Supporting Information

McGary et al. 10.1073/pnas.0910200107

SI Materials and Methods

Collection of Phenotypes. We collected gene-phenotype associations from the literature for five species (worm, yeast, mouse, human, *Arabidopsis*).

For human phenotypes, we used employed human diseases from the Online Mendelian Inheritance in Man (OMIM) database (1), using the compressed OMIM disease categories previously described in McGary et al. (2), such that multiple variants of a disease were grouped together. (For example "Corneal dystrophy, hereditary polymorphous posterior" and "Corneal dystrophy, lattice type I," reduce to a single category of corneal dystrophies).

Mouse gene-phenotype associations were downloaded from Mouse Genome Informatics (MGI) (3) (MGI_PhenoGenoMP.rpt; downloaded on April 21, 2008). Gene-phenotype associations involving more than one locus or that could not be linked to an Entrez Gene were removed. MGI identifiers were converted to Entrez GeneIDs using MGI_Coordinate.rpt (downloaded April 25, 2008). MGI mouse phenotype descriptions were from VOC_MammalianPhenotype.rpt, downloaded May 7, 2008. All MGI data were downloaded from ftp://ftp.informatics.jax.org/pub/reports/index.html. MGI associations were supplemented with a small number of broadly defined mouse phenotypes, http://hugheslab.med.utoronto.ca/supplementary-data/mouseFunc_I/MGI_phenotype.txt, but which are ultimately derived from MGI data.

Worm gene-phenotype associations were assembled from the literature-reported RNAi studies assembled in Lee et al. (4) supplemented by the additional phenotype data from WormBase 188 (5) (ftp://ftp.wormbase.org/pub/wormbase/acedb/WS188/). Worm gene-phenotype association data come from phenotype_association.WS188.wb, phenotype descriptions from phenotype_ontology.WS188.obo, and gene information from geneIDs.WS188.gz, accessed March 26, 2008. Wormbase phenotypes were filtered for positive associations only. All allelic variants and RNAi data were reduced to gene-phenotype pairs. Gene IDs (e.g., WBGene00044645) were translated to sequence names (e.g., Y51H7BR.8) using geneIDs.WS188.gz. Of ≈22K gene-phenotype pairs, 384 could not be linked to a sequence name. These derived primarily from uncloned genes and were thus omitted from further analysis.

Yeast gene-phenotype associations were obtained from McGary et al. (2) [a literature compilation plus *Saccharomyces* Genome Database (SGD) (6)], supplemented with associations from a recent set of genome-wide screens of drug sensitivity (7) (homozygous and heterozygous screens, het.z_tdist_pval_nm.goodbatch.pub, hom.z_tdist_pval_nm.pub, downloaded from http://chemogenomics.stanford.edu/supplements/global/download/data/). All gene-phenotype associations from the drug screens were filtered using the authors' recommended cutoff of $P < 1 \times 10^{-5}$.

Arabidopsis gene-phenotype associations were downloaded from the *Arabidopsis* Information Resource (TAIR) (8) (ftp://ftp.arabidopsis.org/home/tair/Ontologies/Gene_Ontology/ATH_GO_GOSLIM.txt) on December 9, 2008. Most gene ontology (GO) terms are not phenotypes, so only GO biological processes were retained. A mapping between symbol and locus was obtained from TAIR (ftp://ftp.arabidopsis.org/home/tair/Genes/gene_aliases.20080716) on the same date. Other symbol-locus mappings and a set of descriptions for the genes were extracted from the proteome file (see below). This mapping was used to convert from symbols to loci in the gene-phenotype association list. The gene-description mapping was enhanced by inclusion of alternate gene symbols and names. Phenotype pairs whose sets of associated genes overlapped by greater than 90% were combined, provided that the phenotypes each had more than one associated gene.

For the purposes of calculating phenologs from mouse, worm, yeast, and humans, we considered only a subset of the gene-phenotype associations plotted in Fig. 1*A*, analyzing only those implicating single genes (i.e., not genetic interactions or traits requiring simultaneous mutation of multiple loci), and only those phenotypes in which a defect was observed (i.e., omitting genes associated with the phenotype "normal," "wild-type," "no effect," or other such cases.) All gene-phenotype sets are available from the supporting web site (http://www.phenologs.org).

Identification of Nonredundant Phenotype Sets. To minimize the number of redundant comparisons performed, all phenotype-associated gene sets within a single organism were tested for significant overlap and nonredundant sets were selected for subsequent analyses. Within each organism, phenotypes were identified that reciprocally covered ≥ 80% of each other's genes; for each such pair of phenotypes, only the phenotype with the greater number of genes was retained. (For example, in mouse, genes associated with defects in the small petrosal ganglion and small nodose ganglion overlap considerably. The former has nine associated genes, of which a subset of eight is also associated with the latter phenotype; only the former was retained.)

Orthologs. ***Proteomes.*** Orthologs between species were calculated using the following translated genomes:

Human, ftp://ftp.ncbi.nih.gov/genomes/H_sapiens/protein/protein.fa.gz, downloaded Feb. 7, 2008.
Mouse, ftp://ftp.ncbi.nih.gov/genomes/M_musculus/protein/protein.fa.gz, downloaded Oct. 13, 2007.
Worm, ftp://ftp.wormbase.org/pub/wormbase/data_freezes/WS170/sequences/wormpep170.tar.gz, downloaded Feb. 19, 2007.
Yeast, ftp://genome-ftp.stanford.edu/pub/yeast/sequence/genomic_sequence/orf_protein/orf_trans.fasta.gz, downloaded Feb. 19, 2007.
Arabidopsis, ftp://ftp.arabidopsis.org/home/tair/Sequences/blast_datasets/TAIR8_blastsets/TAIR8_pep_20080412, downloaded Dec. 10, 2008.

For human and mouse proteomes, we analyzed only sequences with protein refseq identifiers (NP_ only). For humans, 43 genes without Gene IDs were removed (mostly hypothetical proteins). For mouse, three proteins without current records were removed.
INPARANOID calculation. To identify orthologous genes in different species, orthologs were calculated using INPARANOID v. 1.35 (9) and blastall 2.2.15, both with default parameters. All genes assigned as orthologs (strictly speaking, ortholog groups or orthogroups because of inclusion of in-paralogs) by INPARANOID were kept regardless of their INPARANOID score. Using orthogroups, rather than bidirectional best hits, captures the many-to-many relationships that exist for gene duplicates that exist in more than one copy in one or both species. To prevent isoform variations from resulting in skewed blast results, mouse and human sequences with the same Entrez GeneID but separate RefSeqIDs were treated separately in INPARANOID. Following INPARANOID analysis, orthologs sharing GeneIDs were combined so that gene variants would be considered together in subsequent analyses.

Calculation of Phenologs. For each pair of species, we first converted gene-phenotype associations to ortholog-phenotype associations using the orthologs calculated by INPARANOID. In cases where paralogous genes within an organism result in the same phenotype, multiple gene-phenotype associations thus collapse to a single ortholog-phenotype association, which eliminates artificial inflation of the significance of ortholog overlap. Second, we compared the set of orthologs associated with a given phenotype in one species (species 1) to the set of orthologs associated with a given phenotype in the second species (species 2), repeating this analysis for all pair-wise comparisons of phenotypes from species 1 and species 2. For each pair of phenotypes in which the ortholog sets overlapped (shared members), we calculated the probability of the overlap because of chance using the cumulative hypergeometric distribution, where N is the total number of orthologs shared between the two species; n and m are the number of orthologs linked to the species 1 and species 2 phenotypes, respectively; and c is the number of common orthologs (that is, those linked to both phenotypes). The probability is given by:

$$\sum_{\kappa=c}^{\min(m,n)} \frac{\binom{m}{\kappa}\binom{N-m}{n-\kappa}}{\binom{N}{n}}$$

The hypergeometric probability does not correct for multiple comparisons, so we estimated the false-discovery rate (FDR) with an empirical permutation test. We performed 1,000 random permutations of the ortholog-phenotype associations, for each permutation repeating the all versus all phenotype comparison using ortholog set sizes identical to those associated with the actual phenotypes (i.e., shuffling ortholog identities on a per phenotype basis, thus maintaining the phenotype set size distribution). Significant phenologs were identified at a FDR of 0.05 by ranking real and permuted phenologs on the basis of the associated hypergeometric probabilities and selecting a threshold of probability where the proportion of permuted phenologs above the cutoff accounted for 5% of the phenologs.

Cross-Validated Prediction of Disease Genes. For the set of human genetic diseases, we predicted specific genes associated with each disease using 10-fold cross-validation, evaluating performance by standard receiver-operater characteristic (ROC) analysis (Fig. S2). These tests employed an alternate formalism from that described above to discover significant phenologs, and were performed as follows:

A binary gene-disease association matrix was generated for each species, where the columns represent phenotypes. The rows in the human (or prediction) matrix each represent a single human gene; a true value in cell (i, j) indicates an association has been observed between gene i and disease j. Genes that have no identifiable orthologs in any species are excluded. False values in cells indicate that no association has been observed.

The rows in other species' matrices (the source matrices) are also described in terms of human genes: if the human gene has no ortholog in that species, the row is absent; but if the human gene has one or more orthologs in that species, a single row represents the whole set of orthologs. The presence of a true value in cell (i,j) indicates that a species-specific ortholog of human gene i is observed as associated with species-specific phenotype j. False values indicate no observed association.

Phenologs correspond to mappings between a prediction matrix column and the most similar source matrix columns. To compute intercolumn distances, a submatrix of the prediction matrix is generated, its rows limited to those shared by the source matrix. Treating each phenotype or disease as a column vector, a distance is computed between each of the phenotypes in the source matrix and each of the diseases in the prediction matrix.

As for the calculation of phenologs described above, we defined our distance function as the hypergeometric probability of observing c or more common genes between source phenotype u and prediction disease v, with n total observations in one and m total observations in the other. The cardinality of the vectors u and v is N, the total number of human genes with orthologs in the source species. Thus, the probability is calculated as in the equation given above.

For each prediction disease v, we selected the source phenotype with the smallest distance as the top hit (best performing phenolog), then predicted genes' associations with the human disease according to their associations (true or false) with the source phenotype.

Predictive accuracy was evaluated by 10-fold cross-validation, omitting 10% of the prediction matrix rows for each of ten successive tests, and only evaluating predictions on the with-held 10% test set of genes, repeating for 10 unique test sets, and measuring true and false-positive prediction rates using ROC analysis.

We observed that those phenologs ranked just below the best (smallest distance) hit often provided additional valuable information about a disease. One simple method for integrating predictions across phenologs is to combine information from the k nearest neighbors (the top hit would be $k = 1$). In some cases, distance to the k^{th} neighbor is equal to that of additional neighbors, representing a tie; in which case we included all neighbors tied with item k.

A simple weighting scheme was used to integrate evidence from the k (and tied with k^{th}) nearest neighbors, calculating a score for each human gene (row) as:

$$p(\text{gene} \in \text{disease} \mid k \text{ disease phenologs}) =$$
$$1 - \prod_{i=1}^{k} (1 - p(\text{gene} \in \text{disease} \mid \text{phenolog } i \text{ is correct})$$
$$x\ p(\text{phenolog } i \text{ is correct}))$$

We define the probability that the phenolog is correct (the final term) as one minus the hypergeometric probability given previously. For the probability of the gene being associated with the disease given that the phenolog i is correct, we use the following empirical score: for a true source observation, as the ratio of the phenolog intersection (the size of set $u \cap v$, defined above) to the size of set u; for a false source observation, as zero. Thus, although observations are binary (true or false), predictions are represented by scores (between 0 and 1), which are essentially weighted averages of the predictions of the k nearest orthologous phenotypes.

Null distributions were calculated by repeating the cross-validated analysis with 10 randomizations of the prediction matrix. Randomization was accomplished by shuffling the true values in each prediction matrix column, to ensure that the phenotype gene set size distribution was maintained.

Tests of Subnetwork Modularity. We measured the degree of network interconnectivity among orthologs involved in phenologs from yeast and worms using a modification to a recently developed measure of the network clustering of a set of genes (2, 4). Given a query set of genes, their interconnectivity in a functional gene network [a gene network with edge weights corresponding to the log likelihood of the linked genes functioning in the same biological process (4, 10)] is calculated as the area under a ROC curve (AUC) for predicting back members of the query gene set when rank-ordering all genes in the network by each gene's sum of edge weights to the query gene set (corresponding to the naïve Bayes probability of participating in the same process as genes in the query set), performing the test using cross-validation (each query gene is omitted in turn from the query set for purposes of its evaluation). AUC ranges from 0 to 1. A high AUC (near 1) indicates that query genes are more tightly connected in the

network to each other than to other genes; an intermediate AUC (near 0.5) corresponds to no better than random recovery of query genes, indicating negligible interconnectivity of the query gene set in the network. (AUC values in the range of 0 to near 0.5 indicate worse than random expectation, e.g., systematically lower connectivity of the query set.)

To analyze phenolog gene sets, we modified the method by converting the gene-centric functional yeast network (10) into a network of orthologs based upon INPARANOID ortholog assignments. We retained only yeast gene-gene network edges connecting orthologs present in both yeast and worm. In the case that multiple genes are assigned to a single ortholog, multiple network edges could exist between a pair of orthologs; we retained only the edge with the greatest weight (confidence). The resulting yeast network contains ortholog-ortholog functional associations, rather than gene-gene associations. Using this network, we calculated AUC as in (2, 4): for a given ortholog query set (e.g., the set of orthologs in the intersection of a phenolog), we rank ordered all orthologs shared between yeast and worm by the sum of the edges connecting them to the query set, then calculated AUC for recovery of the query ortholog set using cross-validation.

We calculated network AUC for genes (orthologs) within and outside of phenolog intersections (Fig. S8), considering all significant (5% FDR) yeast-worm phenologs with at least four genes in both the phenolog intersection ortholog set and the ortholog set outside the intersection. To correct for possible query gene size effects, we subsampled the larger of the two sets. For example, if the intersection of a worm phenotype and a yeast phenotype has 30 orthologs and the yeast phenotype has 15 additional orthologs, we calculated the AUC of the 15 additional orthologs, then randomly sampled 15 genes at a time from the intersection set, calculating the AUC of each subset of 15 genes, taking the median value of 100 such samplings as the AUC for the intersection set.

Tests of Deep Paralogy. In principle, deep paralogs or gene families could be responsible for significant phenologs, rather than modules of nonsequence related genes. We reasoned that the deep paralog hypothesis predicts that the overlapping intersection of orthogroups involved in a phenolog should have more significant pair-wise BLAST E-values than the nonintersecting genes that are involved in the same phenotype.

To test this hypothesis, we compared genes in each phenolog set [either intersection (I) or unique (D_{1or2})] (Fig. S9) in pair-wise fashion using default BLASTP settings to all other genes in the set. The most significant BLAST E-values were collected for each orthogroup pair in each set. (BLAST E-values between genes within the same orthogroup were omitted, as genes from the same orthogroup do not contribute separately to calculating the phenolog. When multiple genes from the same orthogroup belong to a set, only the single most significant BLAST E-value to a gene outside the orthogroup was included for that orthogroup.)

For the significant (5% FDR) phenologs of each species pair, we separately collected the BLAST values for the orthogroups in either the intersecting sets or the unique sets, comparing the E-value distributions on a species-by-species basis (Fig. S9). BLAST E-values less significant than 1e-3 were truncated at 1e-3; BLAST E-values more significant than 5e-313 were truncated at 5e-313.

***Xenopus laevis* Embryo Manipulations.** Female *Xenopus laevis* were ovulated overnight after injecting human chorionic gonadotrophin, and eggs were squeezed out for fertilization in vitro. At the two-cell stage, the jelly layer of embryos was removed by swirling in 3% cysteine (pH 7.9) in 1/3× MMR medium and washed in 1/3× MMR five times. For microinjections, embryos were placed in 2% Ficoll in 1/3× MMR, and injected using forceps and an Oxford universal micromanipulator, then reared in 2% Ficoll in 1/3× MMR to stage 9, then washed and reared in 1/3× MMR.

Whole-mount in situ hybridization was performed using a modified method omitting acetylation steps from the standard method (11). For all experiments, morpholino antisense oligonucleotides (MOs) were injected at 20 to 60 ng per blastomere. The posterior cardinal vein and intersomitic veins were targeted by injecting into the two ventral cells equatorially at the four-cell stage. Neural crest cells were targeted by injecting into one dorsolateral blastomere in 16-cell stage embryos.

For whole-mount in situ hybridization for *erg* and *agtrl1*, embryos were fixed in MEMFA medium at stages 34 to 36. The hemorrhage phenotype was photographed at stage 45 after anesthetizing with Benzocaine. For in situ screening for genes expressed in blood vessels, all probes were in vitro transcribed using PCR-amplified cDNA fragments as templates. The T7 promoter sequence is inserted to the 5′ of each reverse primer for the in vitro transcription reaction. In situ hybridizations were initially performed using an InsituPro Vsi automated hybridization station (Intavis) and positive genes were confirmed by a manual protocol.

PCR primers for genes expressed in blood vessels are listed below:

hmha1-F: 5′-TTTTCAAGGAAGAAGCGGGAAC-3′
hmha1-R: 5′-GCGATTTAGGTGACACTATAGCCACCAC-ACAGACTTTCCATTGAC-3′
rab11b-F: 5′-TGGGAGCCAGAGATGACGAATAC-3′
rab11b-R: 5′-GCGATTTAGGTGACACTATAGTGCTGG-ATTTCTGTCCATCCG-3′
tcea1/3-F: 5′-CATTGGAGCTGCTTCAGTCCAC-3′
tcea1/3-R: 5′-GCGATTTAGGTGACACTATAGGTCAGG-TTTTTCCGCATCTCTTTC-3′
tbl1xr1-F: 5′-CCCATTCAGCATTCACTTTTGG-3′
tbl1xr1-R: 5′-GCGATTTAGGTGACACTATAGAAGCCA-TCATAAGACCCAGTTGC-3′

Images of embryos were obtained with a Leica MZ16FA stereomicroscope using ImageProPlus software.

Morpholino Oligonucleotides and cDNA Clones. *sox13*, *erg*, and *agtrl1* cDNAs were obtained from Open BioSystem (*sox13*: IMAGE:6636177, *erg*: IMAGE:5512670, *agtrl1*: IMAGE:8321886). Translation blocking antisense morpholinos for *sox13* and *sec23ip* were designed based on the sequences from the National Center for Biotechnology Information database (*sox13*: BC068647.1 *sec23ip*: BC079740.1). MOs were obtained from Gene Tools. All MO sequences are listed below:

sox13-MO: 5′-TCACCCTGTATGGTATCCATTTAAG-3′
sox13-MM: 5′-TCAGCCTCTATGCTATGCATTCAAG-3′
sec23ip-MO: 5′- CCCCGTTCGGTACCTTCCCCGCCAT-3′

Tube Formation Assays. Human umbilical vein endothelial cells (HUVECs) were purchased from Clonetics, and were used between passages 4 and 5. HUVECs were cultured on 0.2% gelatin-coated (Sigma) plates in endothelial growth medium-2 (EGM-2; Clonetics) in tissue culture flasks at 37 °C in a humidified atmosphere of 5% CO_2. Next, 10^5 cells were replated in six-well plates 1 day before transfection. Scrambled siRNA (Ambion, cat #4611), siRNA (5′- UGCUGAGAAUGAGAGCGGC-3′) corresponding to the human *HOXA9* sequence (12), or human *SOX13*–specific siRNA (Dharmacon Smartpool L-020038–01-0005, specific sequences provided below) were transfected into HUVECs using Lipofectamine RNAiMAX (Invitrogen) according to the manufacturer's instructions. Transfected HUVECs (10^4 cells) were seeded into 96-well plates coated with 50 μL of ECMatrix

(Chemicon) according to the manufacturer's instructions. Cells were incubated for 16 h, monitoring tube formation by microscopy. Branched structures were quantified using the program ImageJ (http://rsb.info.nih.gov/ij).

J-020038–09: 5′-CCUUUAGGGCUUAUGGCCA-3′
J-020038–10: 5′-GUAAACAUAGAUAGUGCUU-3′
J-020038–11: 5′-CAGCAGAUCCAGCAGGUUA-3′
J-020038–12: 5′-CUGCCAUGCUGUUUGAGAA-3′

1. Amberger J, Bocchini CA, Scott AF, Hamosh A (2008) McKusick's Online Mendelian Inheritance in Man (OMIM). *Nucleic Acids Res* 37(Database issue)D793–D796.
2. McGary KL, Lee I, Marcotte EM (2007) Broad network-based predictability of *Saccharomyces cerevisiae* gene loss-of-function phenotypes. *Genome Biol* 8:R258.
3. Eppig JT, et al.; Mouse Genome Database Group (2005) The Mouse Genome Database (MGD): from genes to mice—a community resource for mouse biology. *Nucleic Acids Res* 33(Database issue)D471–D475.
4. Lee I, et al. (2008) A single gene network accurately predicts phenotypic effects of gene perturbation in *Caenorhabditis elegans*. *Nat Genet* 40:181–188.
5. Chen N, et al. (2005) WormBase: a comprehensive data resource for *Caenorhabditis* biology and genomics. *Nucleic Acids Res* 33(Database issue)D383–D389.
6. Dwight SS, et al. (2002) *Saccharomyces* Genome Database (SGD) provides secondary gene annotation using the Gene Ontology (GO). *Nucleic Acids Res* 30:69–72.
7. Hillenmeyer ME, et al. (2008) The chemical genomic portrait of yeast: uncovering a phenotype for all genes. *Science* 320:362–365.
8. Swarbreck D, et al. (2008) The *Arabidopsis* Information Resource (TAIR): gene structure and function annotation. *Nucleic Acids Res* 36(Database issue)D1009–D1014.
9. Remm M, Storm CE, Sonnhammer EL (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J Mol Biol* 314:1041–1052.
10. Lee I, Li Z, Marcotte EM (2007) An improved, bias-reduced probabilistic functional gene network of baker's yeast, *Saccharomyces cerevisiae*. *PLoS One* 2:e988.
11. Harland RM (1991) In situ hybridization: an improved whole-mount method for *Xenopus* embryos. *Methods Cell Biol* 36:685–695.
12. Bruhl T, et al. (2004) Homeobox A9 transcriptionally regulates the EphB4 receptor to modulate endothelial cell migration and tube formation. *Circ Res* 94:743–751.

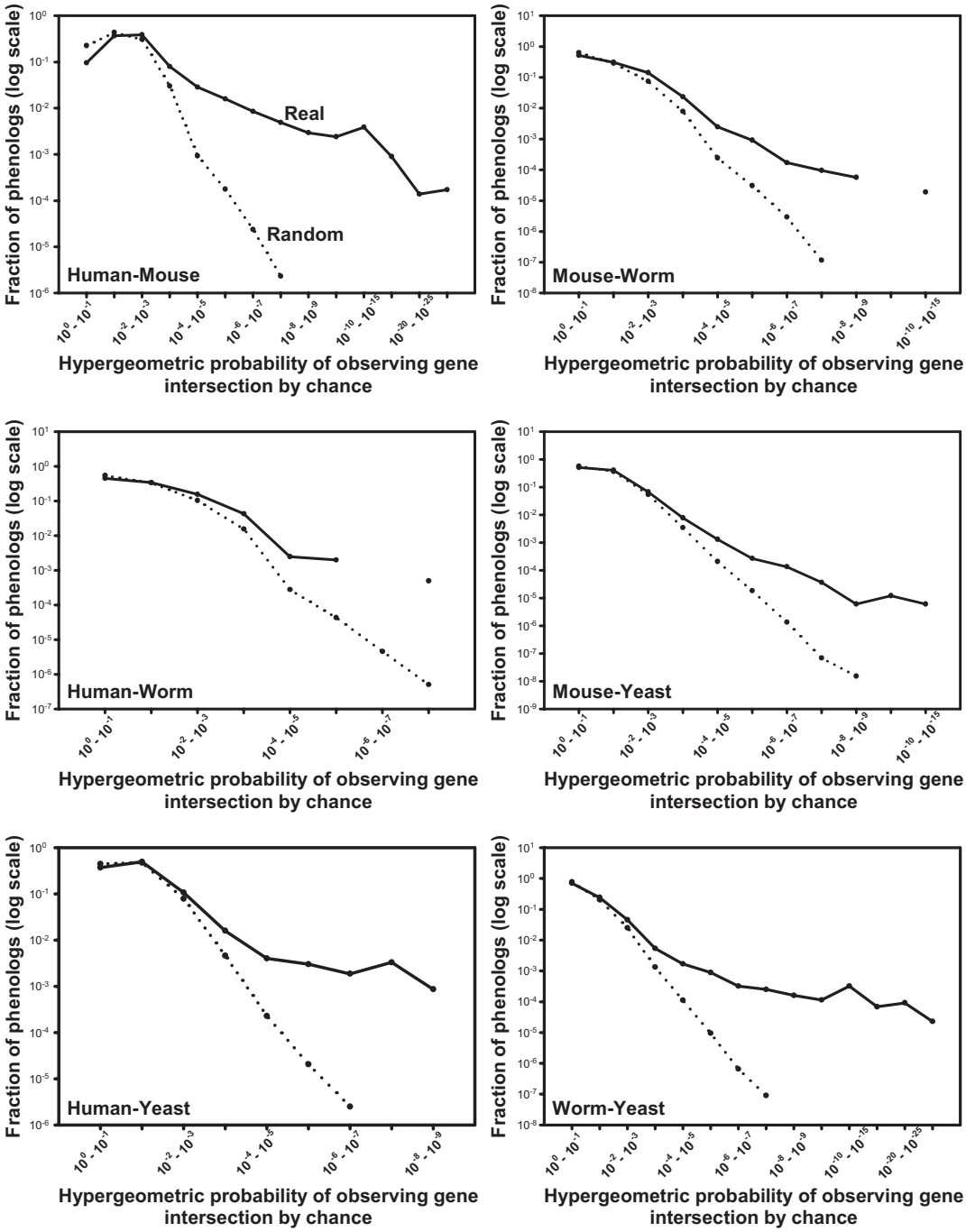


Fig. S1. Enrichment for phenologs above random expectation can be seen following all pair-wise comparisons of the mutational phenotypes from mouse, human, yeast, or worm. Histograms are plotted as in Fig. 2*B*.

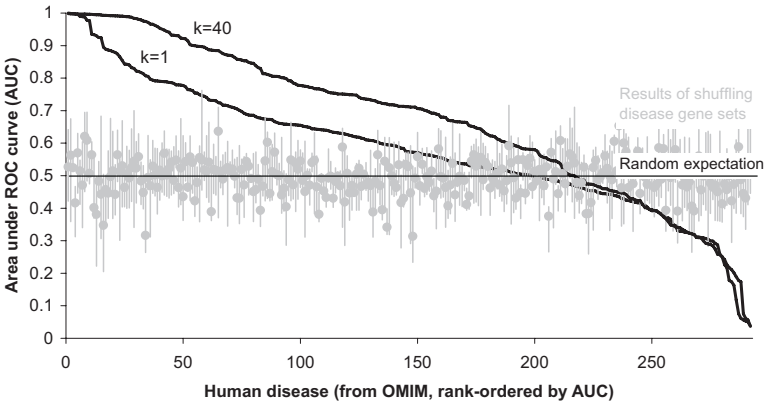
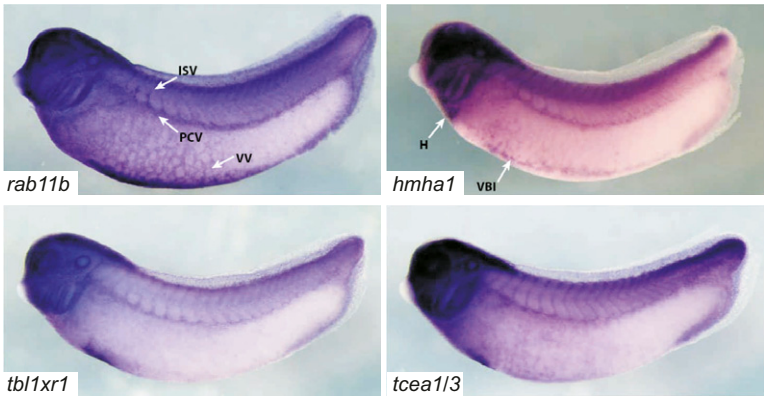


Fig. S2. Ten-fold cross-validated tests show strong disease gene prediction by single phenologs for approx. one-sixth to one-fifth of tested diseases; simple weighted combinations of phenologs (e.g., evaluating the $k = 40$ best phenologs) provide strong predictability for approximately one-third to one-half of the tested diseases. Predictability is measured as the area under a receiver-operater characteristic (ROC) curve as described in *SI Materials and Methods* and evaluated separately for each human genetic disease with ≥ 2 associated genes. An area under the ROC curve (AUC) of 1 indicates perfect prediction of known disease genes in a cross-validated test; an AUC of 0.5 indicates performance no better than chance. Error bars indicate first quartile, median, and third quartile of predictions of shuffled disease gene sets from the $k = 1$ test; score distributions from shuffling tests are similar for both $k = 1$ and $k = 40$ and center around AUC = 0.5, as expected by chance. OMIM, Online Mendelian Inheritance in Man.



ISV, intersomitic vein; PCV, posterior cardinal vein; VV, vitellin vein; VBI, ventral blood island; H, heart

Fig. S3. In situ hybridization shows vascular expression of four candidate angiogenesis genes in stage 32 *Xenopus* embryos.

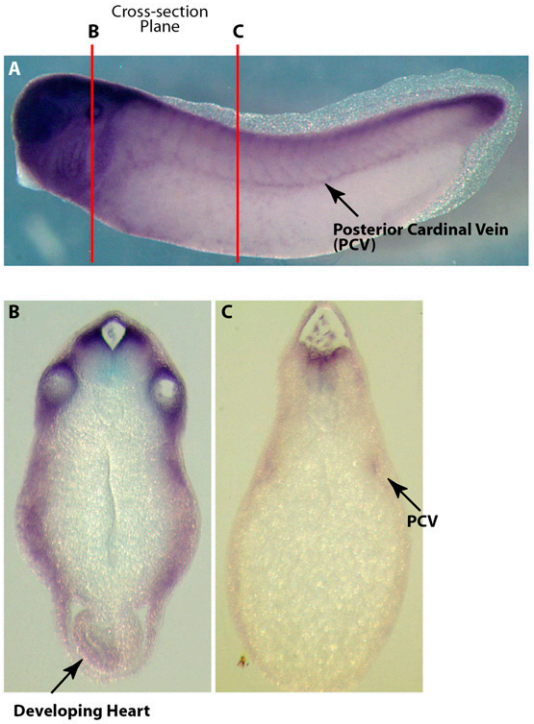


Fig. S4. (*A–C*) In situ hybridization shows *sox13* expression in veins and developing heart of a stage 32 *Xenopus* embryo.

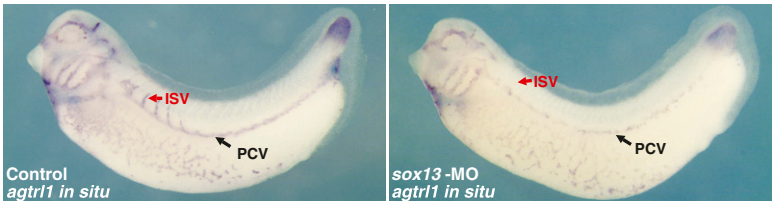


Fig. S5. Morpholino (MO) knockdown of *sox13* induces defects in vasculature, measured using in situ hybridization versus an independent marker of the vasculature, the angiotensin receptor homolog *agtrl1* (12 of 19 animals tested). Such defects are rare in untreated control animals (0 of 22 control animals tested with *agtrl1*).

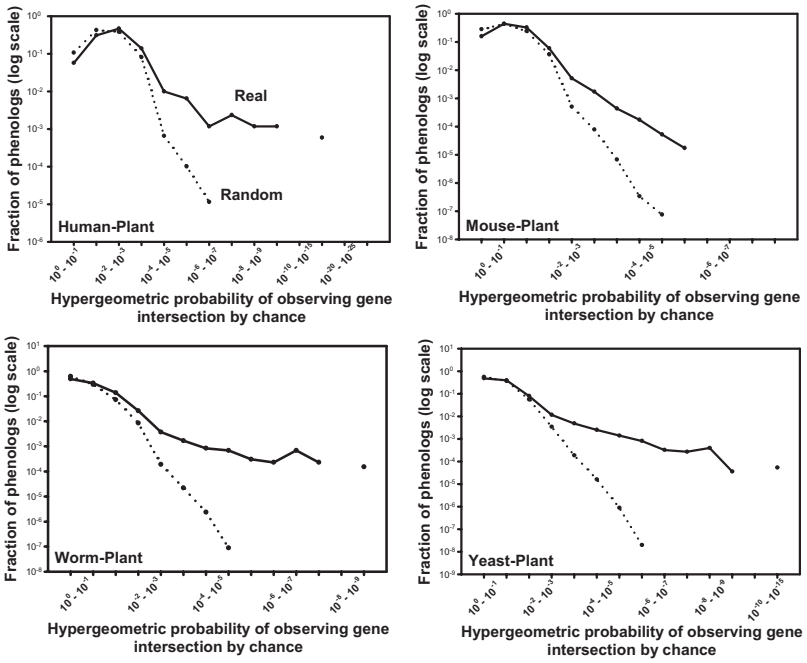


Fig. S6. Enrichment for phenologs above random expectation can be seen following all pair-wise comparisons of *Arabidopsis* phenotypes with those from mouse, human, yeast, or worm. Histograms are plotted as in Fig. 2B and Fig. S1.

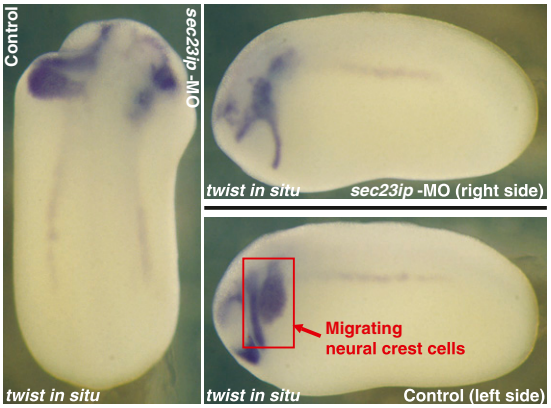


Fig. S7. Morpholino (MO) knockdown of *sec23ip* induces defects in neural crest cell migration, measured using in situ hybridization versus *twist*, an independent marker of the neural crest cells (8 of 14 animals tested). Such defects are rare in untreated control animals (0 of 14 control animals tested with *twist*).

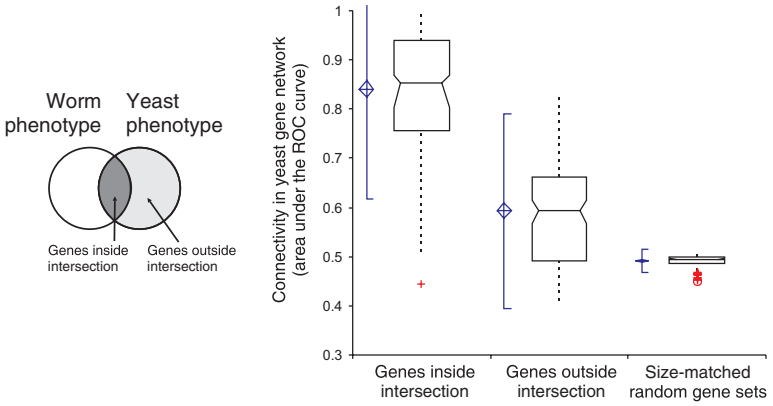


Fig. S8. Genes involved in phenologs show enhanced interconnectivity in a gene network, shown here for yeast genes (10). All significant yeast-worm phenologs with at least four orthologs in both the "intersection" and "nonintersection" sets (*SI Materials and Methods*) were tested for network connectivity, measured as the area under a ROC plot as described in ref. 2, with values ranging from 0.5 (random network connectivity) to 1 (high network connectivity). Genes from phenolog intersections show significantly higher network connectivity than genes associated with a phenolog, but outside of the intersection, which in turn show significantly higher connectivity than size-matched random gene sets. Thus, phenologs capture subnetworks or network modules informative about a given phenotype pair, and carry predictive value for additional genes relevant to the phenotypes. At the left of each box-and-whisker plot, the center of the blue diamond indicates the mean AUC across phenologs, the top and bottom of the diamond indicate the 95% confidence interval, and the accompanying solid vertical line indicates $\pm$ 2 SD. The bottom, middle, and top horizontal lines of the box-and-whisker plots represent the first quartile, the median, and the third quartile of AUCs, respectively; whiskers indicate 1.5 times the interquartile range. Red plus signs represent individual outliers.

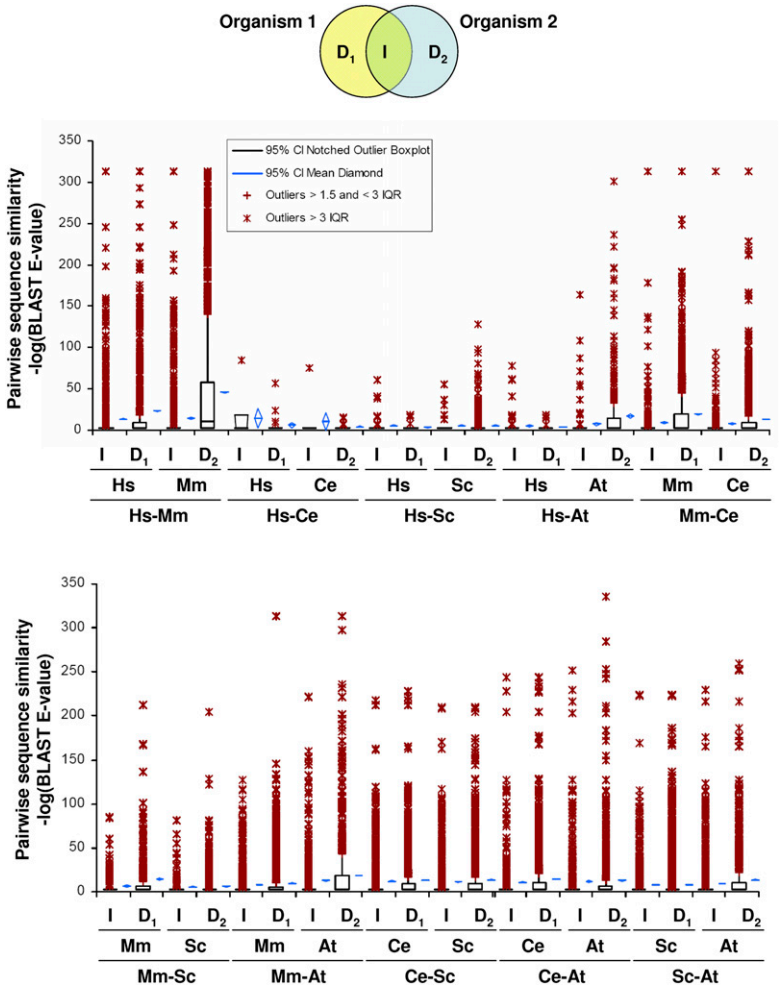


Fig. S9. To rule out the possibility that phenolog intersections arise predominantly from "deep paralogs," we measured the pair-wise BLAST E-values between genes in phenolog intersections (I) and genes in the same phenotypes but not the phenolog intersections (the differential gene sets D_1 or D_2), comparing the E-value distributions on a species-by-species basis for each species pair (see *SI Materials and Methods* for details). Distributions are plotted above; box plots represent first quartile, median, and third quartile, whiskers 1.5 × interquartile range, and stars represent outliers >3 IQR [thus, the majority of pair-wise sequence comparisons show no significant similarity with $-\log_{10}(\text{Eval}) = 3$]. In general, genes in phenolog intersections were no more likely to encode similar protein sequences than genes in the phenologs but outside the intersections (i.e., associated with the phenotype in only one species), indicating that deep paralogy is not a dominant factor in identifying phenologs. Across 20 such comparisons (10 species pairs, performing tests on a per species basis for each comparison), in 14 cases gene pairs in the I sets showed less significant BLAST E-values than those in the D sets (one-tailed $P < 0.0001$ for each; Wilcoxon-Mann-Whitney); in 3 cases gene pairs in the I sets showed more significant BLAST E-values than in the D sets (1 tailed $P < 0.0001$, $P < 0.02$, $P < 0.03$); and in 3 cases the sets were not significantly biased in either direction.

Table S1. Literature evidence for worm *him* genes involvement in human breast/ovarian cancer

Human gene (alias)	Worm gene	Comment
FAM82B	*F33H2.6*	Copy number variation and differential gene expression in primary breast tumors (1)
HMG20A,B (Braf35)	*W02D9.3*	DNA-binding protein in complex with protein encoded by breast cancer susceptibility gene BRCA2 (2)
HORMAD2,1	*him-3,htp-1,2*	Copy number variation and differential gene expression in basal breast cancer (3)
KIF15 (NY-BR-62)	*klp-10,18*	Overexpression in breast cancer (4)
MRE11A	*mre-11*	Mis-sense mutation in breast cancer tumor (5)
RAD1	*mrt-2*	Overexpression/phosphorylation of 911 complex (RAD1, RAD9, HUS1) in breast cancer, as detected for complex partner Rad9 (6)
RAD21	*coh-1*	Genetic association study found 3 significant polymorphisms associated with familial breast cancer (7)
SVIL	*viln-1*	Up-regulated in brain metastases of breast cancer (8)
TSPO, BZRPL (PBR)	*C41G7.3*	Overexpression in breast cancer and change in localization correlates with aggressive tumor growth (9)

1. Chin SF, et al. (2007) High-resolution aCGH and expression profiling identifies a novel genomic subtype of ER negative breast cancer. *Genome Biol* 8:R215.
2. Marmorstein LY, et al. (2001) A human BRCA2 complex containing a structural DNA binding component influences cell cycle progression. *Cell* 104:247–257.
3. Adélaïde J, et al. (2007) Integrated profiling of basal and luminal breast cancers. *Cancer Res* 67:11565–11575.
4. Scanlan MJ, et al. (2001) Humoral immunity to human breast cancer: antigen definition and quantitative analysis of mRNA expression. *Cancer Immun* 1:4.
5. Fukuda T, et al. (2001) Alterations of the double-strand break repair gene MRE11 in cancer. *Cancer Res* 61:23–26.
6. Chan V, et al. (2008) Localization of hRad9 in breast cancer. *BMC Cancer* 8:196.
7. Sehl ME, et al. (2009) Associations between single nucleotide polymorphisms in double-stranded DNA repair pathway genes and familial breast cancer. *Clin Cancer Res* 15:2192–2203.
8. Nishizuka I, et al. (2002) Analysis of gene expression involved in brain metastasis from breast cancer using cDNA microarray. *Breast Cancer* 9:26–32.
9. Hardwick M, et al. (1999) Peripheral-type benzodiazepine receptor (PBR) in human breast cancer: correlation of breast cancer cell aggressive phenotype with PBR expression, nuclear localization, and PBR-mediated cell proliferation and nuclear transport of cholesterol. *Cancer Res* 59:831–842.