mixMint: A multivariate integrative approach to identify a reproducible biomarker signature across multiple experiments and platforms

F. Rohart[a], **A. Eslami**[b], **S. Bougeard**[c], **C. Wells**[d] and **K-A. Lê Cao**[a]

[a]The University of Queensland Diamantina Institute
Translational Research Institute, QLD 4102, Australia
f.rohart@uq.edu.au, k.lecao@uq.edu.au

[b]University of British Columbia, Canada
aida_eslami61@yahoo.com

[c] Department of Epidemiology
Agence nationale de sécurité sanitaire de l'alimentation, de l'environnement et du travail
22440 Ploufragan, France
Stephanie.BOUGEARD@anses.fr

[d]Department of Anatomy and Neuroscience, the University of Melbourne
30 Royal Parade Parkville, VIC 3010, Australia
wells.c@unimelb.edu.au

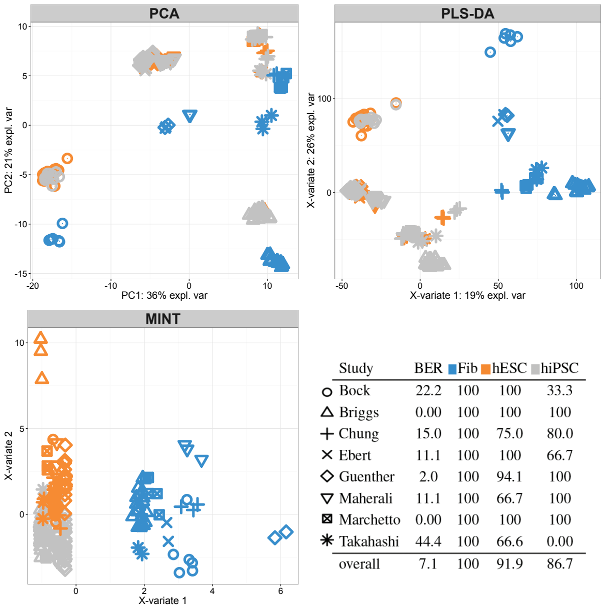
Mots clefs : Statistique, Biologie, Intégration, Multivariate.

With the advent of high-throughput technologies, thousands of transcriptomics studies are now made available through public repositories such as GEO or ArrayExpress, enabling to share data and results amongst the research community. However, biomarker signatures reported from those studies are often not reproducible from one similar study to another, as a consequence of a small sample size per experiment (i.e. in the range of 5-60) compared to several thousands of genes that are measured. Combining raw data from independent experiments, also known as integrative analysis gives the potential to increase sample size, statistical power and reproducibility across studies. However, the major analytical hurdle to overcome is to accommodate for disparities among studies that may use different protocols but also different technological platforms from different manufacturers. Therefore, integrative analyses suffer from the so-called 'batch effect', a.k.a cross-study, cross-platform or unwanted systematic variation [2, 3]. While we expect biological variability to be greater than technological and unwanted variability, this systematic variation must be accounted for when combining independent studies.

We introduce our novel integrative method MINT (*Multivariate INTegrative*) that combines several independent biological experiments while addressing three aims simultaneously: (1) accommodating for unwanted variability, (2) classifying samples in a supervised learning framework and (3) identifying key discriminant variables. *MINT* is based on the PLS framework and is a sparse extension of mgPLS-DA [1]. *MINT* selects a combination of variables on each PLS-component, which upgrade the classical single biomarker identification to a biomarker panel identification. We apply *MINT* to two case studies. First, we analyse a combination of 15 transcriptomics studies (8 as a learning set and 7 as an independent test set) from 5 different platforms that all include three type of human stem cells (Fibroblasts, hESC and hiPSC), resulting in more than 200 samples and $13,000$ variables in the learning set. Second, we analyse

4 cohorts of breast cancer (3 as a learning set and 1 as a test set) for a total of almost 3,000 samples and 16,000 genes in the learning set. We show that *MINT* is more accurate, more reproducible and faster than other procedures that address (1)-(3).

Figure 1 highlights the problem of common analyses when combining transcriptomics studies and displays satisfactory results for *MINT*. Our approach also provides study-specific outputs, e.g. Figure 2, which enable assessment and validation of a study against others (benchmarking, future-proofing), which can also serve as a quality control step.



	Study	BER	🟦Fib	🟧hESC	⬜hiPSC
○	Bock	22.2	100	100	33.3
△	Briggs	0.00	100	100	100
+	Chung	15.0	100	75.0	80.0
×	Ebert	11.1	100	100	66.7
◇	Guenther	2.0	100	94.1	100
▽	Maherali	11.1	100	66.7	100
⊠	Marchetto	0.00	100	100	100
✳	Takahashi	44.4	100	66.6	0.00
	overall	7.1	100	91.9	86.7

Figure 1: Stem cell study. (A) PCA and (B) PLSDA on the concatenated data. (C) *MINT* sample plot show that each cell type is well clustered, (D) *MINT* performance: BER and classification accuracy for each cell type and each study.

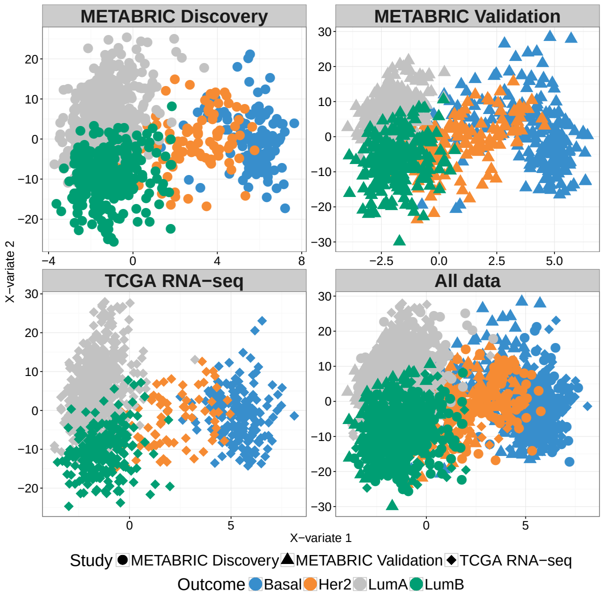


Figure 2: Breast cancer study. *MINT* study-specific sample plots for the first two components for (A) METABRIC Discovery, (B) METABRIC Validation, (C) TCGA-RNA-seq experiments and (D) overall (integrated) outputs.

MINT is currently implemented in the development version of the R-package mixOmics and will be released in the next update (v6.0.0) early May 2016. Both Figure 1 and 2 are direct outputs from mixOmics.

Références

[1] Aida Eslami, El Mostafa Qannari, Achim Kohler, and Stéphanie Bougeard. Algorithms for multi-group PLS. *J. Chemometrics*, 28(3):192–201, 2014.

[2] C. Lazar, S. Meganck, J. Taminau, D. Steenhoff, A. Coletta, C. Molter, D. Y.Weiss-Solis, R. Duque, H. Bersini, and A. Nowé. Batch effect removal methods for microarray gene expression data integration: a survey. *Brief. Bioinform*, 14(4):469–490, 2012.

[3] Kim A. Lê Cao, F. Rohart, L. McHugh, O. Korm, and Christine A. Wells. YuGene: A simple approach to scale gene expression data derived from different platforms for integrated analyses. *Genomics*, 103:239–251, 2014.