Briefings in Bioinformatics, 20(4), 2019, 1269–1279

doi: 10.1093/bib/bbx167
Advance Access Publication Date: 18 December 2017
Paper

Multi-omics integration—a comparison of unsupervised clustering methodologies

Giulia Tini, Luca Marchetti, Corrado Priami and Marie-Pier Scott-Boyer

Corresponding author: Marie-Pier Scott-Boyer, Genomics Center, Centre Hospitalier Universitaire de Québec – Université Laval Research Center, Quebec, Canada. E-mail: mariepier.scottboyer@crchudequebec.ulaval.ca

Abstract

With the recent developments in the field of multi-omics integration, the interest in factors such as data preprocessing, choice of the integration method and the number of different omics considered had increased. In this work, the impact of these factors is explored when solving the problem of sample classification, by comparing the performances of five unsupervised algorithms: Multiple Canonical Correlation Analysis, Multiple Co-Inertia Analysis, Multiple Factor Analysis, Joint and Individual Variation Explained and Similarity Network Fusion. These methods were applied to three real data sets taken from literature and several *ad hoc* simulated scenarios to discuss classification performance in different conditions of noise and signal strength across the data types. The impact of experimental design, feature selection and parameter training has been also evaluated to unravel important conditions that can affect the accuracy of the result.

Key words: molecular-level interaction; biological systems; unsupervised classification; data preprocessing

Introduction

Technological advances in high-throughput biological data generation, such as next-generation sequencing [1], mass spectrometry [2] and nuclear magnetic resonance spectroscopy [3], now allow the simultaneous collection of information from multiple molecular levels and biological systems. Usually, molecular levels (i.e. -omics) have been investigated in isolation for their association with a phenotypic trait of interest. This concept is, however, challenged by many, as it views biology linearly and does not consider the interactions between different molecular levels at the basis of the central dogma of biology [4]. Noble [5] recently proposed a multilevel causality theory with feedback cycles among biochemical layers, where interactions within and across different omics are acknowledged. The growing availability of multi-omics data and the emerging biological phenotypes originating from complex traits and interactions increased the need for adequate multi-omics integration methods [6].

Some reviews and theoretical classifications have recently defined general pipelines to combine omics data. They focused on specific data types or biological systems [7–10] and computational differences among methods [11, 12]. In this work, we will focus on statistical methods that simultaneously combine more than two different omics [13, 14], in line with the hypothesis that multiple biomolecular levels interact nonlinearly to contribute to a given phenotype [9]. We will provide a classification of those methods based on how data are handled before

Giulia Tini is a PhD student in Mathematics at the University of Trento and at COSBI.
Luca Marchetti, PhD, is a contract professor at the University of Verona and head of the computational biology team at COSBI.
Corrado Priami, PhD, has been a professor of Computer Science at the University of Trento since 2001. The results of his PhD thesis on stochastic pi-calculus were the basis for the foundation of COSBI. He has published over 190 scientific papers, given >90 invited talks and lectures, and regularly serves in related advisory, scientific and reviewing boards.
Marie-Pier Scott-Boyer, PhD, is now a research scientist at the Computational Biology Laboratory, Research center of Quebec CHU-Université Laval and was previously a project leader at COSBI in charge of research projects in systems biology and nutrition.
The Microsoft Research-University of Trento Centre for Computational and Systems Biology (COSBI) is a bioinformatics company jointly owned by Microsoft Research and the University of Trento. COSBI operates in the fields of systems pharmacology and systems nutrition with main focus to metabolic and neurodegenerative diseases. The center is an almost unique bioinformatics facility that can manage the whole systems biology pipeline starting from the data produced by wet labs and going through knowledge expansion, biomarker identification, contextual enrichment analysis and functional annotation, identification of the most active subnetworks with respect to specific experimental conditions or phenotypes and modeling of these networks and their simulation.

Submitted: 2 September 2017; **Received (in revised form):** 6 November 2017

performing integration, and we will explore the effects of factors, such as data preprocessing, number of considered omics and signal strength on resulting omics integration.

Statistical integration methods can be used to solve several types of biological questions by reinforcing common signal from different platforms (e.g. genomics and transcriptomics, miRNA and transcriptomics, transcriptomics and proteomics or proteomics and metabolomics) or by combining complementary information potentially carried by data that do not interact directly (e.g. transcriptomics and metabolomics). Multi-omics integration has been used for the discovery of molecular mechanisms [15, 16], biomarkers [17, 18] and sample/patient classification [19–21]. New methods are constantly developed to challenge these biological questions: recently, Singh *et al.* [18] have introduced DIABLO, an expansion to more than two data types of the integrOmics supervised integration method [22], which found biomarkers for three different breast cancer subtypes (Basal, Her2 and Luminal A). This article will focus on the sample classification case by comparing statistical unsupervised multi-omics integration methods that deal simultaneously with more than two data types.

Review of statistical multi-omics integration approaches

Statistical integration approaches can be classified as multivariate, concatenation-based and transformation-based methods according to how data are manipulated before applying the algorithm. Multivariate methods [7] are usually based on partial least squares (PLS) [23, 24] or canonical correlation analysis (CCA) [25, 26], and they treat different omics separately to find associations between them. We focus here on CCA-based approaches [22, 27, 28], which, differently to PLS-based methods [18, 22, 29–33], do not imply any hierarchy between data. An example of multivariate CCA-based approach is the Multiple Canonical Correlation Analysis (MCCA) [27], an extended sparse CCA [34].

Concatenation-based integration [9] is performed by combining omics data in a single matrix, used as input for low-rank-based approximation [35] or latent factor analysis [36], to combine the data into a single low-dimensional space [19, 37–43]. Lock *et al.* [37] proposed Joint and Individual Variation Explained (JIVE), a method based on the decomposition of omics data in the sum of a low-rank joint variation matrix, a low-rank individual matrix and the residual noise. This method applied to gene expression and microRNA (miRNA) from glioblastoma multiforme (GBM) samples revealed differences in GBM subtypes involving both the considered omics. Another concatenation-based method is Multiple Co-Inertia Analysis (MCIA) [39], an extension of Co-Inertia Analysis [44] to more than two data types. Following covariance optimization between the global score derived from the concatenated matrix and single omics scores, this method was applied to mRNA, miRNA and proteomics data, and succeeded in distinguishing profiles from melanoma, leukemia and central nervous system cell lines [45]. Furthermore, multiple factor analysis (MFA) [43, 46] is a concatenation-based method whose strategy is instead based on the principal component analysis (PCA) of the concatenated matrix. MFA was applied in [43] to copy-number measurements and gene expression from a glioma data set to study differences between different tumor subtypes.

Finally, the transformation-based methods integrate omics data after their transformation into an intermediate and common form, like a graph or a kernel matrix [47–52]. The main advantage of a transformation step is to preserve individual omics characteristics that can be lost otherwise [9]. For example, the Similarity Network Fusion (SNF), described by Wang *et al.* [48], creates patient similarity networks from the omics data of interest. The method recognized three GBM subtypes with different survival profiles from the integration of DNA methylation, mRNA and miRNA expression.

The methods selected for the comparison are MCCA [27], JIVE [37], MCIA [39], MFA [43] and SNF [48] (Table 1 and Supplementary Material for their description). The chosen methods are well-known unsupervised algorithms, representative of the different classes of statistical integration approaches and already considered in reviews focused on specific theoretical characteristics of the methods (unsupervised/supervised [53], use of networks [11], cluster computation [12], dimension reduction [45]). Our classification, based on how methods handle data, takes into account all these aspects by providing a direct comparison of the methodologies, which, although suggested in [53], has never been presented in literature. Moreover, these methods can be applied to different types of omics without any required previous knowledge about the phenotype of interest. Interestingly as well, these methods are all provided as R packages, making them suitable for a direct comparison inside the same computing environment. Finally, this article will also address the impact of experimental design, data preprocessing and parameter training on the multi-omics integration outcomes.

A graphical overview of the article structure is presented in Figure 1, describing the comparison pipeline, method classification, the tested data sets and result organization.

Material and methods

Real data sets

Methods were tested on three real data sets (murine liver (BXD) [54], platelet reactivity [55] and Breast Cancer (BRCA) [56] data sets),

Table 1. Summary of the multi-omics integration methods reviewed

Method	Integration approach	Description	Data scaling	R package
MCCA [27]	Multivariate	Seeks linear combination of correlated features from different data	Columns normalization (mean=0; SD = 1)	PMA
JIVE [37]	Concatenation	Separates signal common to all data from individual one	Columns normalization (mean=0; SD = 1)	r.jive
MCIA [39]	Concatenation	Projects data on a common lower dimensional space	Nonsymmetric correspondence analysis	omicade4
MFA [43]	Concatenation	Projects data on a common lower dimensional space	Columns normalization (mean=0; SD = 1)	FactoMineR
SNF [48]	Transformation	Builds a fused network from single ones	Columns normalization (mean=0; SD = 1)	SNFtool

Note: The column ‘data scaling’ indicates which scaling has been applied to data before integration.

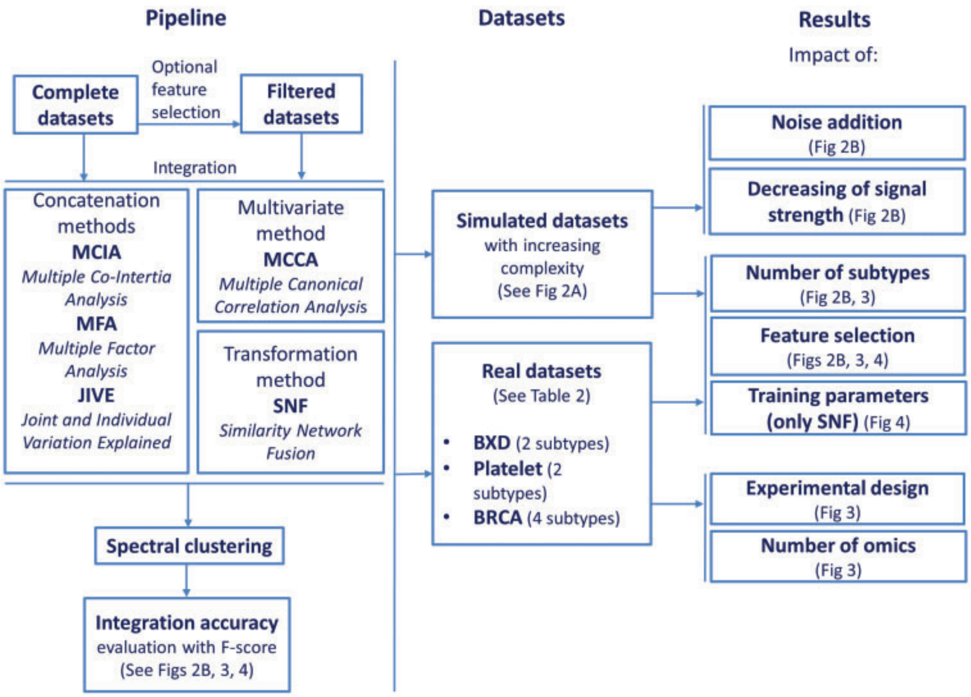
Pipeline
Datasets
Results
Impact of:
Complete datasets
Optional feature selection
Filtered datasets
Integration
Concatenation methods
MCIA
Multiple Co-Intertia Analysis
MFA
Multiple Factor Analysis
JIVE
Joint and Individual Variation Explained
Multivariate method
MCCA
Multiple Canonical Correlation Analysis
Transformation method
SNF
Similarity Network Fusion
Spectral clustering
Integration accuracy evaluation with F-score (See Figs 2B, 3, 4)
Simulated datasets with increasing complexity (See Fig 2A)
Real datasets (See Table 2)
BXD (2 subtypes)
Platelet (2 subtypes)
BRCA (4 subtypes)
Noise addition (Fig 2B)
Decreasing of signal strength (Fig 2B)
Number of subtypes (Fig 2B, 3)
Feature selection (Figs 2B, 3, 4)
Training parameters (only SNF) (Fig 4)
Experimental design (Fig 3)
Number of omics (Fig 3)

Figure 1. Graphical overview of the multi-omics integration method comparison discussed in the review. The schema is divided in three areas: (i) method classification and comparison pipeline, (ii) data sets and (iii) result organization. In the left side of the schema, the different pipeline steps are presented, together with the method classification. This is represented in the pipeline step called 'Integration', where each block collects the methods belonging to the same integration approach. Data sets are represented in the middle according to the division simulated/real data sets (BXD: murine liver data set; Platelet: platelet reactivity data set; BRCA: breast cancer data set). Finally, in the right part of the diagram, results are organized according to how they are presented and discussed in the article. Arrows linking data sets and results indicate which data set has been used to produce the corresponding result.

Table 2. Overview of the three real data sets used to compare integration methods

Data set	Phenotype	Number of subjects	Subtypes	Omics	Platform
BXD [54]	Mitochondrial metabolism	66	High fat diet (31)	Transcriptomics	Affymetrix Mouse Gene 1.0 ST microarrays
			Chow diet (35)	Proteomics	SWATH-MS quantification
				Metabolomic	MS signatures
Platelet [55]	Platelet reactivity	12	High (6)	Transcriptomics	Affymetrix GeneChip Human Genome U133 Plus 2.0 arrays
			Low (6)	Proteomics	MS quantification
				miRNA	NanoString
BRCA [56]	Breast cancer	491	Luminal A (225)	Epigenomics	Illumina Infinium
			Luminal B (120)	Transcriptomics	Agilent microarrays
			HER2 enriched (56)	miRNA	Illumina sequencing
			Basal-like (90)		

Note: Columns provide the studied phenotype, the number of subjects (total and for each subtype) and the omics data included in each data set.

each one composed by three different data types including transcriptomics, proteomics, metabolomics, miRNA and epigenomics (see details in Table 2; PCA data visualization and correlation analysis in Supplementary Figures S1–S3 and Supplementary Tables S1–S3). Some data preprocessing was performed before integration. The averaged values across the measured probes of the BXD data set were retained for each gene. Missing values for proteins and metabolites measurements were substituted by their median over all the cohorts. To obtain the proteomics data for the Platelet data set, the averaged ratio of all peptides for a given protein was considered (available for download from the Omics Discovery Index, http://www.omicsdi.org/). Then, quantile normalization was applied to reduce the batch effect (function normalizeBetweenArrays from limma R package [57]). For the BRCA data set [56], the eight subjects with a normal-like cancer subtype were excluded from the analysis because of the small number of samples. Moreover, 80 randomly selected subjects were considered for the two largest subtypes, Luminal A and B, to avoid a bias with respect to these cancer subtypes: missing values for gene expression and methylation data were then substituted with their median values across the subjects.

Simulated data sets

Several simulated data sets (Figure 2A) (composed of three matrices 60×500 (60 subjects and 500 features) were created to

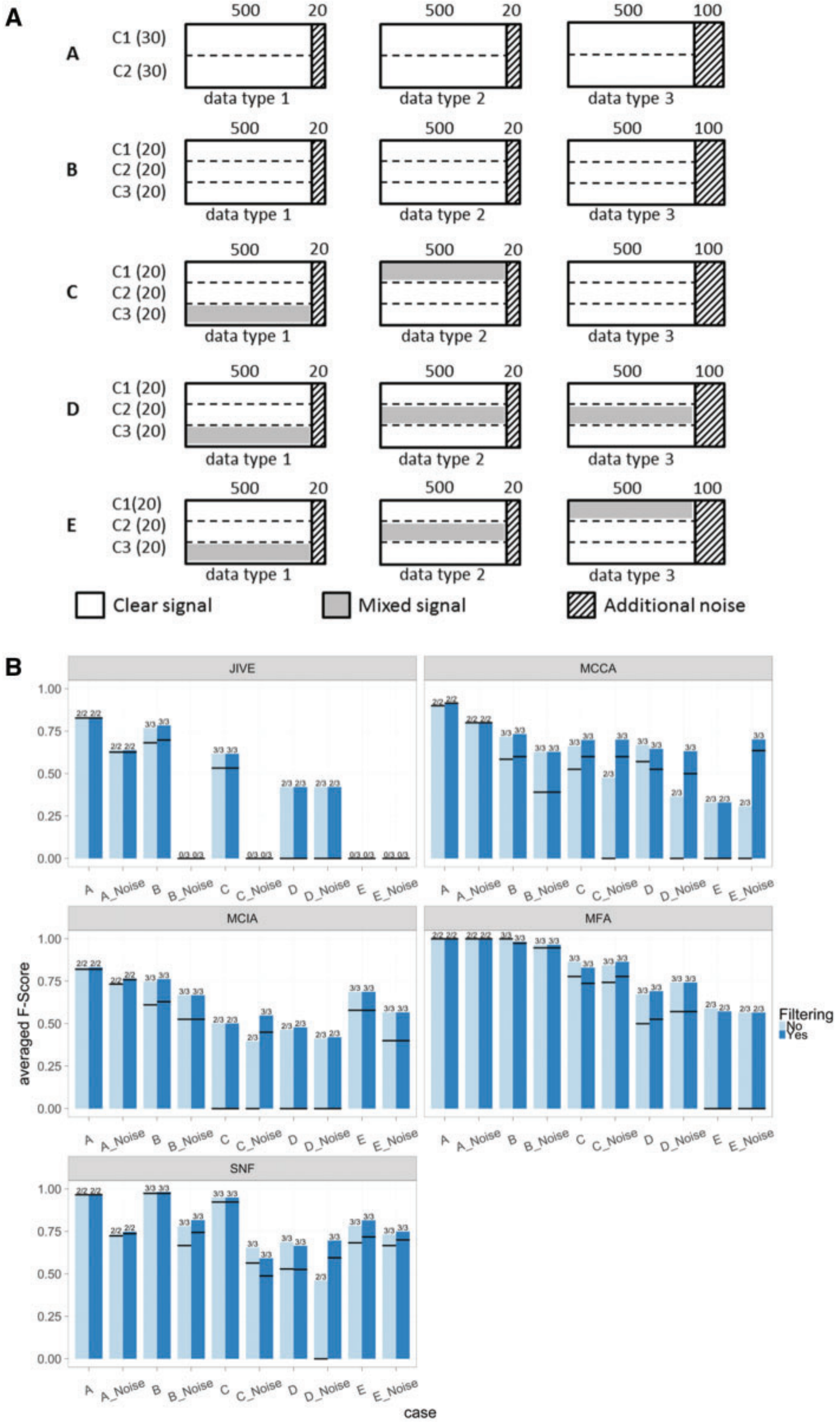

Figure 2. (**A**) Visualization of the simulated scenarios: for each of them, three data types (d1, d2 and d3) of 500 features were generated by creating 60 samples divided in two (Case A) or three groups (Cases B–E). When the group is colored in white, it has been generated to be clearly detectable in the data matrix; otherwise, it is colored in gray. 20 (d1 and d2) or 100 (d3) columns of noise were also added to the data. (**B**) Comparison of the integration methods (JIVE, MCCA, MCIA, MFA and SNF) applied to the simulated scenarios, with and without noise addition. Light- and dark-shaded bars represent the averaged *F*-scores obtained before and after the feature selection step, respectively. The solid black lines represent the minimum *F*-scores. A minimum *F*-score equal to 0 means that not all the groups have been recognized. The number of subtypes recognized for each trial is added above the bars.

evaluate the performances of the considered algorithms in a more controlled context. In each data set, sample profiles were generated following normal distributions with mean and SD derived from randomly selected gene expressions, methylation levels and miRNA from the BRCA data set [56] (details in the Supplementary Material, section 'Generation of the Simulated Datasets'). Independent noise $\sim N(0,\ 0.4)$ was also added to the data matrices during the generation. Finally, as not all the molecules in a system usually contribute to a significant signal, the considered methods were also tested after adding noisy columns to the matrices. Sample profiles on those columns were again computed as normal distributions derived from the BRCA data set features.

Samples have been generated to reproduce the following scenarios, represented in Figure 2A: (A) two groups of 30 samples clearly distinguishable in each data matrix; (B) three groups of 20 samples clearly distinguishable in each data matrix; (C) three groups of 20 samples, with only one matrix over three generated to distinguish all of them. One group is created to be detectable in all the three matrices; (D) two groups over three clearly distinguishable in each matrix, with one of them common to the three matrices. One group is created to be detectable only in one matrix; (E) two groups over three clearly distinguishable in each matrix, without a common detectable one (see Supplementary Figures S4–S8 and Supplementary Tables S4–S8 for PCA visualization and correlation analysis).

Methods

As input for statistical methods, omics data measured on a common set of *n* samples are thought as matrices with *n* rows (samples) but different number of columns (e.g. the number of genes, proteins, etc.). The details of the methods used can be found in the Supplementary Material. To check the effect of preprocessing on integration accuracy, a feature selection step was performed (subsection 'Feature selection' of Supplementary Material). The features showing a coefficient of variation (CV) [58] (a standard way to compare variability in different data types, as it is independent from the scaling) lower than a selected threshold were removed from the analysis (Table 3). Thresholds were selected for each data set accordingly to the omics-specific CV distributions as the common value indicating low variability across all the data. Three-omics and pairwise integration for all the different omics couples were then computed for each data set with the considered methods, both before and after the feature selection step (Supplementary Material, subsection 'Datasets multi-omics integration'). Once integration was completed, spectral clustering, known to outperform other clustering algorithms [59], was applied to classify samples (subsection 'Spectral clustering analysis', Supplementary Material). The *F*-score index ($Fscore = 2\frac{P*R}{P+R} \in [0,\ 1]$, with $P = \frac{\text{True positives}}{\text{True positives+false positives}}$ and $R = \frac{\text{True positives}}{\text{True positives+false negatives}}$) [60], a standard measure assessing the optimality in binary classifications [61, 62], was used to evaluate the agreement between computed clusters and real subtypes. A 0 *F*-score was assigned to subtypes for which no cluster has been identified. The minimum and the averaged *F*-scores were considered to assess the performance of the overall classification. The accuracy index has also been computed for each analysis (Supplementary Figures S9 and S10). The SNF method has also been applied after performing a training/validation procedure on its parameters, to explore the gain of this procedure on the classification performances. Here, the value of SNF parameters σ and *K* were trained to obtain the highest minimum *F*-score on 80% of the samples. The trained parameters were then validated on the remaining 20% of subjects. *F*-scores from integration performances before and after feature selection with default parameters were also computed on the validation set. Owing to the small number of subjects (12), this analysis was not applied to the Platelet reactivity data set.

Results

Statistical multi-omics integration methods have been applied to both simulated and real data sets to assess and compare their performances in sample classification. All results, resumed in Figure 1, are presented in Figure 2B, Figure 3 and Figure 4 where classification results have been indicated in terms of *F*-score results based on the accuracy index, and *F*-scores tables are provided in the Supplementary Material (Supplementary Figures S9 and S10 and Supplementary Tables S9–S16).

Influence of signal strength, noise and feature selection in the simulated scenarios

Influence of signal strength

We observed a general decrease in classification performance in all the simulated scenarios when the signal strength across the data types diminished (Figure 2B, from A to E, light-shaded bars without noise addition). As expected, all methods obtained the highest classification accuracy in Scenario A (easiest situation, see Figure 2A and Supplementary Figure S4), with averaged *F*-scores ranging from a minimum of 0.833 to the best value of 1 obtained by MFA. Classification performances decreased step by step in Scenarios from B to E, where only SNF was able to distinguish all the sample groups in all the considered scenarios (see Figure 2B: only SNF has minimum *F*-score (solid black lines) >0). The method with the worst performance resulted to be JIVE, especially in Scenario E where no clear signal was common to

Table 3. Number of features for each data set before (complete) and after (filtered) feature selection step, with the corresponding threshold on the CV used to filter the data

Data set	Omics	Complete (number of features)	CV threshold	Filtered (number of features)
BXD	Transcriptomic	21 836	0.015	17 036
	Proteomic	976		976
	Metabolomic	2607		2607
Platelet	Transcriptomic	54 675	0.02	39 888
	Proteomic	663		661
	miRNA	490		407
BRCA	Epigenomic	14 443	0.2	12 474
	Transcriptomic	17 814		16 419
	miRNA	1010		942

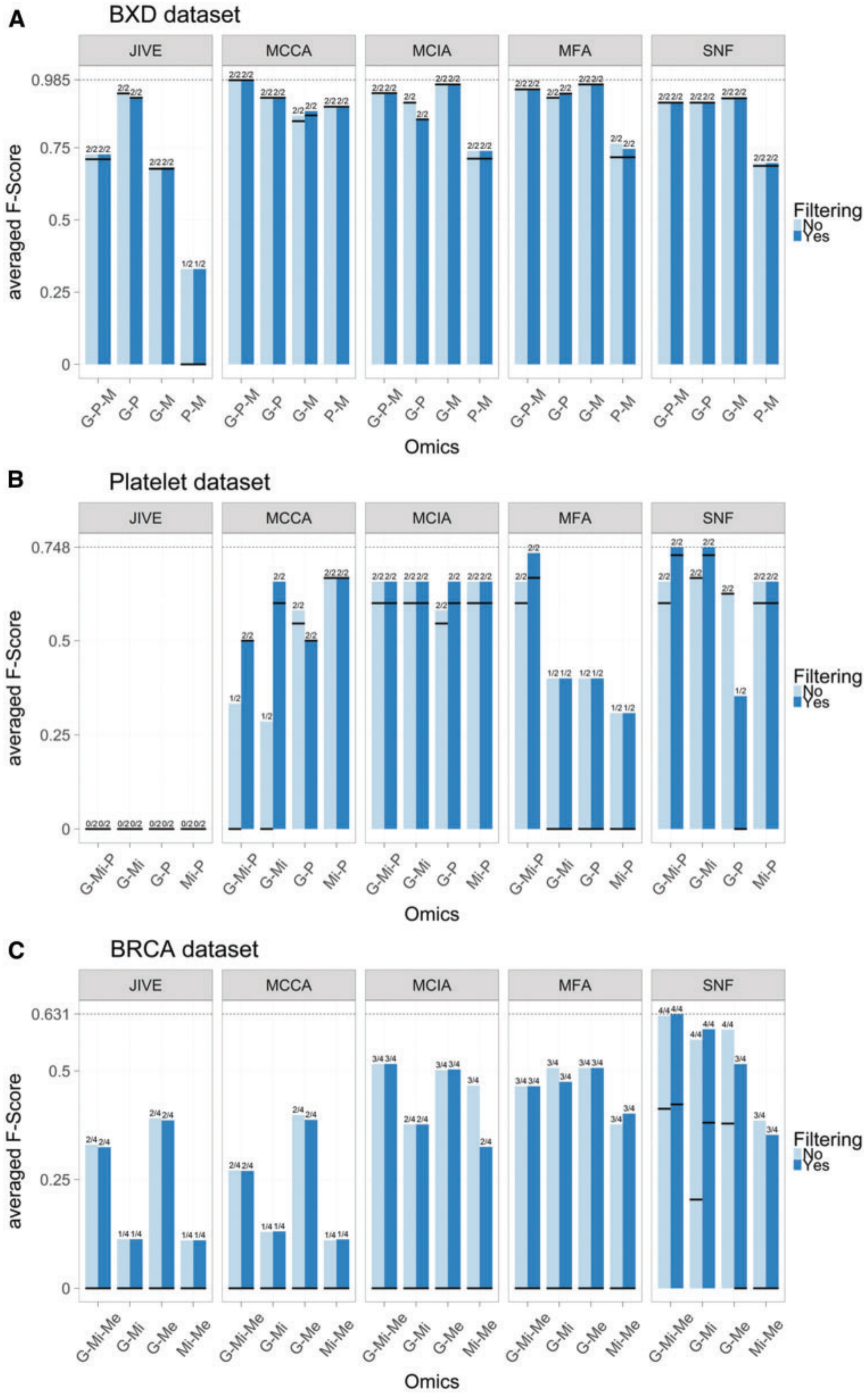

Figure 3. Comparison of the different integration methods (JIVE, MCCA, MCIA, MFA and SNF) applied to the real data sets on all the possible omics combinations (provided in the *x*-axis). The light- and dark-shaded bars represent the averaged *F*-scores obtained before and after feature selection, respectively. For each method, the first two bars represent the results from three-omics integration. The thick black lines represent the minimum *F*-score obtained for each trial: a minimum value equal to 0 means that not all the subtypes have been recognized. The number of subtypes recognized for each trial is added above the bars. The horizontal dashed lines give the highest *F*-scores reached for the data set. (**A**) BXD data set (G: gene expression, P: proteins, M: metabolites. (**B**) Platelet data set (G: gene expression, Mi: miRNA, P: proteins). (**C**) BRCA data set (G: gene expression, Mi: miRNA, Me: methylation).

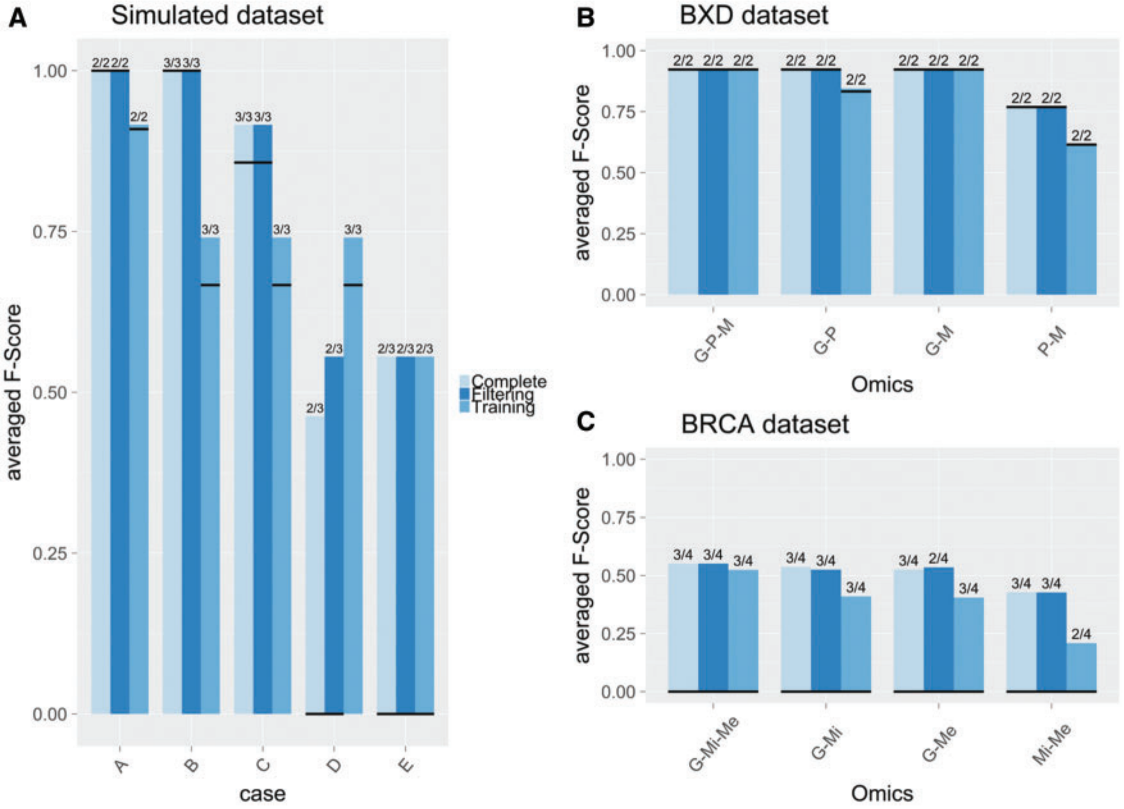

Figure 4. Comparison of SNF results on the validation sets by using default parameters before and after feature selection and by using trained parameters without feature selection. Averaged *F*-scores obtained from the three analyses are represented with light-, dark- and medium-shaded bars, respectively. Minimum *F*-scores are represented with black lines. A minimum *F*-score equal to 0 indicates that not all the subtypes have been recognized. The number of subtypes recognized for each trial is added above the bars. (**A**) Simulated scenarios. (**B**) BXD data set (G: gene expression, P: proteins, M: metabolites). (**C**) BRCA data set (G: gene expression, Mi: miRNA, Me: methylation).

the three data matrices, and therefore, no joint pattern was found by the method.

Influence of noise

All methods exhibited a general decrease in performance when noise has been added to the data sets without applying a feature selection step (Figure 2B, from A to E, light-shaded bars with noise addition). MFA was the method less affected by noise in the simpler scenarios, however, in the most complex case (Scenario E), only SNF was still able to distinguish all the sample groups. JIVE resulted to be the method most affected by noise because of its inability to detect common signal in Scenarios B, C and E. As discussed in [37], noise can overwhelm the low-rank signal, affecting the permutation testing approach used by JIVE.

Influence of feature selection

To understand the impact of data preprocessing on multi-omics integration, a preliminary feature selection step was also applied to the simulated scenarios with and without noise addition (Figure 2B, from A to E, dark-shaded bars). After feature selection, 25 of the 50 considered trials did not change *F*-score; 19 improved and 6 diminished. Feature selection did not improve the accuracy for JIVE, in line with the method description: its strategy is natively able to separate residual noise in an additional matrix without influencing the joint signal. Conversely, performances of MCCA were the most positively affected by feature selection, with 6 improved trials over 10. This result is in line with the paper by Witten and Tibshirani [27], where a fused lasso penalty has been used to reduce samples noise before applying MCCA. Although not all the performances benefited from feature selection, the classification accuracy lost by noise addition has been generally recovered by applying this preprocessing step.

Three-omics integration versus pairwise integration in real data sets

Methods comparison was repeated on the real data sets described in Table 2: the BXD (Figure 3A), the Platelet (Figure 3B) and the BRCA (Figure 3C) data sets. The best classifications were obtained by three-omics integration in all the considered data sets even if this result was obtained by different methods (MCCA for BXD, SNF for Platelet and BRCA). This highlights the importance of considering additional omics when possible. Interestingly, we also observed a sort of general agreement on the omics couple more difficult to integrate: proteins and metabolites for the BXD data set; gene expression and proteins for the Platelet; miRNA and methylation for the BRCA.

BXD data set

Three-omics integration allowed a good separation of mice with different diets (Figure 3A) with an averaged *F*-score ranging from 0.727 to 0.985, value obtained by MCCA. JIVE obtained the best classification result by considering the omics couple genes–proteins (F-score = 0.939) while SNF, MCIA and MFA by considering genes–metabolites (*F*-scores of 0.925, 0.97 and 0.97, respectively). The omics couple proteins–metabolites was the most difficult to integrate for all the approaches except for MCCA. This result could be related to the fact that some cross-dimensional correlations have been observed between metabolites and adjacent enzymes in known metabolic pathways (e.g. tricarboxylic acid cycle) [54].

Platelet data set

SNF was the method performing better (Figure 3B), reaching the same *F*-score of 0.748 both with three-omics integration and with the omics couple genes–miRNAs. SNF obtained the worst performance with the omics couple genes–proteins. This could be explained by the weak Spearman correlation observed between the platelet transcriptome and proteome [63]. MFA obtained the highest *F*-score of 0.657 for three-omics integration. Except for genes–proteins integration, MCIA always obtained the same *F*-score of 0.657, while MCCA reached the highest *F*-score by integrating miRNAs and proteins (*F*-score=0.667). JIVE could not recognize common signal in any of the tested omics sets. This indicates that the amount of signal dividing the two extreme phenotypes across the different data types is not strong, in line with what observed by Zufferey *et al.* [55].

BRCA data set

Differently to previous cases, where a binary classification was required, here samples have been classified in four clinical subtypes: Luminal A, Luminal B, Basal-like and HER2-enriched (Figure 3C). As expected, the obtained *F*-scores were generally lower than those of the other studies, confirming the increased level of uncertainty with respect to binary classification. As for the Platelet data set, SNF provided the highest performance (best result with three-omics integration, *F*-score = 0.631). SNF was also the only method able to recognize all the four clinical subtypes. The classification accuracy of MCIA, the second-best method after SNF, was also highest in three-omics integration (averaged *F*-score of 0.516), but the method failed to recognize the HER2-enriched subtype, which resulted to be the most difficult subtype to recognize, in line with [56]. Also MFA recognized three subtypes (HER2-enriched excluded), but with lower *F*-scores: the highest (0.507) was obtained for genes–miRNA and genes–methylation integration. The latter case provided the highest, but indeed low, *F*-scores for JIVE and MCCA, which could distinguish only Basal-like and Luminal A. The poor result obtained with JIVE could be motivated by the fact that, as the HER2-enriched subtype does not provide a signal common to the three omics, the method could not recognize a shared pattern. This is also in agreement with [64], where JIVE applied on mRNA, methylation and miRNA from another breast cancer data set, separated Basal-like and Luminal A samples from the others. Basal-like and Luminal A subtypes were recognized by all the methods, especially with three-omics integration. This agrees with the literature, as the Basal-like subtype is known to be clearly separated from the Luminal one [56].

Feature selection

The effect of feature selection on the real data sets was also evaluated (Figure 3A–C, dark-shaded bars). A different threshold for the CV was selected for each data set (Table 3), to reduce data dimensions without losing too much signal. Three-omics integration performances were not diminished by feature selection (BXD data set because of the mild filtering, which reduced only transcriptomic features), and in some cases were improved by it, as for Platelet and BRCA data sets (Figure 3B and C). In the latter case, although, *F*-scores were only slightly improved.

Influence of parameter training

All classification results presented so far have been computed by applying the reviewed methods with default parameters. Here, we investigate the gain in classification accuracy obtained by training parameters according to a training/validation procedure (see 'Methods' section and Supplementary Material for details). In this analysis, classification results have been computed by SNF, the method that performed better, on average, in all previous results. The training procedure was applied to all the simulated and real data sets (Figure 4), with the exception of the Platelet data set, where the limited number of samples prevents the reliability of the analysis. According to the literature, all classification results refer to the sample subset devoted to validation (20% of the data set samples). This is an important aspect to consider because in all the scenarios, the classification performances obtained by training parameters outperformed those obtained by using default parameters when considering also samples included in the training set (data not shown). This result, however, has been rarely confirmed in the validation set. Indeed, the training/validation procedure demonstrated an obvious advantage with respect to the standard unsupervised procedure only in simulated Scenario D.

Simulated data sets

Integration with default parameters outperformed training/validation in Cases A, B and C (Figure 4A). An effective gain in training parameters was observed in Scenario D, emphasizing the advantage of training parameters when the signal in the data set becomes weak (see Supplementary Figure S7). Such a result has not been confirmed in Scenario E, but this could be motivated by the high complexity of classifying samples when a clear common signal is missing between the three data matrices (see Figure 2A and Supplementary Figure S8).

Real data sets

Integration with default parameters outperformed training/validation in all cases (Figure 4B and C). On the validation set of the BXD data set (seven Chow diet (CD) and six High Fat diet (HFD) samples), three-omics and genes–metabolites integration after parameter training ($K = 12$, $\sigma = 0.59$ and $K = 7$, $\sigma = 0.8$, respectively) reached the same averaged *F*-score (0.923) of integration with default parameters (Figure 4B). Training parameters for the other omics couples resulted in lower accuracies. On the validation set of the BRCA data set (11 HER2-enriched, 45 Luminal A, 24 Luminal B and 18 Basal-like samples), training parameters never improved integration results (Figure 4C), but this analysis could be influenced by the different number of samples of the clinical subtypes, which can affect the estimation of some parameters of the method.

Influence of multiclass classification and experimental design

The results presented so far indicate that data characteristics, such as the number of sample subtypes and the experimental design, could influence multi-omics integration performances.

Multiclass classification

When samples belong to two subtypes (simulated Scenario A, BXD and Platelet data sets), all the methods, excluded JIVE for Platelet, identified the two sample groups. This confirms the relative simplicity of recovering information when the signal is given by two phenotypes and strong across all the data. Similarly, the two different diets in the BXD data set induced a strong signal when single omics were individually assessed for significance. Conversely, multi-omics integration of data sets with multiple subtypes (BRCA data set and simulated Scenarios from B to E) resulted more challenging, with generally low *F*-scores.

Experimental design

The high classification performances observed for the BXD samples can also be explained in terms of experimental design. Transcriptomics, proteomics and metabolomics were assessed on the same mice livers, with a split-sample study, which Cavil *et al.* [7] suggested being the best experimental design for multi-omics integration. Also in the Platelet data set, the omics couple providing the best classification performance was genes–miRNAs, and this could be motivated by the fact that both gene expression and miRNAs were assessed from the same RNA. This provides an important advantage with respect to the way in which proteomics data have been obtained: proteins were quantified with different preparations and three technical replicates for each patient. Moreover, they were separated in two groups, thus presenting a batch effect on sampling timing that needed to be corrected, and which could have negatively influenced proteomic integration with the other omics.

Discussion

Five multi-omics integration methods, representative of multivariate, concatenation-based and transformation-based approaches, were selected for comparison of their ability to integrate more than two omics data in unsupervised way.

In general, our analysis showed that the integration of three different omics results in better sample classification than pairwise omics integration. This demonstrates that the additional knowledge brought by considering multiple omics data at once is essential to increase the understanding of the mechanisms underlying the characteristics of sample subtypes. Furthermore, *F*-scores obtained with SNF, the transformation-based method, were the highest in 9 of the 22 trials considered and among the highest in the other cases. MFA performed the best in six trials (simulated data sets), MCIA and MCCA in three and JIVE only in one. Additionally, by also considering the accuracy index and the single-class *F*-scores (Supplementary Figures S9 and S10, Supplementary Tables S9–S16), SNF demonstrated to be the best method when the data set complexity increases.

In addition, the comparisons revealed that the outcome of multi-omics integration is data-dependent and influenced by the experimental design as suggested by Cavil *et al.* [7]. This could thus warrant some preliminary examination of the data at hand to determine the appropriate integration method to use. Omics data should be separately analyzed and visualized (e.g. with PCA) to quantify how much signal is carried by each omics and how much of it is shared across omics types. Recently, Ciucci *et al.* [65] proposed an algorithm able to detect the optimal normalization method to be applied and the most discriminative dimensions. In cases where PCA is not powerful enough to segregate samples, more advanced techniques (such as Minimum Curvilinear Embedding [66, 67]) based on nonlinear dimension reduction can be tested to inspect each omics data type at a time [68]. Computing correlations between the omics could also help in assessing and quantifying inter-omics relationships.

If preliminary analysis reveals shared signals across data (simulated Scenario A), a method like JIVE able to separate noise and to provide the common pattern in an already computed matrix could be a good choice. The multivariate method MCCA could instead reinforce visible intra-omics signal, when no evident inter-omics signal (necessary for JIVE) is present (simulated Scenarios C and D). As MCCA is correlation-based, it could also be applied when data sets show well correlated features across omics: it obtained the most precise sample classification for the BXD data set, where 25% of transcript–proteins pairs correlated significantly in the CD subtype (P-value < 0.05), 137 of those with Spearman's $\rho = |0.65|$ ([54]).

For more complex cases (simulated Scenario E: multiple subtypes and noisy data set), SNF, MCIA and MFA, methods based on subjects' similarities and dissimilarities, can be better options. Indeed, these methods recognized all the subtypes demonstrating their ability to recover not only shared but also complementary signals across omics. Moreover, the data transformation step applied by SNF succeeded in distinguishing all the tumor subtypes of the BRCA data set, including the HER2-enriched (weakest signal), while MCIA and MFA could distinguish three of the four subtypes.

Additionally, feature selection is a popular preprocessing step and, according to our analysis, it can be useful to integrate omics not showing a strongly shared signal. However, the thresholds for feature selection need to be carefully selected. Features can carry signals not detected in single omics analysis but that can make the difference when more omics are integrated. To define a unique threshold of low variability across data types, we used a general method to filter out noise. This can be substituted by more specific methods considering the data and the problem at hand (*e.g.* supervised/unsupervised) or whether the relationships among features should be considered while filtering (see [69] for a review of feature selection methods).

Some limitations of our work must be acknowledged. First, methods were tested only in situations known to potentially affect results of multi-omics integration such as when considering omics that bring reinforced or complementary signal, with increasing numbers of subtypes when noise is present. Second, we considered only Gaussian distributed data: integrating more heterogeneous data (e.g. microbiome) would require the study of more specific methods. Additionally, the type of noise we generated in the simulated scenarios and the feature selection approach that was applied could have influenced observations. The influence on omics integration of other types of noise and filtering methods should be tested in the future. Finally, we investigated and compared results related to only one of the possible applications of multi-omics integration (i.e. sample classification) because of the possibility to clearly assess classification results against real subtypes. One needs to keep in mind that the methodologies tested here can be applied to solve other questions with potentially different outcomes.

Conclusions

The addition of biological knowledge obtained by considering multiple molecular levels (omics) to the analysis increases the knowledge extracted from the available data, in the present case, sample classification precision. Simultaneous omics integration should thus be considered in future studies with more omics data available. Noise was also shown to influence integration results, an effect that can be mitigated by adding a feature selection step before proceeding with data integration. This is especially recommended when dealing with complex design (such as those having more than two different omics data, or with low signal strength, or multiple cellular subtypes). However, we believe that statistical integration methods could still be improved, for example by adding a priori information about relationships between the different omics data, which could diminish false-positive results, while enhancing the relevance of true molecular interactions.

Key Points

- Multi-omics integration is sensitive to noise increasing and signal strength decreasing across the omics data.
- Combining more omics data increases the integration precision. However, this process can add noise.
- A complex biological problem (many subtypes, many omics data, low signal) can benefit from a feature selection step.
- SNF resulted to be the most robust method when the complexity of the data set increases.

Supplementary Data

Supplementary data are available online at https://academic.oup.com/bib.

Acknowledgement

The authors would like to thank Federico Reali and Marie-Laure Martin-Magniette for their comments.

Funding

This work was partially supported by the PAT (Provincia Autonoma di Trento) project S503/2015/664730 Zoomer.

References

1. van Vliet AHM. Next generation sequencing of microbial transcriptomes: challenges and opportunities. *FEMS Microbiol Lett* 2010;**302**(1):1–7.
2. Edwards JR, Ruparel H, Ju J. Mass-spectrometry DNA sequencing. *Mutat Res Mol Mech Mutagen* 2005;**573**(1–2):3–12.
3. Fürtig B, Richter C, Wöhnert J, *et al.* NMR spectroscopy of RNA. *Chembiochem* 2003;**4**(10):936–62.
4. Crick F. Central dogma of molecular biology. *Nature* 1970; **227**(5258):561–3.
5. Noble D. A theory of biological relativity: no privileged level of causation. *Interface Focus* 2012;**2**(1):55–64.
6. Nardini C, Dent J, Tieri P. Editorial: multi-omic data integration. *Front Cell Dev Biol* 2015;**3**:46.
7. Cavill R, Jennen D, Kleinjans J, *et al.* Transcriptomic and metabolomic data integration. *Brief Bioinform* 2016;**17**(5):891–901.
8. Moyon T, Le Marec F, Qannari EM, *et al.* Statistical strategies for relating metabolomics and proteomics data: a real case study in nutrition research area. *Metabolomics* 2012;**8**(6):1090–101.
9. Ritchie MD, Holzinger ER, Li R, *et al.* Methods of integrating data to uncover genotype–phenotype interactions. *Nat Rev Genet* 2015;**16**(2):85–97.
10. Castellani GC, Menichetti G, Garagnani P, *et al.* Systems medicine of inflammaging. *Brief Bioinform* 2016;**17**(3):527–40.
11. Bersanelli M, Mosca E, Remondini D, *et al.* Methods for the integration of multi-omics data: mathematical aspects. *BMC Bioinformatics* 2016;**17**(S2):S15.
12. Wang D, Gu J. Integrative clustering methods of multi-omics data for molecule-based cancer classifications. *Quant Biol* 2016;**4**(1):58–67.
13. Gomez-Cabrero D, Abugessaisa I, Maier D, *et al.* Data integration in the era of omics: current and future challenges. *BMC Syst Biol* 2014;**8(Suppl 2)**:I1.
14. Ebbels TMD, Cavill R. Bioinformatic methods in NMR-based metabolic profiling. *Prog Nucl Magn Reson Spectrosc* 2009;**55**(4): 361–74.
15. Glass K, Huttenhower C, Quackenbush J, *et al.* Passing messages between biological networks to refine predicted interactions. *PLoS One* 2013;**8**(5):e64832.
16. Wang L, Xiao Y, Ping Y, *et al.* Integrating multi-omics for uncovering the architecture of cross-talking pathways in breast cancer. *PLoS One* 2014;**9**:e104282.
17. Wahl S, Vogt S, Stückler F, *et al.* Multi-omic signature of body weight change: results from a population-based cohort study. *BMC Med* 2015;**13**:48.
18. Singh A, Gautier B, Shannon CP, *et al.* DIABLO—an integrative, multi-omics, multivariate method for multi-group classification. *bioRxiv* 2016:67611.
19. Meng C, Helm D, Frejno M, *et al.* moCluster: identifying joint patterns across multiple omics data sets. *J Proteome Res* 2016; **15**(3):755–65.
20. Mo Q, Wang S, Seshan VE, *et al.* Pattern discovery and cancer gene identification in integrated cancer genomic data. *Proc Natl Acad Sci USA* 2013;**110**(11):4245–50.
21. Shen R, Mo Q, Schultz N, *et al.* Integrative subtype discovery in glioblastoma using iCluster. *PLoS One* 2012;**7**(4):e35236.
22. Le Cao KA, Gonzalez I, Dejean S. integrOmics: an R package to unravel relationships between two omics datasets. *Bioinformatics* 2009;**25**(21):2855–6.
23. Wold H. Estimation of principal components and related models by iterative least squares. In: Krishnajah PR (ed), *Multivariate Analysis*. New York: Academic Press, 1966, 391–420.
24. Le Cao KA, Rossow D, Robert-Granié C, *et al.* A sparse PLS for variable selection when integrating omics data. *Stat Appl Genet Mol Biol* 2008;**7**:35.
25. Hotelling H. Relations between two sets of variates. *Biometrika* 1936;**28**(3–4):321–77.
26. Gonzalez I, Déjean S, Martin PGP, *et al.* Highlighting relationships between heteregeneous biological data through graphical displays based on regularized canonical correlation analysis. *J Biol Syst* 2009;**17**(02):173–99.
27. Witten DM, Tibshirani RJ. Extensions of sparse canonical correlation analysis with applications to genomic data. *Stat Appl Genet Mol Biol* 2009;**8**(1):28. DOI: 10.2202/1544-6115.1470 4.
28. Lin D, Zhang J, Li J, *et al.* Group sparse canonical correlation analysis for genomic data integration. *BMC Bioinformatics* 2013;**14**:245.
29. Conesa A, Prats-Montalbán JM, Tarazona S, *et al.* A multi-way approach to data integration in systems biology based on Tucker3 and N-PLS. *Chemom Intell Lab Syst* 2010;**104**(1): 101–11.
30. Li W, Zhang S, Liu CC, *et al.* Identifying multi-layer gene regulatory modules from multi-dimensional genomic data. *Bioinformatics* 2012;**28**(19):2458–66.
31. Löfstedt T, Hanafi M, Mazerolles G, *et al.* OnPLS path modelling. *Chemom Intell Lab Syst* 2012;**118**:139–49.

32. Löfstedt T, Hoffman D, Trygg J. Global, local and unique decompositions in OnPLS for multiblock data analysis. *Anal Chim Acta* 2013;**791**:13–24.
33. Bouhaddani SE, Houwing-Duistermaat J, Salo P, *et al*. Evaluation of O2PLS in omics data integration. *BMC Bioinformatics* 2016;**17**(S2):S11.
34. Witten DM, Tibshirani R, Hastie T. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics* 2009;**10**(3): 515–34.
35. Markovsky I. *Low Rank Approximation: Algorithms, Implementation, Applications*. Springer, 2012.
36. Jolliffe I. *Principal Component Analysis*, 2nd edn. Springer Series in Statistics. New York: Springer, 2002, 487.
37. Lock EF, Hoadley KA, Marron JS, *et al*. Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. *Ann Appl Stat* 2013;**7**(1):523–42.
38. Shen R, Olshen AB, Ladanyi M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics* 2009;**25**(22):2906–12.
39. Meng C, Kuster B, Culhane AC, *et al*. A multivariate approach to the integration of multi-omics datasets. *BMC Bioinformatics* 2014;**15**:162.
40. Wu D, Wang D, Zhang MQ, *et al*. Fast dimension reduction and integrative clustering of multi-omics data using low-rank approximation: application to cancer molecular classification. *BMC Genomics* 2015;**16**(1):1022.
41. Schouteden M, Van Deun K, Wilderjans TF, *et al*. Performing DISCO-SCA to search for distinctive and common information in linked data. *Behav Res Methods* 2014;**46**(2):576–87.
42. Liu Y, Devescovi V, Chen S, *et al*. Multilevel omic data integration in cancer cell lines: advanced annotation and emergent properties. *BMC Syst Biol* 2013;**7**:14.
43. de Tayrac M, Le S, Aubry M, *et al*. Simultaneous analysis of distinct omics data sets with integration of biological knowledge: multiple factor analysis approach. *BMC Genomics* 2009; **10**(1):32.
44. Culhane AC, Perrière G, Higgins DG. Cross-platform comparison and visualisation of gene expression data using co-inertia analysis. *BMC Bioinformatics* 2003;**4**:59.
45. Meng C, Zeleznik OA, Thallinger GG, *et al*. Dimension reduction techniques for the integrative analysis of multi-omics data. *Brief Bioinform* 2016;**17**(4):628–41.
46. Pages J. Multiple factor analysis: main features and application to sensory data. *Rev Colomb Estad* 2004;**27**:1–26.
47. Chen Y, Wu X, Jiang R. Integrating human omics data to prioritize candidate genes. *BMC Med Genomics* 2013;**6**:57.
48. Wang B, Mezlini AM, Demir F, *et al*. Similarity network fusion for aggregating data types on a genomic scale. *Nat Methods* 2014;**11**(3):333–7.
49. Speicher NK, Pfeifer N. Integrating different data types by regularized unsupervised multiple kernel learning with application to cancer subtype discovery. *Bioinformatics* 2015; **31**(12):i268–75.
50. Li W, Liu CC, Zhang T, *et al*. Integrative analysis of many weighted co-expression networks using tensor computation. *PLoS Comput Biol* 2011;**7**:e1001106.
51. Daemen A, Gevaert O, Ojeda F, *et al*. A kernel-based integration of genome-wide data for clinical decision support. *Genome Med* 2009;**1**(4):39.
52. Mariette J, Villa-Vialaneix N. Unsupervised multiple kernel learning for heterogeneous data integration. *Bioinformatics* 2017:1–7. doi: 10.1093/bioinformatics/btx682.
53. Huang S, Chaudhary K, Garmire LX. More is better: recent progress in multi-omics data integration methods. *Front Genet* 2017;**8**:84.
54. Williams EG, Wu Y, Jha P, *et al*. Systems proteomics of liver mitochondria function. *Science* 2016;**352**(6291):aad0189.
55. Zufferey A, Ibberson M, Reny JL, *et al*. New molecular insights into modulation of platelet reactivity in aspirin-treated patients using a network-based approach. *Hum Genet* 2016; **135**(4):403–14.
56. Koboldt DC, Fulton RS, McLellan MD, *et al*. Comprehensive molecular portraits of human breast tumours. *Nature* 2012; **490**(7418):61–70.
57. Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol* 2004;**3**(1):1–26.
58. Li W, Fan M, Xiong M. SamCluster: an integrated scheme for automatic discovery of sample classes using gene expression profile. *Bioinformatics* 2003;**19**(7):811–17.
59. von Luxburg U. A tutorial on spectral clustering. *Stat Comput* 2007;**17**(4):395–416.
60. Van Rijsbergen CJ. Foundation of evaluation. *J Doc* 1966;**22**: 266–8.
61. Larsen B, Aone C, Fast and effective text mining using linear-time document clustering. In: *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining—KDD '99*. ACM Press, New York, NY, 1999, 16–22.
62. Shah N, Mahajan S. Document clustering: a detailed review. *Int J Appl Inf Syst* 2012;**4**(5):30–8.
63. Londin ER, Hatzimichael E, Loher P, *et al*. The human platelet: strong transcriptome correlations among individuals associate weakly with the platelet proteome. *Biol Direct* 2014;**9**(1):3.
64. O'Connell MJ, Lock EF. R.JIVE for exploration of multi-source molecular data. *Bioinformatics* 2016;**32**:2877–9.
65. Ciucci S, Ge Y, Durán C, *et al*. Enlightening discriminative network functional modules behind principal component analysis separation in differential-omic science studies. *Sci Rep* 2017;**7**:43946.
66. Cannistraci CV, Ravasi T, Montevecchi FM, *et al*. Nonlinear dimension reduction and clustering by minimum curvilinearity unfold neuropathic pain and tissue embryological classes. *Bioinformatics* 2011;**27**:i531–9.
67. Cannistraci CV, Alanis-Lobato G, Ravasi T. Minimum curvilinearity to enhance topological prediction of protein interactions by network embedding. *Bioinformatics* 2013;**29**(13): 199–209.
68. Alanis-Lobato G, Cannistraci CV, Eriksson A, *et al*. Highlighting nonlinear patterns in population genetics datasets. *Sci Rep* 2015;**5**(1):8140.
69. Hira ZM, Gillies DF. A review of feature selection and feature extraction methods applied on microarray data. *Adv Bioinformatics* 2015;**2015**:198363. DOI: 10.1155/2015/198363.