Published online 21 July 2020

NAR Genomics and Bioinformatics, 2020, Vol. 2, No. 3 **1**
doi: 10.1093/nargab/lqaa050

Model-based joint visualization of multiple compositional omics datasets

Stijn Hawinkel[1,*], Luc Bijnens[2,3], Kim-Anh Lê Cao[4] and Olivier Thas[1,3,5]

[1]Department of Data Analysis and Mathematical Modelling, Ghent University, 9000 Ghent, Belgium, [2]Quantitative Sciences, Janssen Pharmaceutical companies of Johnson and Johnson, 2340 Beerse, Belgium, [3]Data Science Institute, I-BioStat, Hasselt University, 3500 Hasselt, Belgium, [4]Melbourne Integrative Genomics, School of Mathematics and Statistics, University of Melbourne, 3010 Melbourne, Victoria, Australia and [5]National Institute for Applied Statistics Research Australia (NIASRA), University of Wollongong, 2500 Wollongong, New South Wales, Australia

Received December 01, 2019; Revised May 20, 2020; Editorial Decision June 30, 2020; Accepted July 05, 2020

ABSTRACT

The integration of multiple omics datasets measured on the same samples is a challenging task: data come from heterogeneous sources and vary in signal quality. In addition, some omics data are inherently compositional, e.g. sequence count data. Most integrative methods are limited in their ability to handle covariates, missing values, compositional structure and heteroscedasticity. In this article we introduce a flexible model-based approach to data integration to address these current limitations: COMBI. We combine concepts, such as compositional biplots and log-ratio link functions with latent variable models, and propose an attractive visualization through multiplots to improve interpretation. Using real data examples and simulations, we illustrate and compare our method with other data integration techniques. Our algorithm is available in the R-package *combi*.

INTRODUCTION

With the latest advances in high-throughput technologies, an increasing number of omics data types is arising that require statistical analysis and data integration tools. These tools must be tailored to the data types under study, whilst being user-friendly with fast computation and easily interpretable results. Here we define data integration as the combination of at least two different types of biological measurements ('views') carried out on the same samples. The underlying idea is that their common origin engenders some relationship between the measurements, i.e. the biological state of the organism is reflected from the different views. The goals of such data integration can be very diverse. In this study, we adopt an explorative approach to unearth patterns that extend across different datasets, and identify relationships between features from different views. Using dimension reduction and visualization that focus on the strongest biological patterns in several high-dimensional datasets, our aim is to give researchers a first insight into the data structure and to highlight sample clusters and feature relationships that can be further investigated in follow-up studies.

The simplest way to integrate data is by simply concatenating the different data matrices measured on the same samples, then analyse this data matrix using classical techniques. Its simplicity notwithstanding, the risk is to mix data types that are heterogeneous and thus fail to account for their differences in distribution and noise levels. Other contemporary tools rely on variance partitioning, correlations or other measures based on sums of squares. A common approach is canonical correlation analysis (cca), which finds linear combinations of variables with maximal correlation between views (1), whilst partial least squares (pls) finds linear combinations with maximal covariance (2). These methods, however, can be ill-suited to deal with many types of omics data, such as overdispersed sequence count data. Moreover, these methods rely on the singular value and eigenvalue decompositions that require imputation of missing values and lack flexibility to include covariates, such as patient baseline.

Latent variable models have recently gained traction for the analysis of genomics data (3–8). They are based on the principle that a low number of latent sample variables can capture the most important signals in the dataset. These latent variables are included as components in a regression model, and thus enable great flexibility to specify any outcome distribution and include observed sample-specific variables. Moreover, their estimation naturally handles missing values. When the latent variables are orthogonal, they can be plotted in multiplots to effectively represent the whole model in a single graph (3,5,8).

A popular way to obtain interpretable results from high-dimensional datasets is to enforce *sparsity* during the data

*To whom correspondence should be addressed. Tel: +32 9 264 61 79; Email: stijn.hawinkel@ugent.be

integration process (9–11), therefore assuming that only a small fraction of features carry meaningful signal. Usually ℓ_1-norm penalties are imposed on the feature parameters to set the loadings of many features to zero (e.g. the 'lasso'). In addition to potential computational difficulties, this approach may struggle with correlated features. In this paper, we address instead the interpretability problem by visualizing features with the strongest signal using thresholding. All features are included in the model, so that loadings of the features with weaker signal can then still be consulted.

Sequence count data quantify the composition of mixtures of nucleic acids, for instance in transcriptomics and microbiomics. The resulting outcomes are integers, but with high variance and many zero observations. As a result, classical count models such as the Poisson and negative binomial distributions do not provide an accurate fit to such data (12–16). However, it is crucial to account for the particular mean-variance structure of sequence count data (17–19). Since the total number of sequences obtained in a particular experiment is mostly unrelated to the original biomass or number of cells, sequence count data are compositional, as are many other types of omics measurements (20). This means that they only contain information on the composition of a nucleic acids mixture (and hence of the specimen under study), and not on its total size or concentration. As a result, changes in one feature affect the proportion of at least another feature. The statistical analysis should take this dependence into account. The most common approach for compositional data is to perform a log-ratio transformation (21), then consider the transformed data as homoscedastic normal. Applying this approach to sequence count data yields to two limitations. Firstly, sequence count data (and count data in general) have a strong mean-variance relationship: features with high means have highly variable counts. The information on these variances is lost in the log-ratio transform. As a result, analysis of the log-ratio transformed counts may be affected by artefacts which are unrelated to the composition, such as library sizes (3). Secondly, sequence count data often have high zero frequencies. This is particularly the case for microbiome and single cell RNA-sequencing data. As logarithms of zero and division by zero are undefined, some imputation is usually applied to replace these zeroes by pseudocounts. An extensive theoretical framework was proposed to impute these zeroes (22), but this imputation results in a data matrix that is a mixture of observed counts and inferred pseudocounts. The uncertainty in the latter is usually ignored in the subsequent analysis. A recent approach proposed to use log-ratio transforms as dedicated link functions for compositional data in a regression model. The parameters are then inverse transformed to obtain the mean model of a composition, rather than transforming the data (23–25). This alternative addresses the two limitations mentioned above, as it allows for the model to be augmented by a suitable count distribution that is appropriate for zero counts and overdispersion. However, this approach has not yet been widely adopted for the analysis of compositional data.

In this work, we combine latent variable modelling and log-ratio link functions with innovative mean-variance modelling to obtain a new model for the integration of multiple views. Our model, called COMBI (Compositional Omics Model-Based Integration) is embedded into a regression framework and can easily incorporate sample-specific variables. Thanks to appropriate parameter restrictions, the final data integration model can be represented graphically in multiplots to facilitate data exploration of multiple views simultaneously. The explicit formulation of the estimating equations unlocks classical diagnostics tools, and allows the model to be fitted on datasets with missing data. We compare our proposed approach with existing approaches, and discuss pros and cons in a simulation study and in real data analysis.

MATERIALS AND METHODS

Model description

Data structure. Suppose at least two different data matrices or *views*, $\mathbf{X}_{(n \times p)}$ and $\mathbf{Y}_{(n \times q)}$ with p and q features, have been measured on the same source material or individuals (n samples). In practice, there is no limit to the number of views that can be included in the analysis, but we explain the concept here with only two matrices. An additional design matrix of sample-specific variables $\mathbf{c}_{(n \times d)}$ may also be available. Contrary to $\mathbf{X}$ and $\mathbf{Y}$, the elements of $\mathbf{c}$ are not treated as random variables.

Basic latent variable model. The core of our data integration model is based on a set of mean models sharing a set of latent variables, comparable to the *MOFA* model (4):

$$g_x\big[E(\mathbf{X}|\mathbf{Z})\big] = \mathbf{U}_x + \mathbf{Z}\boldsymbol{\Gamma} \tag{1}$$

$$g_y\big[E(\mathbf{Y}|\mathbf{Z})\big] = \mathbf{U}_y + \mathbf{Z}\boldsymbol{\Theta}, \tag{2}$$

where g_x and g_y are link functions defined according to the data type. $\mathbf{U}_x$ and $\mathbf{U}_y$ are offset matrices that correct for baseline differences, e.g. array intensity or sequencing depths. They define an *independence model* where all samples have an identical composition (i.e. the feature composition is independent of the sample). $\mathbf{Z}_{(n \times M)}$ is a low dimensional matrix of sample scores on M latent variables and $\boldsymbol{\Gamma}_{(M \times p)}$ and $\boldsymbol{\Theta}_{(M \times q)}$ are view-wise parameter matrices. High values for the sample scores indicate samples that differ strongly from the average sample, whereas large loadings in the parameter matrices indicate features that discriminate between these samples. M is usually set to 2 or 3 in view of making interpretable multiplots (as described below). Restrictions are needed to render this model identifiable. In particular, the latent variables are restricted to be orthogonal: $\mathbf{Z}^T\mathbf{Z} = \text{diag}(\boldsymbol{\psi})$ with diag() defining a diagonal matrix with $\boldsymbol{\psi}$ the non-negative diagonal entries. The coefficient matrices are restricted to be orthonormal: $\boldsymbol{\Gamma}\boldsymbol{\Omega}_x\boldsymbol{\Gamma}^T = \boldsymbol{\Theta}\boldsymbol{\Omega}_y\boldsymbol{\Theta}^T = \mathrm{I}_M$, with $\boldsymbol{\Omega}_x$ and $\boldsymbol{\Omega}_y$ view-specific, diagonal weight matrices (see Supplementary Section 1.1.3) and I_M the identity matrix of dimension M. Hence our base mean model is identical to *MOFA* but with several substantial improvements: the link functions are better suited for compositional data, the outcome distributions are better suited for sequence count data and the restrictions imposed on the parameters allow to output insightful multiplots to ease interpretation. In addition, since the dimensions are fitted sequentially, the estimates of lower dimensions do not depend on the total number of dimensions required.

Sample-specific variables. There are two ways to include sample variables into the analysis, either by considering them as confounders (e.g. batch or sequencing center) and filter out their effect, or by examining them explicitly and interpret their biological relationship (a constrained analysis) (3). In the case of confounding variables, their effect is eliminated by conditioning on them prior to the estimation of the latent variables. Let $\mathbf{R}$ and $\mathbf{S}$ denote the design matrices of the confounding variables in views $\mathbf{X}$ and $\mathbf{Y}$. We define the following model:

$$g_x\big[E(\mathbf{X}|\mathbf{Z},\mathbf{R})\big] = \mathbf{U}_x + \mathbf{R}\boldsymbol{\Phi} + \mathbf{Z}\boldsymbol{\Gamma} \tag{3}$$

$$g_y\big[E(\mathbf{Y}|\mathbf{Z},\mathbf{S})\big] = \mathbf{U}_y + \mathbf{S}\boldsymbol{\Xi} + \mathbf{Z}\boldsymbol{\Theta}. \tag{4}$$

where $\boldsymbol{\Phi}$ and $\boldsymbol{\Xi}$ are parameter matrices. The estimates of $\mathbf{Z}$, $\boldsymbol{\Phi}$ and $\boldsymbol{\Gamma}$ will then be free of the effect of the confounders.

In case of a constrained analysis, the sample scores in $\mathbf{Z}$ are no longer unrestricted, but become metavariables that are linear combinations of the observed sample variables. In particular $\mathbf{Z} = \mathbf{c}\boldsymbol{\Lambda}$, where $\boldsymbol{\Lambda}$ is a $d \times M$ matrix with the *environmental gradients* (3,26) in the columns. Each environmental gradient consists of loadings that reflect the importance of the sample variables in shaping both views and their relationship with the features. The gradients are restricted to be orthonormal: $\boldsymbol{\Lambda}^t\boldsymbol{\Lambda} = \mathbf{I}_M$. If the design includes categorical variables, then $\mathbf{c}$ will be constructed with dummy (indicator) variables for every level without using a reference level. The elements of $\boldsymbol{\Lambda}$ corresponding to the dummy variables of the same categorical variable are restricted to have zero sum. This arrangement avoids dependence on the choice of reference variable in view of the normalization restriction above. It also leads to informative plots with all levels of the categorical variables shown, without hidden reference levels. The continuous variables in $\mathbf{c}$ are normalized to have standard deviation 1.

Compositionality. The link function determines the range of the expected outcomes of a regression model. For compositional views, we choose the centered log-ratio transform (clr) as a link function, which is defined as:

$$\mathrm{clr}(\mathbf{x}) = \log\left(\frac{x_1}{(\prod_{j=1}^{p} x_j)^{1/p}}, ..., \frac{x_p}{(\prod_{j=1}^{p} x_j)^{1/p}}\right). \tag{5}$$

The inverse transformation (clr^{-1}, also known as the *softmax*) is defined as:

$$\mathrm{clr}^{-1}(\mathbf{x}) = \left(\frac{\exp(x_1)}{\sum_{j=1}^{p}\exp(x_j)}, ..., \frac{\exp(x_p)}{\sum_{j=1}^{p}\exp(x_j)}\right). \tag{6}$$

The result of this inverse transformation is a vector with values between 0 and 1 that sum to 1, i.e. a *composition*. The advantage of the clr over other log-ratio transforms is that it allocates a single parameter to each feature, which is a crucial property for making interpretable biplots (25,27). The mean model for the outcome of feature j in sample i is then:

$$\begin{aligned} E(X_{ij}|\mathbf{Z}_i) &= \left[\mathrm{clr}^{-1}(\mathbf{u}_x + \mathbf{Z}_i^t\boldsymbol{\Gamma})\right]_j s_i \\ &= \pi_{ij}(\boldsymbol{\pi}^{\text{indep}}, \mathbf{Z}_i, \boldsymbol{\Gamma}) s_i, \end{aligned} \tag{7}$$

with s_i an estimate of the baseline sample intensity in sample i (e.g. the sequencing depth), $\boldsymbol{\pi}^{\text{indep}} = \mathrm{clr}^{-1}(\mathbf{u}_x)$ the proportion vector under the independence model and $\pi_{ij}(\boldsymbol{\pi}^{\text{indep}}, \mathbf{Z}_i, \boldsymbol{\Gamma})$ the j-th feature proportion under the full model. This model can be regarded as follows: each dimension m 'perturbs' the lower dimensional composition of sample i with direction $\boldsymbol{\Gamma}_m$ and strength $\mathbf{Z}_{im}$ to form a new composition, which also sums to 1 (23,24). The perturbation operator $\oplus$ is defined for a composition $\mathbf{u}$ and a strictly positive vector $\mathbf{v}$ as proposed by Aitchison (21):

$$\mathbf{u}\oplus\mathbf{v} = \frac{1}{\sum_{j=1}^{p} u_j v_j}\Big(u_1 v_1, ..., u_p v_p\Big). \tag{8}$$

The composition of sample i is then (23):

$$\boldsymbol{\pi}_i = \mathrm{clr}^{-1}(\mathbf{u}_\mathbf{x})\oplus\mathrm{clr}^{-1}(\boldsymbol{\Gamma}_1)^{\mathbf{Z}_{i1}}\oplus\ldots\oplus\mathrm{clr}^{-1}(\boldsymbol{\Gamma}_M)^{\mathbf{Z}_{iM}}. \tag{9}$$

So far we have only specified mean models. To allow these models to be estimated, more information is needed on the outcome distributions. This can be done by specifying a parametric distribution, or by specifying a variance model as described in the next subsection.

Quasi likelihood estimation for sequence count data. Count distributions that are appropriate for sequence count data are not available, hence we chose to estimate the parameters of these views through quasi-likelihood. Quasi-likelihood is a semiparametric estimation technique, whereby only the mean and the variance of an outcome are modelled, and the higher moments are left unspecified (28). The estimating equations for κ have the following general form

$$\sum_{i=1}^{n} \frac{\partial E(X_{ij})}{\partial \kappa} \frac{X_{ij} - E(X_{ij})}{V\big[E(X_{ij})\big]} = 0, \tag{10}$$

with κ some parameter that is part of the mean model and $V[E(X_{ij})] \propto \mathrm{Var}(X_{ij})$. Intuitively, observations with a large raw residual, a low variance and whose expectation varies strongly with the parameter have a strong influence on the parameter estimate. In some special cases of $V[E(X_{ij})]$, these estimating equations correspond to score equations from maximum likelihood estimation, but in general there is no underlying likelihood function being maximized by solving these equations.

The model for the variance in sequence count data can be inspired by a parametric assumption, e.g. assuming the variance equal to the mean as for the Poisson distribution. Alternatively, the mean-variance trend can be estimated nonparametrically from the data, as is often done in genomics (17–19). Yet, as sequence count data are compositional, we are modelling mean compositions and treat the sequencing depths as ancillary statistics. Therefore we chose to model the trend between the mean relative abundance and the variance for every dimension m (henceforth called the 'abundance-variance trend' v_m), and include the sequencing depth only as a constant. More formally we assume that

$$\begin{aligned} &\mathrm{Var}(X_{ij}|\pi_j^{\text{indep}}, \mathbf{Z}_{.1,m}, \boldsymbol{\Gamma}_{1,m.}, s_i) \\ &\quad = v_m\left(\pi_{ijm}(\boldsymbol{\pi}^{\text{indep}}, \mathbf{Z}_{.1,m}, \boldsymbol{\Gamma}_{1,m.})\right) s_i, \end{aligned} \tag{11}$$

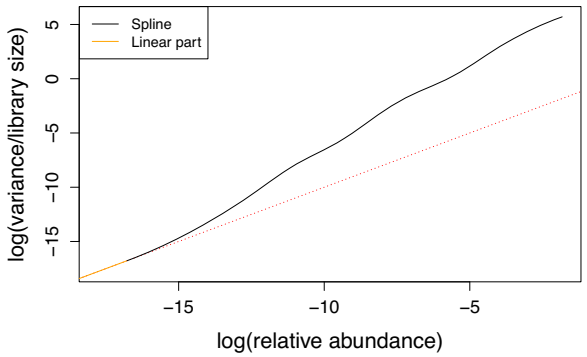

Figure 1. Abundance-variance trend (in black) as estimated from the HMP2 microbiome dataset under the independence model. The red dashed line corresponds to a variance proportional to the mean relative abundance. The linear line in orange reflects the heuristic that the variance can never drop below the mean.

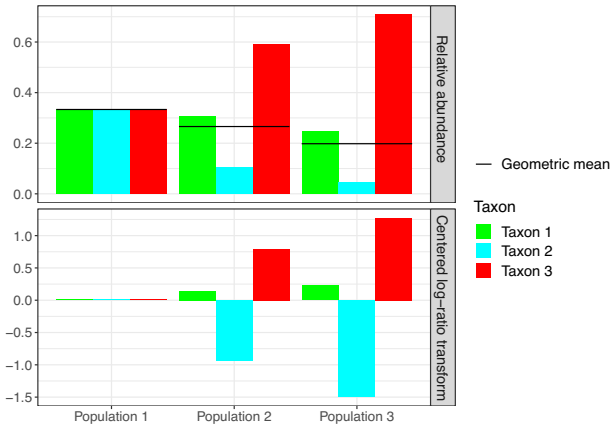

Figure 2. Toy example illustrating the difficulties in interpreting the centered log ratio (clr) transform. In the top panel, three toy populations of three taxa are shown, varying from even on the left to uneven on the right. Horizontal lines represent corresponding geometric means of the relative abundances. The bottom panels shows the clr transforms of these populations. Notice how taxon 1 decreases in abundance from left to right, whilst its clr transform increases, because the geometric mean drops faster than the relative abundance of taxon 1 as the population becomes less even.

where $\mathbf{Z}_{.1,m}$ indicates the first m columns of $\mathbf{Z}$ and $\mathbf{\Gamma}_{1,m.}$ the first m rows of $\mathbf{\Gamma}$. The variance is thus calculated conditional on the first m dimensions of the model fitted so far.

The smooth function v_m needs to be estimated from the data, based on the relationship over all features between the relative abundances under the independence model $\boldsymbol{\pi}^{\text{indep}}$ and the feature-wise weighted variances of under the model of dimension m. The latter are estimated as

$$\overline{\text{Var}(X_j|\mathbf{Z}_{.1,m})} = \frac{\sum_{i=1}^{n} \frac{\left(X_{ij} - E(X_{ij}|\mathbf{Z}_{.1,m})\right)^2}{s_i}}{n-1}. \qquad (12)$$

A cubic smoothing spline on the log-scale is chosen for v_m. As a heuristic, it is restricted to coincide with the diagonal line for low abundances, which corresponds to the variance model of the Poisson distribution (17,29) (detailed in Supplementary Section 1.2.3). Figure 1 shows an example of such abundance-variance trend in microbiome data. Given an estimated sequencing depth s_i and a modelled feature proportion π_{ij}, the predicted variance $v_m(\pi_{ij})s_i$ is then inserted into (10). Of course, $\boldsymbol{\pi}^{indep}$ remains constant, but $\overline{\text{Var}(X_j|\mathbf{Z}_{.1,m})}$ changes as the model is fitted, such that the abundance-variance trend v_m needs to be iteratively re-estimated for each dimension m.

Continuous data. For modelling other data types with approximate (log-)normal distributions (e.g. microarray), we mainly follow the tracks of the popular *limma* package (30). The data may be log-transformed, and is then modelled using a simple linear model with identity link. The estimates of the feature-wise variances are shrunken towards a common value using an empirical Bayes procedure (31).

Missing data. Missing data are a common issue in genomics data. However, in data integration the problem is exacerbated as measurements may be missing for some subjects in one view and for other subjects in another view. This problem is often tackled by removing subjects for whom information is not available for all views (a 'complete cases' analysis). Yet this approach throws away useful data, and is only valid under the missing completely at random assumption. Another strategy is to impute missing measurements from available data, but this adds complexity to the method and may affect the results. Model-based methods can naturally deal with missing observations by excluding missing values from the estimating equations. For example, in Equation (10), index i only runs over observations that are not missing. Such approach is advantageous, as it optimally extracts the information from a given dataset without the need for omission or imputation of data, and valid when data are missing at random (MAR). However, the approach is not valid when values are not missing at random (MNAR or informative missingness).

Influence functions. Noise levels may vary considerably between different views, as different technologies are designed to measure distinct biological processes. Once the data integration model has been fitted, it may be interesting to examine which datasets contributed most to the estimation of the latent variables or the environmental gradient. Because the estimating equations are stated explicitly, these contributions can be extracted directly through their corresponding influence functions. Figure 4 illustrates how influence functions can reveal the views that contribute most to the estimation of the different model components; see Supplementary Section 1.3 for an exhaustive discussion.

Construction and interpretation of multiplots

Once fitted, the low dimensional mean model can be plotted in an integrated graph. First, the scores of the latent variables are plotted as dots in two (or three) dimensions, i.e. the pair (z_{i1}, z_{i2}) determines the location of sample i $(1,\ldots, n)$. The distances between those dots reflect the dissimilarities between the samples across all views. For a constrained integration, the loadings $\boldsymbol{\lambda}$ are added as labels at locations $(\lambda_{k1}, \lambda_{k2})$, $(k = 1,\ldots, d)$. Their distance from the origin reflects the importance of sample-specific variable k in explaining the variability across all views. However, note that no direct

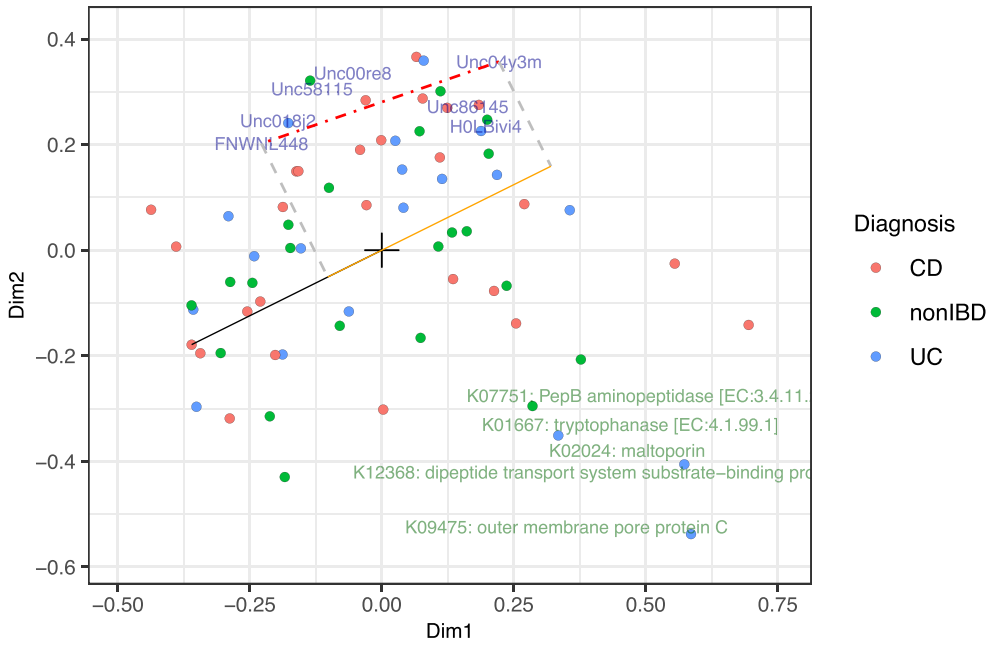

Figure 3. Data integration triplot of the microbiome and proteome datasets from the HMP2 project. Coloured dots represent patients, labels represent features of microbiome (blue) and proteome (green). We focus our interpretation on the link between taxa *Unc04y3m* and *FNWNL488* (shown with a red dashed line). The black line connects the CD sample on the left with the origin, and the dashed grey lines illustrate the projection of the taxa *Unc04y3m* and *FNWNL488* link onto this black line; the projection itself is shown as an orange line. This projection is large, implying that the ratio FNWNL448/Unc04y3m is much larger in this sample than in the average sample. On the contrary, taxa *Unc018j2* and *FNWNL488* lie close together in the top left corner, and have a short link. This implies that their ratio differs very little over all samples, and these features' abundances are thus correlated.

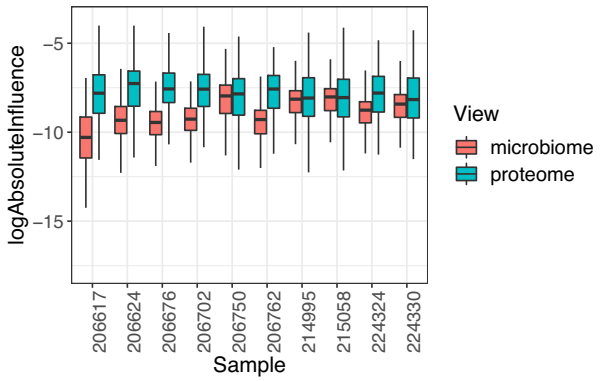

Figure 4. Data integration of the microbiome and proteome datasets from the HMP2 project. Boxplots of log absolute influence on the estimation of the latent variables of the first 10 samples in dimension 1.

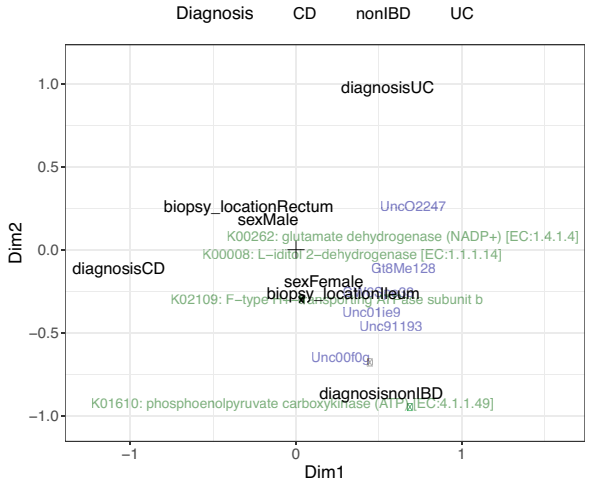

Figure 5. Constrained ordination of the microbiome and proteome datasets from the HMP2 project. Coloured dots represent patients, labels represent features of microbiome (blue) and proteome (green). Black labels represent patient variables.

interpretation of their distances to the sample locations is available, and that distances between variable labels are not meaningful either. Finally, the feature loadings in $\mathbf{\Gamma}$ and $\mathbf{\Theta}$ of the different views are added as labels. Because of the high dimensionality, one may choose to plot features with the largest loadings only (i.e. furthest away from the origin) to avoid overplotting ('thresholding'). The interpretation of these feature labels depends on whether the dataset is compositional.

For non-compositional data types, the interpretation of the loading vector $\boldsymbol{\gamma}_a$ for a given feature a is simple: feature a has a higher mean than in the average sample in a sample i that lies on the same side of the origin (i.e. $\boldsymbol{\gamma}_a^t \mathbf{Z}_i > 0$), and a lower mean otherwise. The outcomes of these features are also positively associated with the sample variable k whose label lies on the same side of the origin as the feature label (i.e. $\boldsymbol{\gamma}_a^t \boldsymbol{\lambda}_k > 0$). Moreover, two features plotted on the same side of the origin are positively correlated, whilst features on opposite sides indicate a negative correlation, regardless of their respective views.

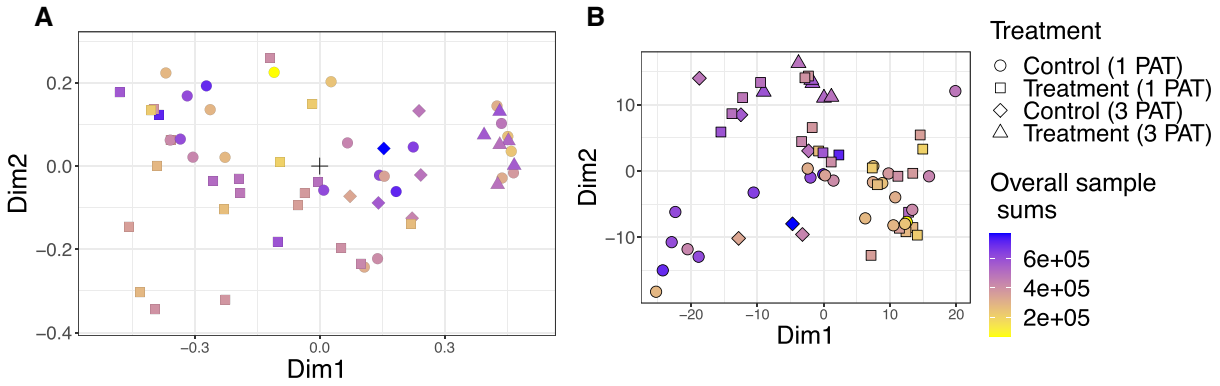

Figure 6. Unconstrained sample ordination of Zhang microbiome and immunological data through (A) our COMBI method and (B) PCA on clr-transformed data. Colours reflect overall sample sums, shapes represent T1D status: control (PATCON) or treatment group (PAT) in experiment with 1 (1 PAT) or 3 (3 PAT) pulsed antibiotic treatments.

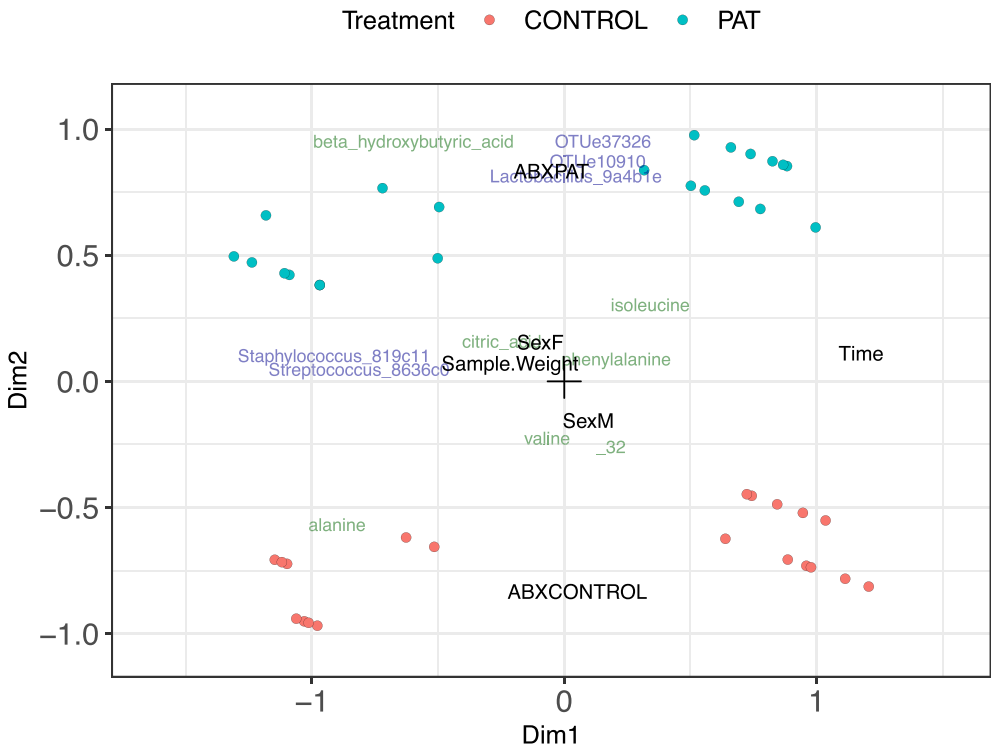

Figure 7. Quadriplot of the constrained integration of Zhang microbiome and metabolomics data. Coloured dots represent mice, blue and green labels are microbial taxa and metabolites, respectively. Black labels represent components of the environmental gradient.

Interpretations involving features from compositional datasets (i.e. compositional biplots as introduced by Aitchison and Greenacre (27)) are less straightforward, since compositionality imposes a certain dependence between features. Hence, feature loadings should not be interpreted individually, but always in combination with at least one other feature. A mathematically convenient quantity is the log-ratio of the feature proportion to the geometric mean (gm) of all feature proportions: $\log\left(\frac{\pi_j}{\mathrm{gm}(\boldsymbol{\pi})}\right)$. When the feature label lies for instance on the same side of the origin as the sample, this means that the log-ratio of this feature is larger in this sample than in the average sample. Yet this may not be very meaningful biologically, because the geometric mean of feature proportions may be a very intractable summary. The geometric mean can be regarded as a measure of evenness, similar to the Shannon index (32). It is maximal (equal to $1/p$) for a perfectly even composition, but decreases irregularly as the composition departs from perfect evenness. As a result, despite the fact that the log-ratio evolves linearly with the latent variables, the feature proportion often varies non-monotonically. These capricious effects are demonstrated in Figure 2.

Therefore, we consider ratios of two (or more) features (27). Denoting $\boldsymbol{\gamma}_b$ the feature loading of feature b, we

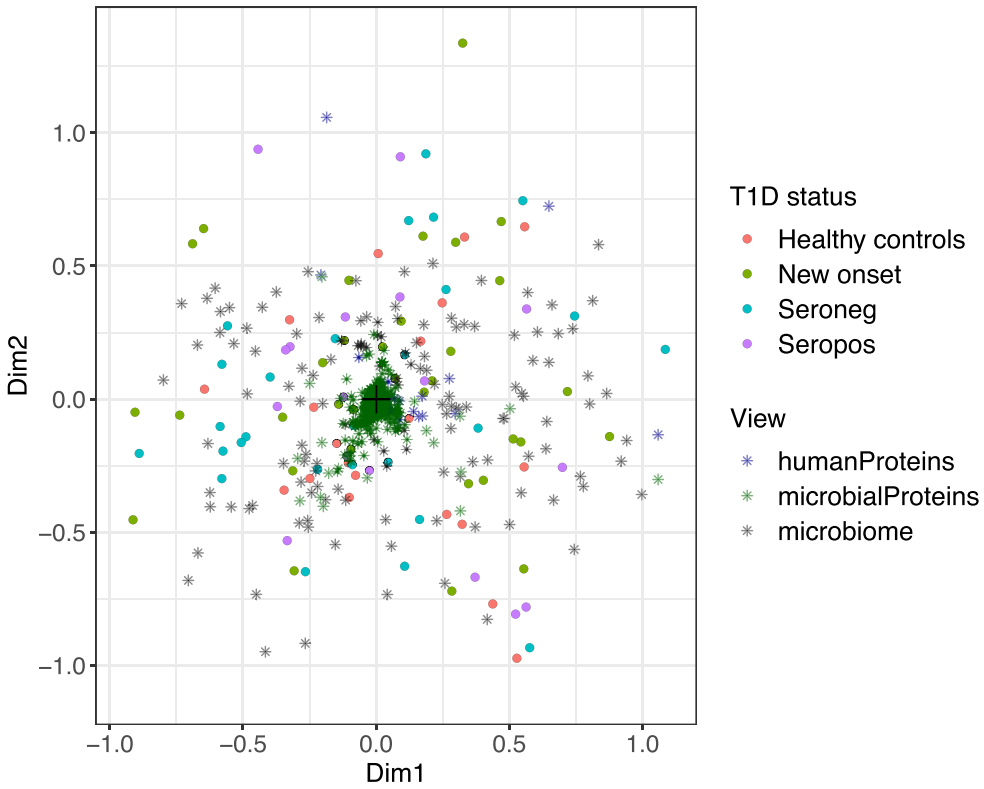

Figure 8. Unconstrained integration of Gavin microbiome and human and microbial proteome data. Dots represent samples, stars represent all features of the different views, without thresholding.

consider:

$$\log\left(\frac{\pi_{ia}}{\pi_{ib}}\right) - \log\left(\frac{\pi_a^{\text{indep}}}{\pi_b^{\text{indep}}}\right) = \mathbf{Z}_i^t(\boldsymbol{\gamma}_a - \boldsymbol{\gamma}_b). \qquad (13)$$

Note that the geometric mean has been eliminated from the expression. The difference $(\boldsymbol{\gamma}_a - \boldsymbol{\gamma}_b)$ between vectors is known as the *link* in a plot, i.e. the straight line connecting the points defined by $\boldsymbol{\gamma}_a$ and $\boldsymbol{\gamma}_b$. This difference is small when the features labels are close on the multiplot (i.e. the Euclidean distance between the loadings is small). In that case, this means that for *any* sample, the ratio of the relative abundances $\frac{\pi_{ia}}{\pi_{ib}}$ and the ratio of the relative abundances under the independence model $\frac{\pi_a^{\text{indep}}}{\pi_b^{\text{indep}}}$ do not differ by much. In a compositional setting, a stable ratio across samples indicates that the features are strongly correlated (21). When this link is large, the projection of the latent variable scores $\mathbf{Z}_i$ onto the link (i.e. $\mathbf{Z}_i^t(\boldsymbol{\gamma}_a - \boldsymbol{\gamma}_b)$) indicates how much and in which direction the ratio $\frac{\pi_a}{\pi_b}$ differs from that under the independence model in sample i (27). This implies that features lying on the same radius from the origin but far apart (i.e. we have $\frac{\boldsymbol{\gamma}_a^t\boldsymbol{\gamma}_b}{||\boldsymbol{\gamma}_a||||\boldsymbol{\gamma}_b||} \approx 1$ but $||\boldsymbol{\gamma}_a|| \neq ||\boldsymbol{\gamma}_b||$) are not necessarily strongly correlated in all samples! These interpretations are illustrated in Figure 3. The interpretation of combinations of features from different compositional views is very difficult.

Importance of the dimensions. Many ordination methods output measures of importance of the dimensions, typically as the fraction of total variability. Yet this is difficult for many non-normal data types, and hard to compare across different views. Moreover, as part of the variability in any stochastic dataset is noise, it unknown which fraction of the total variability this noise represents, and hence hard to know which fraction of total variability the ordination should strive to explain. For these reasons, our COMBI method does not yield any measures of variability explained by the fractions, but the axes are forced to be square. As a result, the euclidean distances between sample dots truthfully reflect dissimilarities between the samples, and a large spread of the sample scores in one dimension indicates a large variability in this dimension.

Real case studies and analysis

We considered three studies. The Human Microbiome Project 2 (HMP2), or integrative HMP (iHMP), aimed to investigate the relationship between the microbiome and host responses. One branch focused on healthy and IBD patients (with either Crohn's disease (CD) or ulcerative colitis (UC)) (33), which constitutes our first study. The second study is of Zhang et al. (34) who investigated the effect of pulsed antibiotic treatments (PAT) on the onset of Type I diabetes (T1D) in mice. The gut microbiome composition, as well as many host genomics measurements were recorded over time. We refer to this study as the Zhang data. A third study (the Gavin data) contain microbiome data and human and microbial proteome data from patients with T1D, as well as from healthy controls (35).

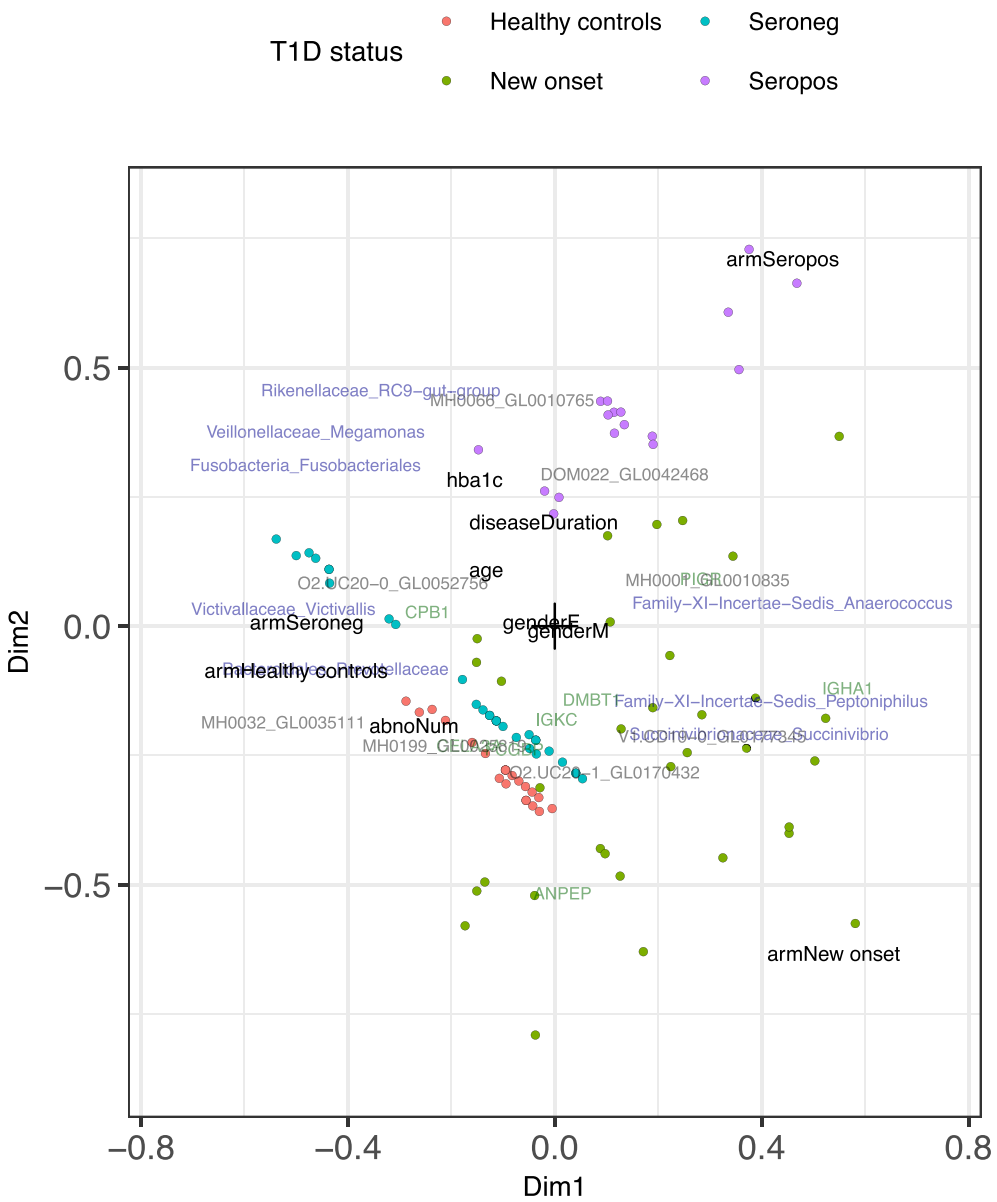

Figure 9. Constrained integration of Gavin microbiome and human and microbial proteome data. Blue labels represent taxa, red labels microbial proteins and green labels human proteins. Black labels represent components of the environmental gradient. Healthy and seronegative states are similar, as expected since they represent similar physiological states.

Real data analysis. All datasets were subjected to unconstrained analysis with default settings of the *combi* package. For the HMP2 package, all unconstrained two-way integrations of the microbiome, proteome and virome datasets were fitted, as well as a three-way integration. Constrained integrations were fitted on the same dataset using the 'biopsy location', 'diagnosis' and 'sex' variables. On the Zhang data, the two-way integrations of microbiome-metabolome and microbiome-immunological data were fitted. Constrained integrations were fitted on the same dataset using the 'Treatment' (PAT), 'Sex', 'Time' and 'Sample.Weight' variables. On the Gavin data the unconstrained and constrained three-way integrations of microbiome and human and microbial proteome data were fitted, the constrained integration using the 'T1D status', 'age', 'Number of auto antibodies' (abnoNum), 'disease duration' and 'hba1c' variables. All datasets were treated as compositional and modelled using quasi-likelihood, except for the Zhang metabolome and Gavin proteome data, for which Gaussian models were fit without compositionality constraint.

Simulation study

Data generation. To evaluate and benchmark the performance of our method, data were generated according to three different paradigms. The first data generation strategy assumes that the sequence count data follow the negative bi-

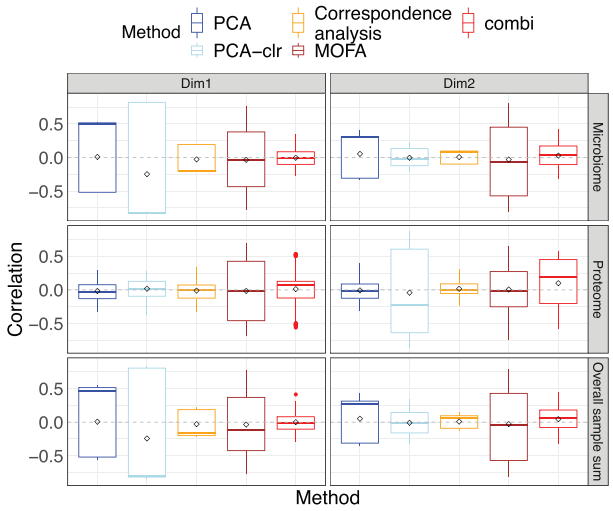

Figure 10. Simulation study. Boxplots of correlation between sample scores and overall and view-wise sample sums for permuted microbiome and proteome data simulated from the HMP2 dataset (strategy 3).

nomial distribution. For metabolome and Gavin proteome data, a Gaussian distribution is assumed. The parameters of these distributions were estimated from the real datasets through maximum likelihood. Parameter values were then sampled from this pool of parameter estimates, and random data were drawn from the corresponding distributions. The samples were split into two equally sized groups, and for 10% of the features, a fold change was introduced in one of the groups. For compositional data this happened both with and without compensation. Compensation means that the abundance of some of the features is increased and for others the abundance is decreased, such that the abundance of the remaining features is left unaltered (12). For Gaussian data the fold change was 0.1, for sequence count data it was 4. In the second strategy, data were generated using the *SimSeq* procedure (14). The IBD status was used as grouping variable for the HMP2 data, the treatment group for the Zhang data, and the T1D status for the Gavin data. The same samples were used to draw observations for both views, in order to preserve correlations between views. In the third strategy, real data were reshuffled by permuting the samples of different views independently. This breaks the correlation between features from the different views, which provides a useful null setting with real data characteristics. In all cases, the number of samples was $n = 40$, and only the $P = 1000$ most abundant features were used. In each setting, 100 Monte-Carlo runs were executed.

Benchmark methods. We compared our method with following other integration methods or approaches. JIVE and MOFA were run using the *r.jive* (36) and *MOFA* (37) packages, respectively. Canonical correlation analysis was applied both with and without shrinkage, and with and without prior clr transformation, using the *PMA* package (38). We also considered concatenating the view matrices by row to perform principal component analysis to either the raw or clr-transformed data, as well as to perform correspondence analysis on the raw data. Partial least squares with canonical mode was applied to raw and clr-transformed data as implemented in the *mixOmics* package (39). Prior to all clr transformations, zero counts were imputed using the *cmultRepl*() function in the *zCompositions* package (40). Unless mentioned otherwise, default settings were used for all packages, with two dimensions. All analyses were run in R programming language, version 3.6.1 (41). Details on the software and package versions used can be found in the Supplementary Section 6.

Method evaluation. The methods were evaluated based on the correlation of the sample scores with the sample-wise sums of each view separately and with the overall sum. To quantify how well the methods identify correlated features, the inner products of all feature loadings were calculated, and the Wilcoxon rank sum test was performed to assess whether truly correlated features had a higher inner product than uncorrelated taxa. We choose the rank approximation because only the features with the strongest signal will be plotted, and hence their correct ranking is crucial. The standardized Wilcoxon rank sum test statistic was then used as a measure of discrimination between correlated and non-correlated taxa. This test statistic was calculated for all feature combinations as well as for between-view combinations only. The pseudo-F statistic was used to evaluate the clustering of samples from the same group (3,42).

RESULTS

Real data results

We summarize in this section the most important findings; the remainder of the analyses are presented in the Supplementary Material.

HMP2 data. The unconstrained integration of HMP2 microbiome and proteome datasets is shown in Figure 3. No clear clusters of patients with different disease statuses are visible, as the variability in this dataset is large. The influence plot in Figure 4 reveals that the proteome view has the largest influence on the estimation of the sample variables. Yet, the constrained ordination of the same data (Figure 5) identifies the disease status as more important in driving variability than gender or biopsy location.

Zhang data. The unconstrained integration of the Zhang microbiome and immunological data (Figure 6A and Supplementary Figure S11) reveals that in these two views, the variability is much larger in the single PAT than in the triple PAT group. For comparison, the sample ordination by PCA with clr-transform is shown in Figure 6B. In this ordination, the separation of the different treatment groups is less clear, as the effect of the overall sample sum distorts the ordination.

The constrained ordination of the microbiome and metabolomics data in Figure 7 shows that time is the most important sample variable in driving variability, followed by antibiotics treatment group. Citric acid, isoleucine and valine can be seen to be associated to the antibiotics treatment, as was also discovered by the authors of the original study (34) and can be seen in the unconstrained integration (Supplementary Figure S15) as well.

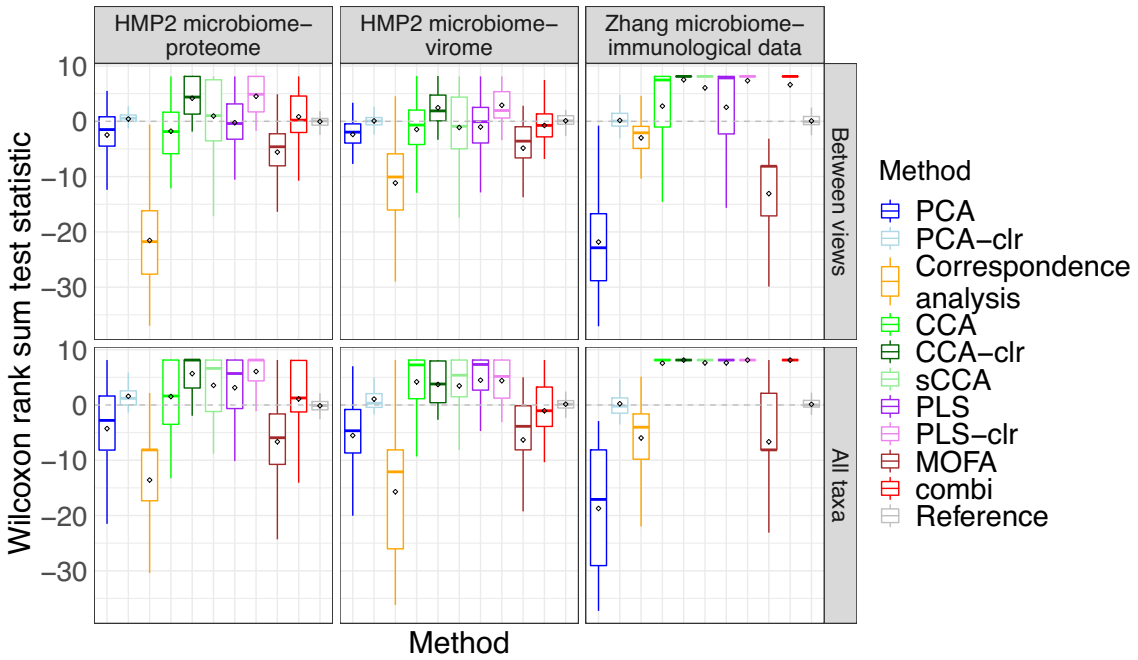

Figure 11. Simulation study. Wilcoxon rank sum test statistic quantifying correlated taxon identification for different methods (*x*-axis) and templates (top panels) on parametrically generated data based on the real case studies (strategy 1).

Gavin data. The unconstrained ordination of the Gavin data (Figure 8) did not show any clear sample clusters, as noise levels are too high. The constrained integration of the Gavin data (Figure 9 and Supplementary Figure S17) confirm that the healthy control and seronegative statuses are most similar. Apart from T1D status, we noted hba1c (a measure for average past blood glucose levels), disease duration and number of auto antibodies (abnoNum) as strong drivers of variability in proteome and microbiome, whereas gender was unimportant. The *PIGR*, *IGKC* and *IGHA1* human proteins are markers of inflammation that are more abundant in seropositive and diseased patients, which may point to abnormal immune response in their gut. Chymotrypsin Like Elastase 3A (*CELA3A*) was higher in abundance in healthy and seronegative patients, as was also found by the authors of the original study.

Simulation results

In parametric as well as non-parametric simulations, the sample scores of PCA (with and without clr-transformation) and MOFA were correlated with the library sizes in some scenarios (see Figure 10; Supplementary Figures S12–14 and 18–35), confirming our observation from the real case study in Figure 6. Methods based on covariances (cca and pls) were best at identifying correlated taxa, both within and between datasets (see Figure 11 and Supplementary Figure S36). These methods perform especially well on clr-transformed data. Our COMBI method was best at identifying clusters of samples over the different views (see Figure 12). Our COMBI method and PCA with clr-transformation perform especially better in the 'no compensation' scenario, indicating how crucial it is to account for compositionality. Detailed results can be found in Supplementary Section 5.

DISCUSSION

Data integration is challenging in statistical data analysis. Most statistical methods focus on a single dataset at a time, but integration requires finding common patterns across different views. Data integration can answer important biological questions, but sets a complicated task due to differences in measurement technology, outcome distributions and signal-to-noise levels. We have used dedicated regression models and outcome distributions for each view separately, whilst including common latent variables to discover signals across the different views. This results in a method enabling graphical exploration of multiple datasets. All features are included in the model, but visualization displays only features with the strongest signal. Thus, feature selection (e.g. by imposing sparsity) is deferred to later stages of the analysis with other methods. Alternatively, penalization could be included in a future version of *COMBI* for automated feature selection.

Another advantage of our method is its ability to handle missing data, that frequently arise in data integration problems—some views may not be measured on some samples. COMBI avoids the need of a complete case analysis or imputation thanks to explicit estimating equations. Thus, our method naturally ignores missing values in the fitting process without the need to drop samples. As with all other methods that account for missingness, this approach may still be biased when data are not missing at random. However, COMBI focuses on visualization rather than formal inference, and we believe this problem is less severe here in such framework. Nevertheless, all data analysts should al-

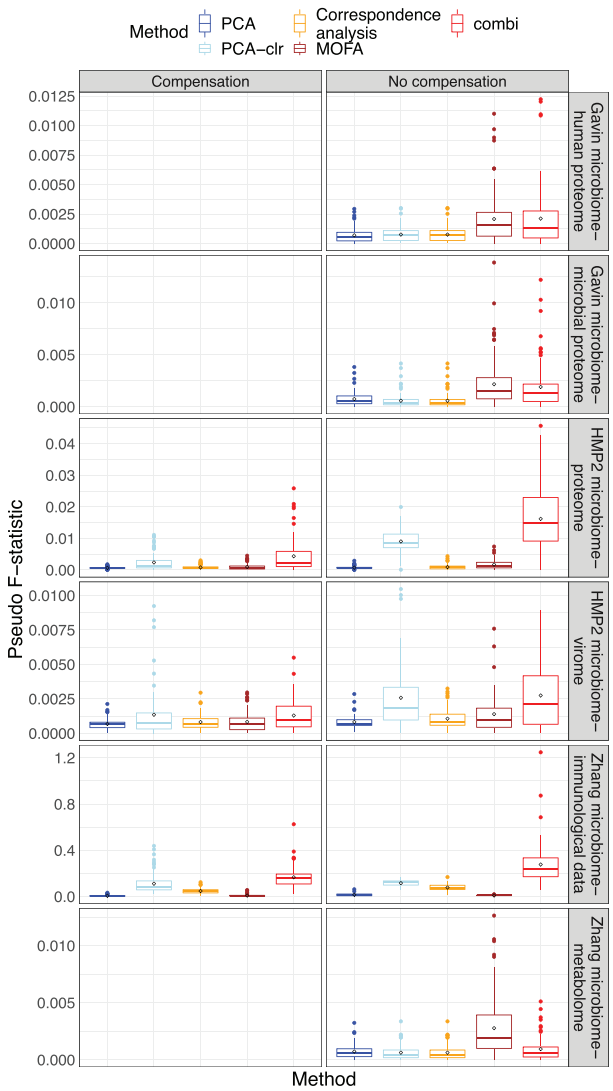

Figure 12. Simulation study. Boxplots of pseudo-F statistic (*y*-axis) quantifying sample separation for different methods (*x*-axis) and templates (top panels) under parametric simulation (Strategy 1).

ways think carefully about the consequences of the study design and the causes of missingness.

In our method, we considered log-ratio link functions in regression models as a promising avenue for modelling compositional (count) data. Such models combine compositional effects with adequate mean-variance modelling and flexibility to include covariates. However we also face the same limitation as any classical compositional methods in terms of the interpretation of the results, which should be done with caution (27). Moreover, fitting these complex regression models on sparse omics datasets can be numerically challenging. Future investigations into their numerical properties, choice of link function and fitting algorithms are needed.

Finding an adequate distribution for sequence count data has proven difficult. Yet it is necessary to account for the unmistakable mean-variance trend of count data in ordination methods, to avoid undue interference of sequencing depths. We proposed a distribution-free approach for estimating the trend between relative abundance and variance by leveraging from the high dimensionality of the data. Contrary to most other data integration methods, our approach successfully avoids any influence of the sequencing depths on the samples scores.

Our COMBI method performed well at clustering samples with similar properties across different views, and is insensitive to differences in sequencing depth. Covariance-based methods such as canonical correspondence analysis and partial least squares are best at identifying correlated features, but do not yield overall sample ordinations, only distinct sample ordinations per view. Thus, the data analysts must choose their methods according to their research question and intended outputs. Our COMBI method offers an interesting alternative to existing data integration approaches when the focus is on exploring the relationship between samples and revealing which features contribute to differences between samples. Its constrained variant allows to include sample variables in the analysis, an appealing feature to deal with complex studies. As such, COMBI is a powerful new tool for the simultaneous exploration of multiple datasets.

SOFTWARE

The R-package *combi*, which implements the integration algorithm, is available from BioConductor, with a detailed exemplar vignette.

SUPPLEMENTARY DATA

Supplementary Data are available at NARGAB Online.

ACKNOWLEDGEMENTS

We are grateful to Xuesong Zhang, Antoine Bodein and Patrick Gavin for sharing their data and providing useful feedback.

FUNDING

Johnson and Johnson (to S.H.).
Conflict of interest statement. L.B. is currently employed by Janssen Pharmaceutical companies of Johnson and Johnson.

REFERENCES

1. Hotelling,H. (1935) The most predictable criterion. *J. Educ. Psychol.*, **26**, 139–142.
2. Wold,S., Ruhe,A., Wold,H. and Dunn,W.D. III (1984) The collinearity problem in linear regression. The Partial Least Squares (PLS) approach to generalized inverses. *SIAM J. Sci. Stat. Comp.*, **5**, 735–743.
3. Hawinkel,S., Kerckhof,F.-M., Bijnens,L. and Thas,O. (2019) A unified framework for unconstrained and constrained ordination of microbiome read count data. *PLoS One.*, **14**, e0205474.
4. Argelaguet,R., Velten,B., Arnol,D., Dietrich,S., Zenz,T., Marioni,J.C., Buettner,F., Huber,W. and Stegle,O. (2018) Multi-omics factor analysis—a framework for unsupervised integration of multi-omics data sets. *Mol. Syst. Biol.*, **14**, e8124.
5. Hui,F., Taskinen,S., Pledger,S., Foster,S. and Warton,D. (2015) Model-based approaches to unconstrained ordination. *Methods Ecol. Evol.*, **6**, 399–411.

6. Pledger,S. and Arnold,R. (2014) Multivariate methods using mixtures: correspondence analysis, scaling and pattern-detection. *Comput. Stat. Data An.*, **71**, 241–261.
7. Sohn,M.B. and Li,H. (2018) A GLM-based latent variable ordination method for microbiome samples. *Biometrics*, **74**, 448–457.
8. Yee,T.W. and Hadi,A.F. (2014) Row–column interaction models, with an R implementation. *Comput. Stat.*, **29**, 1427–1445.
9. Wilms,I. and Croux,C. (2016) Robust sparse canonical correlation analysis. *BMC. Syst. Biol.*, **10**, 72–72.
10. Tenenhaus,A., Philippe,C., Guillemot,V., Cao,K.-A.L., Grill,J. and Frouin,V. (2014) Variable selection for generalized canonical correlation analysis. *Biostatistics.*, **15**, 569–583.
11. Singh,A., Shannon,C.P., Gautier,B., Rohart,F., Vacher,M., Tebbutt,S.J. and Cao,K.-A.L. (2019) DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays. *Bioinformatics*, **35**, 3055–3062.
12. Hawinkel,S., Mattiello,F., Bijnens,L. and Thas,O. (2019) A broken promise: Microbiome differential abundance methods do not control the false discovery rate. *Brief. Bioinform.*, **20**, 210–221.
13. Assefa,A.T., Paepe,K.D., Everaert,C., Mestdagh,P., Thas,O. and Vandesompele,J. (2018) Differential gene expression analysis tools exhibit substandard performance for long non-coding RNA-sequencing data. *Genome Biol.*, **19**, 96.
14. Benidt,S. and Nettleton,D. (2015) SimSeq: a nonparametric approach to simulation of RNA-sequence datasets. *Bioinformatics.*, **31**, 2131–2140.
15. Reeb,P.D. and Steibel,J.P. (2013) Evaluating statistical analysis models for RNA sequencing experiments. *Front. Genet.*, **4**, 178.
16. Hawinkel,S., Rayner,J. C.W., Bijnens,L. and Thas,O. (2020) Sequence count data are poorly fit by the negative binomial distribution. *PLoS One.*, **15**, e0224909.
17. Robinson,M.D. and Smyth,G.K. (2007) Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics.*, **23**, 2881–2887.
18. Anders,S. and Huber,W. (2010) Differential expression analysis for sequence count data. *Genome Biol.*, **11**, R106–R106.
19. Wright,F.A., Xia,K. and Zhou,Y.-H. (2011) A powerful and flexible approach to the analysis of RNA sequence count data. *Bioinformatics.*, **27**, 2672–2678.
20. Gloor,G.B., Macklaim,J.M., Pawlowsky-Glahn,V. and Egozcue,J.J. (2017) Microbiome datasets are compositional: and this is not optional. *Front. Microbiol.*, **8**, 2224.
21. Aitchison,J. (1982) The statistical analysis of compositional data. *J. R. Stat. Soc.*, **44**, 139–177.
22. Martín-Fernández,J.-A., Hron,K., Templ,M., Filzmoser,P. and Palarea-Albaladejo,J. (2015) Bayesian-multiplicative treatment of count zeros in compositional data sets. *Stat. Model.*, **15**, 134–158.
23. Billheimer,D., Guttorp,P. and Fagan,W.F. (2001) Statistical interpretation of species composition. *J. Am. Stat. Assoc.*, **96**, 1205–1214.
24. Xia,F., Chen,J., Fung,W.K. and Li,H. (2013) A logistic normal multinomial regression model for microbiome compositional data analysis. *Biometrics*, **69**, 1053–1063.
25. Chong,F. and Spencer,M. (2018) Analysis of relative abundances with zeros on environmental gradients: a multinomial regression model. *PeerJ.*, **6**, e5643.
26. Zhu,M., Hastie,T. and Walther,G. (2005) Constrained ordination analysis with flexible response functions. *Ecol. Model.*, **187**, 524–536.
27. Aitchison,J. and Greenacre,M. (2002) Biplots of compositional data. *J. R. Stat. Soc. C*, **51**, 375–392.
28. Wedderburn,R. W.M. (1974) Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika*, **61**, 439–447.
29. Marioni,J.C., Mason,C.E., Mane,S.M., Stephens,M. and Gilad,Y. (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.*, **18**, 1509–1517.
30. Ritchie,M.E., Phipson,B., Wu,D., Hu,Y., Law,C.W., Shi,W. and Smyth,G.K. (2015) Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.*, **43**, e47.
31. Smyth,G.K. (2004) Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.*, **3**, doi:https://doi.org/10.2202/1544-6115.1027.
32. Shannon,C.E. (1948) A mathematical theory of communication. *Bell. Syst. Tech. J.*, **27**, 379–423.
33. Lloyd-Price,J., Arze,C., Ananthakrishnan,A.N., Schirmer,M., Avila-Pacheco,J., Poon,T.W., Andrews,E., Ajami,N.J., Bonham,K.S., Brislawn,C.J. *et al.* (2019) Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature*, **569**, 655–662.
34. Zhang,X.-S., Li,J., Krautkramer,K.A., Badri,M., Battaglia,T., Borbet,T.C., Koh,H., Ng,S., Sibley,R.A., Li,Y. *et al.* (2018) Antibiotic-induced acceleration of type 1 diabetes alters maturation of innate intestinal immunity. *eLife.*, **7**, e37816.
35. Gavin,P.G., Mullaney,J.A., Loo,D., Cao,K.-A.L., Gottlieb,P.A., Hill,M.M., Zipris,D. and Hamilton-Williams,E.E. (2018) Intestinal metaproteomics reveals host-microbiota interactions in subjects at risk for Type 1 diabetes. *Diabetes Care*, **41**, 2178–2186.
36. O'Connell,M.J. and Lock,E.F. (2017) R.Jive: perform JIVE decomposition for multi-source data. R package version 2.1.
37. Argelaguet,R., Velten,B., Arnol,D., Buettner,F., Huber,W. and Stegle,O. (2019) MOFA: multi-omics factor analysis (MOFA). R package version 1.0.0.
38. Witten,D., Tibshirani,R., Gross,S. and Narasimhan,B. (2019) PMA: penalized multivariate analysis. R package version 1.1.
39. Rohart,F., Gautier,B., Singh,A. and Cao,K.-A.L. (2017) MixOmics: an R package for 'omics feature selection and multiple data integration. *PLoS. Comput. Biol.*, **13**, e1005752.
40. Palarea-Albaladejo,J. and Martin-Fernandez,J.A. (2015) ZCompositions–R package for multivariate imputation of left-censored data under a compositional approach. *Chemom. Intell. Lab. Syst.*, **143**, 85–96.
41. R Core Team (2020) In: *R: A Language and Environment for Statistical Computing R Foundation for Statistical Computing*. Vienna, Austria.
42. Anderson,M.J. (2001) A new method for non-parametric multivariate analysis of variance. *Austral. Ecol.*, **26**, 32–46.