Rearrangement of incomplete multi-omics datasets combined with ComDim for evaluating replicate cross-platform variability and batch influence

Francesc Puig-Castellví, Delphine Jouan-Rimbaud Bouveresse, Laurent Mazéas, Olivier Chapleur, Douglas N. Rutledge

Rearrangement of incomplete multi-omics datasets combined with ComDim for evaluating replicate cross-platform variability and batch influence

Francesc Puig-Castellví,[†,‡] Delphine Jouan-Rimbaud Bouveresse,[‖] Laurent Mazéas,[‡] Olivier Chapleur,[‡] and Douglas N. Rutledge*[,†,§]

[†]Université Paris-Saclay, INRAE AgroParisTech, UMR SayFood, 75005 Paris, France

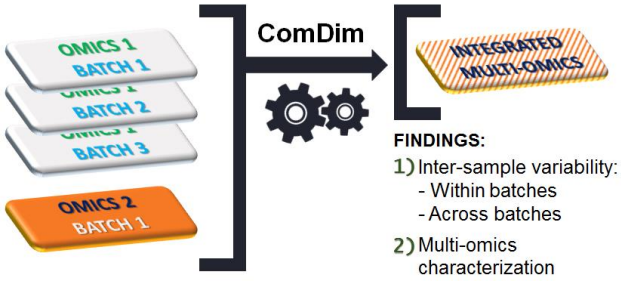
[‡]Université Paris-Saclay, INRAE, PRocédés biOtechnologiques au Service de l'Environnement, 92761 Antony, France

[‖]Université Paris-Saclay, AgroParisTech, INRAE, UMR PNCA, 75005 Paris, France

[§]National Wine and Grape Industry Centre, Charles Sturt University, 2650 Wagga Wagga, Australia

*Corresponding Author. **Tel: +33-(0)144081843,** email address: rutledge@agroparistech.fr

ABSTRACT



Multi-omics studies can highlight the interrelationships among data across different layers of biological information. However, methods for the unsupervised analysis of multi-block data do not take the individual variability across batches into account and cannot deal with omics datasets when they present different numbers of replicates. We have explored three different data arrangement strategies to tackle these limitations. Several multi-block methods can be used to decipher the common variations across blocks and to determine the contribution of each block to each common component. In this study the ComDim method was used to compare these rearrangement strategies for three multi-omics datasets. We found that arranging the data using the 'replicate by blocks' strategy, where each block comprises data from only one replicate independently of its data type, provided the most insightful results. ComDim allowed the evaluation of the variability across the replicate blocks, confirming the existence of batch effects in some of the studies. Moreover, since the contributions of these batch effects were separated from the other contributions, the coordinated biological responses common across the different blocks was characterized for each data type.

1 INTRODUCTION

The advances in high-throughput molecular technologies have enabled the simultaneous analysis of multiple biological layers such as the metabolome [1] the transcriptome (mRNA expression [2] and microRNA [miRNA] expression [3]), the genome [4], and the metataxonome [5]. These layers represent, respectively, the set of metabolites, transcripts, genes, and organisms contained in the investigated biological samples.

Multi-omics analyses promise to reveal how intricate these layers are and to deliver a more comprehensive insight than that to be had by the separate analysis of each layer [6]. As a consequence, multi-omics studies have proliferated over recent years with the aim of deciphering the biological processes related to human diseases [7], cell differentiation [8], ecotoxicology [9], and host-microbiome interactions [10], among others. Analogously, different analytical platforms are sometimes used to increase the coverage of the detected molecules within a specific omics. Multi-platform analyses are particularly popular in metabolomics [11,12]. Since each type of data constitutes a block, both multi-omics and multi-platform studies are in fact multi-block studies.

Several strategies have been developed to integrate different types of data [13]. These strategies are usually based on (1) the exploratory analysis after row-wise (individual-wise) concatenation of the data [14,15], on (2) the search of the interrelationships among data by application of partial least squares (PLS) regression [16] or canonical correlation analyses (CCA) [17], or on (3) the use of transformation-based methods such as kernels [18] or similarity networks [19]. In all these cases, the analyzed datasets must be complete (in other words, all the biological samples must be measured on all of the platforms). However, having complete datasets is not always possible [20–22]. In some cases, for a given platform, the number of sample replicates must be high to compensate for the material heterogeneity or

biological variability; in other cases, the number of sample replicates is limited due to economic constraints or so as to increase sensitivity (e.g. replicates being pooled together to improve the detection of the molecules).

With the existing data integration methods, incomplete multi-block datasets must be analyzed separately, which means that the underlying interrelationships between the different blocks cannot be fully elucidated. Alternatively, complete datasets can be constructed either by reducing the number of replicates in the larger blocks (e.g. by averaging replicates, step 1 in **Fig1B**), or by increasing the number of individuals in the smaller blocks by imputing the missing rows from the existing data (**Fig1C**) [23]. Therefore, in the second method, each data repetition would be considered a replicate sample. However, as a setback, these artificially completed datasets cannot be used to evaluate replicate variability within each block (in the first approach, due to the averaging) or across blocks (in the second approach, since the replicate samples in the block with repeated lines do not show variability while they do in the other blocks of the multi-block structure, **Fig1C**). To overcome this limitation, we propose a different data arrangement consisting of splitting the dataset into smaller blocks, each block comprising data from only one replicate, independently of its data type (**Fig1D**). With this approach, datasets with different numbers of replicates can be made compatible, while retaining all the variability due to the biological replicates in each original dataset.

The three strategies presented above for dealing with incomplete datasets were tested with the chemometrics method ComDim [24,25], a multi-block exploratory method able to extract components related to the maximum variances common to the greatest number of blocks. Although other methods could have been used [26], the objective of this study was to compare the rearrangement methods, not the exploratory multi-block methods.

In this article, we explored three data arrangements strategies, in combination with ComDim, to deal with multi-block datasets where the number of replicates was not consistent across platforms. It must be noted that the scope of this paper is not to study the case of an almost complete multi-block dataset (e.g. a few samples missing in a given platform), since in that situation the amount of missing data is low, and can be imputed using well-known methods which are considered as having negligible impact [23].

2 MATERIAL AND METHODS

2.1. Tested datasets.

In this work, three multi-omics datasets were used.

The first two datasets (Dataset 1 and Dataset 2, respectively) comprise omics data collected for the study of the temporal dynamics during the process of anaerobic digestion, which is the degradation of organic waste mediated by a microbial community and generating biogas in the process. The microbial community in the digesters is composed of Bacteria and Archaea. The Bacteria convert waste biomass into simpler molecules, while some Archaea produce methane from the simpler molecules [27].

The system in Dataset 1 was set up to evaluate the effect on the anaerobic digestion of adding a substrate rich in organic matter (either fish waste-FW, or garden grass-GG) to sewage sludge (SS), in different proportions (100%, 75%, 50%, 25% and 0% of sewage sludge). For each mixture, 3 biological replicates were sampled. Samples were collected on the 21^{st} and 28^{th} days after the start of the digestion for the FW-containing digesters, and on the 14^{th} and the 21^{st} days for the GG-containing digesters. Finally, the digester containing only SS was screened at all 3 time-points. Each waste sample was centrifuged and pellets and supernatants were kept

separately. The sample supernatants were analyzed by liquid chromatography coupled to mass spectrometry (LC-MS) metabolomics to identify the temporal dynamics in the degradation of the metabolites contained in the waste, while the sample pellets were investigated by 16S rRNA sequencing to infer the temporal changes in the composition of the microbial communities. LC-MS metabolomics was performed for all the biological replicates, while only one of each triplicate was investigated with 16S rRNA metataxonomics. The LC-MS metabolomics analysis resulted in 476 features (peak areas relative to putative metabolites), while 1,145 different Operational Taxonomic Units (OTUs, counts representative of microorganisms) were detected in the 16S rRNA metataxonomics analysis. More details about the experimental part can be found in Cardona *et al.* (2020) [28].

Dataset 2 reflects the influence of total ammonia-nitrogen (TAN) on the anaerobic digestion of sewage sludge. The degradation was carried out in nine 1 L digester bottles, where each bottle contained a specific TAN concentration (0, 0.5, 1.0, 1.5, 2.5, 5.0, 7.5, 10.0, and 25.0 mg/L). Aliquots from these digesters were collected on days 9, 29, 42, and 57. As in Dataset 1, supernatants were used for metabolomics and pellets for metataxonomics. However, in this case, metabolomics was performed by gas chromatography coupled to mass spectrometry (GC-MS) while metataxonomics was based on the 16S rDNA. GC-MS measurements were done in triplicate (analytical replicates), while the metataxonomics data was only measured once. Two sample-points (7.5 mg/L TAN at day 29, and 25 mg/L TAN at day 42) were not included in the analysis since their GC-MS data was not available. More technical information about the experimental setup can be found in Poirier *et al*. (2016) [29]. The metabolomics data were downloaded from Metabolights [30] (accession number MTBLS2602), while the metataxonomics data was obtained from Poirier et al. (2018) [31]. Regarding the number of variables, Dataset 2 contains data for 351 GC-MS features and 1,435 OTUs.

Before the chemometric analysis of these two datasets, the metataxonomics data were filtered to remove OTUs with spurious behavior [32]. Specifically, OTUs that occurred in less than 20% of the samples were excluded from further analysis. This step reduced the OTU number to 734 for Dataset 1 and to 654 for Dataset 2.

The third dataset (Dataset 3) describes the time-course differentiation of the mouse B3 cell line to the pre-BII stage [33]. This cell line can differentiate after nuclear translocation of the Ikaros transcription factor. To control the nuclear levels of the Ikaros protein, the B3 cell line was previously retrovirally transduced with a vector encoding an Ikaros-Ret2 fusion protein inducible upon exposure to the drug Tamoxifen. The control group consisted of a non-inducible B3 cell line carrying an empty vector. Samples were collected at 6 different time-points (0h, 2h, 6h, 12h, 18h, and 24h). More details about the experimental design can be found in Gomez-Cabrero *et al.* (2019) [8]. This experiment was repeated several times, resulting in the generation of several biological batches (biological replicates) that were used for different omics experiments. For metabolomics, the three biological batches were measured by both LC-MS and GC-MS. Analogously, the 3 samples used in mRNA-seq analyses were also employed in the miRNA-seq analyses. The metabolomics data were downloaded from Metabolights [30] (accession number MTBLS283), while the RNA-seq data were obtained from the GEO database [34] (accession numbers GSE75417 and GSE75394 for the mRNA-seq and miRNA-seq data, respectively). The number of variables for the omics blocks are 15 (GC-MS), 44 (LC-MS), 12,762 (mRNA-seq), and 469 (miRNA-seq).

2.2. Data pre-processing and multi-block arrangements

The metabolomics datasets were PQN-normalized [35] and auto-scaled. The metataxonomics datasets were normalized by the total sample counts and pareto-scaled. The RNA-seq datasets were already pre-processed. Then each of the datasets was reorganized using the three different

data arrangements (**Figure 1**): reducing the number of rows in the blocks containing the replicates by averaging ('Replicate Reduced' or RR, **Figure 1B**); increasing the number of rows in those blocks without replicates by simply repeating the existing data (i.e., if there is only 1 replicate and 3 are required, then each replicate is repeated 3 times [36]), resulting in having the same lines repeated several times within the block ('Replicate Augmented' or RA, **Figure 1C**); and considering each set of replicate data as a data block (i.e. a block containing triplicate data were split into 3 blocks) ('Replicate-Wise' or RW, **Figure 1D**).

For Datasets 1 and 2, RA and RR data arrangements resulted in multi-block structures with 2 blocks. However, in the RW data arrangement, each omics dataset was split into blocks comprised of data for one replicate of each sample. We will refer to the blocks corresponding to the different replicates of the same data type as the replicate blocks. Finally, the blocks (from all the omics data types) were associated row-wise. According to the RW strategy, Datasets 1 and 2 were arranged into multi-block data structures composed of 3 replicate blocks of metabolomics data (LC-MS or GC-MS data, respectively) and 1 block of 16S metataxonomics data.

On the other hand, since Dataset 3 is already a complete dataset (composed of 12 blocks: 3 replicate blocks of LC-MS data, 3 replicate blocks of GC-MS data, 3 replicate-blocks of miRNA-seq data, and 3 replicate blocks of mRNA-seq data), we have investigated it using the RW data arrangement method. In addition, we also tested the 'traditional' implementation of ComDim on this dataset, where each different data type is regarded as a single block; and also examined the effect of averaging the replicates. Thereby, the multi-block structure of the RW arrangement is composed of 12 blocks with 12 rows each, the 'traditional' arrangement has 4 blocks with 36 rows each, and the averaged dataset has 4 blocks with 12 rows each.

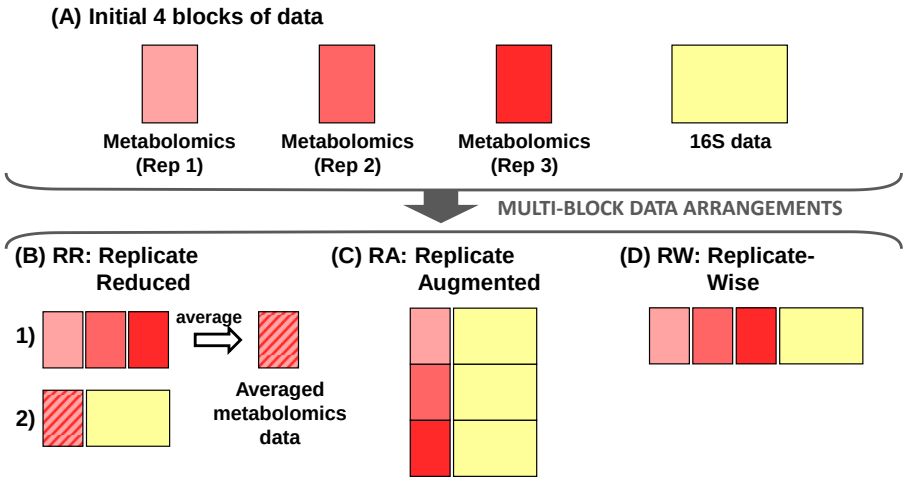


Figure 1. Data arrangements used for Datasets 1 and 2. Three replicate sets of metabolomics data (colored in different shades of red) and one set of 16S metataxonomics data (colored in yellow) were arranged using approaches B-D for multi-block analysis.

2.3. ComDim

ComDim is an unsupervised multi-block method [24,25] that aims to simultaneously consider multiple data tables to find the latent components that are common to all the tables and those that are specific to each data table, along with the contribution of each of the tables to each of these components [37]. ComDim determines a common space describing the dispersion of the samples in all the blocks, each block having a specific weight (salience) associated with each dimension in this common space. Significant differences in the saliences for a given dimension reflect the fact that the dimension contains different amounts of information coming from each block [37]. In addition to the saliences, Local loadings for each analyzed block and two different sets of scores are obtained. The first set corresponds to the Local scores for each analyzed block while the second set is composed of the Global scores, common to all the blocks.

The ComDim algorithm works as follows. First, for a multi-block dataset composed of I blocks ($\mathbf{X_i}$, where i=1,…,I), each of the blocks is centered and norm-scaled ($\mathbf{X_{i,}}_{norm}$). Then, all the blocks are weighted by the square root of their salience and concatenated row-wise [38,39]. The result of this concatenation is the matrix **W**. The salience of each block is optimized by iterative recalculation until convergence (for the first iteration, all saliences are equal to 1), and then the first normed Principal Component scores vector (**q** in **Figure 2**) is extracted. The saliences are recalculated for each block by pre- and post-multiplying the cross-product of $\mathbf{X_i}$, $\mathbf{X_i} \times \mathbf{X_i}^T$, by the vector **q**, and **W** can be calculated again using these new saliences. The first Common Component is the **q** obtained after convergence is attained. The matrices $\mathbf{X_i}$ are then all deflated and the process is repeated until the required number of CC are extracted. The matrix **Q** contains the successive Global Scores, **q**. Local Loadings, $\mathbf{P\_L_i}$, are calculated for each block using the **q** scores and the successive deflated norm-scaled $\mathbf{X_i}$. Local Scores (**T**) can then be calculated by the projection of each deflated $\mathbf{X_i}$ block onto the corresponding Local Loadings. Thus, only one set of **Q** scores is obtained, while there will be as many sets of **T** scores as $\mathbf{X_i}$ blocks.

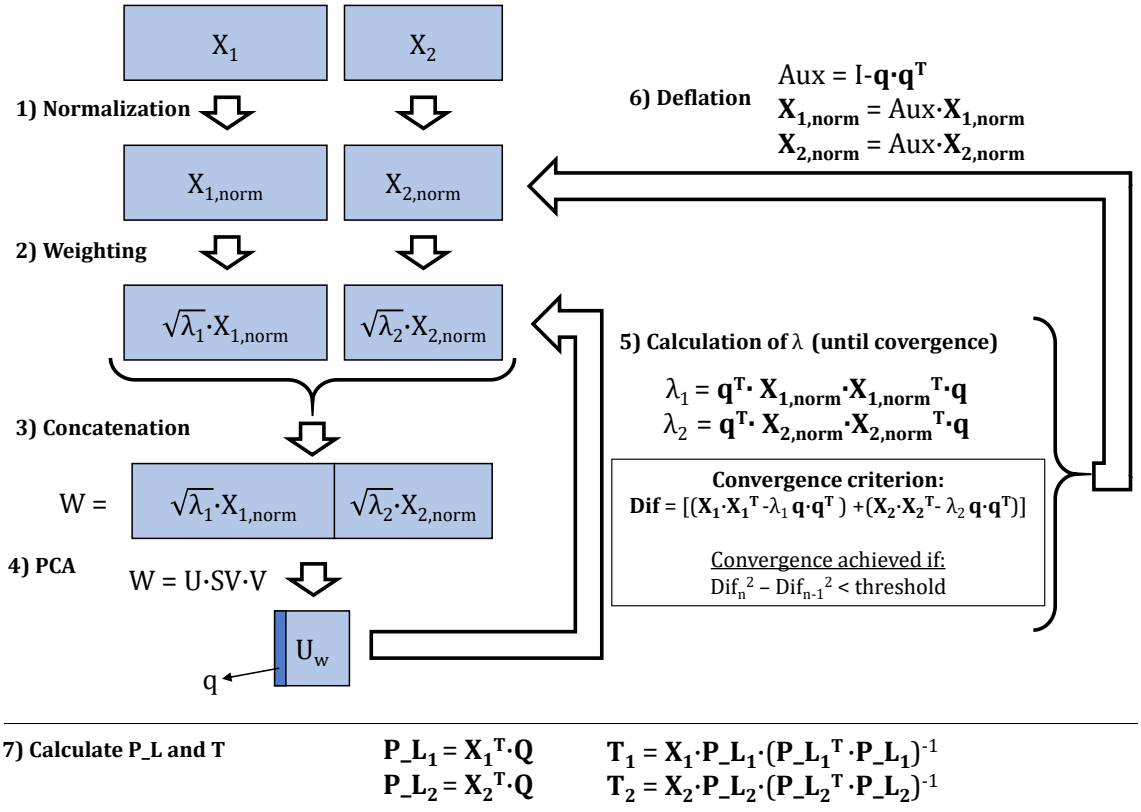


Figure 2. ComDim decomposition for a dataset composed of 2 blocks, $\mathbf{X_1}$ and $\mathbf{X_2}$. Notation used: $\mathbf{X_{1,norm}}$ and $\mathbf{X_{2,norm}}$ are the centered and norm-scaled blocks, λ_1 and λ_2 are the saliences, **Q** is the Global scores matrix, $\mathbf{T_1}$ and $\mathbf{T_2}$ are the Local scores matrices, and $\mathbf{P\_L_1}$ and $\mathbf{P\_L_2}$ are the Local loadings matrices.

To avoid giving too much weight to the data with the higher number of replicates due to its the larger presence in the dataset, each of the replicates blocks was divided by the square root of the number of replicates (i.e., $\sqrt{3}$ in the present case) prior to the weighting step (step 2 in **Figure 2**).

The code for the ComDim method is available at https://github.com/DNRutledge/ComDim (Matlab version) and https://github.com/f-puig/R.ComDim (R version).

2.4. Variable selection

The interesting variables were selected using S-plots [40]. For each data block and CC, the S-plots display the covariance and correlation values calculated from each Global scores vector and each of the normalized blocks. These covariance and correlation values present as many elements as there are variables in these blocks. The variables showing the highest covariance and correlation values were selected. Then for the RW-arranged data only the significant variables common to all replicate blocks and presenting loadings of the same sign were regarded as significant variables.

2.5. Tentative metabolite assignment.

LC-MS features from Dataset 1 were assigned to tentative metabolites using the Metlin database accepting an error tolerance of 10 ppm [41]. GC-MS features from Dataset 2 were assigned by library search in Massbank [42].

3 RESULTS AND DISCUSSION

3.1. Dataset 1

In a first comparative analysis of Dataset 1, we started by evaluating the effect of using the three different dataset arrangements depicted in **Figure 1** on PCA since it is the simplest of the bilinear decomposition methods. Before these analyses, to avoid an unbalanced contribution across blocks, each block was norm-scaled and weighted based on the number of replicates. Following this pre-processing, the replicate blocks were concatenated row-wise.

The PC scores distributions for the three analyses are very similar (**Figure 3A-C**). In all cases, PC1 captured the variations derived from altering the FW and GG contents in the samples while PC2 separated the samples according to the SS content. The time effect could not be appreciated in these two components, suggesting that microbial degradation during the

studied period was limited. The percentage of explained variance for the first two components was also similar in the three PCAs, albeit slightly higher when the RR arrangement was used.

The inspection of the loadings from these 3 analyses (**Figure S1**) revealed that the RR and the RA arrangements resulted in equivalent loadings ($r > 0.990$), despite the different numbers of samples processed, while somewhat different ($r > 0.898$) from those obtained from the RW-arranged data, indicating that the RW arrangement may influence the matrix decomposition. The 3 loadings were also compared with those obtained from the PCA of the metabolomics data alone (**Figure S2**). We observed that the latter showed the highest correlations with the RR block ($r > 0.981$ for the PC1 loadings), followed by the RA block ($r > 0.960$ for the PC1 loadings), and finally the RW block ($r > 0.873$ for the PC1 loadings). The lower correlations found between the loadings of the metabolomics data alone and the three sets of loadings of the 3 replicate blocks suggest that preweighting the 3 metabolomics blocks by $1/\sqrt{3}$ in the RW arrangement, makes the metataxonomics block more relevant in that PCA model.

The same comparative analysis of the three dataset arrangements was repeated for ComDim (**Figure 3D-M**). In all cases, the ComDim analyses were able to explain more than 81% of the dataset variance with the first component alone. The distributions of the Global scores (**Figure 3D-F**) were very similar to the scores obtained in the PCA models (**Figure 3A-C**). Moreover, the comparative analyses of the loadings resulted in interpretations analogous to those from the PCA (**Figure S3**), although the correlation coefficients between the RR- and the RA-arranged data were slightly lower than for the PCA.

In addition to the scores and loadings obtained in both the PCA and ComDim analyses, ComDim also gave the salience values (reflecting the contribution of each block) which showed that these are mainly influenced by the weighting. 16S saliences were similar for the 3

arrangements as they were equally weighted. For the LC-MS blocks, the saliences for the RW-arrangement were 1/3 of those observed in the other two arrangements, in line with the applied weight of $1/\sqrt{3}$ in the RW-arrangement. As well, the CC1 and CC2 saliences were similar for the 3 replicate blocks, indicating that they contained similar sources of structured variability, i.e., not noise. Then, for the RR- and RA-arrangements, CC1 explained in a similar amount the two data types while CC2 was slightly more descriptive of the LC-MS block. For the RW-arrangement, the contribution of the 16S block was stronger than for each of the LC-MS blocks in both CC1 and CC2.

Another information specific to ComDim is the Local scores, which allow to evaluate the data variance similarity between the biological replicates and also across platforms when the RW arrangement was used (**Figure 4**). The similar values of the Local scores for the four blocks compared with the Global scores confirmed that the blocks have comparable sample variance and effects on the dispersion of the samples, even across platforms.

While both PCA and ComDim showed similar results in terms of their loadings and their scores (PCA) / Global scores (ComDim), showing that both methods are equally valid to explore dataset variability, ComDim however also provided information about the contribution of each block in the global model (by the saliences) and how well represented each sample is in the model (by the local scores). For this reason, we consider that the use of ComDim should be preferred rather than PCA for the exploration of incomplete datasets.

In a further step, we also studied the differences in variable selection for the three ComDim analyses tested. The significant variables were selected using S-plots, setting the threshold to 1.5 standard deviations (SD). Next, the results from the three ComDim analyses were compared

employing Venn diagrams depicting the number of selected variables in common (**Figure 3J-M**).

In general, the 16S variables (or OTUs) selected were mostly the same for the three analyses. This can be explained because the data variance encoded by the 16S block is the same in the three analyses. Conversely, some differences were observed for the variable selection for the LC-MS data.

Firstly, when only the RR- and the RA-arrangements are considered, an important variability in the variable selection was noticed, since between ⅔ and ¾ of the selected variables were common to these two analyses (**Figure 3J-M**).

Secondly, the variable selection applied to the RW-arranged dataset was the most restrictive approach, and most of the selected variables were also found to be significant for the other two data arrangements. The stringency and consistency in the results of the RW-arranged dataset came from the fact that the three LC-MS blocks are investigated separately with the S-plots and only the variables in common are considered as truly significant.

In agreement with the previous paragraph, a biological interpretation of the results of the analysis of the RW-arranged data was done. Briefly, the metabolic data from CC1 revealed that FW is richer than SS in pyrrolidine, 3-methylbutanamine, and indole-3-carbinol, among others. Similarly, the metabolic data from CC2 highlighted that the metabolites characteristic of SS are heptanethiol and triethanolamine, among others; while the presence of oxidized organic acids was inversely related to the amount of this type of waste. Overall, these data suggest that waste degradation was promoted by different microorganisms depending on the amount of FW, GG, or SS. Specifically, the metabolic changes described by CC1 were mediated by the reduction of the contribution of microorganisms from the *Cloacimonetes* and *Plantomycetes* phyla in the

microbial community, among others. The same component was also associated with an increase of the *Bacteroidetes* as well as some *Firmicutes* species. Finally, most of the species selected in CC2 (comprising bacterial species from the phyla *Bacteroidetes*, *BRC1*, *Cloacimonentes*, *Chloroflexi*, *Firmicutes* and *Synergistetes*, *Coprothermobacteraeota*, *Proteobacteria*, and *Plantomycetes*; and 3 archaeal species) were more abundant in the samples containing a higher amount of sewage sludge. The full list of selected metabolites and microbial species are given in the **Appendix 2**.

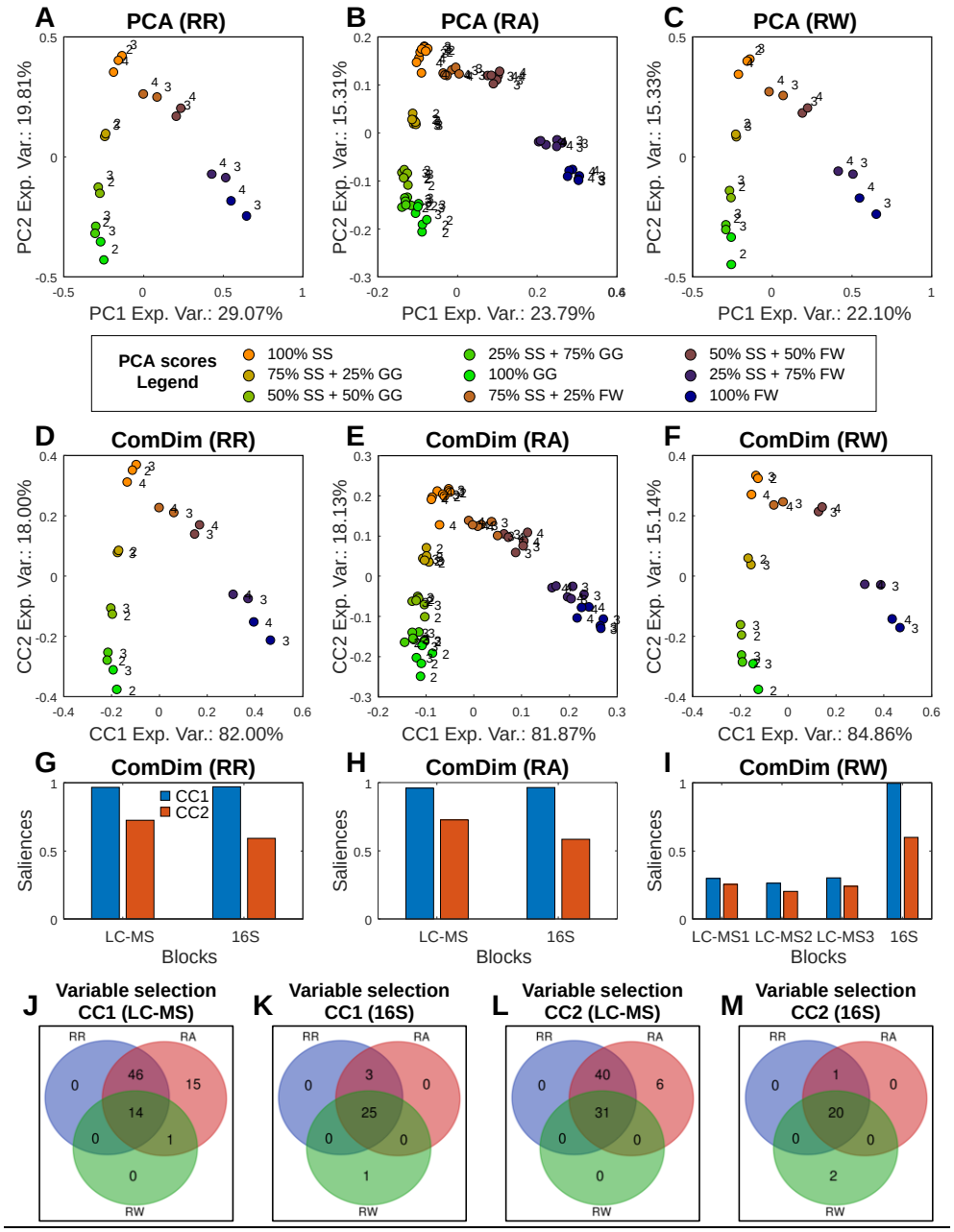


Figure 3. Analysis of Dataset 1 using different data arrangements. PCA analyses: PC1 vs PC2 scores **(A-C)**. ComDim analyses: CC1 vs CC2 scores **(D-F)** and ComDim saliences **(G-I)**. For **(A-I)**, plots from the RR-arranged data are on the left, those from the RA-arranged data in the center,

and those from the RW-arranged data on the right. **G-J)** Venn Diagrams showing the similarity in variable selection among the three ComDim analyses. The numbers in the scores plots represent the time progression in weeks of the anaerobic digestion.

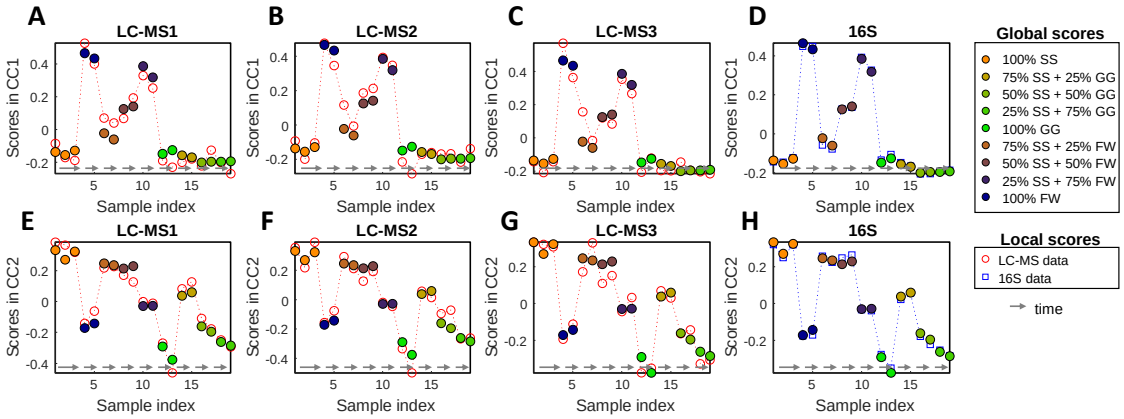


Figure 4. Comparison between the Global ComDim scores and the Local ComDim scores obtained from the ComDim analysis of the RW-arranged Dataset 1. Global ComDim scores were drawn with filled circles, and Local ComDim scores with empty circles (for LC-MS data) or squares (for 16S data). For each replicate block, the grey arrows denote the time progression for each tested condition.

3.2. Dataset 2

Three ComDim analyses were performed on Dataset 2, one for each dataset arrangement (**Figure 5**). In all cases, CC1 alone explained more than 76% of the dataset variance. The three resolved CC1 components are mainly descriptive of time while the three CC2 components relate to the TAN concentration used. Interestingly, this component reveals a proportionality between these scores and the TAN concentration, but only for TAN concentrations up to 10 mg/L. The existence of a different microbial response depending on the level of TAN concentration, in terms of the 16S metataxonomics data, has already been observed by Puig-

Castellví *et al.* (2020) [43]. Then, it must be noted that the ComDim scores from the RA-arrangement are significantly different from those of the other two arrangements, indicating that replicating the samples may in fact have an impact on the resolution in some cases.

Regarding the saliences, for the RR-arrangement, CC1 and CC2 were more descriptive of the LC-MS block than of the 16S block (**Fig5D**). For the RW-arrangement, 16S saliences were similar to those obtained from the RR-arrangement and LC-MS were 1/3 of the corresponding RR-saliences (**Fig5F**), following the same trend observed for Dataset 1 (**Fig3F**). The CC1 saliences for the 3 repetition blocks of GC-MS data are very similar, while the CC2 salience for the first repetition block is somewhat lower than for the other two, indicating that it may contain some noise. A very different result was obtained for the RA-arrangement. In this case CC1 saliences were similar to those obtained from the RR-arrangement, while CC2 saliences followed the opposite trend by being more descriptive of the GC-MS data than of the 16S data. Hence, the artificial augmentation of the dataset with replicate data was a source of variation extracted by CC2.

CC2 local scores obtained in the RW-arrangement (**Fig6**) revealed that there is an important inter-sample variability in the GC-MS data as the scores are significantly different across the three replicate blocks, confirming what was observed for the corresponding saliences. From these results, we can deduce that, in the RA-arrangement, the pairing of the 16S block containing identical data as replicates with the GC-MS block with replicates with low similarity produced an important distortion in the ComDim resolution. This effect was not observed in the RR-arranged data since the GC-MS was averaged, thus reducing the effect of noise. For the RW-arranged data, the phenomenon was not observed since the RW arrangement allows each

set of replicates to be described by the model in a different way, as each replicate block has its own saliences and its own loadings.

We also inspected the loading blocks of the three data arrangements (**FigS5**). As expected, CC1 loadings for RR- and RA-arrangements are very similar ($r > 0.997$), while for CC2 they are weakly correlated ($r > 0.650$). The same trend was observed between the loadings of the RW- and the RA-arranged data. Finally, we observed that the correlations values of each of the replicate blocks in the RW-arranged data differed considerably. For example, between the RR- and the RW- arranged data the correlation ranged between 0.482 to 0.952 (**FigS5E**) as a consequence of the GC-MS sample variability.

We can thereby observe that the RW strategy, as opposed to the other two, is able to confirm whether the inter-sample variability among replicates is low (as in Dataset 1) or not (as in Dataset 2), suggesting in the latter an underlying analytical batch effect. In other words, ComDim can extract at the same time common (in the Global scores) and distinct (or block-specific, in the Local scores) component profiles. This would not be visible from just the inspection of the PCA results since PCA lacks the Local scores, replicate block-specific loadings and saliences. To minimize the effect of the inter-sample variability, we could average the GC-MS replicates. However, we could simply leave out this component from further analyses. The removal of a ComDim component descriptive of a chromatographic drift has been previously used to correct such batch effects [44].

Regarding the variable selection, most of the variables were selected by all the methods (specifically,36 GC-MS variables and 36 16S variables) for CC1. This consistency in the results may be derived from a low inter-sample variability of the GC-MS data in CC1. This was also observed in the comparison of the CC1 loadings (**Figure S5**), as they look identical across

replicates. On the other hand, the analysis of the RW arrangement was again the only one able to demonstrate the low repeatability in CC2 for the GC-MS data, as no variable was selected.

For the 39 selected GC-MS features in CC1 in the analysis of the RW-arranged dataset, 16 eluted at 19.9 min (m/z of 61, 75, 89, 91, 104, 106, 129, 132, 135, 161, 163, 205, 207, 209, 222, and 224 Da), 4 at 16.9 min (m/z of 105, 135, 179 and 181 Da), and 4 at 24.6 min (m/z of 177, 205, 310, and 312 Da), among others. These three groups of GC-MS features were tentatively assigned to phenylpropanoic acid (1TMS), hippuric acid, and 3-(3-hydroxyphenyl)propanoic acid, respectively. Phenylpropanoic acid and 3-(3-hydroxyphenyl)propanoic acid were consumed over time while hippuric acid was produced. These three metabolites belong to the phenylalanine metabolic pathway, suggesting that the temporal component of the microbial activity, which is highlighted by CC1, strongly modifies this pathway. The fact that the different mass fragments of these compounds can be captured in the same component highlights the effectiveness of ComDim for the analysis of this type of data.

For the 16S metataxonomics data, the 38 OTUs selected in CC1 are one Archaea (*Methanosarcina mazei*) and 35 Bacteria. Among them, it includes 18 *Clostridiales*, 9 *Bacteroidales*, two *Anaerolineales*, one *Petrotogales*, one *Synergistales*, and a bacterial OTU from the *Armatimonadetes* phylum. Most of the selected *Bacteroidales* diminished over time, while the contribution of most of the *Clostridiales* accumulated over time. Regarding CC2, as can be seen in **Figures 6E-6H**, only the local scores of the 16S data block are close to the global scores, reflecting that it is this block that contributes to CC2. 27 OTUs were selected for CC2. including 9 *Clostridiales* and 3 *Bacteroidales* that were less abundant at lower TAN concentrations and one *Spirochaeateles* that accumulated at higher TAN levels.

3.3. Dataset 3

To conclude this work, we analyzed Dataset 3 to demonstrate that the RW-arrangement can be extended to any multi-block dataset containing replicates.

The variability in Dataset 3 can be characterized by 4 ComDim components (**Figure 7A-H**). CC1-CC4 components explained 66.68 %, 16.45 %, 10.91 % and 5.96 % of the total variance, respectively (**Figure 7A-D**).

The saliences showed that the highest contributions in CC1 were from the mRNA-seq and miRNA-seq blocks, while the contribution of the metabolomics blocks was between one half and one third of that (**Figure 7E**). Regarding the ComDim Global scores, CC1 showed a stable profile for the control samples (blue), while the Global scores of the differentiated cells samples (yellow) increased linearly with time (**Figure 7A**). A similar response, although not as well defined for the metabolomics blocks, was observed in the Local scores (**Figure S6**).

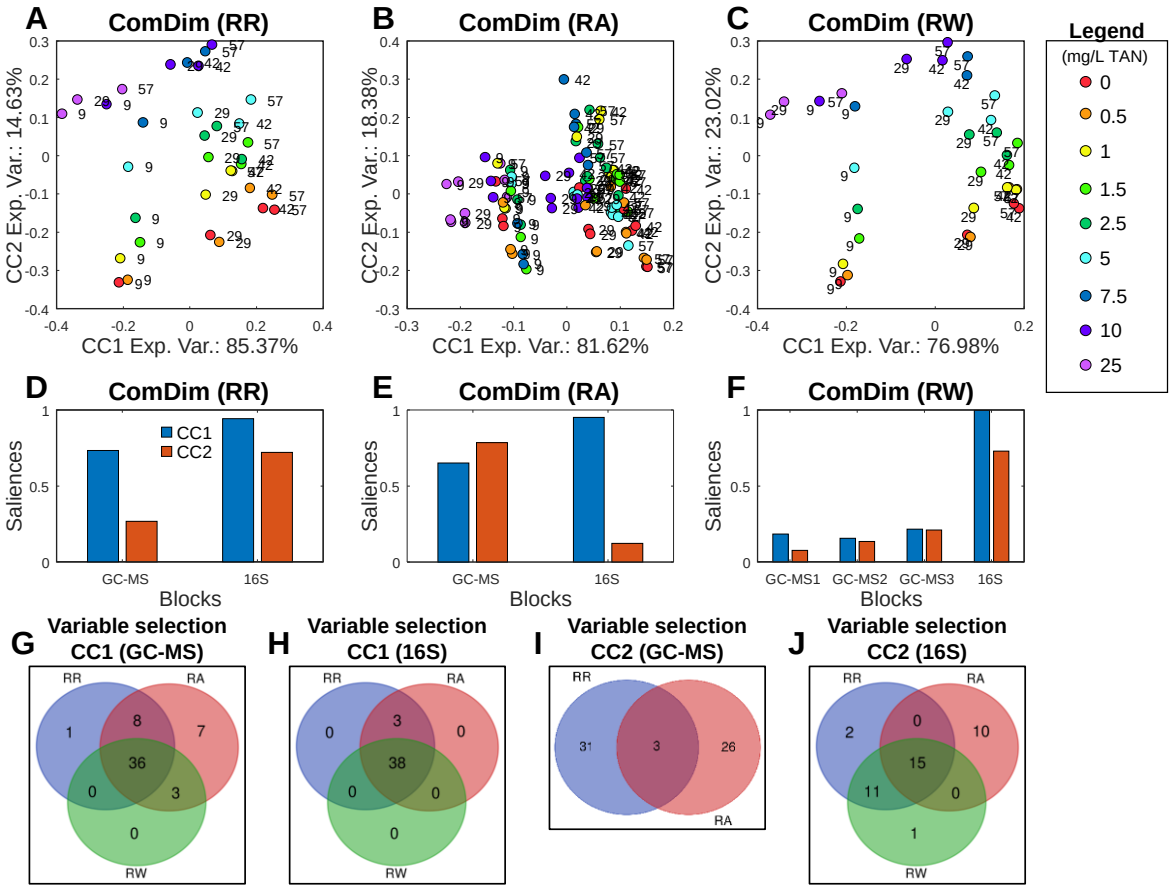


Figure 5. ComDim analyses of Dataset 2 using different data arrangements. **A-C)** CC1 vs CC2 scores. Scores are colored using the same color code as in **Figure 2A-F**. The numbers in the scores plots represent the time progression in days of the anaerobic digestion. **D-F)** ComDim saliences. For (**A-F**), the ComDim results from the RR-arranged dataset are on the left, the ones from the RA-arranged dataset in the center, and those from the RW-arranged dataset on the right. **G-J)** Venn Diagrams showing the similarity in variable selection among the three ComDim analyses.

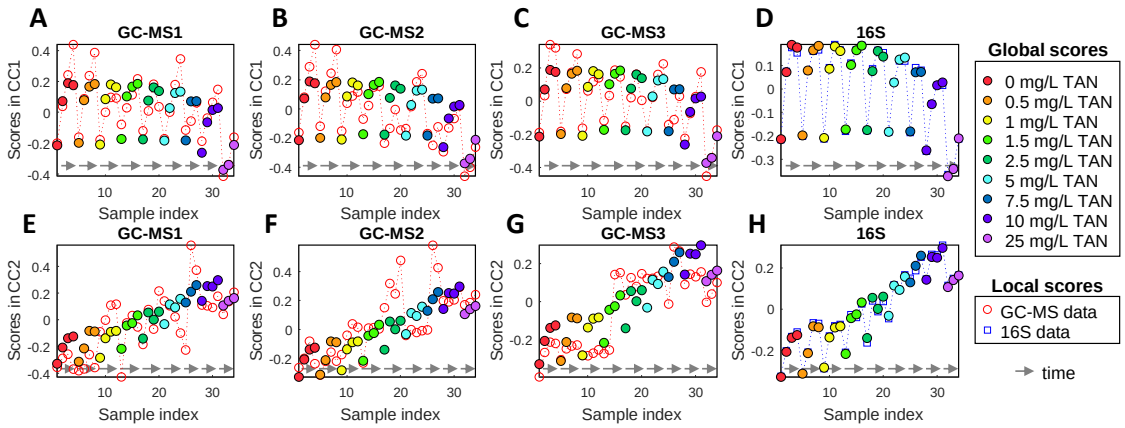


Figure 6. Comparison between the Global ComDim scores and the Local ComDim scores obtained from the ComDim analysis of the RW-arranged in Dataset 2. Global ComDim scores were plotted with filled circles, and Local ComDim scores with empty circles (for LC-MS data) or squares (for 16S data). For each replicate block, the grey arrows denote the time progression for each tested condition.

The temporal pattern for CC2-CC4 Global scores was more complex than for CC1 (**Figure 7B-D**). Interestingly, from the analysis of the saliences it can be deduced that CC2-CC4 components mainly described the GC-MS and LC-MS data types (**Figure 7F-H**) and were all very batch-dependent. That is to say, CC2 explained the 3rd batch of samples of these two data types (**Figure 7F**), CC3 the 1st batch (**Figure 7G**), and CC4 the 2nd batch (**Figure 7H**). In the original manuscript, it is indicated that the cell pellets in 2 of the 3 batches were not completely dried [8], which could have affected the sample stability and the metabolite extraction, thereby explaining the batch-dependent differences. Thus, ComDim detects variations among replicate samples, even when comparing different data types. Moreover, since it extracts the aberrant

variance in separate components, the multi-block data can still be interpreted by looking at the components presenting a similar salience between replicates (in this case, CC1).

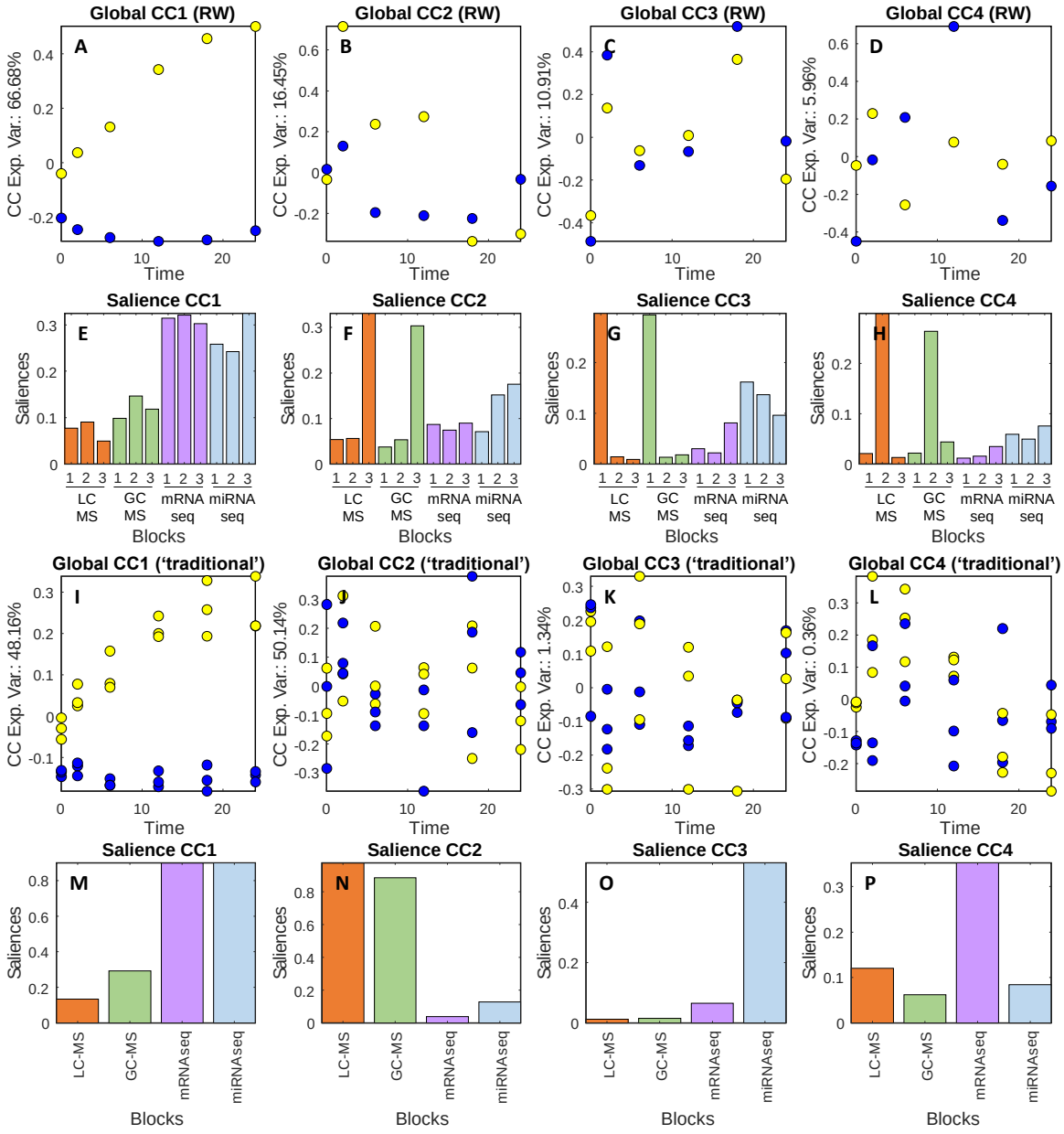


Figure 7. Analysis of Dataset 3**. A-H)** ComDim analysis of Dataset 3 using the RW arrangement. Global ComDim scores plotted over time for CC1-CC4 are given in (**A-D**), while the corresponding

block saliences are presented in (**E-H**). **I-P)** for the ComDim analysis of Dataset 3 using the 'traditional' arrangement. Global ComDim scores plotted over time for CC1-CC4 are given in (**I-L**), while the corresponding block saliences are presented in (**M-P**).

We also investigated Dataset 3 arranged with one data type per block (in the 'traditional' implementation, see methods) using ComDim (**Figure 7I-P**). It must be noted that, when using the 'traditional' data arrangement, it is assumed that each sample has a paired sample in the other data blocks. However, this correspondence across blocks might not in fact exist. For example, for this experiment, the three biological batches used to obtain the metabolomics data (e.g., M1, M2 and M3) are not the same ones used for the RNA-seq data (e.g., T1, T2 and T3) (see methods). Since the 6 batches are independent, M1 can be indistinctly paired to either T1, T2, or T3, and so can M2 and M3. This would lead to different ComDim results depending on the chosen pairings. However, this is not a problem for the RW data arrangement since all the samples from the same biological condition are aligned in the same row (each block is composed of one replicate and data type, therefore 12 blocks). Within each block there is only one sample from each studied condition, and so the pairing across blocks is unique. Due to this particular arrangement, the Global scores are only descriptive of the process studied and the variation between replicates must be inferred from inspecting the saliences or by looking at the Local scores.

CC1 explained the cell differentiation for the 'traditional' data arrangement , accounting for 48.16 % of the explained variance. This is much lower than the 66.68 % of explained variance for CC1 with the RW arrangement.

Moreover, since the 'traditional' arrangement does not allow examining the replicate variability from the saliences (**Figure 7M-P**), other ComDim results must be examined instead.

The Global ComDim scores and the Local ComDim scores are given in **Figure 7I-L** and **Figure S7**, respectively. The batch effect cannot be observed in the Global scores as this effect only exists for the LC-MS and GC-MS data (therefore, it is not a 'global' effect). On the other hand, the CC2 Local scores showed strong correspondences between replicates across the LC-MS and GC-MS data, thus indicating that the two data sets are related. This is also indicated by the similar saliences in **Figure 7N**. Having said that, this observation is much more easily pinpointed from the CC2-CC4 ComDim saliences obtained in the analysis of the proposed RW arrangement (**Figure 7E-H**).

Lastly, we repeated the ComDim strategy in Dataset 3 after averaging the replicates (**Figure S8** and **Figure S9**). Saliences and Global scores from CC1 and CC2 are similar to those obtained from the analysis of the 'traditional' arrangement, as well as the Local scores (if compared to the corresponding averages). The greater resemblance of the results of the averaged dataset to those from the 'traditional' analysis than that from the 'RW' arrangement points out that rearranging from 4 to 12 blocks has a greater influence on the resolution than averaging.

Since the RW-arranged data seems to be more meaningful than when the data is arranged in the 'traditional' way, as it allowed isolating the batch effect of the metabolomics data in separate components, the step of biomarker discovery was performed for the former data arrangement only. Therefore, the biological signatures (the LC-MS metabolites, the GC-MS metabolites, the mRNA transcripts, and the miRNA transcripts) of cell differentiation, represented by CC1, were searched for. To do so, the variables from each block were inspected using S-plots and those exceeding one SD were selected. Then, only the variables found to be significant for the three blocks of replicate data were retained.

This resulted in selecting 6 LC-MS variables (L-cysteine, homoserine, 5-hydroxy-L-tryptophan, putrescine, spermidine, and taurine), 1 GC-MS variables (malic acid), 79 miRNA-seq variables, and 3760 mRNA-seq variables.

The selected metabolites and transcripts were further investigated using bioinformatics tools to assess their role in the biological processes. This analysis can be consulted in the **Appendix 2**.

4 CONCLUSION

In the present study, we examined three different data arrangement strategies designed to cope with inconsistent replicate sample numbers across blocks in multi-block studies, and it was shown that considering each set of single-type replicates as a standalone data block is the most powerful strategy.

The ability of ComDim to deal with multi-omics data sets was validated with three datasets, allowing the identification of the changes in the omics profiles during two anaerobic waste digestions processes and during cell differentiation.

By combining the best data arrangement strategy and ComDim, we were able to assess the variability across platforms and among replicates. Specifically, we detected (1) that CC2 was uninformative in regards to the GC-MS data for Dataset 2, and (2) that there was a batch effect impacting only the (LC-MS and GC-MS) metabolomics blocks in Dataset 3. Despite this, since ComDim components are orthogonal, the sample variances due to the studied factor and to the batch effects were allocated to separate components, thereby allowing a successful analysis and interpretation of the multi-omics data.

Finally, with the proposed RW data arrangement, variable selection methods must be applied to each replicate block and only the compromise biomarkers for all the replicates are retained. This results in a shorter (but more reliable) list of biomarkers, compared to the analyses using the other data arrangement methods.

APPENDICES

Appendix A. Supplementary results: comparative analysis of the loadings from Datasets 1 and 2, Local scores from Dataset 3, analysis of the averaged Dataset 3, most significant changes in Dataset 3 (PDF).

Appendix B. Lists of significant variables for Datasets 1, 2 and 3 (XLSX).

Author Contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENT

This work is part of the DIGESTOMIC project funded by the French National Research Agency (ANR-16-CE05-0014).

REFERENCES

[1] G.J. Patti, O. Yanes, G. Siuzdak, Metabolomics: the apogee of the omics trilogy, Nat. Rev. Mol. Cell Biol. 13 (2012) 263–269. https://doi.org/10.1038/nrm3314.

[2] Z. Wang, M. Gerstein, M. Snyder, RNA-Seq: A revolutionary tool for transcriptomics, Nat. Rev. Genet. 10 (2009) 57–63. https://doi.org/10.1038/nrg2484.

[3] J. Krauskopf, T.M. de Kok, D.G. Hebels, I.A. Bergdahl, A. Johansson, F. Spaeth, H. Kiviranta, P. Rantakokko, S.A. Kyrtopoulos, J.C. Kleinjans, MicroRNA profile for health risk assessment: Environmental exposure to persistent organic pollutants strongly affects the human blood microRNA machinery, Sci. Rep. 7 (2017) 9262. https://doi.org/10.1038/s41598-017-10167-7.

[4] M. Akiyama, Multi-omics study for interpretation of genome-wide association study, J. Hum. Genet. (2020). https://doi.org/10.1038/s10038-020-00842-5.

[5] J.W. Cox, R.A. Ballweg, D.H. Taft, P. Velayutham, D.B. Haslam, A. Porollo, A fast and robust protocol for metataxonomic analysis using RNAseq data, Microbiome. 5 (2017) 7. https://doi.org/10.1186/s40168-016-0219-5.

[6] B.B. Misra, C. Langefeld, M. Olivier, L.A. Cox, Integrated omics: Tools, advances and future approaches, J. Mol. Endocrinol. 62 (2019) R21–R45. https://doi.org/10.1530/JME-18-0055.

[7] Y. Hasin, M. Seldin, A. Lusis, Multi-omics approaches to disease, Genome Biol. 18 (2017) 1–15. https://doi.org/10.1186/s13059-017-1215-1.

[8] D. Gomez-Cabrero, S. Tarazona, I. Ferreirós-Vidal, R.N. Ramirez, C. Company, A. Schmidt, T. Reijmers, V. von Saint Paul, F. Marabita, J. Rodríguez-Ubreva, A. Garcia-Gomez, T. Carroll, L. Cooper, Z. Liang, G. Dharmalingam, F. van der Kloet, A.C. Harms, L. Balzano-Nogueira, V. Lagani, I. Tsamardinos, M. Lappe, D. Maier, J.A. Westerhuis, T. Hankemeier, A. Imhof, E. Ballestar, A. Mortazavi, M. Merkenschlager, J. Tegner, A. Conesa, STATegra, a comprehensive multi-omics dataset of B-cell differentiation in mouse, Sci. Data. 6 (2019) 256. https://doi.org/10.1038/s41597-019-0202-7.

[9] E. Ortiz-Villanueva, L. Navarro-Martín, J. Jaumot, F. Benavente, V. Sanz-Nebot, B. Piña, R. Tauler, Metabolic disruption of zebrafish (Danio rerio) embryos by bisphenol A. An integrated metabolomic and transcriptomic approach, Environ. Pollut. 231 (2017) 22–36. https://doi.org/10.1016/j.envpol.2017.07.095.

[10] L. Nyholm, A. Koziol, S. Marcos, A.B. Botnen, O. Aizpurua, S. Gopalakrishnan, M.T. Limborg, M.T.P. Gilbert, A. Alberdi, Holo-omics: integrated host-microbiota multi-omics for basic and applied biological research, IScience. 23 (2020) 101414. https://doi.org/10.1016/j.isci.2020.101414.

[11] A. Gonzalez-Martinez, A. Rodriguez-Sanchez, M.J. Garcia-Ruiz, B. Muñoz-Palazon, C. Cortes-Lorenzo, F. Osorio, R. Vahala, Performance and bacterial community dynamics of a CANON bioreactor acclimated from high to low operational temperatures, Chem. Eng. J. 287 (2016) 557–567. https://doi.org/10.1016/j.cej.2015.11.081.

[12] L. Li, J. Zhao, Y. Zhao, X. Lu, Z. Zhou, C. Zhao, G. Xu, Comprehensive investigation of tobacco leaves during natural early senescence via multi-platform metabolomics analyses, Sci. Rep. 6 (2016) 1–10. https://doi.org/10.1038/srep37976.

[13] G. Tini, L. Marchetti, C. Priami, M.P. Scott-Boyer, Multi-omics integration-A comparison of unsupervised clustering methodologies, Brief. Bioinform. 20 (2018) 1269–1279. https://doi.org/10.1093/bib/bbx167.

[14] L.N. Rosa, L.C. de Figueiredo, E.G. Bonafé, A. Coqueiro, J.V. Visentainer, P.H. Março, D.N. Rutledge, P. Valderrama, Multi-block data analysis using ComDim for the evaluation of complex samples: Characterization of edible oils, Anal. Chim. Acta. 961 (2017) 42–48. https://doi.org/10.1016/j.aca.2017.01.019.

[15] E.F. Lock, K.A. Hoadley, J.S. Marron, A.B. Nobel, Joint and individual variation explained (JIVE) for integrated analysis of multiple data types, Ann. Appl. Stat. 7 (2013) 523–542. https://doi.org/10.1214/12-AOAS597.

[16] K.A. Lê Cao, D. Rossouw, C. Robert-Granié, P. Besse, A sparse PLS for variable selection when integrating omics data, Stat. Appl. Genet. Mol. Biol. 7 (2008). https://doi.org/10.2202/1544-6115.1390.

[17] A. Singh, C.P. Shannon, B. Gautier, F. Rohart, M. Vacher, S.J. Tebbutt, K.A.L. Cao, DIABLO: An integrative approach for identifying key molecular drivers from multi-omics assays, Bioinformatics. 35 (2019) 3055–3062. https://doi.org/10.1093/bioinformatics/bty1054.

[18] N.K. Speicher, N. Pfeifer, Integrating different data types by regularized unsupervised multiple kernel learning with application to cancer subtype discovery, Bioinformatics. 31 (2015) i268–i275. https://doi.org/10.1093/bioinformatics/btv244.

[19] B. Wang, A.M. Mezlini, F. Demir, M. Fiume, Z. Tu, M. Brudno, B. Haibe-Kains, A. Goldenberg, Similarity network fusion for aggregating data types on a genomic scale, Nat. Methods. 11 (2014) 333–337. https://doi.org/10.1038/nmeth.2810.

[20] D.J. Beale, J. Crosswell, A. V. Karpe, S.S. Metcalfe, P.D. Morrison, C. Staley, W. Ahmed, M.J. Sadowsky, E.A. Palombo, A.D.L. Steven, Seasonal metabolic analysis of marine sediments collected from Moreton Bay in South East Queensland, Australia, using a multi-omics-based approach, Sci. Total Environ. 631–632 (2018) 1328–1341. https://doi.org/10.1016/j.scitotenv.2018.03.106.

[21] Y. Duan, D. Xiong, Y. Wang, H. Li, H. Dong, J. Zhang, Toxic effects of ammonia

and thermal stress on the intestinal microbiota and transcriptomic and metabolomic responses of Litopenaeus vannamei, Sci. Total Environ. 754 (2021) 141867. https://doi.org/10.1016/j.scitotenv.2020.141867.

[22] J. Lloyd-Price, C. Arze, A.N. Ananthakrishnan, M. Schirmer, J. Avila-Pacheco, T.W. Poon, E. Andrews, N.J. Ajami, K.S. Bonham, C.J. Brislawn, D. Casero, H. Courtney, A. Gonzalez, T.G. Graeber, A.B. Hall, K. Lake, C.J. Landers, H. Mallick, D.R. Plichta, M. Prasad, G. Rahnavard, J. Sauk, D. Shungin, Y. Vázquez-Baeza, R.A. White, J. Bishai, K. Bullock, A. Deik, C. Dennis, J.L. Kaplan, H. Khalili, L.J. McIver, C.J. Moran, L. Nguyen, K.A. Pierce, R. Schwager, A. Sirota-Madi, B.W. Stevens, W. Tan, J.J. ten Hoeve, G. Weingart, R.G. Wilson, V. Yajnik, J. Braun, L.A. Denson, J.K. Jansson, R. Knight, S. Kugathasan, D.P.B. McGovern, J.F. Petrosino, T.S. Stappenbeck, H.S. Winter, C.B. Clish, E.A. Franzosa, H. Vlamakis, R.J. Xavier, C. Huttenhower, Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases, Nature. 569 (2019) 655–662. https://doi.org/10.1038/s41586-019-1237-9.

[23] V. Voillet, P. Besse, L. Liaubet, M. San Cristobal, I. González, Handling missing rows in multi-omics data integration: multiple imputation in multiple factor analysis framework, BMC Bioinformatics. 17 (2016) 402. https://doi.org/10.1186/s12859-016-1273-5.

[24] E.M. Qannari, P. Courcoux, E. Vigneau, Common components and specific weights analysis performed on preference data, Food Qual. Prefer. 12 (2001) 365–368. https://doi.org/10.1016/S0950-3293(01)00026-X.

[25] G. Mazerolles, M. Hanafi, E. Dufour, D. Bertrand, E.M. Qannari, Common

components and specific weights analysis: A chemometric method for dealing with complexity of food products, Chemom. Intell. Lab. Syst. 81 (2006) 41–49. https://doi.org/10.1016/J.CHEMOLAB.2005.09.004.

[26] P. Mishra, J.M. Roger, D. Jouan-Rimbaud-Bouveresse, A. Biancolillo, F. Marini, A. Nordon, D.N. Rutledge, Recent trends in multi-block data analysis in chemometrics for multi-source data integration, TrAC - Trends Anal. Chem. 137 (2021) 116206. https://doi.org/10.1016/j.trac.2021.116206.

[27] L. Hao, A. Bize, D. Conteau, O. Chapleur, S. Courtois, P. Kroff, E. Desmond-Le Quéméner, T. Bouchez, L. Mazéas, New insights into the key microbial phylotypes of anaerobic sludge digesters under different operational conditions, Water Res. 102 (2016) 158–169. https://doi.org/10.1016/j.watres.2016.06.014.

[28] L. Cardona, K.A.L. Cao, F. Puig-Castellví, C. Bureau, C. Madigou, L. Mazéas, O. Chapleur, Integrative Analyses to Investigate the Link between Microbial Activity and Metabolite Degradation during Anaerobic Digestion, J. Proteome Res. 19 (2020) 3981–3992. https://doi.org/10.1021/acs.jproteome.0c00251.

[29] S. Poirier, E. Desmond-Le Quéméner, C. Madigou, T. Bouchez, O. Chapleur, Anaerobic digestion of biowaste under extreme ammonia concentration: Identification of key microbial phylotypes, Bioresour. Technol. 207 (2016) 92–101. https://doi.org/10.1016/J.BIORTECH.2016.01.124.

[30] K. Haug, K. Cochrane, V.C. Nainala, M. Williams, J. Chang, K.V. Jayaseelan, C. O'Donovan, MetaboLights: A resource evolving in response to the needs of its scientific community, Nucleic Acids Res. 48 (2020) D440–D444. https://doi.org/10.1093/nar/gkz1019.

[31] S. Poirier, O. Chapleur, Inhibition of anaerobic digestion by phenol and ammonia: Effect on degradation performances and microbial dynamics, Data Br. 19 (2018) 2235–2239. https://doi.org/https://doi.org/10.1016/j.dib.2018.06.119.

[32] F. Puig-Castellví, L. Cardona, D. Jouan-Rimbaud Bouveresse, C.B.Y. Cordella, L. Mazéas, D.N. Rutledge, O. Chapleur, Assessment of the microbial interplay during anaerobic co-digestion of wastewater sludge using common components analysis, PLoS One. 15 (2020) e0232324. https://doi.org/10.1371/journal.pone.0232324.

[33] I. Ferreiros-Vidal, T. Carroll, B. Taylor, A. Terry, Z. Liang, L. Bruno, G. Dharmalingam, S. Khadayate, B.S. Cobb, S.T. Smale, M. Spivakov, P. Srivastava, E. Petretto, A.G. Fisher, M. Merkenschlager, Genome-wide identification of Ikaros targets elucidates its contribution to mouse B-cell lineage specification and pre-B-cell differentiation, Blood. 121 (2013) 1769–1782. https://doi.org/10.1182/blood-2012-08-450114.

[34] E. Clough, T. Barrett, The Gene Expression Omnibus database, in: Methods Mol. Biol., Humana Press Inc., 2016: pp. 93–110. https://doi.org/10.1007/978-1-4939-3578-9_5.

[35] Frank Dieterle, Alfred Ross, and Götz Schlotterbeck, H. Senn*, Probabilistic Quotient Normalization as Robust Method to Account for Dilution of Complex Biological Mixtures. Application in 1H NMR Metabonomics, (2006). https://doi.org/10.1021/AC051632C.

[36] A.M. Gholami, K. Fellenberg, Cross-species common regulatory network inference without requirement for prior gene affiliation, Bioinformatics. 26 (2010) 1082–1090. https://doi.org/10.1093/bioinformatics/btq096.

[37] M. Claeys-Bruno, A. Béal, D.N. Rutledge, M. Sergent, Use of the common components and specific weights analysis to interpret supersaturated designs, Chemom. Intell. Lab. Syst. 152 (2016) 97–106. https://doi.org/10.1016/j.chemolab.2016.01.014.

[38] M. Hanafi, E.M. Qannari, Nouvelles propriétés de l'analyse en composantes communes et poids spécifiques, J. La Soc. Française Stat. 149 (2008) 75–97.

[39] V. Cariou, D. Jouan-Rimbaud Bouveresse, E.M. Qannari, D.N. Rutledge, Chapter 7 - ComDim Methods for the Analysis of Multiblock Data in a Data Fusion Perspective, in: M.B.T.-D.H. in S. and T. Cocchi (Ed.), Data Fusion Methodol. Appl., Elsevier, 2019: pp. 179–204. https://doi.org/https://doi.org/10.1016/B978-0-444-63984-4.00007-7.

[40] S. Wiklund, E. Johansson, L. Sjöström, E.J. Mellerowicz, U. Edlund, J.P. Shockcor, J. Gottfries, T. Moritz, J. Trygg, Visualization of GC/TOF-MS-based metabolomics data for identification of biochemically interesting compounds using OPLS class models, Anal. Chem. 80 (2008) 115–122. https://doi.org/10.1021/ac0713510.

[41] C. Guijas, J.R. Montenegro-Burke, X. Domingo-Almenara, A. Palermo, B. Warth, G. Hermann, G. Koellensperger, T. Huan, W. Uritboonthai, A.E. Aisporna, D.W. Wolan, M.E. Spilker, H.P. Benton, G. Siuzdak, METLIN: A Technology Platform for Identifying Knowns and Unknowns, Anal. Chem. 90 (2018) 3156–3164. https://doi.org/10.1021/acs.analchem.7b04424.

[42] H. Horai, M. Arita, S. Kanaya, Y. Nihei, T. Ikeda, K. Suwa, Y. Ojima, K. Tanaka, S. Tanaka, K. Aoshima, Y. Oda, Y. Kakazu, M. Kusano, T. Tohge, F. Matsuda, Y. Sawada, M.Y. Hirai, H. Nakanishi, K. Ikeda, N. Akimoto, T. Maoka, H. Takahashi,

T. Ara, N. Sakurai, H. Suzuki, D. Shibata, S. Neumann, T. Iida, K. Tanaka, K. Funatsu, F. Matsuura, T. Soga, R. Taguchi, K. Saito, T. Nishioka, MassBank: A public repository for sharing mass spectral data for life sciences, J. Mass Spectrom. 45 (2010) 703–714. https://doi.org/10.1002/jms.1777.

[43] F. Puig-Castellví, L. Cardona, C. Bureau, D. Jouan-Rimbaud Bouveresse, C.B.Y. Cordella, L. Mazéas, D.N. Rutledge, O. Chapleur, Effect of ammonia exposure and acclimation on the performance and the microbiome of anaerobic digestion, Bioresour. Technol. Reports. 11 (2020) 100488. https://doi.org/10.1016/j.biteb.2020.100488.

[44] J. Bouhlel, D. Jouan-Rimbaud Bouveresse, S. Abouelkaram, E. Baéza, C. Jondreville, A. Travel, J. Ratel, E. Engel, D.N. Rutledge, Comparison of common components analysis with principal components analysis and independent components analysis: Application to SPME-GC-MS volatolomic signatures, Talanta. 178 (2018) 854–863. https://doi.org/10.1016/J.TALANTA.2017.10.025.