BIOINFORMATICS ORIGINAL PAPER

Vol. 28 no. 9 2012, pages 1246–1252
doi:10.1093/bioinformatics/bts120

Systems biology

Advance Access publication March 13, 2012

Predicting the lethal phenotype of the knockout mouse by integrating comprehensive genomic data

Yuan Yuan[1,2], Yanxun Xu[3], Jianfeng Xu[4], Robyn L. Ball[5] and Han Liang[2,*]

[1]Graduate Program in Structural and Computational Biology and Molecular Biophysics, Baylor College of Medicine, [2]Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, [3]Department of Statistics, Rice University, Houston, TX 77005, [4]Department of Statistics, University of Illinois at Urbana-Champaign, Champaign, IL 61801 And [5]Department of Statistics, Texas A&M University, College Station, TX 77843, USA

Associate Editor: Olga Troyanskaya

ABSTRACT

Motivation: The phenotypes of knockout mice provide crucial information for understanding the biological functions of mammalian genes. Among various knockout phenotypes, lethality is of great interest because those involved genes play essential roles. With the availability of large-scale genomic data, we aimed to assess how well the integration of various genomic features can predict the lethal phenotype of single-gene knockout mice.

Results: We first assembled a comprehensive list of 491 candidate genomic features derived from diverse data sources. Using mouse genes with a known phenotype as the training set, we integrated the informative genomic features to predict the knockout lethality through three machine learning methods. Based on cross-validation, our models could achieve a good performance (accuracy = 73% and recall = 63%). Our results serve as a valuable practical resource in the mouse genetics research community, and also accelerate the translation of the knowledge of mouse genes into better strategies for studying human disease.

Contact: hliang1@mdanderson.org

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on December 12, 2011; revised on February 21, 2012; accepted on March 6, 2012

1 INTRODUCTION

The mouse is the premier model organism for interpreting the human genome and plays a key role in studying human diseases (Collins *et al*., 2007). Importantly, the mouse is the only vertebrate species in which pre-selected genes can be deliberately mutated (knocked out) such that the phenotypic effect associated with a gene can be defined in a precise manner. Among various phenotypic effects of disrupting a mouse gene, the lethal phenotype is of particular interest for several reasons. First, the lethal phenotype represents the most severe defects for an organism; the disrupted genes that result in the lethality play essential roles. From an evolutionary point of view, the loss of such a gene reduces the fitness of the organism to zero (Liao and Zhang, 2008). Second, computational analyses on the lethal phenotype of gene deletion reveal some key findings in systems biology, such as the 'centrality–lethality' rule (Jeong *et al*., 2001) and functional compensation among duplicated genes (Gu *et al*., 2003). Third, at the practical level, predicting the knockout lethality of mouse genes is highly valuable for the mouse genetics community since mouse knockout experiments are notoriously time-consuming.

Previous analyses on the knockout lethality of mouse genes have focused on one or a few gene features. For example, Liang and Li (2007) showed a weak correlation between the knockout lethality and protein connectivity. Furthermore, whether a gene is a singleton has little effect on the predictive power of knockout lethality (Liang and Li, 2007; Liao and Zhang, 2007), implying a complex relationship between gene duplicability and gene essentiality (Makino *et al*., 2009). More recently, with the availability of various large-scale mouse genomic datasets, there has been wide interest in employing sophisticated computational approaches to predicting gene function by considering many gene features simultaneously. One pioneering attempt was the 'MouseFunc Prediction' project (Hughes and Roth, 2008) in which different bioinformatics teams integrated diverse datasets to predict the biological functions of mouse genes [represented as gene ontology (GO) terms] using advanced machine learning approaches such as ridge regression (Mostafavi *et al*., 2008), support vector machine (SVM) (Guan *et al*., 2008) and random forest (Tasan *et al*., 2008). These studies made reliable predictions for many GO categories and greatly expanded our ability to discover novel biology (Pena-Castillo *et al*., 2008).

The primary biological questions we aim to address are as follows: (i) which genomic features are most correlated with the lethal phenotype of mouse single-gene knockouts; and (ii) through reasonable computational approaches, to what extent can the knockout lethality be predicted from a wide range of genomic features? Figure 1 outlines the overall scheme of our study. We first assembled a comprehensive list of genomic features derived from diverse data sources for each mouse gene, and identified subsets of informative features for predicting the lethality of single-gene knockouts. Then we integrated the selected genomic features to predict the knockout lethality using three machine learning methods and evaluated their predictive power based on cross-validation. Finally, we examined the bias of genes with a known knockout phenotype (the training set) and estimated the bias-corrected predictive power. Our predictive models achieved a reasonably good performance, demonstrating the feasibility of an integrative approach to predicting knockout lethality in complex organisms.

*To whom correspondence should be addressed.

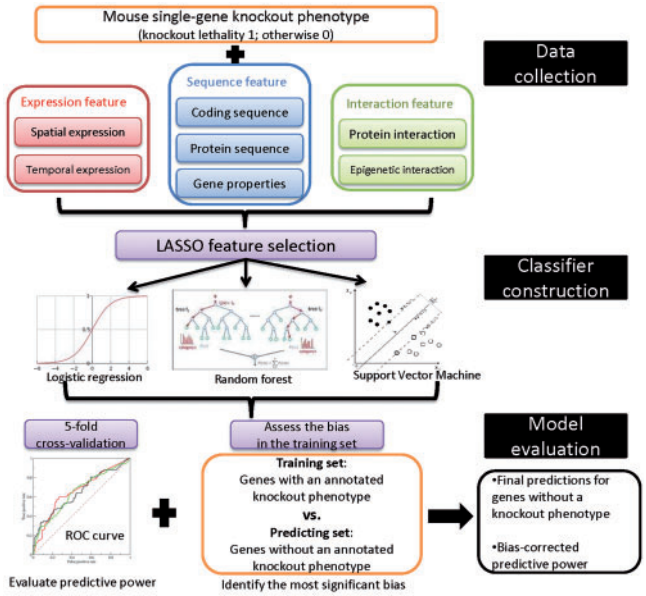

Fig. 1. An overall scheme of the present study.

2 METHODS

2.1 Data collection

Phenotype information of knockout mice was obtained from the Mouse Genome Informatics (MGI) database (http://www.informatics.jax.org). Our analysis included only targeted homozygous knockouts (both copies of a gene were inactivated). The phenotype of a single-gene knockout was classified as lethal if the gene was associated with a lethality annotation (including embryonic/perinatal, prenatal/perinatal and postnatal lethality); otherwise, it was classified as non-lethal.

The gene coding sequence, protein domain, gene homology and structural information were downloaded from Ensembl (release 59), and only those with a one-to-one MGI symbol and Ensembl ID correspondence were retained in the analysis. Gene duplicability was represented by two features: the number of paralogs and paralog sequence identity (the sequence identity with the closest paralog in the mouse genome). The gene evolutionary age was classified into 12 categories based on the presence/absence of the homologs of mouse genes in 11 representative eukaryotic species (including rat, rabbit, human, elephant, dog, chicken, frog, zebrafish, fly, worm and yeast) as previously described (Cai *et al*., 2009). Selective pressure was indexed by d*N*/d*S* between mouse and rat orthologous genes. Hydrophobicity, codon adaptation index, codon bias, effective number of codons, GC content, length of the putative protein in amino acids, frequency of optimal codons and aromaticity were calculated using CodonW (Peden, 1999). For the genes with multiple transcripts, the longest transcript and protein sequences were used in the analysis. The special codons refer to codons with a third-nucleotide mutation from a stop codon, including TAC and TAT encoding Tyr, TGC and TGT encoding Cys, and TGG encoding Trp. The protein helix, sheet and coil content were predicted with PROFphd (Rost and Sander, 1994), and the disordered regions in proteins were predicted by RONN (Yang *et al*., 2005). The protein charge (theoretical isoelectric point) was calculated with ExPAsy Proteomics Server (Gasteiger *et al*., 2003). The length of 5′- or 3′-UTR was calculated based on RefSeq transcripts. The CpG island annotation was obtained from the UCSC Genome Browser, and a CpG island was assigned to a gene if it was within 10-kb upstream from the gene start. MicroRNA target sites were predicted with TargetScanS (Lewis *et al*., 2005) and grouped by microRNA families, and the existence of each type of microRNA target site, as well as the total number of target-site types was calculated for each gene. Mouse recombination rate data were obtained from Cox *et al*. (2009). Information about conserved transcription factor binding sites was downloaded with TRANSFAC track in the UCSC Genome Browser, and the existence of each type of binding site, as well as the total number of binding-site types was calculated for each gene.

The expression data of the mouse transcriptome (Affymetrix GNF1M chips) were downloaded from the NCBI GEO database (Su *et al*., 2004), which included 61 tissue types (spatial expression). The average expression level (both mean and median values), maximum expression level, tissue expression breadth and specificity were calculated as previously described (Liao *et al*., 2006). The temporal expression data for different developmental stages (TS1–TS26, TS28) were obtained from the MGI Gene Expression Database (Finger *et al*., 2011), and for each developmental stage, genes were classified as detected, undetected or unavailable, respectively. A summary feature about whether a gene was expressed before birth (TS1–TS26) was also included.

Mouse protein interaction data are quite incomplete; therefore, human protein interaction data from the Human Protein Reference Database (Keshava Prasad *et al*., 2009) were transferred to mouse genes through one-to-one human-to-mouse orthology, as annotated in Ensembl (Flicek *et al*., 2011). Based on the protein interaction network hereby obtained, betweenness, connectivity and clustering coefficients were calculated. Data about histone modifications (including two types, H3k4me1 and H3k4me3, in 10 tissues) were obtained from the ENCODE Project (Birney *et al*., 2007), and a histone modification peak was assigned to a gene if it was within a gene or 10-kb upstream from the gene start. The existence of histone modifications in each tissue and the total number of histone modifications across all the tissues were calculated for each gene.

2.2 Machine learning methods

We selected informative genomic features using the least absolute shrinkage and selection operator (LASSO) penalty by Tibshirani (1996). An important property of this penalty, generating coefficient estimates of exactly zero (Knight and Fu, 2000), makes it attractive for feature selection. LASSO reduces the estimation variance while providing an interpretable final model. Its application to genomic data (Ghosh and Chinnaiyan, 2005; Shevade and Keerthi, 2003) has shown that selecting a small number of representative features can achieve satisfactory classification. We applied the R package 'glmnet' (Friedman *et al*., 2010) to perform LASSO feature selection. We first obtained the regularization path over a grid of values, through 5-fold cross-validation (*n*-folds = 5), with the response variable as a binary factor (family = 'binomial'), standardize = FALSE (since all the numerical variables in the datasets had been pre-standardized) and the other parameters set by default. The minimum value of the tuning parameter lambda, which was obtained by a previous cross-validation procedure, was then used to fit the model. All the features with non-zero coefficients were retained for subsequent analyses.

Logistic regression is a generalized linear model for binomial regression that has been widely used to model the outcomes of a categorical dependent variable. We used the R package, fitting generalized linear models 'glm', to build logistic regression classifiers with standard settings.

SVM has shown promising empirical performance by providing non-linear boundaries in a large, transformed version of the feature space. According to Schöklkopf *et al*. (1997), SVM achieves higher recognition accuracy in classification compared with some other popular methods. In particular, kernel-based SVM is an easy and efficient way for such a mapping in a higher dimensional space. We used the R package 'e1071' to build the SVM classifiers. Since the effectiveness of SVM largely depends on the selection of kernel and model parameters, we explored the four most common kernel forms (linear, polynomial, Gaussian radial and sigmoid basis). It turned out that the radial kernel with degree 2 achieved the best results. The other parameters were set as defaults.

A random forest is a collection of decision trees such that each tree is built from a random subset of the data. The random forest technique was

first introduced in 2001 (Breiman, 2001), and since then has been shown to be a highly accurate classifier in a number of fields, including genetics (Bureau *et al*., 2005). One important reason we employed the random forest method in this study is because it effectively and validly accommodates a large number of features. This advantage was of particular interest in our study since there were 491 possible informative features. We used the 'randomForest' package in R and chose parameter values according to Breiman's methodology. Although the default number of trees is 500, we chose to build 5000 trees (*n*tree = 5000) to obtain more robust results. As more trees are added, it converges almost surely to the probability that the forest classifies the response correctly. Each tree was grown to its full depth (nodesize = 1) and was not pruned. At each node of each tree, a different random subset of the features was selected, and the Gini criterion was used to determine the feature in this subset that produced the best split of the data. The size of this subset (mtry) was the square root of the number of possible features. The other parameters were set as defaults.

The performance of the various classifiers was evaluated based on complete 5-fold cross-validation: during each of five iterations, feature selection (by LASSO as described earlier) and model selection was based on 80% of the training data and model evaluation was based on the remaining 20% of the data. Since the random forest classifier outperformed the other two in terms of AUC, the final prediction for genes without annotated phenotype was made with the random forest classifier. The predicted values were originally reported as probabilities, and to make binary predictions (0, non-lethal and 1, lethal), we chose the cutoff values by maximizing the overall classification accuracy.

2.3 Bias correction for the predictive power

To assess the bias between genes in the training and predicting sets in terms of a specific genomic feature, mutual information was calculated by the R module 'entropy' via binning, and the feature was treated as if it was a discrete variable. For each of the seven feature sets, the accuracy (Acc), positive predictive value (PPV), negative predictive value (NPV) and recall of a random forest classifier were calculated as follows: through five iterations in the cross-validation, a binary prediction was made on each gene in the training set; given the annotated knockout phenotypes, the 2 × 2 outcome contingency table was obtained, which contained the numbers of true positives (TPs), false positives (FPs), true negatives (TNs) and false negatives (FNs); Acc = (TP + TN)/(P + N), PPV = TP/(TP + FP), NPV = TN/(TN + FN), recall = TP/(TP + FN). The age-bias–corrected estimates were calculated as follows: $\text{Acc} = \Sigma \text{Acc}_i \times f_i$; $\text{PPV} = \Sigma \text{PPV}_i \times f_i$; $\text{NPV} = \Sigma \text{NPV}_i \times f_i$; $\text{recall} = \Sigma \text{recall}_i \times f_i$; $i = 1 \ldots 12$, where Acc_i, PPV_i, NPV_i and recall_i were the measurements for each evolutionary-age gene group in the training set, respectively, and f_i was the fraction of the corresponding gene group in the predicting set. The evolutionary age groups 1–6 in the training set had relatively small gene numbers, so they were combined as a single group in the analysis. For each gene in the predicting set, the phenotype was predicted according to the feature set with the best predictive power (S + I + E > S + I > S + E > S), and the final classifiers were based on the whole training set.

3 RESULTS

3.1 Identification of informative genomic features on the lethality of mouse knockout genes

From the MGI Database, we collected the mouse phenotypic data of targeted single-gene knockouts. Among 19 845 mouse protein-coding genes under survey, the 4670 genes with available knockout phenotypic data were used as the 'training set'; and the 15 175 genes without available knockout phenotype information were designated as the 'predicting set'. For each mouse gene, we compiled 491 genomic features (Table 1) whenever available, and grouped those features into the following three categories according to their data sources. The genomic sequence set (S) contains 373 features, of which eight are inferred from coding sequences, 30 from protein sequences and 335 are based on other gene properties, including the features for gene regulation. The mRNA expression set (E) was derived from two expression datasets, containing 66 features for the tissue expression profiles and 28 features for the developmental stage expression profiles. The interaction set (I) contains three features derived from inferred mouse protein–protein interaction data and 21 features about mouse epigenetic histone modification interactions.

In the training set, the knockout phenotypes of 1645 genes were classified as lethal and those of the remaining 3025 genes as non-lethal. Although genomic sequence features were available for all 4670 genes in the training set, the expression and interaction features were available for only some of the genes. We summarized the gene numbers for which each category of features was available as follows: 4670 genes have sequence features (defined as the S gene set); 4037 genes have expression features (E gene set); 3627 genes have protein interaction features (I gene set); 4036 genes have both sequence and expression features (S + E gene set); 3627 genes have both sequence and interaction features (S + I gene set); 3217 genes have both expression and interaction features (E + I gene set) and 3217 genes have sequence, expression and interaction features (S + E + I gene set). Supplementary Table S1 contains all the phenotype and feature information about the training set.

Among the features we compiled, some may be irrelevant to the knockout lethality, and some are highly correlated (Supplementary Fig. S1) and provide redundant information. Therefore, we next selected the subset of informative features for each of the above seven gene sets using LASSO, which is widely used to simultaneously achieve shrinkage and variable selection (Tibshirani, 1996). (Supplementary Table S2 shows selected features for each gene set). Table 2 shows the top informative genomic features selected by LASSO in the S + E + I gene set. When considering all the features together, the most informative features are some evolutionary features, such as gene evolutionary age and paralog sequence identity; the most informative expression feature is the expression level in utero, followed by the expression level in developmental stage 15; and the most informative interaction feature is protein connectivity. Interestingly, some of the top selected features, such as evolutionary age, were also highly ranked upon evaluation of their individual correlations with the knockout lethality; whereas others, such as paralog sequence identity, did not.

3.2 Assessing the phenotype predictive power of integrated feature assemblies

To address how well different genomic feature sets predict the lethality of knockout mice, we used three common machine learning methods (logistic regression, random forest and SVM) to make the probabilistic prediction as to whether a gene knockout would result in the lethal phenotype. Based on 5-fold cross-validation (including the LASSO feature selection step), we evaluated the predictive power of different feature assemblies by generating ROC and precision-recall curves (Supplementary Fig. S2). The area under the ROC curve (AUC) is a standard measure for the predictive power of different computational methods or different feature assemblies (the diagonal line in the ROC curve has an AUC value of 0.5, representing the predictive power of a random guess). The three

Table 1. Candidate genomic features for mouse knockout phenotype

Category		Feature	Number	Remark	Source/tools
Interaction (I)	Protein interaction				Human Protein
		Betweenness	1		Reference Database
		Connectivity	1		
		Clustering coefficient	1		
	Epigenetic interaction				The ENCODE Project
		Existence of histone modification (H3K4me1/H3K4me3 in 10 tissues)	20	Categorical (binary)	(UCSC Genome Browser)
		Histone modification sum	1		
mRNA expression (E)	Spatial expression				Su *et al.* (2004)
		Tissue expression (61 tissues)	61		
		Mean/median tissue expression	2		
		Maximum tissue expression level	1		
		Tissue expression breadth	1		
		Tissue expression specificity	1		
	Temporal expression	Developmental stage expression (stage TS1–TS26, TS28, before birth)	28	Categorical	MGI Gene Expression Database
Sequence (S)	Coding sequence				Ensembl (release 59)
		Hydrophobicity	1		
		Codon adaptation index (CAI)	1		CodonW
		Codon bias index (CBI)	1		CodonW
		Effective number of codons (Nc)	1		CodonW
		GC content	1		CodonW
		Number of codons	1		
		Special codon content	1		
		Frequency of optimal codons	1		CodonW
	Protein sequence				
		Amino acid content (20 amino acids)	20		
		Predicted secondary structure (helix, sheet, loop content)	3		PROFphd
		Number of domains	1		Biomart
		Existence of domains (Ncoils/signal/transmembrane)	3	Categorical (binary)	Biomart
		Isoelectric point (PI)	1		ExPAsy Proteomics
		Disordered regions in proteins	1		RONN
		Aromaticity	1		CodonW
	Gene				Ensembl (release 59)
		Number of paralogs	1		
		Paralog sequence identity	1		
		Selective pressure (d*N*/d*S* between mouse and rat)	1		
		Gene evolutionary age	1		Cai *et al.* (2009)
		Number of exons	1		
		Average intron size	1		
		Chromosome location	1	Categorical	
		UTR length (5′, 3′)	2		RefSeq transcripts
		CpG island	1		UCSC Genome
		Recombination rate	1		Cox *et al.* (2009)
		Existence of miRNA target sites (152 miRNA families)	152	Categorical (binary)	TargetScanS
		Total number of miRNA target sites	1		
		Existence of transcription factor biding sites (TFBS); 170 types	170	Categorical (binary)	TRANSFAC
		Total number of TFBS	1		
Total			491		

Table 2. Top 20 informative genomic features related to knockout lethality, as selected by LASSO

Features	LASSO coefficient[a]	Rank	Mutual information	Rank
Evolutionary age	0.473	1	4.76E-2	1
Expression in utero	0.436	2	2.15E-3	96
Expression in TS15	0.297	3	1.99E-2	4
Paralog sequence identity	−0.296	4	2.51E-3	85
Total miRNA target sites	0.254	5	9.22E-3	35
Expression in TS11	0.196	6	9.35E-3	33
Connectivity	0.160	7	7.03E-3	48
Expression in TS17	0.146	8	1.96E-2	6
Expression in TS26	0.132	9	1.02E-2	29
Expression in TS18	0.129	10	7.05E-3	47
Total histone modification	0.128	11	2.85E-2	2
Asparagine content	0.113	12	7.35E-3	44
5′-UTR length	−0.110	13	5.08E-4	266
Expression in upper spinal cord	−0.110	14	9.70E-4	171
Leucine content	−0.106	15	1.63E-2	9
Expression in bone	−0.104	16	3.67E-4	307
Expression in TS5	0.100	17	9.83E-3	31
Amino acid length	0.093	18	5.07E-4	267
Expression in ovary	0.089	19	3.08E-3	75
Expression in TS19	0.086	20	1.18E-2	22

[a]Based on the S + E + I gene set (genes with sequence, expression and interaction features).

machine learning methods produced the same trend and showed similar predictive powers across all the feature sets (Fig. 2A). With an increase in selected features, the random forest method outperformed the other two methods (Fig. 2B). Since the random forest method achieved a slightly better performance than the other methods in most cases, we focused on the results by the random forest method in the subsequent analyses.

As shown in Figure 2A, the interaction features alone (AUC_I = 0.598) have limited predictive power for knockout lethality; whereas the sequence features or the expression features showed much better predictive power (AUC_S = 0.738; AUC_E = 0.725). Adding more features to the prediction model consistently improved its performance (AUC_{S+E} = 0.776, AUC_{S+E+I} = 0.782). For each of the seven feature sets, we found that the AUC value of the top two informative features is considerably lower than that of the combination of all the selected features (Fig. 2C). These results indicate that the integrated feature assembly has a substantially increased predictive power for mouse knockout lethality.

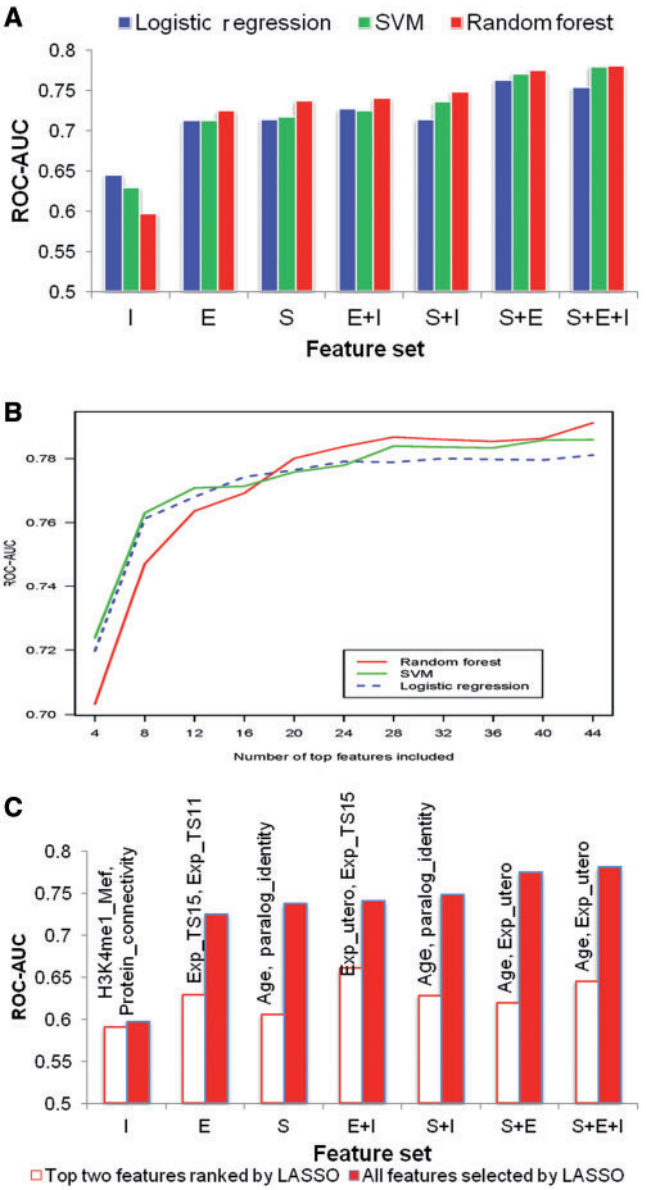

Fig. 2. Predictive power of different feature assemblies and different methods for mouse knockout phenotype. (**A**) Comparison of different machine learning methods across different feature sets; (**B**) Comparison of different methods with an increase in the top selected features (based on S + E + I gene set); and (**C**) Comparison of top two selected features versus combination of all selected features across different gene sets by random forest.

3.3 Estimating the bias-corrected predictive power for the predicting set

If the training set was randomly selected from the mouse genome, we can estimate the predictive power of our models for genes whose knockout phenotypes have not been reported (the predicting set) directly based on cross-validation of the training set. However, this assumption may not be true because the training set was largely collected from individual studies. We therefore evaluated the bias of the surveyed features between genes in the training and predicting sets. We found that evolutionary age shows a much greater bias than any other feature (Fig. 3A, Wilcoxon rank test, $P = 2.2 \times 10^{-16}$). Indeed, the proportions of genes with an ancient evolutionary age are much higher in the training set than in the predicting set (Fig. 3B). Such a bias is not surprising since knockout experiments are time-consuming and researchers tend to choose highly conserved genes for functional investigation.

To more accurately estimate the performance of our random forest classifiers on the predicting set, we performed a bias-correction step upon the raw estimates of the cross-validation (see details in Section 2). As shown in Table 3, considering both accuracy

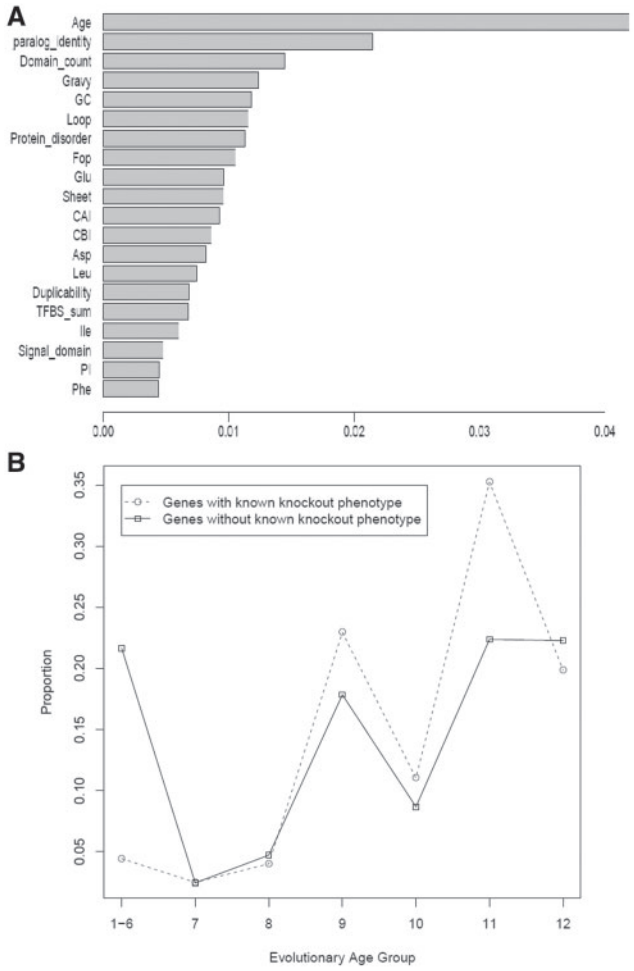

Fig. 3. The bias between genes in the training and predicting sets. (**A**) The bias of a genomic feature was quantified by mutual information based on the S gene set; and (**B**) the distribution of different evolutionary-age gene groups in the training and predicting sets.

and recall, without correcting the bias of evolutionary age, the S + E + I feature set shows the best performance (accuracy = 72.5% and recall = 62.9%), and the I feature set shows the lowest predictive power. After correcting the bias, the predictive power across different feature assemblies remains similar: the S + I + E feature set still shows the best performance (accuracy = 70.9% and recall = 60.7%), followed by the S + I feature set. We predicted the knockout phenotype (lethal or non-lethal) for 15 175 genes in the predicting set, according to the available feature set with the best predictive power for the gene (Supplementary Table S3).

4 DISCUSSION

In this study, we predicted the lethal phenotype for single-gene knockout mice by integrating gene features inferred from various genomic data sources and achieved a relatively good performance. Since mouse knockout experiments are notoriously time-consuming and technically challenging, we expect our results to serve as a valuable resource for the mouse genetics research community, greatly helping to optimize experimental designs and improve biological interpretations. Because of the close genetic and physiological similarities between mice and humans, our results would accelerate the translation of the knowledge of mouse genes into better strategies for studying human disease.

Compared with previous similar studies, our study has some advantages. First, we systematically investigated a wide range of candidate genomic features, most of which have not been examined for their ability to predict the lethal phenotype. Importantly, all the surveyed features were derived from large-scale independent experimental data. To avoid circular reasoning, we did not use GO annotations or phenotype information (e.g. disease association) from other species by orthology. Second, we explored and compared different classifiers before making final predictions, thereby likely maximizing the predictive power. Third, we explicitly considered the bias in the training set when assessing the predictive power for genes without an annotated phenotype.

Our work provides key insights into the genotype-to-phenotype relationship. First, the top selected features for mouse knockout lethality are quite different from those for unicellular eukaryotes (i.e. yeasts). Some dominant predictors for yeast essential genes, such as GC content (Seringhaus *et al.*, 2006) were not selected in our models, suggesting that the genotype-to-phenotype relationship strongly depends on the organismal complexity. Second, many of the features that were highly ranked in our study are compatible with those of previous studies. Among these selected features, gene evolutionary age was the top feature; it was consistently selected across different gene sets, and also showed the highest individual correlation with the knockout phenotype. This result suggests that disrupting an evolutionarily ancient gene tends to result in a more severe functional consequence. Paralog sequence identity is the selected feature with the highest negative coefficient, indicating that removing a gene with a closely related gene copy tends to be non-lethal, which is consistent with the notion of 'functional compensation' among paralogs. Interestingly, this feature alone shows little correlation with the knockout phenotype individually, which is due to the confounding effects of other features, as discussed previously (Liang and Li, 2009). Protein connectivity was another highly ranked feature with a positive contribution to the knockout lethality. Consistent with previous studies (Jeong *et al.*, 2001), this result highlights the critical role of hub nodes in a biological system (Barabasi and Oltvai, 2004). In addition to those previously studied features, we also identified some important novel features, most of which are related to gene expression, such as the expression level in utero and in TS15.

The approach we took also has some limitations. First, like other similar studies, our models were unable to distinguish correlations from causations. Second, some parameters in our machine learning algorithms may not be fully explored. Third, we may still miss some important genomic features, and the feature dataset we compiled may contain some noise, especially for interaction features. Indeed, comparing among the three types of features, the sequence features and expression features show much higher predictive power than the interaction features. Although we did not explicitly model the noise within the feature data, the effect of noise was minimized by considering a comprehensive list of features and employing feature selection. We expect to revisit this topic and to achieve a better performance when more high-quality genomic datasets (such as protein expression data) have become available.

Table 3. Summary of the predictive power of different feature sets using random forest classifiers

Feature set	Select features	Raw estimates through cross-validation					Bias-corrected estimates[a]					Predictable genes #
		Cut-off	Acc	PPV	NPV	Recall	Cut-off	Acc	PPV	NPV	Recall	
S	36	0.48	0.709	0.624	0.737	0.438	0.47	0.744	0.568	0.756	0.331	15 175
S + E	57	0.54	0.727	0.703	0.733	0.391	0.54	0.745	0.646	0.740	0.315	9564
S + I	37	0.49	0.706	0.636	0.738	0.528	0.49	0.696	0.600	0.706	0.499	5187
S + E + I	44	0.46	0.725	0.637	0.777	0.629	0.46	0.709	0.642	0.738	0.607	3965
E	24	0.49	0.705	0.615	0.734	0.432	0.55	0.720	0.599	0.725	0.296	9573
I	20	0.58	0.644	0.598	0.650	0.184	0.49	0.628	0.551	0.634	0.291	5188
E + I	27	0.57	0.697	0.720	0.693	0.323	0.54	0.679	0.670	0.667	0.371	3965

[a]The bias of evolutionary age between the training set and the prediction set is corrected. S, sequence features; S + E, sequence and expression features; S + I, sequence and interaction features; S + E + I, sequence, expression and interaction features; E, expression features; I, interaction features; E + I, expression and interaction features. Accuracy (Acc) is defined as the proportion of genes whose knockout phenotype classification is correctly predicted; PPV, positive predictive value, also known as precision, is defined as the proportion of genes with a positive prediction (lethal) that are correctly predicted; NPV, negative predictive value is defined as the proportion of genes with a negative prediction (non-lethal) that are correctly predicted; recall is defined as the proportion of genes with a lethal phenotype that is correctly predicted.

ACKNOWLEDGEMENTS

We would like to thank LeeAnn Chastain for editorial assistance; Dr Karl Broman (University of Wisconsin at Madison) for providing mouse recombination rate data and the ENCODE Project for making mouse histone modification data publicly available.

Funding: National Institutes of Health (Grants CA143883 and CA016672).

Conflict of Interest: none declared.

REFERENCES

Barabasi,A.L. and Oltvai,Z.N. (2004) Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.*, **5**, 101–113.

Birney,E. *et al*. (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, **447**, 799–816.

Breiman,L. (2001) Random forests. *Machine Learn.* **45**, 5–32.

Bureau,A. *et al*. (2005) Identifying SNPs predictive of phenotype using random forests. *Genet. Epidemiol.*, **28**, 171–182.

Cai,J.J. *et al*. (2009) Similarly strong purifying selection acts on human disease genes of all evolutionary ages. *Genome Biol. Evol.*, **1**, 131–144.

Collins,F.S. *et al*. (2007) A mouse for all reasons. *Cell*, **128**, 9–13.

Cox,A. *et al*. (2009) A new standard genetic map for the laboratory mouse. *Genetics*, **182**, 1335–1344.

Finger,J.H. *et al*. (2011) The mouse Gene Expression Database (GXD): 2011 update. *Nucleic Acids Res.*, **39**, D835–D841.

Flicek,P. *et al*. (2011) Ensembl 2011. *Nucleic Acids Res.*, **39**, D800–D806.

Friedman,J. *et al*. (2010) Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.*, **33**, 1–22.

Gasteiger,E. *et al*. (2003) ExPASy: the proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Res.*, **31**, 3784–3788.

Ghosh,D. and Chinnaiyan,A.M. (2005) Classification and selection of biomarkers in genomic data using LASSO. *J Biomed. Biotechnol.*, **2005**, 147–154.

Gu,Z. *et al*. (2003) Role of duplicate genes in genetic robustness against null mutations. *Nature*, **421**, 63–66.

Guan,Y. *et al*. (2008) Predicting gene function in a hierarchical context with an ensemble of classifiers. *Genome Biol.*, **9** (Suppl. 1), S3.

Hughes,T.R. and Roth,F.P. (2008) A race through the maze of genomic evidence. *Genome Biol.*, **9** (Suppl. 1), S1.

Jeong,H. *et al*. (2001) Lethality and centrality in protein networks. *Nature*, **411**, 41–42.

Keshava Prasad,T.S. *et al*. (2009) Human Protein Reference Database–2009 update. *Nucleic Acids Res.*, **37**, D767–D772.

Knight,K. and Fu,W. (2000) Asymptotics for Lasso-type estimators. *Ann. Stat.*, **28**, 1356–1378.

Lewis,B.P. *et al*. (2005) Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, **120**, 15–20.

Liang,H. and Li,W.H. (2007) Gene essentiality, gene duplicability and protein connectivity in human and mouse. *Trends Genet.*, **23**, 375–378.

Liang,H. and Li,W.H. (2009) Functional compensation by duplicated genes in mouse. *Trends Genet.*, **25**, 441–442.

Liao,B.Y. and Zhang,J. (2007) Mouse duplicate genes are as essential as singletons. *Trends Genet.*, **23**, 378–381.

Liao,B.Y. and Zhang,J. (2008) Null mutations in human and mouse orthologs frequently result in different phenotypes. *Proc. Natl Acad. Sci. USA*, **105**, 6987–6992.

Liao,B.Y. *et al*. (2006) Impacts of gene essentiality, expression pattern, and gene compactness on the evolutionary rate of mammalian proteins. *Mol. Biol. Evol.*, **23**, 2072–2080.

Makino,T. *et al*. (2009) The complex relationship of gene duplication and essentiality. *Trends Genet.*, **25**, 152–155.

Mostafavi,S. *et al*. (2008) GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function. *Genome Biol.*, **9** (Suppl. 1), S4.

Peden,J.F. (1999) *Analysis of Codon Usage*. Department of Genetics, University of Nottingham, Nottingham, United Kingdom.

Pena-Castillo,L. *et al*. (2008) A critical assessment of Mus musculus gene function prediction using integrated genomic evidence. *Genome Biol.*, **9** (Suppl. 1), S2.

Rost,B. and Sander,C. (1994) Combining evolutionary information and neural networks to predict protein secondary structure. *Proteins*, **19**, 55–72.

Schöklkopf,B. *et al*. (1997) Comparing support vector machines with Gaussian kernels to radial basis function classi?ers. *IEEE Trans. Signal Process.*, **45**, 2758–2765.

Seringhaus,M. *et al*. (2006) Predicting essential genes in fungal genomes. *Genome Res.*, **16**, 1126–1135.

Shevade,S.K. and Keerthi,S.S. (2003) A simple and efficient algorithm for gene selection using sparse logistic regression. *Bioinformatics*, **19**, 2246–2253.

Su,A.I. *et al*. (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl Acad. Sci. USA*, **101**, 6062–6067.

Tasan,M. *et al*. (2008) An en masse phenotype and function prediction system for Mus musculus. *Genome Biol.*, **9** (Suppl. 1), S8.

Tibshirani,R. (1996) Regression shrinkage and selection via the Lasso. *J. R. Stat. Soc. B*, **58**, 267–288.

Yang,Z.R. *et al*. (2005) RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins. *Bioinformatics*, **21**, 3369–3376.