PLOS | COMPUTATIONAL BIOLOGY

Functional Knowledge Transfer for High-accuracy Prediction of Under-studied Biological Processes

Christopher Y. Park[1]◑, Aaron K. Wong[1]◑, Casey S. Greene[2]◑, Jessica Rowland[3], Yuanfang Guan[4], Lars A. Bongo[5], Rebecca D. Burdine[3], Olga G. Troyanskaya[1,2]*

1 Department of Computer Science, Princeton University, Princeton, New Jersey, United States of America, **2** Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, New Jersey, United States of America, **3** Department of Molecular Biology, Princeton University, Princeton, New Jersey, United States of America, **4** Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, Michigan, United States of America, **5** Department of Computer Science, University of Tromsø, Tromsø, Norway

Abstract

A key challenge in genetics is identifying the functional roles of genes in pathways. Numerous functional genomics techniques (e.g. machine learning) that predict protein function have been developed to address this question. These methods generally build from existing annotations of genes to pathways and thus are often unable to identify additional genes participating in processes that are not already well studied. Many of these processes are well studied in *some* organism, but not necessarily in an investigator's organism of interest. Sequence-based search methods (e.g. BLAST) have been used to transfer such annotation information between organisms. We demonstrate that functional genomics can complement traditional sequence similarity to improve the transfer of gene annotations between organisms. Our method transfers annotations only when functionally appropriate as determined by genomic data and can be used with any prediction algorithm to combine transferred gene function knowledge with organism-specific high-throughput data to enable accurate function prediction. We show that diverse state-of-art machine learning algorithms leveraging functional knowledge transfer (FKT) dramatically improve their accuracy in predicting gene-pathway membership, particularly for processes with little experimental knowledge in an organism. We also show that our method compares favorably to annotation transfer by sequence similarity. Next, we deploy FKT with state-of-the-art SVM classifier to predict novel genes to 11,000 biological processes across six diverse organisms and expand the coverage of accurate function predictions to processes that are often ignored because of a dearth of annotated genes in an organism. Finally, we perform *in vivo* experimental investigation in *Danio rerio* and confirm the regulatory role of our top predicted novel gene, *wnt5b*, in leftward cell migration during heart development. FKT is immediately applicable to many bioinformatics techniques and will help biologists systematically integrate prior knowledge from diverse systems to direct targeted experiments in their organism of study.

Citation: Park CY, Wong AK, Greene CS, Rowland J, Guan Y, et al. (2013) Functional Knowledge Transfer for High-accuracy Prediction of Under-studied Biological Processes. PLoS Comput Biol 9(3): e1002957. doi:10.1371/journal.pcbi.1002957

Editor: Greg Tucker-Kellogg, National University of Singapore, Singapore

Received May 29, 2012; **Accepted** January 15, 2013; **Published** March 14, 2013

Funding: This work was supported by the National Science Foundation (NSF) CAREER [award DBI-0546275], National Institutes of Health [R01 GM071966, R01 HG005998 and T32 HG003284], and the National Institute of General Medical Sciences (NIGMS) Center of Excellence [P50 GM071508]. Funding for open access charge was provided by the National Institutes of Health. 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: ogt@genomics.princeton.edu

◑ These authors contributed equally to this work

Introduction

Defining the role of proteins in pathways is among the key challenges of human genomics. Many successful approaches have been developed for prediction of protein function and pathway membership [1–6], however they rely on prior knowledge in the organism of interest to make new predictions (i.e. at least some genes in the organism already annotated to the pathway) [7–11]. These approaches rely on identifying characteristic behavioral patterns, in functional genomic datasets, phylogenetic profiles, or genomic feature studies of genes that are known to participate in a pathway, then use these patterns to predict additional pathway members [12–14]. For example, gene expression and protein interaction profiles can be used by machine learning methods to associate novel genes to pathways based on previously known pathway members [15,16]. The potential of such computational approaches to direct experiments has been demonstrated in studies investigating mitochondrial biogenesis [17] and seed pigmentation [18]. Other common exploratory methods, such as hierarchical clustering [19], don't directly use known gene annotations to learn a prediction classifier, however they often use existing annotations to interpret the resulting cluster of genes (e.g. gene enrichment analysis) [20]. However in many organisms including human, pathways and processes where functional annotations of genes are most needed often have few or no prior experimentally confirmed annotations, making novel predictions of genes that may participate in such a process difficult or impossible. Thus, our study describes a method to robustly increase the set of prior gene

Author Summary

Due to technical and ethical challenges many human diseases or biological processes are studied in model organisms. Discoveries in these organisms are then transferred back to human or other model organisms. Traditional methods for transferring novel gene function annotations have relied on finding genes with high sequence similarity believed to share evolutionary ancestry. However, sequence similarity does not guarantee a shared functional role in molecular pathways. In this study, we show that functional genomics can complement traditional sequence similarity measures to improve the transfer of gene annotations between organisms. We coupled our knowledge transfer method with current state-of-the-art machine learning algorithms and predicted gene function for 11,000 biological processes across six organisms. We experimentally validated our prediction of *wnt5b*'s involvement in the determination of left-right heart asymmetry in zebrafish. Our results show that functional knowledge transfer can improve the coverage and accuracy of machine learning methods used for gene function prediction in a diverse set of organisms. Such an approach can be applied to additional organisms, and will be especially beneficial in organisms that have high-throughput genomic data with sparse annotations.

annotations, which has the potential to improve all function prediction methods by increasing the accuracy of their predictions and enabling wider coverage of pathways and biological processes.

Many of these processes are well studied in *some* model organism, but not necessarily in an investigator's organism of interest. Even when applying a conservative examination of only the closely related and heavily studied mammalian species human, mouse, and rat, processes represented in one species are often not well-characterized in another (summarized in Figure 1 and a full list of processes available in Text S1). For example, the process *cellular glucose homeostasis*, an increasingly important process due to the role of cellular metabolism in cancer development, has less than 5 gene annotations in human, yet has 31 in mouse, a commonly used model organism for cancer studies. These processes (referred to in the text as *understudied processes*) are not well studied in a particular organism of interest (i.e. very few genes are annotated) but might be well characterized in some other organism.

A longstanding solution to improving the prediction accuracy of understudied processes has been to transfer functional annotations from organisms where the process is better characterized [21]. The critical challenge in accurately transferring functional knowledge between organisms is identifying the appropriate genes for the transfer: those genes that share not only sequence similarity, but also conserved pathway roles. Large-scale automated methods have so far exclusively used sequence homology to identify

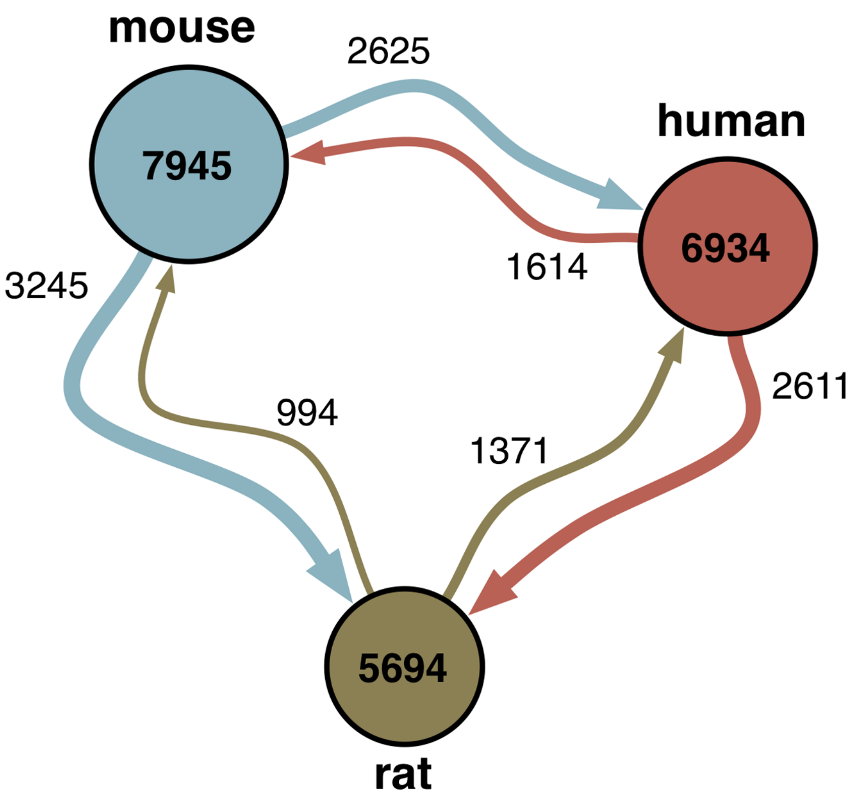


Figure 1. Functional knowledge of biological processes is far from uniform, even among closely related organisms. Each node represents the number of experimentally annotated biological processes in an organism. Each edge value corresponds to the number of experimentally annotated processes in the source organism that lack any experimental annotations in the target organism. Thus, the directed edges between nodes indicate the direction of potential annotation transfer between organisms. For example, 3,245 processes with annotations in mouse have no experimentally annotated genes in rat.
doi:10.1371/journal.pcbi.1002957.g001

functionally conserved genes [22,23]. However, the relationship between sequence similarity and function is not trivial. For example, human angiopoietin-4 (ANGPT4), an important angiogenesis growth factor, has been shown to activate TEK (tyrosine-protein kinase receptor), while the mouse sequence-ortholog (Angpt4) has been shown to inhibit TEK [24].

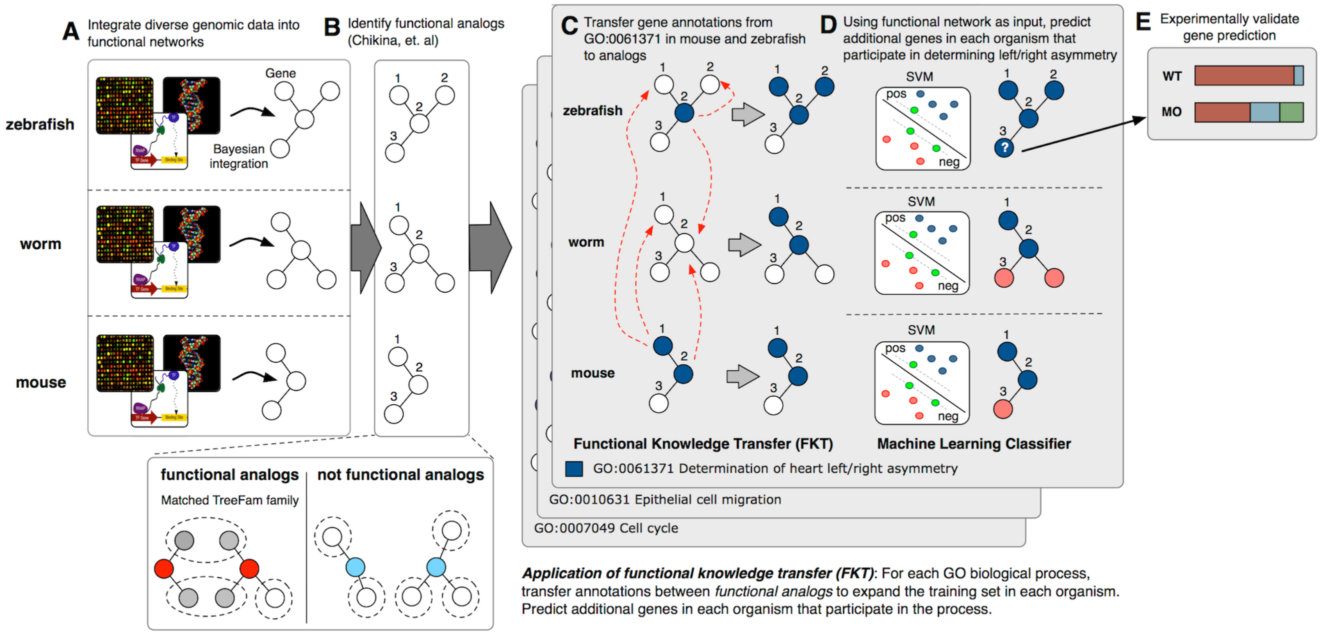
In our previous work [25], we developed a cross-organism gene functional similarity measure, which relied on the concept that functional genomics data can be used to resolve homologous relationships among closely related genes. The approach summarizes the compendium of genomics data in each organism into functional relationship networks to identify genes that do not simply share sequence similarity but also functional behavior in large collections of heterogeneous functional data, and are thus functionally analogous (referred to in text as *functional analogs*). In this current study, we present a novel knowledge transfer method, Functional Knowledge Transfer (also referred to in text as FKT and outlined in Figure 2), which leverages the mapping of functional analogs to direct cross-organism annotation transfer for function prediction. FKT can be especially beneficial for existing and future machine learning methods studying biological processes with sparse annotations in any given organism of interest. By transferring experimental knowledge between genes that have been identified as functional analogs, our method extends beyond simple annotation transfer by sequence similarity. Experimental functional annotations are only transferred for genes that are not just similar in sequence, but also in their functional behavior derived from a large and relatively comprehensive compendium of genomic data.

In this study, we show that FKT improves the prediction accuracy of machine learning algorithms, particularly for biological processes with few existing annotations in an organism of study. We compare FKT to annotation transfer by sequence similarity (BLAST) and demonstrate the superior performance of our method in improving gene function prediction performance. The consistent improvement and high performance across various state-of-the-art classification algorithms demonstrates our approach is robust to different learning models, which is crucial for wide applicability.

We apply FKT to gene function (i.e. biological process) prediction in six metazoan organisms (*Homo sapiens*, *Mus musculus*, *Rattus novegicus*, *Drosophila melanogaster*, *Danio rerio* and *Caenorhabditis elegans*) and show that FKT is robust enough for the automated transfer of annotations among these diverse organisms and accurate function prediction. Finally, we demonstrate an application of FKT to discovering novel biology by coupling the knowledge transfer with a Support Vector Machine (SVM) to predict proteins involved in left-right asymmetry regulation during heart development in *Danio rerio*. We correctly predict several proteins in the pathway and experimentally confirm the first evidence of *wnt5b*'s role in the process. A comprehensive application of FKT to 11,000 biological processes, along with the functional relationship networks for all six organisms, are available through the IMP web-server portal accessible at http://imp.princeton.edu [26].

Results

In Figure 2, we outline the pipeline for FKT and the subsequent gene function predictions (details provided in the Materials and Methods section below). Briefly, we first integrated high and low-throughput experimental data such as gene


A Integrate diverse genomic data into functional networks
B Identify functional analogs (Chikina, et. al)
C Transfer gene annotations from GO:0061371 in mouse and zebrafish to analogs
D Using functional network as input, predict additional genes in each organism that participate in determining left/right asymmetry
E Experimentally validate gene prediction
zebrafish
worm
mouse
Gene
Bayesian integration
1
2
3
SVM
pos
neg
?
WT
MO
Functional Knowledge Transfer (FKT)
Machine Learning Classifier
GO:0061371 Determination of heart left/right asymmetry
GO:0010631 Epithelial cell migration
GO:0007049 Cell cycle
functional analogs
Matched TreeFam family
not functional analogs

Application of functional knowledge transfer (FKT): For each GO biological process, transfer annotations between *functional analogs* to expand the training set in each organism. Predict additional genes in each organism that participate in the process.

Figure 2. Schematic of the functional knowledge transfer. (A) A functional relationship network is constructed for each organism through Bayesian integration of heterogeneous genomic data (e.g. expression, TF motif binding, physical interaction assays). (B) Functionally analogous gene pairs (i.e. functional analogs) are identified by computing a gene pairwise functional similarity score introduced in Chikina et. al between all sequence homologs. Functional similarity is measured by the statistical significance of the number of common TreeFam gene families in the functional relationship network neighbors of each homologous gene pair. (C) Next, experimentally confirmed biological process annotations for each gene are transferred to its functional analogs. (D) For each biological process the extended set of gene annotations (which include direct gene annotations, if available, and cross-annotated genes) can be used as training examples for machine learning methods (SVM used in this study) to make novel gene membership predictions. (E) Top predicted genes are carried over for experimental validation.
doi:10.1371/journal.pcbi.1002957.g002

expression data, protein-protein interaction data, protein domain and transcription factor binding motif information into functional networks for each of seven organisms (*Saccharomyces cerevisiae* was also included as an annotation source). Next, we calculated a network-based functional similarity score as described in our prior work [25] but extended here to additional organisms and data sources, between all ortholog and paralog pairs in a Treefam [22] gene family to identify the targets for annotation transfer. Homologs with high functional similarity scores were determined to be functional analogs. Next, we applied FKT by transferring all gene-process annotations between functional analogs and merge these with existing annotations (if available) in an organism. To test the predictive power of FKT, the set of transferred and organism-specific annotations were used to train a Support Vector Machine (SVM) classifier [27] and predict new genes to all biological processes in six metazoan organisms. Functional network connection weights (i.e. the inferred probability that two genes co-function in the same biological process), were treated as input features to the classifier (see Materials and Methods). Additional state-of-art machine learning methods (L1-regularized logistic regression [28] and Random forest [29]) were trained and evaluated to test the robustness of FKT performance improvement. Finally, we demonstrate the power of our approach with an *in vivo* experiment validating the predicted role of wnt5b in establishing correct heart asymmetry in *Danio rerio*.

Functional knowledge transfer enables accurate gene prediction for pathways with few or no known genes

Most modern machine-learning methods that predict novel members of a biological pathway require a set of genes already known to participate in the pathway. These approaches are therefore limited to predicting genes to biological processes with sufficient prior knowledge in an organism [30]. For example, in the MouseFunc competition [7] (a broad competition focused on the performance of biological process prediction approaches), terms with less than three gene annotations were considered infeasible to predict and not included.

We address this constraint by leveraging knowledge across species, which allows us to take advantage of known biology from a model organism where the pathway of interest may be better studied. We applied our functional cross-annotation strategy (FKT) to biological processes with few known genes (annotation sizes of $<=5$ and $<=15$) in six metazoans and evaluated the predictive performance of an SVM trained with these annotations. To evaluate our performance, we constructed a three-year temporal holdout of experimental annotations. We used only biological process annotations added to Gene Ontology [31] before 5/11/2008 (all dates in mm/dd/yyyy format) in learning the functional networks, transferring annotations across organisms, and predicting gene-process participation. New experimental annotations added to Gene Ontology between 5/11/2008 to 5/11/2011 were held-out and used for evaluation. In total, 3,207 GO biological process terms across the six organisms acquired new gene annotations in the subsequent three years. We evaluated the accuracy of our predictions with the gene-process assignments made during the hold-out time period in Figure 3 (evaluation results of all GO terms in Text S3). We observed substantial improvement using FKT when compared with only using the direct annotations for an organism. Improvement was evident across all six organisms, suggesting that even well characterized model organisms (e.g. mouse) can benefit from genomic-data-driven knowledge transfer. In addition, by holding out gene-process annotations acquired within the last three years, we could evaluate our ability to predict genes to processes which had no known genes in an organism prior to the hold out date (i.e. before 5/11/2008). Even though these processes were uncharacterized at that time, they subsequently became the focus of a directed experiment and thus were deemed biologically relevant and experimentally feasible in the organism. As shown in Figure 3, FKT gene predictions to these processes performed competitively even compared to biological processes with known gene annotations. Furthermore, these results were robust to the evaluation timeframe (1-year temporal holdout shown in Figure S1).

We hypothesized that our transfer method could improve prediction performance for a wide range of machine learning methods. Machine learning algorithms are often based on distinct learning models and assumptions, thus any widely applicable annotation transfer method must be robust to not only the biological variability (e.g. different organisms or pathways) but also to this modeling variability. Thus in addition to SVM, we evaluated two widely used state-of-the-art learning methods: L1-regularized logistic regression [28] and Random forest [29]. We trained both classification methods with and without FKT and evaluated on the held-out set of annotations. FKT improved prediction accuracy across each machine-learning algorithm and organism (Figure 4). In particular, these improvements were consistent across biological process annotation sizes ($<=5$ and $<=15$). Altogether, these results indicated that FKT could recover biological processes that would be otherwise missed by most prediction methods, and that the transfer had wide applicability - improving performance across diverse organisms and machine learning algorithms.

Genes predicted to processes with no prior annotations in the study organism reflect subsequent experimental findings

We coupled FKT with an SVM and applied the machine learning classifier to predicting novel gene functions in six organisms. These predictions included gene-process membership for 8,091 GO biological processes currently without experimental annotations in at least one organism. Supervised machine learning methods would be unable to predict novel genes to these biological processes without annotation transfer. They represent a wide range of biological pathways and processes ranging from development and metabolism to immune response and response to various stimuli (a complete list of these GO terms is in Text S2, categorization and specificity of these terms are shown in Figure S2, S3).

For example, the biological process *regulation of exit from mitosis* (GO:0007096) represents a crucial mitotic cell cycle process that enables cells to regulate their exit from M phase. This process had no experimental annotations in *Danio rerio* at the time of our study, however had been extensively studied in the model organisms *Saccharomyces cerevisiae* [32], *Mus musculus* [33] and *Drosophila melanogaster* [34]. Our functional cross-annotation method has identified a total of 18 genes in *Danio rerio* with functional analogs annotated to this process (11 from yeast, 5 fly, 1 mouse and 1 rat), enabling novel predictions of gene membership to this process.

Our top gene prediction for this process, *cdh2*, has been experimentally confirmed in a recent study examining cell cycle progression in *cdh2* mutant retina cells [35]. Interestingly, *cdh2* is not only a novel prediction in *Danio Rerio* (i.e. this gene function was unknown at the time of our study), but also no *cdh2* homologs are known to be involved in the *regulation of exit from mitosis* in other organisms. *Cdh2* is a member of the cadherin protein family, which are important transmembrane proteins that play a crucial role in

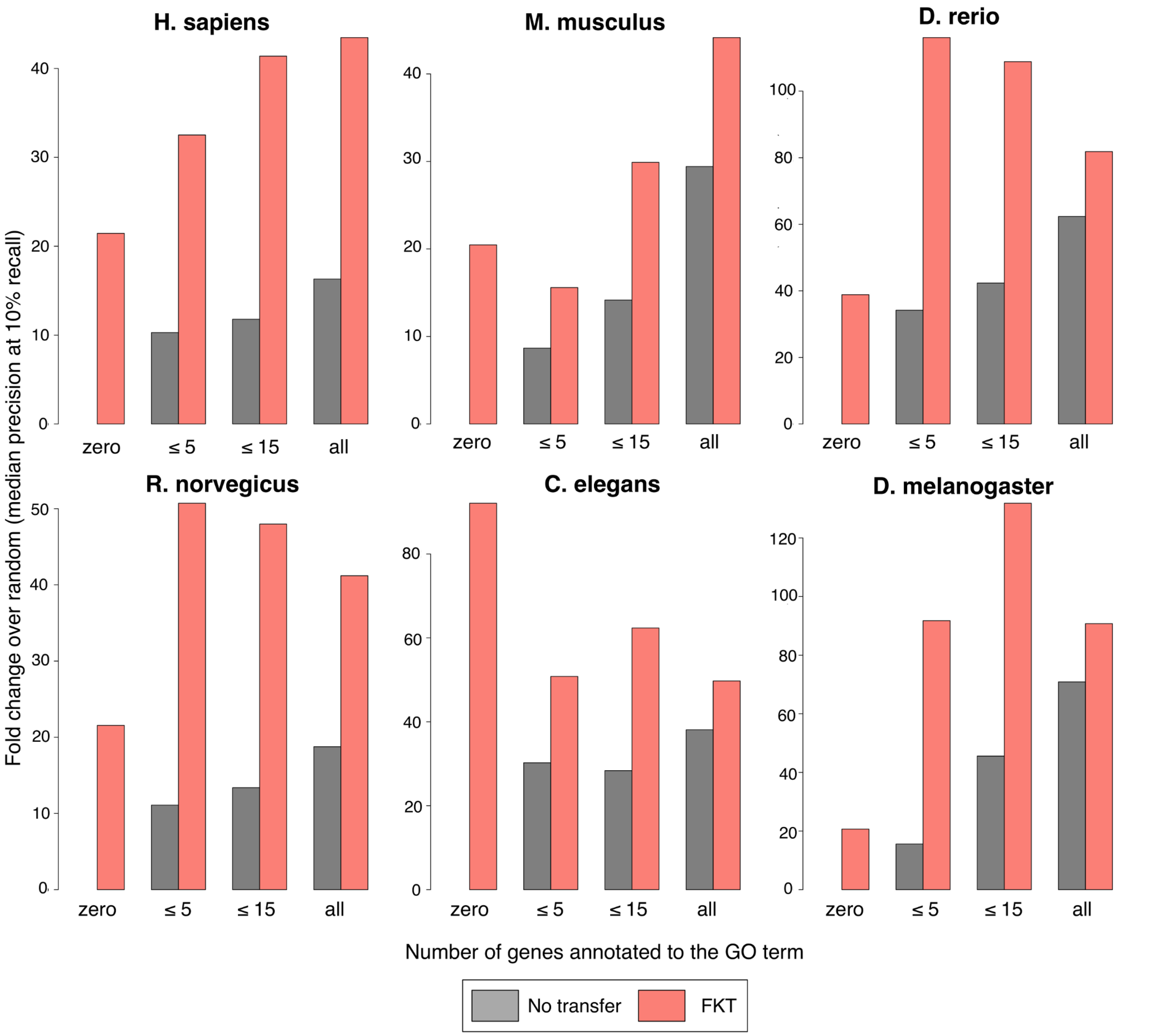


Figure 3. Cross annotation allows accurate recovery of small and unannotated terms. All annotations accumulated after 5/11/2008 are held out from our prediction pipeline (as outlined in Figure 2) and are used for evaluation of prediction performance. 3,207 GO biological processes terms that acquired new annotations subsequent to our holdout date are grouped by organism and by the number of annotations at 5/11/2008 (zero, $<=5$, $<=15$, all). Performances at recapitulating future annotations are compared for a machine learning method (SVM) without (gray) and with (pink) including functional knowledge transfer (FKT) derived examples. For processes with zero annotations before 5/11/2008, no predictions can be made without cross-annotation (shown as absent performance bar). In all six metazoan organisms and for all process sizes, FKT improves prediction performance.
doi:10.1371/journal.pcbi.1002957.g003

cell adhesion in multi-cellular organisms. Methods that employ only sequence similarity would be unable to predict this because *cdh2* homologs have not been annotated to this process in other model organisms. Furthermore, prediction methods without FKT will miss this finding because there are no existing *Danio rerio* annotations to this process. Only methods coupling FKT with a machine learning algorithm can take advantage of information from the single cell model organism *Saccharomyces cerevisiae*, where cell-cycle checkpoints have been extensively studied [36], and successfully predict this finding in the multicellular model organism *Danio rerio*. This *in vivo* experimental result demonstrates FKT's utility for predicting novel genes to understudied processes. In addition, by coupling functional transfer to machine learning methods that leverage organism-specific functional data collections, we can make reliable gene-process predictions even without an annotated sequence-homolog.

Cross-annotation among functional analogs improves prediction accuracy for small processes

To compare our functional transfer method, which applied a more specific annotation transfer, to commonly used methods that used only sequence homology, we evaluated a method that did not

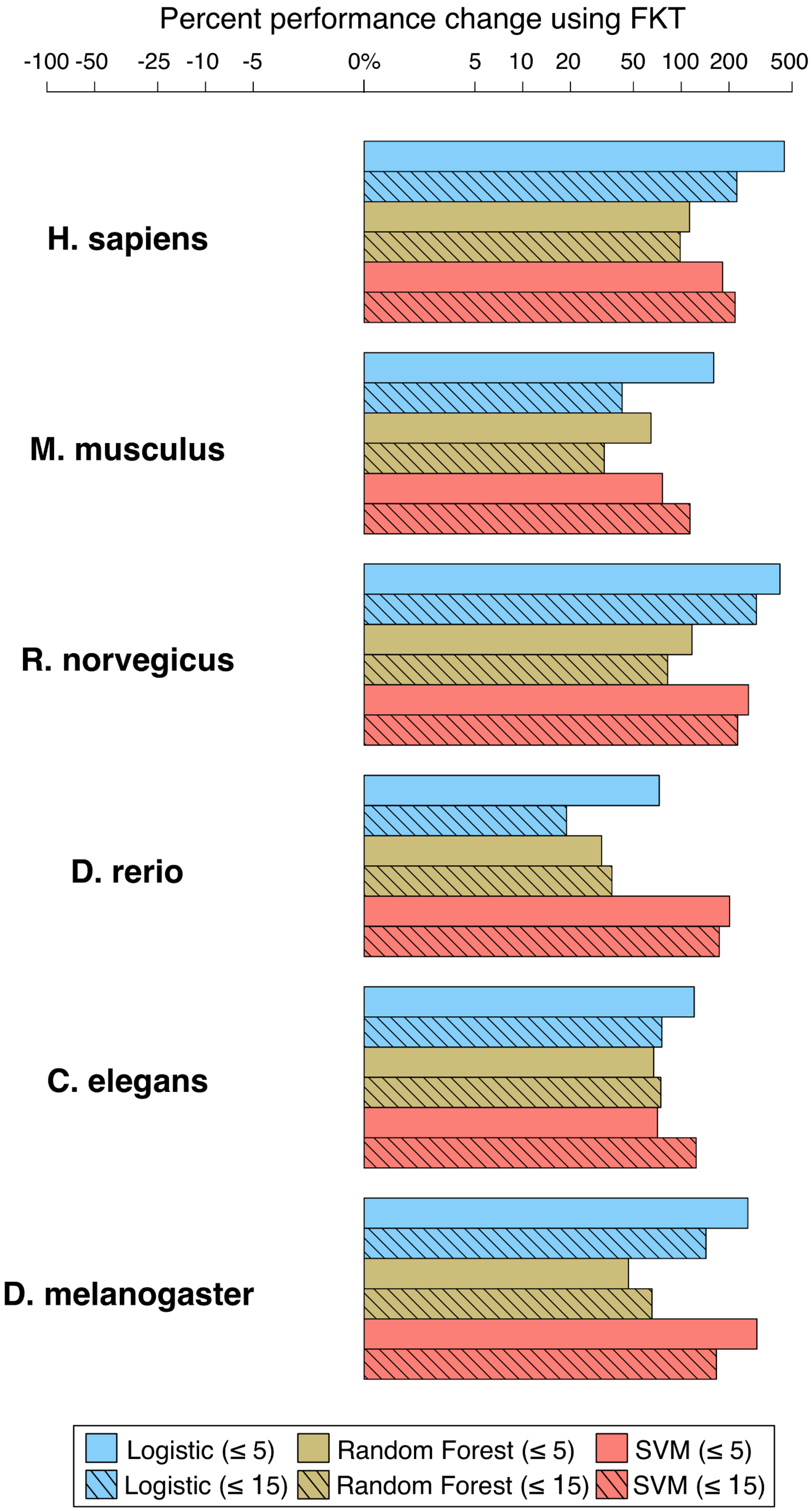

Percent performance change using FKT
-100
-50
-25
-10
-5
0%
5
10
20
50
100
200
500
H. sapiens
M. musculus
R. norvegicus
D. rerio
C. elegans
D. melanogaster
Logistic (≤ 5)
Random Forest (≤ 5)
SVM (≤ 5)
Logistic (≤ 15)
Random Forest (≤ 15)
SVM (≤ 15)

Figure 4. Functional knowledge transfer (FKT) improves prediction accuracy for a wide range of state-of-the-art classification algorithms. The performance change when applying FKT are compared for each of three machine learning algorithms: L1-regularized logistic regression, Random forest and SVM (evaluated based on the ability to recapitulate held-out annotations accumulated after 5/11/2008). 3,207 GO biological process terms are shown, grouped by annotation size at 5/11/2008 (<=5, <=15). The percent change in performance (median fold over random) when applying FKT compared to no FKT with each machine learning algorithm is shown for six diverse organisms. All bars are to the right of zero, indicating a performance improvement when FKT is applied for each machine learning algorithm.
doi:10.1371/journal.pcbi.1002957.g004

leverage functional similarity and a baseline method without any cross-annotation. In this sequence-only method, all homologous gene pairs (reciprocal BLAST best hit gene pairs) were targets for annotation transfer - any biological process annotated to a gene was transferred to its reciprocal best-hit gene in all organisms. To obtain a representative set of gene-process annotations for evaluation, we conducted a threefold cross-validation on genes that had experimental biological process annotations, and evaluated the SVM classifier prediction performance on each corresponding held-out set of biological process annotations. The results of the comparison showed that although both methods improved performance for small processes, FKT showed greater performance gains (Figure 5, evaluation results of all GO terms in Text S3). In a few organisms, the performance gains were substantial - for example, in human and mouse, the median performance (precision at 10% recall) increased more than fivefold.

Upon examining the processes that improved the most when compared to a sequence-only method, many pathways and processes with transcriptional based regulatory control showed improved performance using FKT. *Response to mechanical stimulus*, *ameboidal cell migration*, *regulation of neuron differentiation* and *striated muscle cell development* were among the top improved processes in all organisms using FKT compared to sequence-only. Unsurprisingly, these processes have been well known to be tightly regulated through transcriptional programs (e.g. stress response, developmental TF gradients) [37–39] and have multiple datasets measuring the transcriptional profiles incorporated in our functional networks [40–42].

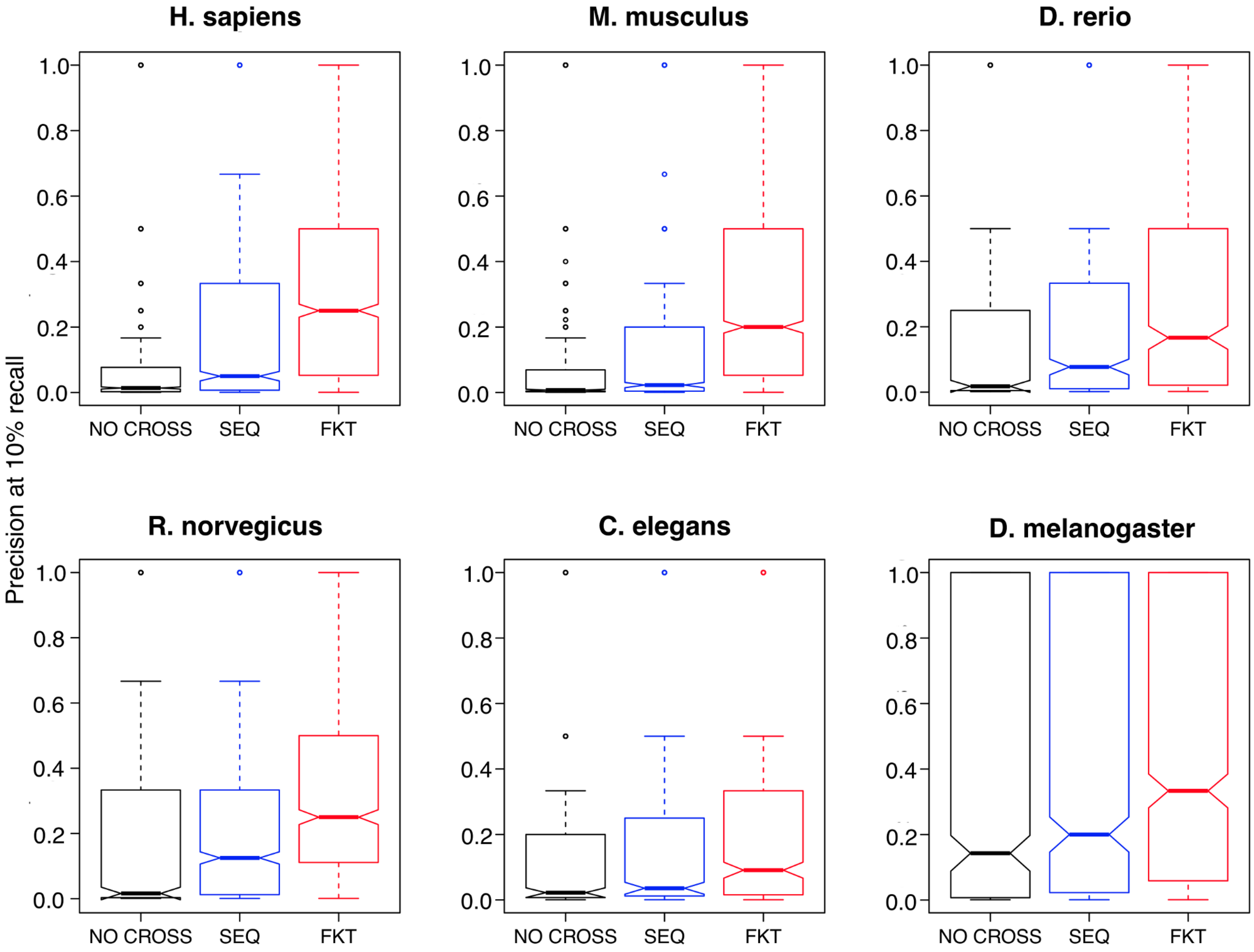


Figure 5. Functional knowledge transfer (FKT) improves performance for predicting small processes. The performance of two knowledge transfer methods (FKT and sequence-only) and a baseline method (with no cross-annotation) are compared. Shown here are results of threefold cross-validation for small processes (<=15) that represent specific or understudied pathways. FKT paired prediction method shows improved performance compared to both sequence-only transfer and the baseline method.
doi:10.1371/journal.pcbi.1002957.g005

We expect that FKT will continue to improve as the functional genomics compendia for many organisms continue to grow, including expression and other types of measurements across multiple perturbations. An additional advantage of a functional genomics similarity approach, as shown in [25], is the ability to differentiate functional differences in tissue specificity between sequence homologs. The example of mouse RNA polymerase II elongation factor *Supt5h* and its direct sequence-ortholog *C. elegans spt-5* highlight this issue. FKT determined these sequence-orthologs as not being functional analogs. Indeed, mouse Supt5h is predominantly neuronal, while *C. elegans* SPT-5 is non-neuronal and primarily expressed in the intestine and pharynx [43–45]. Even though these sequence-orthologs have diverged in tissue specificity, they still share high sequence similarity and a sequence-only method would inappropriately transfer all functional annotations between them.

In vivo validation of *Danio rerio* gene *wnt5b* involvement in the establishment of heart asymmetry

In all vertebrates, the heart develops with a distinct left-right (L-R) asymmetry during embryonic morphogenesis. Deviations in left-right heart development can lead to complex congenital heart defects that are among the most common human neonatal diseases [46,47]. During cardiac morphogenesis in *Danio rerio*, two distinct stages of cell migrations lead to the final asymmetries of the heart. In the first stage, called "heart jogging", myocardial cell migration within the cardiac cone place the ventricular cells to the left side, while atrial cells remain near the midline. In the second stage of "heart looping", the heart tube bends and forms a loop that places the ventricle to the right of the atrium. Although the steps of cell migration progression leading to left-right heart asymmetry are beginning to be explored [48–51], an understanding of how it is achieved mechanistically is still lacking.

In Gene Ontology, the biological process term "determination of heart left right asymmetry" (GO:0061371) represents the developmental pathways regulating heart jogging and looping. To validate our prediction method (FKT coupled with SVM), we investigated the top five predicted genes that had not already been annotated to this GO term: *sox32*, *wnt5b*, *ndr1*, *tbx1* and *lft1*. We found existing literature evidence confirming the involvement of four of the five genes (*sox32* [52–55], *ndr1* [56], *tbx1* [57], and *lft1* [58–60]). Although there existed experimental results confirming the role of these genes in influencing heart asymmetry, these results had not yet been curated by GO annotators. For example, in a knock-out experiment of our top predicted gene (*sox32/casanova*), *Danio rerio* embryos had fewer dorsal forerunner cells which led to defects in Kupffer's vesicle formation and subsequent left-right patterning of the heart, confirming that *sox32* was required for proper establishment of heart asymmetry. The only gene among the top five without existing experimental support was *wnt5b*, our second ranked prediction after *sox32*. Previous work had shown the involvement of *wnt5b* in cell migration during gastrulation [61] but the gene had not been specifically associated with heart left-right asymmetry regulation. To experimentally validate our prediction of *wnt5b* to left-right heart determination, we knocked down its function by means of morpholino antisense oligonucleotides (MO) [62].

A significantly greater proportion of embryos where *wnt5b* was knocked down with a morpholino (Figure 6) had a defective heart jogging phenotype (Fisher's exact test p-value$<$0.001). In total, 48% of morpholino treated embryos showed either right-sided heart jog or midline jog comparable to previous genes known to be involved in this biological process [63–65]. Only 4% of wild type and control-MO treated embryos exhibited this phenotype. This phenotype is likely due to the disruption of asymmetric expression of the TGFbeta member Nodal (data not shown), which is typically asymmetrically expressed on the left side of vertebrate embryos during somitogenesis. Left-sided Nodal in *Danio rerio* myocardial cells directs their subsequent migration during asymmetric cardiac morphogenesis [48,51]. Further investigation would be necessary to understand the mechanistic role of *wnt5b* in left-right heart determination, however our *in vivo* experiment confirmed the regulatory role of *wnt5b* in *Danio rerio* left-right asymmetry determination in heart development, as our method predicted.

Discussion

This study demonstrates that state-of-the-art machine learning methods coupled with our functional knowledge transfer method can accurately prioritize novel genes of understudied processes. Previous methods have focused on incorporating functional genomic data primarily as input data [66–69]. In contrast, here we demonstrate that the prevalence of understudied processes and the abundance of genomic data provide an opportunity to improve the accuracy of cross-organism annotation transfer and extend prediction coverage to processes with no prior annotations. We now integrate FKT into our IMP web-server [26]. This makes IMP a web interface for exploratory analysis covering all organisms included in this study across 10,653 biological processes (http://imp.princeton.edu). Functional knowledge transfer allows IMP to also include gene predictions for processes currently unannotated in an organism. Although in our current study we have experimentally followed up on our top predicted gene, all of our predictions in IMP are shown with estimated probabilities allowing biologist to draw a threshold dependent on how much the assay costs, and how important it is to find true positives (versus not finding false positives). In addition, the website includes the Bayesian functional relationship networks that were used for mapping functional analogs and used as input features to the machine learning methods. In particular, to the best of our knowledge, we include the first zebrafish (*Danio rerio*) functional relationship network.

We anticipate that our approach can be extended beyond the six organisms shown in this study, as it is especially beneficial in organisms that have high-throughput genomic data with sparse annotations (e.g. frog, slime mold). Next-generation sequencing is further increasing the diversity of organisms that are measured on the genome-scale, and functional knowledge transfer can help us understand and annotate the roles of genes in such emerging model systems. Functional knowledge transfer allows for accurate hypothesis generation and experiment guidance even for pathways with no previous experimental knowledge in a given organism, thus benefiting human biology, broadly studied organisms such as mouse and fly, and newly adopted model systems.

Materials and Methods

We developed a functional knowledge transfer method and applied this method to predicting gene functions in six organisms using a functional network based classification strategy. In summary, data integration was performed using a regularized naive Bayes classifier, which summarized the data compendium into organism specific function relationship networks. Edges in functional relationship networks represented, given all collected data from that organism, the posterior probability of a gene pair co-functioning in the same biological process. Next, a collection of organism specific experimental annotations supplemented with cross-annotated gene annotations (based on both sequence and functional similarity) was used as gold standard for each GO biological process to train a GO term specific SVM with the

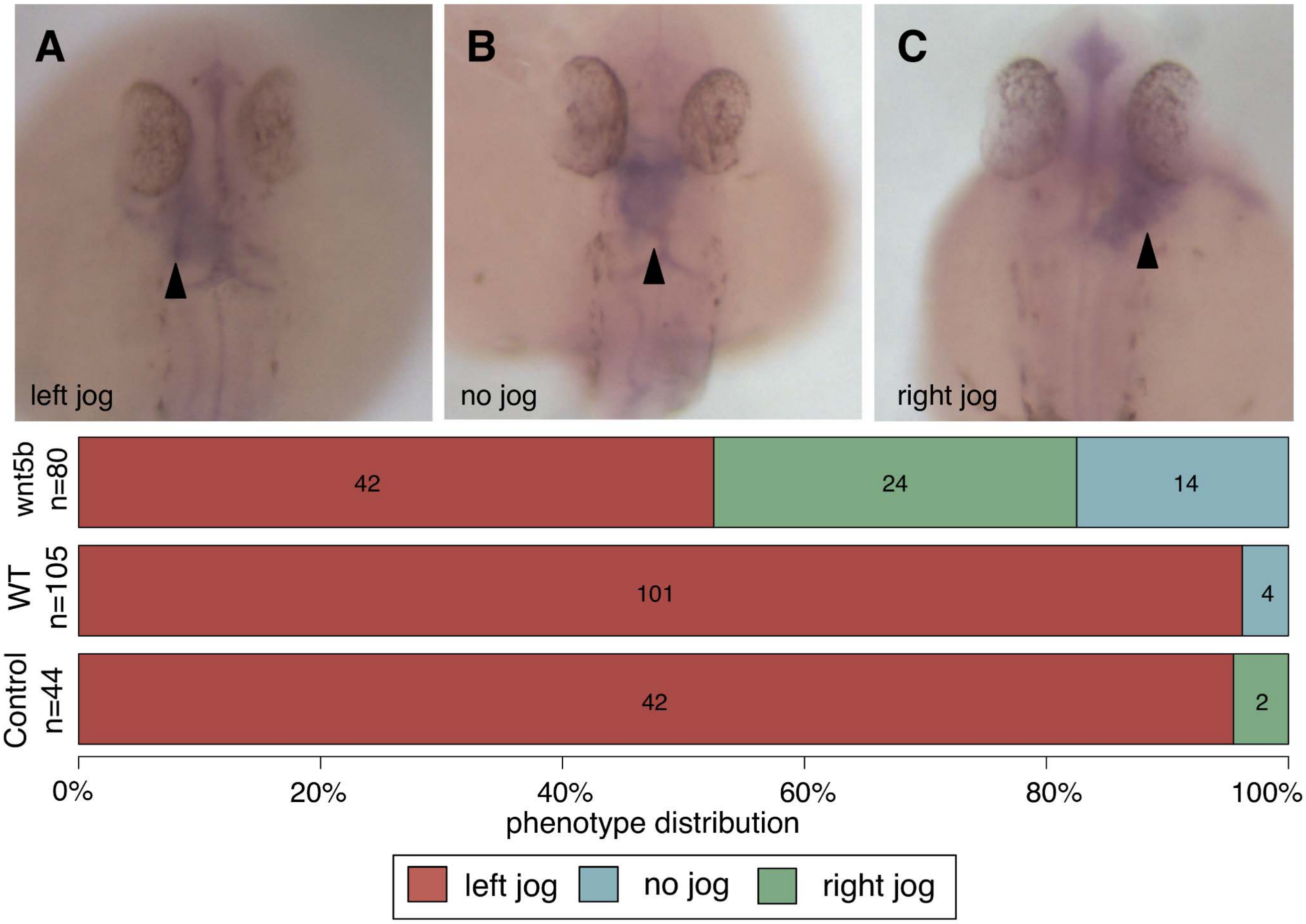


Figure 6. Knockdown of *wnt5b* leads to defects in zebrafish heart asymmetry. Morpholinos (MO) against *wnt5b* were injected into zebrafish embryos at the 1–2 cell stage. Embryos were evaluated for heart jogging at 27 hour post fertilization and scored as either left (C), right (B), or no jog (A). While control MO injected embryos had predominantly left-jogged hearts, embryos injected with the wnt5bMO displayed randomized heart jogging with 48% of embryos displaying right or midline jog.
doi:10.1371/journal.pcbi.1002957.g006

functional relationship network as features. To test for robustness across different machine learning algorithms, L1-regularized logistic regression and Random forest were also evaluated by coupling both algorithms with the functional knowledge transfer method. Final predictions were made on a total of 10,653 unique biological processes. We experimentally validated our method's predictions for the determination of heart left-right asymmetry in *Danio rerio*. Of our top five predictions, four were validated via existing but un-curated experiments from the literature. We validated the fifth, *wnt5b*, using a morpholino knock-down assay.

Integration and summary of organismal data compendia

Data source and pre-processing. We collected 2,444 microarray datasets from NCBI Gene Expression Omnibus (GEO) covering a total of 43,865 conditions across seven model organisms. Probes were collapsed and normalized according to the procedure described in [69] and the Fisher's z-transformed pearson correlation were calculated for each gene-pair as described in [10].

Physical and genetic interaction data from BioGRID [70], IntAct [71], Mint [72], and MIPS [73] were processed as counts of experimental assays that support an interaction between two genes (e.g. a gene pair with evidence from two-hybrid and western blot would receive two counts). Potential transcription factor (TF) to target gene associations were obtained from Yeastract [74] and TF binding site motifs retrieved from Jaspar [75]. Yeastract's predicted TF-gene relations were treated as pair-wise binary scores. For Jaspar, we searched for possible transcription factor binding sites by scanning each TF profile in 1 kb upstream sequence of all genes using FIMO [76]. Motif matches were treated as a binary score (present if p-value<.001 and not-present otherwise) and the final gene pair score was obtained by calculating the pearson correlation between the two genes' binary score vectors.

Phenotype and disease data from SGD [77], MGI [78], Wormbase [79], Flybase [80], GSEA [20], Zfin [81] were incorporated into our functional networks by summing the co-occurrences of gene pairs in all phenotypes/diseases and normalizing by the size of the phenotype/disease. For gene pair, i, j the scoring function is the following:

$$S(i,j) = \sum_{k=1}^{n} \frac{I_k(i) I_k(j)}{N_k}$$

where function $I_k(i)$ and $I_k(j)$ are the indicator functions that have the value 1 when gene i or j is annotated to the phenotype or disease, n is the total number of phenotypes/diseases, and N_k is the total number of genes associated with the phenotype or disease k.

Protein sequence similarity between genes was obtained from Biomart [82], and protein domain data were treated as binary evidence from PfamA [83] and Prosite [84].

Generating functional relationship networks. To summarize the processed heterogeneous genomic data, we generated one global functional relationship network per organism. We applied Bayesian integration, specifically a naïve Bayes classifier, to systematically deal with differences in accuracy and relevance of each data source for predicting gene functional relations. Gene pairs co-annotated to a set of 433 expert selected Gene Ontology [31] biological process fringe terms were used as known functionally related genes (i.e. positive examples) [69,85]. Gene pairs not co-annotated to any terms in the GO fringe, KEGG [86], PID [87] or Biocyc [88] were considered as unrelated (i.e. negative examples) except in the following cases:

1. A gene pair was annotated to terms overlapping with a hypergeometric P-value below 0.05
2. A gene pair was annotated to a set of 'negative' GO terms that define minimal relatedness (as described in [69])

If a gene pair met either of the two conditions, it was excluded from unrelated pair generation (i.e., they were neither related nor unrelated for training). Thus this formed a set of global related and unrelated gene pairs to be used for training and evaluation.

One binary regularized naive Bayes classifier was trained per Gene Ontology fringe term (i.e. biological process/context). Each naive Bayes classifier contains one class node determining the membership of a gene-pair to the biological process and organism specific dataset nodes conditioned on the class node. When integrating large number of genomic datasets, the naive Bayes assumption of conditional independence among datasets can no longer be justified. We have shown that a mutual information based parameter regularization for naive Bayes classifiers can alleviate the conditional dependency among datasets [69]. In this work, we make modifications to our prior method by directly estimating the conditional dependency between a dataset by limiting the mutual information calculation between datasets to gene-pairs that are not functionally related. This heuristic enables us to estimate the conditional dependency between datasets without having to regress out the incomplete functional relation class node information. Specifically, the heuristic sum of shared information U_k is now:

$$U_k = \frac{\sum_{i \neq k} I_{pairs \in negative}(D_k, D_i)}{H(D_k)}$$

$$\alpha_k = 2^{U_k} - 1$$

where $I_{pairs \in negatives}$ is the mutual information between dataset D_k and D_{i} among gene pairs not known to have a functional relationship (i.e. negative gene pair examples) and H is the single dataset entropy. Then we use the exponential decreased ratio (α_{k}) to weight a given dataset's likelihood function. Finally, the naive Bayes functional relationship posterior probability for gene pair i,j is the following:

$$P^*(D_k = d_k(g_i, g_j)|FR=1) = P(D_k = d_k(g_i, g_j)|FR=1) \left(\frac{n_s}{n_s + \alpha_k}\right) + \frac{1}{|D_k|}\left(\frac{\alpha_k}{n_s + \alpha_k}\right)$$

$$P_{g_i, g_j}(FR=1|D) = \frac{P(FR=1) \prod_{k=1}^{n} P^*(D_k = d_k(g_i, g_j)|FR=1)}{P_{g_i, g_j}(D)}$$

where the weighted dataset likelihood function is P^*, $d_k(g_i, g_j)$ is the experimental value for gene pair i,j, $|D_k|$ is the total number of discretization levels and n_s is a pseudocount set to 3 in our integration based on cross-validation results.

Finally, with biological process specific functional relation networks predicted by each naive Bayes classifier, we averaged the edge probabilities from each process specific functional network to generate the final global functional relationship network.

GO biological process gold standard construction through cross-annotation

In total, 10,653 GO biological process terms were predicted for new gene annotations covering six organisms, *Homo sapiens*, *Mus musculus*, *Rattus novegicus*, *Drosophila melanogaster*, *Danio rerio* and *Caenorhabditis elegans*. We limited the positive examples for each GO term to propagated experimental GO annotations with GO evidence codes EXP, IDA, IPI, IMP, IGI and IEP (all "NOT" annotations were removed). In addition, to leverage the research strengths across organisms, we transferred gene annotations among six organisms plus yeast, first based on sequence similarity and second filtered by function similarity. In detail, we start with all sequence paralog and ortholog gene relations within each TreeFam [22] gene family. Next, based on our previous algorithm [25], we filtered for functional analogs among all paralog and ortholog gene pairs using our functional relationship networks. We define a functional analog to be a gene pair that has a significant number of overlapping TreeFam gene families among its closest gene neighbors in the global functional relationship network (a functional network is converted into a binary network by using a probability cutoff of 0.5). We defined a gene pair's score as the following:

$$S_{G1,G2} = \sum_{i=k}^{\min(m,n)} \frac{\binom{m}{i}\binom{N-m}{n-i}}{\binom{N}{n}}$$

where m and n are the number of TreeFam gene families in each gene $G1$ and $G2$'s direct neighborhood in the functional network, k is the number of overlapping TreeFam gene families between gene $G1$ and $G2$ gene neighbors and N is the total number of TreeFam gene families around gene $G1$ and $G2$. The functional similarity score is the probability of observing greater or equal to the number of overlapping gene families by chance, thus can be interpreted as a hypergeometric p-value. We used a score cutoff of $<=0.01$ to consider a gene pair as functional analogs.

Finally, all experimental annotations are propagated between functional analogs. In total, our supervised functional knowledge transfer allowed us to make predictions for 8,091 additional GO biological processes, thus extending our predictions beyond simply well-studied and well annotated processes and pathways.

Biological process prediction with network based SVM

We used the augmented gold standard genes by functional knowledge transfer and functional relation network as features into state-of-art machine learning algorithm Support Vector Machine (SVM) to predict novel biological process gene annotations. Our functional relation network based SVM method has shown to outperform methods that directly input the raw data into the SVM or a simplistic sum of the functional networks to the positive examples [89].

For each biological process, the feature space was constructed as the weights in the functional relation network. Thus for each gene example, all gene edge weights connecting to the example gene were used to create the feature vector. Therefore, each organism feature count will be equal to the number of genes in the organism. The set of feature vectors for training examples were used to train a linear SVM according to the standard formulation:

$$\min_{w,\xi_i \geq 0} \frac{1}{2} w^T w + \frac{C}{n} \sum_{i=1}^{n} \xi_i$$

$$\forall i : y_i \left(w^T x_i \right) \geq 1 - \xi_i$$

where n is the of training example genes, w is the gene weight vector, y_i is the training label of gene i and x_i is the edge weight vector connecting gene i to all genes in the functional network.

Finally, the unbounded SVM prediction scores were transformed into probabilities based on a maximum likelihood sigmoid fit to the SVM outputs [90].

Additional machine learning algorithms

To validate that the observed performance improvement was not specific to any single learning algorithm, we applied the functional knowledge transfer to two additional widely used machine learning methods: L1-regularized logistic regression and Random forest. Regression analysis coupled with regularization has been a broadly used approach to control for the bias-variance trade-off [91]. In particular, L1-regularization has been successfully used in many methods for shrinkage and feature selection applications, most famously in the works of LASSO [92]. By coupling L1-regularization with a logit link function, conditional probabilities of a gene membership to a biological process can be computed based on selected genes of predictive power. The predictive gene weight vector w was obtained by the following regression problem:

$$\arg\min_{w} \sum_{i=1}^{m} \log\left(1 + e^{-y_i w^T x_i}\right) + \lambda \sum_{i=1}^{n} |w_i|$$

where $\lambda > 0$ is the regularization parameter, y_i is the training label of gene i and x_i is the edge weight vector connecting gene i to all genes in the functional network.

Random forest classifiers are a combination of decision trees that are aggregated to make a final prediction. Random forest algorithms have been shown to produce improved prediction accuracy compared to a single decision tree by better estimating the contribution of each predictor through random sampling [29]. In genomic applications, Random forest has gained interest due to the many high-dimensional genomic learning problems [93]. Formally, random forest is defined by the following:

$$RF = \{h(X,d_i), i = 1, \ldots, n\}$$

where the random forest RF is a set of $h(\cdot)$ decision tree functions, trained on training examples X and a bootstrap sample d_i from the original feature space of D. For classification, the votes of each n decision trees are averaged as shown in the following:

$$av_n \sum_{i=1}^{n} I(h_i(X))$$

where $I(\cdot)$ is the indicator function for the prediction class of interest. In our study, for each GO term 61 decision trees were trained on independent bootstrap samples of our original genomic training data.

Performance evaluation

For performance evaluations for GO terms with no prior annotation, we used a three-year temporal holdout set of gene annotations for each GO biological process (one-year holdout results shown in supplemental material). The held-out gene annotations were preserved throughout the prediction pipeline (functional network integration and SVM predictions) to avoid any overestimation of performance. Although we train our SVM classifiers using the augmented cross-annotated gold standard, only the non-transferred experimental GO term annotations were used for evaluation with all transferred annotations excluded.

The GO gene association files used to create our gold standard was downloaded from Gene Ontology [31] on 5/11/2011 (all dates in mm/dd/yyyy format). To generate an accurate temporal three-year holdout we downloaded the GO gene association version archived at 5/11/2008. All annotations were propagated and only experimental examples newly annotated after 5/11/2008 for each GO term was used in the temporal evaluation. Accordingly, any GO term that had no gene annotations on 5/11/2008, but subsequently accumulated new annotations were used to evaluate our performance in predicting terms with no-known prior annotations.

To compare performance between knowledge transfer methods, we conducted an evaluation by performing a threefold cross-validation among genes that had experimental biological process annotations. This set of evaluation annotations represents a random sampling of the current knowledge base as of 5/11/2011. Identical to our temporal holdout, all evaluation annotations for each holdout were withheld from our prediction pipeline to avoid any performance over-estimation.

Implementation

All software used in this study has been implemented in the open source and publicly available Sleipnir library [94] available from http://libsleipnir.bitbucket.org, which interfaces with the SVMperf library [95] for linear kernel SVM classifiers (the error parameter C was set to 100 for these experiments through cross-validation). L1-regularized logistic regression used the LIBLINEAR [28] and Random forest used the MILK (Machine Learning Toolkit) python package implementation with 61 decision trees per GO term.

Morpholino microinjections and scoring of heart left-right asymmetry

The *wnt5b* morpholino (MO) and standard control MO were purchased from GeneTools. The sequence of the *wnt5b* MO used is as follows: 5′-GTCCTTGGTTCATTCTCACATCCAT-3′. Morpholinos were injected at a concentration of 6 ng/uL into the yolk of one-cell stage embryos for whole knockdown in the embryonic cells. Initial assessment (Figure 6) was performed via *in situ* hybridizations on fixed embryos using the standard protocol [96] with *cmlc2/myl7* used as a probe. Images were captured at 4× or 10× magnification using a ProgressC14 digital camera (Jenoptik) on a Leica MZFLIII microscope.

Heart laterality for each treatment (*wnt5b* MO, control MO, wild type) was evaluated in live Tg(*cmlc2::GFP*) embryos at 27 hours post fertilization. Embryos were scored as left/right/no jog based on expression of GFP driven by *cmlc2*'s heart specific promoter using a Leica SP5 confocal microscope (Figure S4).

Supporting Information

Figure S1 FKT cross annotation allows accurate recovery of small and unannotated terms in 1 year temporal holdout (pink: FKT+SVM, gray: SVM). All annotations accumulated after 5/11/2010 are held out from our prediction pipeline (as outlined in Figure 2) and are used to evaluate the predictive power of FKT derived cross annotations (3 year temporal holdout is shown in main text figure 3). GO biological processes terms that acquired new annotations subsequent to our holdout date are grouped by organism and by the number of annotations at 5/11/2010 (zero, < = 5, < = 15, all). Performances at recapitulating future annotations are compared for a machine learning method (SVM) without (gray) and with (pink) learning on functional knowledge transfer (FKT) derived examples. For processes with zero annotations before 5/11/2010, no predictions can be made without cross-annotation (shown as absent performance bar).
(PDF)

Figure S2 The categorization of newly predicted biological processes. In total 8,091 GO biological processes without prior experimental annotation were predicted for novel gene-pathway membership by deploying FKT across our six metazoan organisms (*Homo sapiens*, *Mus musculus*, *Rattus novegicus*, *Drosophila melanogaster*, *Danio rerio* and *Caenorhabditis elegans*). Here we show the nature of these newly predicted biological process terms grouped by each process' parent term in the gene ontology (1 level in the biological process name space).
(PDF)

Figure S3 Specificity of newly predicted biological processes. Here we plot the specificity of 8,091 newly predicted GO biological processes without prior experimental annotations. As an imperfect proxy for biological specificity we use the depth of each process term in the gene ontology biological process name space. As examples of terms for a given depth, depth 2 leukocyte proliferation, depth 4 glomerulus vasculature development, depth 6 intermediate filament cytoskeleton organization, depth 8 purine ribonucleotide biosynthetic process, depth 10 regulation of insulin secretion involved in cellular response to glucose stimulus, depth 12 negative regulation of histone h3 k9 methylation.
(PDF)

Figure S4 Zebrafish wnt5b knockdown in live embryos show significant deviation from wild type heart jogging. Heart laterality for each treatment (wnt5b MO, control MO, wild type) was evaluated in live embryos at 27 hours post fertilization. Embryos were scored as left (C), right (B), or no jog (A) based on the expression of GFP driven by cmlc2's heart specific promoter. In total, 48% of morpholino treated embryos showed either right-sided heart jog or midline/no jog. Only 4% of wild type and control-MO treated embryos exhibited this phenotype. *In situ* results are shown in Figure 6 in the manuscript.
(PDF)

Text S1 GO terms relevant in mammals (mouse, human, rat) but missing in at least one organism.
(TXT)

Text S2 GO terms with no experimental annotations but gene prediction enabled by FKT.
(TXT)

Text S3 All GO terms prediction evaluation results for temporal and random holdout.
(TXT)

Acknowledgments

The authors acknowledge John Wiggins for technical support.

Author Contributions

Conceived and designed the experiments: CYP AKW CSG JR YG RDB OGT. Performed the experiments: CYP AKW CSG JR. Analyzed the data: CYP AKW CSG. Contributed reagents/materials/analysis tools: JR LAB RDB. Wrote the paper: CYP AKW CSG OGT.

References

1. Guan Y, Myers C, Hess D, Barutcuoglu Z, Caudy A, et al. (2008) Predicting gene function in a hierarchical context with an ensemble of classifiers. Genome Biology 9: S3.
2. Mostafavi S, Ray D, Warde-Farley D, Grouios C, Morris Q (2008) GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function. Genome Biology 9: S4.
3. Walker MG, Volkmuth W, Sprinzak E, Hodgson D, Klingler T (1999) Prediction of Gene Function by Genome-Scale Expression Analysis: Prostate Cancer-Associated Genes. Genome Research 9: 1198–1203.
4. Wu LF, Hughes TR, Davierwala AP, Robinson MD, Stoughton R, et al. (2002) Large-scale prediction of Saccharomyces cerevisiae gene function using overlapping transcriptional clusters. Nat Genet 31: 255–265.
5. Ye P, Peyser BD, Pan X, Boeke JD, Spencer FA, et al. (2005) Gene function prediction from congruent synthetic lethal interactions in yeast. Mol Syst Biol 1:2005.0026.
6. Kim W, Krumpelman C, Marcotte E (2008) Inferring mouse gene functions from genomic-scale data using a combined functional network/classification strategy. Genome Biology 9: S5.
7. Pena-Castillo L, Tasan M, Myers C, Lee H, Joshi T, et al. (2008) A critical assessment of Mus musculus gene function prediction using integrated genomic evidence. Genome Biology 9: S2.
8. Pavlidis P, Weston J, Cai J, Grundy WN (2001) Gene functional classification from heterogeneous data. Proceedings of the fifth annual international conference on Computational biology. Montreal, Quebec, Canada: ACM. pp. 249–255.
9. Myers CL, Robson D, Wible A, Hibbs MA, Chiriac C, et al. (2005) Discovery of biological networks from diverse functional genomic data. Genome Biol 6: R114.
10. Mostafavi S, Morris Q (2010) Fast integration of heterogeneous data sources for predicting gene function with limited annotation. Bioinformatics 26: 1759–1765.
11. Greene CS, Troyanskaya OG (2011) PILGRM: an interactive data-driven discovery platform for expert biologists. Nucleic Acids Research 39: W368–W374.
12. Hughes TR, Marton MJ, Jones AR, Roberts CJ, Stoughton R, et al. (2000) Functional Discovery via a Compendium of Expression Profiles. Cell 102: 109–126.
13. Marcotte EM, Pellegrini M, Thompson MJ, Yeates TO, Eisenberg D (1999) A combined algorithm for genome-wide prediction of protein function. Nature 402: 83–86.
14. Jensen LJ, Gupta R, Blom N, Devos D, Tamames J, et al. (2002) Prediction of Human Protein Function from Post-translational Modifications and Localization Features. Journal of Molecular Biology 319: 1257–1265.
15. Barutcuoglu Z, Schapire RE, Troyanskaya OG (2006) Hierarchical multi-label prediction of gene function. Bioinformatics 22: 830–836.
16. Vazquez A, Flammini A, Maritan A, Vespignani A (2003) Global protein function prediction from protein-protein interaction networks. Nat Biotech 21: 697–700.
17. Hess DC, Myers CL, Huttenhower C, Hibbs MA, Hayes AP, et al. (2009) Computationally Driven, Quantitative Experiments Discover Genes Required for Mitochondrial Biogenesis. PLoS Genet 5: e1000407.
18. Lee I, Ambaru B, Thakkar P, Marcotte EM, Rhee SY (2010) Rational association of genes with traits using a genome-scale gene network for Arabidopsis thaliana. Nat Biotech 28: 149–156.
19. Eisen MB, Spellman PT, Brown PO, Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Sciences 95: 14863–14868.

20. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, et al. (2005) Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences of the United States of America 102: 15545–15550.

21. Eisen JA (1998) Phylogenomics: Improving Functional Predictions for Uncharacterized Genes by Evolutionary Analysis. Genome Research 8: 163–167.

22. Li H, Coghlan A, Ruan J, Coin LJ, Heriche JK, et al. (2006) TreeFam: a curated database of phylogenetic trees of animal gene families. Nucleic Acids Res 34: D572–580.

23. O'Brien KP, Remm M, Sonnhammer ELL (2005) Inparanoid: a comprehensive database of eukaryotic orthologs. Nucleic Acids Research 33: D476–D480.

24. Valenzuela DM, Griffiths JA, Rojas J, Aldrich TH, Jones PF, et al. (1999) Angiopoietins 3 and 4: Diverging gene counterparts in mice and humans. Proceedings of the National Academy of Sciences 96: 1904–1909.

25. Chikina MD, Troyanskaya OG (2011) Accurate quantification of functional analogy among close homologs. PLoS Comput Biol 7: e1001074.

26. Wong AK, Park CY, Greene CS, Bongo LA, Guan Y, et al. (2012) IMP: A multi-species functional genomics portal for integration, visualization and prediction of protein functions and networks. Nucleic Acids Research 40: W484–90.

27. Noble WS (2006) What is a support vector machine? Nat Biotech 24: 1565–1567.

28. Fan R-E, Chang K-W, Hsieh C-J, Wang X-R, Lin C-J (2008) LIBLINEAR: A Library for Large Linear Classification. J Mach Learn Res 9: 1871–1874.

29. Breiman L (2001) Random Forests. Machine Learning 45: 5–32.

30. Hwang S, Rhee SY, Marcotte EM, Lee I (2011) Systematic prediction of gene function in Arabidopsis thaliana using a probabilistic functional gene network. Nat Protocols 6: 1429–1442.

31. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25: 25–29.

32. Hofken T, Schiebel E (2002) A role for cell polarity proteins in mitotic exit. EMBO J 21: 4851–4862.

33. Matei V, Pauley S, Kaing S, Rowitch D, Beisel KW, et al. (2005) Smaller inner ear sensory epithelia in Neurog1 null mice are related to earlier hair cell cycle exit. Developmental Dynamics 234: 633–650.

34. Garner M, van Kreeveld S, Su TT (2001) mei-41 and bub1 block mitosis at two distinct steps in response to incomplete DNA replication in Drosophila embryos. Current Biology 11: 1595–1599.

35. Yamaguchi M, Imai F, Tonou-Fujimori N, Masai I (2010) Mutations in N-cadherin and a Stardust homolog, Nagie oko, affect cell-cycle exit in zebrafish retina. Mechanisms of Development 127: 247–264.

36. Hartwell L, Weinert T (1989) Checkpoints: controls that ensure the order of cell cycle events. Science 246: 629–634.

37. Gasch AP, Spellman PT, Kao CM, Carmel-Harel O, Eisen MB, et al. (2000) Genomic Expression Programs in the Response of Yeast Cells to Environmental Changes. Mol Biol Cell 11: 4241–4257.

38. Kuhar SG, Feng L, Vidan S, Ross ME, Hatten ME, et al. (1993) Changing patterns of gene expression define four stages of cerebellar granule neuron differentiation. Development 117: 97–104.

39. Furlong EEM, Andersen EC, Null B, White KP, Scott MP (2001) Patterns of Gene Expression During Drosophila Mesoderm Development. Science 293: 1629–1633.

40. Arlotta P, Molyneaux BJ, Chen J, Inoue J, Kominami R, et al. (2005) Neuronal Subtype-Specific Genes that Control Corticospinal Motor Neuron Development In Vivo. Neuron 45: 207–221.

41. Liu L, Ji C, Chen J, Li Y, Fu X, et al. (2008) A global genomic view of MIF knockdown-mediated cell cycle arrest. Cell Cycle 7: 1678–1692.

42. Mackley JR, Ando J, Herzyk P, Winder SJ (2006) Phenotypic responses to mechanical stress in fibroblasts from tendon, cornea and skin. Biochemical Journal 396: 307–316.

43. Hunt-Newbury R, Viveiros R, Johnsen R, Mah A, Anastas D, et al. (2007) High-Throughput In Vivo Analysis of Gene Expression in *Caenorhabditis elegans*. PLoS Biol 5: e237.

44. Hovatta I, Tennant RS, Helton R, Marr RA, Singer O, et al. (2005) Glyoxalase 1 and glutathione reductase 1 regulate anxiety in mice. Nature 438: 662–666.

45. Carter T, Greenhall J, Yoshida S, Fuchs S, Helton R, et al. (2005) Mechanisms of aging in senescence-accelerated mice. Genome Biology 6: R48.

46. Ramsdell AF (2005) Left–right asymmetry and congenital cardiac defects: Getting to the heart of the matter in vertebrate left–right axis determination. Developmental Biology 288: 1–20.

47. van der Linde D, Konings EEM, Slager MA, Witsenburg M, Helbing WA, et al. (2011) Birth Prevalence of Congenital Heart Disease Worldwide: A Systematic Review and Meta-Analysis. Journal of the American College of Cardiology 58: 2241–2247.

48. Baker K, Holtzman NG, Burdine RD (2008) Direct and indirect roles for Nodal signaling in two axis conversions during asymmetric morphogenesis of the zebrafish heart. Proceedings of the National Academy of Sciences 105: 13924–13929.

49. Smith KA, Chocron S, von der Hardt S, de Pater E, Soufan A, et al. (2008) Rotation and Asymmetric Development of the Zebrafish Heart Requires Directed Migration of Cardiac Progenitor Cells. Developmental Cell 14: 287–297.

50. Rohr S, Otten C, Abdelilah-Seyfried S (2008) Asymmetric Involution of the Myocardial Field Drives Heart Tube Formation in Zebrafish. Circulation Research 102: e12–e19.

51. de Campos-Baptista MIM, Holtzman NG, Yelon D, Schier AF (2008) Nodal signaling promotes the speed and directional movement of cardiomyocytes in zebrafish. Developmental Dynamics 237: 3624–3633.

52. Wang X, Yost HJ (2008) Initiation and propagation of posterior to anterior (PA) waves in zebrafish left–right development. Developmental Dynamics 237: 3640–3647.

53. Liang JO, Etheridge A, Hantsoo L, Rubinstein AL, Nowak SJ, et al. (2000) Asymmetric nodal signaling in the zebrafish diencephalon positions the pineal organ. Development 127: 5101–5112.

54. Essner JJ, Amack JD, Nyholm MK, Harris EB, Yost HJ (2005) Kupffer's vesicle is a ciliated organ of asymmetry in the zebrafish embryo that initiates left-right development of the brain, heart and gut. Development 132: 1247–1260.

55. Alexander J, Rothenberg M, Henry GL, Stainier DYR (1999) casanova Plays an Early and Essential Role in Endoderm Formation in Zebrafish. Developmental Biology 215: 343–357.

56. Rebagliati MR, Toyama R, Fricke C, Haffter P, Dawid IB (1998) Zebrafish Nodal-Related Genes Are Implicated in Axial Patterning and Establishing Left–Right Asymmetry. Developmental Biology 199: 261–272.

57. Hami D, Grimes AC, Tsai H-J, Kirby ML (2011) Zebrafish cardiac development requires a conserved secondary heart field. Development 138: 2389–2398.

58. Feldman B, Concha ML, Saúde L, Parsons MJ, Adams RJ, et al. (2002) Lefty Antagonism of Squint Is Essential for Normal Gastrulation. Current Biology 12: 2129–2135.

59. Lenhart KF, Lin S-Y, Titus TA, Postlethwait JH, Burdine RD (2011) Two additional midline barriers function with midline lefty1 expression to maintain asymmetric Nodal signaling during left-right axis specification in zebrafish. Development 138: 4405–4410.

60. Smith KA, Noël E, Thurlings I, Rehmann H, Chocron S, et al. (2011) Bmp and Nodal Independently Regulate *lefty1* Expression to Maintain Unilateral Nodal Activity during Left-Right Axis Specification in Zebrafish. PLoS Genet 7: e1002289.

61. Goudevenou K, Martin P, Yeh Y-J, Jones P, Sablitzky F (2011) Def6 Is Required for Convergent Extension Movements during Zebrafish Gastrulation Downstream of Wnt5b Signaling. PLoS ONE 6: e26548.

62. Corey D, Abrams J (2001) Morpholino antisense oligonucleotides: tools for investigating vertebrate development. Genome Biology 2: reviews1015.1011–reviews1015.1013.

63. Lopes CAM, Prosser SL, Romio L, Hirst RA, O'Callaghan C, et al. (2011) Centriolar satellites are assembly points for proteins implicated in human ciliopathies, including oral-facial-digital syndrome 1. Journal of Cell Science 124: 600–612.

64. Glazer AM, Wilkinson AW, Backer CB, Lapan SW, Gutzman JH, et al. (2010) The Zn Finger protein Iguana impacts Hedgehog signaling by promoting ciliogenesis. Developmental Biology 337: 148–156.

65. Aamar E, Dawid IB (2010) Sox17 and chordin are required for formation of Kupffer's vesicle and left-right asymmetry determination in zebrafish. Developmental Dynamics 239: 2980–2988.

66. Troyanskaya OG, Dolinski K, Owen AB, Altman RB, Botstein D (2003) A Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae). Proceedings of the National Academy of Sciences 100: 8348–8353.

67. Lee I, Date SV, Adai AT, Marcotte EM (2004) A Probabilistic Functional Network of Yeast Genes. Science 306: 1555–1558.

68. Rhodes DR, Tomlins SA, Varambally S, Mahavisno V, Barrette T, et al. (2005) Probabilistic model of the human protein-protein interaction network. Nat Biotech 23: 951–959.

69. Huttenhower C, Haley EM, Hibbs MA, Dumeaux V, Barrett DR, et al. (2009) Exploring the human genome with functional maps. Genome Research 19: 1093–1106.

70. Stark C, Breitkreutz BJ, Chatr-Aryamontri A, Boucher L, Oughtred R, et al. (2011) The BioGRID Interaction Database: 2011 update. Nucleic Acids Res 39: D698–704.

71. Kerrien S, Aranda B, Breuza L, Bridge A, Broackes-Carter F, et al. (2012) The IntAct molecular interaction database in 2012. Nucleic Acids Res 40: D841–846.

72. Licata L, Briganti L, Peluso D, Perfetto L, Iannuccelli M, et al. (2012) MINT, the molecular interaction database: 2012 update. Nucleic Acids Res 40: D857–861.

73. Mewes HW, Frishman D, Gruber C, Geier B, Haase D, et al. (2000) MIPS: a database for genomes and protein sequences. Nucleic Acids Research 28: 37–40.

74. Abdulrehman D, Monteiro PT, Teixeira MC, Mira NP, Lourenço AB, et al. (2011) YEASTRACT: providing a programmatic access to curated transcriptional regulatory associations in Saccharomyces cerevisiae through a web services interface. Nucleic Acids Research 39: D136–D140.

75. Sandelin A, Alkema W, Engström P, Wasserman WW, Lenhard B (2004) JASPAR: an open access database for eukaryotic transcription factor binding profiles. Nucleic Acids Research 32: D91–D94.

76. Grant CE, Bailey TL, Noble WS (2011) FIMO: scanning for occurrences of a given motif. Bioinformatics 27: 1017–1018.

77. Cherry JM, Adler C, Ball C, Chervitz SA, Dwight SS, et al. (1998) SGD: Saccharomyces Genome Database. Nucleic Acids Research 26: 73–79.
78. Bult CJ, Eppig JT, Kadin JA, Richardson JE, Blake JA, et al. (2008) The Mouse Genome Database (MGD): mouse biology and model systems. Nucleic Acids Research 36: D724–D728.
79. Stein L, Sternberg P, Durbin R, Thierry-Mieg J, Spieth J (2001) WormBase: network access to the genome and biology of Caenorhabditis elegans. Nucleic Acids Research 29: 82–86.
80. Drysdale RA, Crosby MA, Consortium TF (2005) FlyBase: genes and gene models. Nucleic Acids Research 33: D390–D395.
81. Sprague J, Clements D, Conlin T, Edwards P, Frazer K, et al. (2003) The Zebrafish Information Network (ZFIN): the zebrafish model organism database. Nucleic Acids Research 31: 241–243.
82. Kasprzyk A (2011) BioMart: driving a paradigm change in biological data management. Database 2011: bar049
83. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, et al. (2004) The Pfam protein families database. Nucleic Acids Research 32: D138–D141.
84. Hulo N, Bairoch A, Bulliard V, Cerutti L, De Castro E, et al. (2006) The PROSITE database. Nucleic Acids Research 34: D227–D230.
85. Myers C, Barrett D, Hibbs M, Huttenhower C, Troyanskaya O (2006) Finding function: evaluation methods for functional genomic data. BMC Genomics 7: 187.
86. Kotera M, Hirakawa M, Tokimatsu T, Goto S, Kanehisa M (2012) The KEGG databases and tools facilitating omics analysis: latest developments involving human diseases and pharmaceuticals. Methods Mol Biol 802: 19–39.
87. Schaefer CF, Anthony K, Krupa S, Buchoff J, Day M, et al. (2009) PID: the Pathway Interaction Database. Nucleic Acids Res 37: D674–679.
88. Caspi R, Altman T, Dreher K, Fulcher CA, Subhraveti P, et al. (2012) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Res 40: D742–753.
89. Guan Y, Ackert-Bicknell CL, Kell B, Troyanskaya OG, Hibbs MA (2010) Functional genomics complements quantitative genetics in identifying disease-gene associations. PLoS Comput Biol 6: e1000991.
90. John CP (1999) Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. MIT Press.
91. Hoerl AE, Kennard RW (2000) Ridge regression: biased estimation for nonorthogonal problems. Technometrics 42: 80–86.
92. Tibshirani R (1994) Regression shrinkage and selection via the lasso. J R Statist Soc B 73: 273–282
93. Diaz-Uriarte R, Alvarez de Andres S (2006) Gene selection and classification of microarray data using random forest. BMC Bioinformatics 7: 3.
94. Huttenhower C, Schroeder M, Chikina MD, Troyanskaya OG (2008) The Sleipnir library for computational functional genomics. Bioinformatics 24: 1559–1561.
95. Joachims T (2006) Training linear SVMs in linear time. Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. Philadelphia, PA, , USA: ACM. pp. 217–226.
96. Huang C-J, Tu C-T, Hsiao C-D, Hsieh F-J, Tsai H-J (2003) Germ-line transmission of a myocardium-specific GFP transgene reveals critical regulatory elements in the cardiac myosin light chain 2 promoter of zebrafish. Developmental Dynamics 228: 30–40.