

# Towards an Evolutionary Model of Transcription Networks

Dan Xie<sup>1,9</sup>, Chieh-Chun Chen<sup>1,9</sup>, Xin He<sup>2</sup>, Xiaoyi Cao<sup>3</sup>, Sheng Zhong<sup>1,3\*</sup>

**1** Department of Bioengineering, University of Illinois at Urbana-Champaign, Urbana, Illinois, United States of America, **2** Department of Biochemistry and Biophysics, University of California, San Francisco, California, United States of America, **3** Center for Biophysics and Computational Biology, University of Illinois at Urbana-Champaign, Urbana, Illinois, United States of America

## Abstract

DNA evolution models made invaluable contributions to comparative genomics, although it seemed formidable to include non-genomic features into these models. In order to build an evolutionary model of transcription networks (TNs), we had to forfeit the substitution model used in DNA evolution and to start from modeling the evolution of the regulatory relationships. We present a quantitative evolutionary model of TNs, subjecting the phylogenetic distance and the evolutionary changes of *cis*-regulatory sequence, gene expression and network structure to one probabilistic framework. Using the genome sequences and gene expression data from multiple species, this model can predict regulatory relationships between a transcription factor (TF) and its target genes in all species, and thus identify TN re-wiring events. Applying this model to analyze the pre-implantation development of three mammalian species, we identified the conserved and re-wired components of the TNs downstream to a set of TFs including Oct4, Gata3/4/6, cMyc and nMyc. Evolutionary events on the DNA sequence that led to turnover of TF binding sites were identified, including a birth of an Oct4 binding site by a 2nt deletion. In contrast to recent reports of large interspecies differences of TF binding sites and gene expression patterns, the interspecies difference in TF-target relationship is much smaller. The data showed increasing conservation levels from genomic sequences to TF-DNA interaction, gene expression, TN, and finally to morphology, suggesting that evolutionary changes are larger at molecular levels and smaller at functional levels. The data also showed that evolutionarily older TFs are more likely to have conserved target genes, whereas younger TFs tend to have larger re-wiring rates.

**Citation:** Xie D, Chen C-C, He X, Cao X, Zhong S (2011) Towards an Evolutionary Model of Transcription Networks. PLoS Comput Biol 7(6): e1002064. doi:10.1371/journal.pcbi.1002064

**Editor:** Amos Tanay, Weizmann Institute of Science, Israel

**Received:** December 10, 2010; **Accepted:** April 8, 2011; **Published:** June 9, 2011

**Copyright:** © 2011 Xie et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Funding:** This work is supported by NIH DP2-OD007417, NSF DBI 08-45823, and NSF DBI 09-60583. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing Interests:** The authors have declared that no competing interests exist.

\* E-mail: szhong@illinois.edu

**9** These authors contributed equally to this work.

## Introduction

Biologists have long sought to dissect what genes and what changes to their coding and regulatory sequences are responsible for the diversity of life. It has been argued that morphological traits evolve to a large extent through changes in transcription networks (TNs) that regulate gene expression patterns [1,2]. However, except on a relatively small set of well characterized enhancers, it appears difficult to quantitatively analyze *cis* contribution to TN evolution (see reviews [2,3,4]). This is in part due to the computational difficulties in finding *cis*-regulatory sequences or enhancers [5], assessing the binding affinity of an enhancer to a set of transcription factors (TFs) [6], and associating enhancer binding affinities with gene expression levels [7]. Still lacking are principled approaches and evolutionary models to quantitatively analyze the effects of changes in *cis*-regulatory sequences, gene expression, and TNs (Table S1).

The question that inspired us to model TN evolution is the conservation of early embryonic development in mammals. The earliest stages of embryonic development are thought to be highly conserved among placental mammals, because these species all progress through the same morphologic stages before implantation. This traditional view is challenged by recent reports on large

inter-species differences in gene expression [8] and TF binding patterns [9]. During pre-implantation development (PED), an unexpected fraction of 40% of orthologous gene triplets exhibited different expression patterns among humans, mice and cattle [8], accompanied by an even more unexpected fraction of 95% of the binding sites of the core TFs, Oct4 and Nanog, in human and mouse embryonic stem (ES) cells not being located in orthologous genomic regions [9]. Large interspecies differences of gene expression [10] (Figure S1) and transcription factor binding sites [11] were also observed in matched tissues in vertebrates. How can we understand the discrepancy between morphological conservation and molecular differences across species? We hypothesized that the structure of TN, *i.e.* the regulatory relationship between transcription factors (TFs) and target genes, may be more conserved than suggested by TF binding site (TFBS) or gene expression data. For instance, some TFBSs turn over quickly during vertebrate evolution, without necessarily changing TF-target relationships. To test this hypothesis and to provide a general tool for studying TN evolution, we set off to develop an evolutionary model for TN structure based on multi-species genome sequence and gene expression data.

Previous work has made excellent progress in modeling the evolution of regulatory genomic sequences. Earlier attempts were

## Author Summary

DNA evolution models made invaluable contributions to comparative genomic studies. Still lacking is an evolutionary model of transcription networks (TNs). To develop such a model, we had to forfeit the substitution model used in DNA evolution and to start from modeling the evolution of the regulatory relationships, and then subject the phylogenetic distance and the multi-species DNA sequence and gene expression data to one probabilistic framework. This model enabled us to infer the evolutionary changes of transcriptional regulatory relationships. Applying this model to analyze three yeast species, we found the anaerobic phenotype in two species was associated with the evolutionary loss of a larger *cis*-regulatory motif than previously thought. Analyzing three mammalian species, we found increasing conservation levels from genomic sequences to transcription factor-DNA interaction, gene expression, TN, and finally to morphology, suggesting that evolutionary changes are larger at molecular levels and smaller at functional levels. We also found that evolutionarily younger TFs are more likely to regulate different target genes in different species.

focused on identifying putative regulatory sequences highly conserved across multiple species [12,13] or containing conserved TF binding motifs [14,15]. Recent efforts extended the earlier work by accommodating lineage-specific changes and alignment errors [16,17], incorporating the modular structure of regulatory sequences [5,18], and direct modeling of TF occupancy [6,19]. Although these models have been evolved to accommodate many evolutionary events on DNA, it is difficult to extend these models to incorporate the evolutionary changes of non-genomic features. This difficulty can be partially appreciated by noticing that even in a single species, the state-of-the-art models resort to a regression strategy to incorporate DNA sequence and gene expression into one model (see [20] and references within).

In parallel to sequence evolution models, evolutionary models for gene expression are being developed and tested. A neutral evolutionary model for gene expression was proposed [21], enabling statistical tests for evolutionarily selected genes [22]. A major challenge in comparing expression data between organisms is that gene expression is not static and the level of expression is influenced by external conditions. A prominent approach to circumvent this difficulty is to compare co-expression patterns rather than the expression of individual genes [23,24]. This approach has recently been formalized into evolutionary models of co-expression networks, which are based on explicitly stated probabilistic rules [25,26]. These evolutionary models did not explicitly study the evolution of genomic regulatory relations. In fact, while expression data can be useful in predicting co-regulation (in particular, among target genes of the same TF), such data alone can hardly predict which gene is regulated by which TF as the correlation between expression patterns of a TF and that of its targets may not be pronounced [27,28].

In order to build an evolutionary model of TNs, we had to forfeit the substitution model used in DNA evolution and the co-expression model used in expression evolution, and to start from modeling the evolution of the regulatory relationships. In this paper we present a quantitative evolutionary model of TNs, subjecting the changes of *cis*-regulatory sequences, gene expression and network structure to one probabilistic framework. This model aims to address the following question: with the genome sequences and gene expression data from multiple species, can the TF-target

gene relationships be derived in all these species? Taking advantage of the multi-species data and based on the maximum-likelihood principle, this model predicts the evolutionary changes of TF-target relationships, *i.e.* re-wiring of TNs (Figure 1). The major benefits of this model include: 1) it takes advantage of the possible synergism between genome sequence and gene expression data. For instance, if a gene is predicted to be a TF target from sequence data, other genes co-regulated with this gene, according to the expression data, may also be a target. 2) By analyzing data in the evolutionary context, the model is still able to utilize pattern of conservation to predict regulatory relationship (similar to the evolutionary models of regulatory sequences we discussed earlier), while allowing for lineage-specific changes. 3) The unified probabilistic framework allows us to quantify the extent of changes and the uncertainty of the inference results.

## Results

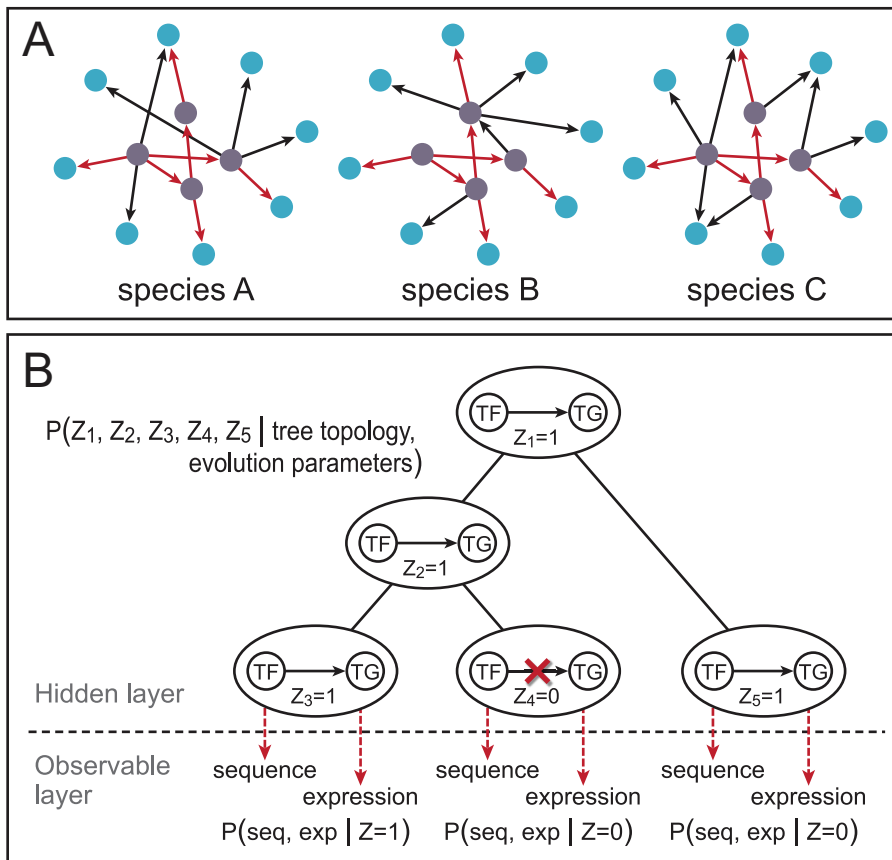
### Simulation study

We simulated a series of six synthetic datasets and used 5-fold cross validation to test the model (see Text S1). Cross-validation results showed almost identical prediction accuracies in training and testing datasets, across all simulated noise levels in sequence and in expression data, and across all choices of the weight parameter ( $\beta$ ), suggesting that it is fairly difficult to make model overfit (Figure S2).

### Transcriptional re-wiring in yeasts – a test of the model

To test the TN evolution model with real data, we used the recent discovery of a re-wiring event in yeast species. We wanted to use the analysis of this relatively well described re-wiring event to test the validity and precision of the new model.

Genes coding for mitochondrial and cytoplasmic ribosomal proteins display a strongly correlated expression pattern in *Candida albicans*, but this correlation is lost in the fermentative yeast *Saccharomyces cerevisiae*. Ihmels *et al.* associated this change in gene expression with the loss of a specific *cis*-regulatory element, AATTTT, from dozens of mitochondrial ribosomal protein (MRP) genes [29]. We attempted to reproduce this finding and potentially explore it in greater details with new data and the TN evolution model. Because the inferred loss of the *cis*-regulatory element happened after the separation of aerobic and anaerobic yeast species, we chose to analyze two anaerobic species, *Saccharomyces cerevisiae*, and *Candida glabrata*, and one aerobic species *Candida albicans*. We identified in the three species the orthologs of 51 MRP genes, 58 rRNA genes, and 73 stress response (STR) genes. While the three gene sets formed their individual expression clusters in *S. cerevisiae* and *C. glabrata*, MRP and rRNA genes appeared to be co-expressed in *C. albicans* (Figure S3-A). An enriched sequence motif was found by MEME [30] in the promoters of rRNA genes of all three species, as well as in the promoters of the MRP genes in *C. albicans* (Figure S3-B New Motif). We hypothesized that this motif may represent the binding specificity of a conserved TF, which we termed TF<sub>a</sub>. If Ihmels *et al.*'s finding can be reproduced, there should exist a TF that regulates MRP genes only in *C. albicans* but not in anaerobic species. We let our evolutionary model predict the transcriptional targets of TF<sub>a</sub> in all three species. We compared the model-predicted regulatory relationships to Ihmels' theory and found strong consistency (Figure S3-B) (Methods). This suggests that the TN evolution model captured the re-wiring of MRP genes as Ihmels *et al.* reported and provides additional support to the hypothesis that the re-wiring event is correlated with the divergence of aerobic and anaerobic species.



**Figure 1. A model for TN evolution.** (A) An example of evolving TNs. Transcription factors and target genes are depicted in purple and blue nodes. Conserved and species-specific regulatory relationships are depicted with red and black arrows. (B) The regulatory states are modeled as a continuous time Markov chain, and the regulatory sequence and the gene expression data are emitted from the hidden regulatory states. doi:10.1371/journal.pcbi.1002064.g001

Since the TN evolution model enabled the analysis to include the expression and the sequence data of a third species, *C. glabrata*, which was not present in Ihmels' analysis, we expected the model-based analysis to reveal more details regarding  $TF_a$  and its regulatory rules. To this end, we asked whether the 6bp *cis*-regulatory element, AATTTT, identified by Ihmels *et al.* was optimal. We compared AAATTTT (new) and AATTTT (Ihmels) by the number of target genes they can correctly predict in the three species. The new motif was more informative in predicting both the target genes and the non-target genes (Figure S3-B, panel ALL, Table S2). To assess the robustness of this result, we investigated every gene group (MRP, rRNA, STR), and we varied the weight of expression data ( $\beta$ ) used in the model. The new motif better distinguished the target and the non-target genes in all the settings tested (Figure S3-B), suggesting the new motif is a more faithful representation of  $TF_a$ 's binding preference. These results corroborated our expectation that the precision of the model is suitable for making discoveries.

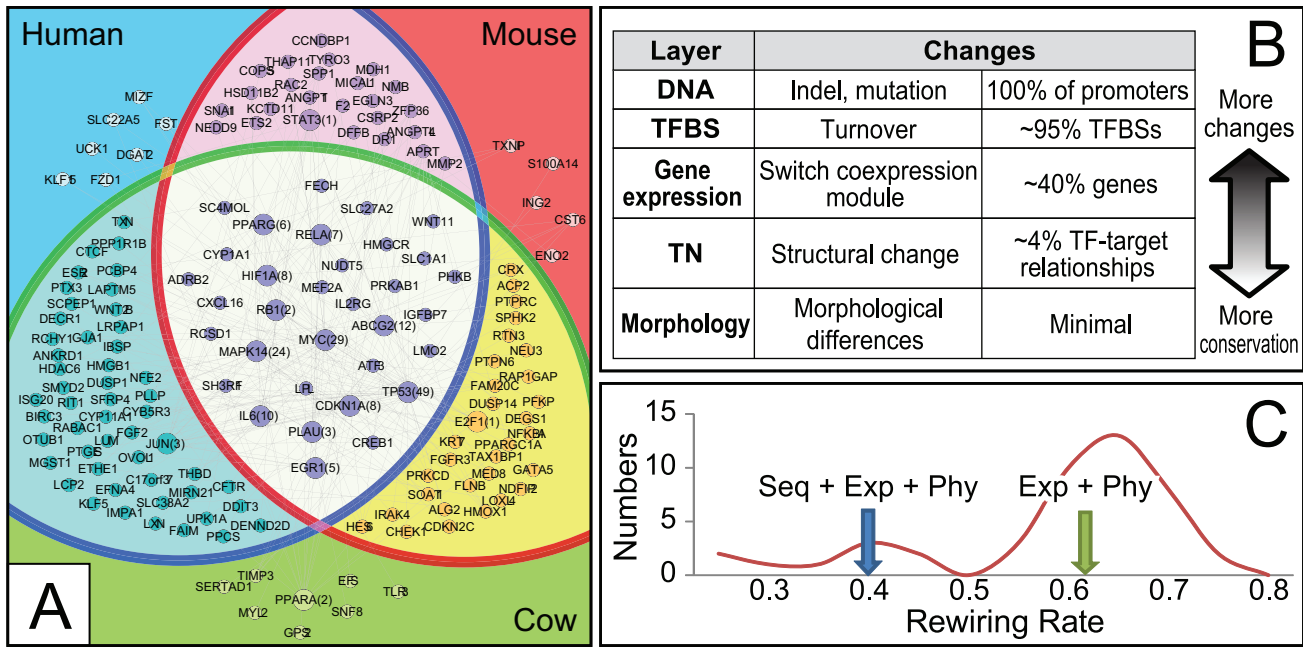
### How different are the TNs among mammals during early embryonic development?

To investigate the discrepancy between morphological conservation and molecular differences in mammals, we hypothesized that the TN structure does not evolve as quickly as the TF binding sites and gene expression. In other words, we hypothesized that although there are substantial amounts of TFBS turnovers and

gene expression changes across mammals, there are fewer changes in TN structure, *i.e.* TF-target regulatory relationships.

To test this hypothesis, we applied the TN evolution model to analyze the sequence and expression data in PED of humans, mice and cattle. Out of a total of 7046 orthologous gene triplets in the three species, 1489 of them fell into some co-expression modules (had non-constant expression and were not clustered as singletons). We chose Oct4 as the TF of focus in this study. The output of our evolutionary model is the regulatory relationship between a TF, Oct4 in this case, and every gene in every input species. Among the 1489 orthologous gene triplets, 823 (55.3%) were predicted to be regulated by Oct4 in all three species, and 113 (7.6%) were predicted to be only regulated in one species [nodes, Figure 2A]. In particular, only 40 (2.7%) orthologous triplets were regulated by Oct4 specifically in humans. This estimated fraction (2.7%) is much smaller than the fraction of genes with human-specific PED expression (45%,  $p\text{-value} < 1E-10$ ) [8], which in turn is much smaller than the fraction of human-specific Oct4 binding sites (95%,  $p\text{-value} < 1E-10$ ) [9] [Figure 2B].

To assess whether the model-inferred *smaller* interspecies difference in TF-target relationship than the previously reported interspecies difference of gene expression is due to model priori, we did two control experiments. First, the gene expression data alone was fed to the same evolutionary model, which led to an estimated 62.3% re-wiring rate [Figure 2C, Exp+Phy]. In this case the estimated re-wiring rate should be interpreted as the percentage of genes with unconserved expression patterns among



**Figure 2. Conservation and rewiring of Oct4 downstream TN.** (A) Venn diagram of model-inferred Oct4 target genes in three species. Each node is a gene and each edge represents a transcriptional regulatory link. When a set of genes are linked to a regulatory hub gene, this set of genes is suppressed into one node, annotated with the name of the hub gene followed by the number of other genes transcriptionally linked to the hub. (B) Inferred conservation levels for DNA, TFBS, gene expression, TN, and morphology. (C) Comparison of model-inferred re-wiring rates with (Exp+Seq+Phy) and without (Exp+Phy) DNA sequence data. The re-wiring rate of TF was calculated as the percentage of non-conserved TF-target relationships among all TF-target relationships. A background distribution (red curve) of the re-wiring rate is derived from randomly permuting the rows and columns of the Oct4 position specific score matrix (PSSM) and feeding the permuted PSSM to the model with unperturbed sequence and gene expression data. doi:10.1371/journal.pcbi.1002064.g002

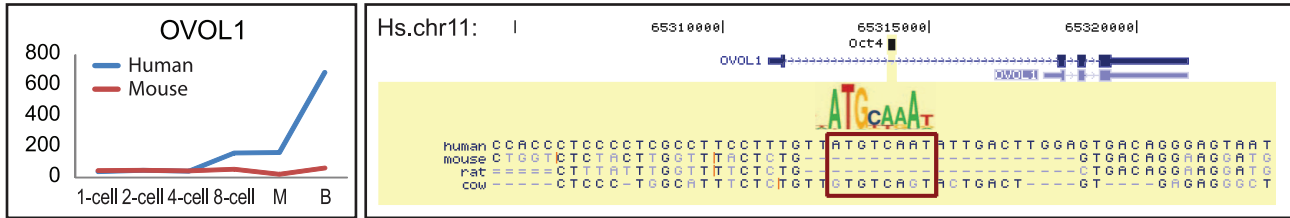
the three species, which is consistent with previously reported 55% [8]. Second, both sequence and expression data were fed to the evolutionary model together with randomly permuted Oct4 DNA binding motifs, which led to a distribution of the estimated re-wiring rates [Figure 2C, red curve]. The mode of this distribution was 0.66, and 82.2% of the estimated rates lay between 0.5 and 0.8. Both control experiments subjected gene expression data to the same model priori as the TF-target analysis, and consistently reported larger estimated interspecies changes of gene expression than TF-target relationship, considering the model inferred 44.7% (100%-55.3% conserved targets) re-wiring rate for Oct4.

We then asked to what extent the interspecies difference in Oct4 target genes affects the downstream gene regulatory networks. We mapped protein-protein interactions and transcriptional regulatory relationships among all the genes [Figure S8, Figure 2A]. Among all the possible protein-protein interactions (PPIs) [31,32,33,34], 134 interactions were found between the genes that were predicted to be conserved targets in all three species, whereas 0 interactions were found between the genes that were specifically regulated by Oct4 in any one species [gray edges, Figure S8] (Chi-square test p-value = 6.6E-210). Among all the transcriptional regulatory relationships [35,36], 270 regulatory relationships were found between the genes that were predicted to be conserved targets in all three species, whereas 1 interaction was found between the gene that were specifically regulated by Oct4 in only one species [gray edges, Figure 2A] (Chi-square test p-value = 8.7E-307). In summary, the data above suggest that compared to the changes in TFBS and gene expression, the transcriptional targets of Oct4 are more conserved, and so are the interaction and regulatory relationships of these Oct4 target genes [Figure 2B, Figure S6].

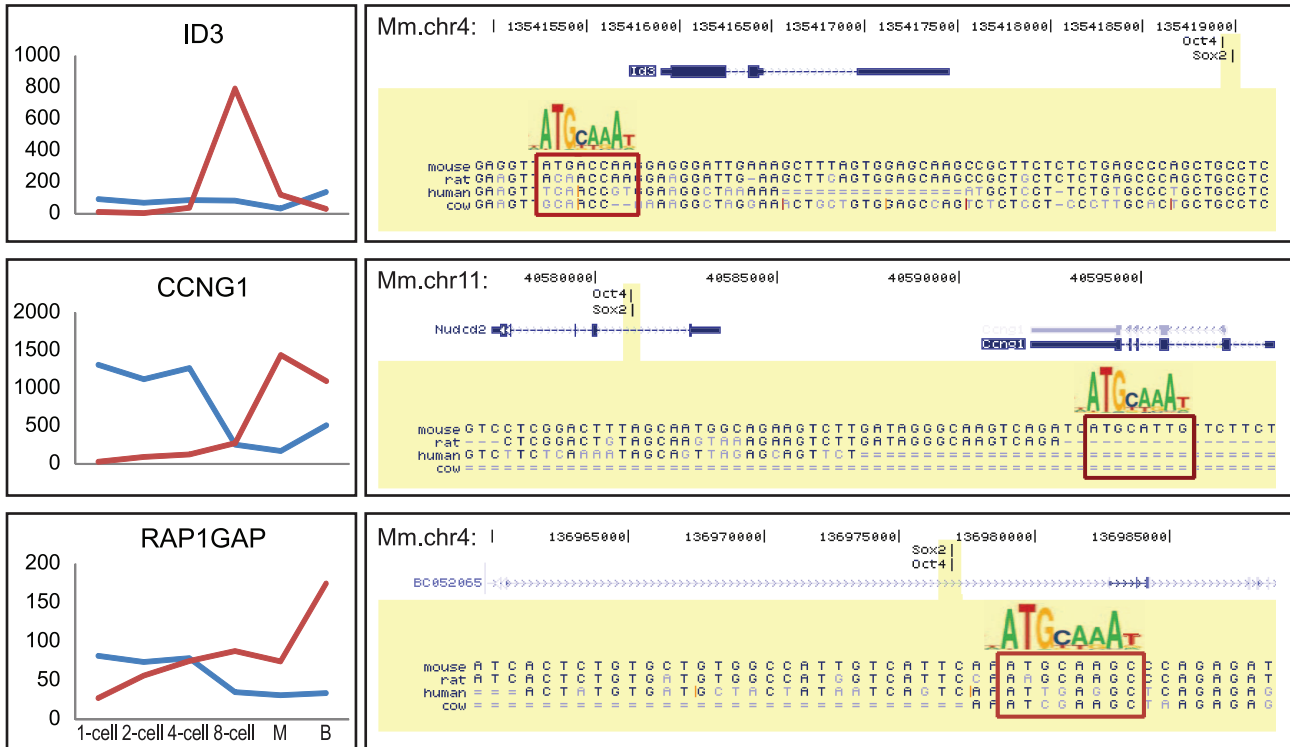
**Evidence for re-wiring of TNs in PED**

There is as yet no proven example of TN re-wiring events reported for early development in mammals, like the example Ihmels *et al.* demonstrated in yeast species. We wanted to identify a few concrete cases of TN re-wiring events. The TN evolution model provides a systematic approach to look for TN re-wiring events (Text S2 and Figure S4). Our model predicted 40 genes as Oct4 targets in humans but not in mice, and vice versa for 24 genes [Figure S5]. We applied two further criteria to these 64 genes to select for re-wiring events with the strongest evidence. First, because *Oct4* itself shows an upward trend of expression during PED, peaking at the blastocyst stage in humans and at the morula stage in mice [8], we selected the genes with clear up-regulation in the late stages of PED. Second, we selected the genes whose predicted Oct4 binding regions harbored clear gain or loss of Oct4 binding motifs. These selections produced four TN re-wiring events associated with the human-specific regulation of *OVOL1* by Oct4 and mouse-specific regulation of *Id3*, *Cngl1*, and *Rap1gap* [Figure 3]. Because embryonic stem (ES) cells were derived from the inner cell mass of blastocyst stage embryos (the last developmental stage of PED), we speculated that TN re-wiring events could be corroborated by gain or loss of Oct4 binding in ES cells. ChIP-seq data in human and mouse ES cells were consistent with this hypothesis [37,38] [“TFBS” track, Figure 3]. By reconstructing the ancestral sequence, we identified the evolutionary events including indels and mutations that mediated the TFBS turnovers [Figure S5]. In particular, the birth of the Oct4 TFBS near *Id3* gene appears to be mediated by a 2bp deletion from ACAgtACCGTG (ancestral) into ACAACCGTG (murine) [Figure S5-B]. TFBS birth by deletion has rarely been reported in vertebrates.

A. Regulated only in human



B. Regulated only in mouse



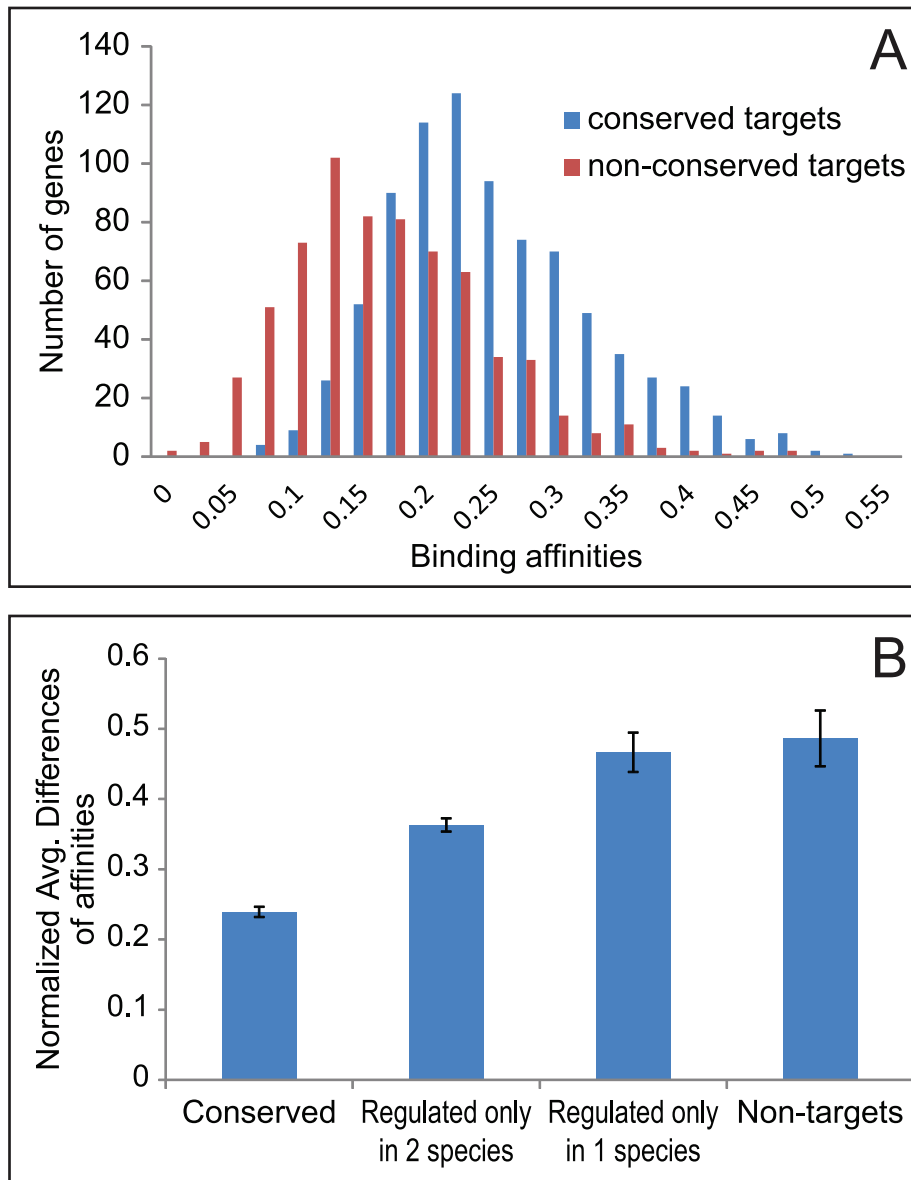
**Figure 3. Examples of model-inferred species-specific target genes of Oct4 in humans (A) and mice (B).** Oct4 is zygotically expressed (4–8 cells), and it is strongly increased at the late stages of pre-implantation development, including morula (M) and blastocyst (B). The target genes' expression is zygotically activated in a species-specific manner (Left panel). The species-specific binding of Oct4 to target genes is observed in ChIP-seq experiments (TFBS track, right panel). The genomic sequences of the Oct4 binding regions are shown in yellow, with Oct4 binding motifs shown in red boxes.

doi:10.1371/journal.pcbi.1002064.g003

Quantitative clues about cis changes that affect TN re-wiring

A long standing question is to what extent the evolutionary changes of the TF-target relationship are associated with cis changes. In other words, for a conserved TF, can we use the cis changes to infer changes in regulatory relationships? Except for testing done on a small set of experimentally characterized enhancers [39], genome-wide analysis attempts seemed to provide negative answers. For example, changes in DNA binding motifs do not seem to correlate well with changes in TF-DNA binding [40], and loss (gain) of in vivo TFBS may not affect target gene expression because they can be compensated by gain (loss) of in vivo TFBS in other regulatory regions of the same target gene [9]. The TN evolution model enabled us to revisit this question. In the case of Oct4 regulated genes, the conserved target genes (regulated by Oct4 in three species) harbored *cis*-regulatory regions (20 k bp

flanking TSS) with larger binding affinities to Oct4, as compared to the non-conserved target genes [Figure 4A]. This is consistent with the fact that motif information was used in the model. Moreover, the interspecies difference in the binding affinities, as determined by the Oct4 motif and the *cis*-regulatory regions, is inversely correlated with the conservation level of the TF-target relationship [Figure 4B]. More specifically, the average interspecies cis difference in the target genes that are conserved in all three species is 55% of the average interspecies cis difference of the target genes conserved in two species (not a target gene in the third species) (p-value = 5.15407E-76). The latter difference is in turn 75% of the average interspecies cis difference of the target gene in only one species (not a target in the other two species) (p-value = 2.877E-17); However, it is not statistically different from the interspecies cis difference of non Oct4 target genes (p-value = 0.1204).



**Figure 4. Relationship between conservation and binding affinities.** (A) The distribution of the average binding affinities between Oct4 and orthologous regulatory sequences in three species. (B) Interspecies difference in binding affinities of the 20 kb upstream sequences of orthologous genes. The orthologous genes are put into four categories of conservation. Error bars show 95% confidence intervals. doi:10.1371/journal.pcbi.1002064.g004

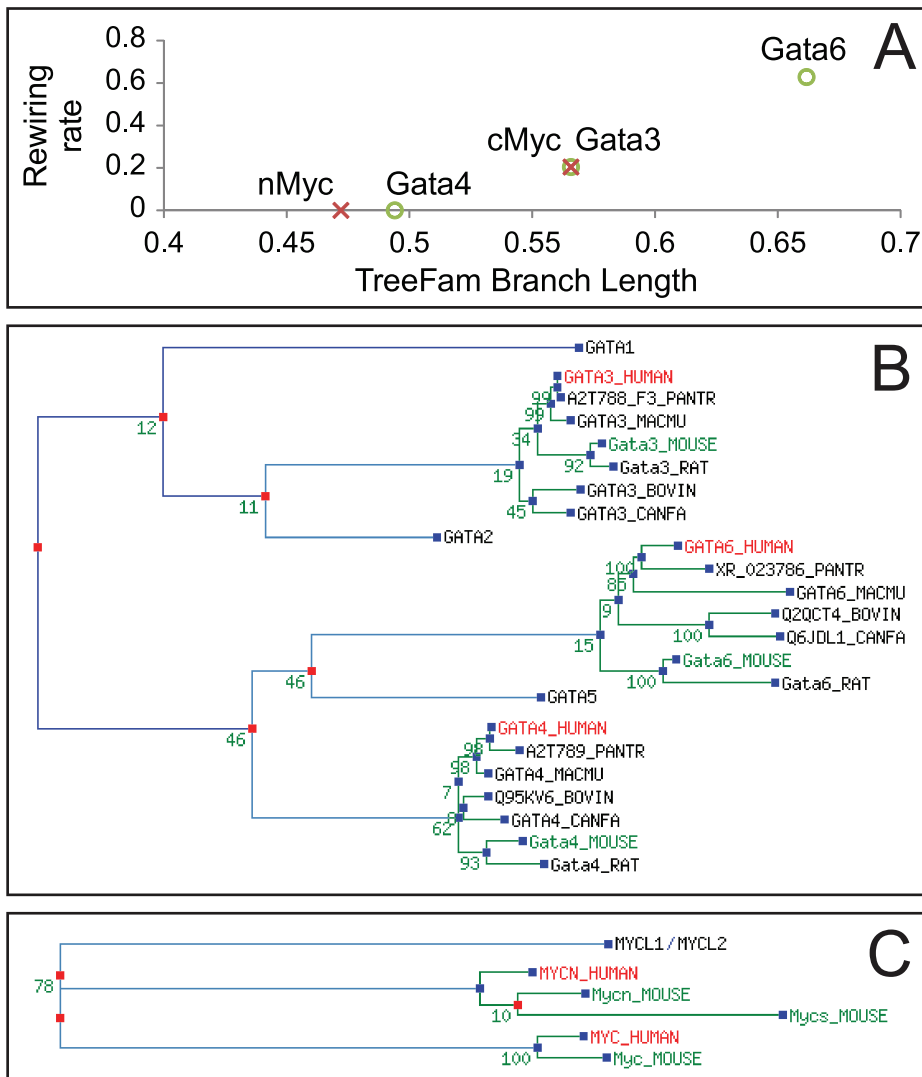
#### Do younger TFs have larger re-wiring rates?

Some TFs are evolutionarily younger than others. We asked whether the divergence time of a TF from its ancestor is correlated with its re-wiring rate. The re-wiring rate of a TF is defined as the percentage of non-conserved TF-target links among all TF-target links of a TF. Since our model infers the target genes of a TF in every species fed to the model, re-wiring rate can be directly derived from the model output. An extensive survey of TFs that may regulate cell fate decisions in mouse PED reported 29 TFs [41], among which there were two sets of paralogous genes including Gata3/Gata4/Gata6 and cMyc/nMyc [Figure 5]. Using the HKY DNA substitution model [42] implemented in TreeFam [43], we inferred the divergence time of each TF to the closest common ancestor of the paralogous group [Figure 5B,C]. In all four comparisons (cMyc vs. nMyc, Gata4 vs. Gata3, Gata6

vs. Gata3, Gata6 vs. Gata4), the TF with a shorter divergence time always showed a smaller re-wiring rate (the largest pairwise comparison  $p$ -value  $< 1E-10$ ), suggesting in mammalian PED, evolutionary younger TFs are more likely to change their regulatory targets.

#### Discussion

A major goal in biology is to understand the evolution of complex traits, such as morphology and behavior [44]. Pursuit of this goal may be catalyzed by evolutionary models for the molecular substrates, including the coding sequences, the *cis*-regulatory sequences [45,46], the epigenome, and the transcriptome [26], as well as by evolutionary models for the interaction or regulatory networks of these molecular substrates.



**Figure 5. Re-wiring rate against divergence time.** (A) Re-wiring rate against phylogenetic distance. Circles: Gata3/Gata4/Gata6 genes. Cross: cMyc(MYC)/nMyc(MYCN) genes. TreeFam branch length of a gene is the distance between the speciation event of the gene and the first duplication event in the paralogous family. (B, C) Phylogenetic trees of the Gata and the Myc families in mammals. Blue dot: gene speciation event. Red dot: gene duplication event. Branch lengths are estimated for the consensus tree of bootstrapped trees using HKY model. Numbers on branching events are the support numbers to the consensus tree in 100 bootstrapped trees. doi:10.1371/journal.pcbi.1002064.g005

A challenge in modeling these networks lies in the question of how to quantitatively associate gene expression with the “strengths” of *cis*-regulatory regions. From a single species perspective, this challenge has been approached by using sequence rules to predict gene expression under a classification scheme [47] and a regression scheme [48]. Because these schemes both resorted to statistical association, at least in some of their analysis steps, we could not generalize them onto an evolutionary model.

Instead, to link expression and sequence data [49], we introduced the notion of “regulatory state.” Conditional on the regulatory state, the probabilistic forms of both the sequence and the expression data were derived. This enabled a generative probabilistic model for the expression and the sequence data on a phylogenetic tree. We chose to model the expression data in a *soft* way, in the sense that we only modeled the difference of the co-expression patterns between the genes with different regulatory states. This choice suppressed a lot of information from the

expression data, but the model seemed to cope well with the noisy nature of the expression data and seemed to capture the essential information on TN in a robust way.

The model can potentially be generalized to treat the combinatorial control of multiple TFs. To do so, the total number of regulatory states should be extended to  $2^{\text{TF}\#}$ , where  $\text{TF}\#$  is the number of TFs in consideration. The evolution of the regulatory states can be modeled as a continuous Markov chain with  $2^{\text{TF}\#}$  states. The conditional probabilities of the sequence and the expression data should be derived from proper assumptions [50] and recently available information on combinatorial transcriptional regulation [51].

Our evolutionary model predicted that TF-target interactions are more conserved than expression patterns and TF binding events. These are conceptually sensible in several perspectives. First, the modular structure of GRNs allows a small change in the upstream regulators to manifest large changes in the expression of

downstream genes without perturbing the *cis*-*trans* interactions between these genes. For example, the change of the expression pattern or function, such as interactions with other partner proteins, of a master transcriptional regulator may change expression patterns of many downstream genes, while preserving their TF-target relationships by either unperturbed *cis*-regulatory sequences and the DNA binding domain on the TF, or compensatory *cis* and *trans* changes that retains regulatory control [51]. This view is consistent the theory of facilitated variation [52]. A recent example is the human transcription factor FoxP2, important for language, on which a small change in the TF (two amino acids, outside DNA binding domain) lead to differential regulation of hundreds of downstream genes [53]. Second, changes of TF binding events do not often lead to TN changes or gene expression changes. The binding sites of a TF may be gained and lost quickly during evolution, but these do not necessarily lead to change of TF regulation of a target gene, as the loss of one site may be compensated by the gain of another site elsewhere in the regulatory region of that gene [11]. Second, regulatory rewiring may happen where the regulator of a gene is switched to another TF in a different species without changing gene expression pattern [54].

The four predicted re-wired Oct4 target genes *OVOL1*, *Id3*, *Ccng1*, and *Rap1gap* were related to development, transcriptional control and signaling, enticing us to speculate that the predicted TN re-wiring events may contribute to interspecies differences during animal development. It turned out that PED is highly conserved among mammals: all progress through the same morphologic stages. Perhaps the most marked difference is the amount of time spent at each stage – a human zygote takes about one day to divide into a 2-cell embryo, whereas a mouse zygote only takes half a day. The predicted re-wiring event on Cyclin G1 (*Ccng1*), a cyclin regulating cell cycle, may be associated with the interspecies timing difference. This speculation led us to re-examine all of the 64 predicted re-wired target genes, which led to another gene *Orc5l*. *Orc5l* encodes a subunit of the origin recognition complex, essential for the initiation of the DNA replication in eukaryotic cells. Further experimentation is needed to test whether the differential regulation of *Ccng1* and *Orc5l* could lead to difference in cell cycle time.

Because changes in mRNA quantities may precede changes in protein levels, the inferred conservation and changes in TN may contribute to interspecies differences after the blastocyst embryonic stage. Notable differences include that bovine blastocysts initiate the gastrulation process before implantation, as well as the general formation and functions of the placenta and yolk sac [55]. We therefore mapped the Oct4 targets onto all developmental signaling pathways and identified the three pathways (Tgfb, WNT, mTOR) that contained more than two Oct4 targets [Figure S7]. The three species appear to use Oct4 to regulate conserved components in these three pathways, with a few exceptions. CtBP, a canonical inhibitor of the WNT signal pathway, appears to be regulated by Oct4 in humans and mice, but not in cattle. This might be correlated with the bovine specific role of WNT in placenta development [56].

The positive correlation between the evolutionary age of a TF and the conservation of its regulatory targets has seldom been exploited. The other studies with a similar flavor are on the analyses of the developmental hourglass model [57]. The developmental hourglass refers to the appearance of embryos in related species converges midway (called phylotypic stage) through development and diverges thereafter. Major supports of the model are that genes expressed during the phylotypic stage are both

evolutionarily older and more conserved across the genus than those expressed at other stages [58]. PED precedes the phylotypic stage in mammalian development. Still the data showed strong correlation between a TF's evolutionary age and its re-wiring rate. Now that in two separate developmental stages, there are evidences that evolutionarily older TFs serve to regulate more conserved sets of target genes. It would be interesting to see if this relationship checks out in other developmental stages and biological processes.

## Materials and Methods

### Symbols

**Indices**,  $i$ : observed species;  $l$ : ancestral species;  $m$ : gene.  $k$ : sequence locations;  $\varphi$ : nucleotide,  $\varphi = \{A, C, G, T\}$ ;  $c$ : gene clusters. **Observed data**,  $S$ : regulatory sequences;  $E$ : gene expression;  $N$ : total number of species;  $M$ : total number of orthologous gene groups;  $T$ : phylogenetic tree;  $t$ : phylogenetic distances (divergence time) between any two nodes. **Data derived from observed data**,  $n^0$ : number of background bases in  $S$ ;  $n^1$ : number of TF binding motifs in  $S$ ;  $LR$ : product of likelihood ratio scores of all motifs on  $S$ ;  $C$ : gene cluster index derived from  $E$ ;  $|c|$ : number of gene clusters. **Hidden variables**,  $Z$ : regulatory states.  $X$ : sequence states (background or motif). **Pre-computed parameters**,  $b_\varphi$ : nucleotide frequencies.  $\beta$ : a tuning parameter, representing the weight of the expression data in the likelihood function. **Model parameters**,  $\lambda, \mu$ : transition probabilities of regulatory states in unit time;  $\alpha$ : vector of the binary probabilities of the regulatory state in the root node;  $\omega$ : marginal probability of the motif state;  $p = \{p_1, \dots, p_{|c|}\}$ : vector of the multinomial probabilities of the cluster index of a target gene.  $q = \{q_1, \dots, q_{|c|}\}$ : vector of the multinomial probabilities of the cluster index of a non-target gene.  $\theta$ : the collection of all model parameters.  $\theta = \{\lambda, \mu, \alpha, p, q\}$ .

### Model

**Data and inference.** The data required by this model are genome sequences, gene expression data, a list of candidate TFs and their DNA binding motifs, and estimated divergence time. The model does not require prior information on the exact locations of TFBS or TF-target relationships. The TF-target relationships, *i.e.* the target genes of a TF in every species, are to be inferred by the model.

**Regulatory states and overall modeling strategy.** Let  $t$  denote the evolutionary time (Figure 1B). We call the regulatory relationship between a set of TFs ( $TF_x$ ) and a target gene  $m$  as the regulatory state of this target gene, denoted as  $Z_{i,TF_x,m}$ , where  $i$  is the species indicator. Here  $TF_x$  can be one TF or a few interacting TFs. Without loss of generalizability, we suppress the subscript  $TF_x$  in  $Z_{i,m}$ . Denote  $Z_{i,m} = 1$  when gene  $m$  is regulated by  $TF_x$  in species  $i$ , and  $Z_{i,m} = 0$  otherwise. Denote  $Z_m = \{Z_{1,m}, \dots, Z_{N,m}\}$ , where  $N$  is the total number of species. The variable  $Z_m$  thus indicates whether the regulatory link between the TF and the target gene is conserved or changed over evolutionary time. The general strategy of inferring  $Z_m$  is: if a gene is regulated by a TF, it is likely to contain the binding sites of this TF in its regulatory region and also likely to be co-regulated with other target genes of this TF.

We describe a probabilistic approach to estimate  $Z_{i,m}$ . The main idea is to express the joint probability of all sequences  $S_{i,m}$  and gene expression data  $E_{i,m}$  as a product of their conditional probabilities to  $Z_{i,m}$ . In other words, the likelihood of all sequence and expression data is:



$$P(S,E) = \prod_{m=1}^M P(S_m, E_m) \quad \text{and}$$

$$P(S_m, E_m) = \sum_{Z_m} P(Z_m) P(S_m, E_m | Z_m) = \sum_{Z_m} \left[ P(Z_m) \prod_{i=1}^N P(S_{i,m} | Z_{i,m}) P(E_{i,m} | Z_{i,m}) \right] \quad (1)$$

where  $N$  is the total number of species and  $M$  is the total number of orthologous gene groups. The model assumption is that in every species, *conditional on* the regulatory state of a target gene, the regulatory sequence of this gene and the expression level of this gene are independent.

This key assumption enabled us to write down the joint probability of the regulatory sequence and the expression level of every gene. We will describe the models for  $P(Z_m), P(S_{i,m} | Z_{i,m})$  and  $P(E_{i,m} | Z_{i,m})$  in the following sections.

**An evolutionary model of the regulatory states.** The evolution of the regulatory state  $Z_{i,m}$  is modeled as a two-state continuous-time Markov chain, with transition probabilities in unit time being  $\lambda(0 \rightarrow 1)$  and  $\mu(1 \rightarrow 0)$ . The transition probabilities of regulatory states between two species diverged for time  $t$  is:

$$\begin{pmatrix} P_{00}(t) & P_{01}(t) \\ P_{10}(t) & P_{11}(t) \end{pmatrix} = \begin{pmatrix} \frac{1}{\lambda + \mu} (\lambda e^{-(\lambda + \mu)t} + \mu) & 1 - P_{00}(t) \\ 1 - P_{11}(t) & \frac{1}{\lambda + \mu} (\mu e^{-(\lambda + \mu)t} + \lambda) \end{pmatrix},$$

where  $P_{x,y}(t)$  is the transition probability from state  $x$  to state  $y$  in time  $t$ ;  $x, y = \{0, 1\}$ . Given a phylogenetic tree  $T$ , the probability of the regulatory state on the leaf node is:

$$P(Z_{i,m}) = P(Z_{\text{root},m}) \prod_{l=\text{root}}^i P_{z_l, z_{\text{parent}(l)}}(t_{z_l, z_{\text{parent}(l)}}),$$

is the parent node of  $l$ ; and  $\prod_{l=\text{root}}^i P_{z_l, z_{\text{parent}(l)}}(t_{z_l, z_{\text{parent}(l)}})$  is the product of transition probabilities on the path starting from the root and ending at leaf node  $i$ .  $Z_m$  is the collection of the regulatory states in all leaf nodes, descending from the *same* root. It follows that

$$P(Z_m) = P(Z_{\text{root},m}) \prod_{i=1}^N \prod_{l=\text{root}}^i P_{z_l, z_{\text{parent}(l)}}(t_{z_l, z_{\text{parent}(l)}}) \quad (2)$$

Where  $N$  is the total number of observed species (leaf nodes). We introduce another model parameter  $\alpha$ , and let  $P(Z_{i=\text{root},m}) = \alpha$ .

**Sequence model.** The conditional probability of the regulatory sequence of the target gene  $S_{i,m}$  given the regulatory state  $Z_{i,m}$  is modeled as a Hidden Markov Model (HMM) [59]. The hidden layer, denoted as  $X_{i,m,k}$ , is a two-state (background, motif) Markov chain, where  $k$  is the index of the state random variable  $X_{i,m}$ . When  $Z_{i,m} = 0$ , the hidden layer stays at the background state with probability 1, *i.e.*  $P(X_{i,m,k} = 0 | Z_{i,m} = 0) = 1$  for any  $k$ . When  $Z_{i,m} = 1$ , the hidden layer transits between the two states with non-zero probabilities. We denote the marginal probability of an  $X_k$  being a motif as  $\omega$ , *i.e.*  $P(X_{i,m,k} = 1 | Z_{i,m} = 1) = \omega$ . We approximate the HMM by effectively assuming the hidden variables  $X$  can be inferred from a motif scanning procedure. Denote  $n_{i,m}^0$  and  $n_{i,m}^1$  as the number of observed background bases and the number of observed motifs,

respectively.  $n_{i,m}^0$  and  $n_{i,m}^1$  are determined by running a motif scan on  $S_{i,m}$ . The motif scan calls a segment on  $S_{i,m}$  as a motif when the likelihood ratio score of this segment reaches a pre-defined threshold. Under these model assumptions, it follows that

$$P(S_{i,m} | Z_{i,m} = 0) = \prod_k b_{k,\varphi} \quad (3.1)$$

where  $\varphi = \{A, C, G, T\}$ ,  $k$  is the index of DNA bases, and  $b_{k,\varphi} = b_\varphi$  is the background probability of a base being  $\varphi$ . It can be shown that

$$P(S_{i,m} | Z_{i,m} = 1) = P(S_{i,m} | Z_{i,m} = 0) \cdot LR_{i,m} \cdot \omega^{n_{i,m}^1} \cdot (1 - \omega)^{n_{i,m}^0} \quad (3.2)$$

where  $LR_{i,m}$  is the product of the likelihood ratio scores of all the reported motifs in  $S_{i,m}$ . Thus, we explicitly expressed  $P(S_{i,m} | Z_{i,m})$  by introducing one extra parameter,  $\omega$ , to the model. When there are multiple TFs to be considered, each with its own weight matrix,  $LR_{i,m}$  becomes the product of the likelihood ratio scores of all the motifs, reported from the scans of every matrix;  $n_{i,m}^0$  and  $n_{i,m}^1$  becomes the number of background bases and the total number of motifs for every TF.

**Expression model.** To model the conditional probability of the expression data, we first considered what makes a sensible and quantifiable difference in the expression data between the two regulatory states. We hypothesized that the transcriptional targets of a TF or a set of interacting TFs are likely to co-appear in co-expression modules. We implemented this idea by first clustering the expression data of all the genes. Let  $C_{i,m}$  be the cluster index of the  $m^{\text{th}}$  gene, and  $|c|_i$  be the total number of clusters.  $C_{i,m} | Z_{i,m} = 0$  follows a (background) multinomial distribution with parameters  $q_i = \{q_{i1}, \dots, q_{i|c|}\}$ , *i.e.*

$$P(C_{i,m} = c | Z_{i,m} = 0) = q_{ic} \quad (4.1)$$

where  $c = \{1, \dots, |c|\}$ . When  $Z_{i,m} = 1$ , the subset of genes, which are transcriptionally regulated by the TF or the set of interacting TFs, would tend to concentrate in a subset of the clusters. Thus,  $C_{i,m} | Z_{i,m} = 1$  follows another multinomial distribution with parameters  $p = \{p_1, \dots, p_{|c|}\}$ , *i.e.*

$$P(C_{i,m} = c | Z_{i,m} = 1) = p_{ic} \quad (4.2)$$

Thus, by inserting probabilities (2) - (4.2) into (1), we derived the complete likelihood of all data.

**Weighing sequence and expression data.** The likelihood model in (1), although it is completely specified, assumes equal weights of the sequence and the expression data. We further introduced a tuning parameter  $\beta$  to adjust the relative weights of the two data types. Thus, the model becomes:

$$P(S,E) = \prod_{m=1}^M P(S_m, E_m) \text{ and}$$

$$P(S_m, E_m) = \sum_{Z_m} P(Z_m) P(S_m, E_m | Z_m) = \sum_{Z_m} P(Z_m) \prod_{i=1}^N P(S_{i,m} | Z_{i,m}) P(E_{i,m} | Z_{i,m})^\beta \quad (5)$$

The larger  $\beta$  is, the more weight is given to the expression data.

**Model fitting.** We developed an estimation-maximization (EM) algorithm to estimate the model parameters. The E-step estimates  $Z_m$ , and the M-step maximizes  $\theta = \{\lambda, \mu, \alpha, p, q\}$ . Denote

$\hat{\theta}$  as the maximum likelihood estimator (MLE) of  $\theta$ . The complete E-M algorithm is available in Text S2.

**Model inference.** To identify the most likely regulatory states of all the genes in an orthologous group  $m$ , we use  $\max_{Z_m} P(Z_m | S_m, E_m, \hat{\theta})$  to simultaneously estimate  $Z_m = \{Z_{1,m}, \dots, Z_{N,m}\}$ .

## Methods

**Yeast data.** Gene expression data of *C. albicans* and *S. cerevisiae* were downloaded from the online supplementary material of Ihmels *et al.* [29]. Gene expression data of *C. glabrata* were collected from GEO (GSE6626, GSE6582, and GSE6058). The map of orthologous genes was obtained from the Yeast Gene Order Browser v.3 [60]. The genes that do not have orthologs in all the three yeast species were eliminated. The remaining 51 mitochondrial ribosomal protein (MRP), 58 rRNA, and 73 stress-related (STR) orthologous gene triplets were used in the TN evolution analysis. The 300bp upstream sequences to the TSSs were obtained from genome databases (www.yeastgenome.org, www.candidagenome.org, wolfe.gen.tcd.ie/ygob/Cglabrata\_sequence.fsa (v.3)). We used the phylogenetic topology and the evolutionary distances estimated by Tuch *et al.* [61].

**Preprocessing yeast data.** The 182 genes in each species were independently clustered by their expression data, using k-means clustering with 3 centers. An enriched sequence motif was identified by applying MEME [30] to the upstream sequences of *C. albicans*, *S. cerevisiae*, and *C. glabrata* rRNA genes, as well as *C. albicans* MRP genes (Figure S3-B New Motif).

**Comparing model prediction with Ihmels' theory.** Ihmels *et al.* suggested that the rRNA genes in all three yeast species and all the MRP genes in *C. albicans* are regulated by the conserved transcription factor TFa; moreover, none of the STR genes in any species and none of the MRP genes in any anaerobic species is regulated by TFa (Table S2) [29]. Our prediction is compared with this theory. For every gene in each species, the TN evolution model predicts whether it is regulated by TFa. We assumed a prediction to be correct if it matched Ihmels' theory, by checking Table S2. By matching Ihmels' theory we meant consistent to the notion that the MRP genes have been re-wired but other genes are not. For example, if a MRP gene in *C. albicans* was predicted to be a target of TFa, this prediction was considered correct. The prediction accuracy of the model was defined as the ratio between the total number of correct predictions and the total number of genes in the three species.

**Mammalian data.** The gene expression data of PED in humans, mice, and cattle were obtained from Xie *et al.* (GEO: GSE18290, GSE18319) [8]. The orthologous gene map was obtained from Xie *et al.* [8]. The upstream sequence of a gene was defined as the 20 k bp sequence flanking the TSS. These sequences were obtained from UCSC Genome Browser. The Oct4 motif was obtained from Chen *et al.* [37]. The phylogenetic distances of the three species were obtained from [62].

**Preprocessing mammalian data.** The orthologous gene triplets were filtered out, if none of the triplets had a clear change of expression levels (coefficient of variation >1.26) during PED. After the filtering, 1,509 orthologous gene triplets were passed onto clustering analysis.

The clustering of genes was performed independently within each of the three species, using a recently developed clustering method based on a Dirichlet Process [51]. Without requiring a pre-determined cluster number, this clustering algorithm automatically determines the optimum cluster number supported by the data. After removing singletons in the clustering result, 1,489

genes formed 27, 22, and 26 clusters in humans, mice, and cattle, respectively.

The probabilities  $P(S_{i,m} | Z_{i,m} = 1)$  and  $P(S_{i,m} | Z_{i,m} = 0)$  were computed for every 500bp sliding window on the 20 k bp sequence of a gene. The sliding window with the largest  $P(S_{i,m} | Z_{i,m} = 1) / P(S_{i,m} | Z_{i,m} = 0)$  was selected as the representative window of the 20 k bp sequence. The  $P(S_{i,m} | Z_{i,m} = 0)$  and  $P(S_{i,m} | Z_{i,m} = 1)$  from this representative window were used as the probabilities in (3.1) and (3.2) for the 20 k bp sequence [Figure S4].

**Choosing the weight ( $\beta$ ) parameter in PED analysis.** Having a weight ( $\beta$ ) parameter in the likelihood function is a commonly used approach to hybridize probability functions of heterogeneous data types. In our case the likelihood calculated from sequence data and the likelihood calculated from expression data can be in different scales. The weight serves as a scaling adjustment to make contributions from both sides comparable. In the analysis of PED data, we assumed that real data were very noisy and therefore were most similar to the "weak sequence signal and weak expression signal" case in the simulation. We directly assigned the weight ( $\beta = 35$ ) which seems to work fine in simulation on "both weak" data in the analysis of real data. We checked two other cases ( $\beta = 5$  and 50) and found in those cases the model predictions were the same as the model predictions when ignoring the expression and ignoring the sequence data, respectively, as expected.

## Supporting Information

**Figure S1** Histogram of human-mouse gene expression correlations. The human-mouse orthologous genes were identified by best blast bi-directional hits (BBH). Gene expression data were obtained from human and mouse gene atlas project, which used gene-chip microarrays to assay various tissues. In gene atlas data contained a total of 28 human-mouse matched tissues, and a total of 2,534 human-mouse BBH gene pairs on the microarrays. For each orthologous gene pair, a Pearson correlation  $\rho$  of their two-species gene expression was calculated, based on their expression levels in 28 matched tissues. 39.7% of the BBH orthologous pairs are negatively correlated; 64.2% had a correlation <0.1, and 91.2% had a correlation <0.6, suggesting large interspecies expression differences. (TIF)

**Figure S2** Prediction accuracies for simulated datasets. The prediction accuracy was plotted against the relative weight of the expression data ( $\beta$ ). Training and testing datasets were separate datasets. The prediction accuracies on training and testing datasets are almost identical, rendering the accuracy curves to overlap. (EPS)

**Figure S3** Rewiring of TNs among three yeast species. (A) Clustering of gene expression data in each species. The functional gene groups, including mitochondria protein genes (MRP), rRNA genes (rRNA), and stress response genes (STR), are correlated with gene clusters. (B) Prediction accuracy of regulatory relationships using the new motif (blue) and using the Ihmels *et al.* reported motif (red). (TIF)

**Figure S4** Preparing the sequence and the expression data for analysis by the evolutionary model of TNs. (EPS)

**Figure S5** Binding site turnover. Rex box indicates the predicted transcription factor binding sites (TFBS) in human (A) and mouse (B-D). Nucleotide sequences in red on top of the red boxes

represent common ancestral sequence reconstructed by parsimonious reconstruction. Mouse and rat experienced a deletion event that removed 4 bp out of the 8 bp TFBS, probably causing a death of the TFBS (A). Another 2 bp deletion from ACAgtACCGTG (ancestral) into ACAACCGTG gave birth to a murine specific binding site (B). Species-specific insertion (C) and mutation (D) could also lead to births of TFBSs.  
(TIF)

**Figure S6** Numbers of predicted conserved Oct4 target genes, protein-protein interactions, and gene regulatory relationships.  
(TIF)

**Figure S7** Conserved and alternatively regulated signaling pathway components. Canonical components of TGF $\beta$ , WNT, and mTOR pathways are shown. A gene in white is not a transcriptional target of OCT4-SOX2. A gene is colored with blue, red, or green when its human, mouse, or bovine ortholog is a transcriptional target of OCT4-SOX2, respectively. A gene with two or three colors is a target in two or three species. For example, c-Myc is colored blue, red, and green, and its orthologs in all three species are OCT4-SOX2 targets.  
(TIF)

**Figure S8** Model-inferred Oct4 target genes and the protein-protein interactions among the gene products.  
(TIF)

## References

- King M, Wilson A (1975) Evolution at two levels in humans and chimpanzees. *Science* 188: 107–116.
- Carroll SB (2008) Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. *Cell* 134: 25–36.
- Hoekstra HE, Coyne JA (2007) The locus of evolution: evo devo and the genetics of adaptation. *Evolution* 61: 995–1016.
- Wray GA (2007) The evolutionary significance of *cis*-regulatory mutations. *Nat Rev Genet* 8: 206–216.
- Xie D, Cai J, Chia NY, Ng HH, Zhong S (2008) Cross-species de novo identification of *cis*-regulatory modules with GibbsModule: application to gene regulation in embryonic stem cells. *Genome Res* 18: 1325–1335.
- He X, Chen CC, Hong F, Fang F, Sinha S, et al. (2009) A biophysical model for analysis of transcription factor interaction and binding site arrangement from genome-wide binding data. *PLoS One* 4: e8155.
- Xing Y, Ouyang Z, Kapur K, Scott MP, Wong WH (2007) Assessing the conservation of mammalian gene expression using high-density exon arrays. *Mol Biol Evol* 24: 1283–1285.
- Xie D, Chen CC, Ptaszek LM, Xiao S, Cao X, et al. (2010) Rewirable gene regulatory networks in the preimplantation embryonic development of three mammalian species. *Genome Res* 20: 804–815.
- Kunars G, Chia NY, Jeyakani J, Hwang C, Lu X, et al. (2010) Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nat Genet* 42: 631–634.
- Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, et al. (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci U S A* 101: 6062–6067.
- Doetschman TC, Eistetter H, Katz M, Schmidt W, Kemler R (1985) The in vitro development of blastocyst-derived embryonic stem cell lines: formation of visceral yolk sac, blood islands and myocardium. *J Embryol Exp Morphol* 87: 27–45.
- King DC, Taylor J, Eltnitski L, Chiaromonte F, Miller W, et al. (2005) Evaluation of regulatory potential and conservation scores for detecting *cis*-regulatory modules in aligned mammalian genome sequences. *Genome Res* 15: 1051–1060.
- Siepel A, Haussler D (2004) Combining phylogenetic and hidden Markov models in biosequence analysis. *J Comput Biol* 11: 413–428.
- Adjaye J, Huntriss J, Herwig R, BenKahla A, Brink TC, et al. (2005) Primary differentiation in the human blastocyst: comparative molecular portraits of inner cell mass and trophectoderm cells. *Stem Cells* 23: 1514–1525.
- Siddharthan R, Siggia ED, van Nimwegen E (2005) PhyloGibbs: a Gibbs sampling motif finder that incorporates phylogeny. *PLoS Comput Biol* 1: e67.
- He X, Ling X, Sinha S (2009) Alignment and prediction of *cis*-regulatory modules based on a probabilistic model of evolution. *PLoS Comput Biol* 5: e1000299.
- Ray P, Shringarpure S, Kolar M, Xing EP (2008) CSMET: comparative genomic motif detection via multi-resolution phylogenetic shadowing. *PLoS Comput Biol* 4: e1000090.
- Siddharthan R (2008) PhyloGibbs-MP: module prediction and discriminative motif-finding by Gibbs sampling. *PLoS Comput Biol* 4: e1000156.
- Ward LD, Bussemaker HJ (2008) Predicting functional transcription factor binding through alignment-free and affinity-based analysis of orthologous promoter sequences. *Bioinformatics* 24: 1165–1171.
- Segal E, Widom J (2009) From DNA sequence to transcriptional behaviour: a quantitative approach. *Nat Rev Genet* 10: 443–456.
- Khaitovich P, Weiss G, Lachmann M, Hellmann I, Enard W, et al. (2004) A neutral model of transcriptome evolution. *PLoS Biol* 2: E132.
- Gilad Y, Oshlack A, Smyth GK, Speed TP, White KP (2006) Expression profiling in primates reveals a rapid evolution of human transcription factors. *Nature* 440: 242–245.
- Tirosh I, Bilu Y, Barkai N (2007) Comparative biology: beyond sequence analysis. *Curr Opin Biotechnol* 18: 371–377.
- Oldham MC, Horvath S, Geschwind DH (2006) Conservation and evolution of gene coexpression networks in human and chimpanzee brains. *Proc Natl Acad Sci U S A* 103: 17973–17978.
- Mustonen V, Lassig M (2005) Evolutionary population genetics of promoters: predicting binding sites and functional phylogenies. *Proc Natl Acad Sci U S A* 102: 15936–15941.
- Cai J, Xie D, Fan Z, Chipperfield H, Marden J, et al. (2010) Modeling co-expression across species for complex traits: insights to the difference of human and mouse embryonic stem cells. *PLoS Comput Biol* 6: e1000707.
- Herrgard MJ, Covert MW, Palsson BO (2003) Reconciling gene expression data with known genome-scale regulatory network structures. *Genome Res* 13: 2423–2434.
- Quackenbush J (2003) Genomics. Microarrays—guilt by association. *Science* 302: 240–241.
- Ihmels J, Bergmann S, Gerami-Nejad M, Yanai I, McClellan M, et al. (2005) Rewiring of the yeast transcriptional network through the evolution of motif usage. *Science* 309: 938–940.
- Bailey TL, Williams N, Misleh C, Li WW (2006) MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res* 34: W369–373.
- Mishra GR, Suresh M, Kumaran K, Kannabiran N, Suresh S, et al. (2006) Human protein reference database—2006 update. *Nucleic Acids Res* 34: D411–414.
- Ulitsky I, Shamir R (2007) Identification of functional modules using network topology and high-throughput data. *BMC Syst Biol* 1: 8.
- Ewing RM, Chu P, Elisma F, Li H, Taylor P, et al. (2007) Large-scale mapping of human protein-protein interactions by mass spectrometry. *Mol Syst Biol* 3: 89.
- Wang J, Rao S, Chu J, Shen X, Levasseur DN, et al. (2006) A protein interaction network for pluripotency of embryonic stem cells. *Nature* 444: 364–368.
- Vailaya A, Bluvus P, Kincaid R, Kuchinsky A, Creech M, et al. (2005) An architecture for biological information extraction and representation. *Bioinformatics* 21: 430–438.

36. Cline MS, Smoot M, Cerami E, Kuchinsky A, Landys N, et al. (2007) Integration of biological networks and gene expression data using Cytoscape. *Nat Protocols* 2: 2366–2382.
37. Chen X, Xu H, Yuan P, Fang F, Huss M, et al. (2008) Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell* 133: 1106–1117.
38. Lister R, Pelizzola M, Downen RH, Hawkins RD, Hon G, et al. (2009) Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* 462: 315–322.
39. Marcellini S, Simpson P (2006) Two or four bristles: functional evolution of an enhancer of scute in *Drosophilidae*. *PLoS Biol* 4: e386.
40. Borneman AR, Gianoulis TA, Zhang ZD, Yu H, Rozowsky J, et al. (2007) Divergence of transcription factor binding sites across related yeast species. *Science* 317: 815–819.
41. Guo G, Huss M, Tong GQ, Wang C, Li Sun L, et al. (2010) Resolution of cell fate decisions revealed by single-cell gene expression analysis from zygote to blastocyst. *Dev Cell* 18: 675–685.
42. Hasegawa M, Kishino H, Yano T (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol* 22: 160–174.
43. Ruan J, Li H, Chen Z, Coghlan A, Coin LJ, et al. (2008) TreeFam: 2008 Update. *Nucleic Acids Res* 36: D735–740.
44. Toth AL, Robinson GE (2007) Evo-devo and the evolution of social behavior. *Trends Genet* 23: 334–341.
45. Majoros WH, Ohler U (2010) Modeling the evolution of regulatory elements by simultaneous detection and alignment with phylogenetic pair HMMs. *PLoS Comput Biol* 6: e1001037.
46. Tanay A, Regev A, Shamir R (2005) Conservation and evolvability in regulatory networks: the evolution of ribosomal regulation in yeast. *Proc Natl Acad Sci U S A* 102: 7203–7208.
47. Beer MA, Tavazoie S (2004) Predicting gene expression from sequence. *Cell* 117: 185–198.
48. Segal E, Raveh-Sadka T, Schroeder M, Unnerstall U, Gaul U (2008) Predicting expression patterns from regulatory sequence in *Drosophila* segmentation. *Nature* 451: 535–540.
49. Holmes I, Bruno WJ (2000) Finding regulatory elements using joint likelihoods for sequence and expression profile data. *Proc Int Conf Intell Syst Mol Biol* 8: 202–210.
50. Wagner GP, Lynch VJ (2008) The gene regulatory logic of transcription factor evolution. *Trends Ecol Evol* 23: 377–385.
51. Kuo D, Licon K, Bandyopadhyay S, Chuang R, Luo C, et al. (2010) Coevolution within a transcriptional network by compensatory trans and cis mutations. *Genome Res* 20: 1672–1678.
52. Gerhart J, Kirschner M (2007) The theory of facilitated variation. *Proc Natl Acad Sci U S A* 104 Suppl 1: 8582–8589.
53. Konopka G, Bomar JM, Winden K, Coppola G, Jonsson ZO, et al. (2009) Human-specific transcriptional regulation of CNS development genes by FOXP2. *Nature* 462: 213–217.
54. Tsong AE, Tuch BB, Li H, Johnson AD (2006) Evolution of alternative transcriptional circuits with identical logic. *Nature* 443: 415–420.
55. Gossler A (1992) Early mouse development. *Results Probl Cell Differ* 18: 151–201.
56. Ledgard A, Lee RS, Couldrey C, Peterson J (2009) Dickkopf-1 expression during early bovine placentation and its down-regulation in somatic cell nuclear transfer (SCNT) pregnancies. *J Reprod Dev* 55: 467–474.
57. Kalinka AT, Varga KM, Gerrard DT, Preibisch S, Corcoran DL, et al. (2010) Gene expression divergence recapitulates the developmental hourglass model. *Nature* 468: 811–814.
58. Domazet-Lošo T, Tautz D (2010) A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns. *Nature* 468: 815–818.
59. Capp e O, Moulines E, Ryden T (2005) Inference in hidden Markov models. New York: Springer. xvii, 652 p.
60. Gordon JL, Byrne KP, Wolfe KH (2009) Additions, losses, and rearrangements on the evolutionary route from a reconstructed ancestor to the modern *Saccharomyces cerevisiae* genome. *PLoS Genet* 5: e1000485.
61. Tuch BB, Galgoczy DJ, Hernday AD, Li H, Johnson AD (2008) The evolution of combinatorial gene regulation in fungi. *PLoS Biol* 6: e38.
62. Nikolaev S, Montoya-Burgos JI, Margulies EH, Rougemont J, Nyffeler B, et al. (2007) Early history of mammals is elucidated with the ENCODE multiple species sequencing data. *PLoS Genet* 3: e2.