IMP 2.0 Methods

1) Inference of functional relationship networks

A. Data processing

We collected 2,444 microarray datasets from NCBI Gene Expression Omnibus (GEO) covering 43,865 conditions in seven organisms. Probes were collapsed and normalized according to the procedure described in (1) and Fisher's z-transformed pearson correlations were calculated for each gene pair as described in (2).

Physical and genetic interaction data from BioGRID (3), IntAct (4), Mint (5), and MIPS (6) 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 (7) and TF binding site motifs retrieved from Jaspar (8). Yeastract's predicted TF-gene relationships were treated as pair-wise binary scores. For the Jaspar database, we searched for the presence of transcription factor binding site motifs by scanning for each TF profile in the 1 kb upstream sequence of all genes using FIMO (9). 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 gene's binary score vectors.

Phenotype and disease data from SGD (10), MGI (11), Wormbase (12), Flybase (13), GSEA (14) and Zfin (15) were processed by summing the co-occurrences of gene pairs in all phenotypes/diseases and normalizing by the size of the phenotype/disease.

Specifically, for a gene pair, *i, j* the scoring function was the following:

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

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

Protein sequence similarity between genes were obtained from Biomart (16), and protein domain data were treated as binary evidence (indicating the presence or absence of a shared domain) from PfamA (17) and PROSITE (18).

B. Gold standard construction

We summarized the processed genomic data into a global functional relationship network for each organism (*Homo sapiens, Mus musculus, Rattus novegicus, Drosophila melanogaster, Danio rerio, Saccharomyces cerevisiae* and *Caenorhabditis*

elegans) using a regularized naïve Bayes classifier method as described in (1). First, we created a gold standard for training from genes that were co-annotated to a set of 973 expert-selected Gene Ontology biological process terms. These gene pairs were considered functionally related genes. Gene pairs *not* co-annotated to any term in this set of GO terms, KEGG (19), PID (20) or Biocyc (21) were considered 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 (22))

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.

C. Generating global functional relationship networks

One binary regularized naive Bayes classifier was trained per Gene Ontology term (out of 433), if there were at least 10 genes annotated to the term in a particular organism. Gene pairs co-annotated to the term were used as positive examples for training. Gene pairs from the global unrelated pairs where one member of the pair was annotated to the term were used as negative examples for training.

Each classifier contained one class node determining the binary presence or absence of a functional relationship between two genes and organism specific dataset nodes conditioned on the class node.

D. Regularization of naive Bayes classifier

When integrating a large number of genomic datasets, the naive Bayes assumption of conditional independence among datasets might not hold. We have previously shown that a mutual information based regularization of naive Bayes classifiers can alleviate the conditional dependency among datasets (1). We modified this method by applying a conditional mutual information (MI) calculation between datasets. MI was calculated from the set of gene pairs that were considered functionally unrelated, generated as described above. The sum of shared information U_k for dataset D_k is calculated as follows:

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

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

where $I_{pairs \in negatives}$ is the mutual information between dataset D_k and D_i for unrelated gene pairs and H is the single dataset entropy. The weight (α_k) given to a dataset exponentially decreases as its shared information increases.

E. Functional relationship inference

The final regularized naive Bayes functional relationship posterior probability for gene pair *i,j* is the following:

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

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

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, we average the edge probabilities from each process specific functional network to generate the final global functional relationship network.

2) Knowledge transfer of GO biological process annotations

To leverage the research strength across organisms, we systematically applied a knowledge transfer for machine learning as described in (23) by identifying functionally similar homologs as described in (24).

A. Functional similarity score

We started with sequence similarity derived TreeFam (25) gene families which cover both paralog and ortholog gene relationships. Next, we filtered the set of paralogs and orthologs for gene pairs that were also functionally similar (which we name '*functional analogs*'). We defined a functional analog to be a gene pair that had 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* were the number of TreeFam gene families in each gene *G1* and *G2*'s direct neighborhood in the functional network, *k* was the number of overlapping TreeFam gene families between gene *G1* and *G2* gene neighbors and *N* was 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, and thus can be interpreted as a hypergeometric p-value.

B. Annotation transfer

We used a functional similarity score cutoff of <= 0.01 (corresponding to a p-value <= 0.01) to consider a gene pair as functional analogs. All experimental annotations (GO evidence codes: EXP, IDA, IPI, IMP, IGI, IEP) were propagated between functional analogs.

3) Biological process predictions with network based SVM

Previous work has shown that functional relationship network based Support Vector Machines (SVM) can outperform methods that directly use raw data as input to an SVM or methods that simply sum network edge weights of the positive examples (26). Additionally, we have shown that functional knowledge transfer is robust to machine learning method (23).

A. Constructing an SVM classifier with a functional network

We trained an SVM classifier for each biological process term using the direct and knowledge-transferred genes for a process as positive examples. The feature space was constructed as the weights in the functional relationship network. Thus for each gene example, edge weights to every other gene make up its feature vector. The total number of features in an organism will be equal to the number of genes in that organism. The set of feature vectors for the 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.

B. Converting SVM scores to probabilities

Finally, the unbounded SVM prediction scores were transformed into probabilities using a maximum likelihood sigmoid fit to the SVM outputs (27).

4) Biological process enrichment analysis

A. Calculating p-values for enrichment

Biological process enrichment p-values were calculated using the hypergeometric distribution. A p-value for a process *t* was calculated as:

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

where *n* was the number of genes in the network, *k* was the number of genes in the network that were also annotated to process *t*, *m* was the total number of genes in *t*, and N was the total number of genes annotated to any process.

B. Multiple test correction

Multiple test correction was performed by controlling the false discovery rate (FDR) using the procedure described in (28). Briefly, for all biological processes tested for enrichment, the corresponding p-values (P_t) were sorted in ascending order. The FDR corrected p-value was calculated as:

$$P_{FDR(t)} = \frac{N}{r_t} P_t$$

where r_t was the rank of the p-value for biological process *t*, P_t was the raw p-value for process *t*, and *N* was the total number of tested processes. A process was decided to be enriched if $P_{FDR(t)} < .05$.

5) Custom SVM Predictions

A. Constructing an SVM classifier from a user's gene set

The user provides gold standard positives genes to IMP using the 'My Predictions' feature. IMP maps these gene symbols to Entrez GeneIDs. A set of 1,000 random genes not included in the positive gene set is selected as negative

examples. These gene sets are used to label positive and negative gene examples from the corresponding functional network, as described in section 3. These labeled data are analyzed using the SVM^{perf} library through the SVMperfer wrapper from the Sleipnir library (available from the source control repository at http://bitbucket.org/libsleipnir/sleipnir). We use a linear kernel and optimize for the area under the curve (AUC). We evaluate performance with seven different penalty factors (C in SVM^{perf}) varying from 1000 to 0.01. When the penalty factor is too high, it can lead to models that generalize poorly and when it is too low, it can lead to models that do not sufficiently fit the data. This process allows IMP to perform well across a wide range of problems. Five-fold cross validation is used to assess the model. Unlabeled genes are classified based on their distance to the maximum-margin hyperplane.

B. Post-processing and evaluation

IMP post-processes the classified genes from GeneIDs back to symbols. We use the cross validation results for genes in the gold standard to calculate an area under the TPR/FPR curve (AUC). The AUC is a commonly used metric to evaluate the *overall* results of a machine-learning strategy. The AUC is equal to the probability that an input (i.e. positive) gene is ranked higher than other non-positive genes. A perfect classifier will result in an AUC of 1.0, while a random classifier will result in an AUC of 0.5. Additionally, we estimate the probability for each prediction by applying the method described in section 3B. For individual predictions, the provided probabilities should be used to decide whether there is sufficient confidence to follow up with experiments.

5) Disease predictions

A. Disease-gene annotations

Disease-gene annotations were downloaded and processed from OMIM [29]. OMIM terms and annotated genes were mapped to their respective Disease Ontology (DO) [30] terms using the OMIM to DO mapping provided by the Disease Ontology OBO file as cross-references. The method for mapping OMIM to DO terms is described and evaluated in [30] and [31]. In cases where multiple OMIM terms were mapped to the same DO term, the genes corresponding to all of the OMIM terms were added to the DO term. Gene-disease annotations were propagated to parent terms according to the DO structure.

B. Disease predictions with network based SVM

Human:
For each disease with sufficient known genes (>= 3 genes), we trained an SVM as described in section 3A. In short, known disease genes are treated as positive examples and a gene's features are its neighbors to every other gene in the organism's

functional network. The subsequent SVM scores are converted to probabilities as described in section 3B.

Other organisms:
We mapped the human disease genes to their homologs in other organisms using the process described in section 2. These transferred annotations were then treated as positive examples for disease gene prediction using the SVM-based classification, with the non-human network, as described before. For example, to predict candidate diabetes genes in mouse, the homologs of known human disease genes (according to FKT, section 2), are positive examples in an SVM classifier whose features are built form the mouse functional network.

Disease-gene predictions with probabilities < .01 are not shown in IMP unless the gene is known to be involved in the disease.

1. Huttenhower,C., Haley,E.M., Hibbs,M.A., Dumeaux,V., Barrett,D.R., Coller,H.A. and Troyanskaya,O.G. (2009) Exploring the human genome with functional maps. *Genome Res.*, **19**, 1093–1106.

2. Huttenhower,C., Hibbs,M., Myers,C. and Troyanskaya,O.G. (2006) A scalable method for integration and functional analysis of multiple microarray datasets. *Bioinformatics*, **22**, 2890–2897.

3. Stark,C., Breitkreutz,B.-J., Chatr-Aryamontri,A., Boucher,L., Oughtred,R., Livstone,M.S., Nixon,J., Van Auken,K., Wang,X., Shi,X., et al. (2011) The BioGRID Interaction Database: 2011 update. *Nucleic Acids Res.*, **39**, D698–704.

4. Kerrien,S., Aranda,B., Breuza,L., Bridge,A., Broackes-Carter,F., Chen,C., Duesbury,M., Dumousseau,M., Feuermann,M., Hinz,U., et al. (2012) The IntAct molecular interaction database in 2012. *Nucleic Acids Res.*, **40**, D841–6.

5. Licata,L., Briganti,L., Peluso,D., Perfetto,L., Iannuccelli,M., Galeota,E., Sacco,F., Palma,A., Nardozza,A.P., Santonico,E., et al. (2012) MINT, the molecular interaction database: 2012 update. *Nucleic Acids Res.*, **40**, D857–61.

6. Pagel,P., Kovac,S., Oesterheld,M., Brauner,B., Dunger-Kaltenbach,I., Frishman,G., Montrone,C., Mark,P., Stümpflen,V., Mewes,H.-W., et al. (2005) The MIPS mammalian protein-protein interaction database. *Bioinformatics*, **21**, 832–834.

7. Abdulrehman,D., Monteiro,P.T., Teixeira,M.C., Mira,N.P., Lourenço,A.B., Santos,dos,S.C., Cabrito,T.R., Francisco,A.P., Madeira,S.C., Aires,R.S., et al. (2011) YEASTRACT: providing a programmatic access to curated

transcriptional regulatory associations in Saccharomyces cerevisiae through a web services interface. *Nucleic Acids Res.*, **39**, D136–40.

8. Sandelin,A., Alkema,W., Engström,P., Wasserman,W.W. and Lenhard,B. (2004) JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.*, **32**, D91–4.

9. Grant,C.E., Bailey,T.L. and Noble,W.S. (2011) FIMO: scanning for occurrences of a given motif. *Bioinformatics*, **27**, 1017–1018.

10. Cherry,J.M., Hong,E.L., Amundsen,C., Balakrishnan,R., Binkley,G., Chan,E.T., Christie,K.R., Costanzo,M.C., Dwight,S.S., Engel,S.R., et al. (2012) Saccharomyces Genome Database: the genomics resource of budding yeast. *Nucleic Acids Res.*, **40**, D700–5.

11. Bult,C.J., Eppig,J.T., Kadin,J.A., Richardson,J.E., Blake,J.A.Mouse Genome Database Group (2008) The Mouse Genome Database (MGD): mouse biology and model systems. *Nucleic Acids Res.*, **36**, D724–8.

12. Stein,L., Sternberg,P., Durbin,R., Thierry-Mieg,J. and Spieth,J. (2001) WormBase: network access to the genome and biology of Caenorhabditis elegans. *Nucleic Acids Res.*, **29**, 82–86.

13. Drysdale,R.A., Crosby,M.A.FlyBase Consortium (2005) FlyBase: genes and gene models. *Nucleic Acids Res.*, **33**, D390–5.

14. Subramanian,A., Tamayo,P., Mootha,V.K., Mukherjee,S., Ebert,B.L., Gillette,M.A., Paulovich,A., Pomeroy,S.L., Golub,T.R., Lander,E.S., et al. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 15545–15550.

15. Bradford,Y., Conlin,T., Dunn,N., Fashena,D., Frazer,K., Howe,D.G., Knight,J., Mani,P., Martin,R., Moxon,S.A.T., et al. (2011) ZFIN: enhancements and updates to the Zebrafish Model Organism Database. *Nucleic Acids Res.*, **39**, D822–9.

16. Kasprzyk,A. (2011) BioMart: driving a paradigm change in biological data management. *Database (Oxford)*, **2011**, bar049.

17. Punta,M., Coggill,P.C., Eberhardt,R.Y., Mistry,J., Tate,J., Boursnell,C., Pang,N., Forslund,K., Ceric,G., Clements,J., et al. (2012) The Pfam protein families database. *Nucleic Acids Res.*, **40**, D290–301.

18. Sigrist,C.J.A., Cerutti,L., de Castro,E., Langendijk-Genevaux,P.S., Bulliard,V., Bairoch,A. and Hulo,N. (2010) PROSITE, a protein domain database for

functional characterization and annotation. *Nucleic Acids Res.*, **38**, D161–6.

19. Kotera,M., Hirakawa,M., Tokimatsu,T., Goto,S. and Kanehisa,M. (2012) The KEGG databases and tools facilitating omics analysis: latest developments involving human diseases and pharmaceuticals. *Methods Mol. Biol.*, **802**, 19–39.

20. Schaefer,C.F., Anthony,K., Krupa,S., Buchoff,J., Day,M., Hannay,T. and Buetow,K.H. (2009) PID: the Pathway Interaction Database. *Nucleic Acids Res.*, **37**, D674–9.

21. Caspi,R., Altman,T., Dreher,K., Fulcher,C.A., Subhraveti,P., Keseler,I.M., Kothari,A., Krummenacker,M., Latendresse,M., Mueller,L.A., et al. (2012) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.*, **40**, D742–53.

22. Myers,C.L., Barrett,D.R., Hibbs,M.A., Huttenhower,C. and Troyanskaya,O.G. (2006) Finding function: evaluation methods for functional genomic data. *BMC Genomics*, **7**, 187.

23. Park,C.Y., Wong,A.K., Greene,C.S., Rowland,J., Guan,Y., Burdine,R.D. and Troyanskaya,O.G. (2012) Functional knowledge transfer for high-accuracy prediction of under-studied biological processes. *PLoS Comput. Biol. (submitted).*

24. Chikina,M.D. and Troyanskaya,O.G. (2011) Accurate quantification of functional analogy among close homologs. *PLoS Comput. Biol.*, **7**, e1001074.

25. Li,H., Coghlan,A., Ruan,J., Coin,L.J., Hériché,J.-K., Osmotherly,L., Li,R., Liu,T., Zhang,Z., Bolund,L., et al. (2006) TreeFam: a curated database of phylogenetic trees of animal gene families. *Nucleic Acids Res.*, **34**, D572–80.

26. Guan,Y., Ackert-Bicknell,C.L., Kell,B., Troyanskaya,O.G. and Hibbs,M.A. (2010) Functional genomics complements quantitative genetics in identifying disease-gene associations. *PLoS Comput. Biol.*, **6**, e1000991.

27. Platt,J. (1999) Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*.

28. Benjamini,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B …*.

29. Amberger,J.S., Bocchini,C.A., Schiettecatte,F., Scott,A.F. and Hamosh,A. (2014) OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Res.* , 10.1093/nar/gku1205.

30. Kibbe, W. A., Arze, C., Felix, V., Mitraka, E., Bolton, E., Fu, G., … Schriml, L. M. (2014). Disease Ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data. Nucleic Acids Research . doi:10.1093/nar/gku1011

31. Schriml,L.M., Arze,C., Nadendla,S., Chang,Y.W.W., Mazaitis,M., Felix,V., Feng,G. and Kibbe,W.A. (2012) Disease ontology: A backbone for disease semantic integration. Nucleic Acids Res., 40.