


Data and text mining

# MOSS: multi-omic integration with sparse value decomposition

Agustin Gonzalez-Reymundez <sup>1,\*</sup>, Alexander Grueneberg<sup>1</sup>, Guanqi Lu<sup>1</sup>,  
Filipe Couto Alves<sup>1</sup>, Gonzalo Rincon<sup>2</sup> and Ana I. Vazquez<sup>1</sup>

<sup>1</sup>Department of Epidemiology and Biostatistics, Michigan State University, East Lansing, MI 48824, USA and <sup>2</sup>Genus PLC Inc., Genome Sciences R&D, De Forest, WI 53532, USA

\*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on November 18, 2021; revised on March 7, 2022; editorial decision on March 19, 2022; accepted on March 23, 2022

## Abstract

**Summary:** This article presents multi-omic integration with sparse value decomposition (MOSS), a free and open-source R package for integration and feature selection in multiple large omics datasets. This package is computationally efficient and offers biological insight through capabilities, such as cluster analysis and identification of informative omic features.

**Availability and implementation:** <https://CRAN.R-project.org/package=MOSS>.

**Contact:** [agugonrey@gmail.com](mailto:agugonrey@gmail.com)

**Supplementary information:** [Supplementary information](https://github.com/agugonrey/GonzalezReymundez2021) can be found at <https://github.com/agugonrey/GonzalezReymundez2021>.

## 1 Introduction

Omic data are characterized by many features from multiple layers of data (e.g. genome, transcriptome and proteome). Thus, traditional methods (e.g. ordinary least squares) are insufficient to obtain significant insights from this multi-layer, high-dimensional data. To effectively integrate multi-omic data, novel methods have been developed (González-Reymundez *et al.*, 2017; Lock *et al.*, 2013; Rohart *et al.*, 2017; Shen *et al.*, 2009, 2016; Zhang *et al.*, 2016). These methods have profoundly contributed to our understanding of variation in complex traits across diverse levels of regulation (e.g. mutations in coding genes and epigenetic regulation) (Hasin *et al.*, 2017; Ritchie *et al.*, 2015).

Thanks to ongoing data collection efforts, omic data increase in the number of features and available samples. This increase in sample size provides more opportunity for inference and prediction of characteristics of interest (Müller *et al.*, 2020). However, more extensive data sizes can make computations progressively lengthier and impossible to perform in some cases (Mangul *et al.*, 2019). Moreover, extensive data sizes also compromise parallelizing complex algorithms (e.g. convolutional neural networks) (Chiroma *et al.*, 2019).

We developed ‘multi-omic integration with sparse value decomposition’ (MOSS) to handle these limitations. MOSS is a free and open-source R package that performs data integration and feature selection on large datasets. It combines the flexibility of sparse value decomposition (SVD) with parallel and in-disk computations to accommodate data sizes reaching biobank dimensions.

## 2 Implementation

The package’s primary function is called `moss`. Omic data are given to `moss` as a list where each element corresponds to a

different omic (see help pages for function `moss`). Each omic enters the function as a numeric array. The rows of each array represent samples (e.g. a subject per row) and the column of each array an omic feature (e.g. expression of a gene). The rows of the different numeric arrays on the list need to be sorted in the same order (i.e. each row belongs to the same sample across omic blocks). Integration of omic blocks occurs by appending them, column-wise, into an extended matrix. Before making the extended matrix, blocks are normalized and standardized. If missing values are present, they are imputed by the mean. The effects of potential confounders can be internally adjusted by giving `moss` a data frame, vector or matrix with covariates. When omic blocks are too big to be handled in memory, File-backed Big Matrix (FBM) (Privé *et al.*, 2018) can be passed to `moss`. For this task, the package `bigstatsr` (Privé *et al.*, 2018) must be installed. Suppose the omic blocks fit in memory but are still too large to be handled in a reasonable time. In that case, `moss` allows turning the omic blocks into FBM objects internally.

MOSS performs a sparse singular value decomposition (sSVD) on the integrated omic blocks to obtain latent dimensions as sparse factors (i.e. with zeroed out elements), representing variability across subjects and features. Sparsity is imposed via Elastic Net (Zou *et al.*, 2005) (EN) on the sSVD solutions. MOSS allows an automatic tuning of the number of elements different from zero, adapting the procedure in Shen and Huang (2008). The primary output of MOSS is a list with the results of standard (dense) and sSVD. However, a flexible set of arguments extends the output to include cluster analysis, non-linear embedding and accompanying visualizations (Supplementary Information). Further statistical and algorithmic details and a description of `moss`’ arguments, plus examples of usage, are provided in Supplementary Information.

### 3 Moss identifies informative omic features as competently as existing methods

MOSS matches the performance of current analogous methods (Fig. 1A). To illustrate this point, we compared MOSS against existing methods of omic integration and feature selection. This comparison was done in terms of the methods ability to detect informative features. The methods included iCluster (Shen *et al.*, 2009), NMF (Gaujoux and Seoighe, 2010), SNFtool (Wang *et al.*, 2014), mixOmics (Rohart *et al.*, 2017) and OmicsPLS (el Bouhaddani *et al.*, 2018). The data consisted of simulations on top of gene and protein expression profiles from breast tumors from The Cancer Genome Atlas (TCGA; Chang *et al.*, 2013) repository (see Supplementary Information) and supplied within mixOmics. In each simulation, omic features were decorrelated by randomly shuffling tumors, one feature at a time. To define informative features in each simulation, a subgroup of randomly chosen features was left intact. These features conserved the naturally occurring correlation present in the data. The two scenarios compared used 10% and 80% of the total features to define the signal. A total of 1000 random simulations were run by scenario. Figure 1A shows MOSS's ranking

amongst the best performance methods. When using strict variable selection (EN parameter equal to 1), MOSS's performance is inversely related to the number of informative features. In scenarios with a larger number of informative features, methods like NMF, more suitable for dense solutions, are more sensible. However, MOSS can compensate for the loss in sensitivity by compromising variables selection in favor of shrinkage (e.g. by setting EN parameter to values between 0 and 1).

### 4 Moss requires less computational time than existing methods and scales to datasets reaching biobank sizes

One of MOSS's essential capabilities is the handling of big data. While other tools demonstrate similar analytical performance (Fig. 1A), MOSS is specifically designed for big data. As a result, even when regular R matrices are used (i.e. omic data handled in RAM), MOSS can still perform in a short amount of time compared to other omic integration and feature selection methods (Fig. 1B).

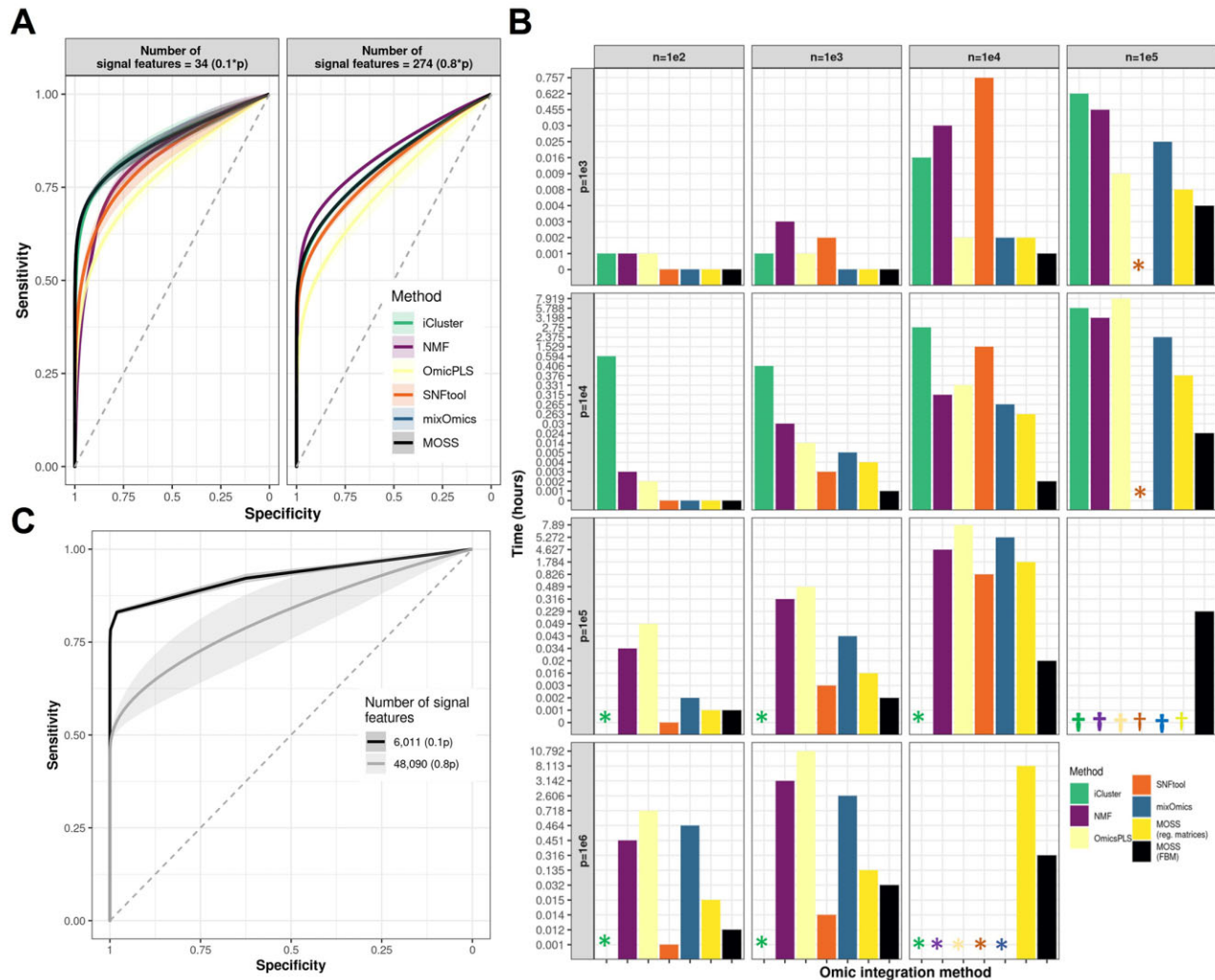


Fig. 1. (A) Performance of MOSS and existing omic integration and features selection methods. Each panel represents a different proportion of informative features. Each curve represents the average specificity and sensitivity of features selection across 1000 random simulations for increasing sparsity degrees (e.g. null effect features). Confidence bands represent inter-simulations noise. (B) Comparison of computational time between MOSS and other methods. The plot shows the computational time taken by MOSS and five other omic integration methods. Scenarios corresponded to a different combination of samples ( $n$ ) and features ( $p$ ) in simulated data. Column panels represent the number of samples, and row panels represent the number of features. Each bar represents a different omic integration method. The y-axis shows the time in hours. The symbols '\*' and '+' represent a method running for more than a day or crashing, respectively. MOSS was used with dense matrices (reg. matrices) or filed-backed big matrices (FBM). (C) Performance of MOSS on real high-dimensional data. The plot shows the performance of MOSS on simulations using data presented in (González-Reymúndez and Vázquez 2020). Different colors represent alternative proportions of features with signals

For huge datasets (e.g. scenario  $n = 1e5$  and  $p = 1e6$  in Fig. 1B), tuning of degree of sparsity with MOSS becomes prohibitive. However, dense solutions are still possible (i.e. without imposing sparsity).

## 5 Moss can be applied to high-dimensional real datasets

In González-Reymundez and Vázquez (2020), we showed that MOSS could also retrieve biologically meaningful results from real data. Figure 1C shows the results of applying the above simulation scheme to data used in González-Reymundez and Vázquez (2020), consisting of ~60 000 features from whole-genome gene expression profiles, DNA methylation and copy numbers across ~5000 tumors from 33 different cancer types.

## 6 Conclusions

Omic integration emerged as a group of techniques to collectively analyze multiple omic data layers and retrieve helpful information of shared biological processes (Hasin et al., 2017). However, the computational and statistical tools used to carry out these tasks are constantly challenged by the vast amount of data generated (Conesa and Beck, 2019; Gomez-Cabrero et al., 2014). As a result, omic integration can become a vast and challenging problem. Consequently, existing algorithms can become painfully slow or impossible to run.

As a features selection tool, MOSS performance is best as the number of signal features decreases (e.g. some signaling pathways affected in cancer, such as canonical MAPK pathway; Braicu et al., 2019). However, lower performance for a larger number of signal features is an unsolved challenge among omic integration and feature selection methods (Tini et al., 2017). In MOSS, this performance could be increased by compromising variable selection in favor of shrinking by varying the value of the EN parameter. For instance, in González-Reymundez and Vázquez (2020), a EN parameter value of 0.5 was used to show MOSS's ability to detect clusters of tumors beyond original diagnoses and molecular signatures of potential therapeutic use. The training of this additional parameter, however, can drastically increase computational time, particularly for large datasets. More sophisticated alternatives might involve the use of different penalties by omic block or set of features, a capability that we are considering for future versions of MOSS.

Despite its benefits as a data integration and mining tool, MOSS lacks statistical inference to support feature selection. Future versions of MOSS can deal with these limitations by adopting fast bootstrap techniques applied to high-dimensional SVD (Fisher et al., 2016). In addition to unsupervised analysis, MOSS can fit supervised analyses via partial least squares, linear discriminant analysis and low-rank regressions. Nevertheless, these options are currently limited by the lack of cross-validation schemes to evaluate supervised models and address their performance.

In sum, MOSS is a flexible and fast tool to perform data integration. It shares capabilities with popular methods, including estimation of latent data dimensions, feature selection and convenient graphical displays. Nevertheless, unlike these methods, MOSS integrates datasets too large to be handled in RAM and requires considerably shorter amounts of time.

## Acknowledgements

The authors acknowledge funding from the Research Alliance Interests grants provided by Zoetis. Results shown here are in part based upon data generated by the TCGA Research Network: <https://www.cancer.gov/tcga>.

## Funding

This work was supported by the Research Alliance Interests grants provided by Zoetis.

*Conflict of Interest:* none declared.

## Data availability

The data underlying this article are available in “Mendeley Data” at <https://data.mendeley.com/datasets/r8p67nfjc8/1>, and in package “MixOmics” at <http://mixomics.org/wp-content/uploads/2016/08/TCGA.normalised.mixDIABLO.RData.zip>.

## References

- Braicu, C. et al. (2019) A comprehensive review on MAPK: a promising therapeutic target in cancer. *Cancers (Basel)*, **11**, 1618.
- Chang, K. et al. (2013) The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.*, **45**, 1113–1120.
- Chiroma, H. et al. (2019) Progress on artificial neural networks for big data analytics: a survey. *IEEE Access*, **7**, 70535–70551.
- Conesa, A. and Beck, S. (2019) Making multi-omics data accessible to researchers. *Sci. Data*, **6**, 1–4.
- el Bouhaddani, S. et al. (2018) Integrating omics datasets with the OmicsPLS package. *BMC Bioinformatics*, **19**, 371.
- Fisher, A. et al. (2016) Fast, exact bootstrap principal component analysis for  $p > 1$  million. *J. Am. Stat. Assoc.*, **111**, 846–860.
- Gaujoux, R. and Seoighe, C. (2010) A flexible R package for nonnegative matrix factorization. *BMC Bioinformatics*, **11**, 367.
- Gomez-Cabrero, D. et al. (2014) Data integration in the era of omics: current and future challenges. *BMC Syst. Biol.*, **8 Suppl 2**, I1.
- González-Reymundez, A. et al. (2017) Prediction of years of life after diagnosis of breast cancer using omics and omic-by-treatment interactions. *Eur. J. Hum. Genet.*, **25**, 538–544.
- González-Reymundez, A. and Vázquez, A.I. (2020) Multi-omic signatures identify pan-cancer classes of tumors beyond tissue of origin. *Sci. Rep.*, **10**, 8341.
- Hasin, Y. et al. (2017) Multi-omics approaches to disease. *Genome Biol.*, **18**, 1–15.
- Lock, E.F. et al. (2013) Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. *Ann. Appl. Stat.*, **7**, 523–542.
- Mangul, S. et al. (2019) Systematic benchmarking of omics computational tools. *Nat. Commun.*, **10**, 1–11.
- Müller, H. et al. (2020) Biobanks for life sciences and personalized medicine: importance of standardization, biosafety, biosecurity, and data management. *Curr. Opin. Biotechnol.*, **65**, 45–51.
- Privé, F. et al. (2018) Efficient analysis of large-scale genome-wide data with two R packages: bigstatsr and bigsnpr. *Bioinformatics*, **34**, 2781–2787.
- Ritchie, M.D. et al. (2015) Methods of integrating data to uncover genotype–phenotype interactions. *Nat. Rev. Genet.*, **16**, 85–97.
- Rohart, F. et al. (2017) mixOmics: an R package for ‘omics feature selection and multiple data integration. *PLoS Comput. Biol.*, **13**, e1005752.
- Shen, H. and Huang, J.Z. (2008) Sparse principal component analysis via regularized low rank matrix approximation. *J. Multivar. Anal.*, **99**, 1015–1034.
- Shen, R. et al. (2009) Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics*, **25**, 2906–2912.
- Tini, G. et al. (2019) Multi-omics integration – a comparison of unsupervised clustering methodologies. *Brief Bioinform.*, **20**, 1269–1279.
- Vázquez, A.I. et al. (2016) Increased proportion of variance explained and prediction accuracy of survival of breast cancer patients with use of whole-genome multiomic profiles. *Genetics*, **203**, 1425–1438.
- Wang, B. et al. (2014) Similarity network fusion for aggregating data types on a genomic scale. *Nat. Methods*, **11**, 333–337.
- Zhang, C. et al. (2016) Integration of multiple heterogeneous omics data. In: *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, pp. 564–569.
- Zou, H. et al. (2005) Regularization and variable selection via the Elastic Net. *J. R. Stat. Soc. Ser. B*, **67**, 301–320.