University Library



A gateway to Melbourne's research publications

Minerva Access is the Institutional Repository of The University of Melbourne

Author/s:
Campos, TL;Korhonen, PK;Young, ND;Chang, BCH;Gasser, RB

Title:
Inference of essential genes in Brugia malayi and Onchocerca volvulus by machine learning and the implications for discovering new interventions

Date:
2024-12-01

Citation:
Campos, T. L., Korhonen, P. K., Young, N. D., Chang, B. C. H. & Gasser, R. B. (2024). Inference of essential genes in Brugia malayi and Onchocerca volvulus by machine learning and the implications for discovering new interventions. Computational and Structural Biotechnology Journal, 23, pp.3081-3089. https://doi.org/10.1016/j.csbj.2024.07.025.

Persistent Link:
https://hdl.handle.net/11343/358337

License:
CC BY-NC-ND

Computational and Structural Biotechnology Journal 23 (2024) 3081–3089

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

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

Inference of essential genes in *Brugia malayi* and *Onchocerca volvulus* by machine learning and the implications for discovering new interventions

Túlio L. Campos [a,b,*], Pasi K. Korhonen [a], Neil D. Young [a], Bill C.H. Chang [a], Robin B. Gasser [a,**]

[a] Department of Biosciences, Melbourne Veterinary School, Faculty of Science, The University of Melbourne, Parkville, Victoria 3010, Australia
[b] Núcleo de Bioinformática, Instituto Aggeu Magalhães, Fiocruz., Av. Professor Moraes Rego, s/n, Cidade Universitária, Recife, PE CEP 50740–465, Brazil

ARTICLE INFO

Keywords:
Essential genes
Machine learning
Brugia malayi
Onchocerca volvulus
Filarioid
Nematodes

ABSTRACT

Detailed explorations of the model organisms *Caenorhabditis elegans* (elegant worm) and *Drosophila melanogaster* (vinegar fly) have substantially improved our knowledge and understanding of biological processes and pathways in metazoan organisms. Extensive functional genomic and multi-omic data sets have enabled the discovery and characterisation of 'essential' genes that are critical for the survival of these organisms. Recently, we showed that a machine learning (ML)-based pipeline could be utilised to predict essential genes in both *C. elegans* and *D. melanogaster* using features from DNA, RNA, protein and/or cellular data or associated information. As these distantly-related species are within the Ecdysozoa, we hypothesised that this approach could be suited for non-model organisms within the same group (phylum) of protostome animals. In the present investigation, we cross-predicted essential genes within the phylum Nematoda – between *C. elegans* and the parasitic filarial nematodes *Brugia malayi* and *Onchocerca volvulus*, and then ranked and prioritised these genes. Highly ranked genes were linked to key biological pathways or processes, such as ribosome biogenesis, translation and RNA processing, and were expressed at relatively high levels in the germline, gonad, hypodermis and/or nerves. The present *in silico* workflow is hoped to expedite the identification of drug targets in parasitic organisms for subsequent experimental validation in the laboratory.

1. Introduction

Parasitic roundworms (nematodes) can cause chronic, debilitating diseases in humans that are challenging to prevent, treat and control, particularly those transmitted by arthropod vectors [1–4]. For instance, *Onchocerca volvulus* is a filarioid nematode that is transmitted by a blackfly (*Simulium* spp.) and causes river blindness (onchocerciasis) in humans [5,6]. In addition, *Brugia malayi*, *Brugia timori, Mansonella* spp. and *Wuchereria bancrofti* are related filarioids that are transmitted by mosquitoes (*Aedes*, *Anopheles* and *Culex*) and cause elephantiasis (lymphatic filariasis) [7–9]. Collectively, this latter form of filariasis affects ~ 120 million people worldwide and is amongst the most neglected tropical diseases as recognised by the World Health Organization (WHO), with elimination from endemic countries expected in the next few years [10–12].

Although mass drug administration (MDA) programs [12–14] have been successful at reducing the prevalence and intensity of filariasis [15] (https://apps.who.int/neglected_diseases/ntddata/lf/lf.html), there is a concern about drug resistance [13,16]. This concern emphasizes the need to search for new drugs or vaccines, built on a sound understanding the molecular biology and biochemistry of these nematodes and/or their relationship with their hosts [17,18]. Thus, deep insights into the genomes, transcriptomes and proteomes of filarial species should allow an enhanced understanding of the molecular processes, mechanisms and/or pathways that govern essential biological as well as infection and disease processes in the host animal. Ultimately, such knowledge should also assist in identifying possible mechanisms of drug resistance and provide avenues for the discovery and development of new interventions [19,20]. Detailed and accurate analyses of nucleic acid and protein sequence data sets, usually by comparison with reference organisms, will be critical in providing biologically meaningful information. In addition, advanced bioinformatic workflow systems are assisting scientists in their analyses of such data sets, enabling the discovery of new intervention targets.

* Corresponding author at: Department of Biosciences, Melbourne Veterinary School, Faculty of Science, The University of Melbourne, Parkville, VIC 3010, Australia.
** Corresponding author.
E-mail addresses: tulio.campos@unimelb.edu.au (T.L. Campos), robinbg@unimelb.edu.au (R.B. Gasser).

https://doi.org/10.1016/j.csbj.2024.07.025
Received 20 May 2024; Received in revised form 30 July 2024; Accepted 31 July 2024
Available online 2 August 2024
2001-0370/© 2024 Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

The discovery of novel anthelmintic targets using conventional approaches is challenging, time-consuming and costly [20–22]. To circumvent this, we have been exploring, evaluating and promoting the use of *in silico* methods for the prediction and prioritisation of essential genes for subsequent validation as drug target candidates. Our exploratory studies [23,24] focused on assessing and employing machine learning (ML)-based approaches for the prediction of such genes in the most intensively studied multicellular model organisms – *Caenorhabditis elegans* (free-living nematode) and *Drosophila melanogaster* (vinegar fly). We have taken this focus because these organisms can be maintained and studied in culture in the laboratory, and, importantly, because chromosome-continuous genomes and extensive functional genomic, transcriptomic, proteomic, biochemical, physiological, biological, morphological, developmental and reproductive data sets and information are publicly available via well-curated databases including WormBase and FlyBase [25–29]. This wealth of resources has enabled deep and meaningful investigations of gene essentiality for these two ecdysozoan species. Our ML-based studies [30,31] have shown that informative features can be extracted/engineered from such data sets, allowing the confident (statistically valid) prediction and prioritisation of known essential genes both within and between *C. elegans* and *D. melanogaster*.

As the complete life cycles of species of filarioid nematodes are long and cannot be readily maintained in vitro, and laboratory culture conditions vary from those in nature (i.e. in the arthropod vector and in the definitive host), establishing high-throughput functional genomic assays for different developmental stages and sexes of these parasitic nematodes has been a major obstacle to evaluating the essentiality of genes, and to inferring or prioritising intervention target candidates. Now that we have demonstrated the feasibility of an ML-based bioinformatic approach for the reliable prediction and prioritisation of essential genes in each *C. elegans* and *D. melanogaster*, and across these two species [30,

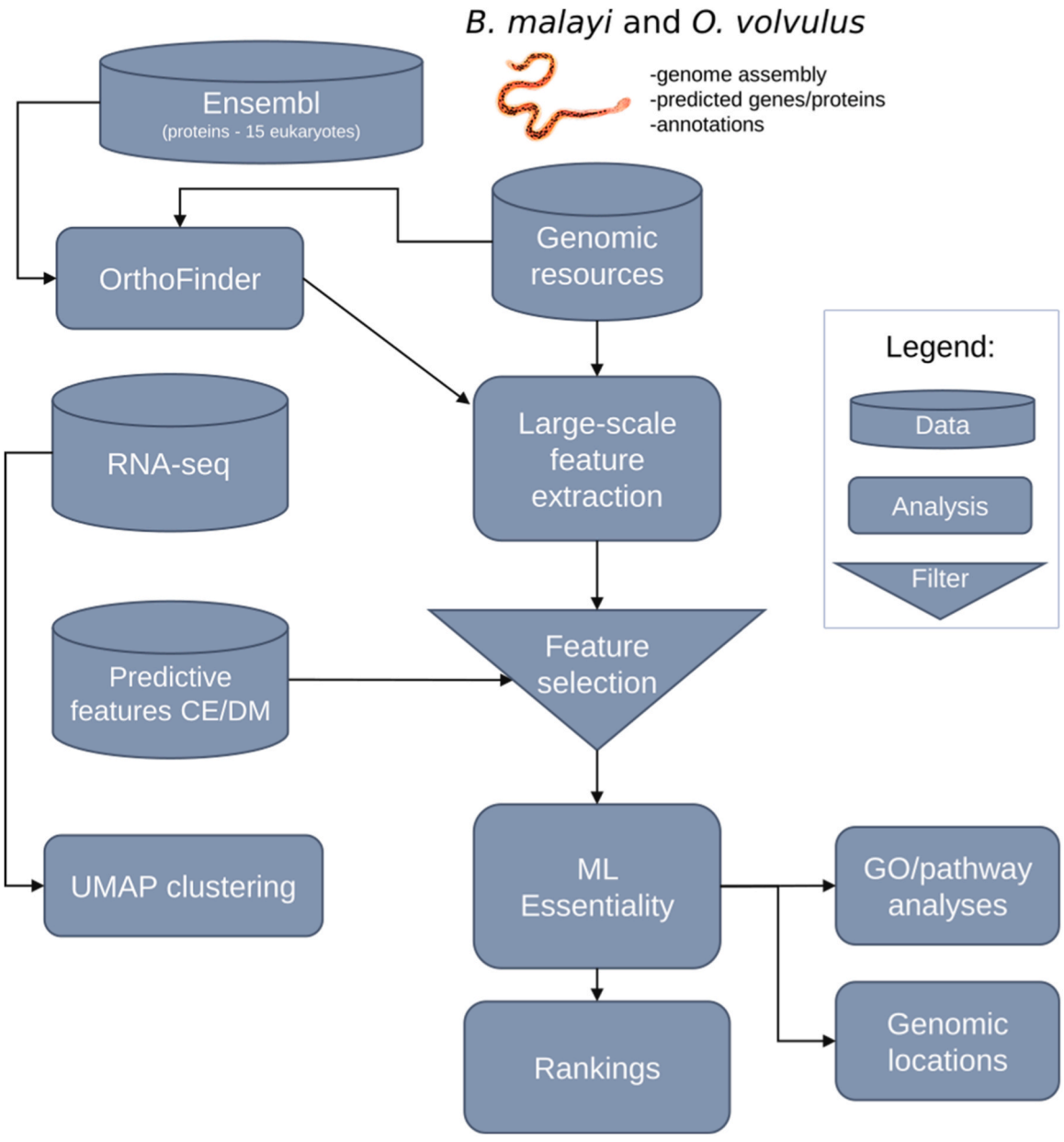


Fig. 1. The workflow used for the prediction of essential genes in *Brugia malayi* and *Onchocerca volvulus* using machine learning (ML) and complementary analyses. A range of features (see Section 2. Materials and methods) were extracted from *B. malayi* and *O. volvulus* genes, and selected features were used to train ML models and predict essential genes. The relationship between essentiality and transcription was investigated by clustering analyses; other complementary analyses included GO/pathway enrichments and the mapping of gene essentiality probability to genomic locations.

31], we propose that this approach will be applicable to cross-species prediction between related nematodes or arthropods (ecdysozoans), provided that suitable, high-quality genomic, transcriptomic and/or proteomic data sets are available for analyses and inferences. Extending our recent work (reviewed by [31]), we now apply our ML workflow to predict and prioritise essential genes of *B. malayi* and *O. volvulus*, employing *C. elegans* and *D. melanogaster* data for algorithm-training purposes and comparative analyses, and we explore the relationship between gene essentiality and transcription in both parasitic nematodes. These filarioid species were selected because of their global importance worldwide and because well-assembled, annotated genomes as well as extensive transcriptomic data sets are available for these pathogens), providing a solid foundation for detailed bioinformatic analyses.

2. Materials and methods

2.1. Data sets and pre-processing

First, we obtained published genomic and transcriptomic data sets for *B. malayi* and *O. volvulus* (Table S1; [32,33]) from WormBase ParaSite (release WBPS19; cf. [25]) and employed a workflow (Fig. 1) for the prediction of essential genes. The reference genome assemblies for *B. malayi* (BioProject PRJNA10729; Bmal-4.0; 88.2 Mb; four autosomes [I to IV] and one sex chromosome [X]) and *O. volvulus* (representing four autosomes; sex chromosome unresolved) were independently assessed for size, ploidy and order of assembled contigs. Six RNA-seq data sets representing 152 samples from *B. malayi* (whole worms) and one data set comprising 10 samples representing different developmental stages and both sexes of *O. volvulus* were obtained from WormBase ParaSite [25]. These RNA-seq data sets had been pre-processed previously [25] and made publicly available on the web. Briefly, reads that mapped to individual genes were enumerated for each sample and normalised using transcripts per million (TPM). This information was loaded into data frames in R (https://www.r-project.org) for subsequent use and analyses. Genes without evidence of transcription (mapped read counts = 0) in a sample were removed.

2.2. Feature extraction and selection

From the 10,842 protein-coding genes predicted/annotated for the genome of *B. malayi* (see [32]) and 12,109 genes of *O. volvulus* (see [33]), we extracted 9569 features linked to: DNA sequences, protein sequences and subcellular localisation (inferred using DeepLoc 1.0; [34]) using established methods [23,24]. Then, we selected 26 features of essential genes in both *C. elegans* and *D. melanogaster* that were previously identified as predictors [30]. One additional feature (OrthoFinder_species), i.e. protein sequence conservation between species, was added. In brief, predicted proteomes (FASTA files) representing 15 eukaryotic species from divergent branches of the Tree-of-Life [35] were obtained from the Ensembl genome database (https://asia.ensembl.org/index.html; [36]); orthologous groups were identified in these 15 species as well as in *B. malayi* and *O. volvulus* using the tool OrthoFinder [37] employing default parameters. Then, we identified the number of species represented within individual orthologous protein groups for *C. elegans*, *D. melanogaster, B. malayi* and *O. volvulus*. The 27 features selected for both species (Table 1) were used in subsequent analyses.

Table 1
Features (n = 27) used to predict essential genes in *Brugia malayi* and *Onchocerca volvulus*; these features are predictive for essential genes in both *Caenorhabditis elegans* and *Drosophila melanogaster* [30].

Feature	Description	Source
OrthoFinder_species	Orthologs in other species	OrthoFinder analysis
exons	Number of exons	BioMart (WomBase ParaSite)
exons_total_length	Total length of exons	BioMart (WomBase ParaSite)
Cytoplasm	Subcellular localisation	DeepLoc analysis
Mitochondrion	Subcellular localisation	DeepLoc analysis
Nucleus	Subcellular localisation	DeepLoc analysis
AAC_S	Protein sequence feature	Extracted using protR*
APAAC_Pc2. Hydrophobicity.2	Protein sequence feature	Extracted using protR*
CTDC_secondarystruct. Group1	Protein sequence feature	Extracted using protR*
CTDD_prop4. G2.residue0	Protein sequence feature	Extracted using protR*
CTDD_prop4. G2.residue25	Protein sequence feature	Extracted using protR*
CTriad_VS153	Protein sequence feature	Extracted using protR*
CTriad_VS431	Protein sequence feature	Extracted using protR*
CTriad_VS613	Protein sequence feature	Extracted using protR*
DC_HA	Protein sequence feature	Extracted using protR*
DC_MP	Protein sequence feature	Extracted using protR*
DC_MS	Protein sequence feature	Extracted using protR*
DC_VF	Protein sequence feature	Extracted using protR*
DC_LA	Protein sequence feature	Extracted using protR*
Geary_CHOC760101.lag7	Protein sequence feature	Extracted using protR*
Moran_CHAM820102.lag7	Protein sequence feature	Extracted using protR*
GC	DNA sequence feature	BioMart (WormBase ParaSite)
kmer_3_GCT	DNA sequence feature	Extracted using rDNAse*
PseKNC_3_Xc1. CCC	DNA sequence feature	Extracted using rDNAse*
PseKNC_5_Xc1. CGT	DNA sequence feature	Extracted using rDNAse*
PseKNC_5_Xc1. GCT	DNA sequence feature	Extracted using rDNAse*
TACC_Nucleosome.lag2	DNA Sequence feature	Extracted using rDNAse*

* Refer to documentation on the R packages protR (https://cran.r-project.org/web/packages/protr/vignettes/protr.html) and rDNAse (https://github.com/wind22zhu/rDNAse) for further information regarding sequence features.

2.3. Predicting and ranking gene essentiality by ML

We assessed the individual and collective powers of the 27 features selected to predict essential genes within *C. elegans* and *D. melanogaster* employing six distinct machine-learning (ML) models (Gradient Boosting Machine - GBM, Generalised Linear Model - GLM, Neural Network - NN, Random Forest - RF, Support Vector Machine – SVM, and Extreme Gradient Boosting Machine - XGB) [23,24]. For this evaluation, features with low variance were removed, the remaining features were normalised and a feature selection approach was employed (cf. [23,24]). Then, subsets with 10 to 90 % (using 10 % increments) of the essential/non-essential genes and their features were used to train the ML models; remaining data were used for testing and evaluation using ROC-AUC and PR-AUC metrics. The best-performing ML models established for *C. elegans* and *D. melanogaster*, based on ROC-AUC and PR-AUC, were employed to predict essential genes in *B. malayi* and *O. volvulus*. We ranked all genes of both parasitic nematodes based on their probabilities (descending) of being essential (defined using the two best-performing ML models). This approach was used to obtain the two lists of high-priority essential genes.

2.4. Gene clustering based on transcription

For each *B. malayi* and *O. volvulus*, genes were assigned to eight clusters, according to their transcription profiles in different

developmental stages and/or conditions (samples). For this analysis, we used unsupervised clustering, employing uniform manifold approximation and projection (UMAP; "umap" package for R), with random initialisation - all the other settings were kept default. Following assignment, gene clusters were displayed using "ggplot2″ for R. We also assessed the association of essential genes with specific clusters using Fisher's exact tests in R.

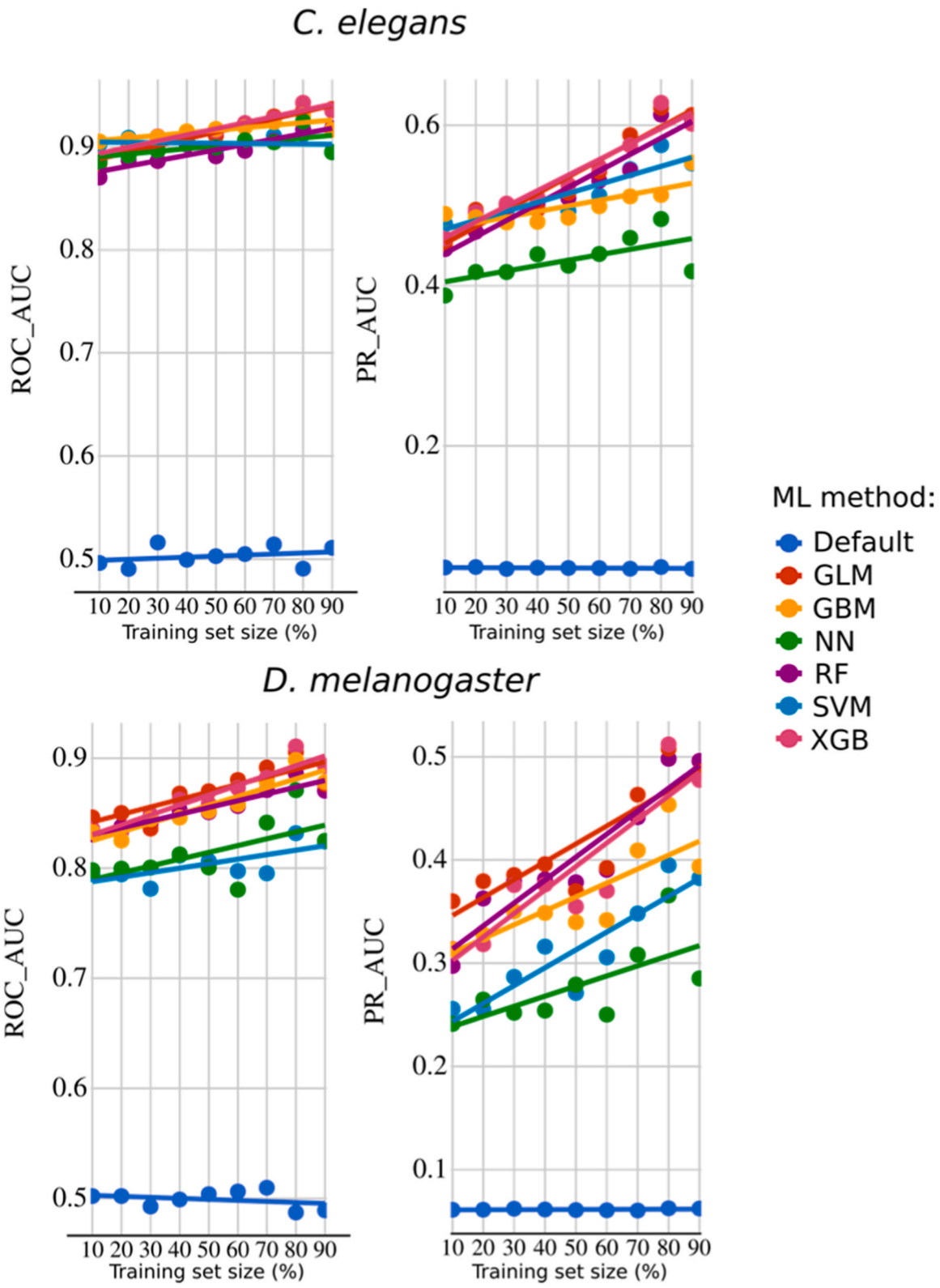


Fig. 2. Machine learning (ML) performance metrics (ROC-AUC and PR-AUC) for the prediction of essential genes in *Caenorhabditis elegans* (top) or *D. melanogaster* (bottom) using features available for *Brugia malayi*. ML methods used: Gradient Boosting Machines (GBM), Generalised Linear models (GLM), Neural Networks (NN), Random Forest (RF), Support-Vector Machines (SVM) and Extreme Gradient Boosting Machines (XGB).

2.5. Inferring genome, tissue and/or cell locations as well as functional annotation of essential genes of B. malayi or O. volvulus

For each gene of *B. malayi* and *O. volvulus* inferred from our prediction/annotation approach, we associated the genomic location in the General Feature Format (GFF) annotation file with the corresponding ML-based probability of being essential. The probability of each gene being essential (defined by ML) was mapped to the largest scaffolds (representing the chromosomes) of each filarioid species using "chromoMap" for R. We assessed and compared the presence of essential versus non-essential genes along chromosomes using density plots as well as Kolmogorov-Smirnov tests in R. We also inferred cell types and tissues in which the "essential gene orthologs" of *B. malayi* or *O. volvulus* predicted were abundantly transcribed in *C. elegans*. For these analyses, we used existing singe-cell RNA sequence (scRNA-seq) data available (Cao_et_al_2017_vignette.RData file; cf. [38]). The genes predicted as essential (probability: >0.5) were also subjected to gene ontology (GO) and pathway enrichment analyses based on their respective orthologs in *C. elegans* using g:Profiler [39] and the Reactome Pathway Database [40].

3. Results

3.1. Identification of predictors of essential genes in B. malayi and O. volvulus

In total, we defined 27 features for protein-coding genes in the nuclear genome of *B. malayi* and *O. volvulus*. These features had been identified previously as strong predictors of gene essentiality within each *C. elegans* and *D. melanogaster* (see [23,24]) and between these two species (see [30]); we also selected "evolutionary conservation" among 15 divergent eukaryotic species as well as *B. malayi* and *O. volvulus*.

Prior to the prediction of essential genes, we evaluated the predictive power of this set of 27 features for *C. elegans* and for *D. melanogaster* using ML approaches and a subsampling strategy for training, testing and evaluation (ROC-AUC and PR-AUC metrics). For *C. elegans* (Fig. 2), the ROC-AUC value was > 0.87 for all six ML models assessed (i.e. GBM, GLM, NN, RF, SVM and XGB), achieving ~0.93 for GBM and XGB. The PR-AUC value was usually > 0.4, achieving close to 0.6 for the best performers (GBM and XGB) using 90 % of the data to train the models. For *D. melanogaster* (Fig. 2), the ROC-AUC value was > 0.8, achieving ~0.9 for the best performers (GBM, GLM, RF and XGB). The PR-AUC value was variable, depending on the ML model used, and ranged from ~0.3 to ~0.5 for the best performers (GBM, RF and XGB), using 90 % of the data in the training set. Of the 26 features employed, the most important predictors of essential genes in both *C. elegans* and *D. melanogaster* were: OrthoFinder_species and exon numbers (exons), followed by two subcellular localisation predictions (nucleus and cytoplasm), and GC content. The relative contributions of individual features to the prediction of essential genes for each species, using each of the six machine learning (ML) models, are given in Table S2.

Following the prediction of essential genes in *B. malayi* using the best-performing models (GBM and XGB, trained with *C. elegans* and *D. melanogaster* features), 246 of 10,842 annotated protein-coding genes had a probability of > 0.5 of being essential, whereas the 5991 genes with a probability of < 0.05 were classified as non-essential (Tables S3 and S4). For *O. volvulus* (Tables S3 and S4), 110 of 12,109 protein-coding genes had a probability of > 0.5 of being essential, whereas the 7237 genes with a probability of < 0.05 were classified as non-essential. In total, 217 (> 88.2 %) of the top 246 essential genes predicted for *B. malayi* have orthologs in *C. elegans*, and 234 (95.1 %) were inferred to be single-copy genes based on an analysis of orthologs using g:Profiler (Table S5). For *O. volvulus*, 96 genes (~87.3 %) in the top 110 genes inferred to be essential have orthologs in *C. elegans*, and 97 (~88.2 %) were identified as single-copy (Table S5).

3.2. Association between essential genes and transcription profiles

To investigate the relationship between gene essentiality and transcription in each *B. malayi* and *O. volvulus*, we identified UMAP gene clusters based on their transcription (RNA-seq) in distinct samples representing multiple developmental stages and sexes (Fig. 3). After filtering out genes that were not transcribed in all 152 samples, 5920 of 10,842 genes (~54.6 %) remained for *B. malayi*. Of these genes, 221 of 246 (~89.8 %) of the high-priority essential genes were present, and most of them clustered together (Fisher's exact test $p < 2.2 \times 10^{-16}$ for the cluster containing most essential genes; Fig. 3). On the other hand, following the filtering step, only 2857 of 5991 (~47.7 %) of the most likely non-essential genes remained; most of the latter genes clustered to the exclusion of the essential genes (Fig. 3). For *O. volvulus*, 8159 genes remained following the filtering step (10 samples), such that 97 of 110 (~88.1 %) of the genes predicted as essential, and 4519 of the 7237 (~62.4 %) genes inferred to be non-essential were retained. Predicted essential genes of *O. volvulus* also clustered together in the UMAP plot (Fisher's exact test $p < 3.4 \times 10^{-10}$ for the cluster containing most essential genes; Fig. 3).

3.3. Essential genes of B. malayi are inferred to be involved predominantly in ribosome biogenesis, translation, RNA binding/processing and signalling.

GO enrichment analysis (Table S6) for the prioritised list of the top essential genes for each *B. malayi* and *O. volvulus* inferred the same molecular functions (MFs), biological processes (BPs) and cellular components (CCs). MFs ($p < 10^{-5}$), including structural molecule activity, structural constituent of ribosome or chromatin, protein heterodimerization activity and nucleic acid binding; BPs ($p < 10^{-4}$) included peptide biosynthetic/metabolic, amide biosynthetic/metabolic and/or cellular nitrogen compound metabolic process; CCs ($p < 10^{-15}$) included ribosome, nucleosome, intracellular non-membrane-bound organelle, intracellular anatomical structure, ribonucleoprotein complex and/or ribosomal subunit. Pathway enrichment analysis via Reactome (Table S7) revealed that *C. elegans* orthologs of genes inferred to be essential in both *B. malayi* and *O. volvulus* were significantly ($p < 10^{-5}$) linked to functions including: (i) the assembly of the ribosome (e.g., GTP hydrolysis, the joining of the 60 S ribosomal subunit and the formation of free 40 S subunits); (ii) translation initiation (e.g., eukaryotic and cap-dependent); and (iii) signalling and regulatory roles (e.g., L13a-mediated translation silencing of ceruloplasmin expression, SRP-dependent co-translational protein targeting to membrane, and nonsense-mediated decay).

3.3. Linking essential genes to genome locations, and their transcription to cell type or tissue

First, we plotted the ML-based gene essentiality probabilities along individual *B. malayi* and *O. volvulus* chromosomes (Fig. 4). For *B. malayi*, most of the high-priority essential genes predicted were linked to the sex chromosome X (n = 61; 24.7 %; Fisher's exact test; p = 0.39), followed by two autosomal chromosomes I (n = 52; 21.1 %; p = 0.14) and III (n = 49; 20 %; p = 0.035). Overall, essential genes were in "hotspots" that were relatively evenly distributed on each of the chromosomes. For *O. volvulus*, most genes inferred to be essential were located to chromosome Ov1b (n = 45; 40.9 %; Fisher's exact test; p = 0.02), followed by chromosomes Ov2 (n = 25; 22.7 %; p = 0.03) and Ov3 (n = 24; 21.8 %; p = 0.19) (Fig. 4).

Second, we studied the distribution densities of the top essential and non-essential genes on the chromosomes for each species (Fig. 5). For *B. malayi*, slight differences in densities between essential and non-essential genes were detected – being skewed to the right arm on chromosomes I, III, and X, and to the left arm on chromosome IV, but clustering of essential genes was only significant for the X chromosome (Kolmogorov-Smirnov test; p = 0.003). For *O. volvulus*, there was no apparent clustering of predicted essential genes on chromosomes Ov1b

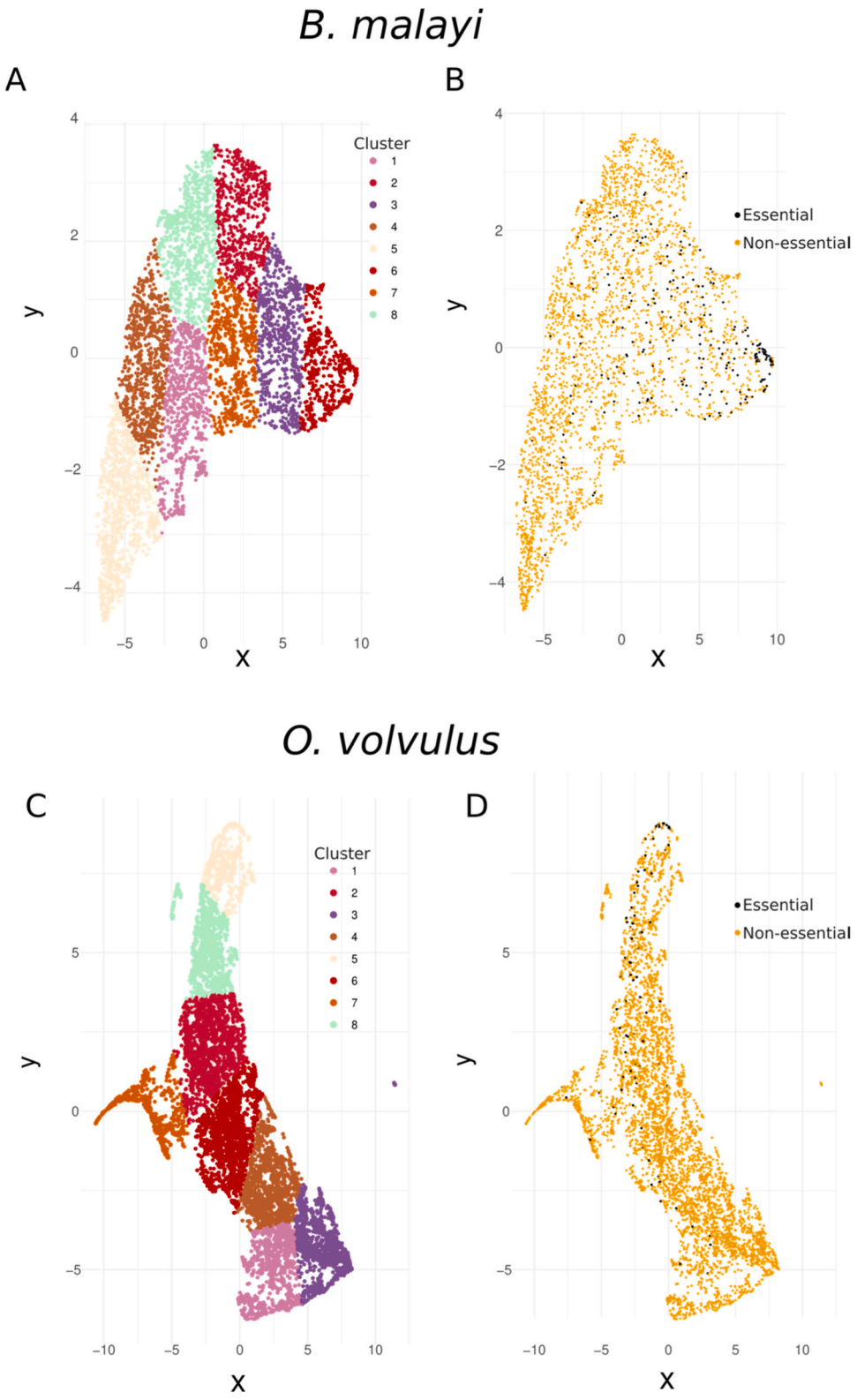


Fig. 3. Establishing the relationship between transcription profiles and essential genes using uniform manifold approximation and projection (UMAP: x- and y-axes). **A.** A selection of 5920 *Brugia malayi* genes was UMAP-clustered based on the level of transcription in 152 samples (RNA-seq; only the genes being transcribed in all samples included). **B.** UMAP plot with 221 of 246 essential (black) and 2857 of 5991 non-essential (orange) genes of *B. malayi* overlayed. **C.** Similarly, a selection of 8159 *Onchocerca volvulus* genes was clustered based on the level of transcription in 10 RNA-seq samples using UMAP and presented in a plot. **D.** UMAP plot with 97 of 110 essential (black) and 4519 of 7237 non-essential (orange) genes of *O. volvulus* overlayed.

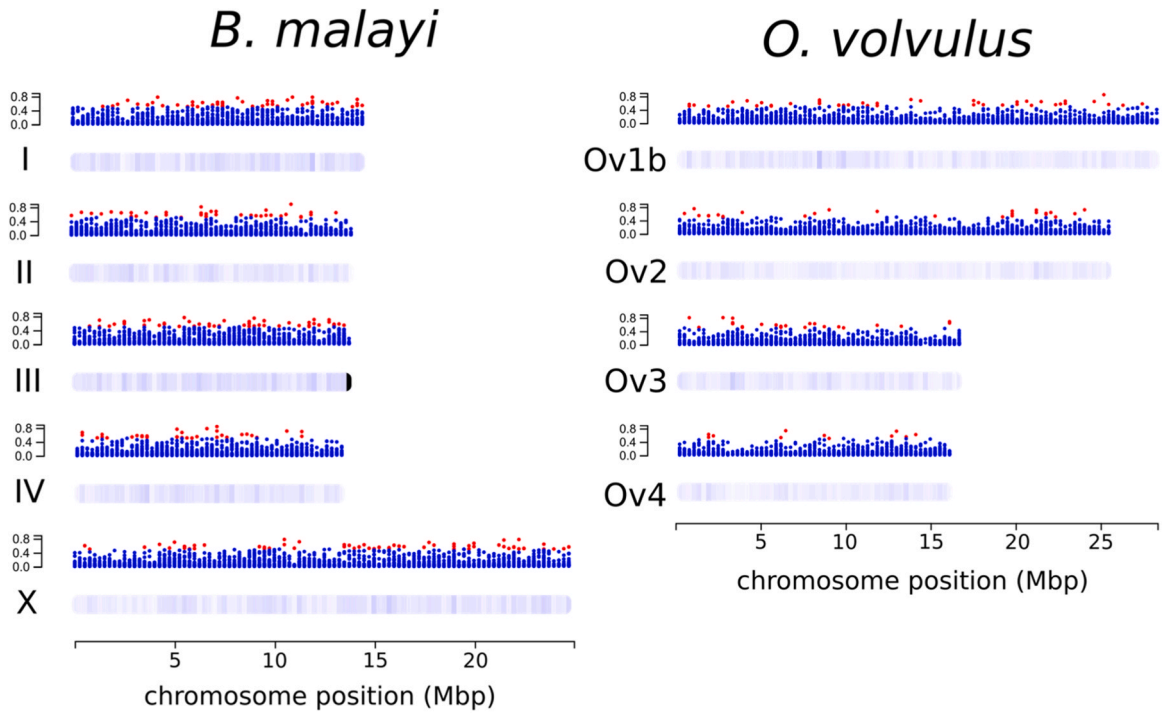


Fig. 4. The probabilities of all individual genes in *Brugia malayi* or *Onchocerca volvulus* being essential – as defined by machine learning (ML) – mapped to their respective genomic coordinates in chromosomes (largest segments for *O. volvulus*). The ordinate (y-axis) indicates ML-based prediction value for essential (red dots; probability > 0.5); non-essential or undefined (blue dots; probability < 0.5). The abscissa (x-axis) is chromosome location in megabase pairs (Mbp). The heatmap under each chromosome shows the density of genes; dark blue indicates a high density of genes; and black indicates the absence of genes in a particular region.

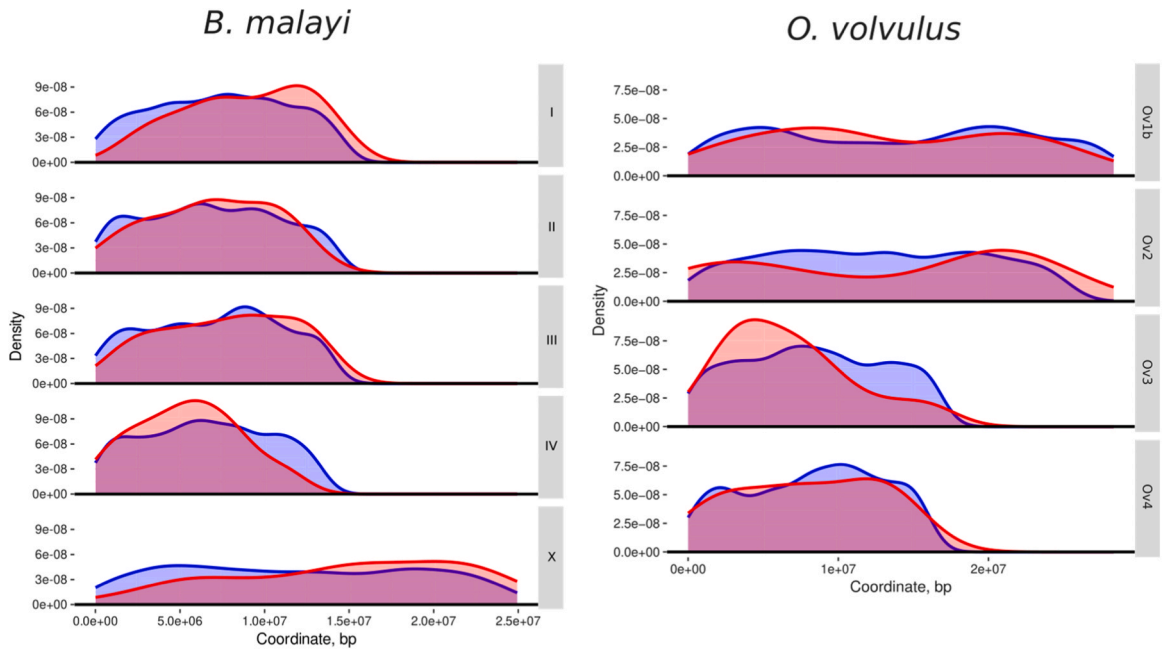


Fig. 5. The distribution densities of the 'top' inferred/prioritised essential genes (red) and non-essential genes (blue) in *Brugia malayi* (red: n = 246; blue: n = 5991) and in *Onchocerca volvulus* (red: n = 110; blue: n = 7237) – as inferred by machine learning (ML).

or Ov4, whereas the density of some of such genes was higher on the left arm of chromosome Ov3 or both arms of chromosome Ov2 (Fig. 5). However, Kolmogorov-Smirnov testing indicated that the distribution densities of essential versus non-essential were not significantly different ($p > 0.05$) in these chromosomes.

Third, using information available for *C. elegans* (see Cao et al., 2017), we inferred cell types or tissues in which essential genes were highly transcribed. To achieve this, we mapped the transcription of *C. elegans* orthologs of the predicted/prioritised essential genes of *B. malayi and O. volvulus* to known cell and tissue types in *C. elegans*. For genes predicted to be essential in *B. malayi*, *C. elegans* orthologs were highly transcribed in the germline (118 genes), somatic gonad precursors (89) or sex myoblasts (80); considering nerve cells only, 99 genes were inferred to be highly transcribed in amphid neurons with finger-

like ciliated endings (AFD), 83 in asymmetric sensory neurons ASE/L and 67 in ASE/R chemosensory neurons. For tissues, 241 genes were abundantly transcribed in the gonad, 135 in the hypodermis, followed by 134 in the body wall muscle and/or glia. For genes predicted to be essential in *O. volvulus*, *C. elegans* orthologs were highly transcribed in the germline (n = 65 genes), followed by Am/PH sheath cells (60), and body wall muscles (45). Considering nerve cells alone, 94 genes were transcribed in AFD, 80 in ASE/L and 63 in ASE/R chemosensory neurons. For tissues, 132 genes transcribed in the gonad, 102 in the glia, followed by 81 in the hypodermis and 78 in body wall muscle.

4. Discussion

Extending on previous work on the gene essentiality, particularly in the model ecdysozoans *C. elegans* and *D. melanogaster* (see [31]), this study provides the first comprehensive, large-scale prediction and investigation of essential genes in the parasitic nematodes *B. malayi* and *O. volvulus* using ML, and includes relevant, complementary analyses. We provide evidence of a relationship between essential genes and transcription and define a feature set that should be a useful resource for identifying essential genes in related filarioid nematodes.

To predict essential genes in *B. malayi* and *O. volvulus,* we used 26 features that had been shown to be strong predictors of essential genes within and between the model organisms *C. elegans* and *D. melanogaster* [30], and we defined and assessed one additional feature linked to sequence conservation which was inferred to be informative. Collectively, these 27 features were reliable predictors of essentiality in each and both model organisms, and were thus employed to predict essential genes in the two parasitic worms. Our analyses yielded some genomic sequence, annotation (exons), conservation and subcellular localisation characteristics that appear to be key for gene essentiality predictions, corroborating some previous studies [30,31,41–44].

Some features that link to essential genes of *C. elegans* and *D. melanogaster*, such as histone modification markers (e.g., H3K4me3 and H3K27me3; [23,24]), and have been shown to be important predictors of essentiality could not be assessed herein, as comparable data were not available for *B. malayi* or *O. volvulus* at the time of this study. Nonetheless, we inferred subsets of high-priority "essential genes" (Table S5) that appear to be exclusive to *B. malayi* or *O. volvulus*, which underpins the proposal that these genes or gene products represent novel intervention targets. The next step needs to test this hypothesis by experimentally verifying the essential function of these genes and their gene products using gene knockout and/or knockdown tools (cf. [45,46]).

We propose a strong relationship between essentiality and transcription profile. We showed that selections of essential genes clustered according to transcription profiles and that "essential genes" usually grouped together to the exclusion of "non-essential" genes predicted for each *B. malayi* and *O. volvulus* (Fig. 3). Through comparative analysis of scRNA-seq data for essential genes of *C. elegans*, we observed that > 50 % of the gene orthologs in *B. malayi* or *O. volvulus* were highly transcribed in tissues and cells of the reproductive tract (germline and associated tissues and cells). These findings pave the way for future studies of the functions, structures and/or interactions of essential proteins encoded in key cell types as a starting point for anthelmintic target validation.

For *O. volvulus*, the proportion of essential genes was higher ($p < 0.05$) on autosomes Ov1b and Ov2. For *B. malayi*, essential genes were more likely found on chromosome X than other chromosomes, but this association was not significant ($p > 0.05$). These findings contrast with those for the free-living nematode *C. elegans*, in which essential genes were less likely to be found on the sex chromosome X [23]. The relatively even distribution of essential genes on individual chromosomes of these filarioids was also distinct from that seen in *C. elegans* (in or near the centre of chromosomes; ref. [23]) or *D. melanogaster* (away from the centre/centromeres; ref. [24]). The distinction between *B. malayi*/*O. volvulus* and *C. elegans* might relate to the different genome and/or centromere organisation and/or gene regulatory mechanisms (genetic versus epigenetic) [47] and, obviously, their distinct biology. Using GO and pathway analyses, we inferred that many essential genes of *B. malayi* and *O. volvulus* are involved in transcriptional regulation and particularly in RNA-binding, ribosome formation and/or translation initiation functions, which supports previous findings for *C. elegans* and *D. melanogaster* (see [31]). Ribosome formation and translation initiation are biologically crucial and very energy-demanding [48–50], suggesting that the disruption or interruption of these processes and associated pathways in *B. malayi* or *O. volvulus* would lead to serious detrimental effects on this species.

In conclusion, the first genome-wide ML-based prediction of essential genes in *B. malayi* and *O. volvulus* provides hypotheses and a foundation to undertake functional investigations to assess the validity of these essentiality predictions, with the potential of determining the complement of genes that sustains life in these parasites. There is a prospect for the functional assessment of the genes predicted and prioritised here as essential using CRISPR-Cas9 [51] and potentially other functional genomic or chemical knockdown tools (e.g., [52]). Given the challenges associated with the treatment and control of filariases, it is particularly important to harness 'omic data sets for filarial worms, and to identify drug or vaccine targets using artificial intelligence (AI) methods. To this end, ML and other AI-based approaches are likely to contribute to accelerating fundamental and applied investigations of essential genes and to evaluating their suitability as drug targets, thus enabling the future design of novel and effective treatments. We are confident that the approach employed here can be extended to explore gene essentiality in other ecdysozoan parasites for which high-quality genomes and transcriptomic data sets are available.

CRediT authorship contribution statement

Robin Gasser: Writing – review & editing, Writing – original draft, Supervision, Resources, Methodology, Funding acquisition, Conceptualization. **Bill Chang**: Validation. **Neil Young**: Writing – review & editing, Validation. **Pasi Korhonen**: Writing – review & editing, Validation, Data curation. **Tulio Campos**: Writing – original draft, Methodology, Formal analysis, Data curation, Conceptualization.

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.

Data Availability

The data and code used are publicly available: https://bitbucket.org/tuliocampos/essential_bmalayi and https://bitbucket.org/tuliocampos/essential_ovolvulus.

Acknowledgements

This work was supported by funding from the Australian Research Council (ARC; LP180101085 and LP220200614).

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2024.07.025.

References

[1] Tolle MA. Mosquito-borne diseases. Curr Probl Pediatr Adolesc Health Care 2009; 39:97–140.

[2] Khan S, Hall TC, Land WG, Dovč-Drnovšek T, Klemenc P, Toplak N, et al. Arbonematodes - nematode infections transmissible by arthropods. Transfus Med Hemother 2013;40:50–62.
[3] Cholewinski M, Derda M, Hadas E. Parasitic diseases in humans transmitted by vectors. Ann Parasitol 2015;61:137–57.
[4] Berenger J-M, Parola P. Arthropod Vectors of Medical Importance. Infect Dis (Auckl). forth ed.,, 1. Amsterdam: Elsevier; 2016. p. 104–12.
[5] Plaisier AP, van Oortmarssen GJ, Habbema JD, Remme J, Alley ES. ONCHOSIM: a model and computer simulation program for the transmission and control of onchocerciasis. Comput Methods Prog Biomed 1990;31:43–56.
[6] Burnham G. Onchocerciasis. Lancet 1998;351:1341–6.
[7] Edeson J, Wilson T. The epidemiology of filariasis due to *Wuchereria bancrofti* and *Brugia malayi*. Ann Rev Entomol 1964;9:245–68.
[8] Fischer P, Supali T, Maizels RM. Lymphatic filariasis and *Brugia timori*: prospects for elimination. Trends Parasitol 2004;20:351–5.
[9] Naing C, Whittaker MA, Tung WS, Aung H, Mak JW. Prevalence of zoonotic (brugian) filariasis in Asia: a proportional meta-analysis. Acta Trop 2024;249: 107049.
[10] Kamgno J, Djeunga HN. Progress towards global elimination of lymphatic filariasis. Lancet Glob Health 2020;8:e1108–9.
[11] Dixon R, Lar L, Dean L. Neglect in the numbers: leaving no voice behind in disease elimination. Lancet Glob Health 2021;9:e22.
[12] Lupenza ET, Gasarasi DB, Minzi OM. Lymphatic filariasis elimination status: *Wuchereria bancrofti* infections in human populations and factors contributing to continued transmission after seven rounds of mass drug administration in Masasi District, Tanzania. PLoS One 2022;17:e0262693.
[13] Hotez PJ, Lo NC. Neglected tropical diseases: public health control programs and mass drug administration. In Hunter's Tropical Medicine and Emerging Infectious Diseases 2020;27:209–13.
[14] Abdul Halim AFN, Ahmad D, Miaw Yn JL, Masdor NA, Ramly N, Othman R, et al. Factors associated with the acceptability of mass drug administration for filariasis: a systematic review. Int J Env Res Pub Health 2022;19:12971.
[15] WHO. Global programme to eliminate lymphatic filariasis: progress report, 2021 - Programme mondial pour l'élimination de la filariose lymphatique: rapport de situation, 2021. Weekly Epidemiological Record - Relevé épidémiologique hebdomadaire 2022;97:513–524
[16] Maddren R, Phillips A, Gomez SR, Forbes K, Collyer BS, Kura K, et al. Individual longitudinal compliance to neglected tropical disease mass drug administration programmes, a systematic review. PLoS Negl Trop Dis 2023;17:e0010853.
[17] Chavda VP, Pandya A, Pulakkat S, Soniwala M, Patravale V. Lymphatic filariasis vaccine development: neglected for how long? Expert Rev Vaccin 2021;20: 1471–82.
[18] Ugbe FA, Shallangwa GA, Uzairu A, Abdulkadir I. Theoretical modeling and design of some pyrazolopyrimidine derivatives as *Wolbachia* inhibitors, targeting lymphatic filariasis and onchocerciasis. Silico Pharm 2022;10:8.
[19] Scott AL, Ghedin E. The genome of *Brugia malayi* - all worms are not created equal. Parasitol Int 2009;58:6–11.
[20] Schneider P, Walters WP, Plowright AT, Sieroka N, Listgarten J, Goodnow Jr RA, et al. Rethinking drug design in the artificial intelligence era. Nat Rev Drug Discov 2020;19:353–64.
[21] Geary TG, Sakanari JA, Caffrey CR. Anthelmintic drug discovery: into the future. J Parasitol 2015;101:125–33.
[22] Sepúlveda-Crespo D, Reguera RM, Rojo-Vásquez F, Balaña-Fouce R, Martínez-Valladares M. Drug discovery technologies: *Caenorhabditis elegans* as a model for anthelmintic therapeutics. Med Res Rev 2020;40:1715–53.
[23] Campos TL, Korhonen PK, Sternberg PW, Gasser RB, Young ND. Predicting gene essentiality in *Caenorhabditis elegans* by feature engineering and machine-learning. Comput Struct Biotechnol 2020;15:1093–102.
[24] Campos TL, Korhonen PK, Hofmann A, Gasser RB, Young ND. Combined use of feature engineering and machine-learning to predict essential genes in *Drosophila melanogaster*. NAR Genom Bioinform 2020;22:lqaa051.
[25] Howe KL, Bolt BJ, Shafie M, Kersey P, Berriman M. WormBase ParaSite - a comprehensive resource for helminth genomics. Mol Biochem Parasitol 2017;215: 2–10.
[26] Marygold S.J., Crosby M.A., Goodman JL – FlyBase Consortium. Using FlyBase, a database of Drosophila genes & genomes. In: Dahmann C. (eds) *Drosophila*. Methods Mol Biol 2016;1478:1–31.
[27] Harris TW, Arnaboldi V, Cain S, Chan J, Chen WJ, Cho J, et al. WormBase: a modern model organism information resource. Nucleic Acids Res 2020;48:D762–7.
[28] Kimble J, Nüsslein-Volhard C. The great small organisms of developmental genetics: *Caenorhabditis elegans* and *Drosophila melanogaster*. Dev Biol 2022;485: 93–122.
[29] Sternberg PW, Van Auken K, Wang Q, Wright A, Yook K, Zarowiecki M, et al. WormBase 2024: status and transitioning to Alliance infrastructure. Genetics 2024; 4:iyae050.
[30] Campos TL, Korhonen PK, Young ND. Cross-predicting essential genes between two model eukaryotic species using machine learning. Int J Mol Sci 2021;22:5056.
[31] Campos TL, Korhonen PK, Hofmann A, Gasser RB, Young ND. Harnessing model organism genomics to underpin the machine-learning-based prediction of essential genes in eukaryotes – Biotechnological implications. Biotechnol Adv 2021;54: 107822.
[32] Ghedin E, Wang S, Spiro D, Caler E, Zhao Q, Crabtree J, et al. Draft genome of the filarial nematode parasite *Brugia malayi*. Science 2007;317:1756–60.
[33] Cotton JA, Bennuru S, Grote A, Harsha B, Tracey A, Beech R, et al. The genome of *Onchocerca volvulus*, agent of river blindness. Nat Microbiol 2016;2:16216.
[34] Armenteros JJA, Sønderby CK, Sønderby SK, Nielsen H, Winther O. DeepLoc: prediction of protein subcellular localization using deep learning. Bioinformatics 2017;33:3387–95.
[35] Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ, et al. A new view of the Tree of Life. Nat Microbiol 2016;1:16048.
[36] Howe KL, Achuthan P, Allen J, Allen J, Alvarez-Jarreta J, Amode MR, et al. Ensembl 2021. Nucleic Acids Res 2021;49:D884–91.
[37] Emms DM, Kelly S. OrthoFinder: phylogenetic ortholog inference for comparative genomics. Genome Biol 2019;20:238.
[38] 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.
[39] Raudvere U, Kolberg L, Kuzmin I, Arak T, Adler P, Peterson H, et al. g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). Nucleic Acids Res 2019;47:W191–8.
[40] Gillespie M, Jassal B, Stephan R, Milacic M, Rothfels K, Senff-Ribeiro A, et al. The reactome pathway knowledgebase 2022. Nucleic Acids Res 2022;50:D687–92.
[41] Aromolaran O, Beder T, Oswald M, Oyelade J, Adebiyi E, Koening R. Essential gene prediction in *Drosophila melanogaster* using machine learning approaches based on sequence and functional features. Comput Struct Biotechnol J 2020;10:612–21.
[42] Aromolaran O, Aromolaran D, Isewon I, Oyelade J. Machine learning approach to gene essentiality prediction: a review. Brief Bioinform 2021;22:bbab128.
[43] Beder T, Aromolaran O, Dönitz J, Tapanelli S, Adedeji EO, Adebiyi E, et al. Identifying essential genes across eukaryotes by machine learning. NAR Genom Bioinform 2021;3:lqab110.
[44] Marques de Castro G, Hastenreiter Z, Silva Monteiro TA, Martins da Silva TT, Pereira Lobo F. Cross-species prediction of essential genes in insects. Bioinformatics 2022;6:btac009.
[45] Boettcher M, McManus M. Choosing the right tool for the job: RNAi, TALEN, CRISPR. Mol Cell 2015;58:575–85.
[46] Quinzo MJ, Perteguer MJ, Brindley PJ, Loukas A, Sotillo J. Transgenesis in parasitic helminths: a brief history and prospects for the future. Parasit Vectors 2022;15:110.
[47] Carlton PM, Davis RE, Ahmed S. Nematode chromosomes. Genetics 2022;221: iyac014.
[48] Mayer C, Grummt I. Ribosome biogenesis and cell growth: mTOR coordinates transcription by all three classes of nuclear RNA polymerases. Oncogene 2006;25: 6384–91.
[49] Kressler D, Hurt E, Bassler J. Driving ribosome assembly. Biochim Biophys Acta 2010;1803:673–83.
[50] Zhou X, Liao WJ, Liao JM, Liao P, Lu H. Ribosomal proteins: functions beyond the ribosome. J Mol Cell Biol 2015;7:92–104.
[51] Kwarteng A, Sylverken A, Asiedu E, Ahuno ST. Genome editing as control tool for filarial infections. Biomed Pharmacother 2021;137:111292.
[52] Wheeler NJ, Heimark ZW, Airs PM, Mann A, Bartholomay LC, Zamanian M. Genetic and functional diversification of chemosensory pathway receptors in mosquito-borne filarial nematodes. PLoS Biol 2020;18:e3000723.