ELSEVIER

COMPUTATIONAL
ANDSTRUCTURAL
BIOTECHNOLOGY
JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Predicting host dependency factors of pathogens in *Drosophila melanogaster* using machine learning

Olufemi Aromolaran [a,b,c], Thomas Beder [b,d], Eunice Adedeji [c,e], Yvonne Ajamma [c], Jelili Oyelade [a,b,c], Ezekiel Adebiyi [a,c], Rainer Koenig [b,d,*]

[a] *Department of Computer & Information Sciences, Covenant University, Ota, Ogun State, Nigeria*
[b] *Integrated Research and Treatment Center, Center for Sepsis Control and Care (CSCC), Jena University Hospital, Am Klinikum 1, 07747 Jena, Germany*
[c] *Covenant University Bioinformatics Research (CUBRe), Covenant University, Ota, Ogun State, Nigeria*
[d] *Institute of Infectious Diseases and Infection Control, Jena University Hospital, Am Klinikum 1, 07747 Jena, Germany*
[e] *Department of Biochemistry, Covenant University, Ota, Ogun State, Nigeria*

ARTICLE INFO

Article history:
Received 8 June 2021
Received in revised form 6 August 2021
Accepted 6 August 2021
Available online 9 August 2021

Keywords:
Host factors
Bacteria
Infection
Knockout screen
Machine learning
Drosophila

ABSTRACT

Pathogens causing infections, and particularly when invading the host cells, require the host cell machinery for efficient regeneration and proliferation during infection. For their life cycle, host proteins are needed and these Host Dependency Factors (HDF) may serve as therapeutic targets. Several attempts have approached screening for HDF producing large lists of potential HDF with, however, only marginal overlap.

To get consistency into the data of these experimental studies, we developed a machine learning pipeline. As a case study, we used publicly available lists of experimentally derived HDF from twelve different screening studies based on gene perturbation in *Drosophila melanogaster* cells or *in vivo* upon bacterial or protozoan infection. A total of 50,334 gene features were generated from diverse categories including their functional annotations, topology attributes in protein interaction networks, nucleotide and protein sequence features, homology properties and subcellular localization. Cross-validation revealed an excellent prediction performance. All feature categories contributed to the model. Predicted and experimentally derived HDF showed a good consistency when investigating their common cellular processes and function. Cellular processes and molecular function of these genes were highly enriched in membrane trafficking, particularly in the trans-Golgi network, cell cycle and the Rab GTPase binding family.

Using our machine learning approach, we show that HDF in organisms can be predicted with high accuracy evidencing their common investigated characteristics. We elucidated cellular processes which are utilized by invading pathogens during infection. Finally, we provide a list of 208 novel HDF proposed for future experimental studies.

1. Introduction

Infectious diseases cause a major human and agricultural health burden. They are caused by opportunistic, pathogenic microorganisms such as bacteria, protozoans and fungi, and by viruses. Typically, discovering drugs and drug targets against these pathogens focus on target factors (proteins, cellular structure, membranes) of the pathogens themselves. However, these drugs become ineffective when the respective pathogens develop drug-resistant variants enabled by high mutation rates and evolutionary pressure [1]. A powerful but, however, yet under-explored alternative is to identify factors of the host cells being essential for the pathogen's life cycle. These host dependency factors (HDF) are proteins of the host cell needed by the pathogens to survive and replicate in the host cell or organism. In contrast to factors of the pathogen, HDF are not exposed to mutations and the evolutionary pressure of the pathogen. Identifying HDF has not only the potential to find therapeutic targets, but may also provide valuable insights into microbial pathogenesis and potential mechanisms for manipulation of host pathways [1]. Cheng and colleagues [2] observed that silencing of gene CG3573, a type II inositol 1,4,5-5-phosphatase, myotubularin, the ortholog of the mammalian myotubular myopathy-related protein 2 (MTMR2) and *Sbf*, a regulating partner

[*] Corresponding author at: Integrated Research and Treatment Center, Center for Sepsis Control and Care (CSCC), Jena University Hospital, Am Klinikum 1, 07747 Jena, Germany.

E-mail address: rainer.koenig@uni-jena.de (R. Koenig).

https://doi.org/10.1016/j.csbj.2021.08.010
2001-0370/© 2021 Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology.

of the myotubularin ortholog (MTMR5) led to a decrease of bacterial entry and less-effective vacuolar escape [3]. Besides this, disrupting direct host-pathogen interactions can also disturb the propagation of pathogens [4–6]. For this, not only knowledge of the bacterial, but, more importantly, of the involved host factors is needed [7]. However, studying gene perturbations in human can only base on cell lines but not on the whole organism. Owing to the genetic similarities and conserved pathways between *D. melanogaster* and mammals, the use of the Drosophila model as a platform to unveil novel mechanisms of infection and disease progression has been widely investigated [8] including host-pathogen interaction studies [9–12]. E.g. Akimana et al. [9] performed an RNAi screening experiment using Drosophila cells to identify host factors infected with *Francisella tularensis*. They found CDC27 and USP22 genes to be HDF which they validated in mammalian kidney cells. Knockdown of these genes inhibited the replication of the bacteria also in human host cells throughout the intracellular infection period.

However, when searching the literature and databases of HDF screening experiments in *D. melanogaster*, we observed a high heterogeneity and only a few overlap of the identified HDF. The variation and differential susceptibility could be attributed to the functional genetic diversity of the immune response [13], the different investigated pathogens, mode of infection, the use of different cell lines for experimental studies, the assay time post infection, the procedures used to measure infection, and differing approaches to analyze experimental data [14,15]. This makes it difficult to derive common mechanisms of these host dependencies.

Besides this, modern machine learning has been applied in a plethora of biological research fields aiming to integrate such heterogeneous data, as e.g. for the prediction of essential genes in Drosophila [16] and bacteria [17], and also cancer cells [18–23]. A semi-supervised machine learning approach predicted host dependency factors of Human Immunodeficiency Virus (HIV) in human cells using network topology features from protein interactions [1] and observed high consistency of the prediction results to the defined gold standard (85% precision at 60% recall) evidencing the validity of this approach. We followed this path and set up a machine learning pipeline to identify HDF and their common cellular processes for pathogenic infection in *D. melanogaster.* We employed a well-elaborated assembly of a broad range of features covering intrinsic and extrinsic gene and protein characteristics, gene network topology, molecular function, compartment information, biological processes and evolutionary conservation. We assembled a gold standard for our predictions from an elaborated set of twelve experimental knockdown or knockout screens. To the best of our knowledge, this is the first attempt to use machine learning to identify HDF for pathogenic (non-viral) microorganisms in a host organism.

2. Materials and methods

2.1. Defining the gold standard

We used data from 12 HDF screening studies listed in the GenomeRNAi database [24]. In total, we collected a list of n = 835 HDF (Table 1). The complete list of HDF is provided in Table S1 in the supplementary material. Fig. S1 shows the overlap of the HDF from these 12 studies. According to the number of studies an HDF was found, we defined four different gold standards (GS) ranging from low (GS-1-out-of-12-HDF), moderate (GS-2-out-of-12-HDF), elevated (GS-3-out-of-12-HDF) to high (GS-4-out-of-12-HDF) stringency. GS-1-out-of-12-HDF contained genes (n = 835) which had been found in at least one of the 12 studies. GS-2-out-of-12-HDF contained 123 genes identified as HDF in at least 2 studies. GS-3-out-of-12-HDF and GS-4-out-of-12-HDF contained 44 genes and 15 genes identified in at least 3 and 4 studies, respectively. To avoid ambiguity in the gold standard of the list of non-HDF genes (class of the negative controls), for all stringencies, we listed a gene as a non-HDF if (1) it was part of at least one screen, and (2) was not identified as an HDF in any of the twelve studies, resulting in a list of 13,074 non-HDF.

2.2. Feature generation

A main hypothesis of this study was, that a broad collection of intrinsic and extrinsic gene and protein features enables predicting host factors for pathogen infection in eukaryotes. A total of 50,334 features were generated based on broad range of features derived from (1) gene sequence, (2) protein sequence, (3) functional domains of the proteins, (4) gene sets from Gene Ontology (GO), (5) the number of homologous sequences, (6) topology properties from protein-protein interaction networks, and (7) subcellular localization of the protein (Fig. 1B). Protein and gene sequences were downloaded from Ensembl [33,34] using BioMart [35]. For deriving the protein and gene sequence features (features in categories 1 and 2), various numerical representations characterizing the nucleotide and amino acid sequences and compositions of the query genes were calculated using seqinR [36], protr [37], CodonW [38] and rDNAse [39]. Using seqinR [36] the number and fraction of individual amino acids and other protein sequence features including the number of residues, the percentage of physico-chemical classes and the theoretical isoelectric point were calculated. Further protein sequence features were obtained using protr [37] including autocorrelation, Conjoint Triad Descriptors (CTD), quasi-sequence order and pseudo amino acid composition. CodonW [38] was used to calculate gene characteristics like gene length and GC content but also frequencies of optimal codons (frequency of codons favored by natural selection, see [40]) and the effective number of codons. Using rDNAse [39] gene descriptors like auto covariance or pseudo nucleotide composition, and *kmer* frequencies (n = 2–7) were calculated.

The feature *seq.attribute.distribution* describes the distribution of amino acid attributes in the protein sequence. Amino acids were categorized into three classes according to their attributes. There are seven attributes used in this feature. These are (1) hydrophobicity, (2) normalized van der Waals volume, (3) polarity, (4) polarizability, (5) charge, (6) secondary structure, and (7) solvent accessibility. These attributes were represented by the first digit in the feature name. The second digit represented the class the amino acids belong to, either (1) polar, (2) neutral or (3) hydrophobic. The last three digits were the “distribution descriptor” describing the location of the attribute in the sequence. There are five “distribution” descriptors for each attribute together with their location, i.e. either at the beginning of the sequence (000), around the 25% quantile of residues (025), 50% (050), 75% (075), or at the end of the sequence (100). For example, seq.attribute.distribution. 51000 is the sequence attribute of amino acids having a charge (5), being polar (1) and are located at the beginning of the sequence (000).

For deriving domain features (feature category 3), BioMart was used to obtain protein family (*pfam*) domains, number of coiled coils, the prediction of membrane helices, post-translational modifications, β-turns, cofactor binding, acetylation and glycosylation sites, trans membrane helices and signal peptides. In addition, the number and lengths of UTRs were obtained using BioMart. For features obtained from gene sets defined by Gene Ontology (feature category 4), gene sets of all GO terms including biological process, cellular localization and molecular function were obtained from Ensembl (version 102, released in Nov 2020) [33,34]. Gene sets were removed if they showed high redundancy according to

Table 1
The experimental studies for our gold standards.

Name of the study[*]	Host cell system or organism	Pathogens	Num-ber of HDFs	Number of silenced genes	Method	Reference
Agaisse	SL2 cells	*Listeria mono-cytogenes*[**]	207	~21,300	dsRNAs	[25]
Akimana	S2R+ cells	*Francisella tularensis*[**]	197	~21,300	dsRNAs	[9]
Cheng	S2 cells	*Listeria mono-cytogenes*[**]	82	7216	dsRNAs	[2]
Derre	SL2 cells	*Chlamydia caviae*[**]	175	16,000	dsRNAs	[26]
Ragab	SL2 cells	*Escherichia coli*[**]	34	~21,300	dsRNAs	[12]
Cronin	Gut epithelium, hemocytes	*Serratia marcescens*[**]	97	10,689	Mutant fly lines	[27]
Qin	S2 cells	*Brucella melitensis*[**], *B. abortus*[**]	50	370	dsRNAs	[28]
Philips	S2 cells	*Mycobacterium fortuitum*[**]	83	21,300	dsRNA	[29]
Brandt	*D. melanogaster*, whole organism	*Plasmodium gallinaceum*[***]	14	1452	Mutant fly lines	[30]
Kutten-keular	*D. melanogaster*, whole organism	*Escherichia coli*[**], *Micrococcus luteus*[**]	19	1033	dsRNAs	[10]
Pielage	S2 cells	*Pseudomonas aeruginosa*[**]	28	80	dsRNAs	[31]
Peltan	S2 cells	*Leishmania donovani*[***], *L. major*[***]	34	1920	dsRNAs	[32]

[*] We named the studies according to the name of the first author.
[**] Bacteria, [***] Protozoa.

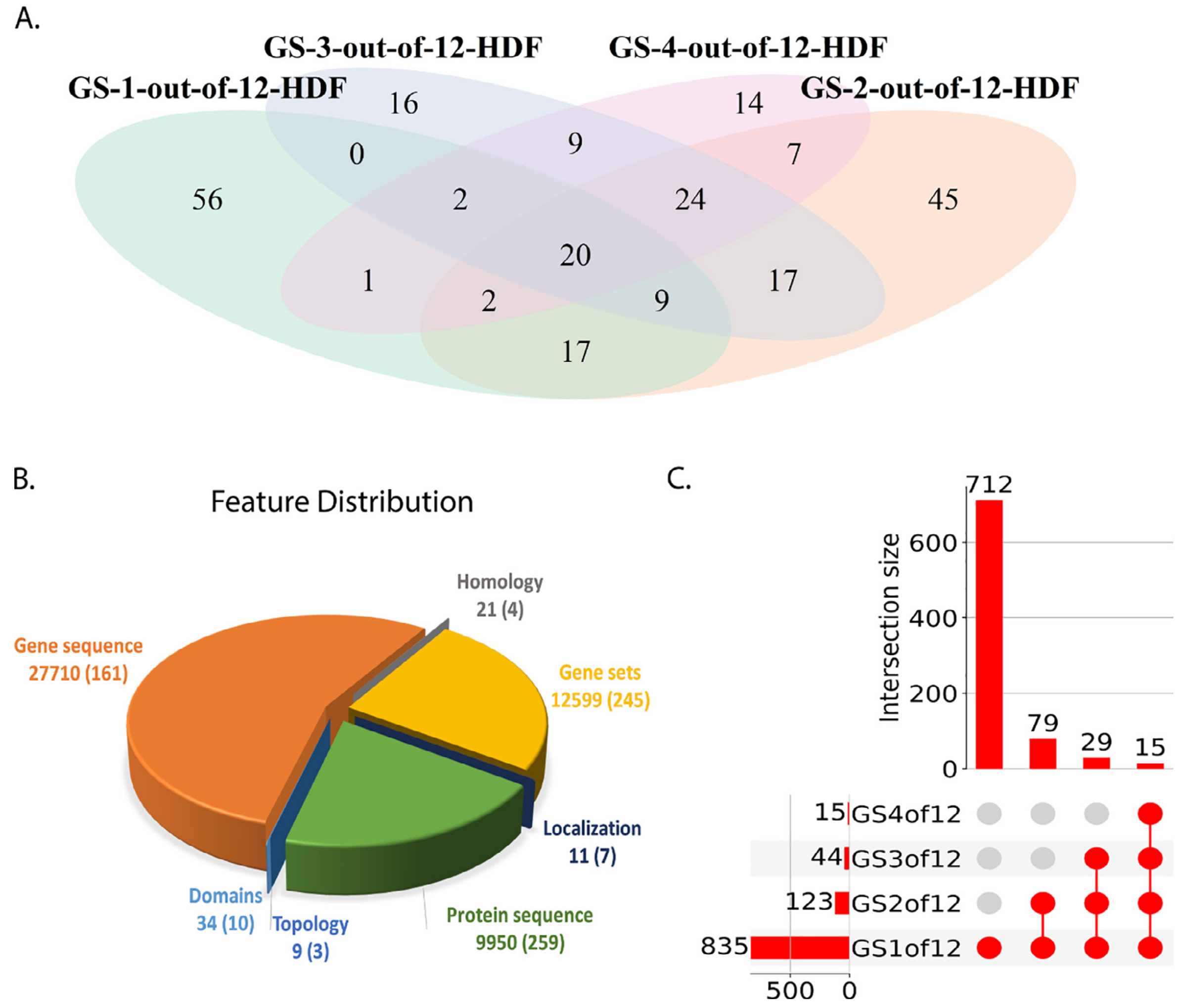

Fig. 1. Statistics of the predictions, the gold standards and the features. (A) Overlap of the predictions from the four classifiers. GS-1-out-of-12-HDF, GS-2-out-of-12-HDF, GS-3-out-of-12-HDF, GS-4-out-of-12-HDF predicted 107, 351, 293 and 191 HDF, respectively. (B) Distribution of the gene features according to the seven major categories. The values in parentheses indicate the number of features selected for machine learning. (C) Visualization of the overlapping HDF among the different investigated gold standards GS-1-out-of-12-HDF (GS1of12) to GS-4-out-of-12-HDF (GS1of12).

the following method. The gene overlap of each pair of gene sets A and B was quantified by Jaccard similarity coefficients,

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{1}$$

Pairs with J(A, B) above a threshold (threshold = 0.3) were included in the model and represented as an undirected graph, G = (X, E), with the gene sets as vertices X and the pairs above the threshold as edges E. A linear model was set up with a constraint to select at most one of the vertices of an edge:

$$\mathrm{X_i} + \mathrm{X_j} \leq 1, \quad \mathit{for\ every}\ \{\mathrm{i,j}\} \in \mathrm{E} \tag{2}$$

$$X_i = 0,\ \text{or}\ X_i = 1, \quad \mathit{for}\ 1 \leq \mathrm{I} \leq n \tag{3}$$

with the objective function to *maximize*

$\Sigma w_i X_i$

where w_i is the weight of a gene set. The weight is derived from its significance (p-value) and calculated as $1 - log10(p\text{-}value)/100$. This maximization was done employing linear integer programming solved using Gurobi (version 7.5.1, https://www.gurobi.com). With this, we formulated the optimization problem to select at most one gene set from each pair in such a way that the overall number of non-redundant gene sets was maximized. This optimization problem was formulated as a mixed integer linear programming problem and solved using Gurobi (version 7.5.1, https://www.gurobi.com). A gene list was generated for each query gene according to a protein association network obtained from the STRING database [41]. The gene list for a gene is the set of all adjacent genes in the protein association network. A gene set enrichment test was performed employing Fisher's exact test and the negative *log10* of the *p*-value was used as a feature.

The number of homologous proteins (feature category 5) was obtained by blasting the protein sequence of the query protein against the complete RefSeq database [42] using PSI-BLAST [43]. The number of proteins found with e-value cutoffs from 1e−5 to 1e−100 were used as features. Topology features (feature category 6) were computed using the NetworkX [44] library in Python. Protein association data was downloaded from STRING [41] and an undirected network was constructed. From this, degree, degree distribution, closeness centrality, eigenvalue centrality, betweenness centrality, harmonic centrality, subgraph centrality, load centrality and Page rank as topological features were computed for each gene. To note, the harmonic centrality of a node *g* is the sum of the reciprocal of the shortest path distances from all other nodes to *g*. The higher the value, the higher the centrality [45]. The subcellular localization of proteins (feature category 7) was derived using DeepLoc [46]. DeepLoc predicts the likely location of a protein within a cell by assigning probability scores to eleven eukaryotic cell compartments (cytoplasm, nucleus, extracellular, mitochondria, plasma membrane, ER, chloroplast, Golgi apparatus, lysosome, vacuole and peroxisome). In total we generated 50,334 features.

2.3. Machine learning

The machine learning procedure is depicted in Fig. 2. Features with low variance were removed (n = 17,681 were removed) using sklearn.feature_selection.*VarianceThreshold* for Python (threshold = 0.01). To improve the training, *z*-score transformation was applied to all features. For cross validation, the dataset was split into training (9/10) and test sets (1/10). Using the training set, we performed two steps for feature selection prior to training of the machines. First, we applied an embedded approach based on Random Forests as suggested by Breiman *et al.* [47] for feature selection. Each tree in the forest was initialized by bootstrapping from the training data to train a baseline model. Its performance was estimated using the out-of-bag (OOB) samples from the training data. Then, the values of one feature was randomly shuffled, keeping all other features the same, yielding permutated data. The permutated dataset was applied to the learned model and its performance was evaluated. Finally, the difference between the benchmark score from the baseline model and the score from the permutated model was calculated to determine the importance of the feature [48]. By this, we ranked all features and selected the features with importance score ≥1 for training the downstream classifier. To avoid overfitting, collinearity was reduced by eliminating highly correlating features with Pearson's correlation coefficients r > 0.70 (step 2). When two features highly correlated, the feature that was less correlated with the target variable was removed [49,50]. A total of 22,889 redundant features were removed. Consequently, we were left with 9764 features after removing low-variance and redundant features. For parameter optimization, the training data was further divided into training and test data using 5-fold cross-validation of the GridSearchCV (a method found in scikit-learn) [51] in an inner loop to obtain optimal hyper-parameters for the classifiers. GridSearch creates a parameter grid where all possible combinations of the hyper-parameter values are evaluated to obtain the optimal hyper-parameter values.

Our data consisted of much more negative than positive class samples, specifically the ratio of dependency factors to non-dependency factors was 1:16. To address this, we used the Synthetic Minority Oversampling Technique (SMOTE) [52]. SMOTE is a frequently used sampling method that creates synthetic, non-duplicated samples of the minority class to balance the number of samples of the classes. For each sample of the minority class, SMOTE selects the *k*-nearest neighbors of the same class and randomly creates multiple synthetic samples between the observation and the nearest neighbors depending on the number of additional samples needed. Six different classification methods were tested to train the model. These classifiers included Random Forest (RF) [47], Extreme Gradient Boosting (XGB) [53], Light Gradient Boosting Model (LGBM) [48], Support Vector Machines (SVM) [54], Artificial Neural Networks (NNET) [55], and Logistic Regression (LREG) [56]. The hyper-parameter settings with the optimal performance was *n_estimators* = 600, *learning_rate* = 0.05, *num_leaves* = 32, *colsample_bytree* = 0.2, *reg_alpha* = 3, *reg_lambda* = 1, *min_split_gain* = 0.01 and *min_child_weight* = 40 for LGBM; *n_estimators* = 600, *max_depth* = 70, *min_samples_leaf* = 4, *min_samples_split* = 10 for RF; *n_estimators* = 600, *max_depth* = 70, *learning_rate* = 0.01, *subsample* = 0.8, *colsample_bytree* = 0.8 for XGB. Default parameter values were used for SVM (RBF kernel, squared L2 penalty was the regularization parameter). The *max_iter* parameter in NNET was set to 2000 and default parameters otherwise. For logistic regression, *elasticnet* penalty was set as the regularization parameter, the algorithm used for the optimization problem was *saga* and the *l1_ratio* was set to 0.5.

Similar data preprocessing techniques were applied to all the four datasets (using the four gold standards GS-1-out-of-12-HDF to GS-4-out-of-12-HDF) yielding four different machine learning models. To improve generalizability, we performed a stratified randomized 10-fold nested cross validation for GS-1-out-of-12-HDF analyses where 90% of the dataset were used for feature selection and training of the classifiers, and 10% for testing. A three-fold cross validation was used for GS-2-out-of-12-HDF, GS-3-out-of-12-HDF and GS-4-out-of-12-HDF due to the small number of positive samples in these datasets (see Fig. 1C) ensuring a reasonable number of positive samples in the test sets during cross-validation. In addition, we repeated these cross validations five times and averaged the results over these five independent runs for each

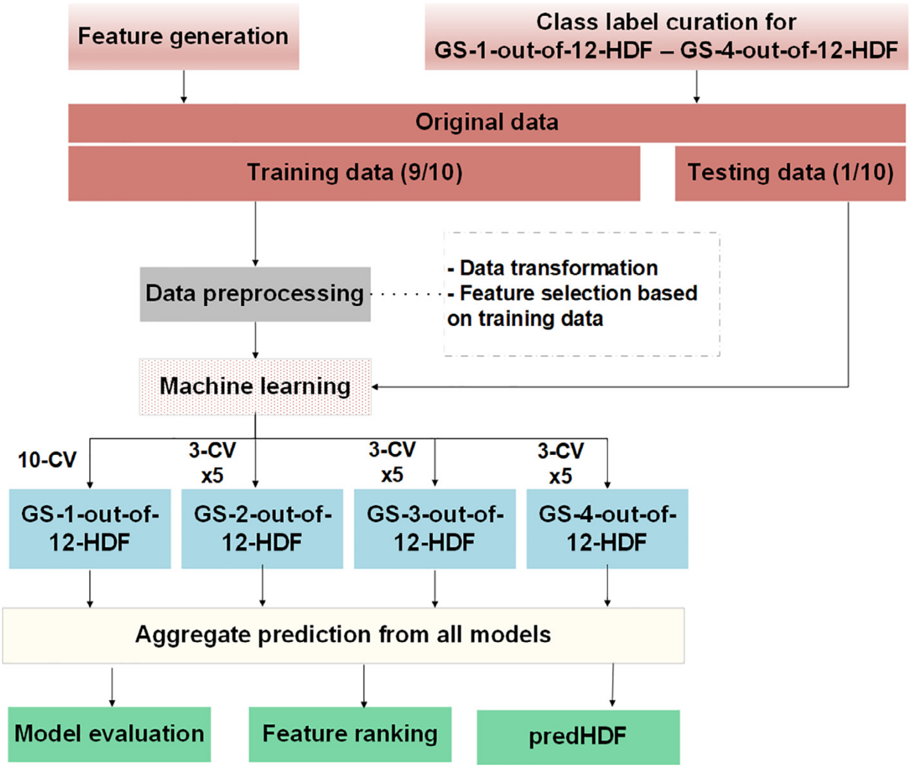

Fig. 2. Schematic overview of the machine learning pipeline. Features were generated from seven sources and four different gold standards (GS) ranging from low (GS-1-out-of-12-HDF), moderate (GS-2-out-of-12-HDF), elevated (GS-3-out-of-12-HDF) to high (GS-4-out-of-12-HDF) stringency. These gold standards were used to train and validate four different classifiers. Predictions from the four classifiers were linearly combined and ranked yielding a combined (aggregated) classifier. The trained classifiers were validated based on cross validation results, the most important feature determined and a list of predicted HDF was given out.

algorithm-dataset combination. To get a combined machine, the four models were *linearly* combined aggregating their predictions and their feature rankings leading to a single list of predictions and feature rankings. The total number of machines used for prediction based on the four gold standard datasets was 16. We performed a single-cross validation run for GS-1-out-of-12-HDF and five independent runs for GS-2-out-of-12-HDF, GS-3-out-of-12-HDF and GS-4-out-of-12-HDF based runs. The list of predicted HDF based on all four models was ranked by the average prediction probability score. For this, we ranked the genes based on the number of classifiers which predicted them as an HDF (first priority) and on the average predicted probability score (second priority). To obtain a list of the most discriminative features, the most important features were selected. For this, we computed the importance of each feature for each classifier employing the feature importance method for ensemble classifiers based on the bootstrapping approach described above [48]. The code for the machine learning procedure including feature generation can be found at the GitHub repository (https://github.com/phemmy2k2/HDF_codes).

2.4. Gene set enrichment analyses of the known and predicted HDF, and of human disease genes

Gene set enrichment analysis was performed using g:Profiler based on the Ensembl version 102 database [57]. The SCS algorithm with default settings was used to correct for multiple testing and the significance threshold was set to P = 0.05. The term size of the selected enriched gene sets was set between 3 and 500 to filter out too specific and too general gene sets. For the comparison of the gold standard and predicted HDF in human, homologous genes were identified using BioMart (for Section 3.4).

3. Results

3.1. Predicting HDF with good accuracy

To identify HDF in *D. melanogaster* by machine learning, 50,334 features from seven different categories were assembled based on protein and gene sequence, gene sets of genes with similar cellular functions or processes, topology of protein interaction networks, evolutionary conservation, functional domains of proteins and subcellular localization of the according proteins. Removing highly correlating and low varying features reduced the number of features to n = 9764. Due to the low overlap of HDF identified among the twelve screening studies (Fig. 1A), we assembled four different gold standards comprising low stringency (an HDF was identified in at least one of the 12 studies, denoted as GS-1-out-of-12-HDF), moderate (at least 2 studies identified an HDF in this list, denoted as GS-2-out-of-12-HDF), elevated and high stringency (at least three and four studies identified an HDF, denoted as GS-3-out-of-12-HDF and GS-4-out-of-12-HDF, respectively). Six machine learning algorithms (LGBM, LReg, NNET, RF, SVM, XGB) were applied to predict HDF. The results from the validation sets from cross-validation are shown in Fig. 3. The machine based on the gold standard with elevated stringency (GS-3-out-of-12-HDF) performed best (accuracy = 0.82), followed by GS-4-out-of-12-HDF (accuracy = 0.78). The classifiers produced good performance when applied to an independent test data set, yielding ROC-AUC = 0.936 and PR-AUC = 0.594 (Fig. 3C). Considering the average performances of the different algorithms across the four datasets, LGBM performed best (Fig. S2). Hence we used the results from LGBM for the further analyses. GS-4-out-of-12-HDF showed a gradual lower performance compared to GS-3-out-of-12-HDF (ROC-AUC = 0.927, Fig. 3D) which may have been due to the low

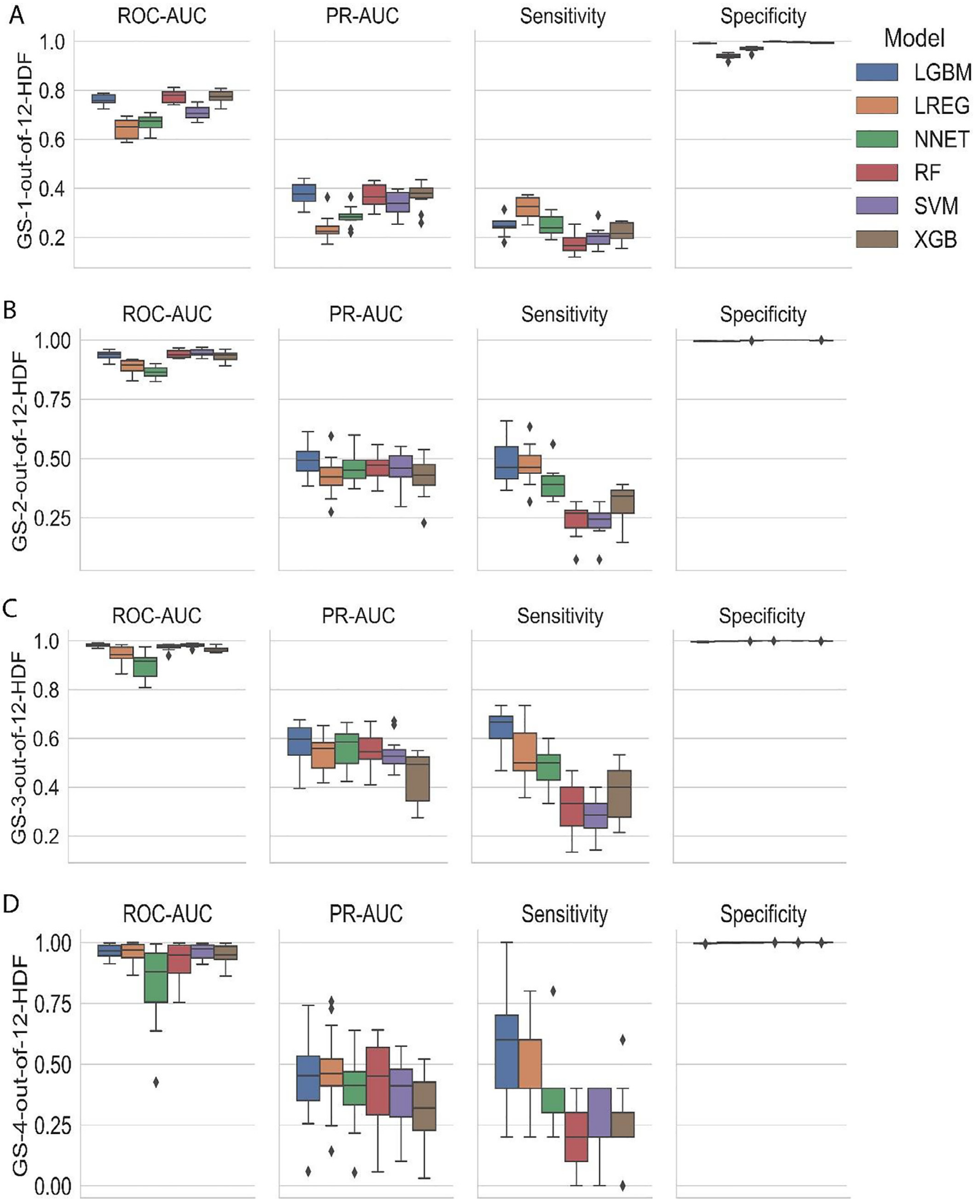

Fig. 3. Results of the machine learning prediction results on the validation sets of GS-1-out-of-12-HDF (A), GS-2-out-of-12-HDF (B), GS-3-out-of-12-HDF (C) and GS-4-out-of-12-HDF (D). We observed performance improvement when two or more studies listed a gene as HDF (GS-2-out-of-12-HDF, GS-3-out-of-12-HDF, GS-4-out-of-12-HDF). The best performance was observed for GS-3-out-of-12-HDF. Reduced performance was observed in the gold standard GS-4-out-of-12-HDF when compared to the gold standard GS-3-out-of-12-HDF, which may be due to the small number of HDF in this gold standard (n = 15).

number of HDF in this gold standard. Next, the classifiers were *linearly* combined to yield a single robust classifier. The combined classifier yielded an ROC-AUC = 0.76, PR-AUC = 0.348, sensitivity = 0.269, specificity = 0.982, and precision = 0.485. The ROC-AUC and sensitivity score of some of the individual classifier were higher than the combined classifier. In turn, the combined classifier yielded the best precision compared to all the individual classifiers (precision = 0.471, 0.462, 0.319 and 0.186 for classifiers GS-1-out-of-12-HDF to GS-4-out-of-12-HDF, respectively). As a good precision is valuable for experimental follow up analysis limiting the number of false positives, we used the results of the combined classifier for the following analyses. By this, 464 genes were predicted to be an HDF of which n = 225 were true positives (part of the gold standard GS-1-out-of-12), i.e. also identified in at least one of the twelve studies from Table 1, and n = 239 were novel predicted HDF.

Drug target investigations aim to identify HDF that are essential for the pathogens, but are not essential to the host cell or lethal to the organism when non-functional. Therefore, we compared the predicted HDF to genes annotated with a lethal loss-of-function phenotype across the developmental stages interrogating several genetic databases (Flybase [58], Database of Essential Genes (DEG) [59] and Online Gene Essentiality database (OGEE) [60]. n = 31 of the predicted HDF were found with a lethal loss-of-function phenotype in at least one of these databases and were hence excluded from our list of predicted HDF. In total, n = 208 predicted HDF (predHDF in the following) remained, listed in Table S2. To obtain a priority list of predHDF, we ranked the predHDF based on the number of classifiers which predicted them and their prediction scores. The top ten ranking predHDF are listed in Table 2 and a complete list is given in Table S2.

3.2. Identifying common cellular processes and functions of the predicted host dependency factors

We performed gene set enrichment analysis to elucidate common biological processes, molecular functions and cellular components of known and predicted HDF. There was a significant overlap (p < 0.0001) in the enriched gene sets of the HDF from the gold standard and the predicted HDF confirming common cellular processes of predicted and known HDF. 467 out of 745 gene sets (from Gene Ontology) of the predicted HDF were also found in the gold standard (Fig. 4B). For the predHDF, we found several transport processes, such as cytosolic and endosomal transport indicating the need for these specific cellular maintenance processes when the micro-organisms are inside the host cells, or Golgi organization and SNAP receptor activity mediating cellular uptake and release. Mitotic cell cycle was identified to be one of the most enriched gene sets of biological processes. Table S3 shows the list of 32 predHDF annotated in Gene Ontology to be involved in mitotic cell cycle. We were interested if we could enlarge the list of predHDF potentially playing a role in this biological process. For this, we compared the hit lists of two publically available gene knockdown screens observing genes being relevant for the cell cycle, performed by Dobbelaere et al. [61] and Goshima et al. [62]. We found further n = 7 genes being hits of these screens in our predHDF suggesting their involvement in cell cycle, as e.g. the genes *Rheb*, *Myb*, *Raptor* (the complete list and a Venn diagram is given in the supplementary material, Table S4, Fig. S4 respectively). Interestingly, several neural related annotated processes were highly enriched in the list of predHDF, such as neuron maturation, retrograde transport, axon, synaptic vesicle and distal axon (Fig. 4A) including several genes of RAB GTPases and Vacuolar Protein Sorting genes which will be discussed below (Discussion). For getting a more comprehensive view on the biology of known and predicted HDF, we compiled these two lists and performed gene set enrichment analysis on this combined list confirming the above described results. The most prominent gene sets of this combination are provided in Fig. 4C. In summary, we observed considerable consistency among the cellular processes and functions and components in which known and predicted HDF are involved, discussed in more detail in Discussion.

Table 2
The ten predicted HDF with the highest scores.

Gene symbol or ID	Gene description	Average predicted probability to be an HDF	Number of models predicting this gene as an HDF
CG41099	Metal ion binding	0.975	15
Auxilin	ATP binding; clathrin binding; protein kinase activity	0.953	15
Mig-2-like	GTP binding; GTPase activity	0.945	15
Secretory 22, Sec22	SNAP receptor activity	0.887	15
Rolled	Protein binding; JUN kinase activity; protein kinase activity	0.872	15
AP-1-2β	Clathrin binding; clathrin adaptor activity	0.908	14
Lrrk2	Protein kinase activity	0.902	14
Pten	Dynein complex binding	0.726	14
Ankyrin	Ion channel binding; spectrin binding; cytoskeletal anchor activity	0.723	14
Act88F	Involved in muscle thin filament assembly and skeletal myofibril assembly	0.933	13

3.3. Investigating the features with high discriminative power

To get an insight into the way how the machines identified HDF, we investigated the features with high discriminative power (obtained by high importance values). Features from all the seven categories constituted to the top 30 discriminative features supporting our approach to assemble features across such a broad spectrum. One feature from the protein category describing the attributes and location of amino acids in the protein sequence was the most important feature (*seq.attribute.distribution.51000*, Fig. S3) addressing proteins which sequences contain charged and polar residues among their first residues (details, see Methods). Amino acid attributes such as hydrophobicity, normalized van der Waals volume, polarity, polarizability, charge, secondary structure and solvent accessibility of the protein sequences were also highly important in discriminating HDF from non-HDF. Interestingly, this compares to a previous study in which these descriptors were essential to predict protein function [63]. Furthermore, harmonic and degree centrality from the topology features were among the highest ranking features (second and third, respectively). They were positively correlated to HDF indicating that HDF are often hubs. Another highly discriminative feature was *prob of N-in*, which is a domain feature that describes the total probability that the n-terminus of the protein is on the cytoplasmic side of the membrane. If the n-terminus of a transmembrane protein is on the cytoplasmic side upon pathogen entry and engulfment to form an endosome, the n-terminus was observed to be excluded from the endosome, making it available for ubiquitin tagging followed by degradation of the protein or entire endosome [64]. This observation reasons the negative correlation of *prob of N-in* to HDF observed in our model, which shows that the higher the total prob-

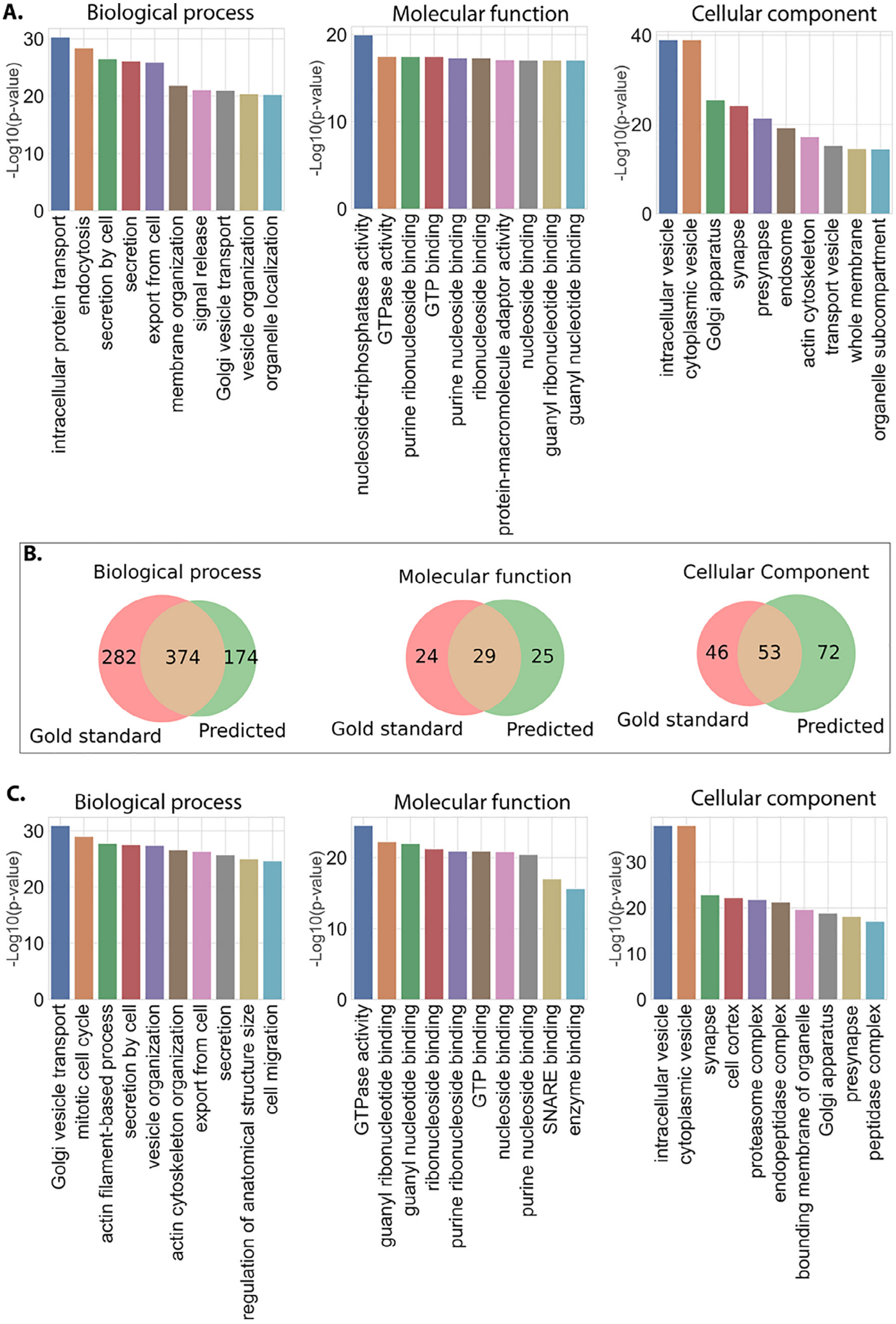

Fig. 4. Results from the gene set enrichment analyses. (A) Gene sets with the most significant enrichment in predicted HDF for the three Gene Ontology domains. (B) Overlap of enriched gene sets between the predicted HDF and the HDF of the gold standard. (C) Gene sets with the most significant enrichment in the predicted HDF together with HDF from the gold standard.

ability of a protein having the n-terminus on the cytoplasmic side of the membrane, the lower the probability of it to act as an HDF.

3.4. Comparing the involvement of HDF from the gold standard and the predicted HDF with a human trafficome screen, and a quantitative assessment of the literature

As described above, we identified several HDF in membrane trafficking (see also Discussion). We were interested how this compares to infected human cells. Kehl et al. [65] performed a focused screen knocking down genes of the trafficome in *Salmonella enterica* infected HeLA cells [65]. Indeed, when comparing our gold standard and our list of predHDF, we found a good overlap, specifically in the list of predHDF (n = 22, 10.6% of predHDF, compared to n = 21 genes, 2.5% of the gold standard, GS-1-out-of-12-HDF). The lists of common genes are given in the supplementary material (Table S6). Furthermore, we performed a statistical literature analysis to test if articles dealing with the predicted HDF were more often associated to infections than articles dealing with non-HDF genes. Hence, for each gene of predHDF (n = 208 genes) we counted the number of articles in PubMed selected by the gene symbol and the word ‘‘infection” and compared these numbers to the numbers of articles using an equal number of randomly selected non-HDF. Based on about 20 million records from Pubmed, predHDF were significantly more often associated with ‘‘infection” compared to non-HDF (P = 1.16 E-07, Wilcoxon rank test). Both computational analyses showed evidence suggesting our predictions to be indeed HDF.

4. Discussion

Due to the high heterogeneity of the gold standard, we investigated if an appropriate machine learning approach can learn distinguishing HDF from non-HDF based on a broad variety of gene features and four different gold standards according to a low, moderate, elevated and high stringency. By this, the machines could well recover these lists. The best performance was obtained for the elevated gold standard (GS-3-out-of-12-HDF) which may best balance between annotation quality of the class labels (HDF versus non-HDF) and the number of HDF. To predict the most precise set of new HDF, we combined all classifiers based on a voting scheme.

To elucidate if the predicted HDF show a consistent pattern to the biology on a statistical view, we performed three investigations. First, we compared their involvement in cellular processes and function with the genes from the gold standard. We found a good agreement. Next, we investigated their associations to diseases and found also similar diseases as for the genes of the gold standard (92% overlap). Thirdly, we performed a statistical literature analysis and found that predHDF were significantly more often associated with ‘‘infection” compared to non-HDF (P = 1.16 E−07). When searching for the predHDF in the literature, we found that many of the predHDF had been described to be important for pathogen infection in the host organism or host cell. Notably, most of these proteins were involved in membrane trafficking or signaling. In the following, we discuss the most interesting findings from our literature study.

A high ranking predHDF is Phosphatase and TENsin homolog (*PTEN*) (Table 1). Expression of *PTEN* was increased in *Trypanosoma cruzi* infected cells to about 300% higher levels compared to controls six hours after infection [66]. In addition, rat myoblast (H9c2 cells) transient transfected cells with rno-miR-190b inhibitor (miR-190b blocks PTEN translation) had increased rates of infection compared to non-transfected controls [66]. This suggests that *PTEN* is necessary for *T. cruzi* infection in host cells. Leucine-rich repeat kinase 2 (*Lrrk2*, FBgn0038816) is another high ranking predHDF (Table 1). *Lrrk2* (also known as *Lrrk*) is a multi-domain protein having two catalytic domains, a GTPase domain and a kinase domain [67]. The role of *Lrrk2* in pathogen infection or clearance might again depend on the pathogen and the host cell type. While *Lrrk2*$^{-/-}$ knockout mice showed increased susceptibility to *Listeria monocytogenes* and *Salmonella typhimurium* [68,69], in another study it was shown that *Lrrk2* deficiency in mice resulted in a significant decrease in *M. tuberculosis* infection [70]. Herbst and Gutierrez suggested that this discrepancy might be due to the different roles of *Lrrk2* in different cell types [71]. This further suggests that proteins acting as host dependency factors depend on the host cell type and/or the infecting pathogen.

Small GTPases regulate transport and fusion of membrane-bound compartments in a cell [72]. They play a central role during intracellular infections. We found *Rap1* as a predHDF. *Rap1* is a small GTPase and required for pathogen vacuole formation of an intracellular bacterial pathogen [73]. *Legionella pneumophila*, a gram negative bacterium which causes the Legionnaires’ disease, a severe pneumonia, exploits *Rap1* for intracellular replication and growth in mammalian macrophages and in the amoebae *Dictyostelium discoideum* [73,74]. *Rap1* is an important host component of the specialized membrane-bound compartment ‘‘Legionella-containing-vacuole” (LCV), within which this bacterium grows and evades the immune response of the host cells. Depletion of *Rap1* by RNAi has been observed to reduce intracellular replication of *L. pneumophila* [74]. LCV supports *L. pneumophila* to grow in host cells (using host cellular components) and prevents them to be cleared. LCV are also formed during *L. pneumophila* infection in *Drosophila* cells [75]. *Rap1* was also studied in infected Drosophila cells. Expression of activated *Rap1* has been found to mimic the effect of the enzymatically active A subunit of Cholera toxin (*CtxA*) in *Drosophila* leading to the reduced expression of *Rab11 (Rab11* was in the gold standard), *Sec15-GFP* and *Delta* (a Notch ligand) [76]. It was noted that *CtxA* exerts its toxic activity by binding the host co-factor *GTP-ARF6* leading to a cascade of signaling events which results in increased *cAMP* concentration. *cAMP* exerts its effects through protein kinase A (*PKA*) and *Epac* (a guanine nucleotide exchange factor that activates *Rap1*). Consequently, *CtxA* activated the expression of *Rap1* to reduce notch signaling and led to increased *V. cholerae* infection [76].

ROCK/Rok (FBgn0026181) is another predicted HDF. Its activity is required for membrane bleb formation and its activation is mediated by Transforming Growth Factor beta (*TGF-β2*), needed for augmented invasiveness of *Theileria* in susceptible Holstein-Friesian macrophages [77]. Another study linked the activity of *ROCK* to contractile force generation, a process necessary for infected cell motility during *Theileria annulata* infection [78]. This suggests *ROCK* to act as a host dependency factor during *Theileria* infection.

We found high enrichment of genes of the Rab GTPase binding protein family in our list of predHDF (*Rab3, Rab9, Rab14, Rab18, Rab40, RabX1, RabX4, RabX5, RabX6*) and the gold standard (*Rab1, Rab2, Rab4, Rab5, Rab7, Rab8, Rab10, Rab11, Rab21, Rab35,*) suggesting their central role for pathogens. Small GTPases belonging to the *Rab* family play an important role in membrane trafficking [79] and intact membrane trafficking in bacteria is crucial for host cell interaction and virulence. A study by Seixas and colleagues [80] examined how bacteria and protozoa modulate the expression of Rab proteins in mouse macrophages. In their study, *Rab9*, a late endosomal Rab protein involved in retrograde trafficking was upregulated during *E. coli* and *Salmonella enterica* infection. It was observed that this increased expression hampered phagocytosis of these bacteria while silencing *Rab9* enhanced their phagocytosis. Similarly, in their study, increased expression of *Rab14* was observed during *Plasmodium berghei* infection [80]. *Rab14* plays an important role in endosomal recycling [81]. Increased expression of *Rab14* was associated with reduced phagocytosis of *P. berghei* and reduced expression of *Rab14* by RNAi led to a significant increase in phagocytosis of *P. berghei* [80]. This suggests that *P. berghei* upregulates host *Rab14* while *E. coli* and *S. enterica* upregulate *Rab9* to escape immune response and enhance their survival in host cells. In a different study, depletion of *Rab9* and *Rab14* reduced the intracellular growth of *S. enterica* [72]. During chlamydia infection, *Rab14* modulates the delivery of endogenously synthesized sphingolipids into the growing bacteria containing vacuole; interfering with Rab14 was observed to reduce bacterial replication and infectivity [82]. Upon this, *Mycobacterium tuberculosis* modulates *Rab14* to block phagosome maturation in infected macrophage cells [83]. This maintains the host cells in an early endosomal phase, preventing the recruitment of late endosomal/lysosomal degradative components, hence enabling the pathogens to escape clearance by host cells. Knockdown of *Rab14* relieved the maturation block, allowing phagosomes with live mycobacteria to progress into phagolysosomes. *Rab18* has been reported to mediate viral replication of classical swine fever virus, CSFV, in swine umbilical vein endothelial cells [84] as well as to mediate assembly and replication of hepatitis C virus [85,86]. It was observed that knockdown of *Rab18* reduced CSFV production while overexpression of *Rab18* increased CSFV production. Thus, *Rab18* was identified as a host factor required for CSFV RNA replication and capsid assembly through its interaction with the viral protein NS5A [84]. Similarly, *Rab18* has been noted as a key component in endosome-ER trafficking of the human polyomavirus BKPyV [87]. In addition, retention of *Rab18* in live Salmonella-containing enabled them to avoid transport to the lysosomes through late endosomes and aiding their proliferation [88]. *Rab3* has been found to co-localize with the bacterium *Neisseria meningitidis* in human lung cancer cells (Calu-3 cells) [89]. Here, it was observed that *N. meningitidis* recruits *Rab3,* a mediator of the host vesicular trafficking to the apical site of infection to aid its replication and survival in host cells. These studies suggest *Rab3*, *Rab18*, *Rab9, Rab14* to be host dependency factors and our analyses suggests future investigation of the entire family.

We identified endosomal transport, Golgi organization and retrograde transport to be highly enriched gene sets in predHDF (P = 6.15E-19, 1.56E-14 and 4.19E-12, respectively), and particularly, several vacuolar protein sorting (VPS) genes were identified as predHDF (*Vps26, Vps35, Vps39, Vps45, Vps52*), or were in the gold standard (*Vps2, Vps28, Vps4*). The vacuolar protein sorting retromer is a heterotrimer complex that mediates the endosome-to-Golgi transport of lysosomal hydrolases receptors [90] and endosomal trafficking processes [91] as e.g. the retrograde transport of specific cargo proteins from endosomes to the trans-Golgi network [92]. It is required by *Brucella* to escape lysosomal degradation in host cells and to establish its intracellular replicative niche [93]. Hence its components have been validated as host dependency factors required for *Brucella* infection. The VPS retromer is composed of *Vps26, Vps35*, and *Vps29* [90]. *Vps35* was predicted in this study as an HDF. Knockdown of *Vps35* significantly reduced *Brucella* infection in HeLa cells [93]. Similarly, silencing *Vps35* reduced intracellular replication of *Coxiella burnetii* [92]. In summary, several VPS proteins are known to act as HDF and we suggest also here further investigation of the entire family.

The most significantly enriched gene set for molecular function was "SNAP receptor activity". The SNARE (soluble-N-ethylmalei mide-sensitive-factor accessory-protein receptor) complex accounts for the major membrane fusion machinery and regulates membrane fusion [94,95]. SNARE complex proteins are crucial for infection of intracellular pathogens as they allow their internalization and establish their niche in the host cell. Several predHDF belong to the SNARE family including *Snap29*, *Sec22* and *Syx16. Sec22* was one of the highest ranking predHDF (Table 1). *Sec22* participates in endoplasmic reticulum (ER)-Golgi trafficking. It is localized in the LCV during *L. pneumophila* infection [96]. Although depleting *Sec22* alone in Drosophila host cells did not reduce *L. pneumophila* replication, depleting a combination of *Sec22* and *Arf1* or members of the transport protein particle (TRAPP) complex, *Bet3*, *Trs23* (both listed in the gold standard) and *Bet5*, markedly reduced *L. pneumophila* replication [75]. *Syx16* participates in the *StxB* retrograde transport and its inhibition prevents *StxB* transport [97].

In the presented approach, the machines learned from a gold standard composed of a comprehensive but quite heterogeneous dataset of twelve screens. These comprised smaller screens of less than 100 genes up to large scale genome wide screens consisting of more than 20,000 genes, of screens investigating cell lines (10 out of 12) and whole organisms of *D. melanogaster* (2 out of 12), and very different studied pathogens, most of which invading the host cells, but some of them not obligatory. Interestingly, we observed that the machines indeed learned and made sense out of these heterogeneous data paving the way for a generic understanding of the need of host factors of infecting pathogens. Restricting to more homogenous datasets may have advantaged from observing a more consistent biology, but, may have drawbacked from higher variance due to less experimental data. We compared the results of our classifier with a more homogenously composed gold standard comprising only data of experiments from (i) Drosophila cell lines, (ii) invading pathogens, and (iii) of large genome wide screens (restricting to the HDF from the studies Agaisse, Akimana, Cheng, Derre and Philips). We found quite comparable results. Compared to n = 225 true positive genes of the complete list of predicted HDF (n = 464), a quite comparable number of n = 190 genes was found in this more homogenous list of experimentally found HDF (Table S5). Still, we suggest further investigations comparing HDF identified in cell lines versus HDF identified in whole organism, and HDF of invasive compared to non-invasive pathogens. To identify drug targets specifically harming the infecting pathogen while keeping the host safe implies avoiding targeting essential genes. In principle, the gold standard data from which we learned was based on experiments of viable cells and organisms after knockdown/-knockout. In our list of 225 predicted HDF we found only 17 genes to be absolute essential (according to the definitions of DEG, OGEE and FlyBase) and removed them from our final list. However, the definition of absolute essential of the investigated databases is very stringent, as such a gene was observed to be essential in each part of the life cycle. In a real setting, one may be interested in genes being essential in only a very focussed part of the life cycle, e.g. in an adult, or child enlarging this set of essential genes. This analysis was out of scope of the presented pilot study and we suggest this as future research.

In summary, the combined model predicted HDF, which were not previously identified as HDF in *Drosophila melanogaster*. Homologs of many proteins predicted as HDF in this study are described in the literature as HDF in other organisms. Several of these proteins were involved in membrane trafficking. Pathogens secrete diverse effector proteins into host cells and manipulate their membrane and vesicle trafficking. More specifically, many pathogens suppress the transport from endosomal compartments to the trans-Golgi network [98]. This seems to be one of the hallmarks particularly for intracellular pathogens. It enables them to form vesicular structures in host cells, to establish and maintain an intracellular replicative niche within the host cell and to prepare for release and spreading. Our study computationally inferred key HDF for *D. melanogaster* guiding further experimental studies to confirm the novel candidates as host dependency factors, also in a human cell culture setting.

5. Conclusion

We show that host dependency factors in *Drosophila melanogaster* can be predicted with high confidence using machine learning. The prediction performance achieved here is attributed to an elaborated assignment of HDF information based on a list of several knockdown screens of infected cells or organisms of *D. melanogaster* and a comprehensive set of a large variety of informative predictive features. Besides confirming genes of the gold standard, a list of 208 genes predicted to be novel host dependency factors showed enrichment in common cellular processes to the gold standard and have been described as HDF in other organisms and cellular contexts. These predicted HDF are proposed for future experimental studies.

CRediT authorship contribution statement

Olufemi Aromolaran: Conceptualization, Methodology, Software, Data curation, Visualization, Writing - review & editing. **Thomas Beder:** Conceptualization, Software, Data curation, Writing - review & editing. **Eunice Adedeji:** Writing - review & editing. **Yvonne Ajamma:** Data curation, Writing - review & editing. **Jelili Oyelade:** Conceptualization, Supervision. **Ezekiel Adebiyi:** Conceptualization, Supervision, Writing - review & editing. **Rainer Koenig:** Conceptualization, Methodology, Supervision, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the Deutsche Forschungsgemeinschaft (https://www.dfg.de/) within the project KO 3678/5-1, and the German Federal Ministry of Education and Research (BMBF) within the project Center for Sepsis Control and Care (CSCC, 01EO1002 and 01EO1502).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2021.08.010.

References

[1] Murali TM, Dyer MD, Badger D, Tyler BM, Katze MG, De Boer RJ. Network-based prediction and analysis of HIV dependency factors. PLoS Comput Biol 2011;7(9):e1002164.

[2] Cheng LW, Viala JPM, Stuurman N, Wiedemann U, Vale RD, Portnoy DA. Use of RNA interference in Drosophila S2 cells to identify host pathways controlling compartmentalization of an intracellular pathogen. Proc Natl Acad Sci 2005;102(38):13646–51.

[3] Kim S-A, Vacratsis PO, Firestein R, Cleary ML, Dixon JE. Regulation of myotubularin-related (MTMR) 2 phosphatidylinositol phosphatase by MTMR5, a catalytically inactive phosphatase. Proc Natl Acad Sci 2003;100(8):4492–7.

[4] Brass AL et al. Identification of host proteins required for HIV infection through a functional genomic screen. Science (80-) 2008;319(5865):921–6.

[5] Zhou H, Xu M, Huang Q, Gates AT, Zhang XD, Castle JC, et al. Genome-scale RNAi screen for host factors required for HIV replication. Cell Host Microbe 2008;4(5):495–504.

[6] König R et al. Global analysis of host-pathogen interactions that regulate early-stage HIV-1 replication. Cell 2008;135(1):49–60.

[7] Monack DM, Mueller A, Falkow S. Persistent bacterial infections: the interface of the pathogen and the host immune system. Nat Rev Microbiol 2004;2(9):747–65.

[8] Younes S, Al-Sulaiti A, Nasser EAA, Najjar H, Kamareddine L, Drosophila as a model organism in host–pathogen interaction studies, Front Cell Infect Microbiol, 10, 2020.

[9] Akimana C, Al-Khodor S, Abu Kwaik Y, Aziz RK. Host factors required for modulation of phagosome biogenesis and proliferation of Francisella tularensis within the cytosol. PLoS ONE 2010;5(6):e11025.

[10] Kuttenkeuler D et al. A large-scale RNAi screen identifies Deaf1 as a regulator of innate immune responses in Drosophila. J Innate Immun 2010;2(2):181–94.

[11] Moser TS, Jones RG, Thompson CB, Coyne CB, Cherry S, Evans DH. A kinome RNAi screen identified AMPK as promoting poxvirus entry through the control of actin dynamics. PLoS Pathog 2010;6(6):e1000954.

[12] Ragab A, Buechling T, Gesellchen V, Spirohn K, Boettcher A, Boutros M. Drosophila Ras/MAPK signalling regulates innate immune responses in immune and intestinal stem cells. EMBO J 2011;30(6):1123–36.

[13] Burgner D, Jamieson SE, Blackwell JM. Genetic susceptibility to infectious diseases: big is beautiful, but will bigger be even better? Lancet Infect Dis 2006;6(10):653–63.

[14] Goff SP. Knockdown screens to knockout HIV-1. Cell 2008;135(3):417–20.

[15] Bushman FD, Malani N, Fernandes J, D'Orso I, Cagney G, Diamond TL, et al. Host cell factors in HIV replication: meta-analysis of genome-wide studies. PLoS Pathog 2009;5(5):e1000437.

[16] Aromolaran O, Beder T, Oswald M, Oyelade J, Adebiyi E, Koenig R. Essential gene prediction in Drosophila melanogaster using machine learning approaches based on sequence and functional features. Comput Struct Biotechnol J 2020;18:612–21.

[17] Wen Q-F, Liu S, Dong C, Guo H-X, Gao Y-Z, Guo F-B. Geptop 2.0: an updated, more precise, and faster Geptop server for identification of prokaryotic essential genes. Front Microbiol 2019;10:1236.

[18] Khan J, Wei JS, Ringnér M, Saal LH, Ladanyi M, Westermann F, et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nat Med 2001;7(6):673–9.

[19] Lee Y, Lee C-K. Classification of multiple cancer types by multicategory support vector machines using gene expression data. Bioinformatics 2003;19(9):1132–9.

[20] Kohlmann A, Schoch C, Schnittger S, Dugas M, Hiddemann W, Kern W, et al. Pediatric acute lymphoblastic leukemia (ALL) gene expression signatures classify an independent cohort of adult ALL patients. Leukemia 2004;18(1):63–71.

[21] Dhanasekaran SM et al. Delineation of prognostic biomarkers in prostate cancer. Nature 2001;412(6849):822–6.

[22] Getz G, Gal H, Kela I, Notterman DA, Domany E. Coupled two-way clustering analysis of breast cancer and colon cancer gene expression data. Bioinformatics 2003;19(9):1079–89.

[23] Sharma AK, Eils R, König R. Copy number alterations in enzyme-coding and cancer-causing genes reprogram tumor metabolism. Cancer Res 2016;76(14):4058–67. https://doi.org/10.1158/0008-5472.CAN-15-2350.

[24] Schmidt EE, Pelz O, Buhlmann S, Kerr G, Horn T, Boutros M. GenomeRNAi: a database for cell-based and in vivo RNAi phenotypes, 2013 update. Nucleic Acids Res 2013;41(D1):D1021–6.

[25] Agaisse H, Burrack LS, Philips JA, Rubin EJ, Perrimon N, Higgins DE. Genome-wide RNAi screen for host factors required for intracellular bacterial infection. Science (80-) 2005;309(5738):1248–51.

[26] Derré I, Pypaert M, Dautry-Varsat A, Agaisse H, Schneider DS. RNAi screen in Drosophila cells reveals the involvement of the Tom complex in Chlamydia infection. PLoS Pathog 2007;3(10):e155.

[27] Cronin SJF et al. Genome-wide RNAi screen identifies genes involved in intestinal pathogenic bacterial infection. Science (80-) 2009;325(5938):340–3.

[28] Qin Q-M, Pei J, Ancona V, Shaw BD, Ficht TA, de Figueiredo P, et al. RNAi screen of endoplasmic reticulum–associated host factors reveals a role for IRE1α in supporting Brucella replication. PLoS Pathog 2008;4(7):e1000110.

[29] Philips JA, Porto MC, Wang H, Rubin EJ, Perrimon N. ESCRT factors restrict mycobacterial growth. Proc Natl Acad Sci U S A 2008;105(8):3070–5.

[30] Brandt SM, Jaramillo-Gutierrez G, Kumar S, Barillas-Mury C, Schneider DS. Use of a Drosophila model to identify genes regulating Plasmodium growth in the mosquito. Genetics 2008;180(3):1671–8.

[31] Pielage JF, Powell KR, Kalman D, Engel JN, Isberg RR. RNAi screen reveals an Abl kinase-dependent host cell pathway involved in Pseudomonas aeruginosa internalization. PLoS Pathog 2008;4(3):e1000031.

[32] Peltan A, Briggs L, Matthews G, Sweeney ST, Smith DF, Kelly BL. Identification of Drosophila gene products required for phagocytosis of Leishmania donovani. PLoS ONE 2012;7(12):e51831.

[33] Yates AD et al. Ensembl 2020. Nucleic Acids Res 2020;48(D1):D682–8.

[34] Howe KL et al., Ensembl genomes 2020—enabling non-vertebrate genomic research, Nucleic Acids Res, 48(D1), D689–D695, 2020.

[35] Smedley D, Haider S, Ballester B, Holland R, London D, Thorisson G, et al. BioMart–biological queries made easy. BMC Genomics 2009;10(1):22. https://doi.org/10.1186/1471-2164-10-22.

[36] Charif D, Lobry JR, SeqinR 1.0-2: a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis, in Structural approaches to sequence evolution, Springer, 2007, pp. 207–232.

[37] Xiao N, Cao D-S, Zhu M-F, Xu Q-S. protr/ProtrWeb: R package and web server for generating various numerical representation schemes of protein sequences. Bioinformatics 2015;31(11):1857–9.

[38] Peden J. CodonW. Nottingham: Univ; 1997.

[39] Zhu M, Dong J, Cao D-S, rDNAse: R package for generating various numerical representation schemes of DNA sequences, 2016.

[40] Hershberg R, Petrov DA, Nachman MW. General rules for optimal codon choice. PLoS Genet 2009;5(7):e1000556.

[41] Szklarczyk D, et al., STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets, Nucleic Acids Res., 2018;47(D1):D607–D613.

[42] K. D. Pruitt, T. Tatusova, and D. R. Maglott, "NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins," Nucleic Acids Res., 2007;35(Database):D61–D65, doi: 10.1093/nar/gkl842.

[43] Altschul SF et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25(17):3389–402.

[44] Hagberg A, Swart P, Chult DS, Exploring network structure, dynamics, and function using NetworkX, Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2008.

[45] Boldi P, Vigna S. Axioms for centrality. Internet Math 2014;10(3-4):222–62.

[46] Almagro Armenteros JJ, Sønderby CK, Sønderby SK, Nielsen H, Winther O. DeepLoc: prediction of protein subcellular localization using deep learning. Bioinformatics 2017;33(21):3387–95.

[47] Breiman L. Random forests. Mach Learn 2001;45(1):5–32.

[48] Ke G, et al., Lightgbm: A highly efficient gradient boosting decision tree, Adv Neural Inf Process Systems, 2017;3146–3154.

[49] Dormann CF, Elith J, Bacher S, Buchmann C, Carl G, Carré G, et al. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. Ecography (Cop) 2013;36(1):27–46. https://doi.org/10.1111/j.1600-0587.2012.07348.x.

[50] Meloun M, Militký J, Hill M, Brereton RG. Crucial problems in regression modelling and their solutions. Analyst 2002;127(4):433–50. https://doi.org/10.1039/b110779h.

[51] Pedregosa F et al. Scikit-learn: machine learning in python. J Mach Learn Res 2011;12:2825–30.

[52] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 2002;16:321–57.

[53] Chen T, He T, Benesty M, Khotilovich V, Tang Y, Xgboost: extreme gradient boosting, R Packag. version 0.4-2, 2015, pp. 1–4.

[54] Cortes C, Vapnik V. Support-vector networks. Mach Learn 1995;20(3):273–97.

[55] Hopfield JJ. Neural networks and physical systems with emergent collective computational abilities. Proc Natl Acad Sci U S A 1982;79(8):2554–8.

[56] Tolles J, Meurer WJ. Logistic regression: relating patient characteristics to outcomes. JAMA 2016;316(5):533–4.

[57] Raudvere U, et al., g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update), Nucleic Acids Res, 2019;47(W1):W191–W198, doi: 10.1093/nar/gkz369.

[58] Thurmond J et al. FlyBase 2.0: the next generation. Nucleic Acids Res 2018;47(D1):D759–65.

[59] Luo H, Lin Y, Gao F, Zhang CT, Zhang R. DEG 10, an update of the database of essential genes that includes both protein-coding genes and noncoding genomic elements. Nucleic Acids Res 2014;42(D1):D574–80. https://doi.org/10.1093/nar/gkt1131.

[60] Chen W-H, Lu G, Chen X, Zhao X-M, Bork P. OGEE v2: an update of the online gene essentiality database with special focus on differentially essential genes in human cancer cell lines. Nucleic Acids Res 2017;45(D1):D940–4.

[61] Dobbelaere J, Josué F, Suijkerbuijk S, Baum B, Tapon N, Raff J, et al. A genome-wide RNAi screen to dissect centriole duplication and centrosome maturation in Drosophila. PLoS Biol 2008;6(9):e224.

[62] Goshima G et al. Genes required for mitotic spindle assembly in Drosophila S2 cells. Science (80-) 2007;316(5823):417–21.

[63] Govindan G, Nair AS, Composition, Transition and Distribution (CTD)—a dynamic feature for predictions based on hierarchical structure of cellular sorting, in 2011 Annual IEEE India Conference, 2011, pp. 1–6.

[64] Ciechanover A, Ben-Saadon R. N-terminal ubiquitination: more protein substrates join in. Trends Cell Biol 2004;14(3):103–6.

[65] Kehl A, Göser V, Reuter T, Liss V, Franke M, John C, et al. A trafficome-wide RNAi screen reveals deployment of early and late secretory host proteins and the entire late endo-/lysosomal vesicle fusion machinery by intracellular Salmonella. PLoS Pathog 2020;16(7):e1008220.

[66] Monteiro CJ, Mota SLA, Diniz LdF, Bahia MT, Moraes KCM. Mir-190b negatively contributes to the Trypanosoma cruzi-infected cell survival by repressing PTEN protein expression. Mem Inst Oswaldo Cruz 2015;110(8):996–1002.

[67] Gilsbach BK, Kortholt A. Structural biology of the LRRK2 GTPase and kinase domains: implications for regulation. Front Mol Neurosci 2014;7:32.

[68] Liu W et al. LRRK2 promotes the activation of NLRC4 inflammasome during Salmonella Typhimurium infection. J Exp Med 2017;214(10):3051–66.

[69] Zhang Q, Pan Y, Yan R, Zeng B, Wang H, Zhang X, et al. Commensal bacteria direct selective cargo sorting to promote symbiosis. Nat Immunol 2015;16(9):918–26.

[70] Härtlova A, Herbst S, Peltier J, Rodgers A, Bilkei-Gorzo O, Fearns A, et al. LRRK2 is a negative regulator of Mycobacterium tuberculosis phagosome maturation in macrophages. EMBO J 2018;37(12). https://doi.org/10.15252/embj.201798694.

[71] Herbst S, Gutierrez MG. LRRK2 in infection: friend or foe? ACS Infect Dis 2019;5(6):809–15.

[72] Kuijl C et al. Rac and Rab GTPases dual effector Nischarin regulates vesicle maturation to facilitate survival of intracellular bacteria. EMBO J 2013;32(5):713–27.

[73] Hilbi H, Kortholt A. Role of the small GTPase Rap1 in signal transduction, cell dynamics and bacterial infection. Small GTPases 2019;10(5):336–42.

[74] Schmölders J et al. Comparative proteomics of purified pathogen vacuoles correlates intracellular replication of Legionella pneumophila with the small GTPase Ras-related protein 1 (Rap1). Mol Cell Proteomics 2017;16(4):622–41.

[75] Dorer MS, Kirton D, Bader JS, Isberg RR. RNA interference analysis of Legionella in Drosophila cells: exploitation of early secretory apparatus dynamics. PLoS Pathog 2006;2(4):e34.

[76] Guichard A et al. Cholera toxin disrupts barrier function by inhibiting exocyst-mediated trafficking of host proteins to intestinal cell junctions. Cell Host Microbe 2013;14(3):294–305.

[77] Chaussepied M et al. TGF-b2 induction regulates invasiveness of Theileria-transformed leukocytes and disease susceptibility. PLoS Pathog 2010;6(11):e1001197.

[78] Ma M, Baumgartner M. Filopodia and membrane blebs drive efficient matrix invasion of macrophages transformed by the intracellular parasite Theileria annulata. PLoS ONE 2013;8(9):e75577.

[79] Stenmark H. Rab GTPases as coordinators of vesicle traffic. Nat Rev Mol cell Biol 2009;10(8):513–25.

[80] Seixas E, Ramalho JS, Mota LJ, Barral DC, Seabra MC. Bacteria and protozoa differentially modulate the expression of Rab proteins. PLoS ONE 2012;7(7):e39858.

[81] Stein M, Müller MP, Wandinger-Ness A. Bacterial pathogens commandeer Rab GTPases to establish intracellular niches. Traffic 2012;13(12):1565–88.

[82] Capmany A, Leiva N, Damiani MT. Golgi-associated Rab14, a new regulator for Chlamydia trachomatis infection outcome. Commun Integr Biol 2011;4(5):590–3.

[83] Kyei GB et al. Rab14 is critical for maintenance of Mycobacterium tuberculosis phagosome maturation arrest. EMBO J 2006;25(22):5250–9.

[84] Zhang L et al. Rab18 binds to classical swine fever virus NS5A and mediates viral replication and assembly in swine umbilical vein endothelial cells. Virulence 2020;11(1):489–501.

[85] Dansako H, Hiramoto H, Ikeda M, Wakita T, Kato N. Rab18 is required for viral assembly of hepatitis C virus through trafficking of the core protein to lipid droplets. Virology 2014;462:166–74.

[86] Salloum S, Wang H, Ferguson C, Parton RG, Tai AW. Rab18 binds to hepatitis C virus NS5A and promotes interaction between sites of viral replication and lipid droplets. PLoS Pathog 2013;9(8):e1003513.

[87] Zhao L, Imperiale MJ, Identification of Rab18 as an essential host factor for BK polyomavirus infection using a whole-genome RNA interference screen, Msphere, 2017;2(4).

[88] Hashim S, Mukherjee K, Raje M, Basu SK, Mukhopadhyay A. Live Salmonella modulate expression of Rab proteins to persist in a specialized compartment and escape transport to lysosomes. J Biol Chem 2000;275(21):16281–8.

[89] Barrile R et al. N eisseria meningitidis subverts the polarized organization and intracellular trafficking of host cells to cross the epithelial barrier. Cell Microbiol 2015;17(9):1365–75.

[90] Verges M, Retromer in polarized protein transport, in International Review of Cell and Molecular Biology, vol. 323, Elsevier, 2016, pp. 129–179.

[91] Collins BM. The structure and function of the retromer protein complex. Traffic 2008;9(11):1811–22.

[92] McDonough JA, Newton HJ, Klum S, Swiss R, Agaisse H, Roy CR, Host pathways important for Coxiella burnetii infection revealed by genome-wide RNA interference screening, MBio, 2013;4(1).

[93] Casanova A et al. A role for the VPS retromer in Brucella intracellular replication revealed by genomewide siRNA screening. Msphere 2019;4(3):e00380–19.

[94] Stow JL, Manderson AP, Murray RZ. SNAREing immunity: the role of SNAREs in the immune system. Nat Rev Immunol 2006;6(12):919–29.

[95] Matte C, Descoteaux A. Exploitation of the Host cell membrane fusion machinery by leishmania is part of the infection process. PLoS Pathog 2016;12(12):e1005962.

[96] Kagan JC, Stein M-P, Pypaert M, Roy CR. Legionella subvert the functions of Rab1 and Sec22b to create a replicative organelle. J Exp Med 2004;199(9):1201–11.

[97] Wang Y, Tai G, Lu L, Johannes L, Hong W, Luen Tang B. Trans-Golgi network syntaxin 10 functions distinctly from syntaxins 6 and 16. Mol Membr Biol 2005;22(4):313–25.

[98] Personnic N, Bärlocher K, Finsel I, Hilbi H. Subversion of retrograde trafficking by translocated pathogen effectors. Trends Microbiol 2016;24(6):450–62.