Advance Access Publication Date: Day Month Year
Application Note

Systems biology

promor: a comprehensive R package for label-free proteomics data analysis and predictive modeling

Chathurani Ranathunge [1,*], Sagar S. Patel [1], Lubna Pinky [1], Vanessa L. Correll [2], O. John Semmes [2], Robert K. Armstrong [1,3], C. Donald Combs [1], and Julius O. Nyalwidhe [2]

[1]School of Health Professions, Eastern Virginia Medical School, Norfolk, VA 23501, USA.
[2]The Leroy T. Canoles Jr. Cancer Research Center, Eastern Virginia Medical School, Norfolk, VA 23501, USA.
[3]Sentara Center for Simulation and Immersive Learning, Eastern Virginia Medical School, Norfolk, VA 23501, USA.

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Summary: We present promor, a comprehensive, user-friendly R package that streamlines label-free (LFQ) proteomics data analysis and building machine learning-based predictive models with top protein candidates.
Availability and implementation: promor is freely available as an open source R package on the Comprehensive R Archive Network (CRAN)(https://CRAN.R-project.org/package=promor) and distributed under the Lesser General Public License (version 2.1 or later). Development version of promor is maintained on GitHub (https://github.com/caranathunge/promor) and additional documentation and tutorials are provided on the package website (https://caranathunge.github.io/promor/).
Contact: caranathunge86@gmail.com

1 Introduction

Label-free quantification (LFQ) approaches are fast becoming popular in mass spectrometry-based proteomics. One of the most widely used software tools for protein identification and quantification is MaxQuant (Tyanova *et al.*, 2016). However, the downstream analysis of MaxQuant output files can be complex and challenging. Some tools available for this purpose are implemented as graphical user interface (GUI) applications, while others, such as MSstats (Choi *et al.*, 2014), are primarily implemented as R packages facilitating greater analytical flexibility and reproducibility.

In recent years, machine learning (ML) has made its presence felt in the field of proteomics. Particularly in biomarker research, ML algorithms are being widely employed to build proteomics-based predictive models of disease prognosis and diagnosis. When building a proteomics-based predictive model, choosing a robust panel of protein candidates can greatly improve the accuracy of the model. In this regard, ML-based predictive models could benefit from narrowing down protein features to those that show significant differences in abundance between groups of interest. In the current landscape of proteomics data analytical tools the capability to seamlessly transition from differential expression analysis to predictive modeling is lacking. Realizing this need, we developed promor, a comprehensive, user-friendly, R package that streamlines differential expression analysis and predictive modeling of label-free proteomics data.

2 Overview

2.1 Implementation

promor is implemented in R ($\geq$ 3.5.0) and relies on packages such as imputeLCMD (Lazar *et al.*, 2016), limma (Ritchie *et al.*, 2015), and caret (Kuhn, 2008) for back-end pre-processing, differential expression analysis, and machine learning-based modeling. As input, promor uses the MaxQuant-produced "proteinGroups.txt" and a user-generated tab-delimited text file containing the experimental design. For visualization, promor employs the popular ggplot2 (Wickham *et al.*, 2016) architecture and produces ggplot objects, which allows for further customization (Figure 1).

2.2 Proteomics data analysis

Proteomics data analysis workflow in promor (Figure 1A) consists of multiple filtration steps to filter out contaminant proteins, reverse-sequence proteins, proteins identified "only-by-site", and proteins identified by fewer than a user-specified number of unique peptides. User can also filter out technical replicates that show weak correlation of intensities. promor provides five missing data imputation methods (minProb,

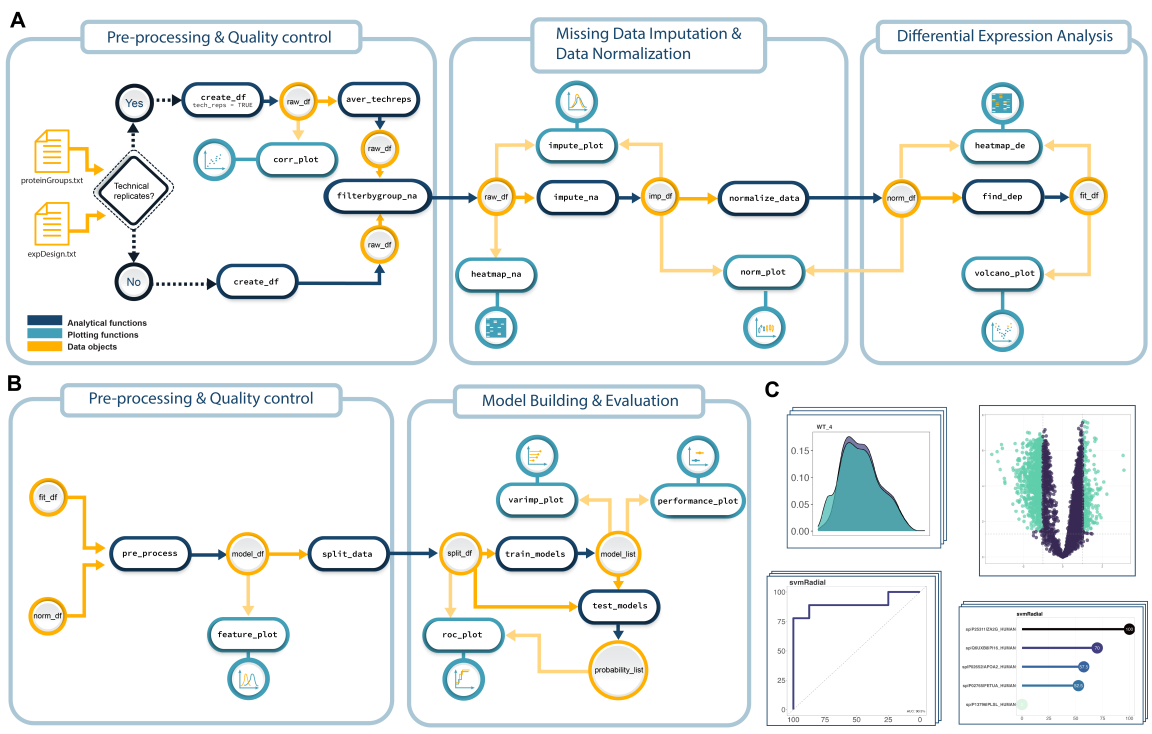


Fig. 1. An overview of promor workflows. (A) Proteomics data analysis workflow includes analytical functions for pre-processing, quality control, missing data imputation, data normalization, and differential expression analysis. (B) Modeling workflow includes analytical functions for pre-processing the output of differential expression analysis, model building, and model evaluation. (C) Several plotting functions are provided to visualize data and produce publication-ready figures using color blind-friendly palettes.

minDet, kNN, SVD, and Random Forest) and three data normalization methods (quantile, scale, and cyclicloess). Differential expression analysis can be performed using the moderated t-test implemented in the limma package (Ritchie *et al.*, 2015). We provide two tutorials to demonstrate promor's proteomics data analysis workflows here: https://caranathunge.github.io/promor/articles/. More information on the data used in these tutorials and the results obtained are provided in Supplementary Note 1, Supplementary Tables S1 - S2, and Supplementary Figures S1 - S7.

2.3 Building predictive models

Differentially expressed proteins can be directly used in the modeling workflow to build predictive models using multiple machine learning algorithms simultaneously (Figure 1B). Over 200 machine learning algorithms are made accessible through the caret package (Kuhn, 2008). For users inexperienced in complex machine learning algorithms, promor provides a default list of four widely used classification-based algorithms ("rf", "glm", "svmRadial", and "xgbLinear"), chosen to represent a variety of machine learning model types. Additionally, promor provides a range of functions to assess and visualize protein (feature) variation among conditions (classes), variable (feature) importance, model performance, and the predictive power of models (Figure 1B). A tutorial demonstrating the modeling workflow is provided here: https://caranathunge.github.io/promor/articles/. More information on the data used in the modeling workflow tutorial and the results obtained are provided in Supplementary Note 2 and Supplementary Figures S8 - S11.

3 Conclusions

We present promor, a user-friendly, comprehensive R package that facilitates seamless transition from differential expression analysis of proteomics data to building predictive models with top protein candidates; a feature that could be particularly useful in clinical and biomarker research.

4 Acknowledgements

We wish to thank Asitha I. Senanayake for his helpful comments and discussions on software development.

5 Funding

This work was supported by the Hampton Roads Biomedical Research Consortium.

References

Choi,M., Chang,C.Y., Clough,T., Broudy,D., Killeen,T., MacLean,B. and Vitek,O. (2014) Msstats: an r package for statistical analysis of quantitative mass spectrometry-based proteomic experiments. *Bioinformatics,* **30** (17), 2524–2526.

Kuhn,M. (2008) Building predictive models in r using the caret package. *Journal of statistical software,* **28**, 1–26.

Lazar,C., Gatto,L., Ferro,M., Bruley,C. and Burger,T. (2016) Accounting for the multiple natures of missing values in label-free quantitative proteomics data sets to compare imputation strategies. *Journal of proteome research,* **15** (4), 1116–1125.

picture(0,0)(-35,0)(1,0)30 (0,35)(0,-1)30 picture

picture(0,0)(35,0)(-1,0)30 (0,35)(0,-1)30 picture

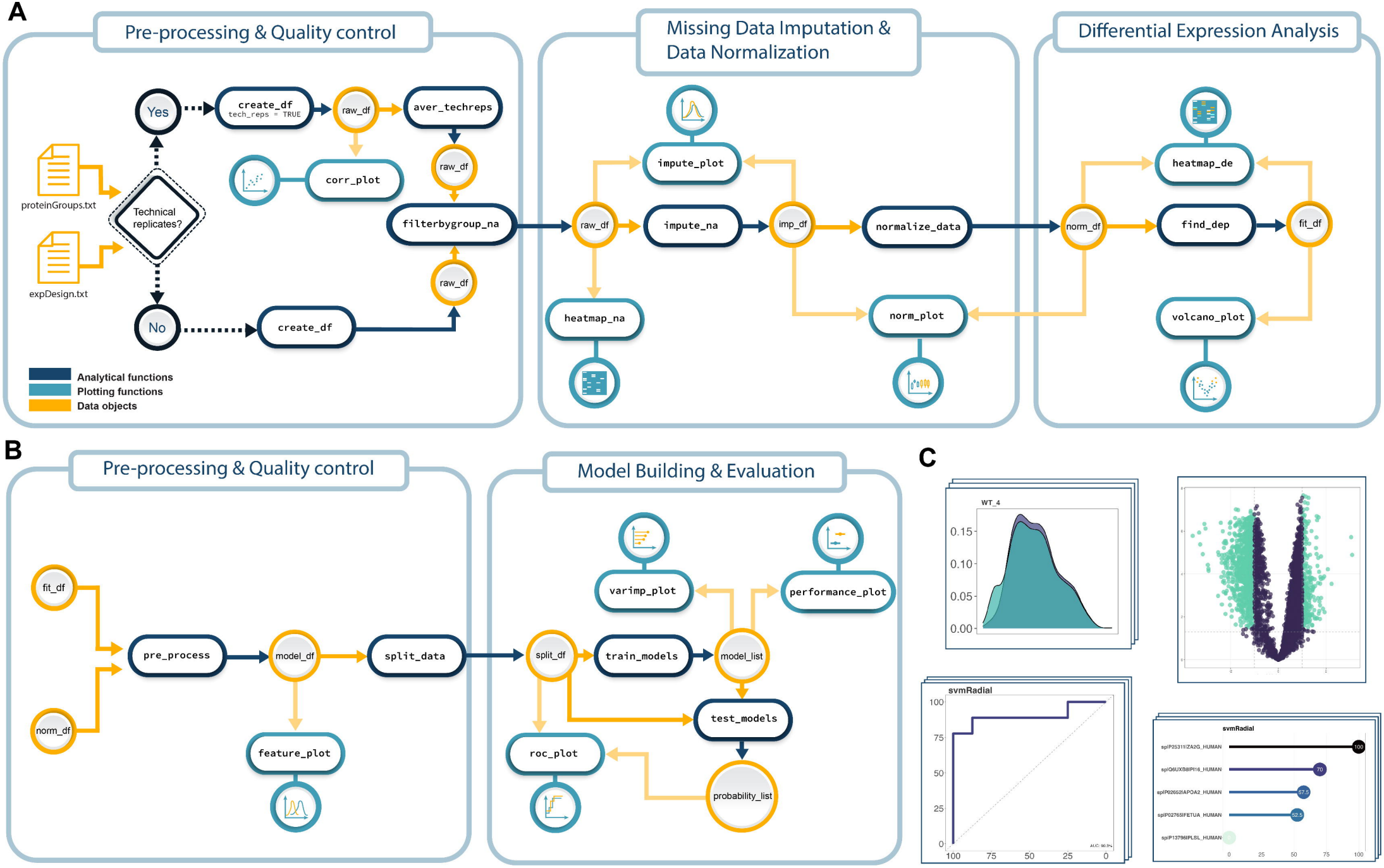
Ritchie,M.E., Phipson,B., Wu,D., Hu,Y., Law,C.W., Shi,W. and Smyth,G.K. (2015) limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic acids research,* **43** (7), e47–e47.

Tyanova,S., Temu,T. and Cox,J. (2016) The maxquant computational platform for mass spectrometry-based shotgun proteomics. *Nature protocols,* **11** (12), 2301–2319.

Wickham,H., Chang,W. and Wickham,M.H. (2016) Package 'ggplot2'. *Create elegant data visualisations using the grammar of graphics. Version,* **2** (1), 1–189.

picture(0,0)(-35,0)(1,0)30 (0,-35)(0,1)30 picture

picture(0,0)(35,0)(-1,0)30 (0,-35)(0,1)30 picture


A
Pre-processing & Quality control
Missing Data Imputation & Data Normalization
Differential Expression Analysis
Yes
create_df
tech_reps = TRUE
raw_df
aver_techreps
raw_df
corr_plot
filterbygroup_na
raw_df
proteinGroups.txt
Technical replicates?
expDesign.txt
No
create_df
impute_plot
raw_df
impute_na
imp_df
normalize_data
heatmap_na
norm_plot
norm_df
heatmap_de
find_dep
fit_df
volcano_plot
Analytical functions
Plotting functions
Data objects
B
Pre-processing & Quality control
Model Building & Evaluation
fit_df
pre_process
model_df
split_data
norm_df
feature_plot
varimp_plot
performance_plot
split_df
train_models
model_list
test_models
roc_plot
probability_list
C
WT_4
0.15
0.10
0.05
0.00
svmRadial
100
75
50
25
0
100
75
50
25
0
svmRadial
sp|P25311|ZA2G_HUMAN
sp|Q6UXB8|PI16_HUMAN
sp|P02652|APOA2_HUMAN
sp|P02765|FETUA_HUMAN
sp|P13796|PLSL_HUMAN
100
70
67.5
62.5
0
25
50
75
100