Bioinformatics, 2017, 1–3
doi: 10.1093/bioinformatics/btx795
Advance Access Publication Date: 11 December 2017
Applications Note

Gene expression

DaMiRseq—an R/Bioconductor package for data mining of RNA-Seq data: normalization, feature selection and classification

Mattia Chiesa, Gualtiero I. Colombo and Luca Piacentini*

Immunology and Functional Genomics Unit, Centro Cardiologico Monzino, IRCCS, 20138 Milan, Italy

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on April 11, 2017; revised on December 2, 2017; editorial decision on December 4, 2017; accepted on December 6, 2017

Abstract

Summary: RNA-Seq is becoming the technique of choice for high-throughput transcriptome profiling, which, besides class comparison for differential expression, promises to be an effective and powerful tool for biomarker discovery. However, a systematic analysis of high-dimensional genomic data is a demanding task for such a purpose. DaMiRseq offers an organized, flexible and convenient framework to remove noise and bias, select the most informative features and perform accurate classification.

Availability and implementation: DaMiRseq is developed for the R environment (R $\geq$3.4) and is released under GPL ($\geq$2) License. The package runs on Windows, Linux and Macintosh operating systems and is freely available to non-commercial users at the Bioconductor open-source, open-development software project repository (https://bioconductor.org/packages/DaMiRseq/). In compliance with Bioconductor standards, the authors ensure stable package maintenance through software and documentation updates.

Contact: luca.piacentini@ccfm.it

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The analysis of high-dimensional RNA-Seq data is a challenging task that researchers may tackle through data mining and statistical learning methods (Libbrecht and Noble, 2015). The basic task of many software packages is the differential expression analysis of genomic features for conventional class comparison studies. Conversely, comprehensive tools for systematic analysis of RNA-Seq data that identify relevant features to classify phenotypes are still lacking. Here, we propose the DaMiRseq package—a structured and convenient workflow to effectively identify transcriptional biomarkers and exploit them for classification purposes.

2 DaMiRseq philosophy

Using a thoughtful decision-making process, DaMiRseq guides the user to select the best putative predictors for classification. It is structured into three main processes: normalization, feature selection and classification.

Normalization includes (i) basic preprocessing of raw counts, (ii) removal of hypervariant genes, i.e. genes with anomalous read counts by comparing to the mean value across samples of the same group and (iii) data transformation. An integral and distinctive part of DaMiRseq data normalization is the removal of the effects of 'unwanted variation.' Canonical normalization methods may still retain systematic sources of variation in expression data (Qiu *et al.*, 2005). Regardless of the protocols and platforms employed, the intrinsic complexity of RNA-Seq method easily introduces technical biases, which may mask biological variability. DaMiRseq identifies putative hidden factors and uses them to adjust expression data to produce more accurate and robust prediction models (Jaffe *et al.*, 2015).

RNA-Seq data usually consists of tens of thousands of features most of which are either irrelevant or redundant for classification purposes. DaMiRseq provides a 'Feature Selection' procedure to extract a small subset of informative features from the original data. Identifying such a subset presents clear benefits since it improves classification performance, reduces training processing time, and produces more cost-effective models (Saeys *et al.*, 2007).

Classification exploits informative features to draw a 'meta-learner' through a 'stacking' ensemble learning approach (Rokach, 2010). The underlying idea is that single, 'weaker' classifiers may have different or insufficient generalization performances leading to incorrect predictions; conversely, weighting and combining the prediction of several classifiers may reduce overall classification errors. The 'weighted voting' method (Littlestone and Warmuth, 1994) we propose herein assesses the performance of each classifier and allows the meta-learner to reach high classification accuracies, better than or comparable with the best single, weak classifiers.

3 Implementation

Details and comparisons with other methods on normalization, feature selection and classification are available in the online-only Supplementary Material. Section 1 of the Supplementary Material describes sample data used for comparisons.

3.1 Reading the data

DaMiRseq requires as input a SummarizedExpression object (Morgan *et al.*, 2016). It incorporates read counts and sample annotations, and ensures interoperability among other R/Bioconductor packages. If data are stored in matrices or data frames, users may exploit *DaMiR.makeSE* to check data suitability and easily construct a SummarizedExpression object.

3.2 Normalization

DaMiR.normalization filters non-expressed and hypervariant genes and applies either vst or rlog read count transformation (Love *et al.*, 2014). The *DaMiR.sampleFilt* introduces a sample quality checkpoint. It computes the mean absolute correlation of each sample so that users may filter out low correlated samples. The *DaMiR.SV* identifies hidden factors subsequently used to adjust expression data. We propose a novel approach, i.e. the weighted 'fraction of variance explained (fve)', to identify the optimal number of surrogate variables (sv) along with the Surrogate Variable Analysis (SVA) by Leek *et al.* (2012). In section 2 and 3 of the Supplementary Material, we describe our implementation and provide comparisons with the methods included in SVA. The *DaMiR.corrplot* helps to visually inspect the correlation between sv and, if available, other known variables. Users may thus evaluate those correlations to prevent inappropriate removal of secondary signals of interest. Finally, *DaMiR.SVadjust*, which embeds the *removeBatchEffect* function of the limma package (Ritchie *et al.*, 2015), removes from the expression matrix the 'unwanted variation' associated with the user-selected number of sv.

3.3 Feature selection

DaMiRseq presents a singular procedure by combining existing techniques to efficiently extracting the most informative class-related variables in a time-saving fashion. *DaMiR.FSelect* implements a method based on the backward variable elimination-Partial Least Squares to remove the less informative variables with respect to class (Ildiko, 1987). *DaMiR.FReduct* assesses the presence of redundant, highly correlated features that may decrease classification performance. It produces a pair-wise absolute correlation matrix and filters out those features that display large mean absolute correlations. Finally, *DaMiR.FSort* sorts the features by RReliefF score (Robnik-Šikonja and Kononenko, 1997), whereas *DaMiR.FBest* extracts a small subset of ranked features to test as predictors for classification. We provide deeper explanations in section 4 of the Supplementary Material and show that our approach outperforms other popular methods for feature selection, namely random forest and naïve Bayes.

3.4 Classification

DaMiR.EnsembleLearning performs a novel procedure by weighting and combining the outputs of Random Forest, Naïve Bayes, 3-Nearest Neighbours, Logistic Regression, Linear Discriminant Analysis and Support Vectors Machines classifiers. Our method estimates an accuracy-based weight for each classifier that, together with predictions on unseen data, produces the final prediction. The function assesses the robustness of predictors by bootstrap resampling. Overall, better weak classifiers influence meta-learner more than worse classifiers, which ensure the 'Ensemble' approach reaching high classification accuracy independently of the experimental settings and assumptions about data. Section 5 of the Supplementary Material details our classification strategy and shows improvement over tree-based and support vector machine methods of other R implementations.

3.5 Visualizing the data

The helper function *DaMiR.Allplot*. generates several diagnostic plots, i.e. clustering/heatmaps, multidimensional scaling and relative log expression plots that users may exploit to evaluate DaMiRseq performance, e.g. the suitability of normalization procedures. *DaMiR.Clustplot*, draws a clustering/heatmap of a gene (rows) × sample (cols) expression matrix. See examples in section 6 of the Supplementary Material.

3.6 Note on generalization

Resampling is used in the classification step to accurately estimate prediction error and choose the most robust set of biomarkers. Normalization and feature selection steps exploit, instead, a full dataset. For a more generalizable approach, the user might consider to split the dataset at the very beginning of the entire procedure, find biomarkers on a training set and validate them on the normalized and adjusted test set.

Acknowledgement

We wish to thank Dr. Aoife Gowran for her writing review and editing.

Funding

Institutional Research Funds (Italian Ministry of Health, RC2016 n. 2627619; to G.I.C.) supported this work.

Conflict of Interest: none declared.

References

Ildiko,E.F. (1987) Intermediate least squares regression method. *Chemometr. Intell. Lab.*, **1**, 233–242.

Jaffe,A.E. *et al.* (2015) Practical impacts of genomic data "cleaning" on biological discovery using surrogate variable analysis. *BMC Bioinformatics*, **16**, 372.

Leek,J.T. *et al*. (2012) The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*, **28**, 882–883.

Libbrecht,M.N., and Noble,W.S. (2015) Machine learning applications in genetics and genomics. *Nat. Rev. Genet*., **16**, 321–332.

Littlestone,N., and Warmuth,M.K. (1994) The weighted majority algorithm. *Inform. Comput*., **108**, 212–261.

Love,M.I. *et al*. (2014) Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome Biol*., **15**, 1.

Morgan,M. *et al*. (2016) SummarizedExperiment: SummarizedExperiment container. *Bioconductor*, R package version 1.4.0. DOI: 10.18129/B9.bioc.SummarizedExperiment.

Qiu,X. *et al*. (2005) Correlation between gene expression levels and limitations of the empirical bayes methodology for finding differentially expressed genes. *Stat. Appl. Genet. Mol. Biol*., **4**, Article34.

Ritchie,M.E. *et al*. (2015) Limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic Acids Res*., **43**, e47.

Robnik-Šikonja,M., and Kononenko,I. (1997) An adaptation of relief for attribute estimation in regression. In *Machine Learning: Proceedings of the Fourteenth International Conference* (ICML'97), Morgan Kaufmann Publishers Inc., pp.296–304.

Rokach,L. (2010) Ensemble-based classifiers. *Artif. Intell. Rev*., **33**, 1–39.

Saeys,Y. *et al*. (2007) A review of feature selection techniques in bioinformatics. *Bioinformatics*, **23**, 2507–2517.