NCS Non-Clinical Statistics Conference
Padua, IT / 23 – 25 September 2026

TITLE: Dimensionality reduction of proteomics and transcriptomics data with omicsGMF

SPEAKER and COAUTHORS: Alexandre Segers, Cristian Castiglione, Christophe Vanderaa, Lennart Martens, Davide Risso, Lieven Clement

ABSTRACT: Technical advancements in RNA-sequencing and mass spectrometry have enabled large-scale transcriptomics and proteomics studies, even at the single cell level. However, the huge number of zero counts or missing values, as well as technical batch effects result in challenges for data analysis. This hinders a first key step in the data analysis workflow, i.e., dimensionality reduction, important for data exploration, visualization, and quality control, as well as for downstream applications such as clustering cells. Current methods that tackle these challenges simultaneously become extremely slow, while chained workflows that address each of these challenges sequentially are harder to interpret and have a performance that depends on the order of the tools used.

We here present omicsGMF [1], a Bioconductor R package that builds on sgdGMF [2], which integrates dimensionality reduction, batch correction and missing value imputation within a single framework, and can deal with all standard exponential family distributions, such as Gaussian, Poisson or negative binomial distributions. It uses a stochastic gradient descent for generalized matrix factorization to scale well to huge datasets. In this contribution, we show how omicsGMF can be used for scalable dimensionality reduction and visualization of both RNA-sequencing and proteomics data, while simultaneously addressing batch effect removal and missing values. We first show that omicsGMF is a scalable alternative for dimensionality reduction scRNA-seq and proteomics data, without requiring prior normalization of the data, batch effect removal or imputation of missing values. Second, omicsGMF can assist the user to select the optimal dimensionality by cross-validation, which improves downstream analysis. Furthermore, we illustrate how it can be used for imputation of missing values, resulting in superior performance than state-of-the-art imputation tools, which in turn leads to superior sensitivity and specificity in downstream differential abundance analyses. By providing an all-in-one solution for dimensionality reduction, batch correction and imputation that is highly interpretable, omicsGMF addresses a critical gap in RNAsequencing and proteomics data analysis for single cell and large-scale applications.

[1] Segers, A., Castiglione, C., Vanderaa, C., Martens, L., Risso, D., & Clement, L. (2025). omicsGMF: a multi,-tool for dimensionality reduction, batch correction and imputation applied to bulk-and single cell proteomics data. *bioRxiv.*
[2] Castiglione, C., Segers, A., Clement, L., & Risso, D. (2024). Stochastic gradient descent estimation of generalized matrix factorization models with application to single-cell RNA sequencing data. *arXiv preprint arXiv:2412.20509.*

BRIEF SPEAKER BIO: Alexandre Segers obtained his Master's in Chemical Engineering in 2019, and his PhD in Statistical Data Analysis in November 2025, working towards computationally efficient dimensionality reduction for transcriptomics and proteomics data. Now, he works as a Senior Scientist in Manufacturing Statistics for Johnson & Johnson. He will here discuss his work on fast dimensionality reduction and imputation of missing values for proteomics data, a research topic he worked on during his PhD.