Preimplantation development regulatory pathway construction through a text-mining approach

Elisa Donnard[1§], Adriano Barbosa-Silva[2], Rafael L. M. Guedes[1], Gabriel R. Fernandes[1], Henrique Velloso[1], Matthew J Kohn[3], Miguel A. Andrade-Navarro[2], J. Miguel Ortega[1]

[1] Laboratório Biodados, Dept. de Bioquímica e Imunologia, Universidade Federal de Minas Gerais, 31270-901, Belo Horizonte - MG, Brazil.

[2] Computational Biology and Data Mining Group, Max-Delbrück Center for Molecular Medicine, Robert-Rössle-Strasse 10, D-13125, Berlin, Germany.

[3] New York State Stem Cell Science, New York State Department of Health Wadsworth Center, Rm C345.

[§]Corresponding author

Email addresses:

ED: edonnard@compbio.ludwig.org

ABS: adriano.barbosa@mdc-berlin.de

RLMG: rafaelmguedes@ufmg.br

GRF: fernandes.gabriel@gmail.com

HV: hvmelo@gmail.com

MJK: mjk09@health.state.ny.us

MAAN: miguel.andrade@mdc-berlin.de

JMO: miguel@ufmg.br

Abstract

Background

The integration of sequencing and gene interaction data and subsequent generation of pathways and networks contained in databases such as KEGG Pathway is essential for the comprehension of complex biological processes. We noticed the absence of a chart or pathway describing the well-studied preimplantation development stages; furthermore, not all genes involved in the process have entries in KEGG Orthology, important information for knowledge application with relation to other organisms.

Results

In this work we sought to develop the regulatory pathway for the preimplantation development stage using text-mining tools such as Medline Ranker and PESCADOR to reveal biointeractions among the genes involved in this process. The genes present in the resulting pathway were also used as seeds for software developed by our group called SeedServer to create clusters of homologous genes. These homologues allowed the determination of the last common ancestor for each gene and revealed that the preimplantation development pathway consists of a conserved ancient core of genes with the addition of modern elements.

Conclusions

The generation of regulatory pathways through text-mining tools allows the integration of data generated by several studies for a more complete visualization of complex biological processes. Using the genes in this pathway as “seeds” for the generation of clusters of homologues, the pathway can be visualized for other organisms. The clusters of homologous genes also allow determination of the ancestry

of the genes involved in the process, leading to a better understanding of the evolution of such process.

Background

Bioinformatics tools currently allow research to focus on the integration of large-scale data generated by sequencing, differential expression analysis, gene interaction studies and others. Several initiatives exist to organize this knowledge in secondary databases, thus allowing easier access and visualization. Databases containing interaction information are a good source for novel research. iHOP [1] allows users to tag gene names of interest and browse through the related PubMed literature with highlighted keywords. Another interaction database is STRING [2], which contains physical interactions and functional associations between proteins and integrates data retrieved from literature (PubMed), genomic context, large scale experiments and conserved co-expression. Text-mining, therefore, has a fundamental role in these tools and allows access to interactions spread throughout the literature. The extraction of biological events from literature through text-mining tools is essential to not only update the interaction databases but also for the creation and annotation of pathways.

Metabolic and regulatory pathways are an example of organized knowledge that allow a better visualization of a complex system and can be found in databases such as iPath [5], BioCyc [6] or KEGG Pathways [7]. When orthology information is added to pathways, the same process can be represented in different organisms. Orthology is also an important tool for sequence annotation. Current orthologue databases such as COG and KOG [3], eggNOG [4], OrthoMCL [5] and KEGG Orthology [6] all provide a good source for manually curated clusters of orthologues defined for organisms with complete genomes. We developed a procedure to enrich

the COG database with UniRef50 clusters from the UniProt database [7], creating the UECOG database [8]. Recently, a similar procedure was applied to the KEGG Orthology database creating the enriched UEKO database (unpublished, Fernandes *et al*.).

The available tools described raise the possibility of integrating current information and generating complex regulatory pathways. Previous publications individually reported the regulatory interactions that control preimplantation embryo development. However, a complete preimplantation development regulatory pathway has never been built.

In humans, the preimplantation phase of embryonic development is a period of approximately six days after fertilization prior to attachment of the embryo to the uterine wall. Implantation can occur before or in the seventh embryonic day (E7), a time during which the uterus is receptive [9]. Mammalian embryonic development has been thoroughly studied in mice and the blastomeres remain totipotent, able to generate any other cell, up to the eight-cell stage, unlike other animals [10]. After fertilization, successive cleavages take place during the first two days of development, resulting in the eight-cell embryo. The next stage of development is called the morula stage. An increase in cell-cell contact results in formation of a compacted morula. The subsequent divisions increase the complexity of the embryo and cells may be located on the inside, surrounded by other cells, or on the outside, in contact with the environment. The identification of the initial cells for each lineage has shown that the trophectoderm (TE) is derived mostly from the outer cells, whereas the inner cells give rise to the inner cell mass (ICM). Later, the ICM divides into the primitive endoderm (PE) and the epiblast (EPI). During the differentiation of the TE from the ICM, the blastocoel is formed through a process of cavitation. The embryo is called a

blastocyst when all three structures are present (TE, ICM and blastocoel). Twenty-four hours after blastocyst formation occurs, the last stage of preimplantation development takes place when the PE differentiates from the ICM. The three lineages thus formed in preimplantation development present different fates during subsequent embryonic development. While the epiblast, which forms from the ICM following implantation, is still undifferentiated and will give rise to the fetus itself, the trophectoderm will become the fetal portion of the placenta and the primitive endoderm (as part of the extraembryonic endoderm) will form the yolk sac [11]. Complex regulatory processes such as animal development are a result of the interaction of many different gene products and elements that control the expression of these genes. Traditional experiments that determine the function of one or a few genes are essential, but do not result in a comprehensive view of complex systems. A complex regulatory network should be able to portray specific and general aspects of development, such as the embryonic fate of certain cells [12].

In this work, we noticed the absence in databases of a pathway describing the preimplantation phase of embryo development and sought to develop the given pathway using text-mining tools, complementing it with orthology information. The resulting pathway comprises 86 genes and the interactions between them. Clusters of orthologous groups were generated for each gene represented in the pathway and provided the necessary information to determine the last common ancestor. This determination revealed that the preimplantation development pathway is an ancient Chordata pathway with addition of modern elements throughout evolution.

Results

Text-mining

Initially, we used the PubMed platform to search for articles related to the embryo preimplantation development (query: “preimplantation development”) and obtained 3524 entries as a result. To obtain a more efficient set of articles with relevancy to our work, the result entries were submitted to MedlineRanker [13]. This software computes discriminating words by comparing a set of user selected abstracts indicated as highly relevant to a background set and then scores any abstracts in terms of their content of those discriminating words. After the classification, we selected the top 1000 abstracts for further analysis. Since human and mouse embryo development are highly similar, we decided to use abstracts from work on both organisms as source of information for the preimplantation pathway construction.

Using these 1000 highly informative abstracts as our input, PESCADOR (manuscript in preparation, Barbosa-Silva *et al*.) an online platform for the LAITOR software [14] was used for tagging of gene names and biointeractions extraction from each abstract. As a result, 722 gene names were tagged and 223 type 1 biointeractions were highlighted as well as other informative biointeractions. Biointeractions are classified by LAITOR as type 1 when in the same sentence the software encounters a gene name, a biointeraction word and another gene name, in that order (e.g.: CDX2 downregulates NANOG). From these tagged abstracts we manually curated the information and constructed the pathway for the preimplantation embryo development describing 86 genes and numerous interactions between them during the early developmental stages, trophectoderm differentiation from the inner cell mass and posterior extraembrionary endoderm differentiation. A sample abstract tagged by PESCADOR and the manual extraction of the information it contains is exemplified

in Figure 1. The pathway shown in Figure 2 was constructed according to KGML (KEGG Markup Language).

Preimplantation Pathway

The pathway obtained after the analysis of all the abstracts from the PESCADOR output is represented in Figure 2 and the regulations are reviewed below.

First embryonic cleavages

The oncogene c-MYC is an important transcriptional regulator and its expression is observed in the initial stages of development, where it is present in embryonic cells until the morula stage and repressed thereafter [15]. Two additional genes recently associated with these early developmental stages are BORIS and ECSA. BORIS is involved in early development following fertilization and soon afterwards repressed, and ECSA, expression begins in the blastocyst exclusively in the cells of the inner cell mass (ICM). The presence of these genes was compared to the expression pattern of the Oct4 transcription factor, which is present in the early cleavages, repressed after this initial stage, and then its expression is afterwards stimulated again in the blastocyst [16]. The expression of the gametogenesis associated gene Gse was also recently identified in cells of the early embryo; later this protein is found only in the ICM, suggesting a role in the specification of cell lineage [17].

Methylation patterns and correct preimplantation development

Genomic methylation patterns in mammalian cells depend on Dnmt1 (DNA methyltransferase-1). In the mouse, an embryo-specific variant called Dnmt1o is

expressed in the early stages of development. In the 8-cell stage this protein relocates to the cell nucleus where maintains essential methylation patterns, allowing embryos to complete early developmental events [18]. It was recently shown [19] that the inability of Dnmt1o to properly relocate not only results in a developmental arrest at the 5-7 cell stage, but is also responsible for the downregulation of five genes involved in the formation of gap and tight junctions (Cx31, Cx43, Cx45, Cdh1 and Ctnnb1). These junctions are crucial for early processes such as compaction of the 8-cell embryo and cavitation of the blastocoel.

TE versus ICM dichotomy: key role of Lats controlling Tead4 co-activator Yap

Cells destined to become part of the ICM are marked by repression of two genes (aPKC and Par3) [20] and by upregulation of Sox2 [21]. In these cells, the major pluripotency transcription factors, including Nanog and Oct4, remain active due to the expression of an important player and member of the Hippo signaling pathway: Lats. This serine/threonine protein kinase is responsible for phosphorylating Yap, leading to its cytoplasmic localization and thus preventing its association with the transcription factor Tead4.

Triggering TE differentiation: Tead4/Yap target Cdx2 to repress Nanog and Oct4

Conversely, in the outer cells that will differentiate and form the trophectoderm, Yap is unphosphorylated, remains in the nucleus and associates with Tead4, leading to the activation of Cdx2, a key repressor of Nanog and Oct4 [22]. Repression of Oct4 and Nanog transcription by Cdx2 then releases the inhibition that these two key factors were exerting on many different genes, in turn activating these targets [23, 24]. Activation of Cdx2 requires release from basal repression; Nanog

[25] and Oct4 [26] repress basal levels of Cdx2 and induction of higher levels of Cdx2 by Tead4/Yap overcomes this repression, allowing Cdx2 to play its role [23]. Tead4 was also recently determined to activate another trophectoderm differentiation factor, GATA3 [27], which acts alongside Cdx2 and affects transcription of a number of genes independent of Cdx2. The Tead4-dependent activation of GATA3 seems to be independent of Yap, suggesting Tead4 interacts with another partner as well as Yap. Also required for high level expression of Cdx2 in trophectoderm cells is the cell motility protein Arp3; experiments with complete knockdown of this protein show trophoblast cells unable to develop properly, possibly undergoing apoptosis as a result of loss of Cdx2 [28]. The TGFbeta pathway is another important pathway for trophectoderm differentiation; TGFbeta signaling is stimulated by BMP4, which leads to the activation of SMAD proteins. These proteins can also stimulate transcription of Cdx2 [29], and BMP4 is known to inhibit Id2, an inhibitor of differentiation [30], and to activate Hand1, which is involved in trophoblast cell differentiation [31].

In the absence of Oct4 and Nanog

The downregulation of Oct4 in the outer cells of the embryo leads to the activation of a positive regulator of TE cell fate, Eomes (T-box protein eomesodermin) [24, 32], which is also a possible Cdx2 target [33]. The subsequent differentiation of these cells into trophectoderm is accompanied by the expression of several genes, such as the glycoprotein PSG2 [34] and the marker KRT18. PSG2 and KRT18 expression are among the first signs that a blastomere has lost its totipotent competence, prior to any visible differentiation [28]. Removal of Oct4-dependent repression also results in activation of genes such as ETIF2B and Rps14 [35], allowing these cells to engage in an intense translation routine. Knockdown studies

targeting Oct4 also show that it represses the expression of Gcm1, which is normally placenta specific [36], and of the hCG hormone's beta chain [37].

Concurrently, Nanog downregulation allows the expression of a number of genes associated with both trophectoderm (GATA2, hCG-alpha and hCG-beta) and extraembryonic endoderm (GATA4, GATA6, LAMB1 and AFP) [25]. These latter genes will in turn initiate the formation of tissues such as the primitive endoderm, a component of the yolk sac. From the early blastocyst stage on, desmosomes are assembled in the trophectoderm in response to desmocollin (DSC2), which is also not expressed in the ICM [38].

Thus, Tead4/Yap activation of Cdx2, accompanied by the subsequent repression of Nanog and Oct4, describes a scenario for the TE differentiation.

Underneath the maintained activation of Oct4 and Nanog

Back in the ICM, the main pluripotency genes remain active and form a complex regulation pathway. Recently it was discovered that transcription of Nanog is further stimulated by the presence of compounds such as retinol [39]. Klf2, Klf4 and Klf5 exert a redundant role in the activation of Nanog. These krüppel-like factors were described as essential for the maintenance of pluripotency. Indeed, Klf4 was already known for this role and is commonly used in reprogramming of differentiated cells into induced pluripotent stem cells. However, only the simultaneous depletion of Klf4, 2 and 5 results in the differentiation of stem cells, indicating functional redundancy [40]. Other proteins known to activate Nanog include the two other main pluripotency regulators, Oct4 [32, 41] and Sox2 [42]. The estrogen receptor ESRRB is also reported to be involved in the activation of Nanog by Oct4 and Sox2 [42].

Conversely, Nanog can activate Oct4 [41], and ESRRB is necessary to maintain Oct4 promoter activity [43].

Each of the three key factors, Oct4, Sox2 and Nanog, also act as self-activators, e.g. the partners Oct4 and Sox2 bind and activate Oct4 transcription [44]. Another key transcription factor involved in the maintenance of cell pluripotency is Sall4 [45]. Sall4 binds to the conserved regulatory region in the Pou5f1 (the Oct4 gene) distal enhancer and activates its transcription [26]. Studies with microRNA interference of Sall4 show that the loss of this factor leads to reduction of Oct4 mRNA levels and significant expression of Cdx2 in the ICM [26]. b-MYB, a gene expressed in proliferating cells, is also a positive regulator of Oct4 and studies report early differentiation of ICM in the absence of b-MYB [46].

The Notch signaling pathway is a conserved pathway that is involved in cellular communication processes and correct cell fate decisions that also has a role in ICM development [47]. Nle protein, a direct regulator of this pathway, is essential for survival of the ICM [48]. Another protein associated with development and survival of the ICM is Tbn (Taube nuss), whose absence promotes cell apoptosis in the ICM [49].

Expression of the platelet and endothelial cell adhesion molecule (PECAM1 or CD31) was detected by immunofluorescence confocal microscopy in the blastocyst and restricted to the ICM cells. Subsequently, PECAM1 remains only in the pluripotent epiblast cells, disappearing the moment these cells undergo differentiation [50], and indicating a new role for this molecule during embryo development.

Activation, but with moderation

Other control pathways maintain expression of these genes at a steady-state concentration and balance these many mechanisms for activation and upregulation of transcription. A complex regulation feedback loop consists of FOXD3, Nanog and Oct4 [41]. To keep Oct4 and Nanog expression within steady-state levels, these three genes interact so that (i) expression of Nanog activates FOXD3 and Oct4 but not above steady-state levels due to Oct4 exerted repression; and (ii) FOXD3 and Nanog activate Oct4 expression but not above steady-state levels due to Oct4 self-repression. Dax1 is an orphan nuclear hormone receptor recently identified as a repressor of Oct4 transcription [51].

Dax1 expression was also capable of reducing Nanog and Rex1 expression. Assays show that Dax1 binds to Oct4 and abolishes its DNA binding activity, thus decreasing the transcription of Nanog and Rex1, targets of Oct4 activation.

Another repressor in the ICM is Tcf3, a Wnt signaling pathway effector. TLE2 (a Groucho family protein) and CtBP (C-terminal binding protein) are key partners of Tcf3 in mediating this repressive effect. Tcf3 binds to and represses the Oct4 promoter, and this repressive effect requires both the Groucho and CtBP interacting domains of Tcf3 [52]. Tcf3 also limits the steady-state levels of Nanog mRNA, protein, and promoter activity in self-renewing embryonic stem cells (ESCs); the Tcf3 Groucho domain is involved in this repression [53]. Thus, Tcf3 is critical for maintaining the appropriate levels of both Oct4 and Nanog in ESCs. Experiments show that loss of Tcf3 by RNA interference (RNAi) knockdown blocks the ability of ESCs to differentiate [52], emphasizing the importance of this interaction.

Downstream of Oct4 and Sox2

Oct4 activates embryonic stem cell-specific gene 1 (Esg1), which encodes an RNA binding protein present in the ICM that is responsible for regulating several specific target transcripts [54]. Oct4 and Sox2 are also responsible for the regulation of the fibroblast growth factor 4 (FGF4) [44]. Expression of FGF4, therefore, requires the combined activity of these two transcription factors that bind to adjacent sites on the FGF4 enhancer DNA region [55]. Once expressed, the FGF4 protein can interact with its receptor FGFR2 and activate ICM and adjacent TE cell proliferation, activating extraembryonic endoderm cells as well in later stages.

Several other genes with important functions in embryonic development are also targets of Oct4-dependent activation. These include growth factor TDGF1, growth inhibitor SAP18, regulator of nonsense transcripts RENT1, two proteins involved in stem cell self-renewal DPPA4 and DPPA1 (developmental pluripotency associated), anterior visceral endoderm (AVE) markers LEFTY1 and LEFTY2, surface antigen THY1, and other genes encoding proteins involved in specialized cellular processes (DPP3, ATP6AP2, DDB1) and hypothetical proteins (GK003, hRscp) [32, 35].

The master regulation exerted by Sox2 and Oct4 during mammalian embryogenesis is believed to operate through their cooperative binding to DNA regulatory regions composed of adjacent HMG and POU motifs (HMG/POU cassettes) [56]. Exemplifying this arrangement, DPPA4 is one such gene with the presence of an HMG/POU cassette in its promoter region [57].

Downstream of Nanog and STAT3

Activation of JAK/STAT pathway also has an important contribution to pluripotency. In mice, the LIF/STAT3 pathway [39, 58, 59] for maintenance of cell pluripotency comprises LIF and LIF receptor, which deliver intracellular signaling

through STAT3. STAT3, a signal transducer and activator of transcription is activated by the JAK1 kinase and binds to several promoters inducing transcription of pluripotency related genes [60]. Nanog and Stat3 were found to bind to and synergistically activate Stat3-dependent promoters [60]. Nanog also functions as a transcriptional inhibitor to NFκB, a factor known to have pro-differentiation activity [60]. Nanog is also responsible for SMAD1 repression, thereby preventing BMP4-induced differentiation through the TGFbeta signaling pathway, for which SMAD1 is a key signal transducer [61].

Extraembryonic Endoderm Differentiation from ICM cells

Prior to embryo implantation one more differentiation takes place. Certain cells from the ICM give rise to the primitive endoderm, the first morphologically distinct cell type of the extraembryonic endoderm. The extraembryonic endoderm comprises the primitive, parietal and visceral endoderm components and will become the yolk sac during posterior development stages.

Wnt6 was recently identified as an inducer of primitive endoderm and this induction is accompanied by translocation of beta-catenin (CTNNB1) and Snail1 to the nucleus [62]. This study also showed that up-regulation of protein kinase A (PKA) induces markers of parietal endoderm. Another Wnt family member, Wnt9a, is expressed only in ICM cells that surround the bastocoel [63] and induces repositioning of the cells expressing GATA6, which is necessary for formation of primitive endoderm [64].

Sox7 plays a major role in parietal endoderm differentiation. Through studies with short interfering RNA molecules, it was established that Sox7 is responsible for transcription induction of GATA4 and GATA6 [65]. Individual or combined silencing

of Sox7, GATA4 and GATA6 result in suppression of cell shape changes and production of laminin-1 (LAMB1), characteristic changes present in parietal endoderm differentiation [65]. Gata4 was previously identified as a transcription factor responsible for the activation of FGF3 [66]. Sox7 also activates the FGF3 promoter. Conversely, Sox2 can negatively modulate the GATA4-dependent activation of FGF3, which is supported by the role of this factor in ICM pluripotency [67]. Another Sox family member, Sox17, is responsible for the differentiation of the extraembryonic endoderm in the final steps of preimplantation development [68]. The Runx1 factor is associated with the expression of Sox17 and is also specific for the extraembryonic endoderm [69]. HNF4 is a transcription factor specific of the extraembryonic endoderm with subsequent roles in post-implantation development and organogenesis [70]. Its expression may result from BMP4-induced differentiation [71]. Finally, the Dab2 protein is indispensable for the development of visceral endoderm; though its exact role is still not established, it is perhaps related to correct cell positioning [72, 73]. The expression of Cer1, a marker of the anterior visceral endoderm (AVE), commences before embryo implantation in the subset of cells that comprise the primitive endoderm. This ancestral population includes both cells expressing Cer1 together with cells in which Cer1 expression begins after implantation and formation of the AVE [56].

Search for Homologues

To establish an ortholog database and provide sequence information to the genes contained in the preimplantation pathway, aminoacid sequences corresponding to the human and mouse gene products were used as seed for the software SeedServer (Guedes *et al*., unpublished, see Methods for details). In fact, only the UniProt identifier for these proteins is necessary to execute SeedServer - gene symbols were

verified in the NCBI Gene database and converted to the corresponding geneID, and the desired identifiers were obtained afterwards from the UniProt database. For each gene a cluster of homologues was generated comprising from 2 to 260 sequences.

The recruited sequences contained in each cluster can be Swiss-Prot annotated or unrevised TrEMBL sequences. In total, 25% of the cluster sequences are Swiss-Prot, the great majority of clusters being comprised of TrEMBL sequences (75%). The search for homologues through SeedServer provides therefore a large amount of candidates for manual curation in Swiss-Prot. Furthermore, SeedServer can recruit sequences from organisms without a complete genome due to its use of UEKO and bidirectional best hit (BBH) searches conducted by SeedLinkage [74], and in fact only 27% of the sequences present in all clusters are from organisms with a complete genome. The ortholog clustering by SeedServer was only performed for genes that had a corresponding SwissProt annotated gene product to be used as seed, therefore hRSCP and DPPA1, which are described in the pathway, did not go through this analysis.

Pathway Ancestry

We then focused on the putative origin of these genes, determining which clade in the human lineage (e.g. class, order, family) shares each gene. The generation of ortholog clusters allowed for the determination of the last common ancestor (LCA) for each of the genes in the pathway. Figure 3 shows the genes according to their origin. Genes were arbitrarily considered ancient for this analysis if their last common ancestor originated before the divergence of the clade Euteleostomi and are coloured grey. Genes with a LCA belonging to the clade Euteleostomi or originated after divergence of Euteleostomi are considered recent genes and are coloured blue. Ancient origin genes with an ortholog in *Drosophila melanogaster* are marked with a

red asterisk. This arbitrary classification was meant to attract attention to the two key pluripotency controlling genes, Nanog (ancient) and Oct4 (modern).

The graph shown in Figure 4 represents the distribution of all the genes in the pathway according to their origin respect to clades of the human lineage. It may be observed that a large quantity of genes originates in certain periods as seen in Eumetazoa, Coelomata, Euteleostomi and Eutheria. The reasons for this wavelike origin need to be further analysed. On the other hand, the apparent origin of complex structures, that characterize all descendents from a certain moment of evolution, might have occurred simultaneously to the specialization of gene groups. . On the other hand, the coverage of genomic sequences in the database is far from homogeneous and can influence the shape of this graph [75]. In any case, the pattern observed agrees with the expansion of protein families related to stem cell markers observed in the ray-finned fish, that is, after divergence of the Euteleostomi [76].

Furthermore, we searched for functional information related to the *D. melanogaster* orthologues in order to determine if these functions are somehow similar or related to the functions of the corresponding pathway genes. This was done through a second text mining approach similar to the first and from the information recovered a secondary pathway was generated simply to illustrate the ortholog genes and their relative functional roles (Additional file 1). The regulatory pathways in which these genes are involved show us that these genes are all related to some part of *Drosophila* embryo development, some of them with highly conserved functions still observed in the preimplantation pathway described. An example is the Hippo signalling pathway, which is extremely conserved, showing Wts (Lats ortholog) phosphorylating Yki (Yap ortholog); this modification prevents Yki interaction with Sd (Tead4 ortholog). The correlation between the human gene names and

corresponding *D. melanogaster* ortholog names can be found in Additional file 2 and also the PMID reference for the gene function in Drosophila development.

Discussion

The use of text-mining tools for the generation of regulatory pathways is an effective approach and it is important for the current interest of gathering data related to an organism or biological process. The search for information related to a specific concept such as “preimplantation development” resulted in the selection of data related to this process only. When other tools such as iHOP [1] and STRING [2] are used for the search of biointeractions, it is necessary to know the names for the genes you are interested on and the information is then retrieved. Moreover in the case of iHOP, the information retrieved consists of a large list of papers related to the gene of interest, which need to be manually analysed to extract the information related to the specific process. In the case of STRING, the result of a query is a network of direct associations to other genes, which can be activations, repressions, or or unknown, but for which it is not possible to perform a search restricting the query to a specific process for which you seek to determine the involvement of a given gene.

The approach described in this work (using PubMed, MedlineRanker, PESCADOR) summarized in Figure 5, allows the researcher to initiate the study of a pathway without knowing exactly the genes involved, simply by selecting the published information related to the process of interest. The manual curation required to create a pathway through this approach is significantly smaller. However, the verification of all the interactions highlighted by the tool is essential. Text-mining is not able to eliminate the selection of false interaction pairs; in the case of LAITOR

(contained in the PESCADOR platform), the type 3 and 4 interactions can present genes with no association specified in the text [14].

The text-mining data contribute the complete description of the pathway in the form of a literature review, a necessary step for the validation of the regulations represented, and for the inclusion of the pathway in a specific database, such as KEGG Pathway. The establishment of this procedure for pathway generation allows future work to enlarge the knowledge on subjects still not approached, such as regulatory pathways for several types of cancer, mechanisms of pathogen resistance in plants and response to abiotic stresses in plants, among other themes of interest.

The inclusion of the preimplantation pathway in databases such as the KEGG database will allow automatic annotation for several other organisms, as it is usually done in this database. Concurrently, a laboratory with a specific interest can promptly build a similar Pathway for its local use. From the 86 genes present in the pathway, 20 do not possess entries in KEGG Orthology and would constitute important additions. Considering that the contribution of KEGG for the sequence recruitment in the SeedServer clusters is only 25% of the total number of sequences, some organisms evolutionarily divergent from the ones represented in KEGG begin to play a more relevant role for a more efficient annotation of new sequences. It is relevant to stress that only the SeedLinkage and UEKO components of SeedServer are capable of clustering sequences proceeding from organisms without a complete genome project.

Another important contribution from the ortholog clustering by SeedServer is the identification of candidates for Swiss-Prot Annotation. Swiss-Prot annotation depends on the correct association of sequences to gene families and proteins with known function, using the available literature as a reference. The annotation is facilitated since each of the genes is associated with PubMed Identifiers (PMIDs)

stored in the PESCADOR tool, which are important references for the related orthologs.

The search for functional information for the *D. melanogaster* orthologues revealed the involvement of the genes in processes related to the embryonic development and was also a good validation for the clustering by SeedServer, since none of the sequences from *D. melanogaster* clustered to the initial human and mouse genes presents an unexpected function.

Generation of correct clusters is essential for the correct determination of gene ancestry, but it is not the sole limiting factor. Sequencing of key organisms from taxonomic outgroups relative to the ones with complete genome sequences available will be a crucial source of sequences that will allow a revaluation of gene ancestry. Meanwhile, additional sequences clustered by software (SeedLinkage) and database enrichment (UEKO) improve the inspection of ancestry.

Determination of the ancestry for the genes in the preimplantation pathway was nonetheless a central analysis, given the expectancy that this pathway would be mainly formed by more contemporary components. Our data suggest that an ancient fraction of the pathway including Nanog and Sox2 originated before Chordata, whereas a modern fraction including Oct4 and LIF has appeared near the origin of Eutheria, the placentary organisms. Thus, an important transcriptional pathway comprising ancient and modern members has been characterized with text mining, and homologues search with SeedServer promptly allowed LCA determination.

Conclusions

Generation of regulatory pathways through text-mining tools allows integration of data generated by previous studies for a more complete view of a biological process. If the genes present in this pathway are associated with clusters of

orthologues this information is added to the pathway making the visualization of the same process available for different organisms. The analysis of orthology also permits determination of the ancestry of the genes involved in the process leading to a better understanding of the evolution of such process.

Methods

Text-mining and Pathway construction

NCBI's PubMed database was used as a source of available literature (http://www.ncbi.nlm.nih.gov/pubmed) for the text-mining approach. The search query used was "preimplantation development" and the PubMed identification numbers of the selected papers (PMIDs) were saved as a text file. Ten papers were selected manually by us to be used in the Medline Ranker software ([13]; http://cbdm.mdc-berlin.de/tools/medlineranker/). These papers, (references [17, 20, 23, 24, 26, 45, 55, 77, 78]), were considered by us as highly informative because they described numerous gene regulations concerning preimplantation development. We used the PMIDs retrieved by the PubMed search as the background set and the 10 manually selected PMIDs as the training set. After classification by order of relevance we selected the 1000 best classified abstracts for further analysis. These abstracts were then submitted through PESCADOR (manuscript under preparation, Barbosa-Silva *et al*.), an online platform for the software LAITOR [14]. Afterwards PESCADOR results were manually curated and the gene biointeractions recovered were used to build a regulatory pathway in Keynote MacOS according to the markup language used by KEGG for pathway construction (KGML can be found at http://www.genome.jp/kegg/xml/docs/).

SeedServer search for homologues

UniProt IDs for human and mouse gene products corresponding to each of the genes represented in the preimplantation pathway were used as seed in the SeedServer software (not published, Guedes et al.). SeedServer is a web application (seedserver.cenabid.org) which searches for homologous sequences through two components: the program SeedLinkage [74] and the databases KEGG Orthology (KO) [6] and its enriched version UEKO (unpublished, developed by Fernandes *et al.* by application of the procedure described to enrich COG [8] to the KEGG Orthology database). Clustering was verified by PSI-BLAST searches using seed sequences as query and the recruited proteins as database, and eventual false positives were discarded (1.5% of the recruited sequences).

LCA determination

Clusters generated for each of the pathway genes were used to determine the Last Common Ancestor (LCA) of each gene. Each cluster provided a list of Taxonomy IDs corresponding to the organisms in which orthologs of the pathway genes were found. The clade in the human lineage that comprised these Taxonomy IDs as leaves in the Taxonomy Tree was considered to bear the LCA.

Authors' contributions

ERD and JMO conceived the project and wrote the paper. ERD performed the research and pathway construction. MJK curated the pathway biointeractions. ABS (author of SeedLinkage and LAITOR) and MAAN designed the PESCADOR platform. RLMG designed the SeedServer software and conducted the ortholog search. HV was responsible for the LCA determination. GRF constructed the UEKO database. All authors read and approved the final manuscript.

References

1. Hoffmann R, Valencia A: **A gene network for navigating the literature.** *Nat Genet* 2004, **36:**664.
2. Jensen LJ, Kuhn M, Stark M, Chaffron S, Creevey C, Muller J, Doerks T, Julien P, Roth A, Simonovic M, et al: **STRING 8--a global view on proteins and their functional interactions in 630 organisms.** *Nucleic Acids Res* 2009, **37:**D412-416.
3. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, et al: **The COG database: an updated version includes eukaryotes.** *BMC Bioinformatics* 2003, **4:**41.
4. Muller J, Szklarczyk D, Julien P, Letunic I, Roth A, Kuhn M, Powell S, von Mering C, Doerks T, Jensen LJ, Bork P: **eggNOG v2.0: extending the evolutionary genealogy of genes with enhanced non-supervised orthologous groups, species and functional annotations.** *Nucleic Acids Res* 2010, **38:**D190-195.
5. Li L, Stoeckert CJ, Jr., Roos DS: **OrthoMCL: identification of ortholog groups for eukaryotic genomes.** *Genome Res* 2003, **13:**2178-2189.
6. Kanehisa M, Goto S: **KEGG: kyoto encyclopedia of genes and genomes.** *Nucleic Acids Res* 2000, **28:**27-30.
7. Suzek BE, Huang H, McGarvey P, Mazumder R, Wu CH: **UniRef: comprehensive and non-redundant UniProt reference clusters.** *Bioinformatics* 2007, **23:**1282-1288.
8. Fernandes GR, Barbosa DV, Prosdocimi F, Pena IA, Santana-Santos L, Coelho Junior O, Barbosa-Silva A, Velloso HM, Mudado MA, Natale DA, et

al: **A procedure to recruit members to enlarge protein family databases--the building of UECOG (UniRef-Enriched COG Database) as a model.** *Genet Mol Res* 2008, **7:**910-924.

9. Wang H, Dey SK: **Roadmap to embryo implantation: clues from mouse models.** *Nat Rev Genet* 2006, **7:**185-199.

10. Johnson MH, McConnell JM: **Lineage allocation and cell polarity during mouse embryogenesis.** *Semin Cell Dev Biol* 2004, **15:**583-597.

11. Yamanaka Y, Ralston A, Stephenson RO, Rossant J: **Cell and molecular regulation of the mouse blastocyst.** *Dev Dyn* 2006, **235:**2301-2314.

12. Davidson EH, Rast JP, Oliveri P, Ransick A, Calestani C, Yuh CH, Minokawa T, Amore G, Hinman V, Arenas-Mena C, et al: **A genomic regulatory network for development.** *Science* 2002, **295:**1669-1678.

13. Fontaine JF, Barbosa-Silva A, Schaefer M, Huska MR, Muro EM, Andrade-Navarro MA: **MedlineRanker: flexible ranking of biomedical literature.** *Nucleic Acids Res* 2009, **37:**W141-146.

14. Barbosa-Silva A, Soldatos TG, Magalhaes IL, Pavlopoulos GA, Fontaine JF, Andrade-Navarro MA, Schneider R, Ortega JM: **LAITOR--Literature Assistant for Identification of Terms co-Occurrences and Relationships.** *BMC Bioinformatics* 2010, **11:**70.

15. Suzuki T, Abe K, Inoue A, Aoki F: **Expression of c-MYC in nuclear speckles during mouse oocyte growth and preimplantation development.** *J Reprod Dev* 2009, **55:**491-495.

16. Monk M, Hitchins M, Hawes S: **Differential expression of the embryo/cancer gene ECSA(DPPA2), the cancer/testis gene BORIS and**

the pluripotency structural gene OCT4, in human preimplantation development. *Molecular Human Reproduction* 2008, **14:**347-355.

17. Mizuno S, Sono Y, Matsuoka T, Matsumoto K, Saeki K, Hosoi Y, Fukuda A, Morimoto Y, Iritani A: **Expression and subcellular localization of GSE protein in germ cells and preimplantation embryos.** *J Reprod Dev* 2006, **52:**429-438.

18. Chung YG, Ratnam S, Chaillet JR, Latham KE: **Abnormal regulation of DNA methyltransferase expression in cloned mouse embryos.** *Biol Reprod* 2003, **69:**146-153.

19. Yu JN, Xue CY, Wang XG, Lin F, Liu CY, Lu FZ, Liu HL: **5-AZA-2'-deoxycytidine (5-AZA-CdR) leads to down-regulation of Dnmt1o and gene expression in preimplantation mouse embryos.** *Zygote* 2009, **17:**137-145.

20. Plusa B, Frankenberg S, Chalmers A, Hadjantonakis AK, Moore CA, Papalopulu N, Papaioannou VE, Glover DM, Zernicka-Goetz M: **Downregulation of Par3 and aPKC function directs cells towards the ICM in the preimplantation mouse embryo.** *J Cell Sci* 2005, **118:**505-515.

21. Guo G, Huss M, Tong GQ, Wang C, Li Sun L, Clarke ND, Robson P: **Resolution of cell fate decisions revealed by single-cell gene expression analysis from zygote to blastocyst.** *Dev Cell* 2010, **18:**675-685.

22. Ralston A, Rossant J: **Cdx2 acts downstream of cell polarization to cell-autonomously promote trophectoderm fate in the early mouse embryo.** *Dev Biol* 2008, **313:**614-629.

23. Nishioka N, Inoue K, Adachi K, Kiyonari H, Ota M, Ralston A, Yabuta N, Hirahara S, Stephenson RO, Ogonuki N, et al: **The Hippo signaling pathway**

components Lats and Yap pattern Tead4 activity to distinguish mouse trophectoderm from inner cell mass. *Dev Cell* 2009, **16:**398-410.

24. Strumpf D, Mao CA, Yamanaka Y, Ralston A, Chawengsaksophak K, Beck F, Rossant J: **Cdx2 is required for correct cell fate specification and differentiation of trophectoderm in the mouse blastocyst.** *Development* 2005, **132:**2093-2102.

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

26. Zhang J, Tam WL, Tong GQ, Wu Q, Chan HY, Soh BS, Lou Y, Yang J, Ma Y, Chai L, et al: **Sall4 modulates embryonic stem cell pluripotency and early embryonic development by the transcriptional regulation of Pou5f1.** *Nat Cell Biol* 2006, **8:**1114-1123.

27. Ralston A, Cox BJ, Nishioka N, Sasaki H, Chea E, Rugg-Gunn P, Guo G, Robson P, Draper JS, Rossant J: **Gata3 regulates trophoblast development downstream of Tead4 and in parallel to Cdx2.** *Development* 2010, **137:**395-403.

28. Vauti F, Prochnow BR, Freese E, Ramasamy SK, Ruiz P, Arnold HH: **Arp3 is required during preimplantation development of the mouse embryo.** *FEBS Lett* 2007, **581:**5691-5697.

29. Hayashi Y, Furue MK, Tanaka S, Hirose M, Wakisaka N, Danno H, Ohnuma K, Oeda S, Aihara Y, Shiota K, et al: **BMP4 induction of trophoblast from mouse embryonic stem cells in defined culture conditions on laminin.** *In Vitro Cell Dev Biol Anim* 2010, **46:**416-430.

30. Kondo M, Cubillo E, Tobiume K, Shirakihara T, Fukuda N, Suzuki H, Shimizu K, Takehara K, Cano A, Saitoh M, Miyazono K: **A role for Id in the regulation of TGF-beta-induced epithelial-mesenchymal transdifferentiation.** *Cell Death Differ* 2004, **11:**1092-1101.

31. Riley P, Anson-Cartwright L, Cross JC: **The Hand1 bHLH transcription factor is essential for placentation and cardiac morphogenesis.** *Nat Genet* 1998, **18:**271-275.

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

33. Marikawa Y, Alarcón VB: **Establishment of trophectoderm and inner cell mass lineages in the mouse embryo.** *Mol Reprod Dev* 2009, **76:**1019-1032.

34. Adjaye J, Herwig R, Brink TC, Herrmann D, Greber B, Sudheer S, Groth D, Carnwath JW, Lehrach H, Niemann H: **Conserved molecular portraits of bovine and human blastocysts as a consequence of the transition from maternal to embryonic control of gene expression.** *Physiol Genomics* 2007, **31:**315-327.

35. Shin MR, Cui XS, Jun JH, Jeong YJ, Kim NH: **Identification of mouse blastocyst genes that are downregulated by double-stranded RNA-mediated knockdown of Oct-4 expression.** *Mol Reprod Dev* 2005, **70:**390-396.

36. Yamada K, Ogawa H, Tamiya G, Ikeno M, Morita M, Asakawa S, Shimizu N, Okazaki T: **Genomic organization, chromosomal localization, and the complete 22 kb DNA sequence of the human GCMa/GCM1, a placenta-**

specific transcription factor gene. *Biochem Biophys Res Commun* 2000, **278:**134-139.

37. Matin MM, Walsh JR, Gokhale PJ, Draper JS, Bahrami AR, Morton I, Moore HD, Andrews PW: **Specific knockdown of Oct4 and beta2-microglobulin expression by RNA interference in human embryonic stem cells and embryonic carcinoma cells.** *Stem Cells* 2004, **22:**659-668.

38. Collins JE, Lorimer JE, Garrod DR, Pidsley SC, Buxton RS, Fleming TP: **Regulation of desmocollin transcription in mouse preimplantation embryos.** *Development* 1995, **121:**743-753.

39. Chen L, Yang M, Dawes J, Khillan JS: **Suppression of ES cell differentiation by retinol (vitamin A) via the overexpression of Nanog.** *Differentiation* 2007, **75:**682-693.

40. Jiang J, Chan YS, Loh YH, Cai J, Tong GQ, Lim CA, Robson P, Zhong S, Ng HH: **A core Klf circuitry regulates self-renewal of embryonic stem cells.** *Nat Cell Biol* 2008, **10:**353-360.

41. Pan G, Li J, Zhou Y, Zheng H, Pei D: **A negative feedback loop of transcription factors that controls stem cell pluripotency and self-renewal.** *FASEB J* 2006, **20:**1730-1732.

42. van den Berg DL, Zhang W, Yates A, Engelen E, Takacs K, Bezstarosti K, Demmers J, Chambers I, Poot RA: **Estrogen-related receptor beta interacts with Oct4 to positively regulate Nanog gene expression.** *Mol Cell Biol* 2008, **28:**5986-5995.

43. Zhang X, Zhang J, Wang T, Esteban MA, Pei D: **Esrrb activates Oct4 transcription and sustains self-renewal and pluripotency in embryonic stem cells.** *J Biol Chem* 2008, **283:**35825-35833.

44. Okumura-Nakanishi S, Saito M, Niwa H, Ishikawa F: **Oct-3/4 and Sox2 regulate Oct-3/4 gene in embryonic stem cells.** *J Biol Chem* 2005, **280:**5307-5317.

45. Cauffman G, De Rycke M, Sermon K, Liebaers I, Van de Velde H: **Markers that define stemness in ESC are unable to identify the totipotent cells in human preimplantation embryos.** *Hum Reprod* 2009, **24:**63-70.

46. Holzinger M, Bouffier L, Villalonga R, Cosnier S: **Adamantane/beta-cyclodextrin affinity biosensors based on single-walled carbon nanotubes.** *Biosens Bioelectron* 2009, **24:**1128-1134.

47. Adjaye J, Huntriss J, Herwig R, BenKahla A, Brink TC, Wierling C, Hultschig C, Groth D, Yaspo ML, Picton HM, et al: **Primary differentiation in the human blastocyst: comparative molecular portraits of inner cell mass and trophectoderm cells.** *Stem Cells* 2005, **23:**1514-1525.

48. Cormier S, Le Bras S, Souilhol C, Vandormael-Pournin S, Durand B, Babinet C, Baldacci P, Cohen-Tannoudji M: **The murine ortholog of notchless, a direct regulator of the notch pathway in Drosophila melanogaster, is essential for survival of inner cell mass cells.** *Mol Cell Biol* 2006, **26:**3541-3549.

49. Voss AK, Thomas T, Petrou P, Anastassiadis K, Schöler H, Gruss P: **Taube nuss is a novel gene essential for the survival of pluripotent cells of early mouse embryos.** *Development* 2000, **127:**5449-5461.

50. Robson P, Stein P, Zhou B, Schultz RM, Baldwin HS: **Inner cell mass-specific expression of a cell adhesion molecule (PECAM-1/CD31) in the mouse blastocyst.** *Dev Biol* 2001, **234:**317-329.

51. Sun C, Nakatake Y, Akagi T, Ura H, Matsuda T, Nishiyama A, Koide H, Ko MS, Niwa H, Yokota T: **Dax1 binds to Oct3/4 and inhibits its transcriptional activity in embryonic stem cells.** *Mol Cell Biol* 2009, **29:**4574-4583.

52. Tam WL, Lim CY, Han J, Zhang J, Ang YS, Ng HH, Yang H, Lim B: **T-cell factor 3 regulates embryonic stem cell pluripotency and self-renewal by the transcriptional control of multiple lineage pathways.** *Stem Cells* 2008, **26:**2019-2031.

53. Pereira L, Yi F, Merrill BJ: **Repression of Nanog gene transcription by Tcf3 limits embryonic stem cell self-renewal.** *Mol Cell Biol* 2006, **26:**7479-7491.

54. Tanaka TS, Lopez de Silanes I, Sharova LV, Akutsu H, Yoshikawa T, Amano H, Yamanaka S, Gorospe M, Ko MS: **Esg1, expressed exclusively in preimplantation embryos, germline, and embryonic stem cells, is a putative RNA-binding protein with broad RNA targets.** *Dev Growth Differ* 2006, **48:**381-390.

55. Ambrosetti DC, Schöler HR, Dailey L, Basilico C: **Modulation of the activity of multiple transcriptional activation domains by the DNA binding domains mediates the synergistic action of Sox2 and Oct-3 on the fibroblast growth factor-4 enhancer.** *J Biol Chem* 2000, **275:**23387-23397.

56. Torres-Padilla ME, Richardson L, Kolasinska P, Meilhac SM, Luetke-Eversloh MV, Zernicka-Goetz M: **The anterior visceral endoderm of the mouse embryo is established from both preimplantation precursor cells and by de novo gene expression after implantation.** *Dev Biol* 2007, **309:**97-112.

57. Chakravarthy H, Boer B, Desler M, Mallanna SK, McKeithan TW, Rizzino A: **Identification of DPPA4 and other genes as putative Sox2:Oct-3/4 target genes using a combination of in silico analysis and transcription-based assays.** *J Cell Physiol* 2008, **216:**651-662.

58. Saito S, Liu B, Yokoyama K: **Animal embryonic stem (ES) cells: self-renewal, pluripotency, transgenesis and nuclear transfer.** *Hum Cell* 2004, **17:**107-115.

59. De Felici M, Farini D, Dolci S: **In or out stemness: comparing growth factor signalling in mouse embryonic stem cells and primordial germ cells.** *Curr Stem Cell Res Ther* 2009, **4:**87-97.

60. Torres J, Watt FM: **Nanog maintains pluripotency of mouse embryonic stem cells by inhibiting NFkappaB and cooperating with Stat3.** *Nat Cell Biol* 2008, **10:**194-201.

61. Suzuki A, Raya A, Kawakami Y, Morita M, Matsui T, Nakashima K, Gage FH, Rodríguez-Esteban C, Izpisúa Belmonte JC: **Nanog binds to Smad1 and blocks bone morphogenetic protein-induced differentiation of embryonic stem cells.** *Proc Natl Acad Sci USA* 2006, **103:**10294-10299.

62. Krawetz R, Kelly GM: **Wnt6 induces the specification and epithelialization of F9 embryonal carcinoma cells to primitive endoderm.** *Cell Signal* 2008, **20:**506-517.

63. Kemp C, Willems E, Abdo S, Lambiv L, Leyns L: **Expression of all Wnt genes and their secreted antagonists during mouse blastocyst and postimplantation development.** *Dev Dyn* 2005, **233:**1064-1075.

64. Meilhac SM, Adams RJ, Morris SA, Danckaert A, Le Garrec JF, Zernicka-Goetz M: **Active cell movements coupled to positional induction are**

involved in lineage segregation in the mouse blastocyst. *Dev Biol* 2009, **331:**210-221.

65. Futaki S, Hayashi Y, Emoto T, Weber CN, Sekiguchi K: **Sox7 plays crucial roles in parietal endoderm differentiation in F9 embryonal carcinoma cells through regulating Gata-4 and Gata-6 expression.** *Mol Cell Biol* 2004, **24:**10492-10503.

66. Murakami A, Thurlow J, Dickson C: **Retinoic acid-regulated expression of fibroblast growth factor 3 requires the interaction between a novel transcription factor and GATA-4.** *J Biol Chem* 1999, **274:**17242-17248.

67. Murakami A, Shen H, Ishida S, Dickson C: **SOX7 and GATA-4 are competitive activators of Fgf-3 transcription.** *J Biol Chem* 2004, **279:**28564-28573.

68. Shimoda M, Kanai-Azuma M, Hara K, Miyazaki S, Kanai Y, Monden M, Miyazaki J: **Sox17 plays a substantial role in late-stage differentiation of the extraembryonic endoderm in vitro.** *J Cell Sci* 2007, **120:**3859-3869.

69. Kurimoto K, Yabuta Y, Ohinata Y, Ono Y, Uno KD, Yamada RG, Ueda HR, Saitou M: **An improved single-cell cDNA amplification method for efficient high-density oligonucleotide microarray analysis.** *Nucleic Acids Res* 2006, **34:**e42.

70. Duncan SA, Manova K, Chen WS, Hoodless P, Weinstein DC, Bachvarova RF, Darnell JE: **Expression of transcription factor HNF-4 in the extraembryonic endoderm, gut, and nephrogenic tissue of the developing mouse embryo: HNF-4 is a marker for primary endoderm in the implanting blastocyst.** *Proc Natl Acad Sci USA* 1994, **91:**7598-7602.

71. Coucouvanis E, Martin GR: **BMP signaling plays a role in visceral endoderm differentiation and cavitation in the early mouse embryo.** *Development* 1999, **126:**535-546.

72. Morris SM, Tallquist MD, Rock CO, Cooper JA: **Dual roles for the Dab2 adaptor protein in embryonic development and kidney transport.** *EMBO J* 2002, **21:**1555-1564.

73. Yang DH, Smith ER, Roland IH, Sheng Z, He J, Martin WD, Hamilton TC, Lambeth JD, Xu XX: **Disabled-2 is essential for endodermal cell positioning and structure formation during mouse embryogenesis.** *Dev Biol* 2002, **251:**27-44.

74. Barbosa-Silva A, Satagopam VP, Schneider R, Ortega JM: **Clustering of cognate proteins among distinct proteomes derived from multiple links to a single seed sequence.** *BMC Bioinformatics* 2008, **9:**141.

75. Perez-Iratxeta C, Palidwor G, Andrade-Navarro MA: **Towards completion of the Earth's proteome.** *EMBO Rep* 2007, **8:**1135-1141.

76. Krzyzanowski PM, Andrade-Navarro MA: **Identification of novel stem cell markers using gap analysis of gene expression data.** *Genome Biol* 2007, **8:**R193.

77. Scaffidi P, Bianchi ME: **Spatially precise DNA bending is an essential activity of the sox2 transcription factor.** *J Biol Chem* 2001, **276:**47296-47302.

78. Lim CY, Tam WL, Zhang J, Ang HS, Jia H, Lipovich L, Ng HH, Wei CL, Sung WK, Robson P, et al: **Sall4 regulates distinct transcription circuitries in different blastocyst-derived stem cell lineages.** *Cell Stem Cell* 2008, **3:**543-554.

KEGG Markup Language [http://www.genome.jp/kegg/xml/docs/]

MedlineRanker [http://cbdm.mdc-berlin.de/tools/medlineranker/]

NCBI Pubmed Database [http://www.ncbi.nlm.nih.gov/pubmed]

SeedServer [seedserver.cenabid.org]

Figures

Figure 1 - Biointeraction extraction from PESCADOR

Top: Sample abstract tagged by PESCADOR. Gene or protein names (terms) recognized are highlighted in violet and the biointeraction words in yellow. The platform allows users to search for their interactions of interest by terms, abstracts or concepts of interest added initially by the user. Bottom: Manual curation of the information presented in the abstract and its graphical representation in the form of a regulatory pathway.

Figure 2 - Preimplantation Development Pathway

The figure shows a pathway representation of the genes involved in the regulation of the preimplantation development and interactions between them. Some functions are also detailed in the grey rectangles. The interactions are described in the text. KEGG Markup Language was used for pathway representation. The developmental stages figures were adapted from Yamanaka *et al.* 2006 [11].

Figure 3 - Pathway Ancestry

The pathway genes are represented according to their ancestry based on the determination of their Last Common Ancestor. Genes considered recent are shown in blue while genes of more ancient origin are shown in grey. Genes that present an ortholog in *D. melanogaster* are marked (*).

Figure 4 - Gene Origin in Human Evolution
Distribution of the genes in the preimplantation pathway according to their origin in clades of the human lineage, based on the determination of the Last Common Ancestor for the ortholog clusters generated by SeedServer. The y-axis represents the number of genes and the x-axis represents the taxonomical groups in which the genes originated.

Figure 5 - Pathway Construction Flowchart
The initial step consists of a PubMed search with the subject of interest (e.g. preimplantation development). The list of PubMed identifiers (PMIDs) obtained in the search is then used in the web tool Medline Ranker as the background set along with a list of PMIDs of manually selected abstracts considered informative which form the test set. The tool generates a list of abstracts classified by order of relevance. Best 1000 abstracts are recovered and their corresponding PMID is then introduced in the PESCADOR platform. Abstracts are tagged by PESCADOR and provide a source of biointeractions for manual curation and pathway construction. UniProt IDs for products of the genes present in the final pathway are obtained and used as seed in SeedServer. The software recruits homologues for each gene and creates the final clusters. Taxonomy IDs from each cluster can be used for Last Common Ancestor (LCA) determination.

Additional files

Additional file 1 – Ortholog Functions in *Drosophila melanogaster*
This figure represents the corresponding *D. melanogaster* orthologs found by SeedServer and their respective interactions and functions in fruit fly development. Note that these orthologs are involved in processes related to *D. melanogaster* embryo development. See Additional file 2 for a table with gene name correspondence between the genes in this figure and the ones on Figure 3.

Additional file 2 – Gene Correspondence Table
Human and *Drosophila melanogaster* gene name correspondence for the orthologs grouped by SeedServer. Column 3 lists the PubMed identifiers (PMIDs) from the papers where functions described in Additional file 1 were found.