AMERICAN SOCIETY FOR MICROBIOLOGY mSystems®

# Approximation of a Microbiome Composition Shift by a Change in a Single Balance Between Two Groups of Taxa

Vera E. Odintsova,[a] Natalia S. Klimenko,[a,b] Alexander V. Tyakht[a,b]

[a]Atlas Biomed Group–Knomx LLC, Moscow, Russia

[b]Center for Precision Genome Editing and Genetic Technologies for Biomedicine, Institute of Gene Biology, Russian Academy of Sciences, Moscow, Russia

**ABSTRACT** Linking microbiome composition obtained from metagenomic or 16S rRNA sequencing to various factors poses a real challenge. The compositional approach to such data is well described: a so-called isometric log-ratio (ILR) transform provides correct treatment of relative abundances. Most existing compositional methods differ in the particular choice of the transform. Although this choice does not influence the prediction of a model, it determines the subset of balances between groups of microbial taxa subsequently used for interpreting the composition shifts. We propose a method to interpret these shifts independently of the initial choice of ILR coordinates by the nearest single-balance shift. We describe here application of the method to regression, classification, and principal balance analysis of compositional data. Analytical treatment and cross-validation show that the approach provides the least-squares estimate of a single-balance shift associated with a factor with possible adjustment for covariates. As for classification and principal balance analysis, the nearest balance method provides results comparable to other compositional tools. Its advantages are the absence of assumptions about the number of taxa included in the balance and its low computational cost. The method is implemented in the R package NearestBalance.

**IMPORTANCE** The method proposed here extends the range of compositional methods providing interpretation of classical statistical tools applied to data converted to the ILR coordinates. It provides a strictly optimal solution in several special cases. The approach is universally applicable to compositional data of any nature, including microbiome data sets.

**KEYWORDS** compositional analysis, balances, microbial signatures, microbiome, metagenomics, principal balance analysis, regression analysis

The development of high-throughput DNA sequencing methods has provided vast opportunities to explore complex microbial communities inhabiting various ecological niches such as soil, ocean, and host-associated locations. Hence, it has become particularly important to standardize microbiome data analysis at each step. To date, there is no golden standard of statistical approach to these data. Researchers have to choose a single method from a wide assortment (from the nonparametric Mann-Whitney test to complex probabilistic models) in order to overcome the limitations arising from the specific nature of microbiome composition data, such as compositionality and sparsity.

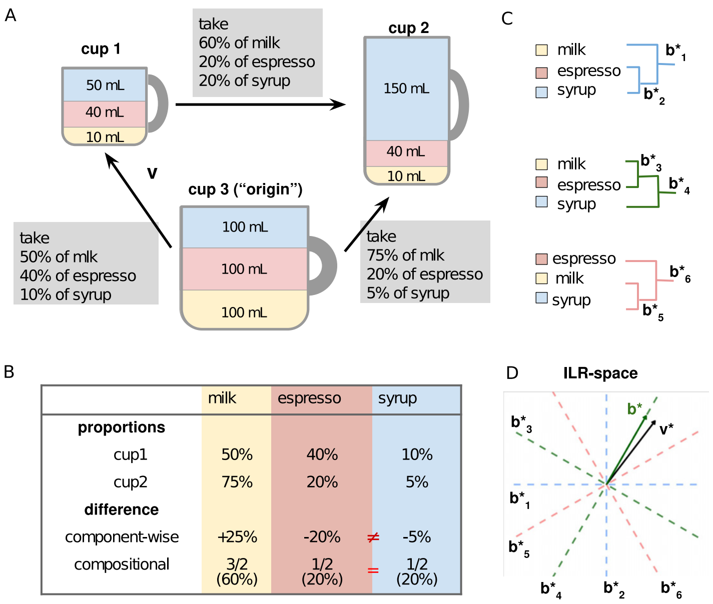
Generally, each statistical method for microbiome data implements one of two approaches: component-wise or compositional. Component-wise methods treat each taxon individually and differ by their underlying probabilistic models. On the other hand, compositional methods treat each species as a part of the whole community and differ by the specific ways to identify patterns in the composition. Several reviews (1, 2) emphasized that only the compositional approach is appropriate for the analysis

**Editor** Jack A. Gilbert, University of California San Diego

Address correspondence to Vera E. Odintsova, vera.odints@gmail.com.

The authors declare no conflict of interest.

**Received** 1 March 2022
**Accepted** 4 April 2022
**Published** 9 May 2022



| | milk | espresso | | syrup |
|---|---|---|---|---|
| **proportions** | | | | |
| cup1 | 50% | 40% | | 10% |
| cup2 | 75% | 20% | | 5% |
| **difference** | | | | |
| component-wise | +25% | -20% | ≠ | -5% |
| compositional | 3/2<br>(60%) | 1/2<br>(20%) | = | 1/2<br>(20%) |

**FIG 1** Compositional approach to the analysis of difference between two compositions: an example of two coffee recipes. (A) Definition of perturbation as a fold change of components' proportions in samples. (B) Comparison of compositional and component-wise definitions of the difference in proportions. (C) Binary trees that may be used for the construction of an ILR system of coordinates. The difference between the drinks may be described by a shift in balance $b^*_1$, while $b^*_2$ is constant across the compositions. (D) Visualization of the suggested method: search for a balance b that is the closest to some vector of interest v in an ILR system of coordinates [$v^* = \mathrm{ILR}(v)$, $b^* = \mathrm{ILR}(b)$].

of proportions, while the component-wise approach leads to biases and unreliable conclusions. However, component-wise methods are still widespread since they are more intuitive and simpler. They yield a list of taxa associated with a factor. In contrast, most compositional methods yield a list of balances: the proportions between groups of taxa calculated in a specific way. As such, this approach requires more efforts to achieve meaningful interpretations for expert biologists and a wider audience.

Here, we propose a method to interpret the results of compositional analysis. This approach aims to approximate the list of balances by a single balance. To make the idea clearer, let us review some details of component-wise approach and current compositional methods (a more rigorous and complete review can be found in reference 3).

In most cases, 16S rRNA sequencing does not give information about the absolute amount of taxa in a sample but only the number of reads mapped on the genome of each taxa from a database (counts). All information contained in the data are related to the proportions between taxa. This sort of information is similar to a recipe for a drink: it may say to mix 50 mL of espresso, 40 mL of milk, and 10 mL of syrup, but making a double portion (100 mL of espresso, 80 mL of milk, and 10 mL of syrup) actually does not change the taste, since it is only the proportions between components that matter. Such data, i.e., providing a vector with positive numbers for which only proportions between components are informative and where neither the order nor the total sum are important, are called a composition. Data obtained from sequencing are compositional.

The most obvious disadvantage of the component-wise approach to such data are missing the interdependence of components' relative proportions. Figure 1 illustrates the intuition of this disadvantage by comparing two coffee drinks. Consider two cups of coffee that both include syrup, espresso, and milk. The only difference between them is that the second cup contains a triple portion of milk: it is the first cup with the addition of two milk portions (Fig. 1A). If one were to taste a spoonful of each drink, the observed dissimilarities would refer to each component: the second drink would be less sweet, less strong, and more milky. While that assessment would be accurate, a more informative answer would be that the proportions between coffee and syrup in the spoons are the same, and it is only the proportion of milk to the other ingredients that differs. Now, let us imagine a study comparing the gut microbiome compositions of healthy subjects and those affected by a certain disease. If the sample size is large enough, some differences in the proportions of all taxa may be observed even if only a single pathogenic taxon overgrows in the patients. As with the coffee example, those differences would be true, but a more important conclusion would be that the commensal microbiome remained unperturbed and that only the proportion of a specific pathogen to the other species changed. However, an insufficient sample size may lead to additional bias, whereby the list of the affected taxa would be shorter and would depend on their initial proportions, as the effect size of observed changes varies between individual species. Nonetheless, it would not imply that some commensal taxa increased in presence, whereas others did not increase at all or did so to a lower extent. Figure 1B illustrates this bias on the coffee example: although the absolute change between the cups is equally absent for both syrup and espresso, the component-wise approach detects differences between changes in the proportions of these two components; the observed change in espresso proportion is almost as high as for the milk, even though milk was the only altered component.

The compositional approach starts from changing the perspective on what a taxon's differential presence between two samples means: it is the ratio of relative abundances in the two samples, rather than the difference. Returning to the coffee example, the fold changes of coffee and syrup are the same and differ from the fold change in milk proportion (Fig. 1B); such a description of dissimilarities is more informative. This kind of difference, thought of as a shift that should be applied to the first sample to obtain the second one, is called a "perturbation."

Interestingly, a perturbation itself may be treated as a composition: only the proportions of its components matter, and normalization to, for instance, 100%, preserves the result. The reverse is also true: composition may be thought of as a perturbation of an "origin," a composition with equal proportions of all components (Fig. 1A). This initial composition, when thought of as a perturbation, represents a zero shift (leaving one-third of all milk, espresso, and syrup does not change the taste of coffee), making this definition similar to the one of a zero vector in a vector space.

Thus, we have a vector space of compositions with a natural definition of the origin and the summation (adding a perturbation). This space may be accomplished with specific definitions of the inner product and Aitchison distance that estimates the differences in proportions between two samples. Together, they represent a Euclidean space (3). Though the Euclidean geometry is quite intuitive, in our specific case operations with vectors are inconvenient (for example, component-wise division is not a natural way of calculating the difference); it obstructs usage of such common methods as linear regression or linear discriminant analysis.

A solution to the problem has been presented in reference 4. The authors of that study proposed the so-called isometric log-ratio (ILR) transform; this preserves the results of major operations with vectors (e.g., inner product and compositional difference) and converts data to a new space, which allows for further calculations to occur in a natural way. Following the "principle of working in coordinates" (5), the data should be transformed to ILR coordinates, all the analysis should be performed there, and then the inverse conversion should be applied to the results for interpretation.

Such ILR transform may be defined in numerous ways: the rotation of the basis around the origin preserves all required properties. The choice of particular ILR coordinates does not influence the results of the analysis when they are transformed back to the initial space of proportions. This approach is convenient for interpretation of vectors that are treated as compositions rather than perturbations. For example, microbiome composition predicted by linear regression in the ILR coordinates can be easily transformed to a vector of taxa proportions. The vector of differences between two samples is harder to interpret in the initial terms. Its meaning is less intuitive: it describes which proportion of each taxon from the first sample should be taken to obtain the microbiome with the proportions observed in the second sample (Fig. 1A).

In reference 6, the authors suggested a method to define an interpretable system of coordinates. It is based on the idea of using simple perturbations as coordinate vectors. Imagine adding the second, third, and fourth portions of milk to a cup of coffee and leaving the syrup/espresso ratio unperturbed. This may be viewed as perturbations of the initial state in the same direction with increasing effect size (only one component of perturbation will change). Thus, compositions of the coffee will shift along a straight line in the ILR space. The idea may be generalized to a more complex case of perturbing proportions between two groups of taxa and leaving unperturbed all other independent proportions. This relation between two groups is called a "balance" and should be calculated in a specific way: it is a proportion between the abundances of the "mean" representative of each group of taxa. The "mean" is understood as a geometric mean, the proportion is treated in logarithmic scale to make it symmetric relative to the choice of groups in the numerator and in the denominator (see details in Materials and Methods). The basis of the ILR coordinate system is constructed of simple unit perturbations called balancing elements. The choice of orthogonal balancing elements is based on a binary tree with components as leaves; each node of the tree is used to define a balance between two groups of taxa represented by the leaves of its two branches (Fig. 1C). Now, any shift of microbiome can be described as several simple perturbations of known effect size.

The binary tree may be constructed in many ways; the specific choice limits the set of simple perturbations used for interpretation of changes in microbiome. Construction of the tree may be based on the phylogenetic affinity of taxa (7), approximate results of principal coordinates analysis (8, 9), or a difference in phenotype between samples (9, 10).

Almost all compositional methods assume that the choice of the coordinates is made before the statistical analysis. In a more complicated algorithm, the choice of just one balance and the analysis are iterated many times to find the ultimate balance that provides the best accuracy (*selbal* algorithm [11]).

Here, we propose to reverse the order: first, perform the analysis in any arbitrary ILR system of coordinates and only then look for interpretation. We present an algorithm (here called "the nearest balance method") that finds the nearest interpretable direction to the vector of interest. The output format of the algorithm is very close to the results provided by component-wise analysis: a list of taxa positively or negatively associated with a factor.

Figure 1D illustrates how the interpretation is obtained for the coffee example. The first cup differs from the "origin" by a complex shift "v" which cannot be described by a change in a single balance (Fig. 1A). An ILR transform converts proportions to a plane (Fig. 1D); any pair of balances constructed by a binary tree (Fig. 1C) corresponds to two orthogonal directions in this plane and may be used as a coordinate system. A simple procedure described in Materials and Methods allows us to find the balance defining the direction closest to the vector of differences. Here, such a balance is the proportion of milk and coffee to syrup.

The proposed algorithm gives an exact solution. In addition, we describe here two similar algorithms: (i) the construction of an ILR system of coordinates based on the nearest balance principle and (ii) the construction of two orthogonal balances for interpreting two ILR vectors. These three algorithms may be combined with any statistical method that provides ILR vectors as a result. Examples of applications (Fig. 2) include interpreting and visualization of various statistical methods' results, such as classification, linear regression, and principal component analysis.

Importantly, the method provides the nearest balance to a microbiome shift vector but does not guarantee that this balance is the best in terms of the initial optimization problem, such as minimization of the loss.

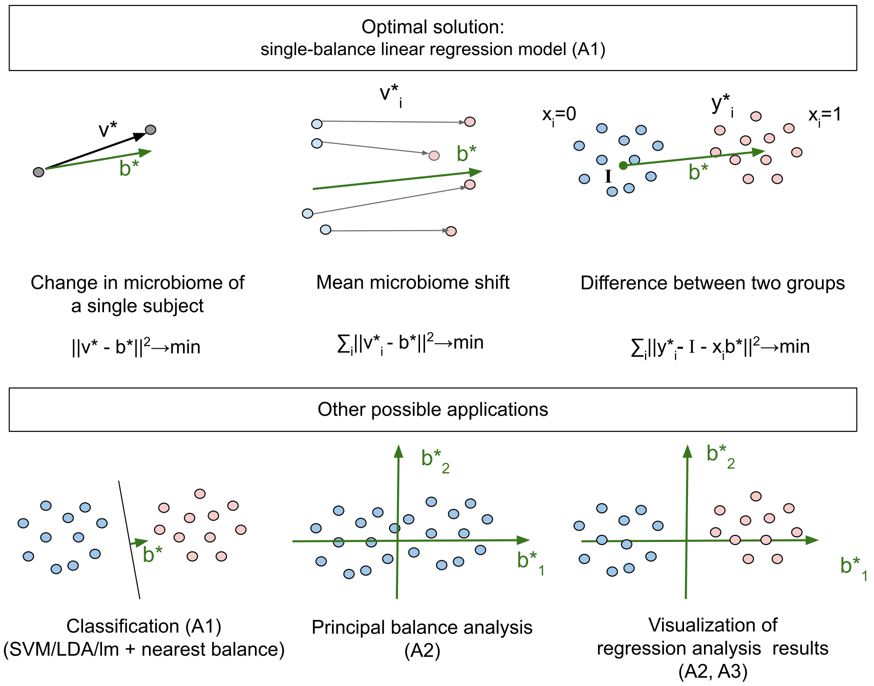

Optimal solution:
single-balance linear regression model (A1)
v*
b*
v*_i
b*
x_i=0
y*_i
x_i=1
I
b*
Change in microbiome of a single subject
||v* - b*||^2→min
Mean microbiome shift
Σ_i||v*_i - b*||^2→min
Difference between two groups
Σ_i||y*_i - I - x_ib*||^2→min
Other possible applications
b*
b*_2
b*_1
b*_2
b*_1
Classification (A1)
(SVM/LDA/lm + nearest balance)
Principal balance analysis
(A2)
Visualization of
regression analysis results
(A2, A3)

**FIG 2** Examples of applications of the nearest balance approach. Here, ||.|| denotes Euclidean norm in the ILR space (which equals the Aitchison norm for the proportions). A1, A2, and A3 denote the applied algorithms from the Materials and Methods section.

However, we show below that the approach provides a least-squares solution for several simple but useful cases. First of all, it is obviously a least-squares approximation of a single subject microbiome shift. To our knowledge, this is the first algorithm to facilitate the interpretation of individual changes. Next, nearest balance provides a least-squares solution for a linear regression model with one of the predictors being associated with a single-balance shift ("single-balance regression"). The method provides accuracy comparable to existing compositional methods for the classification problem and "principal balance analysis" (the search for a sequence of orthonormal balances successively maximizing the explained variance in a data set [8]).

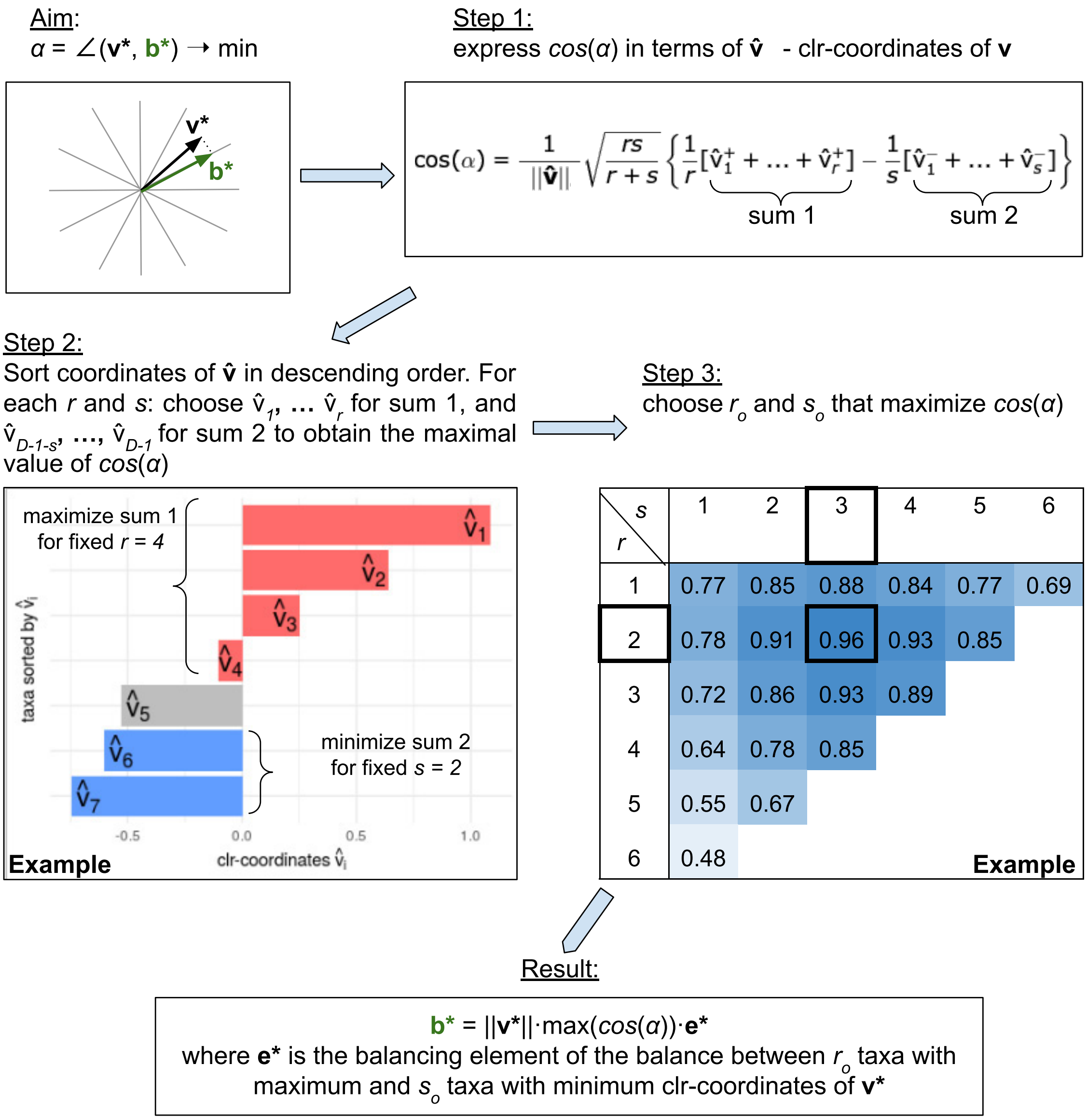
## RESULTS

Here, we propose three algorithms, which are described in detail in Materials and Methods.

The first algorithm (A1) provides the nearest balance "b" to a given microbiome composition shift "v." The second algorithm (A2) is designed for joint interpretation of two ILR-vectors ($v_1$, $v_2$) (such as two first PCA components) by two orthogonal balances ($b_1$ is the nearest to $v_1$; $b_2$ is the nearest balance to $v_2$ among all balances orthogonal to $b_1$). The third algorithm (A3) provides an ILR system of coordinates to decompose a microbiome shift v into a superposition of orthogonal balances ($b_1$, $b_2$, . . ., $b_{D-1}$) sequentially approximating the shift.

The presentation here is organized as follows. In Results we explore several applications: interpretation of individual changes in microbiome composition (A1), regression analysis (A1 for analysis, A3 for validation), classification (A1 for analysis, A2 for visualization, A3 for validation), and principal balance analysis (A2). The Materials and Methods section describes algorithms in detail and contains the proof of optimality for certain applications. The Discussion section summarizes the advantages and disadvantages of the approach and describes directions for future improvements.

**Interpretation of single subjects' microbiome composition changes.** Figure 3 shows the main steps of the main algorithm A1, which provides the nearest balance b to a given microbiome composition shift v. In brief, the cosine between the ILR



Aim:
$\alpha = \angle(\mathbf{v}^*, \mathbf{b}^*) \rightarrow \min$

Step 1:
express $cos(\alpha)$ in terms of $\hat{\mathbf{v}}$ - clr-coordinates of $\mathbf{v}$

$$\cos(\alpha) = \frac{1}{\|\hat{\mathbf{v}}\|}\sqrt{\frac{rs}{r+s}}\left\{\underbrace{\frac{1}{r}[\hat{v}_1^+ + \ldots + \hat{v}_r^+]}_{\text{sum 1}} - \underbrace{\frac{1}{s}[\hat{v}_1^- + \ldots + \hat{v}_s^-]}_{\text{sum 2}}\right\}$$

Step 2:
Sort coordinates of $\hat{\mathbf{v}}$ in descending order. For each $r$ and $s$: choose $\hat{v}_1, \ldots \hat{v}_r$ for sum 1, and $\hat{v}_{D-1-s}, \ldots, \hat{v}_{D-1}$ for sum 2 to obtain the maximal value of $cos(\alpha)$

Step 3:
choose $r_o$ and $s_o$ that maximize $cos(\alpha)$

| r \ s | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 1 | 0.77 | 0.85 | 0.88 | 0.84 | 0.77 | 0.69 |
| 2 | 0.78 | 0.91 | 0.96 | 0.93 | 0.85 | |
| 3 | 0.72 | 0.86 | 0.93 | 0.89 | | |
| 4 | 0.64 | 0.78 | 0.85 | | | |
| 5 | 0.55 | 0.67 | | | | |
| 6 | 0.48 | | | | | |

Result:

$\mathbf{b}^* = \|\mathbf{v}^*\| \cdot \max(cos(\alpha)) \cdot \mathbf{e}^*$
where $\mathbf{e}^*$ is the balancing element of the balance between $r_o$ taxa with maximum and $s_o$ taxa with minimum clr-coordinates of $\mathbf{v}^*$

**FIG 3** Diagram of algorithm A1 to find the nearest balance b* to an ILR vector v*.

components of the shift v* and an arbitrary balance b* is proportional to the difference between the means of centered log-ratio (CLR) coordinates that correspond to the groups of taxa in the numerator and denominator of the balance (step 1 in Fig. 3; see the proof in Materials and Methods). The balance between taxa with *r* maximum and *s* minimum CLR components of microbiome shift v is the closest to the shift (step 2 in Fig. 3). This is the key step of the algorithm, as it allows avoiding a grid search through all balances between fixed-sized groups. A search through all possible *r* and *s* values provides the optimal balance (step 3 and results in Fig. 3).

Although the algorithm exactly finds the closest balance, in practice it has some limitations. Assume the microbiome change is indeed a single-balance change b* and the observed vector of microbiome change v* differs from this single-balance change b* due to technical noise. If the noise level is sufficiently high, the observed change v* may become closer to another balance b*′. In such cases, the nearest balance algorithm will return a wrong balance b*′.

Figure 4 illustrates the effect of the noise level on the ability to reconstruct the correct balance on simulated data. For the simulation, we constructed three ILR systems

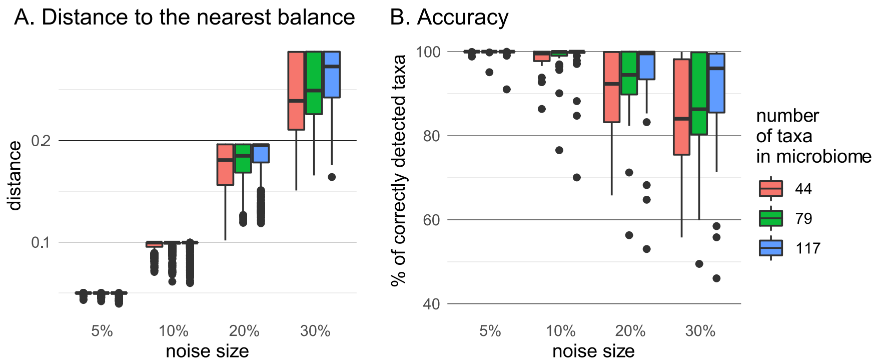


**FIG 4** Stability of the nearest balance search to the noise in observations. (A) Distance from the disturbed vector to the nearest balance. If it is less than the noise size, it means that the disturbed balance becomes closer to a balance different from the original one. (B) Proportion of taxa correctly included in the numerator, denominator, or correctly excluded from the balance.

of coordinates for a healthy human stool microbiome data set; the systems differ by the total number of taxa included in the analysis (see Materials and Methods). Each balancing element was perturbed in an orthogonal direction with a given noise size: 5, 10, 20, and 30% of its length. The procedure was repeated 20 times for each size of perturbation. The nearest balance was identified for each perturbed vector and compared to the initial vector. The accuracy of the method was measured as the proportion of taxa correctly included in the numerator or denominator or correctly excluded from the balance. Distance to the closest balance was compared to the *a priori* known size of the noise (i.e., distance to the disturbed balance).

As expected, the higher the noise, the more often the perturbed vector turns out to be closer not to the initial balancing element, but to some other element (Fig. 4A). The accuracy of detecting balance components decreases as the noise increases, too (Fig. 4B). The result depends on the total number of taxa in composition: in a higher-dimensional space larger noise is acceptable. For the tested values (from 44 to 117 taxa), noise that is $\leq$10% of the initial balance length leads to a rather low ($<$10%) proportion of mistakes.

**Single-balance regression analysis.** Multivariate linear regression in an ILR space is a promising tool for modeling microbiome composition dependence on various scalar factors, such as study group or age. Its coefficients are ILR vectors; they describe direction of changes associated with a unit change of a predictor. The nearest balance method may provide an interpretation of such alterations by changes in a single balance.

It turns out that such interpretation is optimal not only in terms of the difference between the coefficient and the balance but also in terms of the mean squared error of the prediction: the nearest balance is the least-squares estimate of the linear regression coefficient, given that it is restricted to be a balance vector (see Materials and Methods for the proof).

We compared the approach with several other compositional methods during cross-validation on synthetic and real data (see Fig. S1 in the supplemental material). The synthetic data set (1,000 samples per group 20 times randomly split into test and train subsets at 1:1 ratio) simulated a case-control study with a single-balance difference between the groups' mean compositions. The real data set contains microbiomes of healthy subjects and subjects with Crohn's disease (CD; 34 samples per group, 20 times randomly split into test and train subsets at a 1:4 ratio). The case-control design was selected for benchmarking, since it is simple and provides a relatively wide range of applicable compositional methods.

Since none of the existing compositional methods was intended exactly for interpreting linear regression coefficients, we adopted them, respectively. The *balance* R package (9) allows construction of an ILR coordinate system based on differences in microbiome composition between two groups in several ways: *PDBA* (the most

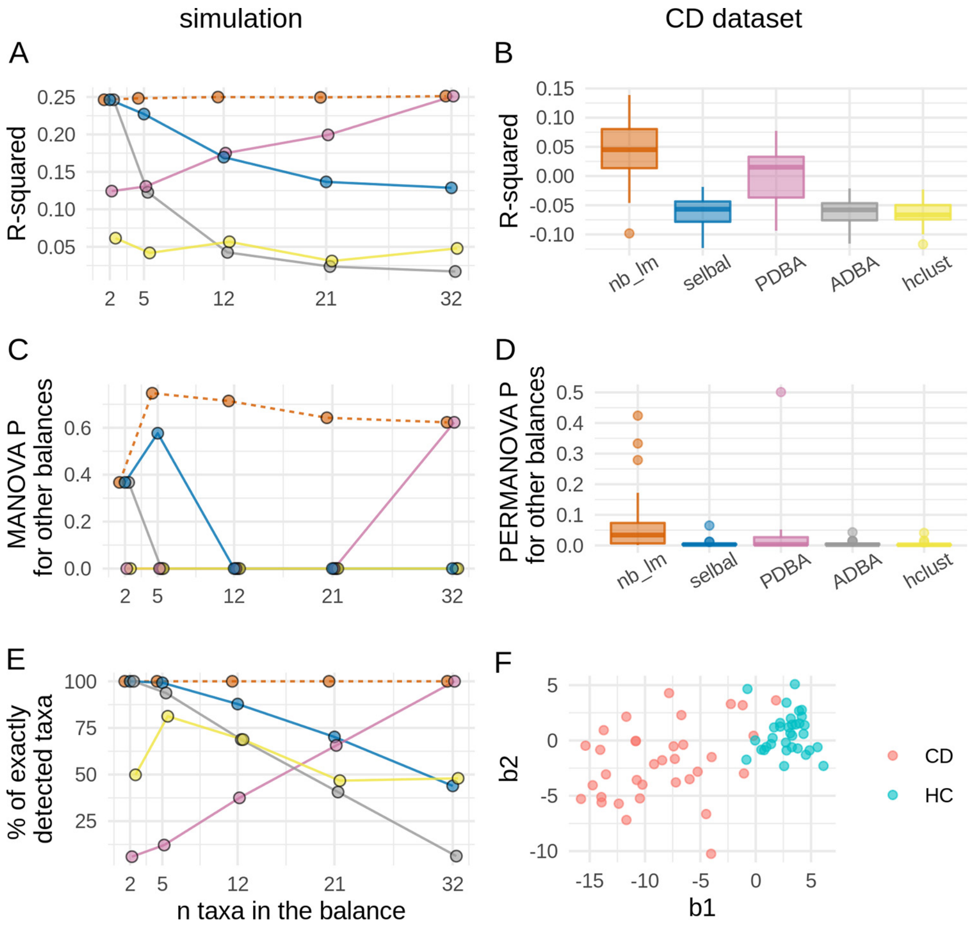


**FIG 5** Results of the cross-validation for the regression problem. The left column illustrates results for the simulation and the right column shows those for the Crohn's disease study (CD, $n = 68$). For the simulation, the mean values through all iterations of cross-validation are plotted; horizontal jitter is added to make points distinguishable. Boxplots are used to illustrate the results on the CD data set. (A and B) Proportions of explained variance on the test set. (C and D) $P$ values for the MANOVA (for simulated data) or PERMANOVA (for CD data) analyses performed on all balances except for the main on the test set; if it is high (>0.05), it means that the main balance incorporates all the differences between two groups. (E) Proportion of taxa that were correctly included in the balance or excluded from it. (F) CD data set in coordinates obtained by *nb_lm* on the whole sample.

differentiating balances include large number of taxa), *ADBA* (the most differentiating balances include small number of taxa), and *hclust* (clustering taxa as it is done in *gneiss* [10]). In addition, we applied the *selbal* algorithm and constructed a random ILR system of coordinates containing the identified balance. For each of the ILR systems, we selected the coordinate with the most explainable by the case-control factor variance. The regression parameters were optimized to describe the difference between groups by change in this single coordinate in the way to provide minimum squared error (see Materials and Methods). We compared the methods by the proportion of explained variance, MANOVA (or PERMANOVA if the sample size was too small) $P$ values for all coordinates except for the main balance and the proportion of correctly detected taxa for the simulation (as described in the previous section).

The proportion of explained variance [the $R^2$ value was measured as 1 – (the mean squared error/the total variance)] is negatively related to the mean squared error; thus, it directly reflects the optimization criterion of linear regression. The nearest balance method guarantees the highest value on the train subset. Figure 5 shows the results on the test subset. The *nb_lm* method showed best results both on simulated and real data sets (Fig. 5A and B). The results of the simulation show that the predictive power of other methods depends on the number of taxa included in the discriminating balance. *ADBA* and *PDBA* coordinates incorporate assumptions about this number and thus work reasonably well if they are satisfied and less accurate otherwise. R-squared for the *selbal* prediction decreased with increasing balance size. This is partly caused by an artificial restriction: by default, the algorithm searches for balances of between ≤20 taxa. Besides the fact that this setting is usually used in practice, another reason to use it was that the increase of the balance size considerably slows down the cross-validation. The *hclust* method showed a relatively low R-squared independently from the number of taxa included in the balance.

The *nb_lm* method was also best for determining the members of the balance (Fig. 5E) on the simulated data. The precision of other balances depended on the balance size.

The MANOVA test was used to test the hypothesis that there is no difference in balances orthogonal to the main one. High *P* values (>0.05) mean that all differences between two groups are incorporated in a single balance. Figure 5C shows that nearest balance was the only method that showed high *P* values independently of the balance size. The sample size of the CD data set was too small to evaluate MANOVA; thus, a nonparametric analogue—PERMANOVA—was applied (Fig. 5D). None of the methods provided a balance incorporating all of the differences between groups in all cross-validation iterations; the *nb_lm* method yielded higher *P* values.

This difference in MANOVA *P* values on simulated and PERMANOVA on the real data may be explained by the difference in the noise size. The noise observed for the shift may be estimated as the standard deviation (SD) of the coefficient of linear regression. For the simulated data, the noise was about 4% of the shift between groups, while for the CD data set it was 25%. As discussed above, the latter value is rather high.

The ILR coordinates constructed by *nb_lm* are useful for the visualization of the regression results (Fig. 5F). For CD data, healthy and control subjects are clearly distinct.

**Classification.** The same simulated and real data sets were used to benchmark classification by a single balance (see the description for Fig. S1 and above). In addition to the methods described above, we performed the nearest balance combined with LDA (*nb_lda*) and SVM (*nb_svm*). The values for the area under the receiver operating characteristic curve (AUC) for simulated data were high for all methods (Fig. 6A). As for the CD data set, all methods except *nb_lda* provided comparable AUC values (Fig. 6B).

In addition, we compared the ILR coordinate systems provided by these methods. For this purpose, the LASSO regression was fitted. A good coordinate choice should result in a high AUC value and a low number of balances remaining in the model. All methods except for *nb_lda* yielded high AUC values on both simulated and real data sets (Fig. 6C and D). The lowest numbers of balances were produced by the *nb_lm*, *nb_svm*, *nb_lda*, and *PDBA* methods on simulated data and by *nb_lm* and *nb_svm* on real data (Fig. 6E and F).

**Principal balance analysis.** We compared application of algorithm 3 to PCA interpretation (*nb_pca*) with several methods of principal balance analysis: "*cluster*" and "*constrained*" implemented in the *pb_basis*() function of R package *coda.base* (8) and anti-principal balance analysis (*ABA*) from the *balance* package on the CD data set. The proportions of variation explained by the first two balances by each method are presented in Table 1. *nb_pca* and *constrained* provided the best variance explained by the first coordinate and, by the first two coordinates, *nb_pca* was slightly better.

**Example: microbiome composition shift associated with Crohn's disease.** To illustrate the biological relevance of our method, we applied single-balance regression to investigate the microbiome shift in patients with Crohn's disease (CD). Comparison data sets from two studies with distinct populations (and likely varying by the experimental protocols) allowed us to evaluate the reproducibility potential of the method.

Data Set 1 is the one used for cross-validation above. Data Set 2 contains microbiome composition data on healthy subjects and patients obtained from reference 12. For consistency, we considered proportions of only those 38 taxa remaining after filtering Data Set 2. Below, the single-balance regression trained on Data Set 1 is denoted model 1, while the one corresponding to the Data Set 2 is denoted model 2.

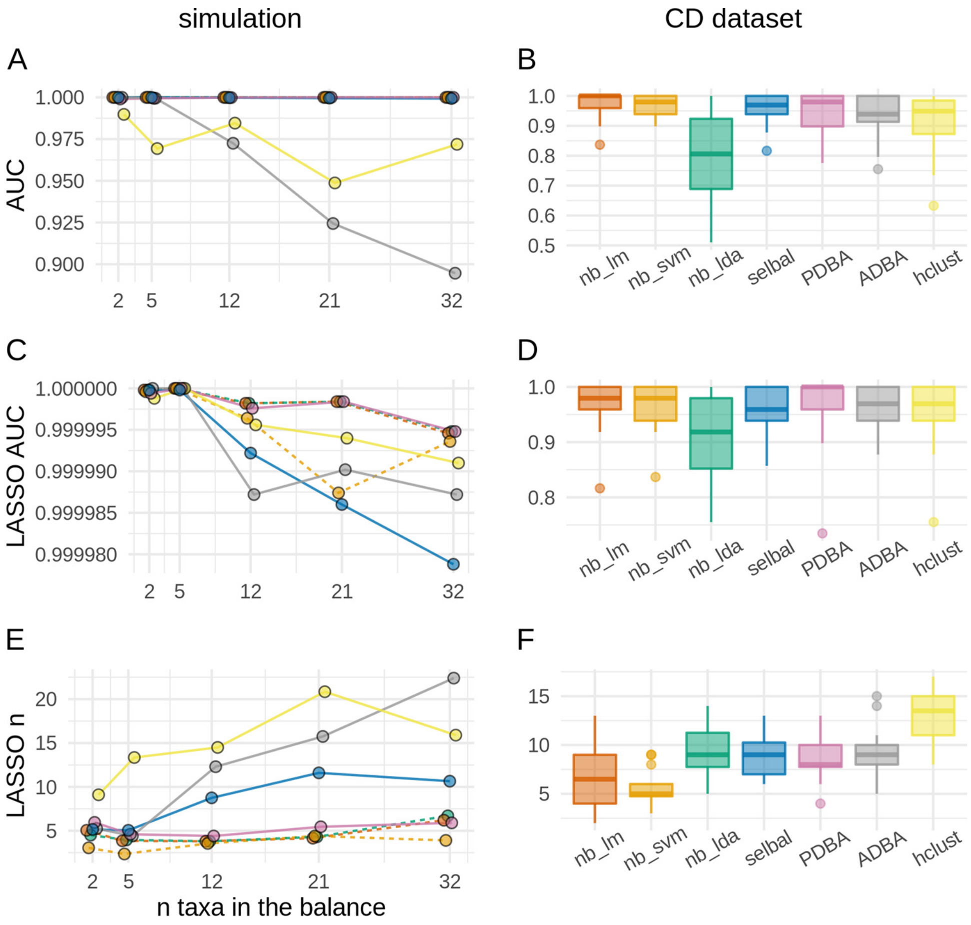


**FIG 6** Results of the cross-validation for the classification problem. The left column illustrates results for the simulation, and the right shows the results for the Crohn's disease study (CD). For the simulation, the mean values through all iterations of cross-validation are plotted. Boxplots are used to illustrate the results on the CD data set. Horizontal jitter is added to the plot to make points more distinguished. (A and B) AUC of classification by a single balance obtained by various methods. (C and D) AUC of classification by a LASSO regression trained in coordinate systems obtained by various methods. (E and F) Number of balances selected by the LASSO regression classifiers.

Figure 7A shows clear differences in ILR space between the two studies, as well as between CD patients (CD) and healthy controls (HC) within each study. The difference between HC and CD samples is clearer in Data Set 1. This fact is supported by the multivariate linear regression analysis: Data Set 2 data set are noisier (36% versus 24%), and the size of the shift is three times smaller (4.1 versus 12.5, Aitchison norm of the shift). As a result, the quality of the single-balance regression model is higher for the less noisy Data Set 1: model 1 has higher R-squared (0.17 versus 0.05, Fig. 7B) and AUC (0.99 versus 0.85) values than model 2. The shift is statistically significant in both data sets ($P$ = 8E–07 in Data Set 1 and $P$ = 6E–08 in Data Set 2, MANOVA for all ILR coordinates). Both models provide MANOVA $P < 0.05$ for the orthogonal to main balance subspace, suggesting that the difference between CD and HC microbiomes is only partially described by a single balance in both data sets. However, the impact of the main balance in this difference is considerable: 83.6% for model 1 and 86.8% for model 2.

Members of the balances obtained via single-balance regression are shown in Fig. 7C. Taxa with an ambiguous classification at the genus level were marked using "/" and augmented with their family name [e.g., "*Lachnospiraceae;(Blautia/unclassified)*"]. Each balance splits taxa into three categories: numerator of the balance, its denominator and taxa excluded from the balance. Only 22 of 38 taxa (57%) are equivalently categorized by the two models. However, the balances are almost noncontradictory except

**TABLE 1** Proportion of variance explained by the first two balances ($b_1$ and $b_2$) obtained by several principal balance analysis methods applied to the Crohn's disease microbiome data set[a]

| Method | $b_1$ (%) | $b_2$ (%) | Sum (%) |
|---|---|---|---|
| *nb_pca* | 25.00 | 9.54 | 34.54 |
| *constrained* | 24.71 | 9.54 | 34.25 |
| *cluster* | 20.52 | 9.96 | 30.48 |
| *ABA* | 8.62 | 7.49 | 16.11 |

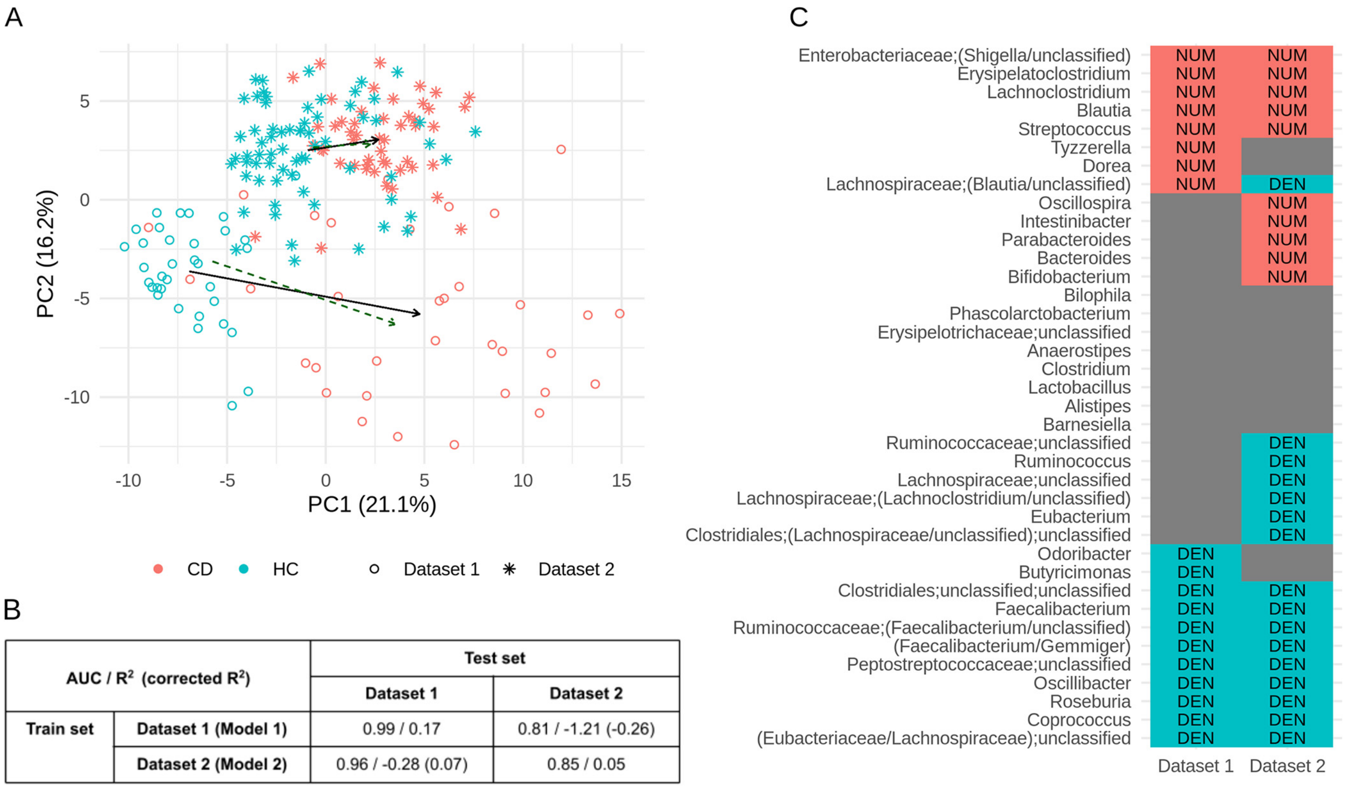
[a] The following notations are used: *nb_pca*, nearest balances approach; *constrained* and *cluster*, methods implemented in R package *coda.base*; *ABA*, anti-principal balance analysis implemented in the R package *balance.*

for one genus-level taxon—"*Lachnospiraceae;(Blautia/unclassified)*"—that is positively associated with the disease in model 1 and negatively associated with the disease in model 2. The noisier Data Set 2 yielded a balance that included a higher number of taxa.

Next, model 1 was cross-tested on Data Set 2, and model 2 was cross-tested on Data Set 1. The predictive power of single-balance regression itself was low (R-squared is negative). It can be in part explained by the systematic shift: R-squared is 0.07 for model 2 tested on the Data Set 1 corrected for this shift. Another source of error is the difference in effect size of the CD-associated shift between Data Sets 1 and 2: as discussed above, it was three times larger in Data Set 2.

The direction of the shift itself had a considerably high predictive power for both models. Model 1 has an AUC of 0.81 for Data Set 2 (versus AUC = 0.85 for model 1), while model 2 has an AUC of 0.96 for Data Set 1 (versus AUC = 0.99 for model 1).

As for the members of balances, the two denominators mostly include genera from the *Clostridiales* order, many of which are known as prominent producers of butyrate, a



| AUC / $R^2$ (corrected $R^2$) | | Test set | |
|---|---|---|---|
| | | Dataset 1 | Dataset 2 |
| Train set | Dataset 1 (Model 1) | 0.99 / 0.17 | 0.81 / -1.21 (-0.26) |
| | Dataset 2 (Model 2) | 0.96 / -0.28 (0.07) | 0.85 / 0.05 |

| | Dataset 1 | Dataset 2 |
|---|---|---|
| Enterobacteriaceae;(Shigella/unclassified) | NUM | NUM |
| Erysipelatoclostridium | NUM | NUM |
| Lachnoclostridium | NUM | NUM |
| Blautia | NUM | NUM |
| Streptococcus | NUM | NUM |
| Tyzzerella | NUM | |
| Dorea | NUM | |
| Lachnospiraceae;(Blautia/unclassified) | NUM | DEN |
| Oscillospira | | NUM |
| Intestinibacter | | NUM |
| Parabacteroides | | NUM |
| Bacteroides | | NUM |
| Bifidobacterium | | NUM |
| Bilophila | | |
| Phascolarctobacterium | | |
| Erysipelotrichaceae;unclassified | | |
| Anaerostipes | | |
| Clostridium | | |
| Lactobacillus | | |
| Alistipes | | |
| Barnesiella | | |
| Ruminococcaceae;unclassified | | DEN |
| Ruminococcus | | DEN |
| Lachnospiraceae;unclassified | | DEN |
| Lachnospiraceae;(Lachnoclostridium/unclassified) | | DEN |
| Eubacterium | | DEN |
| Clostridiales;(Lachnospiraceae/unclassified);unclassified | | DEN |
| Odoribacter | DEN | |
| Butyricimonas | DEN | |
| Clostridiales;unclassified;unclassified | DEN | DEN |
| Faecalibacterium | DEN | DEN |
| Ruminococcaceae;(Faecalibacterium/unclassified) | DEN | DEN |
| (Faecalibacterium/Gemmiger) | DEN | DEN |
| Peptostreptococcaceae;unclassified | DEN | DEN |
| Oscillibacter | DEN | DEN |
| Roseburia | DEN | DEN |
| Coprococcus | DEN | DEN |
| (Eubacteriaceae/Lachnospiraceae);unclassified | DEN | DEN |

**FIG 7** Single-balance regression applied to two studies comparing gut microbiomes of healthy subjects and patients with Crohn's disease. (A) Principal-component analysis (in an ILR space). Arrows indicate predictions of multivariate linear regression (solid black) and single-balance linear regression (dashed green). (B) Results of single-balance linear regressions trained and tested on two data sets. AUC, R-squared, and corrected (for systematic shift between data sets) R-squared results are shown. (C) Members of balances nearest to the CD-associated microbiome shifts for each data set. Red NUM, taxon is in the numerator; blue DEN, in the denominator.

short-chain fatty acid crucial for gut health. Depletion of butyrate production potential is a hallmark of the CD microbiome (13, 14). The list of butyrate producers includes *Faecalibacterium*, *Roseburia*, and *Coprococcus* (15, 16), also previously reported to be negatively associated with CD in noncompositional systematic meta-analyses (14, 17). The mechanism of *Faecalibacterium* participation in intestinal inflammation, particularly during CD, is not limited to butyrate production; the bacteria can produce other microbial anti-inflammatory molecules (18). *Roseburia* is a gut microorganism found to be depleted not only in the microbiome of CD patients but also in healthy subjects with high genetic risk of inflammatory bowel disease (IBD) (19).

On the other hand, the intersection of the two numerators includes taxa for which strong positive associations with the CD and/or other gastrointestinal disorders have been demonstrated in the literature, along with possible mechanisms. For example, the "*Enterobacteriaceae;(Shigella/unclassified)*" taxon is related to *Escherichia coli*, the species often enriched in the guts of CD patients. Its adherent-invasive pathotypes (AIEC) are known to be able to invade epithelial cells in the disease, and a wide genomic diversity of CD-associated *E. coli* strains has been reported as well (13). Other taxa included in the numerator have been implicated in pathologies. For example, *Erysipelatoclostridium*, opportunistic clostridia linked to diseases that include metabolic disorders, gout (20), and IBD (21); in addition, another numerator member, *Lachnoclostridium*, was linked to colorectal adenoma (22) and atherosclerosis (23).

## DISCUSSION

We propose a *post hoc* method to interpret compositional analysis results: the microbiome changes are optimally approximated by a shift in a single balance between two groups of taxa (i.e., the nearest balance). It is compatible with any statistical method that yields a vector in ILR coordinates, such as coefficients of a mixed-effect model, a vector normal to a plane obtained via a linear classifier, PCA axis, or a vector of difference between two compositions.

Such an approach does not guarantee optimality in terms of the initial problem optimization criterion. Thus, we explored the accuracy of several applications—regression, classification problems, and principal balance analysis—with analytical treatment and/or cross-validation.

The approach provides the least-squared error for the single-balance regression problem, i.e., for the search for single-balance changes in a microbiome associated with a factor possibly adjusted for several covariates. Three simple examples of such analysis are the approximation of case-control differences, mean changes in the microbiome of a group of subjects, or a single subject's microbiome shift by a single-balance change. As for the third case, to our knowledge, the nearest balance approach is the first compositional method to solve the problem.

The method is comparable to other compositional approaches for the classification problem and principal balance analysis. Its advantage is lack of a prior on the number of components in the resulting balances and rather low computational complexity (except for algorithm A3).

Application to the gut microbiome in Crohn's disease showed that the method provides biologically meaningful and reproducible results. Two models trained on two data sets from distinct populations provided considerably overlapping balances. These balances reflect a trade-off between commensal butyrate-producing taxa and taxa linked to negative effects on gut health. The single-balance models showed low predictive power in cross-testing (due to differences in populations), but the balances themselves allowed good performance for classifying subjects into healthy ones and patients, suggesting the biological relevance of the results.

The presented method shares the limitations of any compositional analysis: the ILR transform is not compatible with zero abundance values, and the total sum of counts per sample (and thus the information about the precision of the measurement) is lost while converting to proportions. In addition, the amalgamation of taxa into groups presents a separate issue that can influence the results (24). Another limitation concerns the assumption that the nearest balance reveals the biological interpretation of the vector. It is violated if the approximated vector is a superposition of changes in several, possibly nonorthogonal directions. This problem may be partially solved by a careful choice of covariates in the model.

The implementation needs further improvement. First, if several balancing elements are equally close to the approximated vector, only one of them is detected. As a consequence, the search for the second balance orthogonal to the first one, as well as the search for the nearest balance tree, may be suboptimal. Second, the search for two orthogonal balances is based on the assumption that they correspond to the nodes of the same binary tree; other possibilities have not yet been explored.

To summarize these findings, the nearest balance is a compositional method providing optimal solutions for single-balance regression and extending the opportunities for classification, principal balance analysis, and other methods of treating compositional data.

## MATERIALS AND METHODS

**Basic definitions.** The detailed theory of compositional analysis may be found in reference 3. Here, we list several definitions and properties essential for understanding this method.

Composition is a vector that contains the information about relative proportions of components, while the total sum is of no concern. Normalization of a composition to a certain total sum is called a closure operation.

An ILR transform should be applied for the correct treatment of such data. It converts data to a Euclidean space where operations with compositions are calculated as common operations with vectors. The ILR transform is not unique: its particular form yields a choice of coordinates in the space; any orthogonal system is appropriate for the analysis provided its origin corresponds to the composition with equal proportions of all parts. One of the ways to define such a transform is based on the construction of a binary tree with features as leaves (in case of microbiome data the features are microbial taxa). ILR coordinates of a vector v are calculated as balances that correspond to each node *i* of the tree (Fig. 1C) in the following way:

$$ilr(\mathrm{v}) = \mathrm{v}^* = (\mathrm{v}^*{}_1, \mathrm{v}^*{}_2, \ldots \mathrm{v}^*{}_{D-1})$$

$$\mathrm{v}^*{}_i = \sqrt{\frac{rs}{r+s}} \ln \frac{\mathrm{g}_m(\mathrm{Num}_i)}{\mathrm{g}_m(\mathrm{Den}_i)}$$

where $\mathrm{Num}_i$ and $\mathrm{Den}_i$ denote two sets of leaves that correspond to one of the child branches of node *i*, $\mathrm{g}_m$ denotes the geometric mean of relative abundances of taxa that correspond to these leaves and *D* is the dimension of the initial space of proportions (i.e., the total number of taxa). The balances may be considered independently of the ILR system construction, since they characterize the relations between groups of taxa in a composition. The change of a balance defines a direction in the Euclidean space; a unit vector in this direction is called a balancing element.

The centered log-ratio (CLR) transform (which is the reverse *softmax* function) is another transform that preserves all operations with compositional vectors and converts data to a Euclidean space. Components of the transformed vector are calculated as follows:

$$clr(\mathrm{v}) = \hat{\mathrm{v}} = (\hat{\mathrm{v}}_1, \hat{\mathrm{v}}_2, \ldots, \hat{\mathrm{v}}_D)$$

$$\hat{\mathrm{v}}_i = \ln(\mathrm{v}_i) - \sum_{j=1}^{D} \ln(\mathrm{v}_j)/D$$

Unlike the ILR transform, CLR transform does not reduce the dimension of vectors, leaving the problems of components' interdependence unresolved; thus, this method is appropriate for statistical treatment of distances between compositions, but not for component-wise analysis.

The ILR and CLR coordinates are related in the following way:

$$\mathrm{v}^* = \hat{\mathrm{v}}\Psi^{\mathrm{T}}$$

$$\hat{\mathrm{v}} = \mathrm{v}^*\Psi$$

where the superscript "T" denotes transposition, $\Psi$ is a $(D - 1) \times D$ matrix with CLR coordinates of the basis vectors of the ILR system in the rows. Each basis vector represents a balancing element $\mathrm{e}^* = (\mathrm{e}^*{}_1, \ldots, \mathrm{e}^*{}_{K-1})$. Its CLR coordinates $\hat{\mathrm{e}} = (\hat{\mathrm{e}}_1, \ldots, \hat{\mathrm{e}}_{D-1})$ are equal to:

$$a_+(r,s) = \frac{1}{r}\sqrt{\frac{rs}{r+s}}$$

for *r* components corresponding to the numerator of e*,

$$a_-(r,s) = -\frac{1}{s}\sqrt{\frac{rs}{r+s}}$$

for *s* components corresponding to its denominator and 0 for all others.

Since both CLR and ILR transforms preserve the distance and inner product, the following equalities are valid for any two compositions v and w:

$$\langle \hat{v}, \hat{w} \rangle = \langle v^*, w^* \rangle$$

$$\|\hat{v}\| = \|v^*\|.$$

where <,> denotes the inner product, and ||.|| denote the Euclidean norm in CLR or ILR space.

**Suggested algorithms. (i) Algorithm A1 (ILR vector → the nearest balance).** Here, the nearest balancing element to a vector v* is understood as the balancing element that includes the minimum angle to v* in an ILR space. Since the balancing element is a vector of unit length, the cosine between a balancing element e* and vector v* may be calculated as follows:

$$\cos(\alpha) = \frac{\langle v^*, e^* \rangle}{\|v^*\|} = \frac{\langle \hat{v}, \hat{e} \rangle}{\|\hat{v}\|}$$

It reaches its maximum value when $\langle v^*, e^* \rangle_{ilr}$ is maximum. The coordinates of $\hat{v}$ are calculated as $v^*\Psi$ and CLR coordinates $\hat{e}$ of element e take one of the values {0; $a_+(r,s)$; $a\_(r,s)$}. The cosine of the angle between the balancing element and the vector is described by the formula:

$$\cos(\alpha) = \frac{1}{\|\hat{v}\|}\sqrt{\frac{rs}{r+s}}\left\{\frac{1}{r}\left[\hat{v}_1^+ + \dots + \hat{v}_r^+\right] - \frac{1}{s}\left[\hat{v}_1^- + \dots + \hat{v}_s^-\right]\right\}$$

where $\hat{v}^+{}_i$ values are *r* components corresponding to the taxa in the numerator of e*, and $\hat{v}_i^-$ values are *s* components related to its denominator. For a fixed *r* and *s*, the expression reaches the maximum value when the sum $\left[\hat{v}_1^+ + \dots + \hat{v}_r^+\right]$ is the maximum and $\left[\hat{v}_1^- + \dots + \hat{v}_s^-\right]$ is the minimum. For fixed *r* and *s*, the maximum sum of *r* components is provided by summing the *r* maximum components of vector $\hat{v}$, and the minimum sum of *s* components is provided by summing its *s* minimum components. If the components of $\hat{v}$ are descendingly ordered, the minimum cos($\alpha$) equals:

$$\cos(\alpha) = \frac{1}{\|\hat{v}\|}\sqrt{\frac{rs}{r+s}}\left\{\frac{1}{r}\left[\hat{v}_1 + \dots + \hat{v}_s\right] - \frac{1}{s}\left[\hat{v}_{D+1-s} + \dots + \hat{v}_D\right]\right\}.$$

Thus, for each *r* and *s*, the nearest balancing element to v* may be obtained by uniting *r* taxa with maximum CLR coordinates in one group and *s* taxa with minimum coordinates in the other group. Now, a simple search through all possible values of *r* and *s* leads to the optimal solution.

Vector v* may be approximated by its projection on the nearest balancing element. We call this projection b*, the nearest balance. It has the least possible angle to v* and is the closest to it among all balance vectors colinear to e*, thus b* satisfies the equality:

$$b^* = \underset{\beta \in \mathcal{B}^{D-1}}{\operatorname{argmin}} \|\beta - v^*\|^2,$$

where $\mathcal{B}^{D-1}$ denotes a set of all balance vectors in $\mathbb{R}^{D-1}$ (the ILR space).

The impact of b* to the total vector v* is estimated as $\|b^*\|^2/\|v^*\|^2$: the nearest balancing element may be considered as a basis vector of an ILR system, and this definition gives a unit total sum for the impacts of all axes.

**(ii) Algorithm A2 (two ILR vectors → two orthogonal balances).** Let us find balancing element $e^*_2$ that is the closest to an ILR-vector w* among all balancing elements that may be observed in one and the same binary tree with $e^*_1$ (and thus orthogonal to it [6]). Vector w* can be equal to v*, if the aim is a more exact approximation, but it can be a new vector as well if the aim is, for example, an approximation of the PCA axes.

The constraint of $e^*_1$ and $e^*_2$ being from the same binary tree implies that one of the following conditions must be satisfied (see Fig. S2): (case A) node $e^*_2$ is inside one of the child branches of node $e^*_1$; (case B) node $e^*_1$ is inside one of the child branches of node $e^*_2$; and (case C) child branches of nodes $e^*_1$ and $e^*_2$ have no common leaves (the nodes are named after the balancing elements).

In cases A and C, nonzero CLR coordinates of $e^*_2$ are related only for a certain group of taxa: either the numerator or denominator of the first balance in case A and all other taxa in case C. Reduction to a subspace of all nonzero $e^*_2$ components preserves the inner product with w*. Thus, the best balancing element among each of these groups may be found in the same manner that is described in the previous section: w* coordinates that correspond to certain subspace are ordered decreasingly, and inner products of $e^*_2$ and w* are calculated for all variants of *r* and *s* and then compared.

In case B, the procedure is slightly more complicated, since all components related to $e^*_1$ should be considered together. For example, an algorithm to search for the balance containing all taxa related to $e^*_1$ in its numerator includes the following steps: (i) order decreasingly components of $\hat{w}$ that were not included in the first balance; (ii) place the remaining taxa at the beginning of the list; (iii) calculate the inner product with $\hat{w}$ for each *r* and *s* similarly as described for the first balance and $\hat{v}$, with the only difference that only values of *r* greater than the number of taxa in the first balance should be considered; and (iv) choose the optimal *r* and *s*. The first two steps ensure that the condition B is satisfied; the second and third steps provide maximum inner product for the given *r* and *s*. The algorithm of searching for the balance that contains all taxa from the first balance in its denominator is similar; the only difference is that the components that correspond to the first balance should be placed at the end of the list.

Text S1 shows that $b^*_2$ is the least-squares approximation of w* by its superposition with a given orthogonal balance $b^*_1$ among all orthogonal to $b^*_1$ balances.

**(iii) Algorithm A3 (an ILR vector → the nearest balance tree).** The aim of this algorithm is to construct an ILR system of coordinates with basis $\{e^*_1, e^*_2, \ldots, e^*_{D\text{-}1}\}$, such that $e^*_1$ is the nearest one to vector v*, the $e^*_2$ being the nearest to v* among all elements orthogonal to $e^*_1$, the third element being the nearest to v* among all balances orthogonal to $e^*_1$ and $e^*_2$ and so on. This algorithm is constructed recursively similar to algorithm A2: the taxa are stepwise split into groups to provide orthogonality of balances, on each step the cosine is maximized for fixed *r* and *s* (see Fig. S3), and the optimal values are chosen. The main difficulty is in the choice of elements that maximize the cosine for a given *r* and *s*. Similarly to A2, some elements are united in a branch on previous steps and should be included or excluded from the numerator or denominator of the new balance together. In the case of A2 the number of groupings is 1, but in this algorithm it may be larger. To resolve this difficulty, at each step we substitute components of $\hat{v}$ that are included in the same branch by their means and obtain a new vector $\hat{v}'$. Next, we assign weights to its components: "1" for the ones that came from the original vector $\hat{v}$ and the number of substituted components for the others. The sum of $\hat{v}$ components equals the weighted sum of $\hat{v}'$ components. A simple search through all combinations of weights summing to *r* or *s* gives the maximal cosine with $\hat{v}$ among all balances orthogonal to the previously found ones (for fixed *r* and *s*). The *partitions* package (25) is used to implement the search for possible combinations of weights. Figure S3 provides a detailed algorithm scheme, and Text S3 illustrates this with a simple example.

**Single-balance linear regression.** Consider a linear dependence between ILR coordinates of a *D*-part composition y* (e.g., microbiome composition consisting of *D* taxa in various proportions represented by a vector with (*D*-1) ILR components), a univariate predictor x (e.g., a dosage of a medicine or a binary variable representing presence of a certain disease) and several univariate covariates $z_1, \ldots, z_K$ (e.g., $K = 2$, $z_1$ is age, and $z_2$ is an integer value representing gender of a participant):

$$y^* = xv^* + z_1 a^*_1 + \ldots + z_K a^*_K.$$

Coefficients of this linear dependence v* and $a^*_1, \ldots, a^*_K$ are (*D*-1)-dimensional vectors in the ILR space. Each of them represents an ILR shift of composition y* associated with a unit change of a predictor or a covariate (e.g., presence of a disease, a unit change of a medicine dosage, or a unit change in age). These coefficients are often an object of interest. In practice they may be estimated basing on known values of the composition y*, predictor x, and covariates $z_1, \ldots, z_K$ in a sample and assuming a random measurement error in observations.

The ordinary multivariate linear regression model assumes that composition of each sample in ILR space is a random vector from a multivariate normal distribution with a mean being linear dependent on the predictor and the covariates, i.e., for each $i = 1, \ldots, N$:

$$y^*_i \sim \mathrm{N}(x_i v^* + z_{1i} a^*_1 + \ldots + z_{Ki} a^*_K, \sigma^2 \mathbb{I})$$

or in matrix representation:

$$y^*_i \sim N(x_i v^* + A^* z_i, \sigma^2 \mathbb{I})$$

where $y^*_i$ is an (*D*-1)-dimensional ILR vector representing microbial composition of the $i$th sample, $x_i$ is the predictor's value for the sample, and $z_i = [z_{1i}, \ldots, z_{Ki}]^T$ is the vector of covariates' values for it, $\sigma^2$ is the variance parameter for the model, *N* is the sample size, and N(⋯,⋯) denotes a multivariate normal distribution with a given mean vector and covariance matrix.

The suggested single-balance linear regression model assumes a similar distribution but restricts the predictor's coefficient to be a balance vector. It means that a unit change of the predictor is associated with a change in a single balance between two groups of parts (taxa):

$$y^*_i \sim N(x_i b^* + A^*_{sb} z_i, \sigma^2_{sb} \mathbb{I}),$$

where b* is a single-balance shift associated with the predictor x, $A^*_{sb}$ is the matrix of covariates' coefficients, and $\sigma^2_{sb}$ is the variance parameter for the model.

Below we use matrix representation of the sample's characteristics: $Y^* = [y^*_1, \ldots, y^*_N] = [y^*_{ji}]$ is a (*D*-1) × *N* matrix with observed ILR coordinates of the compositions, $Z = [z_1, \ldots, z_N] = [z_{ki}]$ is a $K \times N$ matrix of covariates values, X = [x$_1$, . . ., x$_N$]$^T$ is a 1×*N* matrix of the predictor's values, *j* = 1, . . ., (*D*-1) is an index of ILR coordinate, and *k* = 1, . . ., *K* is an index of a covariate. Using these notifications, we propose a statement about relationship between least-squares solutions of an ordinary multivariate linear regression model and a single-balance regression model.

*Theorem*. Let v*$_{\text{ls}}$ be the least-squares estimate of coefficient v*, $\beta_{\text{ls}}$ be nearest to the v*$_{\text{ls}}$ balance vector, and $\mathcal{A}_{ls}$ = (Y*– $\beta_{\text{ls}}$X$^{\text{T}}$)Z$^T$ (ZZ$^{\text{T}}$)$^{-1}$. Then, [$\beta_{\text{ls}}$, $\mathcal{A}_{ls}$] is the least-squares estimate of the single-balance regression coefficients [b*, *A**$_{\text{sb}}$].

The proof of the theorem in the most common case is provided in Text S2. Below, two simple cases are described: the approximation of the mean shift by a single-balance change and a single-balance model with a single predictor.

It is worth noting that the least-squares solution for the single-balance model coincides with the maximum-likelihood solution for the same reason that it is true for the ordinary linear regression with diagonal covariance matrix.

**(i) Approximation of the mean shift.** If x$_{\text{i}}$ is 1 for all samples and the covariates are absent, the linear regression and single-balance regression reduce to models:

$$\mathrm{v}^*{}_i \sim N(\mathrm{v}^*, \sigma^2\mathbb{I})$$

and

$$\mathrm{v}^*{}_i \sim N(\mathrm{b}^*, \sigma^2_{sb}\mathbb{I})$$

The least-squares (and maximum likelihood) estimate of v* is the mean in the sample. The least-squares estimate of b* is a balance vector $\beta_{\text{ls}}$ that satisfies the following:

$$\beta_{\mathbf{ls}} = \underset{\beta\in\mathcal{B}^{D-1}}{\operatorname{argmin}} \sum\nolimits_{i=1}^{N} \|\mathrm{v}^*{}_i - \beta\|^2.$$

Since the argmin is independent of multiplying the function by a positive number and adding a constant,

$$\beta_{\mathbf{ls}} = \underset{\beta\in\mathcal{B}^{D-1}}{\operatorname{argmin}} \sum\nolimits_{i=1}^{N} (\mathrm{v}^*{}_i - \beta)^T(\mathrm{v}^*{}_i - \beta) =$$

$$= \underset{\beta\in\mathcal{B}^{D-1}}{\operatorname{argmin}} \sum\nolimits_{i=1}^{N} \left(\mathrm{v}^{*\boldsymbol{T}}_i\mathrm{v}^*{}_i - 2\beta^{\boldsymbol{T}}\mathrm{v}^*{}_i + \beta^{\boldsymbol{T}}\beta\right) =$$

$$= \underset{\beta\in\mathcal{B}^{D-1}}{\operatorname{argmin}} \left[N\beta^{\boldsymbol{T}}\beta - 2\beta^{\boldsymbol{T}}\sum\nolimits_{i=1}^{N}\mathrm{v}^*{}_i\right] =$$

$$= \underset{\beta\in\mathcal{B}^{D-1}}{\operatorname{argmin}} \left[\beta^{\boldsymbol{T}}\beta - 2\beta^{\boldsymbol{T}}\sum\nolimits_{i=1}^{N}\mathrm{v}^*{}_i/N\right] =$$

$$= \underset{\beta\in\mathcal{B}^{D-1}}{\operatorname{argmin}} \left(\beta - \sum\nolimits_{i=1}^{N}\mathrm{v}^*{}_i/N\right)^T\left(\beta - \sum\nolimits_{i=1}^{N}\mathrm{v}^*{}_i/N\right) =$$

$$= \underset{\beta\in\mathcal{B}^{D-1}}{\operatorname{argmin}} \|\beta - \sum\nolimits_{i=1}^{N}\mathrm{v}^*{}_i/N\|^2.$$

The vector $\sum_{i=1}^{N} \mathrm{v}^*{}_i/N$ is the mean of the sample. Thus, the nearest balance to the vector of the mean change is the least-squares estimate of the mean change among all balance vectors.

**(ii) Least-squares estimate for linear regression with a single predictor.** For the case of a single predictor, the linear regression and the single-balance model are written as

$$\mathrm{y}^*{}_i \sim N(\mathrm{a}^* + \mathrm{x}_i\mathrm{v}^*, \sigma^2\mathbb{I})$$

and

$$\mathrm{y}^*{}_i \sim N(\mathrm{a}^*{}_{sb} + \mathrm{x}_i\mathrm{b}^*, \sigma^2_{sb}\mathbb{I}).$$

*Theorem*. Let v*$_{\mathbf{ls}}$ be the least-squares estimate of v*, $\beta_{\text{ls}}$ be nearest to v*$_{\text{ls}}$ balance vector and $\mathrm{a}_{\text{ls}} = \overline{y}^* - \overline{x}\beta_{\text{ls}}$, where $\overline{y}^* = \sum_{i=1}^{N}\mathrm{y}_i/N$ and $\overline{x} = \sum_{i=1}^{N}\mathrm{x}_i/N$ are the observation and predictor means. Then [$\beta_{\text{ls}}$, a$_{\text{ls}}$] is the least-squares estimate of the single-balance regression coefficients [b*, a*$_{\text{sb}}$].

*Proof.* For an arbitrary estimate of single-balance regression coefficients the residual sum of squares (RSS) is as follows:

$$\mathrm{RSS}(\mathrm{a},\boldsymbol{\beta}) = \sum_{i=1}^{N} \|\mathrm{y}^*{}_i - \mathrm{a} - \mathrm{x}_{\boldsymbol{i}}\boldsymbol{\beta}\|^2 = \sum_{i=1}^{N} \left(\mathrm{y}^*{}_i - \mathrm{a} - \mathrm{x}_{\boldsymbol{i}}\boldsymbol{\beta}\right)^T \left(\mathrm{y}^*{}_i - \mathrm{a} - \mathrm{x}_{\boldsymbol{i}}\boldsymbol{\beta}\right)$$

Zero value of partial derivatives of the function by a* is a necessary condition of its minimum:

$$\frac{\partial \mathrm{RSS}}{\partial \mathrm{a}}(\mathrm{a}_{\mathrm{ls}}, \boldsymbol{\beta}_{\mathrm{ls}}) = -2\sum_{i=1}^{N} \left(\mathrm{y}^*{}_i - \mathrm{x}_i\boldsymbol{\beta}_{\mathrm{ls}}\right) + 2N\mathrm{a}_{\mathrm{ls}} = 0,$$

where 0 is a zero vector. This leads to a relation between the coordinates of the least-squares estimates of the coefficients:

$$\mathrm{a}_{\mathrm{ls}} = \overline{\mathrm{y}}^* - \overline{\mathrm{x}}\boldsymbol{\beta}_{\mathrm{ls}}.$$

Substituting the expression to RSS leads to requirement

$$\boldsymbol{\beta}_{\mathrm{ls}} = \underset{\boldsymbol{\beta}\in\mathcal{B}^{D-1}}{\operatorname{argmin}} \sum_{i=1}^{N} \|\delta\mathrm{y}^*{}_i - \delta\mathrm{x}_i\boldsymbol{\beta}\|^2 =$$

$$= \underset{\boldsymbol{\beta}\in\mathcal{B}^{D-1}}{\operatorname{argmin}} \sum_{i=1}^{N} \left[\left(\delta\mathrm{y}^*{}_i\right)^T \delta\mathrm{y}^*{}_i + (\delta\mathrm{x}_i)^2 \boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{\beta} - 2\delta\mathrm{x}_i\boldsymbol{\beta}^{\boldsymbol{T}}\delta\mathrm{y}^*{}_i\right] =$$

$$= \underset{\boldsymbol{\beta}\in\mathcal{B}^{D-1}}{\operatorname{argmin}} \left\|\boldsymbol{\beta} - \left(\sum_{i=1}^{N} \delta\mathrm{x}_i\delta\mathrm{y}^*{}_i\right) \Big/ \left(\sum_{i=1}^{N} (\delta\mathrm{x}_i)^2\right)\right\|^2$$

where $\delta\mathrm{y}^*{}_i = \mathrm{y}^*{}_i - \overline{\mathrm{y}}^*$ and $\delta\mathrm{x}_i = \mathrm{x}_i - \overline{x}$ are deviations of the observation and predictor values from the means. Here, we do not use a partial derivative the way it was done for $\partial RSS/\partial \mathrm{a}$. The reason is that $\beta$ is restricted to be a balance vector; thus, *RSS* is not a continuous function of $\beta$.

Similar calculations give the well-known least-squares estimate of the ordinary linear regression coefficient v*:

$$\mathrm{v}^*{}_{\mathrm{ls}} = \underset{\mathbf{v}*\in\mathbb{R}^{D-1}}{\operatorname{argmin}} \|\mathrm{v}^* - \left(\sum_{i=1}^{N} \delta\mathrm{x}_i\delta\mathrm{y}^*{}_i\right) / \left(\sum_{i=1}^{N} (\delta\mathrm{x}_i)^2\right)\|^2 =$$

$$= \left(\sum_{i=1}^{N} \delta\mathrm{x}_i\delta\mathrm{y}^*{}_i\right) / \left(\sum_{i=1}^{N} (\delta\mathrm{x}_i)^2\right)$$

Thus,

$$\mathrm{a}_{\mathrm{ls}} = \overline{\mathrm{y}}^* - \overline{\mathrm{x}}\boldsymbol{\beta}_{\mathrm{ls}},$$

$$\boldsymbol{\beta}_{\mathrm{ls}} = \underset{\boldsymbol{\beta}\in\mathcal{B}^{D-1}}{\operatorname{argmin}} \|\boldsymbol{\beta} - \mathrm{v}^*{}_{\mathrm{ls}}\|^2$$

The last equation means that $\beta_{\mathrm{ls}}$ is the nearest to the $\mathrm{v}^*{}_{\mathrm{ls}}$ vector.

*End of the proof*.

**Data sets. (i) Simulated data for exploitation of stability of the nearest balance method.** Microbiome composition profiles of stool samples collected before and after short-term high-fiber dietary intervention were taken as a start point (26) ($n$ = 368). Three types of filtration were applied to the data: taxa present at the level of ≥2 reads in ≥5, 20, or 50% of samples were included in the analysis. Zeros replacement was done with the *cmultRepl*() function from the *zCompositions* package. Principal balance analysis using the hierarchical clustering of components method was applied to construct an ILR system of coordinates. The mean shift in the coordinates was used as the disturbed vector. Noise was simulated from a multivariate normal distribution with zero mean and identity covariate matrix in the subspace orthogonal to the mean shift. The noise was resized to a certain length: 5%, 10%, 20%, or 30% of the mean shift length. The disturbed vector was calculated as the sum of the mean shift and the noise. For each noise size, 20 disturbed vectors were generated.

**(ii) Simulated data for cross-validation.** A case-control study observations were simulated for cross-validation. The HIV data set (27) included in the *selbal* package was used as the ground for simulation. The low-abundance taxa were filtered from the read counts table; those with >15 reads in >30% of samples were included in the analysis. An ILR system of coordinates was constructed by the PDBA method from the *balance* package. A multivariate linear regression with HIV_Status and MSM predictors was fitted in these coordinates. The prediction of the model for subjects with negative HIV status and "nonMSM" value of MSM factor were used as the mean microbiome composition for one of the simulated groups. The mean value in the other group was obtained by a shift of a single balance in the direction of one of the balancing elements of the ILR system. The size of the shift was equal to the one associated with the MSM factor estimated by the regression model. One thousand samples were simulated for each of the groups from the multivariate normal distribution with dispersion parameter from the linear regression model.

**(iii) CD data sets.** The reads for stool samples of healthy subjects and subjects with Crohn's disease were taken from references 28 and 12. In reference 28, only samples from the Spanish cohort collected at the first time point were included (34 healthy subjects versus 34 patients with CD). In reference 12, only the samples from the Chinese cohort with a single sample per subject were used, and patients under infliximab treatment were excluded from the analysis. The tables of counts were obtained as described previously (29). Only taxa that were present at a level of $>$5 reads in $\geq$50% of samples of the Spanish cohort were included in the analysis. The zero replacement was done with the *cmultRepl*() function from the *zCompositions* package.

**Data availability.** The code for simulations and cross-validation is available at https://bitbucket.org/knomics/nearest_balance_for_paper. The algorithms are implemented in the R library NearestBalance (https://bitbucket.org/knomics/nearestbalance).

## SUPPLEMENTAL MATERIAL

Supplemental material is available online only.

**TEXT S1**, DOCX file, 0.01 MB.

**TEXT S2**, DOCX file, 0.02 MB.

**TEXT S3**, DOCX file, 0.02 MB.

**FIG S1**, EPS file, 0.5 MB.

**FIG S2**, EPS file, 0.02 MB.

**FIG S3**, PDF file, 0.05 MB.

**FIG S4**, EPS file, 0.1 MB.

**TABLE S1**, DOCX file, 0.01 MB.

**TABLE S2**, DOCX file, 0.01 MB.

**TABLE S3**, DOCX file, 0.01 MB.

## ACKNOWLEDGMENTS

A.V.T. and N.S.K. were supported by grant 075-15-2019-1661 from the Ministry of Science and Higher Education of the Russian Federation.

We thank Nikita Dovidchenko and Evgenia Borisenko for helpful discussions.

Author contributions were as follows: V.E.O., idea, implementation and testing of the method, and text writing; N.S.K., idea and implementation of the algorithm A3, biological interpretation, and text writing; A.V.T., impact in the conception and its application to microbiome, biological interpretation, and text writing.

## REFERENCES

1. Gloor GB, Macklaim JM, Pawlowsky-Glahn V, Egozcue JJ. 2017. Microbiome datasets are compositional: and this is not optional. Front Microbiol 8. https://doi.org/10.3389/fmicb.2017.02224.
2. Quinn TP, Erb I, Richardson MF, Crowley TM. 2018. Understanding sequencing data as compositions: an outlook and review. Bioinformatics 34: 2870–2878. https://doi.org/10.1093/bioinformatics/bty175.
3. Pawlowsky-Glahn V, Egozcue J, Tolosana-Delgado R. 2015. Modelling and analysis of compositional data. John Wiley & Sons, Ltd, London, United Kingdom. https://doi.org/10.1002/9781119003144.
4. Egozcue JJ, Pawlowsky-Glahn V, Mateu-Figueras G, Barcelo-Vidal C. 2003. Isometric logratio transformations for compositional data analysis1. Mathemat Geol 35:279–300. https://doi.org/10.1023/A:1023818214614.
5. Pawlowsky-Glahn V, Buccianti A. 2011. Compositional data analysis: theory and applications. John Wiley & Sons, Ltd, London, United Kingdom.
6. Egozcue JJ, Pawlowsky-Glahn V. 2005. Groups of parts and their balances in compositional data analysis. Math Geol 37:795–828. https://doi.org/10.1007/s11004-005-7381-9.
7. Silverman JD, Washburne AD, Mukherjee S, David LA. 2017. A phylogenetic transform enhances analysis of compositional microbiota data. Elife 6. https://doi.org/10.7554/eLife.21887.
8. Martín-Fernández JA, Pawlowsky-Glahn V, Egozcue JJ, Tolosona-Delgado R. 2018. Advaata. Math Geosci 50:273–298. https://doi.org/10.1007/s11004-017-9712-z.
9. Quinn TP, Erb I. 2020. Interpretable log contrasts for the classification of health biomarkers: a new approach to balance selection. mSystems 5: e00230-19. https://doi.org/10.1128/mSystems.00230-19.
10. Morton JT, Sanders J, Quinn RA, McDonald D, Gonzalez A, Vázquez-Baeza Y, Navas-Molina JA, Song SJ, Metcalf JL, Hyde ER, Lladser M, Dorrestein PC, Knight R. 2017. Balance trees reveal microbial niche differentiation. mSystems 2:e00162-16. https://doi.org/10.1128/mSystems.00162-16.
11. Rivera-Pinto J, Egozcue JJ, Pawlowsky-Glahn V, Paredes R, Noguera-Julian M, Calle ML. 2018. Balances: a new perspective for microbiome analysis. mSystems 3:e00053-18. https://doi.org/10.1128/mSystems.00053-18.
12. Zhou Y, Xu ZZ, He Y, Yang Y, Liu L, Lin Q, Nie Y, Li M, Zhi F, Liu S, Amir A, González A, Tripathi A, Chen M, Wu GD, Knight R, Zhou H, Chen Y. 2018. Gut microbiota offers universal biomarkers across ethnicity in inflammatory bowel disease diagnosis and infliximab response prediction. mSystems 3:e00188-17. https://doi.org/10.1128/mSystems.00188-17.
13. Tyakht AV, Manolov AI, Kanygina AV, Ischenko DS, Kovarsky BA, Popenko AS, Pavlenko AV, Elizarova AV, Rakitina DV, Baikova JP, Ladygina VG, Kostryukova ES, Karpova IY, Semashko TA, Larin AK, Grigoryeva TV, Sinyagina MN, Malanin SY, Shcherbakov PL, Kharitonova AY, Khalif IL, Shapina MV, Maev IV, Andreev DN, Belousova EA, Buzunova YM, Alexeev DG, Govorun VM. 2018. Genetic diversity of *Escherichia coli* in gut microbiota of patients with Crohn's disease discovered using metagenomic and genomic analyses. BMC Genomics 19:1–14. https://doi.org/10.1186/s12864-018-5306-5.

14. Iablokov SN, Klimenko NS, Efimova DA, Shashkova T, Novichkov PS, Rodionov DA, Tyakht AV. 2020. Metabolic phenotypes as potential biomarkers for linking gut microbiome with inflammatory bowel diseases. Front Mol Biosci 7:603740. https://doi.org/10.3389/fmolb.2020.603740.
15. Pryde SE, Duncan SH, Hold GL, Stewart CS, Flint HJ. 2002. The microbiology of butyrate formation in the human colon. FEMS Microbiol Lett 217: 133–3139. https://doi.org/10.1111/j.1574-6968.2002.tb11467.x.
16. Louis P, Hold GL, Flint HJ. 2014. The gut microbiota, bacterial metabolites, and colorectal cancer. Nat Rev Microbiol 12:661–672. https://doi.org/10.1038/nrmicro3344.
17. Walters WA, Xu Z, Knight R. 2014. Meta-analyses of human gut microbes associated with obesity and IBD. FEBS Lett 588:4223–4233. https://doi.org/10.1016/j.febslet.2014.09.039.
18. Quévrain E, Maubert MA, Michon C, Chain F, Marquant R, Tailhades J, Miquel S, Carlier L, Bermúdez-Humarán LG, Pigneur B, Lequin O, Kharrat P, Thomas G, Rainteau D, Aubry C, Breyner N, Afonso C, Lavielle S, Grill J-P, Chassaing G, Chatel JM, Trugnan G, Xavier R, Langella P, Sokol H, Seksik P. 2016. Identification of an anti-inflammatory protein from *Faecalibacterium prausnitzii*, a commensal bacterium deficient in Crohn's disease. Gut 65:415–425. https://doi.org/10.1136/gutjnl-2014-307649.
19. Imhann F, Vich Vila A, Bonder MJ, Fu J, Gevers D, Visschedijk MC, Spekhorst LM, Alberts R, Franke L, van Dullemen HM, Ter Steege RWF, Huttenhower C, Dijkstra G, Xavier RJ, Festen EAM, Wijmenga C, Zhernakova A, Weersma RK. 2018. Interplay of host genetics and gut microbiota underlying the onset and clinical presentation of inflammatory bowel disease. Gut 67:108–119. https://doi.org/10.1136/gutjnl-2016-312135.
20. Gurwara S, Ajami N, Jang A, Hessel F, Chen L, Plew S, Wang Z, Graham D, Hair C, White D, Kramer J, Kourkoumpetis T, Hoffman K, Cole R, Hou J, Husain N, Jarbrink-Sehgal M, Hernaez R, Kanwal F, Ketwaroo G, Shah R, Velez M, Natarajan Y, El-Serag H, Petrosino J, Jiao L. 2019. Dietary nutrients involved in one-carbon metabolism and colonic mucosa-associated gut microbiome in individuals with an endoscopically normal colon. Nutrients 11:613. https://doi.org/10.3390/nu11030613.
21. Qiu Z, Yang H, Rong L, Ding W, Chen J, Zhong L. 2017. Targeted metagenome based analyses show gut microbial diversity of inflammatory bowel disease patients. Indian J Microbiol 57:307–315. https://doi.org/10.1007/s12088-017-0652-6.
22. Liang JQ, Li T, Nakatsu G, Chen Y-X, Yau TO, Chu E, Wong S, Szeto CH, Ng SC, Chan FKL, Fang J-Y, Sung JJY, Yu J. 2020. A novel faecal *Lachnoclostridium* marker for the noninvasive diagnosis of colorectal adenoma and cancer. Gut 69:1248–1257. https://doi.org/10.1136/gutjnl-2019-318532.
23. Cai Y-Y, Huang F-Q, Lao X, Lu Y, Gao X, Alolga RN, Yin K, Zhou X, Wang Y, Liu B, Shang J, Qi L-W, Li J. 2022. Integrated metagenomics identifies a crucial role for trimethylamine-producing Lachnoclostridium in promoting atherosclerosis. NPJ Biofilms Microbiomes 8:11. https://doi.org/10.1038/s41522-022-00273-4.
24. Quinn TP, Erb I. 2020. Amalgams: data-driven amalgamation for the dimensionality reduction of compositional data. NAR Genom Bioinform 2:lqaa076. https://doi.org/10.1093/nargab/lqaa076.
25. Hankin RKS. 2006. Additive integer partitions in R. J Stat Soft 16. https://doi.org/10.18637/jss.v016.c01.
26. Klimenko N, Tyakht A, Popenko A, Vasiliev A, Altukhov I, Ischenko D, Shashkova T, Efimova D, Nikogosov D, Osipenko D, Musienko S, Selezneva K, Baranova A, Kurilshikov A, Toshchakov S, Korzhenkov A, Samarov N, Shevchenko M, Tepliuk A, Alexeev D. 2018. Microbiome responses to an uncontrolled short-term diet intervention in the frame of the citizen science project. Nutrients 10:576. https://doi.org/10.3390/nu10050576.
27. Noguera-Julian M, Rocafort M, Guillén Y, Rivera J, Casadellà M, Nowak P, Hildebrand F, Zeller G, Parera M, Bellido R, Rodríguez C, Carrillo J, Mothe B, Coll J, Bravo I, Estany C, Herrero C, Saz J, Sirera G, Torrela A, Navarro J, Crespo M, Brander C, Negredo E, Blanco J, Guarner F, Calle ML, Bork P, Sönnerborg A, Clotet B, Paredes R. 2016. Gut microbiota linked to sexual preference and HIV infection. EBioMedicine 5:135–146. https://doi.org/10.1016/j.ebiom.2016.01.032.
28. Pascal V, Pozuelo M, Borruel N, Casellas F, Campos D, Santiago A, Martinez X, Varela E, Sarrabayrouse G, Machiels K, Vermeire S, Sokol H, Guarner F, Manichanh C. 2017. A microbial signature for Crohn's disease. Gut 66:813–822. https://doi.org/10.1136/gutjnl-2016-313235.
29. Volokh O, Klimenko N, Berezhnaya Y, Tyakht A, Nesterova P, Popenko A, Alexeev D. 2019. Human gut microbiome response induced by fermented dairy product intake in healthy volunteers. Nutrients 11:547. https://doi.org/10.3390/nu11030547.