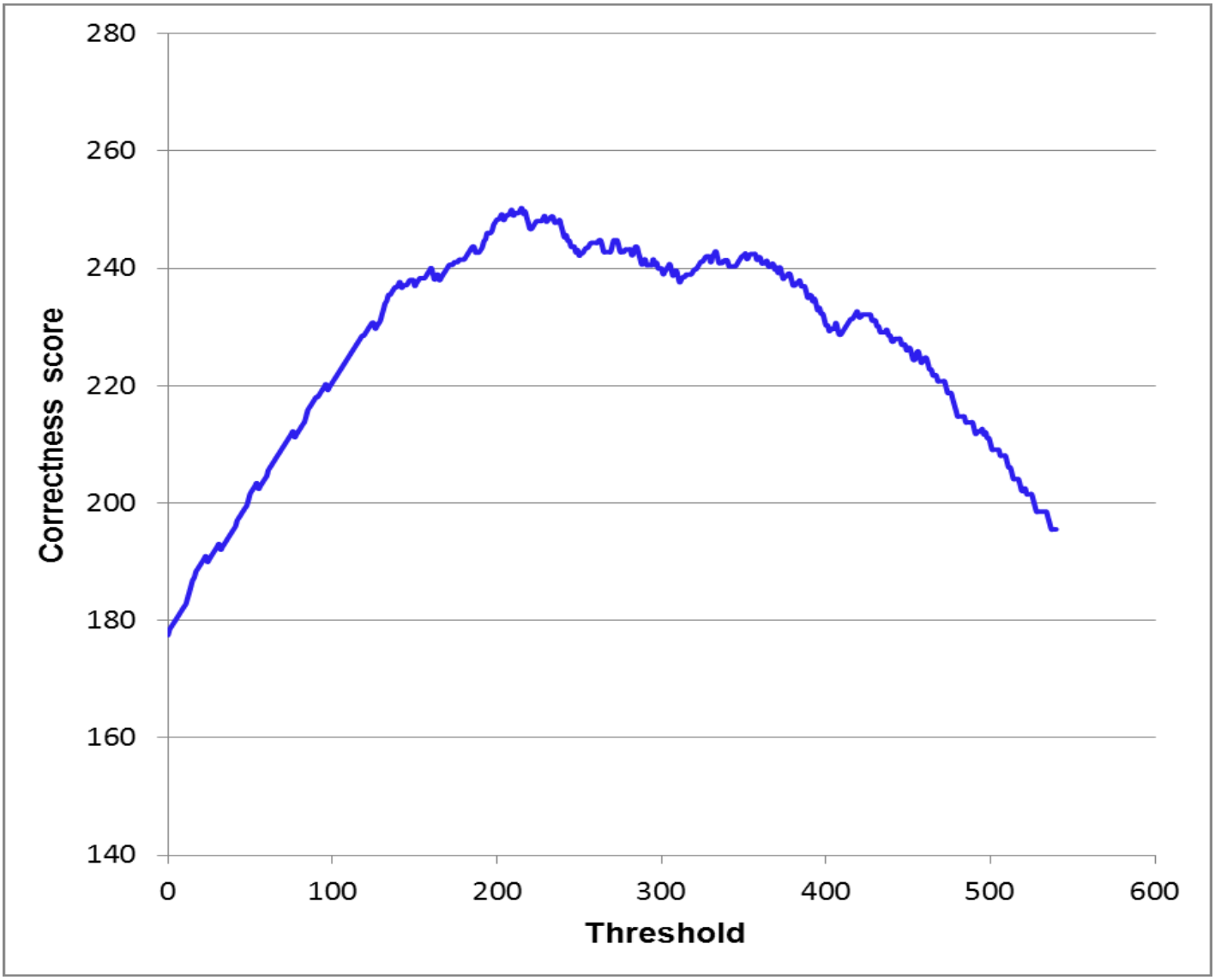

Fig. 4: Threshold versus correctness score for the RNAi method for all the links for which we had RNAi information, which is 540 out of 545 links. The correctness score reaches the highest value (250.29) when the threshold is 215, which indicates the top-ranking 215 links are to be retained in the human network.

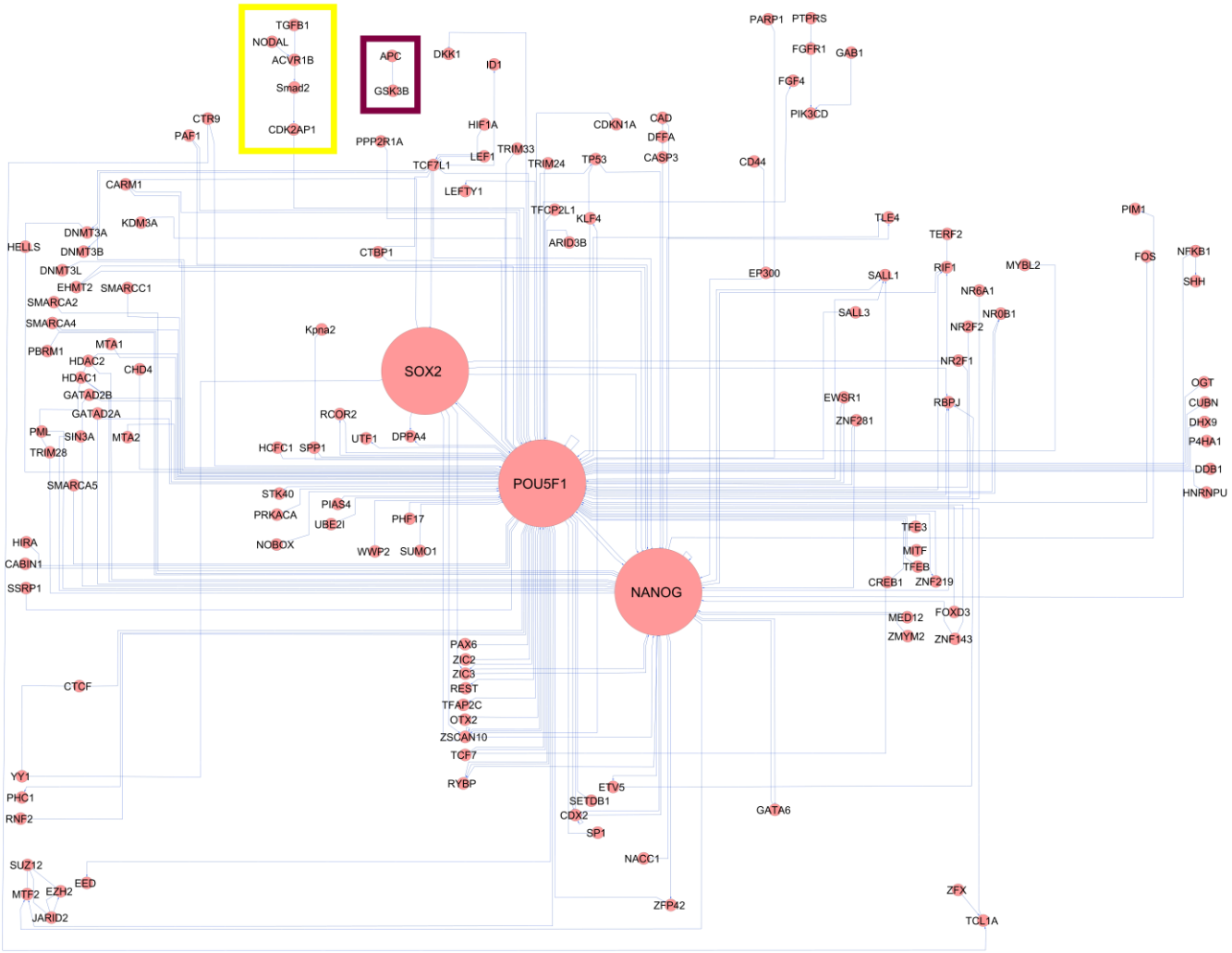

Fig. 5: The derived human pluripotency network after filtering out putative false positives. This network contains 136 nodes (genes/proteins) and 196 links. A high-resolution JPEG image and a Cytoscape version of the network are found in Supplementary Fig. S4.

Supplementary materials

Table S1: The list of human orthologs of mouse pluripotency genes/proteins. The mouse-human orthologous relationships were established by combining the orthologous information from Ensembl, InParanoid, and HomoloGene databases.

Table S2: The list of the 14 genomes used for the phylogenetic profiling method.

Table S3: A set of the 406 links used to evaluate the relative performance of the four methods by their best variants. For each link and for each of the four methods, evolutionary dissimilarity score (EDS), GO semantic similarity score, co-expression LinkScore, and RNAi score and their respective ranks are given. Rank 1 indicates the link has the highest probability that it is involved in the human pluripotency network.

Table S4: The links (196) and nodes (136) retained in the final predicted human pluripotency network.

Fig. S1: (a) A high-resolution JPEG image of the mouse *PluriNetWork*, which is used as the model network and (b) a Cytoscape version of the mouse *PluriNetWork*.

Fig. S2: A Cytoscape version of the initial predicted human pluripotency network (links transferred on the basis of mouse-human orthologous relationships). This network contains 262 nodes (genes/proteins) and 545 links (stimulation, inhibition, and interaction).

Fig. S3: The comparison of the variants of each of the four link evaluation methods. Threshold versus correctness score is plotted for each variant. A higher correctness score indicates that the variant better filters out possible false positive interactions.

Fig. S4: The final human pluripotency network in (a) JEPG image file and (b) Cytoscape format.