Comprehensive benchmark of integrative strategies for analyzing microbiome-metabolome relationships

Loïc Mangnier[1], Margaux Mariaz[1], Neerja Vashist[2], Alban Mathieu[1], Antoine Bodein[1], Marie-Pier Scott-Boyer[1], Matthew S. Bramble[3,4], Arnaud Droit[1,5,*]

[1] Centre de Recherche du CHU de Québec-Université, Laval, Université Laval, G1V 4G2, Québec, Canada
[2] Department of Pathology and Laboratory Medicine, UCLA, USA
[3] Center for Genetic Medicine Research, Children's Research Institute, Children's National Hospital, Washington, DC, USA
[4] Department of Genomics and Precision Medicine, The George Washington University of Medicine and Health Sciences, Washington, DC, USA
[5] Département de Médecine Moléculaire, G1V 0A6, Québec, Canada

Loïc Mangnier: loic.mangnier@crchudequebec.ulaval.ca
Margaux Mariaz: margaux.mariaz@gmail.com
Neerja Vashist: neerjavashist@ucla.edu
Alban Mathieu: alban.mathieu@crchudequebec.ulaval.ca
Antoine Bodein: antoine.bodein@crchudequebec.ulaval.ca
Marie-Pier Scott-Boyer: mariepier.scottboyer@crchudequebec.ulaval.ca
Matthew S. Bramble: mbramble@childrensnational.org
* corresponding author: arnaud.droit@crchudequebec.ulaval.ca

Abstract

Background
The exponential growth of high-throughput sequencing technologies was an incredible opportunity for researchers to combine different -omics within computational frameworks. In particular metagenomics and metabolomics data have gained an increasing interest due to their implication in many complex diseases. However, currently, no standard seems to emerge for jointly integrating both microbiome and metabolome datasets within statistical models.
Results
Thus, in this paper we comprehensively benchmarked fifteen different integrative methods to link microorganisms and metabolites. Methods evaluated in this paper cover most of the researcher's goals such as global associations, data summarization, individual associations and feature selection. Through an extensive simulation study and an application to real gut microbial datasets, we highlighted the best approaches for unraveling complementary biological processes involved between the two omics. We provided general guidelines for practitioners depending on the scientific question and the data at-hand.
Conclusion
In summary, we argue that this paper constitutes a promising avenue for establishing research standards when mutually analyzing metagenomics and metabolomics data, while providing foundations for future methodological developments.

Keywords: multi-omics, metagenomics, metabolomics, benchmark, statistical methods

Background

The recent development of high-throughput sequencing technologies has permitted the generation of omics data at an exponential scale. Combining different high dimensional biological datasets within computational models represents a wonderful opportunity for researchers to better understand the underlying biological mechanisms involved in diseases [1]. In particular, the microorganism-metabolite interactions have gained an increasing interest due to their potential involvement in a large set of traits. It has been demonstrated that shifts in the microbiome-metabolome interactions have important implications on individual health [2, 3]. Indeed, recent studies for cardio-metabolic diseases [4] or autism spectrum disorders [5] have shown that pathoetiology could be explained by a complex interplay between microbes and host metabolites [6] or by disruptions in the microbiota-derived metabolite processes [7]. Thus, efficiently incorporating microbiome and metabolome data within statistical frameworks offers critical insights on the complex relationships occurring between diet or lifestyle factors on the microbe-metabolite recomposition and remains an important challenge in order to adequately identify hence target biological pathways [8]. However, the tremendous amount of available statistical models makes the choice of the right method a daunting task for many researchers.

The statistical joint integration of microbiome and metabolome data can be achieved with different integrative strategies. Standard workflows include various types of analysis, each addressing a specific biological question [2]. Briefly, traditional pipelines include the detection of global associations, data summarization, individual associations and identification of core features. Firstly, researchers are often interested in determining whether a global association is occurring between the two omics. For example, one can look for a global change in metabolome levels due to a microbial recomposition induced by a specific diet or lifestyle [2]. Consistent with recent reports, testing for global associations can be performed using multivariate methods such as the Mantel test [9] or the multivariate microbiome regression-based kernel association test (MMiRKAT) [10]. This step frequently

precedes the application of subsequent analyses such as data summarization methods or the identification of core features [2]. Then, following approaches used for single omics, a common research objective is to summarize information contents in the two omics, facilitating the visualization and interpretation of large scale biological data [1]. The presence of two types of omics allows the exploitation of the intra- and inter- correlation existing between features of the two datasets. Application of data summarization methods including Canonical Correlation Analysis (CCA) [11], Partial Least Square (PLS) [12], Redundancy Analysis (RDA) [13] or more recently Multi-Omics Factor Analysis (MOFA2) [14] is an important step in order to uncover features explaining a large proportion of data variability. Indeed, applications of data summarization methods have allowed the identification of taxonomic groups or metabolites involved in Type 2 diabetes [15]. However, both global association and data summarization methods fail to provide individual relationships between one or several microorganisms and metabolites. This aspect remains central to highlight core features involved in a particular biological context. As an illustration, methods for detecting individual associations may prove relevant for the identification of bacterial genus associated with dietary-impacted metabolites [2]. One strategy is to compute a measure of association between each metabolite-microbiota pair, using either a correlation or a regression model. Although easily implementable and interpretable, these approaches suffer from lack of power induced by the number of models fitted, limiting result transferability. An alternative way is to employ univariate or multivariate feature selection methods to adequately identify key actors at a large scale. The least absolute shrinkage and selection operator (LASSO) is a method initially developed to improve predictability while proceeding to feature selection [16]. Indeed, the LASSO is able to set coefficients to zero, hence facilitating identification of core features. Consistently with this idea, sparse CCA (sCCA) [17] or sparse Partial Least Square (sPLS) [18] are multivariate penalized methods summarizing data variability while proceeding to feature selection. However, due to the complex structure of both microbiome and metabolome data, standard methods fall short of providing unbiased associations, limiting the biological interpretation of results.

On the one hand, because of the sequencing technology, metagenomics data highlight hard-to-analyze characteristics [19, 20]. Indeed, it is now globally accepted that microbiome datasets are over-dispersed, zero-inflated, highly correlated, and compositional. Without adequate transformation the inherent compositionality of the data makes the application of standard methods incorrect, leading to inconsistent results [19–21]. On the other hand, metabolomics data shares some of these features, exhibiting over-dispersion and high correlation structures [21]. Thus, combining these two omics together within statistical frameworks requires particular attention. Approaches to deal with compositional data either as an outcome or explanatory variable have already been proposed [20, 22, 23], covering applications of global association methods, data summarization, individual associations or identification of core features. Conventional strategies include utilization of standard methods after suitable data transformations or purely compositional approaches [24–27]. Subsequently, determining which strategy is the best depending on the research question remains an open problem with major implications for practitioners.

Despite recent efforts to integrate microbiome and metabolome within unified tools [28], to our knowledge there is no systematic framework to evaluate integrative methods to link microbiome with metabolome datasets; constantly pushing researchers to make their choice without any robust comparison. Thus, in this paper, we comprehensively benchmarked fifteen different integrative methods to link microorganisms and metabolites, covering most of the researcher's aims, such as global associations, data summarization, individual associations, or feature selection (Figure 1). Our extensive simulation studies provide insightful lessons on the strengths and limits of methods commonly encountered in practice. Then, we applied best methods to real data on the gut microbiome and metabolome for Konzo disease [29], highlighting a complex interplay between the two omics occurring at different scales. Finally, we provide general guidelines and avenues for future methodological developments, depending on the data at-hand and the research aims.

Results

SIMULATION SETUP AND BENCHMARKED METHODS

Taking advantage of the “Normal to Anything” (NORtA) framework, we generated synthetic microbiome and metabolome datasets mimicking complex data structures and relationships (See Methods). We produced two simulation settings, a low dimensional and a high dimensional setting, both representing different scenarios commonly encountered in practice (Figure 1A). We therefore compared fifteen integrative methods depending on the research question (Figure 1B). Methods were presented as follows. Firstly, in the *global associations* subsection we compared the Mantel test and MMiRKAT with respect to the Type-I error rate and power. Then, in the *data summarization* subsection we evaluated four different models including CCA, PLS, RDA and MOFA2, regarding their capability to recapitulate data variability across latent factors. Third, in the *individual associations* subsection we compared three strategies for performing regression-based approaches between compositional covariates and metabolites, the clr-linear model, the log-contrast and MiRKAT, respectively. Approaches were evaluated based on the Type-I error rate and power. Fourth, in subsections *univariate feature-selection for compositional predictors*, *univariate feature-selection for compositional outcomes* and *multivariate feature-selection* we compared approaches for identifying core microbes and metabolites, leveraging both univariate and multivariate feature selection strategies. For univariate frameworks, depending on the nature of the response, several models were considered. Indeed, when microorganisms are the explanatory variables, we compared three approaches, the clr-LASSO, the clr-MLASSO and CODA-LASSO [23]. Consistently, when microorganisms are the response variables, we evaluated the LASSO, MLASSO, and the sparse Dirichlet regression [27]. Nonetheless, for multivariate feature selection models, we considered sCCA and sPLS. Approaches were evaluated based on sparsity and reliability. Details on the methods and their related performance metrics were provided in the Methods section. Finally, in order to highlight complementary biological insights provided by methods, best approaches were illustrated in

the *real-data application* subsection, exploiting metagenomics and metabolomics data from Konzo disease.

GLOBAL ASSOCIATIONS

A common question in practice for researchers is to find global associations between two omics datasets [2]. Thus, we compared two multivariate methods detecting associations occurring at the global level between microbiome and metabolome, the Mantel test [9] and MMiRKAT [10], respectively. Since these two methods provide frequentist statistical frameworks i.e., p-values, we systematically evaluated their performance with respect to Type-I error rate control and power (See Methods). Firstly, when applying on the ILR transformed microbiome data, the Mantel test provides a good control of Type-I error rate in the high dimensional scenario while MMiRKAT highlights a slightly more conservative behavior (Figures 2A-2B). Secondly, MMiRKAT exhibits strikingly higher power than the Mantel test under our high dimensional scenario. Indeed, at the 0.05 significance threshold MMiRKAT reaches on average 97% of power against 22% for the Mantel test (Figures 2C). This difference is however mitigated in the low dimensional setting, where the two methods exhibit comparable performances (Figures S1-S2) . Importantly, the distance kernel choice seems to strongly impact the Mantel test power, from single to double, while MMiRKAT power remains stable across data transformations (Figure 2C). These results were confirmed in our low dimensional scenario and considering different data normalizations (Figures S3-S15). Interestingly, when the Mantel test was considered, the Canberra distance exhibits the lowest powers, while no clear distinction could be observed between Euclidean and Manhattan distance kernels (Figure 2C). This result suggests the Canberra distance as the poorest choice when using the Mantel test. Collectively, our results suggest comparable performance for the two methods under the low dimensional setting regarding both Type-I error rate and power. However, in the high dimensional scenario MMiRKAT is the most powerful method to find global associations. In addition the method is robust to data transformation and distance kernels.

DATA SUMMARIZATION

Instead of measuring one global association, one can be interested in recapitulating information contained within the two datasets through latent factors, accounting for the between- within-correlation [30] . Thus, we compared Canonical Correlation Analysis (CCA) [11], Regression PLS (PLS-Reg) [12], Canonical PLS (PLS-Can) [12], Redundancy Analysis (RDA) [13], and Multi-Omic Factor Analysis (MOFA2) [14] in our two scenarios with respect to their capability to summarize explained variance through their components (See Methods). Generally, regardless of the considered data normalization, in our two scenarios, MOFA2 was the best method, exhibiting larger explained variances, with a modest variability compared to PLS-Reg, PLS-Can, CCA, and RDA (Figure 2D; Figures S16-S19). Indeed, when ILR transformed microbiome data were considered, in our high dimensional scenario, MOFA2 exhibited an average of explained variance of 86% (sd = 1.37) compared to 44% (sd = 4.35), 14% (sd = 2.03), 21% (sd = 2.34), and 22% (sd = 0.76) for PLS-Reg, PLS-Can, CCA and RDA, respectively. Surprisingly, except for MOFA2 and the PLS-Reg, where the explained variances increase (64% to 86% and 41% to 44%, respectively), all the remaining methods exhibit a smaller explained variance in the high dimensional scenario compared to the low dimensional setting. Aligned with this result, we investigated the behavior of each method with respect to the number of associated features and the effect size and found positive associations in both cases across all methods (Figures S20-S21). Importantly, method performances may vary depending on the considered data transformation (Figure 2D; Figures S16-S19). Our results pointed to MOFA2 as the best model to summarize data variability through latent factors. Finally, our findings suggested that the method is versatile and robust under scenarios commonly encountered in practice.

INDIVIDUAL ASSOCIATIONS

Studying the relationship between metabolites and microorganisms may represent an important challenge in order to account for the compositionality induced by microbiome datasets. Indeed, the perfect correlation brought by the compositionality makes the

application of standard methods incorrect. This is particularly true when microbiota are incorporated as covariates [19, 22]. We therefore compared three equivalent strategies in order to study the global effect of microorganisms on one particular metabolite, the Log-contrast model [22], MiRKAT [10] and a linear regression on the CLR transformed microbiome (referred to as clr-lm), respectively. Methods were evaluated with respect to their capability to adequately control false positives while maintaining a good power (See Methods). Globally, under the null hypothesis, the three methods adequately controlled the Type-I error rate, with the linear log-contrast model exhibiting a slightly conservative behavior across the two scenarios (Figures 3A-3B). Then, under the alternative hypothesis, the linear log-contrast model offers a higher power than MiRKAT or the clr-lm model, on average twice larger across the data transformations considered in the high dimensional setting (Figure 3C). This result was also confirmed when comparing the log-contrast model to Spearman's or Pearson's correlations, while MiRKAT or the clm-lm model do not exhibit clear advantage (Figure S22). Indeed, at a 0.05 significance threshold, the log-contrast model offers 52% of power against 29% for MiRKAT and clr-lm, and 29% and 21% for Pearson's and Spearman's correlations, respectively. This result was confirmed in our low dimensional setting, where smaller discrepancies can be observed (Figure 3C). However, consistent with results observed for MMiRKAT, MiRKAT provided a stable power and a good control of Type-I error rate across data normalizations (Figure S23). Importantly, when evaluating individual association methods for compositional outcomes, we found no clear superiority of the Dirichlet regression or the linear regression on the CLR transformed microbiome data over Spearman's or Pearson's correlations in our low dimensional setting (Figure S24). Collectively, our results suggest that in order to study the global impact of microorganisms on individual metabolites, the linear log-contrast model represents the best method compared to competitor approaches, providing higher power and a suitable control of the Type-I error rate.

UNIVARIATE FEATURE-SELECTION FOR COMPOSITIONAL PREDICTORS

Feature selection methods have gained increasing interest from researchers in order to identify a subset of microbiota associated with a variable of interest [31]. However, due to the compositionality induced by microbiome data, traditional methods have been shown to lead to incorrect results [19]. Thus, we compared univariate feature selection methods accounting for compositional predictors, CODA-LASSO [23], clr-LASSO [23] and clr-MLASSO, respectively. Firstly, we evaluated whether methods were able to provide sparse sets of microorganisms across our two scenarios. In our low dimensional setting, CODA-LASSO highlighted sparser selections, showing average sparsities of 2% against 9% and 14% for clr-LASSO and clr-MLASSO. This result was consistent in our high dimensional setting, where CODA-LASSO showed stable sparsities, while the sparsity of clr-LASSO and clr-MLASSO greatly improves (Figures 4A-4D; CODA-LASSO=2%; clr-LASSO=5%; clr-MLASSO=11%). This result suggests that CODA-LASSO tends to provide a stable sparsity across our two scenarios, selecting only a small proportion of the total microorganism-metabolite interactions compared to the two other methods. Then, we assessed how accurate the methods are in order to find true associations. In the low dimensional scenario, clr-LASSO offered slightly higher classification performances, showing average F1-Scores of 43%, compared to 35% and 30% for CODA-LASSO and clr-MLASSO, respectively (Figure 4A). Nonetheless, in the high dimensional scenario, CODA-LASSO provided higher F1-Scores than clr-LASSO or clr-MLASSO, with accurate classification rates on average 1.40 times higher (Figure 4D). Collectively, our results point to CODA-LASSO as a good trade-off between sparsity and classification performances to accurately select sparse subset of microbiota associated with metabolites.

UNIVARIATE FEATURE-SELECTION FOR COMPOSITIONAL OUTCOMES

Finding a subset of metabolites associated with microbiota may bring important insights into the underlying biological mechanisms involved between the two omics. Thus, consistently

with the previous subsection, we systematically compared three different methods taking into account compositional outcomes with respect to sparsity and F1-Score, the sparse Dirichlet regression [27], LASSO and MLASSO of the CLR transformed microbiome data. Firstly, in the low dimensional setting, the LASSO offered strikingly sparser solutions, showing sparsity scores of 8% compared to 40% and 18% for the sparse Dirichlet regression and MLASSO, respectively (Figure 4C). Except for the sparse Dirichlet regression, where the sparsity was multiplied by roughly 2 between the two scenarios, LASSO and MLASSO exhibit sparser selection in the high dimensional setting compared to the low dimensional scenario (Figure 4D). This result suggests that standard methods applied on the CLR transformed microbiome data seems to provide sparse and consistent solutions across our scenarios. Moreover, regardless of the scenario considered, F1-Scores remained low across methods suggesting poor method performances to accurately classify associations between microorganisms and metabolites (Figures 4C-4D). However, it is worth mentioning that high F1-Scores achieved by the sparse Dirichlet regression in the low dimensional scenario may be explained by weak sparsity scores. Taken together, our results point to poor performance of methods to select accurately metabolites associated with microorganisms; where standard methods applied on the CLR transformed microbiome data correspond to a better trade-off between sparsity and classification performances than a purely compositional penalized method.

MULTIVARIATE FEATURE-SELECTION

Instead of analyzing each feature independently, exploiting information shared across two omics may represent an interesting avenue to select the most contributive features [32]. Thus, we compared three methods taking advantage of both intra- and inter-correlation occurring between features of the two datasets, the regression sparse PLS, the canonical sparse PLS [18] and the sparse CCA [17], respectively. Firstly, in our low dimensional setting the regression sPLS seems to provide high levels of sparsity compared to the two other methods (Figure 4C). Indeed, the method tends to select about 34% of total features

297 compared to 23% or 26% for sCCA or canonical sPLS. This pattern was also observed in our
298 high dimensional setting, even if an increase of sparsity between the two scenarios has to be
299 noted (Figures 4C-4F). This result aligns with a too high number of selected features, since
300 our simulation setup maximally assumes a 10% of associated features. Then, we
301 investigated whether methods were able to accurately discriminate contributive features from
302 uninformative ones. In our low dimensional scenario, the regression sPLS offered higher F1-
303 Scores, showing average values of 76% compared to 70% and 60% for the canonical sPLS
304 and sCCA, respectively (Figures 4C). This result was confirmed in the high dimensional
305 scenario, even if lower scores across the three methods have to be noted (Figures 4F). For
306 example, the average F1-Score for the regression sPLS decreased by 63%, while for the
307 canonical sPLS and sCCA, the decrease is of 53% and 69%, respectively, consistent with
308 lower classification performance as the dimensionality increases. Then, we investigated
309 whether methods are sensitive to data transformation. Interestingly, we found that in the low
310 dimensional scenario CLR transformation offered higher sparsity scores showing equivalent
311 F1-Scores across methods, while in the high dimensional setting absence of microbiome
312 data transformation slightly improved both sparsity and F1-Scores (Figure S25). Finally, our
313 results align with regression sPLS as the preferred choice for selecting features accounting
314 for between and within omics correlation. However, our findings point to modest levels of
315 sparsity across the methods suggesting poor method specificity with inconsistencies of
316 method results across data transformation.
317
318 **REAL-DATA APPLICATION**
319 Our systematic evaluation of strategies to jointly analyze microbiome and metabolome data
320 has permitted the illustration of the best methods depending on the research question. Thus,
321 through an application on metabolomics and metagenomics data of the Konzo disease [29],
322 we applied the more appropriate approaches to highlight different biological patterns
323 occurring between microorganisms and metabolites. We presented the exact workflow in the
324 Konzo data analysis section and Figure S26. Firstly, we used the Mantel test and found a

significant global association between the two omics (Spearman's permutation p-value <= 9.9e-5). Then we applied MOFA2 and found that through the fifteen first latent factors, the model explains 50% and 40% of microbiome and metabolome variability, respectively (Figure S27). Moreover, the top-10 most contributing features on the first factor highlighted relevant microbiota or metabolites associated with intestinal health. For example, MOFA2 identifies the *2,3-Dihydroxy-2-methylbutanoic acid*, a fatty-acid which has been demonstrated to be related to lipid metabolism pathways [33] (Figure 5A). Similarly, *Faecalibacterium prausnitzii* was identified as the most strongly associated microbiota, exhibiting a highly negative contribution (Figure 5B). This microbiota has already been shown to be involved in gut health [34, 35]. Subsequently we used the sPLS regression and were able to identify 249 metabolites and 70 microorganisms significantly contributing to the two first components, where clear clusters of microbiota could be observed (Figure 5C). Consistently with our benchmark, we used the log-contrast regression in order to identify metabolites significantly impacted by microbial communities and found that out of the 249 metabolites, 193 are significantly associated with microbial communities (Bonferroni adjusted p-values <= 2e-04). Then applying CODA-LASSO we detected 234 metabolites with at least one interaction with microorganisms. Interestingly, every microorganism has been selected at least once across the 234 metabolites, with an average of 35 microbiota associated (Figure 5D). For example, the *2,3-Dihydroxy-2-methylbutanoic acid*, previously identified by MOFA2, is associated with 8 microorganisms, mostly involved in gastrointestinal health (Figure 5E). Finally, we checked whether microorganisms exhibit consistent effects across metabolites and we observed 5 microbiota highlighting important variability in their effect (Figure 5F). This result was confirmed at a larger scale by network analysis from log-contrast regression and CODA-LASSO (Figures S28-S29). Our results from metagenomics and metabolomics data from Konzo disease highlight complementary biological interactions between microorganisms and metabolites, where different microbial dynamics seems to be involved.

Discussion

The integration of microbiome and metabolome datasets within statistical frameworks has become an important resource for researchers in order to comprehensively understand the underlying biological mechanisms involved in diseases. Indeed, recent studies in inflammatory bowel disease [36] or cardiometabolic traits [4] have highlighted that pathoetiology may result in disruptions of interactions between microorganisms and host-metabolites interplay or shifts in the microbial-derived metabolite levels. Understanding these interactions represent therefore a critical avenue for unraveling the biology of complex phenotypes. However, currently, there are no standards on how to integrate these two omics together, pushing researchers to constantly reinvent the wheel. Thus, deciding which method fits best for a specific biological question remains a daunting task, critically limiting the result interpretations and replicability. In this paper, we extensively benchmarked fifteen existent integrative methods to study microbiome-metabolome interactions covering most of the researcher aims: global associations, data summarization, individual associations, and feature selection (Figure 1). Based on a comprehensive simulation study and a real data application, we highlighted best methods depending on the research question and data at-hand, providing important insights about statistical good practices (Table 1) and avenues for future methodological developments (Table 2).

When evaluating global association methods, our results have pointed to important lessons for practitioners. Indeed, MMiRKAT represents the most promising method compared to the Mantel test, showing higher power and robustness to data transformations and distance kernels (Figure 2C). We argue this aspect is particularly relevant since choosing the right data transformation or distance metric may represent an important challenge for practitioners. Moreover, MMiRKAT has the possibility to adjust for confounding factors which is an appealing feature for most phenotypes where bias can be induced by

certain individual characteristics, such as age, sex or lifestyle [3, 4]. However, one limitation of MMiRKAT compared to the Mantel test is its incapability to deal with scenarios with a larger number of features than individuals. We therefore recommend filtering out features based on a feature selection approach or to use the Mantel test in order to have a crude idea about the global association occurring between the two omics. Importantly, when using the Mantel test, our results suggest that the Canberra distance on metabolome data is the poorest choice for detecting global associations across all our scenarios (Figures 2B; Figures S1-S15). Thus, applying Euclidean distance on transformed microbiome data while applying Euclidean or Manhattan distances on metabolites should constitute the default usage for most cases.

Data reduction is often used by practitioners in order to summarize information through a small number of components. Having an efficient method which recapitulates variability across two omics is critical for facilitating subsequent analyses such as visualization or clustering [1]. We considered four different methods exhibiting specific features to summarize omics information and found that in addition to being robust to data normalization, MOFA2 is the best method to recapitulate data variability. In our high dimensional setting MOFA2 explains about 80% of metabolome variance when ILR normalization was considered and remains stable across alpha and CLR transformations (Figure 2D, Figure S16). This result may be explained by the capability of the method to capture complex relationships, as suggested by [37]. Thus, we recommend using MOFA2 when researchers want to achieve efficient data reduction. We then applied MOFA2 to our Konzo dataset and found important microbiota and metabolites involved in biologically relevant pathways of gut health, while preserving a great portion of data variability (Figure 5A-5B). For example, MOFA2 identifies *Faecalibacterium prausnitzii* as the most negatively contributive microorganisms on the first factor (Figure 5B). Previous studies have shown that *Faecalibacterium prausnitzii* levels are strongly associated with anti-inflammatory metabolite quantities involved in intestinal health [34, 35]. Similarly, MOFA2 found *2,3-Dihydroxy-2-*

methylbutanoic acid with the strongest positive correlation on the first factor, a fatty-acid which has been demonstrated to be related to lipid metabolism pathways [33] (Figure 5A).

In practice another important question for researchers is to determine the relationship between microbial communities with a variable of interest [29, 38]. However, the underlying compositional structure of microbiome data is an important challenge for model performance. In this paper we have compared three methods accounting for the compositionality of predictors with different strategies, a linear regression applied on the CLR transformed microbiome data, MiRKAT, and the log-contrast model. Compared to correlations, these methods can incorporate confounding factors which is an important feature in practice. Our main finding is that regardless of the method considered here, better performances are achieved compared to correlations, still widely used in practice [5]. However, the linear log-contrast offers higher power across our simulation scenarios compared to MiRKAT and the linear regression (Figure 3C). Also, one important advantage of the log-contrast model over MiRKAT or the linear regression is to not require a choice of a particular data normalization, which can represent an important challenge for most researchers. This is particularly important since the CLR transformation has been shown to provide still-correlated features while sub-compositionally incoherent, limiting result transferability [23, 24]. This result highlights the need for new compositional data transformations, keeping the original number of features while linearly independent (Table 2). Hopefully, MiRKAT performance is robust across data transformations, with stable power and suitable Type-I error rate control (Figure 3C, Figure S23). Additionally, one main difference of the log-contrast or the linear regression over MiRKAT is to provide individual contribution of each microbe. We therefore strongly recommend to use the log-contrast regression when evaluating the association between microorganisms and metabolites. Consistently, out of the 249 metabolites selected by the regression sPLS, the log-contrast model highlights 193 metabolites with significant associations with microbiota in the Konzo dataset. Interestingly, we found that microorganisms exhibit heterogeneous effects across metabolites suggesting different microbial dynamics possibly involved in the disease (Figure S28). Similarly to MDiNe [39],

where authors provided a mechanistic framework to study differential microbial co-occurrence networks, additional work is needed to link microbiome and metabolome from a dynamic perspective at large scale (Table 2; Ongoing work). We argue this aspect is particularly critical in order to pinpoint the underlying biological mechanisms hence facilitating precision medicine applications [40, 41].

Also, one important contribution of this work is to extensively evaluate feature selection methods. This is particularly critical for researchers in order to accurately select metabolites and microorganisms involved in a specific biological context. Our results point to moderate performance of multivariate feature selection methods with inconsistent performances across scenarios and the data transformations considered (Figures 4C-4F, Figure S25). This result is also observed for univariate feature selection models with compositional outcomes (Figures 4B-4E). The best performances are achieved for methods with compositional predictors, with CODA-LASSO exhibiting stable sparsity results with good classification performances (Figures 4A-4D). Thus, we recommend in practice to use CODA-LASSO for scenarios with microbial predictors, while using the LASSO regression after CLR transforming the microbiome data when these latter are the outcome. Then we applied both regression sPLS and CODA-LASSO on the Konzo dataset. Regression sPLS has permitted the detection of 249 metabolites and 70 microorganisms contributing the most to data variability (Figure 5C). From these 249 metabolites, CODA-LASSO has subsetted the most contributing features, highlighting different microbial dynamics of effects (Figures 5F; Figure S29). Further investigations have shown that *Vescimonas fastidiosa* was the most interacting microbiota, significantly connected to 138 metabolites. This result is aligned with the model where microorganisms may be connected to a large set of metabolites. This complex microbiome-metabolome crosstalk has been shown to be associated with diseases [6]. However, associations found may result in artifact signals since most feature selection methods benchmarked in this paper suffer from lack of sparsity and reliability. This result is aligned with previous reports where authors have shown poor performances of traditional feature selection models [42]. Indeed, most penalized methods are mainly built upon cross-

validation where small perturbations in data may yield drastic changements in results. Similarly to [42] extending sparse multivariate or univariate methods to knockoff framework [43] or stability selection [44] should represent interesting avenues for improving both sparsity and reliability for compositional data [45] (Table 2).

Although our simulation setup is able to realistically simulate microbiome and metabolome data, our framework suffers from two limitations that we think it is important to mention here. Firstly, the NORtA algorithm is limited in its capability to generate real correlated compositional data. Indeed, as discussed by [46], simulating pure compositional data from a Dirichlet distribution induced only a small correlation between features, which is often unrealistic regarding the biology of the microbial communities and metabolites. We therefore generated compositional microbiome data post-hoc, promoting the correlation, zero-inflation and overdispersion characteristics over a purely compositional structure. This “hard” compositionality disturbed the original data structure but has several advantages especially in the data generating process (DGP). Indeed, through our simulation we are able to control underlying parameters while providing a DGP-agnostic procedure, not promoting one method over another. We argue that this aspect is central in order to provide systematic objective method comparisons. Also, as a parametric framework the NORtA algorithm is limited in its capability to simulate data with a higher number of microorganisms or metabolites than the number of individuals. Thus, as initially mentioned for global association methods, we suggest filtering out core elements using either an univariate or a multivariate method before using models assuming a sample size bigger than the number of features.

To summarize, in this paper we provide an extensive benchmark of integrative computational methods for incorporating metagenomics and metabolomics data. We hope this work will represent a great opportunity for the multi-omics community in order to improve research standards and practices. This aspect is central for scientific discovery and reproducibility.

Conclusions

In summary, the present study provides to the multi-omics community one of the largest comprehensive benchmarks of statistical frameworks to jointly integrate metagenomics and metabolomics data. Through an extensive simulation study, we systematically compared fifteen integrative approaches across most of the research questions encountered in practice. We identified the best methods and illustrated their capability to highlight complementary biological processes involved at different scales with an application to microbiome and metabolome data for Konzo disease. Overall, our study provides a robust and replicable comparative framework of integrative methods. We hope this work will serve as a foundation for setting research standards and the development of new efficient statistical models to mutually analyze metagenomics and metabolomics data.

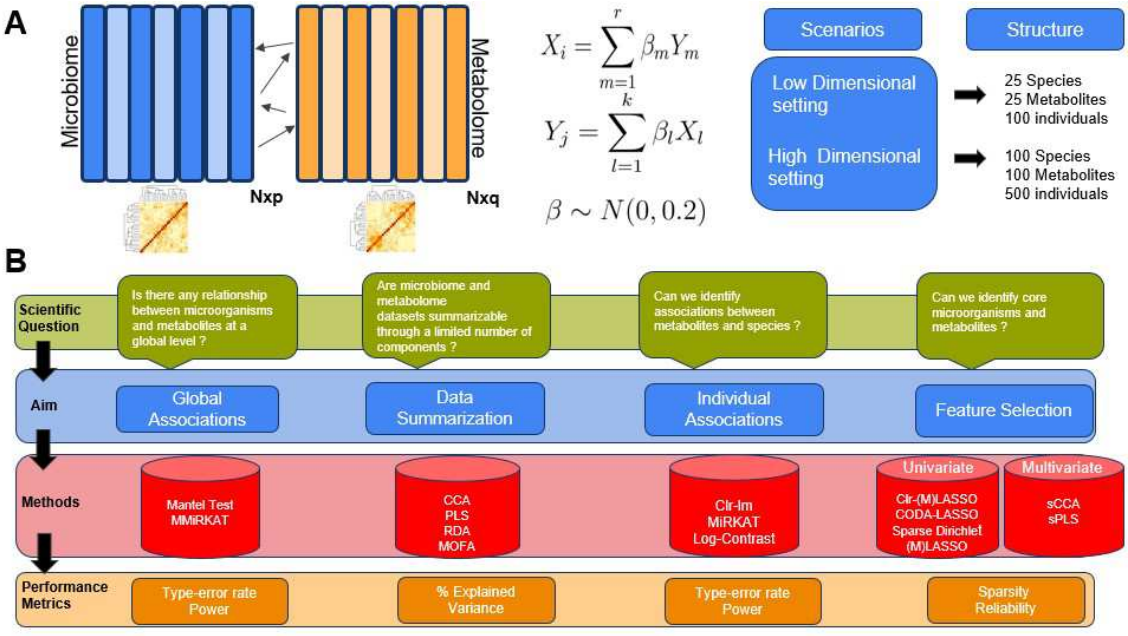
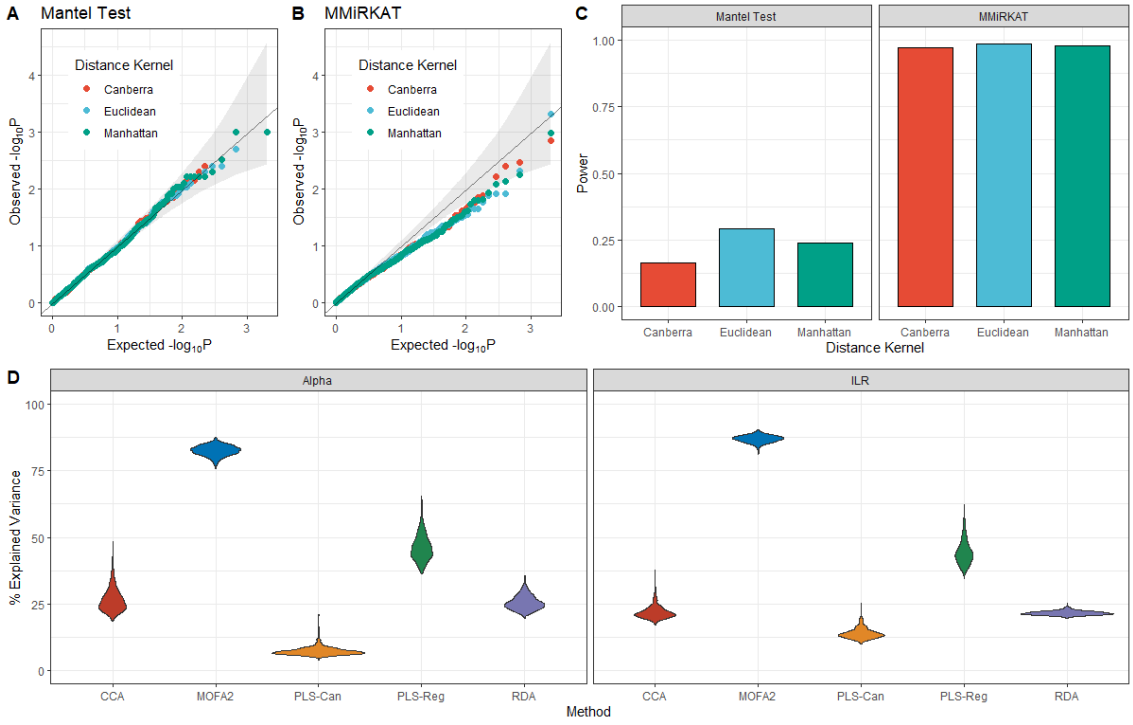


Figure 1 Overview of the simulation setup and integrative methods for analyzing microbiome-metabolome relationships depending on the research question

(**A**) Correlated microbiome and metabolome data were generated using the “Normal to Anything” framework (See Methods). Microbiome data were simulated considering a zero-

508 inflated negative binomial distribution, while metabolome datasets follow a negative binomial
509 distribution. For each dataset, proportions of associated features vary between 1% and 10%,
510 with association strengths randomly picked from a Gaussian distribution.
511 (**B**) Overview of the integrative methods related to the research question.



512

513 **Figure 2 Performance of the multivariate methods for both global association and data**
514 **summarization in the high dimensional scenario.**
515 When control of Type-I error rate is of interest, we are looking for methods providing
516 quantiles of observed p-values similar to quantiles of expected p-values, i.e., following the
517 diagonal line. In other words, the closer the dots to the straight line, the more the method
518 adequately controls the false positives. Similarly, for power, we are looking for methods
519 providing high powers. That is, detecting a significant association when we know there is an
520 association. Explained variance is the data variance contained through latent factors. See
521 Methods for details on performance metrics. **(A)** QQ-Plot of the Mantel test applied on the
522 ILR transformed microbiome and log transformed metabolome data, considering different
523 distance kernels for metabolites. Here we considered Spearman's method for computing the
524 global association between the two datasets. **(B)** QQ-Plot of MMiRKAT applied on the ILR

transformed microbiome and log transformed metabolome data, considering different distance kernels for metabolites. Points below the straight line refer to a conservative behavior in the result section. **(C)** Power of the Mantel test applied on the ILR transformed microbiome and log transformed metabolome data, considering different distance kernels for metabolites for both the Mantel test and MMiRKAT. P-values <= 0.05 were considered as significant. **(D)** Proportion of explained variance for the data summarization methods considering the log transformed metabolome and the the alpha transformed and ILR transformed microbiome data.

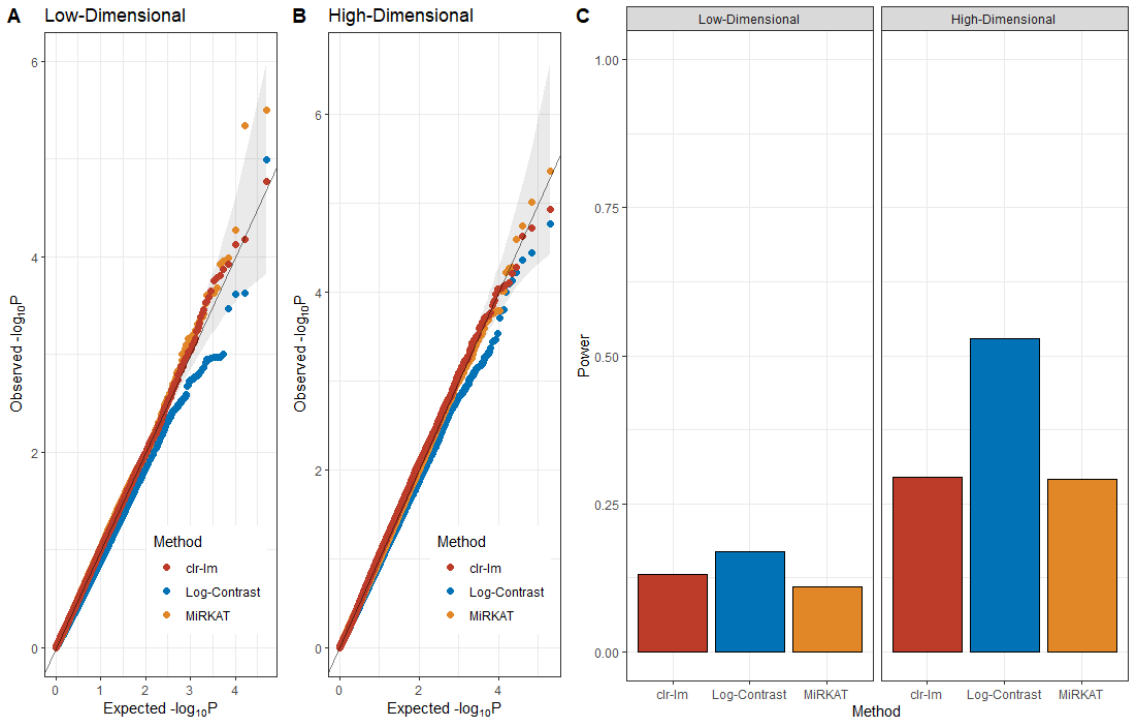
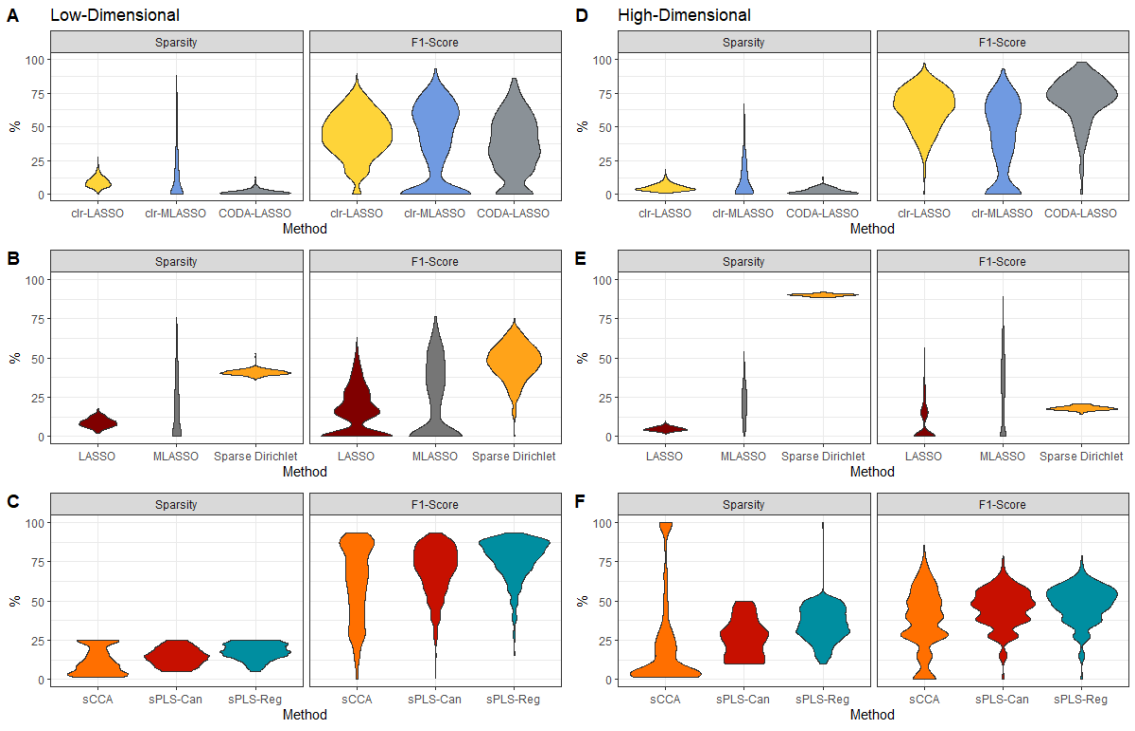


Figure 3 Performance of the individual association methods for compositional predictors

When control of Type-I error rate is of interest, we are looking for methods providing quantiles of observed p-values similar to quantiles of expected p-values, i.e., following the diagonal line. In other words, the closer the dots to the straight line, the more the method adequately controls the false positive. Similarly, for power, we are looking for methods

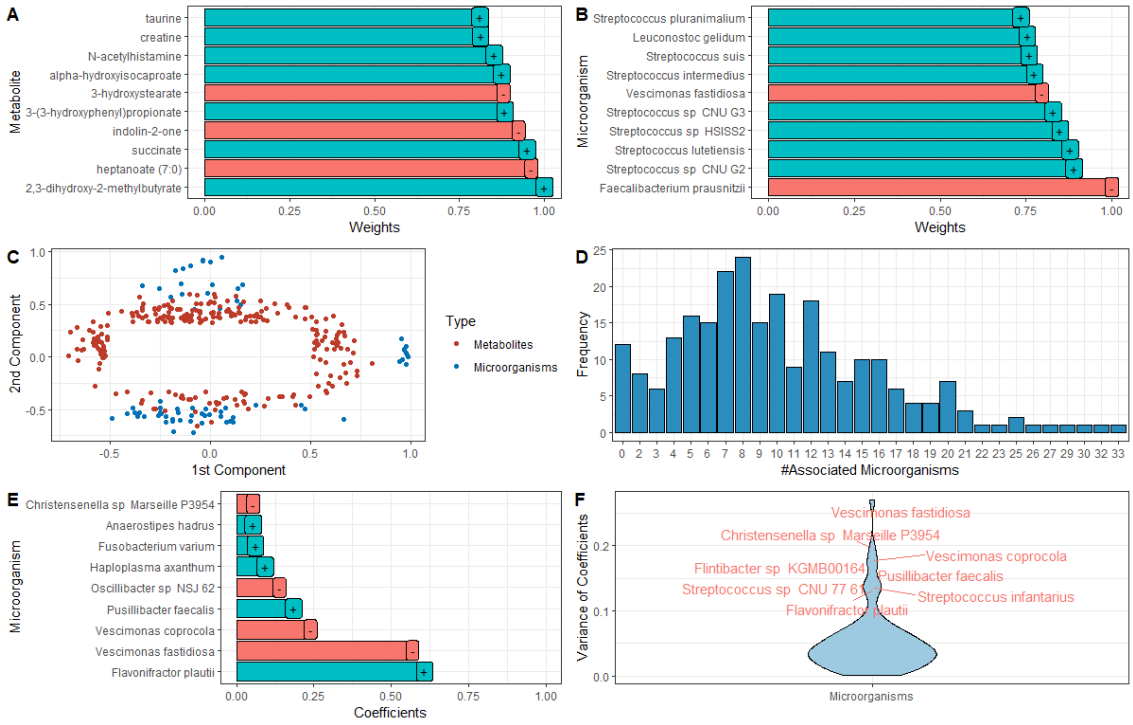
541 providing high powers. That is, detecting a significant association when we know there is an
542 association. See Methods for details on performance metrics.
543 QQplots of the individual association methods in (**A**) the low dimensional scenario and in the
544 (**B**) high dimensional scenario. (**C**) Power of the individual association methods across our
545 two main scenarios. P-values <= 0.05 were considered as significant. For the clr-lm method,
546 p-values were combined using ACAT [47] in order to provide similar comparisons with the
547 log-contrast regression and MiRKAT (See Methods). For MiRKAT, we reported Type-I error
548 rate and power using the ILR transformed microbiome data and the log transformed
549 metabolites.
550



551

552 **Figure 4 Performance of the feature selection methods for providing sparse and**
553 **reliable subset of elements across our two scenarios.**
554 Method performance was evaluated with respect to sparsity and F1-Score. For the former,
555 we are looking for methods providing low values corresponding to a small proportion of
556 selected features, while for the latter, high values of F1-Scores correspond to better
557 classification performances (See Methods).

558 Performance of univariate feature selection methods considering microorganisms as
559 covariates under our **(A)** low dimensional and **(D)** high dimensional scenarios. For CODA-
560 LASSO under the high dimensional setting performances were calculated on 100 replicates.
561 Performance of univariate feature selection methods considering metabolites as covariates
562 under our **(B)** low dimensional and **(E)** high dimensional scenarios. For the sparse Dirichlet
563 regression under the high dimensional setting performances were calculated on 100
564 replicates. Performance of the multivariate feature selection methods considering the CLR
565 transformed microbiome and the log transformed metabolome under our **(C)** low dimensional
566 and **(F)** high dimensional scenarios.



567

568 **Figure 5 Application of best strategies highlights complementary biological**
569 **interactions between microorganisms and metabolites in Konzo data**
570 Top-10 most contributing **(A)** metabolites and **(B)** microbiota on the first factor as identified
571 by MOFA2. Positive correlations were identified by a +, while negative correlations were
572 identified with a - sign **(C)** Projection of metabolites (red) and microorganisms (blue) into the
573 2D regression sPLS space. Features with null loadings were removed from the analysis. **(D)**
574 Distribution of the number of significant microorganisms found by CODA-LASSO across the

575 subset of metabolites identified by the regression sPLS. **(E)** Log-contrast coefficients for the
576 *2,3-Dihydroxy-2-methylbutanoic acid* **(F)** Violin plot of the variance of log-contrast coefficients
577 through the subset of microorganisms identified by the regression sPLS. Red dots
578 correspond to outliers with high coefficient's variability.

Scientific Question	Research Aim	Best Method	Pros	Cons
Is there any relationship between microorganisms and metabolites at a global level?	Global associations	MMiRKAT	Robust to data normalization and distance kernels Allow adjustment for covariates	Unable to deal with scenarios with higher number of features than individuals
Are microbiome and metabolome datasets summarizable through a limited number of components?	Data summarization	MOFA2	Robust to data normalization and distance kernels	Running time
Can we identify associations between metabolites and species?	Individual associations	Log-contrast	Compositional and sub-compositional consistent No need to data transformation Allow adjustment for covariates	Limited to few families of generalized linear models
Can we identify core microorganisms and metabolites?	Feature selection (univariate)	CODA-LASSO (compositional covariates)	Compositional and sub-compositional coherent No need to data transformation Allow adjustment for covariates	Limited to few families of generalized linear models
		LASSO (compositional outcomes)	Flexible framework Allow adjustment for covariates	Need a suitable data transformation
	Feature selection (multivariate)	sPLS	Flexible framework Efficiently account for within- between-	Tuning parameters

			correlation	

579 **Table 1: Summary of best methods depending on the research question**

Objective	Methods and Limits	Methodological avenues
Data normalization	- ***CLR***: provides still-correlated features in the original space - ***ILR***; ***Alpha***: are "black-box" transformations providing uncorrelated features in a restricted space	Data normalizations providing uncorrelated features in the original space for facilitating result interpretation
Mechanistic interpretation	- ***Log-contrast*; *CODA-LASSO*:** unable to provide a mechanistic view of modifications between microbiome and metabolome data	Network-based model to jointly study the modifications of microorganism and metabolite co-occurrence networks [39]
Feature selection	- **CODA-LASSO; Sparse Dirichlet; sPLS; sCCA:** lack of sparse solutions	Extension to knockoff framework [43] or stability selection [44] for improving feature selection performances

Table 2: Overview of avenues for future methodological developments to jointly analyze metagenomics and metabolomics data

Methods

Simulation setups

Microbiome and Metabolome data were simulated using the "*Normal to Anything*" approach (NORtA), already used for different multi-omics analyses [39, 46, 48]. An appealing feature of the NORtA algorithm is to provide a framework capable of simulating data from any marginal distribution while specifying arbitrary correlation structures. Thus, we are able to generate synthetic microbiome data respecting: (1) *correlation structure*, (2) *zero-inflation*, and (3) *over-dispersion*, while metabolome was generated similarly removing the zero-inflation property. This is consistent with real-data characteristics [21]. Moreover, we induced compositionality for microbiome data by dividing the count of each microorganism by the sum over all elements in a given individual. Several data transformations for both microbiome and metabolome were evaluated across our scenarios to account for data structure (See subsection Data and Distance Kernel Transformation). In order to evaluate the Type-I error control, we independently generated two datasets under the null hypothesis of no association between microorganisms and metabolites. Under the alternative hypothesis, we varied both the number of associations between microorganisms and metabolites and the strength of associations, mimicking microbiome-metabolome complex interdependence. Methods were compared under two main scenarios, simulating: (1) 25 microorganisms and 25 metabolites with 100 individuals and (2) 100 microorganisms and 100 metabolites with 500 individuals. Details on the simulation and sensitivity scenarios were provided in the supplementary. Under all scenarios we simulated 1,000 replicates. Simulation setup was summarized in Figure 1.

Data and Distance Kernel Transformation

Most methods used in practice need either a normalization step or a distance-based transformation in order to be applied properly on compositional or over-dispersed data [19]. Thus, we considered in our main analyses three data normalizations for microbiome and one data transformation for metabolome data. The choice of data normalization depends on the research objective.

In order to take into account the compositionality of microbiome data while keeping the original number of features, we considered the centered log-ratio transformation (CLR) [49] applied on the original count data. This normalization was considered across all the different considered methods. Basically, the CLR transformation computes the log ratio of each microbiota count on the geometric mean for a given individual. Formally, the CLR transformation is given by:

$$CLR(X_j) = \log(\frac{X_j}{g(X)})$$

where $g(X)$ is the geometric mean over all the microorganisms for one sample. This transformation projects the simplex onto a D compositional subspace under a zero-sum constraint [24, 50]. By keeping the original number of features the CLR transformation is a one-one transformation, facilitating result interpretation which is an appealing feature in practice. We therefore considered the CLR transformation as the reference normalization when individual associations or feature selection are of interest. However, the CLR transformation does not ensure independence between features and sub-compositionality coherence. This latter represents a major limitation for distance-based methods due to singular covariance matrices. Thus, when distance between features is of interest we considered the isometric log-ratio (ILR) [25] and alpha transformation [24]. Intuitively, these two transformations project the original D-dimensional space into an independent D-1 quasi-orthogonal space, the main difference laying into the transformation used. The ILR transformation projects the original data onto a Euclidean space. Formally:

632 $$ILR(X_j) = \sqrt{(\frac{j}{j+1})} \log(\frac{\prod_{j=1}^{D-1} X_j}{X_j + 1})$$

633 While the alpha transformation is a Box-Cox type transformation, where the transformed data
634 follow a multivariate distribution after a suitable alpha-transformation [24].
635 This facilitates the use of traditional multivariate methods. We therefore considered the ILR
636 and alpha transformations when evaluating global associations, and data summarization
637 methods, since the correspondence with the original features does not really matter.
638 Moreover, since the metabolome data have been shown to be log-normally distributed we
639 applied a natural log transform on the original count data [51].
640 Also, we applied different distance kernel transformations before performing some
641 global association or individual association analyses, highlighting different patterns of
642 relationships occurring among features. Briefly, we considered Euclidean, Canberra and
643 Manhattan distances on metabolome matrices of original and log transformed counts, while
644 considering the Euclidean distance on original and transformed microbiome data.
645 Interestingly, as presented by [19], the Euclidean distance applied on CLR transformed data
646 corresponds to the Aitchison distance. This latter has been shown superior to the Bray-Curtis
647 dissimilarity, representing a true linear relationship, while more stable to data subsetting or
648 aggregating [52], and will be considered as our reference method here. All data and distance
649 kernel transformations depending on the method used were summarized in Table S1.
650
651

652 ## *Statistical Analyses*

653 Let's assume X and Y, a matrix of microbiome and metabolome, collected on the same set of
654 samples, of size n x p and n x q, where n is the number of samples, p the number of
655 microbiota and q the numbers of metabolites, respectively. Xij represents the jth
656 microorganism in the ith sample, with j = 1,2,.., p, while Yik is the kth metabolite in the ith
657 sample, where k=1,2,...,q. For the sake of simplicity we considered the case where p=q.

Global Associations

In this paper we refer to global association methods, the statistical approaches providing global associations between microbiome and metabolome data (Figure 1). We considered two general methods, the Mantel test [9] and MMiRKAT [10], respectively.

The Mantel test [9] is a statistical framework measuring global correlation between two datasets measuring on the same set of samples. Traditionally, the Mantel test is applied on distance or dissimilarity matrices. Here we considered three different distance kernels applied on the metabolome dataset, Euclidean, Canberra and Manhattan distances. Also, we applied the Euclidean distance on the original and transformed microbiome matrix, since this projection leads to more natural interpretations [52] (Table S1). The Mantel test was applied considering either Pearson's or Spearman's correlation. P-values were obtained empirically based on permutations using 10,000 replicates. The Mantel test was performed using the *vegan* R package.

MMiRKAT is the multivariate extension of MiRKAT providing global association between a distance-transformed microbiome dataset and a low dimensional continuous multivariate phenotype [10]. Consistent with distance kernels used in the Mantel test, we considered Euclidean, Canberra and Manhattan distances applied on the original and transformed microbiome data, while the entire original or log transformed metabolome matrix was considered as the outcome (Table S1). MMiRKAT was applied using the *MiRKAT* R package.

Data Summarization

In this benchmark, we considered 4 distinct data summarization methods, encompassing CCA, PLS, RDA, and MOFA2. Briefly, all these methods seek to summarize data information through latent factors.

CCA initially proposed by [11] summarizes the relationship between two datasets by finding linear combinations of the two matrices maximizing the correlation. CCA was performed using the *CCA* R package.

Unlike CCA, PLS seeks for linear combinations maximizing the covariance between the two datasets [12]. Also, in PLS directionality of effect of one matrix on the other can be taken into account, leading to two general forms of PLS, regression and canonical, respectively [13]. Thus, canonical PLS and regression PLS were applied with the *mixOmics* R package.

Moreover, RDA is a two-step procedure, combining multivariate linear regression and PCA [13]. In the first step, a multivariate linear regression is fitted between each element of the matrix of responses and the matrix of predictors. Then a PCA is applied on the matrix of predicted values. RDA was performed using the *vegan* R package.

Finally, MOFA2 is an unsupervised multi-omics framework able to untangle sources of variability shared by different omics [14]. MOFA2 is a Bayesian probabilistic model able to find latent factors linking two omics by putting priors on model parameters. We applied MOFA2 using the related R package *MOFA2* with default parameters.

Except for MOFA2 where the best number of latent factors were chosen by the model, we kept all the components corresponding to the minimal number of features observed in one dataset.

Individual associations

When individual relationships are of interest, we consider different regression models taking into account the compositionality induced by microbiome data as predictors.

Indeed, for microbiota that are explanatory variables, we fitted 3 different models, a log-linear regression on the CLR transformed microbiome, a log-contrast model [22] and MiRKAT [10].

Formally the log-linear model of the CLR transformed microbiome (referred to as clr-lm in the Result section) is given by:

$$E(Y^*_{ik}|X^*_{ij},\beta_j) = \beta_0 + X^*_{ij}\beta_j + \epsilon_i, \forall(j,k)$$

where Y^* is the log transformed metabolome matrix and X^* the CLR transformed microbiome data. Although the compositionality in the microbiome data is taken into account using the CLR transformation, the previous model is not robust to the subset of microorganisms, not preserving the sub-compositionality feature of microbiome data. Thus, the log-contrast model by imposing a zero-sum constraint on regression coefficient preserves the scale invariance property needed to ensure the sub-compositionality characteristic of microbiome data [22]. Formally, the model is given by:

$$E(Y^*_{ik}|X_{i.},\beta) = X_{i.}\beta + \epsilon_i, \sum_{j=1}^{p}\beta_j = 0$$

Under the log-contrast framework, following [22] we applied the global significance F-test in order to determine whether there is an association between at least one microorganism and a given metabolite. The log-contrast model was performed using the *Compositional* R package. Aligned with the idea of global association, MiRKAT is a statistical framework exploiting semi-parametric kernel machine regression framework in order to summarize microbiome relationships [10]. One major feature of MiRKAT compared to other approaches is permitting the use of several distance kernels at the same time. This is particularly appealing since it is often unclear in practice which kernel is the more suitable. In our

734 context, we considered Euclidean, Canberra and Manhattan distances either on original or
735 transformed microbiome data, while considering the original or log transformed metabolome
736 as outcome. MiRKAT was applied with the *MiRKAT* R package.
737
738 ***Feature Selection: Univariate***
739 Adapted from [23] we considered two different models accounting for compositional
740 predictors, when fitting models with metabolites as outcomes. Firstly, we considered the
741 CLR-LASSO, performing the CLR transformation on microbiome data before fitting a
742 univariate or multivariate LASSO log-linear regression [16]. We referred to as LASSO and
743 MLASSO in the Results section. Formally for a metabolite k, the LASSO log-linear model is
744 given by:

$$\sum_{i=1}^{n}(Y_{ik}^{*} - \sum_{j=1}^{p} X_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

745
746
747 with Y^* is the log transformed metabolome matrix and X^* the CLR transformed
748 microbiome data. Best penalty parameters λ were chosen using a 10-fold cross-validation
749 through a 10 step grid-search from 0.01 to 1. LASSO or MLASSO models were fitted using
750 the *glmnet* R package.
751 Then, consistently with the log-contrast model, we applied the coda-LASSO
752 considering a log-linear response of the metabolome level. Briefly, the coda-LASSO is a
753 penalized log-contrast model, permitting to select only the most contributive features, with a
754 zero-sum constraint on regression coefficients, ensuring scale invariance, a property needed
755 for compositional data. The model considered in the coda-LASSO framework is a direct
756 extension of the model initially proposed by [53]. This latter fits a two-stage model on all
757 possible log-ratios between each pair of microbiota, leading to sparse solutions. The R

package *coda4microbiome* with the default parameters were used when applying coda-LASSO.

Then, following the same rationale, when fitting models with microorganisms as outcomes, we considered two different approaches, adjusting a univariate or multivariate LASSO linear model on the CLR transformed microorganisms or taking advantage of the sparse Dirichlet regression framework [27]. For the former, the model for the jth microorganism is given by:

$$\sum_{i=1}^{n}(X_{ij}^{*}-\sum_{k=1}^{p}Y_{ik}\beta_j)^2+\lambda\sum_{j=1}^{p}|\beta_j|$$

where X^{*} is the CLR transformed microbiome data. Here we considered the original and the log transformed metabolome signal as explanatory variables. In the sparse Dirichlet regression we used a multinomial dirichlet distribution. These models are direct extensions of the original LASSO model assuming X following a Dirichlet distribution [27]. Consistently with the methodology used in LASSO, best penalty parameters were chosen from a 0.1 step grid-search between 0.01 and 1 using a 10-fold cross-validation. Sparse Dirichlet regression framework was applied using the *MGLM* R package.

Feature Selection: Multivariate

Sparse Canonical Correlation Analysis (sCCA) [17] and sparse Partial Least Squares (sPLS) [18] are two penalized extensions of CCA and PLS permitting to summarize data information through latent factors while proceeding to feature selection.

For sCCA we used L1 penalty on the two datasets, only keeping features contributing on the two first components. Best penalty parameters were found using 25 permutation-based samples considering a 0.1-step grid search from 0.01 to 1. sCCA were performed using the *PMA* R package.

Consistently, canonical and regression sPLS were tuned using a 10-fold cross validation, considering a 5 step grid search ranging from 5 to 25 in our low dimensional

784 setting and from 10 to 50 in our high dimensional scenario. We maximally kept two
785 components in order to select the most contributive features. sPLS were applied using the
786 *mixOmics* R package. For both sCCA and sPLS, features on the two first components with
787 non-null loadings were considered as informative variables hence were kept to compute the
788 performance metrics.
789

790 ***P-value combinations***

791 In order to provide fair comparisons across our individual association methods with
792 compositional predictors, we combined p-values using the Aggregated Cauchy-based test
793 (ACAT) [47] when CLR-lm was considered. Indeed, for a large number of microbiota and
794 metabolites, applying univariate methods can lead to p x q possible correlations, limiting the
795 statistical power due to multiplicity. Similarly to the log-contrast model or MiRKAT, in practice
796 one can be interested in having the global association between one metabolite and several
797 microorganisms. Thus, in order to provide a powerful method controlling the Type-I error rate
798 well, we combined p-values for all microorganisms in a given metabolite using ACAT [47],
799 resulting from p p-values. We argue that this approach may result in more detected signals,
800 since the multiplicity burden is drastically reduced. Briefly, ACAT is a method combining p-
801 values through a Cauchy distribution.
802 Formally for one metabolite, the aggregated p-values across the p microbiota can be
803 approximated by:
804

805 $$0.5 - \frac{\arctan\left(\frac{T}{w}\right)}{\pi}$$

806 where $$T = \sum_{j=1}^{p} w_j \tan(\{0.5 - p_j\}\pi)$$.

807 One important feature of ACAT compared to other aggregation methods, such as Fisher's
808 method, is that the method can efficiently control the Type-I error rate even in presence of

809 correlated p-values, while maintaining good power [47]. Also, the method does not require
810 any resampling step, facilitating its application to large datasets.
811

812 ## *Performance Metrics*

813 Since all the methods considered in this benchmark exploit different statistical concepts, the
814 outputs cannot be directly compared. Consequently, we opted for several performance
815 metrics depending on the research question.
816 Indeed, for global and individual association methods, we systematically evaluated
817 model performance through Type-I error control and power, since the considered methods
818 are frequentist frameworks. Briefly, Type-I error control assesses whether a method provides
819 a good control of false positives at a given significance threshold. In other words, under the
820 null hypothesis of no association, at a significance threshold equals to 0.05 we maximally
821 expect 5% of false positives for a method that performs well. Type-I error control was
822 evaluated using the quantile-quantile plot of the -log10 of p-values. Similarly, the power is the
823 capability of a method to detect a significant signal (at a given significance threshold) when
824 we know that there is an association. In practice, researchers want methods maximizing the
825 power while accurately controlling the Type-I error.
826 Data summarization methods were compared based on the proportion of the
827 explained variance. We refer to explained variance, the amount of data variability kept by
828 latent factors built by methods.
829 Moreover, inspired from [42] when univariate *and* multivariate feature *selection*
830 methods were evaluated, we considered sparsity and reliability as primary performance
831 metrics. For univariate *methods* sparsity corresponds to the total number of relevant
832 associations found by the method (here with coefficients different from zero), while reliability
833 is the capability of a method to accurately discriminate true from false associations between
834 two features. However, we adapted both sparsity and reliability calculation when considering
835 multivariate feature selection methods. Indeed, sparsity was computed by the total number of

nonzero coefficients on the total number of features while reliability was adapted to capture the model performance to keep true contributive variables within the two datasets. Reliability was evaluated using the F1-Score (harmonic mean of the precision and recall). In practice, researchers are looking for sparse methods with high F1-Score. Performance metrics depending on the considered method were summarized in Figure 1. Technical details on the performance metric calculation and adaptations were provided in the supplementary. Methods.

Konzo data analysis workflow

Stool samples collected from individuals from study populations in Masi-Manimba (n = 65) and Kahemba (n = 106) regions of the Democratic Republic of the Congo were used for metagenomics and metabolomics assessment, where a proportion of the cohort is affected with Konzo. Shotgun metagenomics sequencing was performed on DNA extracted from ~250mg of stool with the goal of generating ~50 million reads per sample. Data was analyzed following similar methodology as described previously using Kracken2 and Bracken for taxonomic classifications [29] . Additionally, stool was analyzed by the company Metabolon, harnessing their large in-house repository of rigorously tested and validated metabolites that are used as reference, to detect metabolites present in the samples. Analysis was performed on the 1,098 microorganisms and 1,340 metabolites across the 171 individuals unconditionally of the disease status. Microbiome data at the genus level were normalized using the CLR transformation while metabolome data were log transformed. The workflow was as follows 1) global association, 2) data summarization 3) univariate and multivariate feature selection and 4) individual associations. Moreover, we considered microorganisms as explanatory variables and the microorganisms as outcomes. For global associations, since the number of features exceeds the number of individuals, we performed the Mantel test instead of MMiRKAT. We further discussed this aspect in the Discussion section. Then, we applied MOFA2 in order to detect the most contributing microorganisms

and metabolites on the first component. Following the same methodology as presented in the Method section, we extracted the core microorganisms and metabolites using the regression sPLS, keeping only the features with nonzero loadings on the two first components. We finally applied the log-contrast and CODA-LASSO in order to highlight contributions of microorganisms on metabolites. We summarized the workflow in Figure S26.

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

Codes to reproduce the analyses are available at: https://github.com/lmangnier/Benchmark_Integration_Metagenomics_Metabolomics. The simulated data are produced using the simulate_data.R script available in the same Github repository. R 4.2.2 is required to reproduce results from the paper. The metagenomics and metabolomics data for Konzo disease are available upon request from Matthew S. Bramble.

Competing interests

879 The authors declare no competing interests.

880 Funding

881 Not applicable

882

883 Authors' contributions

884 LM designed, conducted, performed the data analysis, and wrote the manuscript. MM
885 performed the data analysis. AM and NV wrote the manuscript. AB, MPSB, MSB, and AD
886 revised the manuscript. All authors read and approved the final version of the manuscript.

887

888 Acknowledgments

889 We would like to thank members of the Arnaud Droit Lab, particularly Louis-Maël Gueguen,
890 Thomas Jeanne, and Tania Cuppens for their insightful comments on the manuscript.

891

892

893

894

895

896

897

898

899 References

900 1. Rohart F, Gautier B, Singh A, Cao KAL (11 2017) mixOmics: An R package for 'omics
901 feature selection and multiple data integration. PLoS Comput Biol.

https://doi.org/10.1371/journal.pcbi.1005752

2. Tang ZZ, Chen G, Hong Q, Huang S, Smith HM, Shah RD, Scholz M, Ferguson JF (2019) Multi-omic analysis of the microbiome and metabolome in healthy subjects reveals microbiome-dependent relationships between diet and metabolites. Front Genet. https://doi.org/10.3389/fgene.2019.00454

3. Vernocchi P, Chierico FD, Putignani L (7 2016) Gut microbiota profiling: Metabolomics based approach to unravel compounds affecting human health. Frontiers in Microbiology. https://doi.org/10.3389/fmicb.2016.01144

4. Fromentin S, Forslund SK, Chechi K, et al (2 2022) Microbiome and metabolome features of the cardiometabolic disease spectrum. Nat Med 28:303–314

5. Dan Z, Mao X, Liu Q, et al (9 2020) Altered gut microbial profile is associated with abnormal metabolism activity of Autism Spectrum Disorder. Gut Microbes 11:1246–1267

6. Lee-Sarwar KA, Lasky-Su J, Kelly RS, Litonjua AA, Weiss ST (2020) Metabolome-Microbiome Crosstalk and Human Disease. Metabolites. https://doi.org/10.3390/metabo10050181

7. Lavelle A, Sokol H (2020) Gut microbiota-derived metabolites as key actors in inflammatory bowel disease. Nat Rev Gastroenterol Hepatol 17:223–237

8. Puig-Castellví F, Pacheco-Tapia R, Deslande M, Jia M, Andrikopoulos P, Chechi K, Bonnefond A, Froguel P, Dumas M-E (2023) Advances in the integration of metabolomics and metagenomics for human gut microbiome and their clinical applications. Trends Analyt Chem 167:117248

9. Mantel N (1967) The Detection of Disease Clustering and a Generalized Regression Approach. Cancer Res 27:209–220

10. Zhao N, Chen J, Carroll IM, Ringel-Kulka T, Epstein MP, Zhou H, Zhou JJ, Ringel Y, Li

H, Wu MC (5 2015) Testing in microbiome-profiling studies with MiRKAT, the microbiome regression-based kernel association test. Am J Hum Genet 96:797–807

11. Hotelljng H (1936) RELATIONS BETWEEN TWO SETS OF VARIATES. Biometrika 28:321–377

12. Abdi H (2010) Partial least squares regression and projection on latent structure regression (PLS Regression). Wiley Interdiscip Rev Comput Stat 2:97–106

13. Legendre, Pierre, Louis (2012) Numerical Ecology.

14. Argelaguet R, Velten B, Arnol D, Dietrich S, Zenz T, Marioni JC, Buettner F, Huber W, Stegle O (6 2018) Multi□Omics Factor Analysis—a framework for unsupervised integration of multi□omics data sets. Mol Syst Biol. https://doi.org/10.15252/msb.20178124

15. Al Bataineh MT, Künstner A, Dash NR, Alsafar HS, Ragab M, Schmelter F, Sina C, Busch H, Ibrahim SM (2023) Uncovering the relationship between gut microbial dysbiosis, metabolomics, and dietary intake in type 2 diabetes mellitus and in healthy volunteers: a multi-omics analysis. Sci Rep 13:17943

16. Tibshirani R (1996) Regression Shrinkage and Selection via the Lasso. J R Stat Soc Series B Stat Methodol 58:267–288

17. Witten DM, Tibshirani R, Hastie T (7 2009) A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. Biostatistics 10:515–534

18. Chun H, Kele¸s SK (2009) Sparse partial least squares regression for simultaneous dimension reduction and variable selection. J R Stat Soc Series B Stat Methodol 3–25

19. Gloor GB, Macklaim JM, Pawlowsky-Glahn V, Egozcue JJ (11 2017) Microbiome datasets are compositional: And this is not optional. Front Microbiol.

https://doi.org/10.3389/fmicb.2017.02224

20. Greenacre M (2021) Compositional Data Analysis. Annual Review of Statistics and Its Application 8:271–299

21. Xia Y, Sun J (2022) Statistical Data Analysis of Microbiomes and Metabolomics. American Chemical Society

22. Aitchison J, Bacon-Shone J (1984) Log contrast models for experiments with mixtures. Biometrika 71:323–353

23. Susin A, Wang Y, Cao KAL, Calle ML (6 2020) Variable selection in microbiome compositional data analysis. NAR Genomics and Bioinformatics. https://doi.org/10.1093/nargab/lqaa029

24. Tsagris M, Preston S, Wood ATA, Tsagris M, Preston S, Wood ATA (2016) Improved Classification for Compositional Data Using the α-transformation. J Classification 33:243–261

25. Egozcue JJ, Pawlowsky-Glahn V, Mateu-Figueras G, Barceló-Vidal C (2003) Isometric Logratio Transformations for Compositional Data Analysis. Math. Geol. 35:

26. Hijazi RH, Jernigan RW (2009) Modeling Compositional Data Using Dirichlet Regression Models. Journal of Applied Probability & Statistics 4:77–91

27. Chen J, Li H (3 2013) Variable selection for sparse Dirichlet-multinomial regression with an application to microbiome data analysis. Ann Appl Stat 7:418–442

28. Ni Y, Yu G, Chen H, Deng Y, Wells PM, Steves CJ, Ju F, Fu J (2020) M2IA: a web server for microbiome and metabolome integrative analysis. Bioinformatics 36:3493–3498

29. Bramble MS, Vashist N, Ko A, et al (12 2021) The gut microbiome in konzo. Nat

973 Commun. https://doi.org/10.1038/s41467-021-25694-1

974 30. Clos-Garcia M, Ahluwalia TS, Winther SA, et al (2022) Multiomics signatures of type 1
975 diabetes with and without albuminuria. Front Endocrinol 13:1015557

976 31. Li Y, Mansmann U, Du S, Hornung R (2022) Benchmark study of feature selection
977 strategies for multi-omics data. BMC Bioinformatics 23:412

978 32. Nguyen QP, Karagas MR, Madan JC, et al (2021) Associations between the gut
979 microbiome and metabolome in early life. BMC Microbiol 21:238

980 33. Watson AD (2006) Thematic review series: systems biology approaches to metabolic
981 and cardiovascular disorders. Lipidomics: a global approach to lipid analysis in biological
982 systems. J Lipid Res 47:2101–2111

983 34. Lopez-Siles M, Duncan SH, Garcia-Gil LJ, Martinez-Medina M (4 2017)
984 Faecalibacterium prausnitzii: From microbiology to diagnostics and prognostics. ISME
985 Journal 11:841–852

986 35. Sokol H, dicte Pigneur B né, Watterlot L, et al (2008) Faecalibacterium prausnitzii is an
987 anti-inflammatory commensal bacterium identified by gut microbiota analysis of Crohn
988 disease patients. Proceedings of the National Academy of Sciences 105:16731–16736

989 36. Ning L, Zhou YL, Sun H, et al (12 2023) Microbiome and metabolome features in
990 inflammatory bowel disease via multi-omics integration analyses across cohorts. Nat
991 Commun. https://doi.org/10.1038/s41467-023-42788-0

992 37. Cantini L, Zakeri P, Hernandez C, Naldi A, Thieffry D, Remy E, Baudot A (2021)
993 Benchmarking joint multi-omics dimensionality reduction approaches for the study of
994 cancer. Nat Commun 12:124

995 38. Yang L, Chen J (2022) A comprehensive evaluation of microbial differential abundance
996 analysis methods: current status and potential solutions. Microbiome 10:130

997 39. McGregor K, Labbe A, Greenwood CMT (2020) MDiNE: A model to estimate differential
998 co-occurrence networks in microbiome studies. Bioinformatics 36:1840–1847

999 40. Petrosino JF (2 2018) The microbiome in precision medicine: The way forward. Genome
1000 Med. https://doi.org/10.1186/s13073-018-0525-6

1001 41. Talmor-Barkan Y, Bar N, Shaul AA, et al (2 2022) Metabolomic and microbiome profiling
1002 reveals personalized risk factors for coronary artery disease. Nature Medicine 28:295–
1003 302

1004 42. Hédou J, Marić I, Bellan G, et al (2024) Discovery of sparse, reliable omic biomarkers
1005 with Stabl. Nat Biotechnol. https://doi.org/10.1038/s41587-023-02033-x

1006 43. Candès E, Fan Y, Janson L, Lv J (2018) Panning for Gold: “Model-X” Knockoffs for High
1007 Dimensional Controlled Variable Selection. J R Stat Soc Series B Stat Methodol
1008 80:551–577

1009 44. Meinshausen N, Bühlmann P (2010) Stability selection. J R Stat Soc Series B Stat
1010 Methodol 417–473

1011 45. Srinivasan A, Xue L, Zhan X (2021) Compositional knockoff filter for high-dimensional
1012 regression analysis of microbiome data. Biometrics 77:984–995

1013 46. Hawinkel S, Mattiello F, Bijnens L, Thas O (1 2019) A broken promise: Microbiome
1014 differential abundance methods do not control the false discovery rate. Brief Bioinform
1015 20:210–221

1016 47. Liu Y, Chen S, Li Z, Morrison AC, Boerwinkle E, Lin X (3 2019) ACAT: A Fast and
1017 Powerful p Value Combination Method for Rare-Variant Analysis in Sequencing Studies.
1018 Am J Hum Genet 104:410–421

1019 48. Wang Y, Cao KAL (3 2023) PLSDA-batch: a multivariate framework to correct for batch
1020 effects in microbiome data. Brief Bioinform. https://doi.org/10.1093/bib/bbac622

1021 49. Aitchison J (1986) The statistical analysis of compositional data (Monographs on
1022 statistics and applied probability). Chapman and Hall

1023 50. Tsagris MT, Preston S, Wood ATA (2011) A data-based power transformation for
1024 compositional data. arXiv [stat.ME]

1025 51. Antonelli J, Claggett BL, Henglin M, et al (7 2019) Statistical workflow for feature
1026 selection in human metabolomics data. Metabolites.
1027 https://doi.org/10.3390/metabo9070143

1028 52. Aitchison J, Barceló-Vidal C, Martín-Fernández JA, Pawlowsky-Glahn V (2000) Logratio
1029 Analysis and Compositional Distance 1. Math. Geol. 32:

1030 53. Bates S, Tibshirani R (2019) Log-ratio lasso: Scalable, sparse estimation for log-ratio
1031 models. Biometrics 75:613–624