Computational and Structural Biotechnology Journal xxx (xxxx) xxx

ELSEVIER

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

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

# Predicting gene essentiality in *Caenorhabditis elegans* by feature engineering and machine-learning

Tulio L. Campos [a,b], Pasi K. Korhonen [a], Paul W. Sternberg [c], Robin B. Gasser [a,*], Neil D. Young [a,*]

[a] Department of Veterinary Biosciences, Melbourne Veterinary School, The University of Melbourne, Parkville, Victoria 3010, Australia
[b] Instituto Aggeu Magalhães, Fundação Oswaldo Cruz (IAM-Fiocruz), Recife, Pernambuco, Brazil
[c] Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, CA, United States

ARTICLE INFO

*Article history:*
Received 23 March 2020
Received in revised form 1 May 2020
Accepted 6 May 2020
Available online xxxx

*Keywords:*
*Caenorhabditis elegans*
Machine-learning
Essential genes
Essentiality predictions

ABSTRACT

Defining genes that are essential for life has major implications for understanding critical biological processes and mechanisms. Although essential genes have been identified and characterised experimentally using functional genomic tools, it is challenging to predict with confidence such genes from molecular and phenomic data sets using computational methods. Using extensive data sets available for the model organism *Caenorhabditis elegans*, we constructed here a machine-learning (ML)-based workflow for the prediction of essential genes on a genome-wide scale. We identified strong predictors for such genes and showed that trained ML models consistently achieve highly-accurate classifications. Complementary analyses revealed an association between essential genes and chromosomal location. Our findings reveal that essential genes in *C. elegans* tend to be located in or near the centre of autosomal chromosomes; are positively correlated with low single nucleotide polymorphim (SNP) densities and epigenetic markers in promoter regions; are involved in protein and nucleotide processing; are transcribed in most cells; are enriched in reproductive tissues or are targets for small RNAs bound to the argonaut CSR-1. Based on these results, we hypothesise an interplay between epigenetic markers and small RNA pathways in the germline, with transcription-based memory; this hypothesis warrants testing. From a technical perspective, further work is needed to evaluate whether the present ML-based approach will be applicable to other metazoans (including *Drosophila melanogaster*) for which comprehensive data set (i.e. genomic, transcriptomic, proteomic, variomic, epigenetic and phenomic) are available.





## 1. Introduction

Model organisms, such as the free-living nematode *Caenorhabditis elegans*, have been utilised extensively to explore the biology of multicellular (metazoan) organisms [1–3]. The sequencing of the *C. elegans* genome [4] and subsequent development of functional genomics tools, such as double-stranded RNA interference (RNAi), transgenesis and, more recently, CRISPR/Cas9, combined with genetic mapping, have underpinned studies of gene function [5–9]. A key research focus has been to identify or define *genes* which are functionally essential for life in cells, tissues and/or the organism (thus called '*essential genes*') using such gene knock-down or knock-out approaches [7,10–12]. These efforts have led to a wealth of experimental data and information on essential genes, now publicly available in the WormBase database [13]. While these data are rich and highly informative, there have been some discrepancies in the assignment of gene essentiality among studies using phenotypic data. Such discrepancies can be due to some genes being 'conditionally-essential' [1] depending, for example, on developmental stage, strain or experimental/environmental conditions. However, it is also possible that some discrepancies might relate to possible off-target effects in RNAi [14] and/or human error during large-scale experiments [15]. Despite

*Abbreviations:* ML, machine-learning; RNAi, RNA interference; CRISPR, Clustered Regularly Interspaced Short Palindromic Repeats; SNP, single nucleotide polymorphism; CDS, coding sequence; TSS, transcription start site; EST, expressed sequence tag; VCF, variant call file; GFF, general feature format; ES, Essentiality Score; PPI, protein-protein interaction; SPLS, Sparse Partial Least Squares; GO, gene ontology; GLM, Generalised Linear Model; NN, Artificial Neural Network; GBM, Gradient Boosting Method; SVM, Support-Vector Machine; RF, Random Forest; ROC-AUC, Area Under the Receiver Operating Characteristic Curve; PR-AUC, Area Under the Precision-Recall Curve; TEA, Tissue Enrichment Analysis tool (WormBase).

* Corresponding authors.
*E-mail addresses:* robinbg@unimelb.edu.au (R.B. Gasser), nyoung@unimelb.edu.au (N.D. Young).

https://doi.org/10.1016/j.csbj.2020.05.008
2001-0370/

Please cite this article as: T. L. Campos, P. K. Korhonen, P. W. Sternberg et al., Predicting gene essentiality in *Caenorhabditis elegans* by feature engineering and machine-learning, Computational and Structural Biotechnology Journal, https://doi.org/10.1016/j.csbj.2020.05.008

such variation among experimental studies, there appears to be a consensus set of essential genes in *C. elegans*.

In recent years, computational approaches have been evaluated for the prediction of the complement of essential genes on a genome-wide scale employing functional genomic-phenotypic data sets for *C. elegans*. Such approaches could become important tools for predicting essential genes in less-studied organisms, such as many parasitic helminths, for which extensive genome, transcriptome and/or proteome data are available, but for which genome-wide functional genomic data have been lacking (e.g., [16,17]). Some studies of *C. elegans* data sets have used genome-wide genetic interaction networks [18,19] or single-nucleotide polymorphism (SNP) analyses [20,21]. Others have identified features, such as gene size, evolutionary rate, phyletic retention, transcription level, protein–protein interaction (PPI) network connectivity and/or cellular or subcellular localisation, which correlate with gene essentiality [1,22,23]. Despite the apparent utility or promise of these computational approaches, some discrepancies in experimental results among functional genomic studies, variation in the nature and extent of data sets used, and the limited curation of some data sets can markedly affect the confidence of predicting essential genes [1,24–26]. Here, we tackle this problem by employing a scoring-system to assign essentiality to genes from phenotypic data and by establishing procedures for large-scale extraction/engineering and selection of features associated with those genes from extensive 'omics data sets. Using these essentiality annotations and selected predictive features, we constructed and systematically evaluated a machine-learning (ML)-based workflow for the genome-wide prediction of essential genes in *C. elegans*.

## 2. Materials and methods

### 2.1. Data sets

We obtained extensive data and annotations from three sources (i.e. WormBase [27], the Ensembl database [28] and/or published studies). Functional genomic/phenomic data sets from RNAi studies and annotated data (genomic, transcriptomic, proteomic and epigenetic; in GFF) linked to the *C. elegans* genome were from WormBase (WS270 release – 25/02/2019) [27]. Genomic, coding sequences (CDSs) and proteins (canonical) were from Ensembl. Gene transcription data for different developmental stages [29]; transcription start site (TSS) locations in the genome [30]; multicell or single-cell transcriptomic data [31,32]; Ribo-seq annotations [33]; epigenetic markers (ChIP-seq and ATAC-seq) [34–36]; and variomic data containing genome-wide SNPs (high-quality VCF; release 20180527) [37] were obtained from the peer-reviewed literature.

### 2.2. Scoring of gene essentiality and provisional assignment

From WormBase, we extracted phenotypic data from all published RNAi studies of *C. elegans* and corresponding ontology terms using established scripts (see Data and Code Availability). We extracted all 'lethal' terms and their descendants from the phenotype_ontology.WS270.obo file and all 'not lethal' terms from the association file (phenotype_association.WS270.wb; column 4). We used the latter file to identify individual genes reported (in the peer-reviewed literature) to be linked to 'lethal' or 'not lethal' phenotypes upon RNAi. For each gene, we then calculated an essentiality score (ES), defined as the total number of RNAi experiments reporting essential/lethal (E) terms squared divided by the total number of experiments reporting essential/lethal and non-essential/viable terms (T) squared ($E^2/T^2$). A gene was provisionally assigned as "essential" (ES > 0.9) or "non-essential" (ES < 0.1); any other genes with an ES of $\geq$0.1 and $\leq$0.9 was assigned as "conditionally-essential".

### 2.3. Feature extraction or engineering

For individual genes, features were extracted from six (i.e. genomic, CDSs, overlapping-gene, transcriptomic, protein and 'variome') data sets derived from WormBase, Ensembl and/or published studies; see Data sets, above).

From genomic data, we extracted features including length, number of exons, distance from the chromosome centre (average distance between start codon of the first gene and the stop codon of the last gene in a chromosome), number of isoforms and presence/absence of associated Pfam-domains using "biomaRt" for R. From CDSs, we extracted nucleotide composition and correlation features using rDNAse (R package) as well as codon usage features using codonW (http://codonw.sourceforge.net).

For overlapping gene regions, we engineered new features (e.g., occurrence of chromatin state-domains; [34–36]) using the program BEDTools. The same approach was used to count features of overlapping genes defined in the GFF file (column 2) obtained from WormBase. In addition, we engineered additional features by establishing whether genes overlap outron- and/or exon-mapping transcription starting sites (TSS) (https://wormtss.utgenome.org) [30].

For 'pooled' transcriptomic data, we individually queried all designated 'essential', 'conditionally-essential' or 'non-essential' genes against the WormExp database, and then recorded the presence/absence of each gene in each of the first 30 returned data sets. For developmental transcriptomic data [29], we used the transcription levels of individual genes in each developmental stage as features. For single-cell transcriptomic data [31], we recorded the transcription level of each gene in each cell and enumerated the cells transcribing a particular gene.

From protein sequences, we extracted features using "protr" utilising all descriptors defined in this R package as well as the numbers of predicted transmembrane domains and signal peptides per protein employing TMHMM [38] and SignalP [39], respectively. We also obtained features from predicted protein subcellular localisations using WolfPsort [40] and DeepLoc [41] as well as protein disorder features employing DisEMBL [42].

For the variomes of *C. elegans* (variomics-natural file; see Data sets), we calculated the numbers of SNPs in individual genes using BEDTools"and inferred the effect(s) of individual SNPs on gene function using SnpEff [43] - these data were employed as features. The Ka/Ks ratio was calculated from the SnpEff output using an available script (https://github.com/MerrimanLab/selectionTools/blob/master/extrascripts/kaks.py). The data sets and code used to extract or engineer features are in the "R Markdown" script available at (https://bitbucket.org/tuliocampos/essential_elegans).

### 2.4. Feature sets

We combined all extracted/engineered features with respective genes essentiality annotations and stacked this information into a matrix using R. In this feature matrix, each line represented a gene, each column represented an extracted feature and the last column represented the essentiality annotation ("*essential*" or "*non-essential*"); this matrix contained all data ("FULL"). To create a non-redundant (NR) set of features, we first clustered protein sequences using USEARCH (parameters: -cluster_fast -centroids) [44], obtained gene identifiers and then removed genes and associated features if multiple amino acid sequences had $\geq$25% identity, retaining only the centroid sequences of all individual clusters. Subsequently, we removed features with low variance from both



Please cite this article as: T. L. Campos, P. K. Korhonen, P. W. Sternberg et al., Predicting gene essentiality in *Caenorhabditis elegans* by feature engineering and machine-learning, Computational and Structural Biotechnology Journal, https://doi.org/10.1016/j.csbj.2020.05.008

the "FULL and "NR" feature sets using the *nearZeroVar* method in "caret". For "FULL", we also assessed statistical differences in the features between "*essential*" and "*non-essential*" using two-tailed pairwise *t*-tests (95% confidence interval) in R (*t.test*), recording p-values and Holm-Bonferroni corrected (*p.adjust*) values.

### 2.5. Feature selection, ML training and performance assessment

eatures were selected by random subsampling from 10% to 90% of data representing "*essential*" or "*non-essential*" gene (in 10% stepwise increments) based on a consensus between elasticNet (alpha = 0.5) and ensemble Sparse Partial Least Squares (SPLS) methods using "glmnet" and "enspls" in R, respectively [26]. The features were then used to train each of six ML-models (GBM (Gradient Boosting Machine), GLM (Generalised Linear Model), NN (Neural Network - perceptron), Random Forest (RF), SVM (Support-Vector Machine) [26] and XGB (eXtreme Gradient Boosting – xgbTree) in the "caret" R-package. During the training process, we employed parameter-tuning and 5-fold cross-validation, ultimately selecting the models with highest ROC-AUC. Following subsampling, we employed the remaining data (90%–10%) to evaluate the performance of the final models using ROC-AUC and PR-AUC.

Subsequently, we trained each of the six ML-models with 100% of each set, and calculated the 'importance' of each feature for each ML algorithm for each feature set using the *varImp* method in the "caret" package. For each ML-model, we calculated ROC-AUCs using 5-fold cross-validation and plotted them against the parameters tested. We ranked the predictors according to the median feature-importance among the best three ML-models and selected 40 consensus-features that were highly predictive of gene essentiality employing the "FULL" or "NR" data set. Then, we assessed whether these consensus-features correlated with essentiality using "correlationfunnel", and evaluated pairwise correlations among features using "corrplot" (R). Using this reduced set of consensus-features (NR_SELECTED), we then trained the ML-methods and evaluated their prediction-performance using ROC-AUC and PR-AUC. Finally, we assessed variation in these metrics using bootstrapping (1000-times) employing 90% of the consensus-features used for training and the remaining 10% for testing.

### 2.6. Distribution of gene and SNPs on chromosomes

We counted the number of SNPs per each 1000 bp-window on each chromosome using published variomic data (high-quality VCF; release 20180527) [37]. We established the locations of genes provisionally assigned as "essential", 'non-essential' or 'conditionally-essential' (see Subsection 2.2) using the WormBase GFF file, and generated individual density plots showing the distribution of genes for each chromosome ("ggplot" for R). We compared the distributions of genes by essentiality annotation susing Kolmogorov-Smirnov tests (*ks.test* in R) [36].

### 2.7. Gene ontology (GO), transcription and tissue enrichment analyses

Using the GBM, RF and XGB methods trained with NR_SELECTED data, we identified 500 *C. elegans* genes with the highest median probabilities of being essential and then conducted gene ontology (GO), transcription and tissue enrichment analyses. For these 500 genes, GO enrichment (for biological process, molecular function and cellular component) was carried out using the Gene Set Enrichment Analysis available at WormBase [27], DAVID [45] and WebGestalt ('over-representation analysis') [46] databases, after which WormExp database/website [32] was interrogated for transcription enrichment. Then, we queried WormBase using the Tissue Enrichment Analysis (TEA) tool [47].

### 2.8. Validation of ML predictions using mutant allele data

First, we ranked all genes used in the present study by their final ML predictions (see Sub-section 2.5). Second, a list of all *C. elegans* genes with at least one report of a "lethal" phenotype in the GExplore database [48] was created. Third, we incrementally searched for all genes in GExplore, according to ML probability, in an descending and also in an ascending manner, and then calculated cumulative ratios. These ratios were displayed in a graph using "ggplot" in R.

## 3. Results

We built and then employed a well-defined workflow (Fig. 1) to: (i) annotate genes for essentiality from phenomic data; (ii) extract features predictive of gene essentiality; (iii) train and test ML approaches using selected features; (iv) locate essential genes and SNPs to locations on chromosomes; and (v) explore gene ontology (GO) and transcription enrichments linked to essential genes.

### 3.1. Annotating genes for essentiality from phenomic data

We first categorised sets of genes as 'essential', 'non-essential' or 'conditionally-essential' – with the latter category reflecting discrepant experimental results between or among published studies. For this categorisation, we inspected the hierarchical phenotype ontology for *C. elegans* (in WormBase), obtained 150 ontology-identifiers and then used them to calculate individual essentiality scores (ESs) (Table S1). Using these ESs, we provisionally assigned 670 genes in *C. elegans* as essential, 16,070 as non-essential, and 1721 as conditionally-essential using RNAi data sets (Fig. 2a; Tables S2–S4). A small percentage of genes annotated as essential (23 of 670; 3.4%) or non-essential (1616 of 16,070; 10%) were recorded as having both lethal/essential and viable/non-essential entries in the phenotype association file from WormBase. Most gene annotations were supported by results from at least three published RNAi experiments (via WormBase): 527 (78.6%) for essential, 13,579 (84.5%) for non-essential and 1592 (92.5%) for conditionally-essential.

### 3.2. Predictive features identified from multiple sources

For all individual genes annotated previously, 55,694 features were identified. Following the removal of features exhibiting low variance, 1609 features (per gene) were retained and used in subsequent analyses. After p-value correction (Holm-Bonferroni), 801 features displayed significant differences between essential and non-essential genes (Table S5. More than half (n = 416) of these features were from protein sequences, 193 from nucleotide sequences, 42 from transcriptomic data (from the WormExp database), 16 from SNP data, 14 related to subcellular localisation, 9 to single-cell RNA-seq (scRNA-seq) data, 5 from genomic locations or gene models, and 4 related to evidence of transcription in different developmental stages. In addition, we identified 102 predictive features that overlap with the genomic locations of genes, including 50 features derived from WormBase, 49 from epigenetic markers, 2 from transcription start sites (outron/exon) and 1 from Riboseq (Table S5).



Please cite this article as: T. L. Campos, P. K. Korhonen, P. W. Sternberg et al., Predicting gene essentiality in *Caenorhabditis elegans* by feature engineering and machine-learning, Computational and Structural Biotechnology Journal, https://doi.org/10.1016/j.csbj.2020.05.008

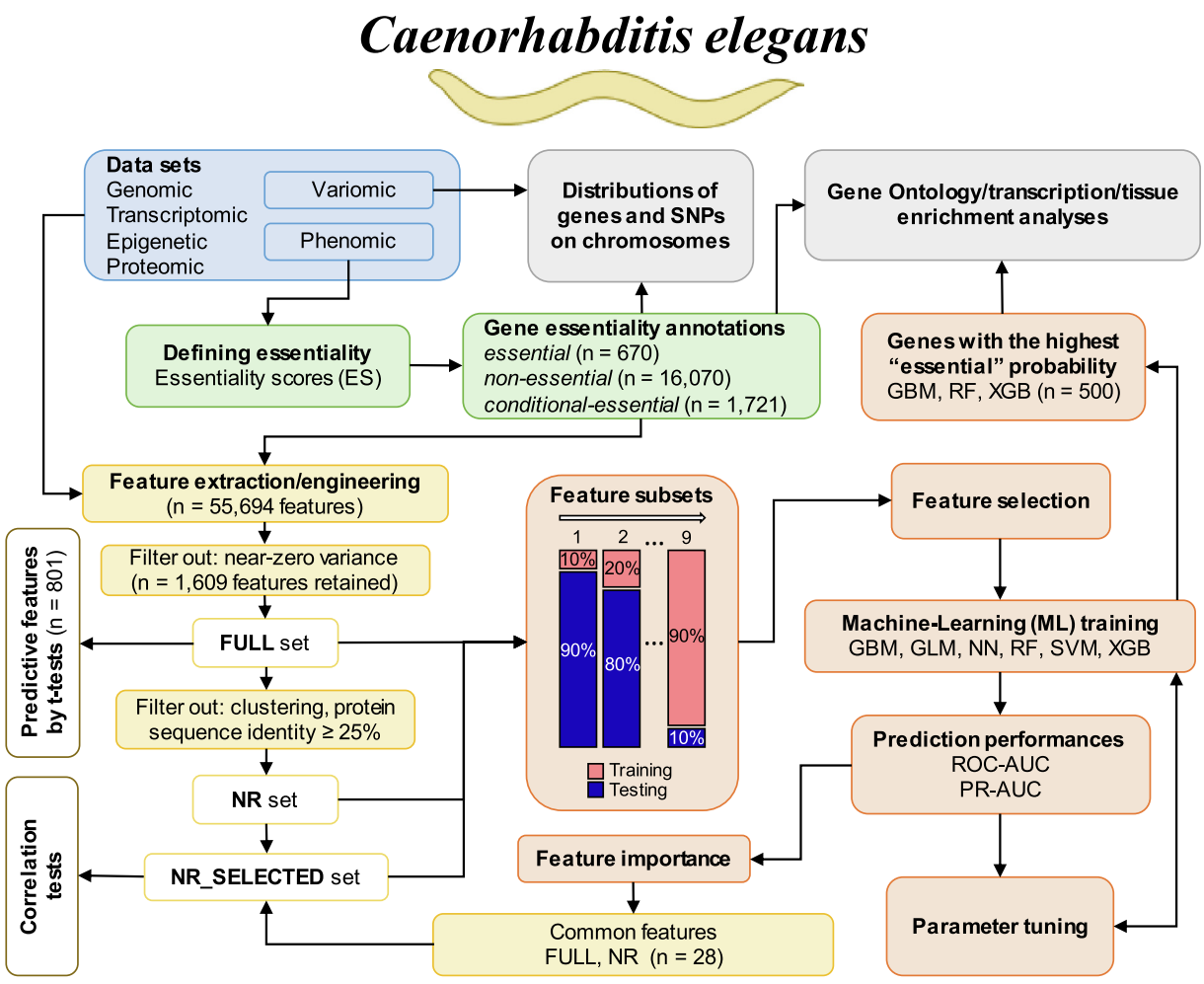


**Fig. 1.** Workflow employed in the present study. First, a wealth of publicly available 'omics data sets for *C. elegans* were obtained (blue). Then, we employed a 'scoring system' to the phenomic data to annotate *C. elegans* genes for essentiality (green). Next, we extracted or engineered features (yellow) from the data sets to establish feature sets (FULL – all features; NR – all features from sequences containing <25% amino acid identity; NR_SELECTED – 28 highly-predictive features of essentiality, selected from the NR data set). These feature sets were used for a systematic evaluation of machine-learning (ML) approaches for essential gene predictions (orange). *T*-tests and correlation tests were performed on the FULL and NR_SELECTED sets, respectively. The performances of the individual ML models, and the importance of the selected features for essentiality predictions were calculated and evaluated (orange). Finally, Gene Ontology (GO), transcription and tissue enrichments were performed, as well as an analysis on the preferential genomic locations of SNPs and genes by essentiality annotations (grey). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

### *3.3. Systematic feature selection, and training/evaluation of ML approaches*

First, we selected a complete (FULL) set of features from 'essential' and 'non-essential' genes (filtered) (n = 1609 per gene). Then, we used subsets of the FULL set (10–90% random samples) to train six individual ML methods (Gradient Boosting Machine, GBM; Generalised Linear Model, GLM; Neural Network, NN; Random Forest, RF; Support-Vector Machine, SVM; and eXtreme Gradient Boosting, XGB) to predict the same subsets, usually achieving high prediction performances (ROC-AUC of ~1 and PR-AUC of ~1; Fig. S1). Nonetheless, NN and GLM did exhibit a decrease in ROC-AUC (~0.97 and ~0.97, respectively) and in PR-AUC (~0.97 and ~0.8, respectively). Having trained individual ML methods, we then predicted gene essentiality from nine independent test-sets (not used for model training). Each of the six ML models achieved a high ROC-AUC of 0.94 to 1.0, with PR-AUCs of 0.75–0.95 for GBM, RF and XGB, and 0.65 to 0.76 for GLM, NN and SVM (Fig. S2). Only the latter model decreased PR-AUC as more data were added to individual training sets. Subsequently, we used the FULL set for the final selection of features and to train each of the six ML methods. Using this approach, we identified 418 predictors of gene essentiality, with the relative importance of these predictors being recorded for each model (Table S6).

Second, we created a non-redundant (NR) set of features by clustering protein sequences, retaining the centroid sequences with <25% identity representing all individual clusters. This NR dataset represented 615 essential and 12,193 non-essential genes, each having 1609 features. We employed this data set for the systematic selection of features as well as the training and testing of all six ML methods. The prediction performances of most ML models were commensurate with those achieved using the FULL data set (Fig. 2b – left), with SVM achieving a superior PR-AUC performance when trained using the NR set (Tables S6 and S7). Following feature-selection and training with the NR data set, 291 features were selected as the 'best' predictors of essentiality (representing a reduction of 30% compared with the FULL set).

Third, we established the minimum number (n = 40) of features that were highly-predictive for essentiality in the FULL or the NR data set (Fig. S3); 28 of these 40 features were shared between the two data sets. These highly-predictive features included: exon number; gene length; GC content; presence of an encoded signal

Please cite this article as: T. L. Campos, P. K. Korhonen, P. W. Sternberg et al., Predicting gene essentiality in *Caenorhabditis elegans* by feature engineering and machine-learning, Computational and Structural Biotechnology Journal, https://doi.org/10.1016/j.csbj.2020.05.008

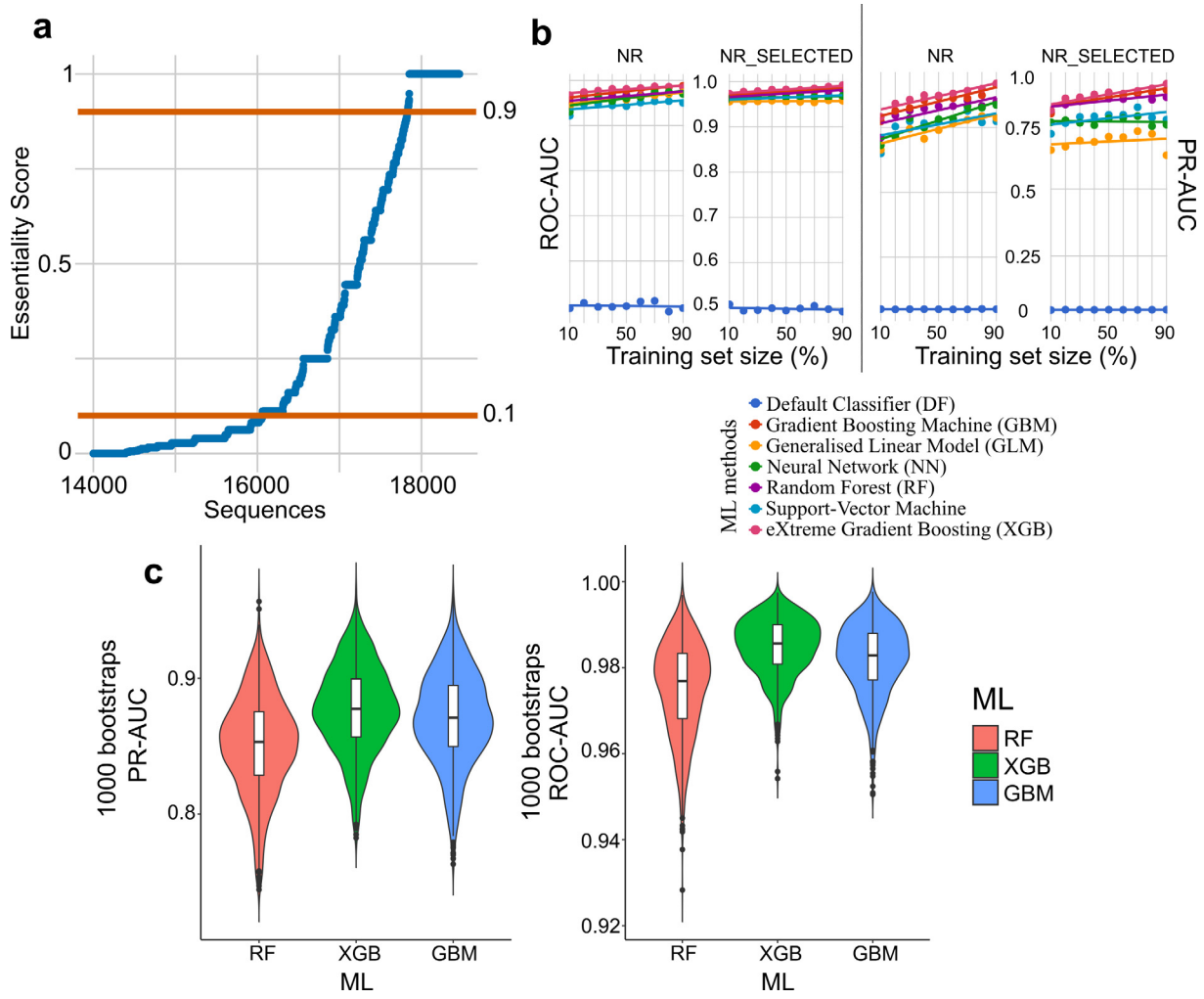


**Fig. 2.** Curation of essential genes from phenotype data and performance of ML methods for essentiality predictions. A. *C. elegans* genes were curated for essentiality using phenotype data available in WormBase. For each gene, an essentiality score (ES) was calculated (y-axis) and ordered using the formula $E^2/T^2$, were "E" is the number of entries relating to lethality/essentiality, and "T" is the total number of entries reported. Genes were annotated as 'essential' if ES was >0.9, or 'non-essential' if was ES <0.1, or 'conditionally-essential' otherwise. B. In the systematic evaluation of gene essentiality predictions ('essential' vs. 'non-essential') the performance of six machine-learning (ML) algorithms and a default classifier were assessed, initially with a data set (FULL) containing all genes curated previously and their features. In addition, a non-redundant (NR) data set with features from sequences that contained <25% amino acid sequence identity was created, and all features identified for these genes were included. Another data set containing the NR genes and a selection of 28 best-predictive features (NR_SELECTED) was also evaluated. For each data set, random subsets of genes (10–90%, 10% increments) were used as training sets (x-axis), and the remaining 90–10% used as independent test sets. At each step, the prediction performance was evaluated using the test set using ROC-AUC (right) and PR-AUC (left) metrics. C. Violin and box plots of ROC-AUC and PR-AUC from 1000 bootstraps of RF, XGB and GBM, with random sampling of 90% of the NR_SELECTED used for training and the remaining 10% of this feature set used for independent testing.

peptide; sequence characteristics (e.g., nucleotide sequence composition, which considers order and physiochemical properties [PseKNC_5_Xc1.CGT] or amino acid triads in a protein sequence [CTriad_VS115]); epigenetic chromatin-state markers relating to promoter regions, three of which associated with early embryo (EE_1, EE_2 and EE_3) and one in the third-stage larva (L3_2); subcellular localisation; expressed sequence tag (ESTs) 'best-hit' by BLAT (BLAT_Caen_EST_BEST in WormBase); RNAi probes (RNAi_primary) and peptide fragments from mass spectrometry (mass_spec_genome); scRNA-seq data (number of cells with transcription – num_cells_expressed) and transcription profiles of selected cells (e.g., cele.010.023.TCGTAGAGAA – in the germline) (Table S8).

Fourth, we assessed the correlation between 28 individual (highly-predictive) features and gene essentiality upon pairwise comparison (Fig. 3a). The correlations ranged between 0.1 and 0.35, showing that no single feature correlated perfectly with essentiality, which justified the use of multivariate methods for prediction using ML models. When we assessed the pairwise correlations among the 28 features (378 tests; Fig. 3b), most (>99%) values were between −0.5 and +0.5, and 12 (<1%) were >0.5. A strong correlation was recorded for chromatin-state markers in EE_1 to EE_3 and L3_2; num_cells_expressed; and scRNA-seq for cele.010.023.TCGTAGAGAA. Interestingly, num_cells_expressed also correlated positively with BLAT_Caen_EST_BEST, and the subcellular localisations 'cytoplasm' and 'nucleus' correlated negatively with 'endoplasmic reticulum' (Fig. 3b).

Fifth, we assessed the performances of the six individual ML models to predict essentiality from the NR data set using the final set of 28 highly-predictive features (NR_SELECTED data set). High ROC-AUCs (>0.95) were achieved for training sets. PR-AUCs were consistently ~1.0 for the XGB, GBM and RF models, compared with performances of ~0.98–0.85 for NN, 0.88–0.84 for SVM and 0.78–0.74 for GLM (Fig. 2b). For test sets, ROC-AUCs were >0.92 for all six ML models, and PR-AUCs were 0.85–0.96 for XGB, GBM and RF, and 0.65–0.77 for SVM, NN and GLM. An evaluation of the med-

Please cite this article as: T. L. Campos, P. K. Korhonen, P. W. Sternberg et al., Predicting gene essentiality in *Caenorhabditis elegans* by feature engineering and machine-learning, Computational and Structural Biotechnology Journal, https://doi.org/10.1016/j.csbj.2020.05.008

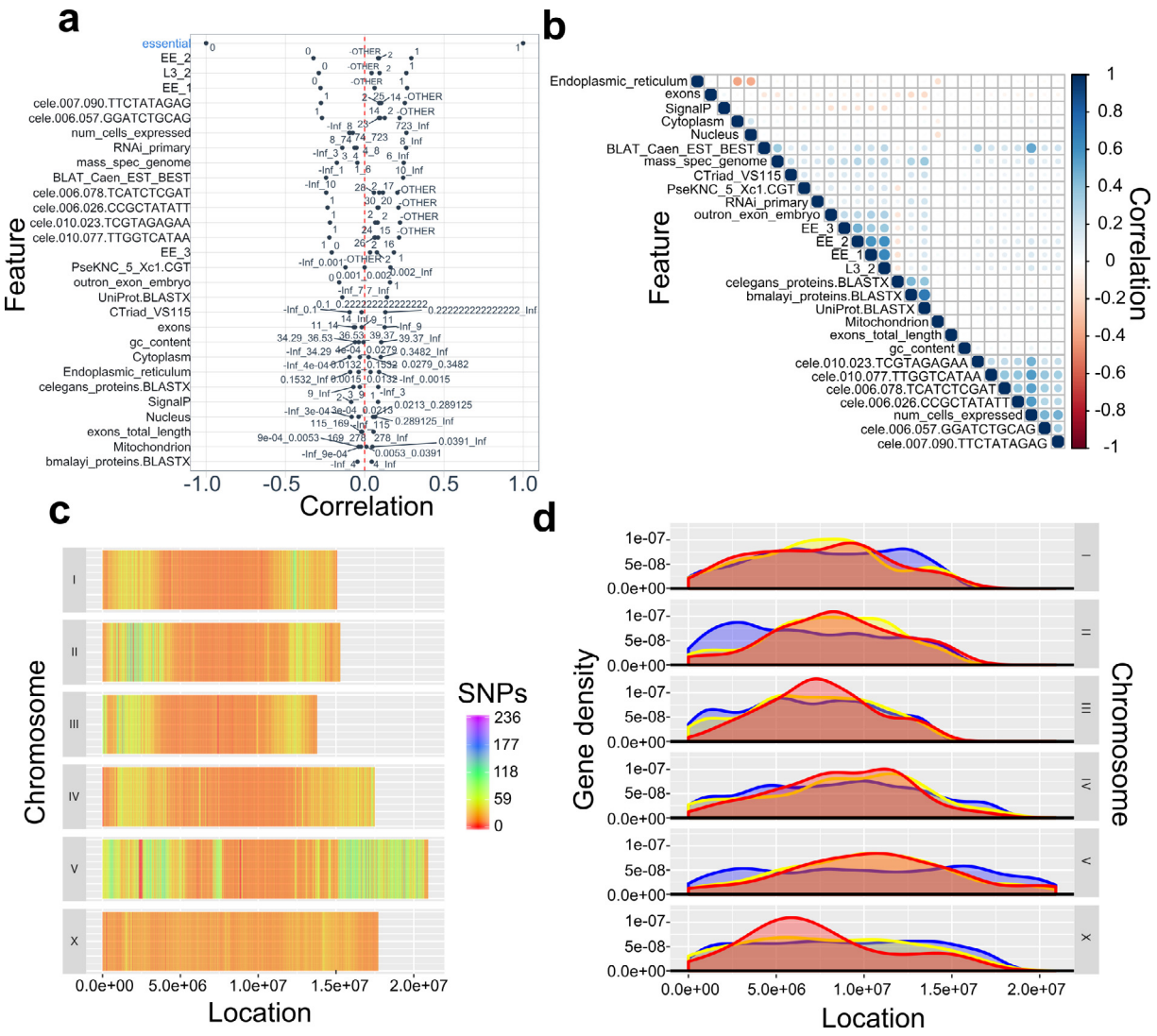


**Fig. 3.** Correlations of features with essentiality; distributions of single nucleotide polymorphisms (SNPs) in and gene essentiality density along *C. elegans* chromosomes. A. The correlations (x-axis) of 28 highly-predictive features (y-axis) with gene essentiality. B, The pairwise correlation among these 28 predictors. C. The distribution of SNPs (1000 bp- windows) along *C. elegans* chromosomes, based on a variant-call file (VCF) derived from whole-genome sequencing of natural *C. elegans* populations [37]. D. Density plots showing the distributions of genes along *C. elegans* chromosomes, stratified by essentiality annotations (red – 'essential'; blue – 'non-essential'; yellow – 'conditionally-essential'). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

ian importance of each of the 28 highly-predictive features for all six ML models showed that 'num_cells_expressed' (71.26), 'BLAT_Caen_EST_BEST' (66.62) and 'RNAi_primary' (54.98) were the strongest predictors using NR_SELECTED data (Table S8). Using the same data set, we assessed variation in the ROC-AUCs and PR-AUCs by bootstrapping (random subsampling; 90% of the data for training; 10% for testing; n = 1000) employing XGB, GBM or RF (Fig. 2c); ROC-AUCs were consistently ≥0.90 for these three ML models, with XGB and GBM each achieving a median ROC-AUC of >0.98. PR-AUCs were consistently >0.7 for these three models, occasionally achieving ~1, with a median of between 0.85 and 0.90.

Sixth, the entire NR_SELECTED data set was used to predict essentiality for each individual gene included here employing each of the six models, and essentiality probabilities calculated (Tables S9 and S10). Using the best performing models (i.e. GBM, RF and XGB), 755 genes were assigned as 'essential' based on high median probabilities (>0.70). Almost 65% of these genes (n = 490) had been annotated previously, based on ESs, as essential, 34% (n = 255) as conditionally-essential and 1% (n = 10) as non-essential. For each of the data sets (i.e. FULL, NR and NR_SELECTED), we then assessed the effects of parameter-tuning on ROC-AUC using a 5-fold cross-validation for each of the six final ML models (Figs. S4–S6). For the parameters tested, we observed that the prediction performance (ROC-AUC) was superior using a regularisation-parameter value of <0.02 for GLM; sigma-parameter of <0.02 for SVM; >1000 boosting iterations and max-tree-depth of ≥3 for both XGB and GBM; >10 hidden-layer units for NN; and randomly selected predictors of 10–50 for RF.

Finally, the validation of the final ML predictions against independent mutant allele data available in the GExplore database [48] (Fig. 4) showed that 7.25% of all *C. elegans* genes studied here have at least one "lethal" phenotype recorded in GExplore. The ratios of genes with a "lethal" phenotype were higher (>20%) for genes with higher ML probabilities (>0.7), and these ratios decreased to 7.25%, as more genes with lower probabilities were included in the search. Conversely, the ratios were consistently low (<5%) for genes with the lowest ML prediction probabilities (<0.1), and increased to 7.25% as more genes with higher ML prediction probabilities were included.

Please cite this article as: T. L. Campos, P. K. Korhonen, P. W. Sternberg et al., Predicting gene essentiality in *Caenorhabditis elegans* by feature engineering and machine-learning, Computational and Structural Biotechnology Journal, https://doi.org/10.1016/j.csbj.2020.05.008

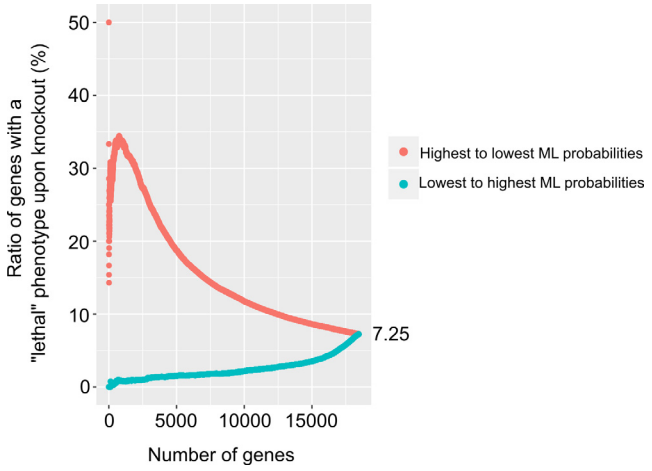


**Fig. 4.** Relationship between ML predictions and the likelihood of a "lethal" phenotype upon knockout. Genes ranked by ML prediction probabilities were searched against a list of genes with at least one "lethal" phenotype reported in the GExplore database. Ratios were calculated cumulatively for genes from the highest to the lowest ML probabilities (red), and from the lowest to the highest (turquoise). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

### *3.4. Essential genes and SNPs are usually located centrally on autosomal chromosomes of C. elegans*

We calculated the numbers of SNPs per 1000 bp and then plotted them on to chromosomes (Fig. 3c). Interestingly, there were considerably more SNPs along chromosome arms than in the centres, except for sex chromosome X where SNPs were evenly distributed. Then, we investigated respective distributions (density plots) of essential, conditionally-essential and non-essential genes on chromosomes (Fig. 3d). We showed that essential genes (usually) have a higher density in the middle of autosomal chromosomes I to V rather than their arms, whereas the density of non-essential genes was higher in the arms of autosomal chromosomes (I–V; Fig. 3d). Interestingly, essential and conditionally-essential genes had similar distributions on all autosomal chromosomes, except chromosome III where the distributions of conditionally-essential and non-essential genes were similar. On sex chromosome X, there appeared to be a preference for essential genes on its left-arm.

The gene density patterns appeared to match SNP densities on chomosomes. Notably, essential genes are preferentially located within regions of low SNP density, as these genes tend to be more conserved than non-essential ones. Moreover, most essential genes are found on autosomal chromosomes (n = 173 on chromosome I; 124 on II; 131 on III; 120 on IV; 103 on V), and only a small number (n = 19) on the sex chromosome. Using Kolmogorov-Smirnov tests, we compared gene densities along chromosomes; there were significant differences between essential and non-essential ($p = 1.243e^{-05}$), and between non-essential and conditionally-essential ($p = 4.79e^{-12}$), but not significant between essential and conditionally-essential genes ($p = 4.651e^{-1}$).

### *3.5. Gene ontology (GO) and transcription enrichments pertaining to essential genes*

Multiple separate GO enrichment analyses (WormBase, WebGestalt and DAVID) revealed information on the biological processes, cellular components and molecular functions for which essential genes play a role. For biological processes, the three most significant terms were 'peptide biosynthetic process' (99 genes), 'cellular macromolecule localisation' (73) and 'embryo development ending in birth or egg hatching' (66) (WormBase; $p \leq 1.3e^{-10}$; Table S11); 'embryo development ending in birth or egg hatching', 'ribonucleoprotein complex biogenesis' and 'translation' (WebGestalt; Fig. S7); 'translation' (88 genes), 'protein transport' (26) and 'intracellular protein transport' (24) (DAVID; $p \leq 2e^{-10}$; Table S12). For cellular components, predominating terms were 'organelle' (412 genes), 'cytoplasm' (325) and 'envelope' (60) (WormBase; $p \leq 1.7e^{-08}$; Table S11); 'cytosolic large ribosomal subunit', 'cytosolic ribosome' and 'large ribosomal subunit' (WebGestalt; Fig. S8); 'intracellular ribonucleoprotein complex' (63 genes), 'ribosome' (62) and 'cytosolic large ribosomal subunit (30)' (DAVID; $p \leq 1.5e^{-07}$; Table S12). For molecular functions, highly-enriched terms were 'structural constituent of ribosome' (62 genes); 'protein heterodimerisation activity' (31) and 'primary active transmembrane transporter activity' (18) (WormBase; $p \leq 2.9e^{-05}$; Table S11); 'ATPase activity, coupled to transmembrane movement of ions', 'structural constituent of ribosome' and 'structural molecule activity' (WebGestalt; Fig. S9); 'nucleotide binding' (104 genes), 'ATP binding' (78) and 'structural constituent of ribosome' (61) (DAVID; $p \leq 1.3e^{-08}$; Table S12).

For transcription (WormEXP database; Table S13), there was an enrichment of targets for small RNAs bound to CSR-1 – an argonaut responsible for chromatin segregation and the protection of germline gene expression [49,50], gene down-regulation in gonad-ablated *C. elegans*, constitutive post-embryonic gene expression as well as matches to orthologues in *D. melanogaster* and *S. cerevisiae* (Table S14). The transcription of most essential genes (92.6% of 500) was enriched in the 'reproductive system' (including germline and gonad tissues) (WormBase; Table S14).

## 4. Discussion

Here, we demonstrate that gene essentiality in *C. elegans* can be reliably predicted using ML models trained using: (i) sets of genes which are well-annotated for essentiality and (ii) features selected and/or engineered from 'omics data. We also reveal highly-predictive features and multiple gene ontology and tissue enrichment analyses to associate with the functions of essential genes in the worm.

The prediction of essentiality from published functional genomic (i.e. RNAi) experiments can be challenging because of ambiguous or contradictory results achieved as a consequence of variations relating to *C. elegans* strains, techniques (soaking vs. injection), experimental conditions used, a lack of repeatability or reproducibility of findings and, in some instances, off-target effects in RNAi [51]. In order to not exclude data for genes that might be essential, we created a scoring system for the inclusion of conditionally-essential genes with ambiguous or variable results from previously published studies. Indeed, the present investigation using well-trained ML models showed that some of these genes provisionally assigned as 'conditionally-essential' (e.g., *dpy-23* [WBGene00001082]; *rpl-7* [WBGene00004418] and *vha-15* [WBGene00020507]) are highly likely to be essential (Table S10). Indeed, "lethal" phenotypes have been recorded for *dpy-23* (WBGene00001082) and *vha-15* (WBGene00020507) in gene knockout data sets in the GExplore database. In addition, 10 genes provisionally assigned as non-essential appear to be essential based on ML predictions. For instance, phenotype information linked to essentiality upon knockout ('L1 arrest' and 'reduced brood size') has been reported for *vav-1* (WBGene00006887) in GExplore. Nonetheless, further work is required to experimentally prove essentiality predictions using classical or modern (e.g., CRISPR/Cas9) gene knock-out methods [52].



Please cite this article as: T. L. Campos, P. K. Korhonen, P. W. Sternberg et al., Predicting gene essentiality in *Caenorhabditis elegans* by feature engineering and machine-learning, Computational and Structural Biotechnology Journal, https://doi.org/10.1016/j.csbj.2020.05.008

Employing large-scale feature engineering, we identified strong essentiality predictors, not previously described, and showed that it is possible to predict gene essentiality reliably without protein–protein interaction network data – which can be error prone [53]. We identified a small number of features (n = 28) that, collectively, contributed to a significant improvement to ML prediction performance. Some of these predictors relate to exon number, GC content and subcellular localisation, identified previously by other workers [23], and novel genomic features such as scRNA-seq or epigenetic markers. Particularly exciting were the four epigenetic markers, EE_1, EE_2, EE_3 and L3_2, identified as being strong predictors of essential genes. For instance, EE_1 and EE_2 corresponded to chromatin states, defined in early embryos by the markers H3K4me3 and H3K4me2, respectively [34]. These markers are known to be involved in cellular differentiation [54], lifespan [55] and/or aging [56], are present in germline cells [57] and are represented throughout the life cycle of *C. elegans* [34]. Interestingly, H3K4me3 has also been associated with gene essentiality in human cells [58]. Previous work [59] has shown that chromatin organisation is highly variable among select metazoans, which would partially explain the distinctiveness in the spectra of essential genes among species [26]. This aspect stimulates studies to explore which features that are predictive of essentiality are common to or distinct among eukaryotic species representing closely and distantly related groups.

The ML models trained using selected features reliably predicted essential genes in *C. elegans* based on a thorough evaluation using multiple independent test sets and threshold-independent metrics (ROC-AUC/PR-AUC). PR-AUC is recognised to be more informative for 'imbalanced' data sets (e.g., markedly more non-essential than essential genes) [60]. In our systematic evaluation, we showed that predictions were quite consistent among the six ML methods and data sets of different sizes, with high prediction performances being achieved using a data set (i.e. NR_SELECTED) that was less prone to sequence bias. Moreover, the ensemble-based ML methods (XGB, GBM and RF) were shown to be most suitable for essentiality prediction, in accordance with other recent findings [26,61]. Here, we calculated probabilities for gene essentiality based on predictions made using high-performing ML methods trained with the NR_SELECTED data set. In addition, a validation conducted using independent functional genomic (mutant allele) data revealed a clear relationship between the ML predictions and the likelihood of a "lethal" phenotype upon knockout. Future work should focus on experimentally confirming our ML-based predictions.

We showed that essential genes in *C. elegans* tend to be located in or near the centre of autosomal chromosomes, and are positively correlated with low SNP densities and epigenetic markers in promoter regions [34,62]. GO results inferred that essential genes in *C. elegans* are involved in protein and nucleotide processing, are transcribed in most cells, are enriched in reproductive tissues and/or are targets for small RNAs bound to the argonaut CSR-1. It has been reported that the CSR-1 and its targets are involved in chromatin segregation [49] and protection of germline cells against piRNA-mediated silencing [50]. This argonaut appears to be responsible for holocentromere organisation [63,64] particularly in nematodes of evolutionary clades V and III [64,65]. Collectively, this information stimulates future investigations of the chromosomal structures and intricate molecular mechanisms linked to gene essentiality, which likely govern the life/survival of nematodes of these clades. Interestingly, selected (non-conserved) essential genes in *C. elegans* are known to be involved in chromatin segregation [66] and exhibit characteristics of house-keeping genes [67], which might suggest an interplay between epigenetic markers and small RNA pathways in the germline [68] linked to a transcription 'memory' profile of gene essentiality that is transmitted to the next generation of cells.

## 5. Conclusion

This study shows that well-trained ML methods can be useful tools to predict essential genes in *C. elegans*. From a biological perspective, our findings show that essential genes tend to be located in or near the centre of autosomal chromosomes; are positively correlated with low SNP densities and epigenetic markers in promoter regions; are involved in protein and nucleotide processing; are transcribed in most cells; are enriched in reproductive tissues or are targets for small RNAs bound to argonaut CSR-1. Based on these results, we speculate that there is an intimate interplay between epigenetic markers and small RNA pathways in the germline, with one or more transcription-based memory profile(s). From an informatic perspective, although the present ML approach seems promising for broader application, it remains to be established whether essentiality can be reliably predicted in distantly related taxa, based on evidence for *C. elegans* (cf. Campos et al., 2019). This aspect requires in-depth evaluation. As a first step, we propose to predict/explore gene essentiality in *D. melanogaster* – for which extensive data and feature sets are available – using the present ML approach, and then to compare findings with those achieved here for *C. elegans*. Such an investigation would establish whether there is a panel of concordant features which are strong predictors of essentiality in both of these model organisms (superphylum Ecdysozoa). If successful, the next step would be to assess the applicability of our approach to a range of metazoan (invertebrate) taxa, for which suitably large and informative genomic, transcriptomic and/or proteomic data sets are available (in the absence of functional genomic and PPI network data sets), so that a panel of "universal" strong predictors of essentiality can be defined for invertebrates.

## 6. Data and code availability

The data used herein, the code developed to perform the systematic ML approaches as well as information regarding software versions and attached libraries are available at: https://bitbucket.org/tuliocampos/essential_elegans. A static version linked to this publication is available at: https://doi.org/10.6084/m9.figshare.11533101.

## CRediT authorship contribution statement

**Tulio L. Campos:** Conceptualization, Methodology, Software, Validation, Data curation, Writing - original draft, Visualization, Investigation, Writing - review & editing. **Pasi K. Korhonen:** Conceptualization, Supervision, Software, Validation, Visualization, Investigation, Writing - review & editing. **Paul W. Sternberg:** Visualization, Investigation, Writing - review & editing. **Robin B. Gasser:** Conceptualization, Supervision, Visualization, Investigation, Writing - review & editing. **Neil D. Young:** Conceptualization, Supervision, Visualization, Investigation, 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 research was funded by grants from the National Health and Medical Research Council (NHMRC) of Australia and the Australian Research Council (ARC) to RBG, PKK and/or NDY. Other sup-

port to RBG was from the Melbourne Water. NDY was supported by a Career Development Fellowship, and PKK by an Early Career Research Fellowship from NHMRC. TLC was a recipient of a Research Training Program Scholarship from the Australian Government and is also supported by the Oswaldo Cruz Foundation (Fiocruz/Brazil). PWS was supported by U.S. National Institutes of Health grant U24-HG002223.

## Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2020.05.008.

## References

[1] Zhan T, Boutros M. Towards a compendium of essential genes - From model organisms to synthetic lethality in cancer cells. Crit Rev in Biochem Mol Biol 2016;51:74–85.
[2] Howe DG, Blake JA, Bradford YM, Bult CJ, Calvi BR, Engel SR, et al. Model organism data evolving in support of translational medicine. Lab Anim (NY) 2018;47:277–89.
[3] Giansanti MG, Fraschini R. Editorial: Model organisms: a precious resource for the understanding of molecular mechanisms underlying human physiology and disease. Front Genet 2019;10:822.
[4] C. elegans Sequencing Consortium. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. Science 1998;282:2012–8.
[5] Clark DV, Rogalski TM, Donati LM, Baillie DL. The unc-22(IV) region of *Caenorhabditis elegans*: genetic analysis of lethal mutations. Genetics 1988;119:345–53.
[6] Kamath RS, Ahringer J. Genome-wide RNAi screening in *Caenorhabditis elegans*. Methods 2003;30:313–21.
[7] Kamath RS, Fraser AG, Dong Y, Poulin G, Durbin R, Gotta M, et al. Systematic functional analysis of the *Caenorhabditis elegans* genome using RNAi. Nature 2003;421:231–7.
[8] Sönnichsen B, Koski LB, Walsh A, Marschall P, Neumann B, Brehm M, et al. Full-genome RNAi profiling of early embryogenesis in *Caenorhabditis elegans*. Nature 2005;434:462–9.
[9] Wang H, Park H, Liu J, Sternberg PW. An efficient genome editing strategy to generate putative null mutants in *Caenorhabditis elegans* using CRISPR/Cas9. G3 (Bethesda) 2018;8:3607–16.
[10] Rogalski TM, Moerman DG, Baillie DL. Essential genes and deficiencies in the unc-22 IV region of *Caenorhabditis elegans*. Genetics 1982;102:725–36.
[11] Meneely PM, Herman RK. Lethals, steriles and deficiencies in a region of the X chromosome of *Caenorhabditis elegans*. Genetics 1979;92:99–115.
[12] Dickinson JD, Goldstein B. CRISPR-Based methods for *Caenorhabditis elegans* genome engineering. Genetics 2016;202:885–901.
[13] Harris TW, Arnaboldi V, Cain S, Chan J, Chen WJ, Cho J, et al. WormBase: a modern model organism information resource. Nucleic Acids Res 2019;8: D762–7.
[14] Zhou X, Xu F, Mao H, Ji J, Yin M, Feng X, et al. Nuclear RNAi contributes to the silencing of off-target genes and repetitive sequences in *Caenorhabditis elegans*. Genetics 2014;197:121–32.
[15] Mohr SE, Perrimon N. RNAi screening: new approaches, understandings, and organisms. Wiley Interdiscip Rev RNA 2012;3:145–58.
[16] Hagen J, Lee EF, Fairlie WD, Kalinna BH. Functional genomics approaches in parasitic helminths. Parasite Immunol 2012;34:163–82.
[17] Castelletto ML, Gang SS, Hallem EA. Recent advances in functional genomics for parasitic nematodes of mammals. J Exp Biol 2020;7:223 (Pt Suppl 1).
[18] Zhong W, Sternberg WP. Genome-wide prediction of *C. elegans* genetic interactions. Science 2006;311:1481–4.
[19] Lee I, Lehner B, Crombie C, Wong W, Fraser AG, Marcotte EM. A single gene network accurately predicts phenotypic effects of gene perturbation in *Caenorhabditis elegans*. Nat Genet 2008;40:181–8.
[20] Qin Z, Johnsen R, Yu S, Chu JS, Baillie DL, Chen N. Genomic identification and functional characterization of essential genes in *Caenorhabditis elegans*. G3 (Bethesda) 2018;8:981–97.
[21] Yu S, Zheng C, Zhou F, Baillie DL, Rose AM, Deng Z, et al. Genomic identification and functional analysis of essential genes in *Caenorhabditis elegans*. BMC Genomics 2018;19:871.
[22] Doyle MA, Gasser RB, Woodcroft BJ, Hall RS, Ralph SA. Drug target prediction and prioritization: using orthology to predict essentiality in parasite genomes. BMC Genomics 2010;11:222.
[23] Dong C, Jin YT, Hua HL, Wen QF, Luo S, Zheng WX, et al. Comprehensive review of the identification of essential genes using computational methods: focusing on feature implementation and assessment. Brief Bioinform 2018;2018. pii: bby116.
[24] Li M, Wang JX, Wang H, Pan Y. Identification of essential proteins from weighted protein-protein interaction networks. J Bioinf Comput Biol 2013;11:1341002.
[25] Zhang X, Acencio ML, Lemke N. Predicting essential genes and proteins based on machine learning and network topological features: a comprehensive review. Front Physiol 2016;7.
[26] Campos TL, Korhonen PK, Gasser RB, Young ND. An evaluation of machine learning approaches for the prediction of essential genes in eukaryotes using protein sequence-derived features. Comput Struct Biotechnol J 2019;17:785–96.
[27] Howe KL, Bolt BJ, Cain S, Chan J, Chen WJ, Davis P, et al. WormBase 2016: expanding to enable helminth genomic research. Nucleic Acids Res 2016;44: D774–80.
[28] Birney E, Andrews TD, Bevan P, Caccamo M, Chen Y, Clarke L, et al. An overview of Ensembl. Genome Res 2004;2004(14):925–8.
[29] Spencer WC, Zeller G, Watson JD, Henz SR, Watkins KL, McWhirter RD, et al. A spatial and temporal map of *C. elegans* gene expression. Genome Res 2011;21:325–41.
[30] Saito TL, Hashimoto S, Gu SG, Morton JJ, Stadler M, Blumenthal T, et al. The transcription start site landscape of *C. elegans*. Genome Res 2013;23:1348–61.
[31] Cao J, Packer JS, Ramani V, Cusanovich DA, Huynh C, Daza R, et al. Comprehensive single-cell transcriptional profiling of a multicellular organism. Science 2017;357:661–7.
[32] Yang W, Dierking K, Schulenburg H. WormExp: a web-based application for a *Caenorhabditis elegans*-specific gene expression enrichment analysis. Bioinformatics 2016;32:943–5.
[33] Kiniry SJ, O'Connor PBF, Michel AM, Baranov PV. Trips-Viz: a transcriptome browser for exploring Ribo-Seq data. Nucleic Acids Res 2019;47:D847–52.
[34] Evans KJ, Huang N, Stempor P, Chesney MA, Down TA, Ahringer J. Stable *Caenorhabditis elegans* chromatin domains separate broadly expressed and developmentally regulated genes. Proc Natl Acad Sci USA 2016;113:E7020–9.
[35] Ikegami K, Egelhofer TA, Strome S, Lieb JD. *Caenorhabditis elegans* chromosome arms are anchored to the nuclear membrane via discontinuous association with LEM-2. Genome Biol 2010;11:R120.
[36] Daugherty AC, Yeo RW, Buenrostro JD, Greenleaf WJ, Kundaje A, Brunet A. Chromatin accessibility dynamics reveal novel functional enhancers in *C. elegans*. Genome Res 2017;27:2096–107.
[37] Cook DE, Zdraljevic S, Roberts JP, Andersen EC. CeNDR, the *Caenorhabditis elegans* natural diversity resource. Nucleic Acids Res 2017;45:D650–7.
[38] Krogh A, Larsson B, von Heijne G, Sonnhammer EL. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol 2001;305:567–80.
[39] Petersen TN, Brunak S, von Heijne G, Nielsen H. 2011. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat Methods 2011;8:785–6.
[40] Horton P, Park K-J, Obayashi T, Fujita N, Harada H, Adams-Collier CJ, et al. WoLF PSORT: protein localization predictor. Nucleic Acids Res 2007;35: W585–7.
[41] Almagro Armenteros JJ, Sonderby CK, Sonderby SK, Nielsen H, Winther O. DeepLoc: prediction of protein subcellular localization using deep learning. Bioinformatics 2017;33:3387–95.
[42] Linding R, Jensen LJ, Diella F, Bork P, Gibson TJ, Russell RB. Protein disorder prediction: implications for structural proteomics. Structure 2003;11:1453–9.
[43] Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. Fly (Austin) 2012;6:80–92.
[44] Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinformatics 2010;26:2460–1.
[45] Huang DW, Sherman BT, Tan Q, Collins JR, Alvord WG, Roayaei J, et al. The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists. Genome Biol 2007;8:R183.
[46] Wang J, Vasaikar S, Shi Z, Greer M, Zhang B. WebGestalt 2017: a more comprehensive, powerful, flexible and interactive gene set enrichment analysis toolkit. Nucleic Acids Res 2017;45:W130–7.
[47] Angeles-Albores D, Lee RYN, Chan J, Sternberg PW. Tissue enrichment analysis for *C. elegans* genomics. BMC Bioinf 2016;17:366.
[48] Claycomb JM, Batista PJ, Pang KM, Gu W, Vasale JJ, van Wolfswinkel JC, et al. The Argonaute CSR-1 and its 22G-RNA cofactors are required for holocentric chromosome segregation. Cell 2009;139:123–34.
[49] Hutter H, Suh J. GExplore 1.4: an expanded web interface for queries on *Caenorhabditis elegans* protein and gene function. Worm 2016;5:e1234659.
[50] Wedeles CJ, Wu MZ, Claycomb JM. Protection of germline gene expression by the *C. elegans* Argonaute CSR-1. Dev Cell 2013;27:664–71.
[51] Fellmann C, Lowe SW. Stable RNA interference rules for silencing. Nat Cell Biol 2014;16:10–8.
[52] Evers B, Jastrzebski K, Heijmans JP, Grernrum W, Beijersbergen RL, Bernards R. CRISPR knockout screening outperforms shRNA and CRISPRi in identifying essential genes. Nat Biotechnol 2016;34:631–3.
[53] Kuchaiev O, Rasajski M, Higham DJ, Przulj N. Geometric de-noising of protein-protein interaction networks. PLoS Comput Biol 2009;5:e1000454.
[54] Benayoun BA, Pollina EA, Ucar D, Mahmoudi S, Karra K, Wong ED, et al. H3K4me3 breadth is linked to cell identity and transcriptional consistency. Cell 2014;158:673–88.
[55] Han S, Schroeder EA, Silva-Garcia CG, Hebestreit K, Mair WB, Brunet A. Mono-unsaturated fatty acids link H3K4me3 modifiers to *C. elegans* lifespan. Nature 2017;544:185–90.

[56] Pu M, Wang M, Wang W, Velayudhan SS, Lee SS. Unique patterns of trimethylation of histone H3 lysine 4 are prone to changes during aging in *Caenorhabditis elegans* somatic cells. PLoS Genet 2018;14:e1007466.
[57] Kelly WG. Transgenerational epigenetics in the germline cycle of *Caenorhabditis elegans*. Epigenetics Chromatin 2014;7:6.
[58] Chen H, Zhang Z, Jiang S, Li R, Li W, Zhao C, et al. New insights on human essential genes based on integrated analysis and the construction of the HEGIAP web-based platform. Brief Bioinform 2019;pii: bbz072.
[59] Ho JW, Jung YL, Liu T, Alver BH, Lee S, Ikegami K, et al. Comparative analysis of metazoan chromatin organization. Nature 2014;512:449–52.
[60] Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS ONE 2015;10.
[61] Zhong J, Sun Y, Peng W, Xie M, Yang J, Tang X. XGBFEMF: An XGBoost-based framework for essential protein prediction. IEEE Trans Nanobiosci 2018;17:243–50.
[62] Garrigues JM, Sidoli S, Garcia BA, Strome S. Defining heterochromatin in *C. elegans* through genome-wide analysis of the heterochromatin protein 1 homolog HPL-2. Genome Res 2015;25:76–88.
[63] Subirana JA, Messeguer X. A satellite explosion in the genome of holocentric nematodes. PLoS ONE 2013;8:e62221.
[64] Wedeles CJ, Wu MZ, Claycomb JM. A multitasking Argonaute: exploring the many facets of *C. elegans* CSR-1. Chromosome Res 2013;21:573–86.
[65] Tu S, Wu MZ, Wang J, Cutter AD, Weng Z, Claycomb JM. Comparative functional characterization of the CSR-1 22G-RNA pathway in *Caenorhabditis* nematodes. Nucleic Acids Res 2015;43:208–24.
[66] Verster AJ, Styles EB, Mateo A, Derry WB, Andrews BJ, Fraser AG. Taxonomically restricted genes with essential functions frequently play roles in chromosome segregation in *Caenorhabditis elegans* and *Saccharomyces cerevisiae*. G3 (Bethesda) 2017;7:3337–47.
[67] Eisenberg E, Levanon EY. Human housekeeping genes, revisited. Trends Genet 2013;29:569–74.
[68] Gushchanskaia ES, Esse R, Ma QC, Lau NC, Grishok A. Interplay between small RNA pathways shapes chromatin landscapes in *C. elegans*. Nucleic Acids Res 2019;47:5603–16.



Please cite this article as: T. L. Campos, P. K. Korhonen, P. W. Sternberg et al., Predicting gene essentiality in *Caenorhabditis elegans* by feature engineering and machine-learning, Computational and Structural Biotechnology Journal, https://doi.org/10.1016/j.csbj.2020.05.008