© 2018. Published by The Company of Biologists Ltd | Disease Models & Mechanisms (2018) 11, dmm034546. doi:10.1242/dmm.034546

RESOURCE ARTICLE

Identifying mouse developmental essential genes using machine learning

David Tian[1,*,¶], Stephanie Wenlock[1,‡,¶], Mitra Kabir[1], George Tzotzos[2,§], Andrew J. Doig[3,4,] and Kathryn E. Hentges[1,**]**

ABSTRACT

The genes that are required for organismal survival are annotated as 'essential genes'. Identifying all the essential genes of an animal species can reveal critical functions that are needed during the development of the organism. To inform studies on mouse development, we developed a supervised machine learning classifier based on phenotype data from mouse knockout experiments. We used this classifier to predict the essentiality of mouse genes lacking experimental data. Validation of our predictions against a blind test set of recent mouse knockout experimental data indicated a high level of accuracy (>80%). We also validated our predictions for other mouse mutagenesis methodologies, demonstrating that the predictions are accurate for lethal phenotypes isolated in random chemical mutagenesis screens and embryonic stem cell screens. The biological functions that are enriched in essential and non-essential genes have been identified, showing that essential genes tend to encode intracellular proteins that interact with nucleic acids. The genome distribution of predicted essential and non-essential genes was analysed, demonstrating that the density of essential genes varies throughout the genome. A comparison with human essential and non-essential genes was performed, revealing conservation between human and mouse gene essentiality status. Our genome-wide predictions of mouse essential genes will be of value for the planning of mouse knockout experiments and phenotyping assays, for understanding the functional processes required during mouse development, and for the prioritisation of disease candidate genes identified in human genome and exome sequence datasets.

[1]Division of Evolution and Genomic Sciences, Faculty of Biology, Medicine and Health, Manchester Academic Health Science Centre, The University of Manchester, Oxford Road, Manchester M13 9PT, UK. [2]Department of Agriculture, Food and Environmental Sciences, Marche Polytechnic University, Ancona 60121, Italy. [3]Manchester Institute of Biotechnology, The University of Manchester, 131 Princess Street, Manchester M1 7DN, UK. [4]Division of Neuroscience and Experimental Psychology, Faculty of Biology, Medicine and Health, The University of Manchester, Manchester M13 9PT, UK.
*Present address: School of Computing, Creative Technologies and Engineering, Leeds Beckett University, Leeds LS1 3HE, UK. ‡Present address: Department of Pathology, Cambridge Genomic Services, University of Cambridge, Cambridge CB2 1TN, UK. §Present address: Ferrogasse 27, 1180 Vienna, Austria.

¶These authors contributed equally to this work

**Authors for correspondence (andrew.doig@manchester.ac.uk; Kathryn.Hentges@manchester.ac.uk)

G.T., 0000-0001-9258-4338; A.J.D., 0000-0003-0346-2270; K.E.H., 0000-0001-8917-3765

Received 14 March 2018; Accepted 19 October 2018

KEY WORDS: Essential genes, Supervised machine learning, Mouse knockout, Essentiality database

INTRODUCTION

Essential genes are those that are required for the survival of an organism. Although studies in unicellular organisms, such as yeast, have experimentally defined the set of essential genes in those species (Kofoed et al., 2015), the large genome size and developmental complexity of animal models have hindered a comprehensive experimental essentiality analysis in these organisms. Knowledge of essential genes in animal species is informative for understanding the biological functions required during development, as well as for identifying candidate genes for human genetic diseases. In particular, the mouse has been a long-standing model for human disease research due to the ability to generate specific genome alterations in mouse embryonic stem cells, allowing the targeted deletion or knockout of individual genes. Mouse knockout experiments have proved useful in identifying a subset of mammalian essential genes (Sung et al., 2012); however, the entirety of the mouse genome has not yet been experimentally examined.

Current efforts to experimentally investigate gene function using mouse models are enhanced by the creation of the International Knockout Mouse Consortium (IKMC) (Bradley et al., 2012), a large global project with the goal of generating knockouts for over 20,000 protein-coding mouse genes. The International Mouse Phenotyping Consortium (IMPC) (Ayadi et al., 2012; Brown and Moore, 2012) builds upon the efforts of IKMC to discover functional insights for every gene by systematically phenotyping over 20,000 knockout mouse strains. In order to optimise knockout experiment design, machine learning algorithms (Yuan et al., 2012) have been used to predict the essentialities of mouse genes based on their genomic features. Moreover, predicting the essentialities of mouse genes using machine learning algorithms can aid in the identification of candidate genes for human genetic diseases, due to the close genetic and physiological similarities between mouse and human (Rosenthal and Brown, 2007). Machine learning methods are also useful in identifying features associated with gene essentiality (Kabir et al., 2017).

A variety of machine learning methodologies have proven useful in predicting essential genes in several organisms. Many studies have sought to identify bacterial and fungal essential genes, because knowledge of gene essentiality in microbial species can reveal potential drug targets (Yu et al., 2017; Hua et al., 2016; Deng, 2015; Ning et al., 2014; Lu et al., 2014; Cheng et al., 2014; Cheng et al., 2013; Deng et al., 2011; Plaimas et al., 2010; Seringhaus et al., 2006; Gustafson et al., 2006; Liu et al., 2017; Nigatu et al., 2017). *Saccharomyces cerevisiae* essential genes have been identified using machine learning classifiers trained on multiple characteristics of protein function, such as physical, metabolic and transcriptional regulatory interactions, gene expression patterns and annotated biological functions (Acencio and Lemke, 2009; Zhong et al., 2013;

Disease Models & Mechanisms (2018) 11, dmm034546. doi:10.1242/dmm.034546

Hwang et al., 2009). Protein interaction network topologies have also been utilised for the prediction of human essential genes (Yang et al., 2014). The lack of functional annotation of the majority of plant genes, and the long generation time required for experimental analysis of mutant plant phenotypes, provided the motivation to implement a random forest machine learning algorithm for the prediction of *Arabidopsis thaliana* essential genes (Lloyd et al., 2015); similar challenges underlie the identification of mammalian essential genes.

In order to provide insights into the gene functions required during mammalian development, we identified a dataset of genes needed for a mouse embryo to survive until the postnatal period, which we define as essential genes (Kabir et al., 2017). Here, we implement a supervised machine learning approach to generate an essentiality classifier, testing a variety of machine learning methods. We found that random forests provided the most accurate classifier and, following feature selection, achieved classification accuracy of greater than 95% during 10-fold cross-validation. The accuracy of our classifier was also assessed against 2 blind test sets, and over 80% accuracy was achieved on these datasets. The classifier was then used to predict the essentiality of the remaining protein-coding genes in the mouse genome. Functions linked to each essentiality class were identified, and the transferability of our classifications was determined by comparing our predictions with experimental data from mouse mutants generated through non-knockout experimental methods and human gene essentiality annotations. We conclude that our predictions have a high degree of accuracy, and thus could facilitate mouse knockout experimental design and contribute to a deeper understanding of biological functions that are essential for mammalian development.

RESULTS

Training and test sets

Manually curated datasets containing 1307 essential genes (those with pre- or perinatal lethal phenotypes in mouse knockout experiments) and 3459 non-essential genes (those with viable phenotypes in mouse knockout experiments) (Kabir et al., 2017) were used as the input to our classifier. In total, 102 features (Tables S1 and S2) were identified from multiple public databases as characteristics that might distinguish between essential and non-essential genes. In total, 75 of the 102 features analysed had statistically significant differences in values between genes in the essential and non-essential training sets (Kabir et al., 2017). Owing to the large number of features with distinct values, we hypothesised that essential and non-essential genes could be differentiated by their properties. We therefore sought to test a variety of machine learning methods to identify the most accurate approach to categorise genes as essential or non-essential. Our original dataset is an imbalanced dataset as the number of non-essential genes is much larger than the number of essential genes. Imbalanced datasets can degrade the classification performance of machine learning classifiers due to their bias towards classifying instances belonging to the majority class (Visa and Ralescu, 2005). Therefore, to develop a machine learning classifier, we generated balanced training sets containing all 1307 essential genes, and 1307 non-essential genes selected at random from the total set of 3459 non-essential genes (Table S3). To remove possible bias, this process was repeated 10 times in order to generate 10 different balanced training datasets containing different sets of non-essential mouse genes (Table S3). We further developed 10 random forest classifiers by implementing 10-fold cross-validation on these training datasets, utilising all features. We found a very small range in the cross-validation accuracies (89.89-91.42%) (Table S4), showing that the choice of genes in the training datasets had little effect. The mean accuracy of these classifiers was 90.90%; therefore, we selected the training dataset that had an accuracy of 90.85% for all further experiments, as this was closest to this mean value. We might have overestimated the overall performance of our classifier if we selected a training dataset for which the cross-validation accuracy was more than the mean value.

In order to evaluate the accuracy of the machine learning classifiers, we assembled test sets. Test set 1 (Table S3) contained 229 essential and 802 non-essential genes, the essentiality status of which was published by the IMPC either in the literature or via their website (Koscielny et al., 2013) after our training sets were compiled. Test set 2 (Table S3) was formed of the 2152 genes in our original non-essential gene dataset that were not incorporated into the balanced essential and non-essential training sets. Test set 4 contained 169 lethal and 441 viable genes, which were added to the IMPC database at the conclusion of the project (April 2018), and were not already included in our training datasets or in Test sets 1 and 2. Test set genes were not used in classifier training.

We also compiled a prediction dataset containing all genes in the mouse genome with no experimental essentiality annotations (Test set 3). MouseMine (Motenko et al., 2015) was used to retrieve all known mouse genes. In total, 22,944 protein-coding mouse genes were identified. After excluding genes with known essentiality that are included in training and test sets, and removing non-mouse genes and duplicate gene names from the MouseMine dataset, 15,495 unique protein-coding genes with unknown essentiality status remained in Test set 3 (Table S3). All the features previously collected for training set genes were then collected for test set genes, following the same methodology used for compiling training set features (Kabir et al., 2017).

Data pre-processing

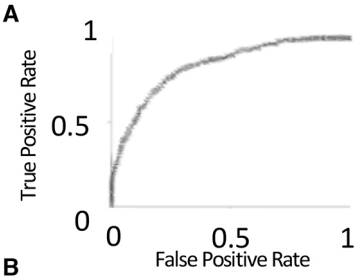
We found that there were no data available for several features for genes in the training and test datasets. We found that 10 features of the protein-protein interaction (PPI) network compiled from known PPIs had missing values for nearly 40% of the genes in the training set, so these features (Table S2) were removed from classifier training. The other 92 features had missing values for fewer than 12% of the genes. For classifier training, the missing values of these features were replaced with the feature mean values. Following the replacement of missing values, features within the training datasets were discretised using the ChiMerge algorithm (Kerber, 1992).

Classifier optimisation

An iterative process was used to test 6 different supervised machine learning classifiers. We assayed random forests, support vector machines (SVMs) with radial basis function (RBF) kernel, polynomial kernel SVMs, logistic regression, naïve Bayes classifier and decision tree classifiers in 10-fold cross-validation on the discretised training sets. We applied information gain feature selection (Yang and Pederson, 1997), and found that 83 features had an information gain greater than 0 (Table S4). These 83 features were ranked in order of significance. Classifiers were tested using increasing numbers of features (ranging from 5 to 83 features) for 10-fold cross-validation on the training sets (Table S4). From these studies, we found that the random forest classifier trained with 80 features had the best performance in 10-fold cross-validation. Using a random forest with 230 trees, we generated a 10-fold cross-validation accuracy of 98.1%. This classifier reached 79.3% accuracy on Test set 1, and the area under the curve (AUC) value of the corresponding receiver operating characteristic (ROC) plot

Disease Models & Mechanisms (2018) 11, dmm034546. doi:10.1242/dmm.034546

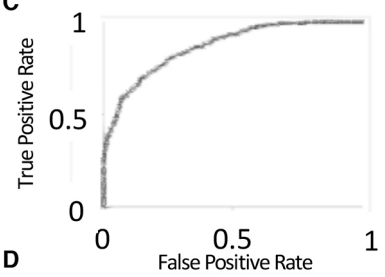
A



B

Predicted Essential	Predicted Non-Essential	
170	59	Known Essential
178	624	Known Non-Essential

C



D

Predicted Essential	Predicted Non-Essential	
170	59	Known Essential
157	645	Known Non-Essential

Fig. 1. Prediction accuracies of the random forest classifiers. Prediction accuracies of the random forest classifiers. (A) ROC plot with AUC 0.803 for the random forest classifier trained on 80 features and tested on Test set 1. (B) Confusion matrix of the random forest classifier trained on 80 features and tested on Test set 1. (C) ROC plot with AUC 0.816 for the random forest classifier trained on the 39 features selected by the genetic algorithm feature selection and tested on blind test set 1. (D) Confusion matrix of the random forest classifier trained on the 39 features selected by the genetic algorithm feature selection and tested on blind test set 1.

was 0.85 (Fig. 1A). A confusion matrix shows that this classifier predicted 59 known essential genes to have a non-essential function, and 178 known non-essential genes to have essential functions (Fig. 1B). This random forest classifier had an accuracy of 85% on Test set 2. Because Test set 1 contains both essential and non-essential genes, we chose the classifier with the best performance on Test set 1 for further studies. None of the other machine learning methods tested achieved a higher AUC on Test set 1 than the random forest classifier (Fig. 2; Table S4), so the random forest method was used henceforth.

We sought to improve the performance of the random forest classifier by implementing feature selection. When implementing a classifier, an individual feature can be irrelevant, strongly relevant (removal of this reduces the overall prediction accuracy) or weakly relevant (not sufficient alone for prediction). Feature selection, therefore, is a very important stage for the classification problem when using datasets comprised of a large number features, in order to select the most informative features and remove those that simply add noise and thus weaken a predictor. A genetic algorithm (GA) feature selection method (Witten et al., 2011) was applied on the training sets as an alternative method to determine whether a smaller set of features would result in random forests with increased prediction accuracy. The GA found a subset of 39 features (Table S4) after 20 generations that improved the classifier performance. These 39 features belong to 9 types: features of the PPI network representing known PPIs and predicted PPIs, features of the PPI network representing known PPIs only, amino acid content of proteins, gene expression, protein types, subcellular localisation, predicted subcellular localisation and enzyme classes. The PPI network features are ranked highest by information gain, which measures the relevance of a feature, and are the most informative features for predicting the essentiality of protein-coding mouse genes. Notably, features such as gene length, GC content, evolutionary age, presence of transmembrane domains and all Gene Ontology (GO) annotations, which we previously identified as statistically different in their distribution between essential and non-essential genes (Kabir et al., 2017), were not found to improve classifier accuracy and were not incorporated into further classifier training. One reason for this surprising result is that the information in these features could be related to or dependent upon information found in other features, so their inclusion adds no value to the classifier. For example, gene length is not needed if protein length is present.

A random forest classifier was subsequently trained on the 39 features identified from GA feature selection, yielding an improved ROC plot AUC of 0.816 on blind test set 1 (Fig. 1C). The random forest has a true-positive class of 170 instances, true-negative class of 645 instances, false-positive class of 157 instances and false-negative class of 59 instances (Fig. 1D). These results are an improvement over a prior study predicting the essentiality of mouse genes (Yuan et al., 2012). On Test set 2, which only contains non-essential genes, the random forest classifier trained on all 92 features had an accuracy of 80.1%. Following GA feature selection, the random forest classifier trained on 39 features showed an accuracy of 79.9% on Test set 2, showing very little decline in accuracy despite the removal of many features, which allows for increased speed of classification. We formed an additional blind test set of mouse knockout phenotypes published by the IMPC in April 2018 (Test set 4, Table S3). Genes already included in our training sets or Test sets 1 and 2 were excluded from Test set 4. Our random forests classifier trained on 39 features produced accurate predictions for 72% of genes with reported lethal phenotypes and 71% of genes with reported viable phenotypes in Test set 4, consistent with our findings from Test set 1, which included IMPC data reported prior to 2018.

We also compared the overlap between our known essential and non-essential genes obtained from searches of the Mouse Genome Informatics (MGI) database and data released by the IMPC (Koscielny et al., 2013). We found a total of 4752 genes in MGI with essentiality data (Table S4). Of these genes, 3467 have not been tested by the IMPC. In comparing the essentiality annotations for each gene with known essentiality, we did find mismatches between the MGI classifications and IMPC classifications. The percentage of mismatches is greatest for genes classified as essential in MGI and as non-essential by the IMPC. A significant proportion of genes falling into this mismatch category have multiple alleles described in MGI, including both essential and non-essential alleles (owing to experimental differences in gene targeting strategy or strain background); in the IMPC, the phenotype analysis of a single

Disease Models & Mechanisms (2018) 11, dmm034546. doi:10.1242/dmm.034546

A Random forest

Predicted Essential	Predicted Non-essential	
170	59	Actual Essential
178	624	Actual Non-essential

B RBF Kernel SVM

Predicted Essential	Predicted Non-essential	
138	91	Actual Essential
197	605	Actual Non-essential

C Linear SVM

Predicted Essential	Predicted Non-essential	
141	88	Actual Essential
216	586	Actual Non-essential

D Logistic regression

Predicted Essential	Predicted Non-essential	
113	116	Actual Essential
192	610	Actual Non-essential

E Naïve Bayes

Predicted Essential	Predicted Non-essential	
163	66	Actual Essential
210	592	Actual Non-essential

F Decision Tree

Predicted Essential	Predicted Non-essential	
65	164	Actual Essential
54	748	Actual Non-essential

Fig. 2. Confusion matrices of the 6 classifiers trained on all 83 features. The machine learning algorithm is listed at the top of each chart: (A) random forest; (B) RBF kernel SVM; (C) linear SVM; (D) logistic regression; (E) naïve Bayes; (F) decision tree.

allele has been reported. We calculate that ∼20% of genes with mismatching essentiality status between MGI and the IMPC have variations in the phenotypes produced due to the existence of multiple knockout experiments. Additionally, the IMPC classifies some genes as subviable, defined as genes with knockout alleles whereby homozygous null pups comprise less than 12.5% of a litter (Koscielny et al., 2013), which is a category that we did not include in our essentiality definitions. Of the 432 subviable genes listed in IMPC, 109 are found in our training sets compiled from MGI. Of these 109 genes, ∼20% were contained within our essential gene training set, with the remaining 80% in our non-essential gene sets. Approximately 92% of the subviable genes found within our essential genes training set had additional experimental alleles reported in MGI, which met our definition of essential genes (Table S5). Based on our analysis of the discrepancies between MGI and IMPC data, we predict that as many as 20% of genes will display conflicting essentiality phenotypes depending upon the experimental analysis performed.

Essentiality predictions

Based on the accurate predictions of genes in Test sets 1 and 2, we used the random forest classifier trained on 39 features (identified from genetic algorithm feature selection) to predict the essentiality status of the remainder of mouse protein-coding genes with no experimental annotations (Table S3). Using this classifier, we found that 28% of genes in the genome are known or predicted essential genes, and 72% of genes in the genome are known or predicted non-essential genes, percentages consistent with mouse knockout experimental results (White et al., 2013; Dickinson et al., 2016). The confidence level for each gene essentiality prediction was determined as a measure of whether or not the prediction is accurate. The confidence level is the fraction of the trees of the random forest that predict an essential gene to be essential, or the fraction of trees that predict a non-essential gene as non-essential. A confidence level of 1 indicates that 100% of trees had the same essentiality status prediction. The confidence levels of the predictions of essential genes are between 0.5 and 0.88, with 1 as the maximum confidence and 0.5 as the minimum confidence. The mean confidence level of essential gene predictions is 0.65. The confidence levels of non-essential gene predictions are between 0.5 and 0.95, with the mean confidence level of non-essential gene predictions being 0.65.

Applicability to point mutation phenotypes

We compared the accuracy of our predictions with experimental data generated by alternative mouse mutagenesis methodologies aside from targeted gene deletions. Data were collected from the MGI database (version 6.07) (Bult et al., 2016), using the search terms ‘Viable’ and ‘Lethal’ and specifying ‘Null/Knockout alleles’, with all chromosomes and generation methods selected other than ‘Targeted’, ‘Transgenic’ and ‘QTL’. We excluded targeted alleles because these are already in our training sets. We excluded transgenic alleles as some of these experiments assess overexpression or misexpression of genes, which are not directly comparable to the null alleles contained in our training sets. Finally, we excluded QTL alleles because these are not single gene effects. The search returned 201 essential genes and 29 non-essential genes. Duplicate entries, genes included in our test sets or genes found in our training sets were excluded from the analysis. Some genes were retrieved from both the essential and non-essential searches; these genes were categorised as either essential or non-essential following consultation of published phenotypes. Our final alternative mutagenesis method dataset included 116 essential and non-essential query genes, with allele types of ‘Gene trapped’, ‘Transposon induced’, ‘Chemically induced’, ‘Spontaneous’ or ‘Endonuclease mediated’, which were checked against our classifier predictions (Table S3). In 72% of cases, the essentiality classifier predicted the correct essentiality of the query genes, with 32 out of 116 genes being incorrectly predicted. The average prediction confidence level for incorrectly predicted genes was 0.608, with the mean confidence level for correct predictions being 0.647 (Fig. 3A). The difference in confidence levels between correct and incorrect predictions was significant (Welch’s 2-sample t-test, P=0.0166), confirming that incorrectly predicted genes had lower confidence predictions and correctly predicted genes had higher prediction confidence levels. We also compared the prediction confidence for Test set 1 genes, and found a similar trend within both the essential and non-essential gene predictions, such that incorrect predictions were of significantly lower confidence than correct predictions (Fig. 3B,C).

Disease Models & Mechanisms (2018) 11, dmm034546. doi:10.1242/dmm.034546

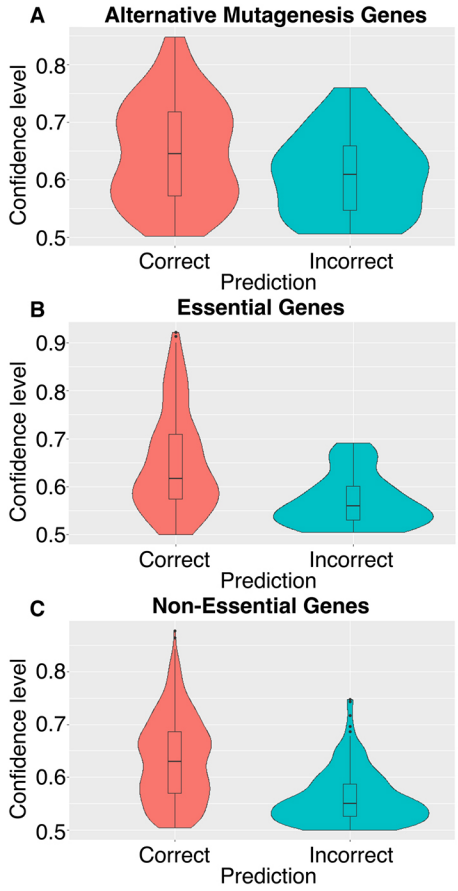


Fig. 3. Differences in 'Essentiality' gene prediction confidence levels for experimentally validated blind and alternative mutagenesis mouse genes. (A-C) A Normal distribution was confirmed for alternative mutagenesis data (n=115 genes) using Shapiro–Wilk test. Welch's 2-sample t-test identified a significant difference between correct and incorrect prediction confidence-levels (P=0.0166) for predictions of alternative mutagenesis genes (A). Both essential (n=229 genes) and non-essential (n=802 genes) blind test set 1 data were not normally distributed (Shapiro–Wilk test). Using Wilcoxon's Rank-Sum 2-sided test, significant differences were found between prediction confidence levels of correct and incorrect predictions for essential (B) and non-essential (C) blind test set 1 genes (P=1.75×10^{-7} and P≤2.2×10^{-16}, respectively).

Thus, we conclude that our classifier predicts essentialities of genes that have been experimentally determined by mutagenesis methods other than targeted deletions, with greater than 72% of essentiality predictions correctly validated. The confidence levels of our predictions reflect their probable accuracy for all datasets examined.

Additionally, a recent publication listed mouse essential genes revealed from experiments to generate a haploid mouse embryonic stem cell biobank (Elling et al., 2017). A total of 23 essential genes were identified through experimental analysis as essential for mouse embryonic stem cell survival. Of these genes, 16 were contained within our prediction dataset. Our classifier accurately predicted the essentiality status for 15 of the 16 genes (94%; Table S6), demonstrating further successful application of our classifier to additional experimental data types.

Enriched features of essential and non-essential genes

To understand the biological functions specific to essential and non-essential genes, we performed functional annotation of known and predicted essential and non-essential mouse genes using 4 distinct web tools to identify enriched features: Database for Annotation, Visualisation and Integrated Discovery (DAVID) v6.8 (Dennis et al., 2003), WebGestalt (2017 update) (Zhang et al., 2005), g:Profiler (2016 update) (Reimand et al., 2007) and PANTHER (v11.1) (Mi et al., 2016). Because our predicted gene datasets are considerably larger than the training sets we have previously analysed, we wished to explore whether or not the functional annotations of the predicted genes were similar to those of the genes with known essentiality status. Consistent with our previous work on experimentally validated mouse essential genes (Kabir et al., 2017), proteins encoded by predicted essential genes were found to be significantly enriched in localisation to intracellular locations, with 50.5% of genes annotated with the cellular component (CC) GO term 'nucleus'. Furthermore, biological process (BP) and molecular function (MF) GO terms relating to translation, chromosome segregation, information processing, RNA splicing, mRNA processing and numerous metabolic process were commonly enriched in predicted essential or non-essential mouse genes (Table 1). Helicase protein domains and helicase-related terms were also frequently significantly enriched (P<0.05) in all webtool outputs for predicted essential genes. These results confirm that essential genes tend to have critical functions in DNA replication, DNA repair, transcription and translation, as helicases are known to be involved in these processes (Sedman et al., 2000). Disease pathways were frequently enriched amongst essential genes, including many cancers, and Huntington's, Alzheimer's and Parkinson's diseases, confirming prior reports that essential genes are disease related (Dickerson et al., 2011).

Conversely, the UniProt keywords 'transmembrane helix' and 'transmembrane' were significantly enriched in the predicted non-essential genes (P=1.10×10^{-154} and P=2.62×10^{-154}, respectively), which is consistent with the significant enrichment of transmembrane proteins found in the known viable mouse genes previously examined (Kabir et al., 2017). Notably, the number of protein transmembrane domains was not a feature that was included in classifier training following GA feature selection, so it is interesting that this feature is prominent amongst the predicted non-essential genes even though it was not used in the classification criteria. We noted that olfactory functions were enriched in the predicted viable gene set, most likely due to the large number of olfactory receptor genes found in the mouse genome. We therefore excluded the olfactory receptors from our predicted viable gene dataset and performed the functional annotation analysis again to identify other features that are enriched once olfactory functions are excluded (Table S7).

Our findings on the functional enrichments of the large predicted gene datasets are consistent with the functions enriched in the smaller training datasets (Kabir et al., 2017), and can therefore identify biological requirements during development and postnatal life. Our classifier did not incorporate GO functional annotations within its selection criteria, so it is striking that there is consistent agreement between the GO functions enriched in genes with known essentiality status and genes with predicted essentiality status. In general, the known and predicted genes of either essentiality category share the same GO Slim annotations for BP, CC, MF and PANTHER protein domains, with deviation from the overall genome distribution for these annotations (Table 2). These findings highlight the functional differences between essential and non-essential genes.

PPI networks of essential and non-essential genes

Since we found protein network features to be highly informative in our classifier, we sought to examine the protein network topology of predicted essential and non-essential genes for comparison with their known essentiality counterparts. All PPI network graphs can

Disease Models & Mechanisms (2018) 11, dmm034546. doi:10.1242/dmm.034546

Table 1. Top 10 enriched GO terms found within DAVID for predicted essential and predicted non-essential mouse genes

Predicted essential			Predicted non-essential		
Cellular component	Biological process	Molecular function	Cellular component	Biological process	Molecular function
Nucleus	mRNA processing	Poly(A) RNA binding	Integral component of membrane	Sensory perception of smell	Olfactory receptor activity
50.54%	5.17%	16.14%	38.89%	10.48%	10.42%
$P=7.97\times10^{-265}$	$P=3.81\times10^{-73}$	$P=1.10\times10^{-207}$	$P=2.61\times10^{-182}$	$P<3.83\times10^{-197}$	$P=3.55\times10^{-289}$
Cytoplasm	Transcription, DNA templated	RNA binding	Plasma membrane	G-protein-coupled receptor signalling pathway	G-protein-coupled receptor activity
46.89%	16.79%	9.72%	23.67%	12.55%	13.64%
$P=1.75\times10^{-125}$	$P=3.70\times10^{-68}$	$P=4.82\times10^{-89}$	$P=9.75\times10^{-20}$	$P=3.83\times10^{-197}$	$P=1.84\times10^{-240}$
Nucleoplasm	RNA splicing	Nucleotide binding	Extracellular region	Detection of chemical stimulus involved in sensory perception	Odorant binding
19.03%	3.90%	15.52%	8.64%	1.93%	3.62%
$P=2.13\times10^{-115}$	$P=5.88\times10^{-55}$	$P=1.14\times10^{-43}$	$P=2.22\times10^{-6}$	$P=7.95\times10^{-58}$	$P=2.56\times10^{-123}$
Nucleolus	Regulation of transcription, DNA templated	Nucleic acid binding	Cornified envelope	Response to pheromone	Pheromone receptor activity
9.91%	17.06%	9.70%	0.44%	0.99%	1.78%
$P=4.10\times10^{-85}$	$P=2.05\times10^{-35}$	$P=3.95\times10^{-23}$	$P=1.81\times10^{-5}$	$P=5.02\times10^{-30}$	$P=2.15\times10^{-41}$
Intracellular ribonucleoprotein complex	Translation	DNA binding	Keratin filament	Detection of chemical stimulus involved in sensory perception of smell	Pheromone binding
4.41%	4.67%	13.07%	0.64%	0.39%	1.01%
$P=2.19\times10^{-49}$	$P=6.01\times10^{-34}$	$P=6.04\times10^{-20}$	$P=0.0542$	$P=8.62\times10^{-10}$	$P=1.16\times10^{-25}$
Spliceosomal complex	Protein transport	Cadherin binding involved in cell-cell adhesion	Acrosomal vesicle	Response to stimulus	Transmembrane signalling receptor activity
2.40%	6.10%	3.09%	0.65%	1.29%	2.24%
$P=2.72\times10^{-40}$	$P=1.70\times10^{-33}$	$P=1.17\times10^{-19}$	$P=0.548$	$P=1.14\times10^{-6}$	$P=1.55\times10^{-23}$
Ribosome	Cell division	Structural constituent of ribosome	Integral component of plasma membrane	Sensory perception of chemical stimulus	Arachidonic acid epoxygenase activity
2.86%	4.13%	2.75%	5.31%	0.25%	0.41%
$P=1.67\times10^{-37}$	$P=3.42\times10^{-26}$	$P=1.62\times10^{-14}$	$P=0.647$	$P=8.83\times10^{-4}$	$P=1.67\times10^{-6}$
Mitochondrion	Ribosomal RNA processing	Ligase activity	Sperm fibrous sheath	Peptide cross-linking	Steroid hydroxylase activity
12.98%	1.99%	3.44%	0.12%	0.42%	0.43%
$P=1.99\times10^{-31}$	$P=6.01\times10^{-25}$	$P=2.40\times10^{-14}$	$P=0.998$	$P=1.99\times10^{-3}$	$P=1.18\times10^{-4}$
Nuclear speck	Cell cycle	mRNA binding	Motile cilium	Epoxygenase P450 pathway	Serine-type endopeptidase inhibitor activity
2.59%	5.80%	1.75%	0.52%	0.26%	0.80%
$P=2.55\times10^{-23}$	$P=1.48\times10^{-24}$	$P=1.87\times10^{-13}$	$P=0.999$	$P=6.52\times10^{-3}$	$P=2.07\times10^{-4}$
Centrosome	Mitotic nuclear division	ATP binding	Outer dynein arm	Cilium movement	Sulfotransferase activity
4.30%	3.14%	10.27%	0.08%	0.29%	0.40%
$P=1.04\times10^{-22}$	$P=1.69\times10^{-20}$	$P=1.49\times10^{-11}$	$P=0.999$	$P=7.22\times10^{-3}$	$P=4.31\times10^{-3}$

The percentage of predicted genes in each term, along with the Bonferroni *P*-value of enrichment, is listed underneath each term. Terms were retrieved using DAVID's default thresholds (EASE=0.1, Count=2).

be represented by a scale-free model (Vella et al., 2017), as shown by the degree distribution of the networks, which fits a power-law curve (Fig. S1). In scale-free models, the degree value (i.e. number of interactions per network node) of most nodes is far from the mean. Only a few nodes in each network have a high number of interactions. However, PPIs of the essential genes datasets (known and predicted) form networks that are denser, having a higher average number of neighbours, a higher tendency to form clusters and less heterogeneity than the corresponding datasets of non-essential genes (Table 3), using network parameters as defined in Hubba (Lin et al., 2008; Dong and Horvath, 2007) and NetworkAnalyzer (Doncheva et al., 2012). We infer from the graph data that the PPI network generated from proteins encoded by essential genes shows higher connectivity than networks generated from non-essential genes, and that essential proteins are more likely to form hubs in the network (Table S8). Network features such as degree do differ between the known and predicted networks of both essentiality classes, indicating that the expectation that known and predicted proteins of a particular essentiality class will have the same properties could be an oversimplification.

Chromosomal distribution of essential and non-essential genes

We examined the distribution of essential and non-essential genes within the mouse genome, partitioned by known and predicted essentiality status (Fig. 4; Table S9). Chromosomes 11, 12 and 18 have the highest proportion of known essential genes, which comprise 9.96%, 9.84% and 9.60% of their entire chromosomal

Disease Models & Mechanisms (2018) 11, dmm034546. doi:10.1242/dmm.034546

Table 2. GO Slim functional annotations for essential and non-essential genes

	Essential mouse genes		Non-essential mouse genes		
Biological process (BP)	Known	Predicted	Known	Predicted	Whole-genome total
Biological adhesion	1.4%	0.9%	2.0%	1.4%	1.2%
Biological regulation	5.9%	4.3%	9.2%	11.7%	10.6%
Response to stimulus	7.7%	4.0%	10.5%	13.2%	11.7%
Cellular component organisation or biogenesis	6.9%	9.2%	4.0%	3.5%	5.4%
Cellular process	26.1%	31.0%	26.2%	25.6%	26.2%
Developmental process	10.4%	4.9%	8.3%	5.0%	6.0%
Immune system process	2.4%	1.4%	5.8%	2.8%	1.6%
Metabolic process	26.8%	32.2%	18.6%	17.6%	20.5%
Multicellular organismal process	4.7%	2.4%	6.5%	11.6%	8.4%
Localisation	5.7%	8.4%	6.6%	5.6%	6.5%
Locomotion	0.4%	0.2%	0.9%	0.3%	0.4%
Reproduction	1.5%	1.0%	1.2%	1.5%	1.2%
Cell killing	0.0%	0.0%	0.0%	0.1%	0.0%
Growth	0.0%	0.0%	0.1%	0.0%	0.0%
	Essential mouse genes		Non-essential mouse genes		
Cellular component (CC)	Known	Predicted	Known	Predicted	Whole-genome total
Cell junction	0.7%	0.8%	0.8%	0.5%	0.6%
Cell part	40.0%	42.2%	35.6%	28.6%	43.1%
Extracellular matrix	1.0%	0.4%	3.0%	1.5%	1.2%
Extracellular region	5.1%	1.0%	10.3%	6.8%	6.4%
Macromolecular complex	15.0%	19.0%	7.9%	7.2%	12.6%
Membrane	11.4%	7.1%	22.1%	35.7%	14.4%
Organelle	26.1%	29.6%	19.5%	19.2%	21.3%
Synapse	0.7%	0.0%	0.8%	0.5%	0.4%
	Essential mouse genes		Non-essential mouse genes		
Molecular function (MF)	Known	Predicted	Known	Predicted	Whole-genome total
Antioxidant activity	0.0%	0.1%	0.4%	0.2%	0.2%
Binding	44.6%	42.5%	36.6%	28.5%	34.0%
Catalytic activity	37.8%	39.9%	33.8%	24.9%	30.8%
Channel regulator activity	0.0%	0.1%	0.3%	0.2%	0.2%
Receptor activity	5.0%	1.9%	11.1%	19.7%	13.1%
Signal transducer activity	1.5%	0.4%	4.8%	14.7%	8.8%
Structural molecule activity	5.4%	9.0%	4.2%	4.3%	5.4%
Translation regulator activity	0.3%	1.5%	0.1%	0.3%	0.5%
Transporter activity	5.4%	4.4%	8.7%	7.3%	6.9%
	Essential mouse genes		Non-essential mouse genes		
Protein class	Known	Predicted	Known	Predicted	Whole-genome total
Calcium-binding protein	2.0%	1.9%	2.7%	2.9%	2.2%
Cell adhesion molecule	1.7%	0.6%	3.5%	3.9%	1.9%
Cell junction protein	0.9%	0.8%	0.9%	1.0%	1.1%
Nucleic acid binding	21.2%	25.9%	9.8%	9.7%	15.1%
Cytoskeletal protein	4.9%	5.4%	3.9%	4.3%	4.9%
Defence/immunity protein	1.4%	0.7%	4.0%	3.3%	3.8%
Enzyme modulator	6.9%	9.5%	7.8%	8.7%	8.3%
Extracellular matrix protein	1.3%	0.6%	2.8%	2.7%	1.6%
Membrane traffic protein	1.7%	3.1%	1.8%	1.8%	2.4%
Transmembrane receptor	0.3%	0.3%	0.5%	0.4%	0.5%
Signalling molecule	6.1%	2.1%	10.0%	6.9%	6.1%
Transcription factor	15.7%	11.6%	7.5%	6.4%	8.9%
Chaperone	0.7%	1.9%	0.8%	0.8%	1.1%
Oxidoreductase	3.6%	3.8%	3.6%	4.1%	3.8%
Receptor	4.7%	2.1%	10.4%	9.8%	6.5%
Hydrolase	7.8%	7.7%	9.6%	10.2%	8.6%
Isomerase	0.8%	1.2%	0.8%	1.2%	1.1%
Ligase	2.8%	4.3%	1.2%	1.9%	2.5%
Lyase	1.0%	1.1%	1.1%	0.8%	1.1%
Structural protein	0.6%	0.5%	1.0%	1.2%	1.1%
Carrier protein	1.5%	2.4%	2.2%	3.0%	2.5%
Transferase	7.9%	8.7%	6.9%	6.4%	8.0%
Transporter	4.2%	3.5%	7.1%	8.2%	6.6%
Viral protein	0.1%	0.0%	0.1%	0.1%	0.1%

Results report the percentage of genes in each group with a particular annotation. Whole-genome values of 0 include terms with a representation lower than 0.1%.

Disease Models & Mechanisms (2018) 11, dmm034546. doi:10.1242/dmm.034546

Table 3. Network statistics of PPIs of known and predicted essential and non-essential datasets

	Known essential	Known non-essential	Predicted essential	Predicted non-essential
Proteins in dataset	1307	3451	4455	12,505
Network nodes	850 (65%)	1663 (48%)	2635 (59%)	2879 (23%)
Degree	5.6	5.3	12.9	8.6
Clustering coefficient	0.17	0.14	0.32	0.22
Edge percolation component (EPC)	31.8	38.3	33.5	94.2
Density	0.007	0.003	0.005	0.003
Heterogeneity	1.20	1.40	1.27	1.78
Diameter	12	20	13	17
Centralisation	0.109	0.071	0.045	0.049
Path length	4.2	5.3	4.3	5.6

The average value of each parameter for each network is presented.

gene content, respectively. Chromosomes 5, 12 and 18 have the highest proportions of predicted essential genes across the whole genome. This finding agrees with previous experimental work, including a balancer chromosome random chemical mutagenesis study that found that ∼60% of mutant phenotypes mapped to mouse Chromosome 11 were homozygous lethal (Kile et al., 2003), and an additional study that reported many embryonic lethal mutations map to mouse Chromosome 5 (Wilson et al., 2005).

The essential and non-essential training set and predicted gene lists were separately uploaded into the bioinformatics database DAVID v6.8 (Dennis et al., 2003), and significantly enriched chromosomes were identified in each dataset. In agreement with our genomic analysis, Chromosome 11 was significantly enriched for both known essential genes and predicted essential genes (Bonferroni-corrected *P*-values of 6.88×10^{-5} and 1.30×10^{-3}, respectively). Chromosome 5 was the most significantly enriched chromosome in the predicted essential genes dataset, with 365 predicted essential genes (8.4% of 4329 genes) located on Chromosome 5 (Bonferroni-corrected *P*-value of 1.17×10^{-3}). Similar results were obtained from WebGestalt (2017 update) (Table S10).

Chromosome 7 is the autosome with the highest combined percentage of known and predicted non-essential genes at over 79%. This result suggests that the majority of genes localised to this chromosome tend not to function in developmentally crucial processes. According to the DAVID functional annotation tool, Chromosome 7 was the most significantly enriched chromosome in the predicted non-essential genes dataset, with a Bonferroni corrected *P*-value of 3.10×10^{-12}, containing 11.2% (1128 of 10,068 DAVID IDs) of predicted non-essential genes. Similar results were obtained with WebGestalt, finding 5 significantly over-represented (false discovery rate <0.05) cytogenetic bands belonging to Chromosome 7 for the predicted non-essential genes (Table S10). Three Chromosome 7 regions were also detected in the top 25 most significantly over-represented chromosomal locations for the known non-essential genes.

Overall, our findings show that there is variation in the distribution of essential and non-essential genes throughout the genome. These findings are consistent with a prior study on gene synteny and density, which found that Chromosome 7 contains far fewer essential genes than other mouse chromosomes, and that Chromosome 11 contains a high proportion of essential genes (Hentges et al., 2007). Additionally, experimental studies interrogating regions of mouse chromosomes through random chemical mutagenesis are consistent with our findings of gene essentiality predictions, indicating the localisation of essential genes on mouse Chromosomes 5 (Wilson et al., 2005) and 11 (Kile et al., 2003).

Database of gene predictions

In order to facilitate searches for essential and non-essential genes, we created a database of mouse essentiality data (MED; http://essentiality.ls.manchester.ac.uk). The essentiality status of all protein-coding mouse genes, and the confidence level of essentiality predictions, is included in the MED database. The database has several search options, including gene symbol, MGI gene ID, Ensembl gene ID and chromosomal location. Additionally, lists of all essential or non-essential genes within the genome can be retrieved and downloaded, or lists of genes by essentiality status within a particular genomic region. The MED database should expedite searches for mouse gene essentiality status, based upon our criteria for essential gene identification (Kabir et al., 2017).

Comparison to human essential and non-essential genes

We evaluated the applicability of our findings on mouse gene essentiality to human genes. We identified 1495 known human non-essential genes from the literature (Table S11) (MacArthur et al.,

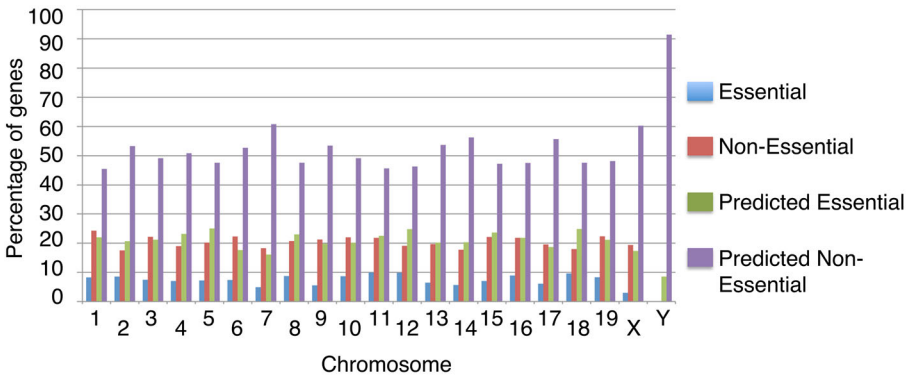


Fig. 4. The genomic distribution of essential and non-essential mouse genes, separated into known and predicted essentiality. The percentages of essential and non-essential genes on each chromosome are compiled from the MED database. In the genome as a whole, we calculate that there are 28% essential genes and 72% non-essential genes when known and predicted essentiality statuses are combined. Data are provided in Table S8.

Disease Models & Mechanisms (2018) 11, dmm034546. doi:10.1242/dmm.034546

2012; Sulem et al., 2015; Kaiser et al., 2015; Saleheen et al., 2017). Manual identification of the mouse orthologues of these human genes was conducted using Homologene, Online Mendelian Inheritance in Man (OMIM), GeneCards and the UCSC Genome Browser. Following this, duplicate genes present in the data were removed, in addition to any read-through genes and non-RefSeq UCSC genes (as annotated in the UCSC genome browser). Human genes without known mouse orthologues were excluded from the analysis. We therefore identified 1260 known non-essential human genes with mouse orthologues. Known human essential genes were also collected from the literature, providing in total 5205 genes from 4 publications (Table S11) (Blomen et al., 2015; Lek et al., 2016; Shamseldin et al., 2015; Wang et al., 2015). As above, mouse orthologues of these essential human genes were identified, and read-through genes, duplicate genes and those without a mouse orthologue excluded from our analysis. We obtained a final dataset of 5084 essential human genes and their mouse equivalents.

We found that 337 of the 1260 human non-essential genes and 1811 of the 5084 human essential genes were contained within our mouse essential or non-essential training sets. We then assessed these human and mouse genes for matching essentiality (Table 4; Table S11) to determine whether a gene that is annotated as essential in humans is also known to be essential in the mouse. We found that 296 (87.83%) known non-essential human genes were found to be non-essential in mouse knockout experiments, with 41 (12.17%) essentiality mismatches (i.e. non-essential in human but essential in mouse). The 1811 known essential human genes had 956 (52.79%) essentiality matches to their mouse orthologues, leaving 855 (47.21%) essential human genes with mismatched essentialities with their mouse equivalent (Table S11). This discrepancy could reflect the physiological, biological and developmental differences between mouse and human. Essentiality mismatches could also be due to the methodology of identifying human essential genes, as 2 publications classified human essential genes as those that caused proliferation failure when knocked down in cell culture cancer lines (Wang et al., 2015; Blomen et al., 2015). Cell culture essential genes might not be required for whole-organism viability, and cancerous cells might require tumour-specific essential genes not essential for healthy cells (Guo et al., 2017). However, when only human essential genes identified by sequencing are compared with mouse essential genes, 54% of these genes have mismatched essentiality with their mouse orthologue (Table S11), suggesting that the methodology for essential gene identification does not play a significant role in explaining the divergent essentiality classifications. Differences in mouse and human physiology and selective pressures since the human-mouse evolutionary split (Thomas et al., 2012) could result in non-essential genes becoming essential and vice versa. Critically, most human studies are unable to be truly comparable to mouse studies due to inabilities to test human embryos experimentally. One study sequenced human embryonic DNA, yet was unable to unequivocally confirm that all mutated genes cause embryonic lethality (Shamseldin et al., 2015). Therefore, genes that are identified as essential in humans from experimental cell culture data or sequence analysis might not necessarily cause lethality during human development.

For the 923 non-essential human genes and 3273 essential human genes which were not contained in either mouse training set, our mouse classifier predictions had a high percentage of essentiality status matches (Table S11). For example, 71.1% (2326/3273) of the human essential genes were also predicted as being essential in mice. Additionally, 79.4% of the 923 human non-essential genes had the same essentiality prediction status as their mouse orthologues (Table S11). Some discrepancies between human and mouse gene essentiality status are expected due to biological differences, rather than inaccurate classifier performance, as it has been reported that at least 20% of shared human and mouse genes result in different phenotypes when functionally deleted (Liao and Zhang, 2008). These results therefore give confidence that our mouse gene predictions can be used to inform future mouse and human genetic research.

To discover whether features enriched in essential and non-essential mouse genes are also enriched in human genes of the same essentiality, the DAVID functional annotation tool was used to retrieve enriched annotations. Overall, enriched terms matched across both species: essential genes had DNA-binding, helicase, transcription and nucleus-related enrichment, with non-essential genes enriched in transport, receptor, signalling, immunity, and membrane and extracellular locations (Table S12). Information processing terms are therefore absolutely fundamental to all organisms for viability, survival and reproduction as they are found to be enriched in minimal gene sets of bacteria (Juhas et al., 2014), yeast (Acencio and Lemke, 2009), mouse and human (Yang et al., 2014). Inconsistencies included protein domains associated with ion channels being enriched in the human essential gene dataset, but also enriched in the mouse non-essential gene dataset. Ubiquitin-related and mRNA processing terms were enriched in human non-essential genes and also in mouse essential genes. This finding was unexpected, as ubiquitin and mRNA processing have key developmental functions (Tu et al., 2012; Vriend et al., 2015); therefore, discrepancies between mouse and human essentiality annotations might be due to reported human cellular essential genes not being essential at the organismal level.

DISCUSSION

We compiled training sets from mouse knockout data to identify essential genes (Kabir et al., 2017), which were utilised to train several classifiers to predict gene essentiality. This work used a wide range of genomic features to predict essentiality, many of which have not been examined in previous studies (Yang et al., 2014). Our methodology has achieved greater 10-fold cross-validation classification accuracy than prior machine learning predictions of mouse knockout phenotypes (Yuan et al., 2012). Our classifier's performance is also more accurate than a support vector machine human essential gene classifier examined in jackknife tests and by 10-fold cross-validation (Yang et al., 2014). A strength of our study is the use of 2 blind test sets to further interrogate the validity of our classifier, which differs from other prior research generating

Table 4. Human and mouse essential gene conservation

	Human essential genes	Human non-essential genes
Known mouse essential genes	52.8% (956/1811)	12.2% (41/337)
Known mouse non-essential genes	47.2% (855/1811)	87.8% (296/337)
Predicted mouse essential genes	71.7% (2326/3273)	20.6% (190/923)
Predicted mouse non-essential genes	28.3% (926/3273)	79.4% (733/923)

mammalian essential gene classifiers (Yang et al., 2014; Yuan et al., 2012), but is similar to methodology utilised in a study to predict plant gene essentiality (Lloyd et al., 2015). The high accuracy of our predictions on the blind test sets, and the strong correlation between the confidence of our predictions and their accuracy, indicates that our classifier is discriminating between essential and non-essential genes. The percentage of genes predicted to be essential in the mouse genome using our classifier is similar to the percentage of genes found to be essential in mouse knockout experimental studies, and the properties we found to be enriched in mouse predicted essential genes are consistent with annotations of known mouse essential genes (White et al., 2013; Dickinson et al., 2016). Notably, we found that ∼20% of genes in our essential gene training dataset had been designated as non-essential genes by the IMPC (Koscielny et al., 2013). Although clearly the IMPC alleles produced viable mice, the majority of these genes had additional experimentally generated alleles reported in the MGI database that displayed lethal phenotypes. The IMPC database only contains reports of alleles generated as part of the IMPC project and not prior experimental data from other laboratories, which presents a limitation for utilising the IMPC data alone in determining the essentiality status of a given gene. The comparison of the MGI and IMPC datasets allows a quantification of the variation in experimental results for essentiality phenotypes that can be obtained from mouse knockout studies.

The 10 most informative features used in the random forest classifier to predict gene essentiality status relate to protein interactions or protein composition (Table S4). A study on human essential genes reported that topological properties of the PPI network are highly informative for predicting essential genes (Yang et al., 2014), and several studies on other organisms also find that protein interaction network features are useful for distinguishing essential and non-essential genes (Acencio and Lemke, 2009; Lloyd et al., 2015; Hwang et al., 2009; Li et al., 2014). In many species, essential genes occupy hubs within protein interaction networks (Lee et al., 2010; Liang and Li, 2007; Hwang et al., 2009); thus, it is understandable that protein network features are highly informative for predicting the essentiality of a gene with unknown essentiality status. Seven features reporting developmental gene expression levels are also highly discriminatory, because genes that are not expressed during development are unlikely to be essential for survival throughout gestation. Subcellular localisation features such as nucleus and plasma membrane were also found to have high information gain, which correlates with our finding that these same features showed significant statistical differences in their distribution amongst our training set genes (Kabir et al., 2017).

A publically available online database has been created to disseminate the essentiality predictions of mouse genes lacking experimental essentiality annotations (http://essentiality.ls.manchester.ac.uk), which is searchable by multiple identifiers and can produce lists of gene essentiality for download. We believe that our mouse gene essentiality status predictions will be useful for researchers seeking to create mouse mutants (a rapidly expanding area of biological research due to genome editing technology) (Singh et al., 2015), because researchers can quickly determine whether their gene of interest is likely to be essential or not. Owing to the conservation of function and essentiality status between mouse and human genes, knowledge of mouse gene essentiality will aid clinical geneticists seeking to interpret the impact of genome sequence variants on phenotype, a need that is rapidly increasing with the expanding use of genome and exome sequencing in clinical diagnostics. Knowledge of the composite set of essential genes of an organism is also of benefit for synthetic biology (Rancati et al., 2018).

Upon comparing our predictions of mouse gene essentiality with human gene essentiality annotations, we found a high degree of correlation between predicted mouse non-essential and essential genes and their human orthologues with known essentiality status. Similarly, we found a strong correlation between experimentally identified mouse non-essential genes and human known non-essential genes. Larger discrepancies were found between mouse known essential genes and human known essential genes, however, which we propose is related to the differing methodologies used to identify mouse and human essential genes, a hypothesis noted by others (Bartha et al., 2018). Given the prominence of mouse models for the study of human diseases (Rosenthal and Brown, 2007), an increased understanding of whether discrepancies in gene essentiality between these species represent biological differences or functional annotation differences will improve the interpretation of mouse model data.

MATERIALS AND METHODS

Compilation of datasets

Our essential and non-essential mouse gene datasets have previously been described (Kabir et al., 2017). We defined an essential gene as a gene causing lethality prior to postnatal day 3 in a single gene knockout experiment. Only single gene knockout (targeted deletion) experiments were considered. If a gene had a lethal phenotype in any knockout experiment, it was considered lethal, even if knockouts of other exons or on other strain backgrounds, or mutations generated by methods other than targeted deletion, did not have a lethal phenotype. IMPC data were retrieved through the 'phenotypes' query on the IMPC website (Koscielny et al., 2013), using the keywords 'embryonic lethality' for essential genes and MP keyword terms previously chosen for MGI searches (Kabir et al., 2017) for non-essential genes. IMPC subviable genes were obtained from the Embryo Development Special Report accessed on their website (Koscielny et al., 2013).

Alternative mouse mutagenesis-methodology data were collected from the MGI database. MGI genes were filtered using terms 'Viable' and 'Lethal' and specifying 'Null/Knockout alleles', with all chromosomes and generation methods selected, apart from 'Targeted', 'Transgenic' and 'QTL'. Publications for genes retrieved with both 'viable' and 'lethal' keywords were manually assessed, allowing verification of genes as essential or non-essential. Duplicate genes and those in training sets were excluded. Genes essential in mouse embryonic stem cells were identified from the literature (Elling et al., 2017).

Human essential genes (Blomen et al., 2015; Shamseldin et al., 2015; Wang et al., 2015; Lek et al., 2016) and non-essential genes (MacArthur et al., 2012; Kaiser et al., 2015; Sulem et al., 2015; Saleheen et al., 2017) were retrieved from the literature. To compare human gene lists with mouse datasets, mouse orthologues were manually retrieved from OMIM (Amberger et al., 2015), HomoloGene at NCBI (NCBI Resource Coordinators, 2016), GeneCards (v4.4.1) (Stelzer et al., 2016) and the UCSC Genome Browser (Casper et al., 2017). Duplicate genes were excluded, as were read-through genes and non-RefSeq UCSC genes [as annotated in the UCSC genome browser (Casper et al., 2017)], along with human genes without mouse orthologues.

Retrieval of gene features

Features including 'gene length', 'transcript count', 'exon count' and 'transcript per million' were computed based on data retrieved from Ensembl BioMart (Yates et al., 2016) and UniGene (Pontius et al., 2003; Stanton et al., 2003). The other genomic and protein-sequence-based features were retrieved directly from Ensembl (Cunningham et al., 2015), UniProt (UniProt Consortium, 2015), Pepstats (Rice et al., 2000) and SignalP (Petersen et al., 2011; UniProt Consortium, 2015). Mouse PPI data were obtained from the I2D database (Brown and Jurisica, 2005). In-depth descriptions of the features collected have previously been described (Kabir et al., 2017).

Disease Models & Mechanisms (2018) 11, dmm034546. doi:10.1242/dmm.034546

Dataset balancing

Because the essential and non-essential mouse gene training sets differed in the number of genes, random subsampling with no replacement (Vitter, 1985) was used to select a class-balanced subset from the training data set with no duplicate instances in the subset.

Discretisation

Discretisation (Han et al., 2011; Witten et al., 2016, 2011) of the numeric features of the training dataset was performed using the ChiMerge algorithm (Kerber, 1992) to remove noise and improve the speed of classifier training. Two adjacent intervals of each feature were merged into bigger intervals repeatedly, based on the chi-squared correlation of the 2 adjacent intervals and the class attribute. Initially, for each numeric value of a feature, an interval was created to contain the numeric value only. Then, a chi-squared test was used to test the hypothesis that the class attribute is independent of the 2 adjacent intervals. If the test was independent of the 2 adjacent intervals, they were merged; otherwise, they remained separate. Merging all pairs of adjacent intervals continued until the chi-squared value of every pair of adjacent intervals was greater than the chi-squared value determined with a significance level of 0.95.

Machine learning classifiers

In this study, the mammalian essential gene prediction problem was formulated as a supervised binary classification problem. Given a mouse gene p, we intended to predict the corresponding class y, such that $p \in y$ (Chen et al., 2012). We used Weka (version 3.6), a publicly available Java-based machine learning software (Hall et al., 2009), to implement the predictive classifier. We used naïve Bayes (Rish, 2001), J48 decision tree (Breiman et al., 1984), SVM (Cortes and Vapnik, 1995), logistic regression and Random Forest (Breiman, 2001) methods implemented in Weka as classifiers. Classifiers were trained on a fixed number of mouse genes labelled as essential or non-essential, each consisting of m features. Separate test datasets were also created that have not been included in the training datasets. We implemented 10-fold cross-validation on the training sets to assess the performance of each classifier, followed by 10-fold cross-validation on Test sets 1 and 2. Calculating the proportion of correctly predicted genes in these test datasets validated the performance of classifiers.

For the RBF kernel SVM, we set C to 50 and experimented with different values of gamma: 0.1, 0.05, 0.01, 0.005, 0.001, 0.0005 and 0.0001. For RBF kernel SVM and polynomial SVM, C is set to 50 because 50 is a common value for cost. For gamma, we tested the values 0.1, 0.05, 0.01, 0.005, 0.001, 0.0005 and 0.0001 to find the best value. Similarly, we tested polynonmial orders of 1, 2, 3 and 4. Polynomial kernel SVMs with the penalty term C of 50 and different orders 1, 2, 3 and 4 were trained using 10-fold cross-validation. For regularisation of logistic regression (LR), we used the default setting that the regulariser is set to w^2 and the ridge (penalty term) is set to 10^{-8}, where w is the weight vector of the LR. The default setting is the most common setting for LR. We treated the categorical features (e.g. subcellular localisations and types of proteins) of the gene essentiality dataset as numeric features and coded the discrete features as integers. For decision trees, we used the default parameter settings such that the confidence factor is set to the default value 0.25 (the confidence factor is used for pruning), and used the default C4.5 pruning instead of reduced error pruning. For naïve Bayes classifier, we assumed that the distribution of each attribute is Gaussian and used the probability density estimation to compute the prior probabilities. We used Bayes theorem to compute conditional probabilities.

Performance measures

Classifier performance was evaluated by 10-fold cross-validation analysis, where each training dataset was randomly partitioned into 10 equal parts with 9 parts being used for model training and the remaining part used for testing. We used the cross-validation method to limit overfitting of the classifier.

The performance of each classifier was determined from the total number of essential genes predicted correctly (TP), essential genes predicted incorrectly (FN), non-essential genes predicted correctly (TN) and non-essential genes predicted incorrectly (FP), presented as a confusion matrix. From the counts of each of these, 3 performance measures, including the true-positive rate (recall or sensitivity; TPR), false-positive rate (FPR) and the overall classification accuracy, as defined by the following equations, were estimated:

$$TPR = \frac{TP}{TP + FN} \quad (1)$$

$$FPR = \frac{FP}{FP + TN} \quad (2)$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}. \quad (3)$$

Further evaluation of classifier performance was achieved through the use of ROC curves, which were generated by plotting the TPRs against the FPRs at various threshold settings to present the probability of predicting true positives as a function of the probability of predicting false positives (Huang and Ling, 2005). The AUC of the ROC curves was used to estimate the overall prediction performance of the classifier, whereby an AUC of 1 represents a perfect prediction and an AUC of 0.5 represents a random guess.

Feature selection algorithms

Feature selection was performed using the GA implemented in Weka. This wrapper method relies on a fitness function, population size, crossover probability, mutation probability and maximum number of generations to select relevant features in relation to the chosen classifier. The fitness function, generally defined as the accuracy of the chosen classifier, measured the quality of the solution. We used the Information Gain feature selection filter in Weka, which selects a subset of features from the pool of all features (Han et al., 2011) to estimate the worth (rank) of a feature by measuring its information gain with respect to a classification target. We did not examine all possible combinations of features, but ranked the features individually in order of significance to identify the most informative features for classification.

Protein interaction network analysis

Four datasets of protein IDs corresponding to (1) known essential genes, (2) predicted essential genes, (3) known non-essential genes and (4) predicted non-essential genes were used to query the STRING database (Jensen et al., 2009) for PPIs. We used the stringApp (v.1.1.0) (Szklarczyk et al., 2017) plugin of Cytoscape (v.3.5.1) (Cline et al., 2007) to retrieve data from the STRING database. We filtered out PPIs for which there is no experimental evidence and those with a confidence score <0.4. Statistical analysis of the resulting networks was conducted using NetworkAnalyser (v.3.3.2) (Doncheva et al., 2012; Assenov et al., 2008) and the Cytoscape plugin cytoHubba (Chin et al., 2014). Unlinked nodes were eliminated prior to network analysis.

Functional classification and annotation of gene sets

Four web-based applications – DAVID (v6.8) (Dennis et al., 2003), WebGestalt (2017 update) (Zhang et al., 2005), g:Profiler (Reimand et al., 2007) and PANTHER (v11.1) (Mi et al., 2016) – were used for functional evaluation of predicted and known genes, all utilising a *Mus musculus* genomic background. For each tool, 4 mouse gene sets were separately uploaded: (1) known essential genes, (2) predicted essential genes, (3) known non-essential genes and (4) predicted non-essential genes.

DAVID's functional annotation tool was employed, applying default thresholds (unless otherwise stated in results). Enrichment data were collected from DAVID's 'Tissue Expression', 'UP_Keywords', 'Chromosome', 'KEGG_Pathway', 'InterPro', 'Pfam', 'BioGrid', 'GOterm_BP_Direct', 'GOterm_CC_Direct' and 'GOterm_MF_Direct' categories, and the top 50 results were analysed for each dataset. DAVID's 'Related Term' tool was implemented, alongside biological knowledge, to place similar terms in groups.

WebGestalt's Over-Representation enrichment Analysis (ORA) tool was utilised (Zhang et al., 2005). Data for the top 25 most significant results for GO BP, CC and MF non-redundant terms, chromosomal location, Wiki and Panther Pathways, and Phenotype were retrieved. For g:Profiler (Reimand et al., 2007), Kyoto Encyclopedia of Genes and Genomes (KEGG)

pathways, and mouse sequence homologs of the Human Phenotype Ontology and GO BP, CC and MF terms were retrieved.

Statistical over-representation was retrieved from PANTHER (Mi et al., 2016) for PANTHER Protein Classes, PANTHER Pathways, GO BP complete, GO CC complete and GO MF complete categories. Results were manually analysed, and terms over-represented in one essentiality and under-represented in either opposing essentiality gene-set were identified as differentiating terms. Additionally, PANTHER and WebGestalt provided visual and text-based GO Slim tools for functional classification of each dataset. GO Slim pie charts representing the whole mouse genome and our selected gene sets were generated from PANTHER, allowing comparative analysis. GO annotations from DAVID, WebGestalt and g:Profiler were combined to identify common significant GO terms enriched across multiple outputs.

Functional annotation for reported essential and non-essential human genes was completed using gene Ensembl IDs uploaded to DAVID. Six gene sets were separately uploaded: (1) essential human genes, (2) non-essential human genes, (3) essential mouse genes, (4) non-essential mouse genes, (5) 'matching essentiality' essential human genes, and (6) 'matching essentiality' non-essential human genes. A *Homo sapiens* background was applied for human gene lists and annotation results were retrieved from the same categories as stated above for mouse genes.

Genomic distribution of essential and non-essential genes

Utilising the MED (http://essentiality.ls.manchester.ac.uk), the total number of genes on each mouse chromosome was retrieved, along with each gene's known or predicted essentiality. Genomic distribution analysis of essential and non-essential genes within the entire mouse genome, partitioned into known and predicted essentiality, was performed, and proportions of lethal and viable genes on each chromosome were determined. Chromosomal location and cytogenetic band enrichment for mouse essential and non-essential genes was identified from WebGestalt and DAVID.

Essentiality model testing

Gene predictions were compared against blind and alternative mouse mutagenesis genes, both with currently validated essentialities, by testing known genes against their equivalent gene's predicted essentiality. Custom-written Python scripts (available on request) compared collated gene lists with model gene predictions.

Statistics

All statistical analyses were carried out using R statistical software (R 3.0.1, The R Foundation for Statistical Computing). For all database functional analyses, the Bonferroni correction was applied to retrieve significantly enriched terms, with a statistical significance threshold of $P<0.05$ (unless otherwise stated). Distributions of plotted data were tested for normality using the Shapiro–Wilk test. For normally distributed data, Welch's 2-sided *t*-test for unequal variance was implemented, whereas for non-normally distributed data, the 2-sided non-parametric Wilcoxon Rank-Sum test was used, to determine statistical significance.

Acknowledgements

We thank David Robertson for useful discussions, Paul Johnston for technical support with the MED database and Rory Luscombe for assistance with figure editing.

Competing interests

The authors declare no competing or financial interests.

Author contributions

Conceptualization: A.J.D., K.E.H.; Methodology: D.T., S.W., M.K., G.T.; Software: D.T., S.W., M.K., G.T.; Validation: S.W.; Formal analysis: D.T., S.W., M.K., G.T., A.J.D., K.E.H.; Investigation: D.T., S.W., M.K., G.T.; Data curation: D.T., S.W., M.K., G.T.; Writing - original draft: D.T., S.W., M.K., G.T., A.J.D., K.E.H.; Writing - review & editing: S.W., M.K., G.T., A.J.D., K.E.H.; Visualization: D.T., S.W., G.T.; Supervision: A.J.D., K.E.H.; Funding acquisition: A.J.D., K.E.H.

Funding

This work was supported by the Biotechnology and Biological Sciences Research Council (BB/L018276/1 to K.E.H. and A.J.D.) and the University of Manchester (British Commonwealth PhD Studentship and Presidential Doctoral Scholarship to M.K.). The funders had no role in study design, data collection, data interpretation or writing of the manuscript.

Supplementary information

Supplementary information available online at
http://dmm.biologists.org/lookup/doi/10.1242/dmm.034546.supplemental

References

Acencio, M. L. and Lemke, N. (2009). Towards the prediction of essential genes by integration of network topology, cellular localization and biological process information. *BMC Bioinformatics* **10**, 290.

Amberger, J. S., Bocchini, C. A., Schiettecatte, F., Scott, A. F. and Hamosh, A. (2015). OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Res.* **43**, D789-D798.

Assenov, Y., Ramírez, F., Schelhorn, S.-E., Lengauer, T. and Albrecht, M. (2008). Computing topological parameters of biological networks. *Bioinformatics* **24**, 282-284.

Ayadi, A., Birling, M.-C., Bottomley, J., Bussell, J., Fuchs, H., Fray, M., Gailus-Durner, V., Greenaway, S., Houghton, R., Karp, N. et al. (2012). Mouse large-scale phenotyping initiatives: overview of the European Mouse Disease Clinic (EUMODIC) and of the Wellcome Trust Sanger Institute Mouse Genetics Project. *Mamm. Genome* **23**, 600-610.

Bartha, I., Di Iulio, J., Venter, J. C. and Telenti, A. (2018). Human gene essentiality. *Nat. Rev. Genet.* **19**, 51-62.

Blomen, V. A., Majek, P., Jae, L. T., Bigenzahn, J. W., Nieuwenhuis, J., Staring, J., Sacco, R., van Diemen, F. R., Olk, N., Stukalov, A. et al. (2015). Gene essentiality and synthetic lethality in haploid human cells. *Science* **350**, 1092-1096.

Bradley, A., Anastassiadis, K., Ayadi, A., Battey, J. F., Bell, C., Birling, M.-C., Bottomley, J., Brown, S. D., Bürger, A., Bult, C. J. et al. (2012). The mammalian gene function resource: the International Knockout Mouse Consortium. *Mamm. Genome* **23**, 580-586.

Breiman, L. (2001). Random forests. *Mach. Learn.* **45**, 5-32.

Breiman, L., Friedman, J., Stone, C. J. and Olshen, R. A. (1984). *Classification and Regression Trees*. CRC Press.

Brown, K. R. and Jurisica, I. (2005). Online predicted human interaction database. *Bioinformatics* **21**, 2076-2082.

Brown, S. D. M. and Moore, M. W. (2012). Towards an encyclopaedia of mammalian gene function: the International Mouse Phenotyping Consortium. *Dis. Model. Mech.* **5**, 289-292.

Bult, C. J., Eppig, J. T., Blake, J. A., Kadin, J. A., Richardson, J. E. and **Mouse Genome Database Group**. (2016). Mouse genome database 2016. *Nucleic Acids Res.* **44**, D840-D847.

Casper, J., Zweig, A. S., Villarreal, C., Tyner, C., Speir, M. L., Rosenbloom, K. R., Raney, B. J., Lee, C. M., Lee, B. T., Karolchik, D. et al. (2017). The UCSC Genome Browser database: 2018 update. *Nucleic Acids Res.* **46**, D762-D769.

Chen, W.-H., Minguez, P., Lercher, M. J. and Bork, P. (2012). OGEE: an online gene essentiality database. *Nucleic Acids Res.* **40**, D901-D906.

Cheng, J., Wu, W., Zhang, Y., Li, X., Jiang, X., Wei, G. and Tao, S. (2013). A new computational strategy for predicting essential genes. *BMC Genomics* **14**, 910.

Cheng, J., Xu, Z., Wu, W., Zhao, L., Li, X., Liu, Y. and Tao, S. (2014). Training set selection for the prediction of essential genes. *PLoS ONE* **9**, e86805.

Chin, C.-H., Chen, S.-H., Wu, H.-H., Ho, C.-W., Ko, M.-T. and Lin, C.-Y. (2014). cytoHubba: identifying hub objects and sub-networks from complex interactome. *BMC Syst. Biol.* **8** Suppl. 4, S11.

Cline, M. S., Smoot, M., Cerami, E., Kuchinsky, A., Landys, N., Workman, C., Christmas, R., Avila-Campilo, I., Creech, M., Gross, B. et al. (2007). Integration of biological networks and gene expression data using Cytoscape. *Nat. Protoc.* **2**, 2366-2382.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Mach. Learn.* **20**, 273-297.

Cunningham, F., Amode, M. R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S. et al. (2015). Ensembl 2015. *Nucleic Acids Res.* **43**, D662-D669.

Deng, J. (2015). An integrated machine-learning model to predict prokaryotic essential genes. *Methods Mol. Biol.* **1279**, 137-151.

Deng, J., Deng, L., Su, S., Zhang, M., Lin, X., Wei, L., Minai, A. A., Hassett, D. J. and Lu, L. J. (2011). Investigating the predictability of essential genes across distantly related organisms using an integrative approach. *Nucleic Acids Res.* **39**, 795-807.

Dennis, G., Jr, Sherman, B. T., Hosack, D. A., Yang, J., Gao, W., Lane, H. C. and Lempicki, R. A. (2003). DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol.* **4**, P3.

Dickerson, J. E., Zhu, A., Robertson, D. L. and Hentges, K. E. (2011). Defining the role of essential genes in human disease. *PLoS ONE* **6**, e27368.

Dickinson, M. E., Flenniken, A. M., Ji, X., Teboul, L., Wong, M. D., White, J. K., Meehan, T. F., Weninger, W. J., Westerberg, H., Adissu, H. et al. (2016). High-throughput discovery of novel developmental phenotypes. *Nature* **537**, 508-514.
Doncheva, N. T., Assenov, Y., Domingues, F. S. and Albrecht, M. (2012). Topological analysis and interactive visualization of biological networks and protein structures. *Nat. Protoc.* **7**, 670-685.
Dong, J. and Horvath, S. (2007). Understanding network concepts in modules. *BMC Syst. Biol.* **1**, 24.
Elling, U., Wimmer, R. A., Leibbrandt, A., Burkard, T., Michlits, G., Leopoldi, A., Micheler, T., Abdeen, D., Zhuk, S., Aspalter, I. M. et al. (2017). A reversible haploid mouse embryonic stem cell biobank resource for functional genomics. *Nature* **550**, 114-118.
Guo, F.-B., Dong, C., Hua, H.-L., Liu, S., Luo, H., Zhang, H.-W., Jin, Y.-T. and Zhang, K.-Y. (2017). Accurate prediction of human essential genes using only nucleotide composition and association information. *Bioinformatics* **33**, 1758-1764.
Gustafson, A. M., Snitkin, E. S., Parker, S. C. J., Delisi, C. and Kasif, S. (2006). Towards the identification of essential genes using targeted genome sequencing and comparative analysis. *BMC Genomics* **7**, 265.
Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. and Witten, I. H. (2009). The weka data mining software. *ACM SIGKDD Explorations Newsletter* **11**, 10-18.
Han, J., Pei, J. and Kamber, M. (2011). *Data Mining: Concepts and Techniques*. Elsevier.
Hentges, K. E., Pollock, D. D., Liu, B. and Justice, M. J. (2007). Regional variation in the density of essential genes in mice. *PLoS Genet.* **3**, e72.
Hua, H.-L., Zhang, F.-Z., Labena, A. A., Dong, C., Jin, Y.-T. and Guo, F.-B. (2016). An approach for predicting essential genes using multiple homology mapping and machine learning algorithms. *Biomed. Res. Int.* **2016**, 7639397.
Huang, M. J. and Ling, C. X. (2005). Using AUC and accuracy in evaluating learning algorithms. *IEEE Trans. Knowledge Data Eng.* **17**, 299-310.
Hwang, Y.-C., Lin, C.-C., Chang, J.-Y., Mori, H., Juan, H.-F. and Huang, H.-C. (2009). Predicting essential genes based on network and sequence analysis. *Mol. Biosyst.* **5**, 1672-1678.
Jensen, L. J., Kuhn, M., Stark, M., Chaffron, S., Creevey, C., Muller, J., Doerks, T., Julien, P., Roth, A., Simonovic, M. et al. (2009). STRING 8–a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res.* **37**, D412-D416.
Juhas, M., Reuss, D. R., Zhu, B. and Commichau, F. M. (2014). Bacillus subtilis and Escherichia coli essential genes and minimal cell factories after one decade of genome engineering. *Microbiology* **160**, 2341-2351.
Kabir, M., Barradas, A., Tzotzos, G. T., Hentges, K. E. and Doig, A. J. (2017). Properties of genes essential for mouse development. *PLoS ONE* **12**, e0178273.
Kaiser, V. B., Svinti, V., Prendergast, J. G., Chau, Y. Y., Campbell, A., Patarcic, I., Barroso, I., Joshi, P. K., Hastie, N. D., Miljkovic, A. et al. (2015). Homozygous loss-of-function variants in European cosmopolitan and isolate populations. *Hum. Mol. Genet.* **24**, 5464-5474.
Kerber, R. (1992). ChiMerge: discretization of numeric attributes. *Proceedings of the Tenth National Conference on Artificial Intelligence*. San Jose: AAAI Press.
Kile, B. T., Hentges, K. E., Clark, A. T., Nakamura, H., Salinger, A. P., Liu, B., Box, N., Stockton, D. W., Johnson, R. L., Behringer, R. R. et al. (2003). Functional genetic analysis of mouse chromosome 11. *Nature* **425**, 81-86.
Kofoed, M., Milbury, K. L., Chiang, J. H., Sinha, S., Ben-Aroya, S., Giaever, G., Nislow, C., Hieter, P. and Stirling, P. C. (2015). An updated collection of sequence barcoded temperature-sensitive alleles of yeast essential genes. *G3* **5**, 1879-1887.
Koscielny, G., Yaikhom, G., Iyer, V., Meehan, T. F., Morgan, H., Atienza-Herrero, J., Blake, A., Chen, C.-K., Easty, R., Di Fenza, A. et al. (2013). The International Mouse Phenotyping Consortium Web Portal, a unified point of access for knockout mice and related phenotyping data. *Nucleic Acids Res.* **42**, D802-D809.
Lee, I., Ambaru, B., Thakkar, P., Marcotte, E. M. and Rhee, S. Y. (2010). Rational association of genes with traits using a genome-scale gene network for Arabidopsis thaliana. *Nat. Biotechnol.* **28**, 149-156.
Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., O'Donnell-Luria, A. H., Ware, J. S., Hill, A. J., Cummings, B. B. et al. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285-291.
Li, M., Zheng, R., Zhang, H., Wang, J. and Pan, Y. (2014). Effective identification of essential proteins based on priori knowledge, network topology and gene expressions. *Methods* **67**, 325-333.
Liang, H. and Li, W.-H. (2007). Gene essentiality, gene duplicability and protein connectivity in human and mouse. *Trends Genet.* **23**, 375-378.
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.
Lin, C.-Y., Chin, C.-H., Wu, H.-H., Chen, S.-H., Ho, C.-W. and Ko, M.-T. (2008). Hubba: hub objects analyzer–a framework of interactome hubs identification for network biology. *Nucleic Acids Res.* **36**, W438-W443.
Liu, X., Wang, B.-J., Xu, L., Tang, H.-L. and Xu, G.-Q. (2017). Selection of key sequence-based features for prediction of essential genes in 31 diverse bacterial species. *PLoS ONE* **12**, e0174638.
Lloyd, J. P., Seddon, A. E., Moghe, G. D., Simenc, M. C. and Shiu, S.-H. (2015). Characteristics of plant essential genes allow for within- and between-species prediction of lethal mutant phenotypes. *Plant Cell* **27**, 2133-2147.
Lu, Y., Deng, J., Rhodes, J. C., Lu, H. and Lu, L. J. (2014). Predicting essential genes for identifying potential drug targets in Aspergillus fumigatus. *Comput. Biol. Chem.* **50**, 29-40.
Macarthur, D. G., Balasubramanian, S., Frankish, A., Huang, N., Morris, J., Walter, K., Jostins, L., Habegger, L., Pickrell, J. K., Montgomery, S. B. et al. (2012). A systematic survey of loss-of-function variants in human protein-coding genes. *Science* **335**, 823-828.
Mi, H., Poudel, S., Muruganujan, A., Casagrande, J. T. and Thomas, P. D. (2016). PANTHER version 10: expanded protein families and functions, and analysis tools. *Nucleic Acids Res.* **44**, D336-D342.
Motenko, H., Neuhauser, S. B., O'Keefe, M. and Richardson, J. E. (2015). MouseMine: a new data warehouse for MGI. *Mamm. Genome* **26**, 325-330.
NCBI Resource Coordinators. (2016). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **44**, D7-D19.
Nigatu, D., Sobetzko, P., Yousef, M. and Henkel, W. (2017). Sequence-based information-theoretic features for gene essentiality prediction. *BMC Bioinformatics* **18**, 473.
Ning, L. W., Lin, H., Ding, H., Huang, J., Rao, N. and Guo, F. B. (2014). Predicting bacterial essential genes using only sequence composition information. *Genet. Mol. Res.* **13**, 4564-4572.
Petersen, T. N., Brunak, S., von Heijne, G. and Nielsen, H. (2011). SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat. Methods* **8**, 785-786.
Plaimas, K., Eils, R. and König, R. (2010). Identifying essential genes in bacterial metabolic networks with machine learning methods. *BMC Syst. Biol.* **4**, 56.
Pontius, J. U., Wagner, L. and Schuler, G. D. (2003). UniGene: a unified view of the transcriptome. In *The NCBI Handbook*. Bethesda: National Center for Biotechnology Information.
Rancati, G., Moffat, J., Typas, A. and Pavelka, N. (2018). Emerging and evolving concepts in gene essentiality. *Nat. Rev. Genet.* **19**, 34-49.
Reimand, J., Kull, M., Peterson, H., Hansen, J. and Vilo, J. (2007). g:Profiler–a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic Acids Res.* **35**, W193-W200.
Rice, P., Longden, I. and Bleasby, A. (2000). EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.* **16**, 276-277.
Rish, I. (2001). An empirical study of the naive Bayes classifier. In *IJCAI 2001 Work. Empir. Methods Artific. Intell.* 3, 41-46.
Rosenthal, N. and Brown, S. (2007). The mouse ascending: perspectives for human-disease models. *Nat. Cell Biol.* **9**, 993-999.
Saleheen, D., Natarajan, P., Armean, I. M., Zhao, W., Rasheed, A., Khetarpal, S. A., Won, H.-H., Karczewski, K. J., O'Donnell-Luria, A. H., Samocha, K. E. et al. (2017). Human knockouts and phenotypic analysis in a cohort with a high rate of consanguinity. *Nature* **544**, 235-239.
Sedman, T., Kuusk, S., Kivi, S. and Sedman, J. (2000). A DNA helicase required for maintenance of the functional mitochondrial genome in Saccharomyces cerevisiae. *Mol. Cell. Biol.* **20**, 1816-1824.
Seringhaus, M., Paccanaro, A., Borneman, A., Snyder, M. and Gerstein, M. (2006). Predicting essential genes in fungal genomes. *Genome Res.* **16**, 1126-1135.
Shamseldin, H. E., Tulbah, M., Kurdi, W., Nemer, M., Alsahan, N., AL Mardawi, E., Khalifa, O., Hashem, A., Kurdi, A., Babay, Z. et al. (2015). Identification of embryonic lethal genes in humans by autozygosity mapping and exome sequencing in consanguineous families. *Genome Biol.* **16**, 116.
Singh, P., Schimenti, J. C. and Bolcun-Filas, E. (2015). A mouse geneticist's practical guide to CRISPR applications. *Genetics* **199**, 1-15.
Stanton, J. A., Macgregor, A. B. and Green, D. P. (2003). Identifying tissue-enriched gene expression in mouse tissues using the NIH UniGene database. *Appl. Bioinformatics* **2**, S65-S73.
Stelzer, G., Rosen, N., Plaschkes, I., Zimmerman, S., Twik, M., Fishilevich, S., Stein, T. I., Nudel, R., Lieder, I., Mazor, Y. et al. (2016). The GeneCards suite: from gene data mining to disease genome sequence analyses. *Curr. Protoc. Bioinformatics* **54**, 1.30.1-1.30.33.
Sulem, P., Helgason, H., Oddson, A., Stefansson, H., Gudjonsson, S. A., Zink, F., Hjartarson, E., Sigurdsson, G. T., Jonasdottir, A., Jonasdottir, A. et al. (2015). Identification of a large set of rare complete human knockouts. *Nat. Genet.* **47**, 448-452.
Sung, Y. H., Baek, I.-J., Seong, J. K., Kim, J.-S. and Lee, H.-W. (2012). Mouse genetics: catalogue and scissors. *BMB Rep.* **45**, 686-692.
Szklarczyk, D., Morris, J. H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., Santos, A., Doncheva, N. T., Roth, A., Bork, P. et al. (2017). The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Res.* **45**, D362-D368.
Thomas, P. D., Wood, V., Mungall, C. J., Lewis, S. E., Blake, J. A. and **Gene Ontology Consortium**. (2012). On the use of gene ontology annotations to

assess functional similarity among orthologs and paralogs: a short report. *PLoS Comput. Biol.* **8**, e1002386.

Tu, Y., Chen, C., Pan, J., Xu, J., Zhou, Z. G. and Wang, C. Y. (2012). The Ubiquitin Proteasome Pathway (UPP) in the regulation of cell cycle control and DNA damage repair and its implication in tumorigenesis. *Int. J. Clin. Exp. Pathol.* **5**, 726-738.

Uniprot Consortium. (2015). UniProt: a hub for protein information. *Nucleic Acids Res.* **43**, D204-D212.

Vella, D., Zoppis, I., Mauri, G., Mauri, P. and Di Silvestre, D. (2017). From protein-protein interactions to protein co-expression networks: a new perspective to evaluate large-scale proteomic data. *EURASIP J. Bioinform. Syst. Biol.* **2017**, 6.

Visa, S. and Ralescu, A. (2005). Issues in mining imbalanced data sets-a review paper. *Proc. 16th Midwest Artificial Intelligence and Cognitive Science Conference*, 67-73.

Vitter, J. S. (1985). Random sampling with a reservoir. *ACM Trans. Math. Softw.* **11**, 37-57.

Vriend, J., Ghavami, S. and Marzban, H. (2015). The role of the ubiquitin proteasome system in cerebellar development and medulloblastoma. *Mol. Brain* **8**, 64.

Wang, T., Birsoy, K., Hughes, N. W., Krupczak, K. M., Post, Y., Wei, J. J., Lander, E. S. and Sabatini, D. M. (2015). Identification and characterization of essential genes in the human genome. *Science* **350**, 1096-1101.

White, J. K., Gerdin, A.-K., Karp, N. A., Ryder, E., Buljan, M., Bussell, J. N., Salisbury, J., Clare, S., Ingham, N. J., Podrini, C. et al. (2013). Genome-wide generation and systematic phenotyping of knockout mice reveals new roles for many genes. *Cell* **154**, 452-464.

Wilson, L., Ching, Y. H., Farias, M., Hartford, S. A., Howell, G., Shao, H., Bucan, M. and Schimenti, J. C. (2005). Random mutagenesis of proximal mouse chromosome 5 uncovers predominantly embryonic lethal mutations. *Genome Res.* **15**, 1095-1105.

Witten, I. H., Frank, E. and Hall, M. A. (2011). *Data Mining Practical Machine Learning Tools and Techniques*. Morgan Kaufmann.

Witten, I. H., Frank, E., Hall, M. A. and Pal, C. (2016). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann.

Yang, Y. and Pederson, J. O. (1997). A Comparative Study on Feature Selection in Text Categorization. *ICML '97 Proceedings of the Fourteenth International Conference on Machine Learning*, 412-420.

Yang, L., Wang, J., Wang, H., Lv, Y., Zuo, Y., Li, X. and Jiang, W. (2014). Analysis and identification of essential genes in humans using topological properties and biological information. *Gene* **551**, 138-151.

Yates, A., Akanni, W., Amode, M. R., Barrell, D., Billis, K., Carvalho-Silva, D., Cummins, C., Clapham, P., Fitzgerald, S., Gil, L. et al. (2016). Ensembl 2016. *Nucleic Acids Res.* **44**, D710-D716.

Yu, Y., Yang, L., Liu, Z. and Zhu, C. (2017). Gene essentiality prediction based on fractal features and machine learning. *Mol. Biosyst.* **13**, 577-584.

Yuan, Y., Xu, Y., Xu, J., Ball, R. L. and Liang, H. (2012). Predicting the lethal phenotype of the knockout mouse by integrating comprehensive genomic data. *Bioinformatics* **28**, 1246-1252.

Zhang, B., Kirov, S. and Snoddy, J. (2005). WebGestalt: an integrated system for exploring gene sets in various biological contexts. *Nucleic Acids Res.* **33**, W741-W748.

Zhong, J., Wang, J., Peng, W., Zhang, Z. and Pan, Y. (2013). Prediction of essential proteins based on gene expression programming. *BMC Genomics* **14**, S7.