Combined use of feature engineering and machine-learning to predict essential genes in *Drosophila melanogaster* and its implications for other arthropods

Tulio L. Campos[1,2*], Pasi K. Korhonen[1], Andreas Hofmann[3], Robin B. Gasser[1], Neil D. Young[1]

*[1] Melbourne Veterinary School, The University of Melbourne, Parkville, Victoria 3010, Australia; [2] Instituto Aggeu Magalhães (IAM-Fiocruz), Pernambuco, Brazil; [3] Griffith Institute for Drug Discovery, Griffith University, Brisbane, Queensland 4111, Australia. * **tulio.campos@unimelb.edu.au***

Background

- *D. melanogaster* is a model to understand the biology of multicellular organisms (metazoans).
- High-quality reference genome and extensive functional genomic data sets are available.
- Major demand for a reliable computational method to infer which genes are essential for life
- Machine learning (ML) could provide the solution.
- Using available data for *D. melanogaster*, we critically assessed the feasibility of a ML-approach for gene essentiality prediction.

Aims

1. To create and employ a scoring-system to provisionally assign essentiality to genes using phenotypic data.

2. Large-scale extraction/engineering and selection of features associated with those genes from extensive 'omic data sets.

3. To construct and systematically evaluate a machine-learning (ML)-based workflow for the genome-wide prediction of essential genes in *D. melanogaster*.

Methodology established, and Results

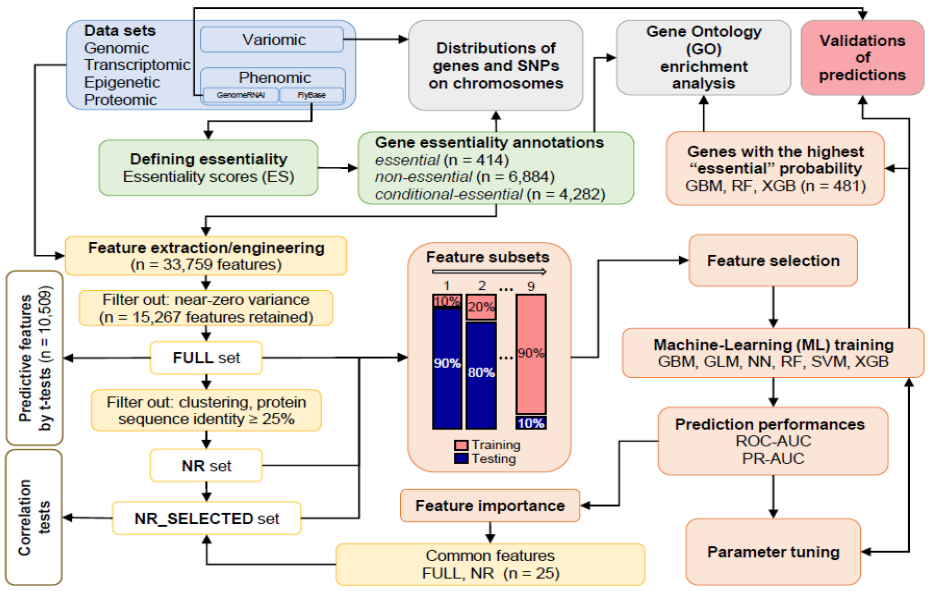

Figure 1 – Workflow employed. Data sets used (blue), provisional annotations of essential genes (green), preparation of feature sets (yellow), systematic feature selection and ML approaches (orange), complementary analyses (grey), independent validation (red). The data used and the code developed are available at: https://bitbucket.org/tuliocampos/essential_melanogaster (*Campos et al., 2020*).

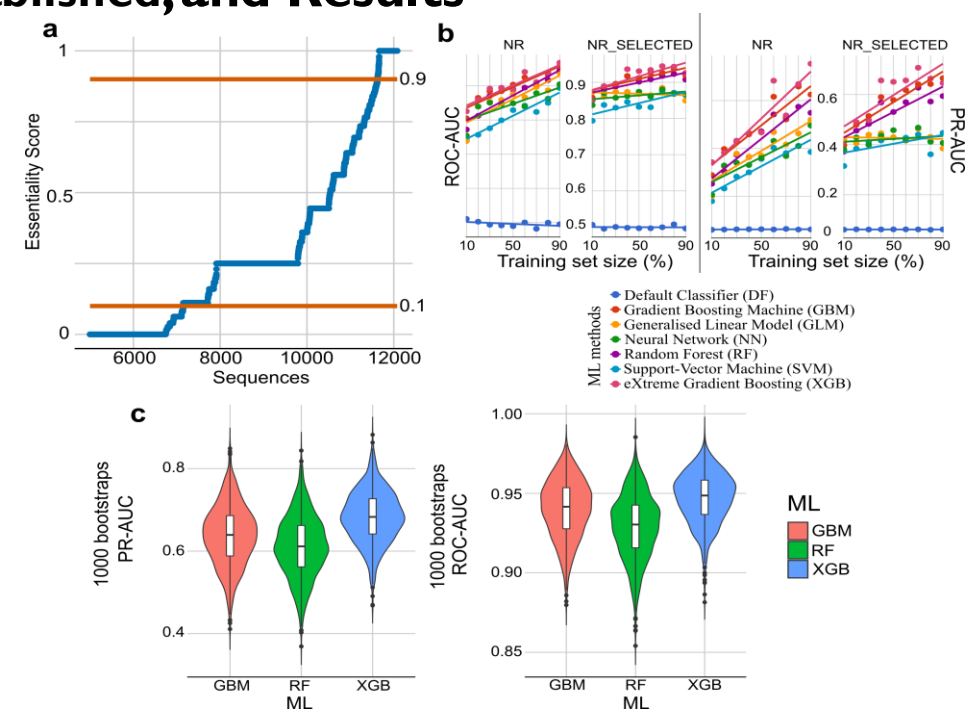

Figure 2 – (a) Scoring system for provisional annotations. (b) Systematic ML approach using subsets of different sizes for training, ROC- and PR-AUC calculated on test sets. (c) Bootstrapping approach (90% training, 10% testing) – RF, XGB and GBM methods.

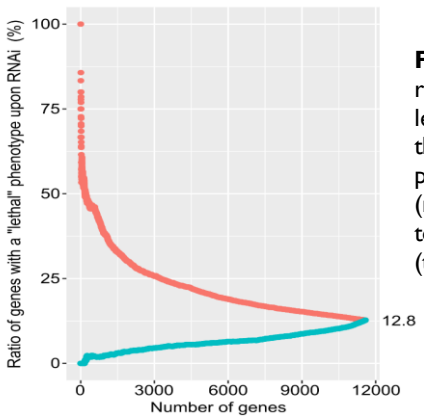

Figure 3 – Cumulative ratios of genes with a lethal phenotype. From the highest to lowest: prediction probabilities (red); from the lowest to the highest (turquoise).

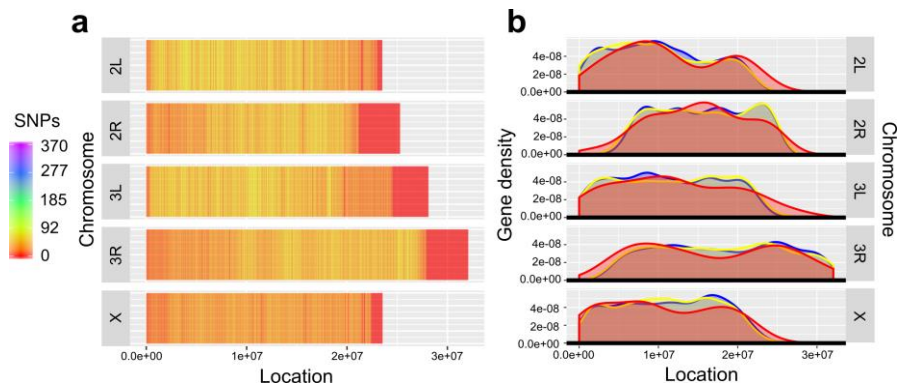

Figure 4 – (a) SNP counts along *D. melanogaster* chromosomes per 1000 bp windows. (b) Distributions of gene locations by essentiality annotations (red: essential, blue: non-essential, yellow: conditional).

Discussion

- Gene essentiality can be predicted reliably in *D. melanogaster* using well-trained machine-learning models.
- A combination of feature engineering/extraction/selection identified strong predictors.
- Best predictors were features derived from nucleotide/protein sequences, subcellular localisation and transcriptomic data (RNA-seq).
- ML methods achieved high prediction performance (XGB, GBM and RF).
- Essential genes located away from telomeric and centromeric regions of chromosomes. Little relationship with regions of low SNP density.
- Functional roles of essential genes: protein/nucleotide processing, reproductive tissues, potential regulation through mTOR pathway.

Conclusions

- ML-based workflow developed here is promising for application in other species.
- Future work: essential gene predictions using ML in arthropod parasites, pests and vectors of infectious diseases.

Acknowledgements

Australian Government
Australian Research Council
Ministério da Saúde
FIOCRUZ
Fundação Oswaldo Cruz
Instituto Aggeu Magalhães
Australian Government
National Health and Medical Research Council
The Australian Society for Parasitology Inc.