A simultaneous feature selection and compositional association test for detecting sparse associations in high-dimensional metagenomic data

Andrew Hinton (andrew84@email.unc.edu)
University of North Carolina at Chapel Hill School of Medicine

Peter J. Mucha
Dartmouth College

Research Article

Keywords: Microbiome Association Study, Sparse Association Signals, Pairwise Log Ratios, Compositional Data, Multivariate Analysis, Feature Selection, Network Analysis

DOI: https://doi.org/10.21203/rs.3.rs-703177/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License
A simultaneous feature selection and compositional association test for detecting sparse associations in high-dimensional metagenomic data

Andrew L. Hinton\textsuperscript{1,2,*} and Peter J. Mucha\textsuperscript{1,3,4}

* Correspondence: andrew84@email.unc.edu

\textsuperscript{1} Curriculum in Bioinformatics and Computational Biology, University of North Carolina, Chapel Hill, NC, USA

\textsuperscript{2} University of North Carolina at Chapel Hill Food Allergy Initiative, School of Medicine, Chapel Hill, NC, USA

\textsuperscript{3} Departments of Mathematics and Applied Physical Sciences, University of North Carolina, Chapel Hill, NC, USA

\textsuperscript{4} Department of Mathematics, Dartmouth College, Hanover, NH, USA
Abstract (350 words)

Background: Numerous metagenomic studies aim to discover associations between the microbial composition of an environment (e.g. Gut, Skin, Oral) and a phenotype of interest. Multivariate analysis (MVA) is often performed in these studies without critical a priori knowledge of which taxa are associated with the phenotype being studied. Consequently, non-parametric MVA methods are applied directly to all taxa surveyed independent of noise. This approach typically reduces statistical power in settings where true associations among only a few taxa are obscured by high dimensionality (i.e. sparse association signals). At the same time, the inclusion of all taxa can confound the extraction of key biological insights. Further, low sample size and compositional sample space constraints exist in these data whereby beyond-study generalizability may be reduced if not properly accounted for. More powerful association tests that are interpretable and directly account for compositional constraints while detecting sparse association signals are needed.

Methods: We developed Selection-Energy-Permutation (SelEnergyPerm), a non-parametric group association test with embedded feature selection. SelEnergyPerm directly accounts for compositional constraints by selecting parsimonious log ratio signatures from the set of all pairwise log ratios (PLR) between features (OTUs, taxa, etc.). To do this, network methods are used to rank, select, and maximize the between-group association of a candidate log ratio subset. This process is then repeated with an appropriate permutation testing design to simultaneously determine the significance of the selected signatures and association.

Results: Simulation results show SelEnergyPerm selects small independent sets of log ratios that capture strong associations in a range of scenarios with small and large dimensional feature spaces. Additionally, our simulation results demonstrate SelEnergyPerm consistently detects/rejects associations in synthetic
data with sparse, dense, or no association signals. We demonstrate the novel benefits of our method in four case studies utilizing publicly available 16S rRNA and whole-genome sequencing datasets.

**Conclusions:** Tools to analyze complex high-dimensional metagenomic datasets with sparse association signals using robust PLR have not been sufficiently developed previously. We propose SelEnergyPerm, a novel framework for the discovery of phenotype-associated, metagenomic log ratio signatures for characterizing and understanding alterations in microbial community structure. SelEnergyPerm is implemented in R, available at https://github.com/andrew84830813/selEnergyPermR.

**Keywords:** Microbiome Association Study, Sparse Association Signals, Pairwise Log Ratios, Compositional Data, Multivariate Analysis, Feature Selection, Network Analysis

**Background**

Metagenomic studies have enabled unprecedented insight into connections between microbes, their functions, and human disease [1]. These insights are a direct result of rapid advances in next-generation sequencing technologies which are critical to metagenomic studies. Specifically, these technologies are leveraged in two popular approaches: 16S ribosomal rRNA amplicon (16S) and whole-genome shotgun (WGS) sequencing [2]. Application of these approaches are widespread and have been used to study associations between the gut microbiome composition and Colorectal Cancer, IBD, Obesity, Cirrhosis, and anxiety/depression in humans via the gut-brain axis to name a few [3-7]. The skin, oral, and nasal microbiomes among other sites have also been studied in connection to disease onset and progression [8-10]. With an increasing number of putative associations between microbial communities from various sites of the human body and disease being reported, microbial compositions are now being explored as diagnostic and screening tools [11, 12]. While exciting, appropriate statistical methods are still needed to overcome methodological challenges in these exceptional data, so that robust microbial biomarkers
and true associations can be discovered among noisy high-dimensional metagenomic data where sample sizes in observational studies are smaller than the number of features discovered.

Before metagenomic data can be used to test for associations, raw sequencing data must be appropriately processed. Taxonomic count tables are created by processing raw 16S or WGS sequencing data through bioinformatics pipelines such as QIIME [13] or mothur [14] for amplicon sequencing data and MetaPhlAn2 [15] or Kraken [16] for WGS data. Sequencing reads are assigned a taxonomy where the resulting count tables are then used to profile and analyze the association between groups under study at various taxonomic levels (Phylum - Species). These data are often sparse and summarize the total number of reads for each taxonomic assignment within each sample. In current practice, total counts in these settings have been widely recognized as being uninformative due to limitations within sequencing technology [17-19]. These data are considered to carry only relative information whereby special statistical techniques and considerations are required. That is, these relative data have a unit-sum simplex sample space where traditional Euclidean-based statistical methods have limited applicability due to geometrical differences between sample spaces. Ignoring these constraints has been shown to increase type I error [19], increase the chance of reporting spurious [20] associations, and thus may limit the ability of biomarkers (discovered from relative abundance) to generalize beyond studies.

A direct way to address simplex sample space constraints imposed by relative data is through a log-ratio transformation. This transformation emerged from the statistical analysis of compositional data [21] and functions by mapping relative data from the unit-sum simplex to traditional Euclidean space. Importantly, the log-ratio transformation is sub-compositionally coherent [21, 22], independent of the number of dimensions (taxa, OTU’s, etc.) observed in a cohort whereby true associations in the log-ratio form are preserved. This is not true for relative abundance where proportions change as new dimensions (OTU, Taxa, etc.) are considered, discovered, or removed. Further, sub-compositional coherence is of practical importance in biomedical studies where biomarker discovery, disease
prediction, and beyond-study generalization are paramount. While the log-ratio transformation is well-known and routinely applied in some fields [23], its use in metagenomic datasets has been limited despite its advantages. Indeed, significant barriers exist when applying a log-ratio transformation to metagenomic data. Core challenges include properly handling zeroes [24, 25], selection and interpretation of various log-ratio forms [21, 26-27], and scale differences in counts [28].

While the importance of the compositional nature of metagenomic data has recently been recognized [17, 29], few multivariate statistical methods have been developed directly for these data. Currently, several univariate methods (ALDEx2, ANCOM, ANCOM-BC) for detecting differential abundance in compositional metagenomic data have been proposed [30-32]. While powerful, these methods are unable to detect multivariate structure within complex interconnected microbial communities [33, 34]. Complicated microbial patterns, which occur in settings where there are significantly more variables than samples, can however be discovered with appropriate network and multivariate statistical methods. Further, multivariate statistical methods, which are appropriate when there exist relationships between a set of variables (i.e. microbial composition) and two or more groups are to be analyzed, have better control over type I error [34].

Currently, several multivariate statistical methods to detect between-group distributional differences or associations in metagenomic data can be used. A subset of these methods are universal and require a suitable beta diversity or between-sample distance (Euclidean, Manhattan, Mahalanobis, etc.) or dissimilarity (Bray-Curtis, weighted/unweighted UNIFRAC, Jaccard, etc.) metric be specified before analysis. PERMANOVA [35], ANOSIM [36], and/or the Energy distance [37], which are nonparametric tests, can then be applied to test distributional differences between groups. Between-group association signals in metagenomic data may be sparse, i.e., result from differences among only a few features (OTU, taxa, etc.) or dense where the association signal is formed by differences between many features.
Importantly, the above-mentioned nonparametric tests lack embedded feature selection and thus can have limited statistical power for detecting sparse signals in high-dimensional data. Feature selection, which is essential to detecting sparse association signals in sparse high-dimensional metagenomic data [38, 39], requires sophisticated methods and care to simultaneously select features and test associations while maintaining reasonable type I error control [40, 41]. To address this need, the adaptive microbiome-based sum of powered score (aMiSPU) method was proposed to test between-group associations with embedded feature selection through the variable ranking of untransformed relative abundances [38]. However, aMiSPU does not inherently account for the compositional structure of the data and requires OTU’s and phylogenetic tree to test associations which may not always be available (e.g. WGS data). Further, aMiSPU lacks automatic feature selection across all tests and was shown to be underpowered when applied to highly sparse data [39]. Similarly, OTU’s and phylogenetic information are also required as input for microbiome higher criticism analysis (MiHC) [39], which like aMiSPU tests sparse associations in 16S data using untransformed relative abundances and lacks automatics feature selection. Moreover, MiHC results report an arbitrary top number of OUT’s to be important thus making it difficult to understand which features specifically contribute to the overall detected association.

Inspired from concepts put forth in the Direction-Projection-Permutation method for assessing statistical significances in high-dimensional settings [43], we introduce here the Selection-Energy-Permutation (SelEnergyPerm) method for testing and understanding associations in sparse 16S and WGS data. SelEnergyPerm is the first method to our knowledge to utilize robust pairwise log ratios to detect and understand parsimonious log ratio signatures from all types of metagenomic data through simultaneous feature selection and association testing. We first show that our novel approach selects smaller subsets of non-redundant log ratios that better maximize between-group associations when compared to popular feature selection methods. Next, we show through an extensive simulation study
(using synthetic and empirical 16S/WGS data distributions) that SelEnergyPerm has, on average, better combined power and false discovery control via the Matthews Correlation Coefficient (MCC) when compared to existing beta-diversity-based approaches. Finally, to demonstrate the utility of SelEnergyPerm in detecting and understanding differences between metagenomic distributions, we apply our method in four case studies utilizing publicly available metagenomic datasets where we test associations between: (1) Ugandan infant’s cerebrospinal fluid microbiomes and post-infectious hydrocephalus, (2) delivery mode and the composition of infant’s gut microbiomes over the first three months of life, (3) adult gut microbiomes and abnormal fecal calprotectin levels, and (4) the gut microbiome composition of infants within the first 6 months of life and future food allergy to egg, milk, or peanuts.

**Results**

**The Selection-Energy-Permutation framework for simultaneous feature selection and group association testing.**

To robustly uncover sparse microbial signatures while simultaneously testing multivariate group associations, we based our SelEnergyPerm framework on a novel network-based feature selection approach combined with permutation testing for sparse high-dimensional low-sample-size compositional metagenomic data. Our framework (Figure 1A), which selects from all pairwise log ratios between variables (Taxa, OTU’s, etc.), first scores the between-group variation of individual log ratios using our Differential Compositional Variation (DCV) scoring measure (see Methods). From this, a weighted DCV log-ratio network is formed and subsequently pruned to reduce redundancy and complexity via a maximum spanning tree. Final subsets are then selected by maximizing the between-group association using a greedy forward stepwise selection procedure. Multivariate test statistics, which measure the strength of the association between groups, are then computed on the final retained
subset. Statistical significance is determined by repeating this process with permuted group labels to obtain the permutation distribution of the test statistic of interest under feature selection. In this way, we determine if the observed association is larger than what would be expected by chance (Figure 1B). To this end, our framework tests the overall null hypothesis of no association between the metagenomic composition and group labels.

**Feature selection comparison to other methods.**

We first benchmarked the multivariate characteristics of subsets selected by our feature selection (FS) approach against a set of popular methods for FS: Boruta [43], Least Absolute Shrinkage and Selection Operator (LASSO) [44], Information Gain Filtering [45], and Random Forest Recursive Feature Elimination (RFE) [46]. The benchmarks were carried out by varying the number of log-ratio dimensions in the full feature set using five simulation scenarios with both balanced and unbalanced sampling designs (see Methods). Specifically, for subsets returned by each method, we studied the number of log ratios selected (as a proxy for model complexity), the clustering coefficient of the log-ratio network (measuring log-ratio redundancy), and the combined F-statistic (strength of association, see Methods), and the computational time required to return the final subset (Additional file 1: Figure S1). In Figure 2, we present results from scenarios with a balanced sampling design. Notably, the results for the unbalanced sampling design scenarios are similar and do not change the overall comparative interpretation (Additional file 2: Figure S2). Examination of the clustering coefficient across all simulation scenarios/dimensions demonstrates that SelEnergyPerm consistently selects linearly independent subsets of log ratios (Figure 2 and Additional file 2: Figure S2, clustering coefficient = 0), in contrast with the subsets observed in other methods tested. Of note, a clustering coefficient > 0 indicates a selected log-ratio subset contains at least one triple of linearly-dependent log ratios (closing a triangle in the log-ratio network), thereby unnecessarily increasing dimensionality and model complexity. (We note that any cycle present in a log-ratio network indicates linear dependence, though we did not test for cycles...
larger than closed triangles, and in particular we emphasize that by construction the SelEnergyPerm-selected subsets do not include any such cycles.) Additionally, the number of log ratios retained by each method across every scenario tested revealed subsets selected by SelEnergyPerm were, on average, 14 to 149 times smaller than other methods (Figure 2 and Additional file 2: Figure S2). Next, the strength of the association, measured by the combined-F statistic (see Methods), indicates SelEnergyPerm-selected subsets typically capture higher between-group variations than other methods tested. In particular, in Scenarios 2–4, SelEnergyPerm subsets were observed to have, on average, higher combined-F values than all other methods across all dimensions tested (Figure 2 and Additional file 2: Figure S2). Meanwhile, in Scenarios 1 and 5, SelEnergyPerm subsets generally performed similarly to the other methods but better as the dimensionality increased. Notably, Scenarios 1 and 5 do not simulate sparse association signals and have strong between-group dispersion effects present. These results indicate SelEnergyPerm returned subsets better capturing sparse associations (Scenarios 2–4) than the other FS methods tested. Computational time experiments show, across all scenarios tested, SelEnergyPerm is on average faster than Boruta and RFE but slower than LASSO and Information Gain filtering (Additional file 1: Figure S1). Overall, SelEnergyPerm subsets were non-redundant, significantly more parsimonious, and captured stronger associations than other methods tested, thereby enabling robust biological interpretation using log ratios in high-dimensional feature spaces.

Detection of sparse associations in synthetic data.

Here, we use data simulated from theoretical distributions to compare the ability of SelEnergyPerm, PERMANOVA, ANOSIM, and the energy test to detect associations in sparse high-dimensional data. That is, we are interested in determining how well each method accepts or rejects the null hypotheses (no difference between groups) when presented with two groups of data that, as ground truth, come from the same (Null Case; Type I error assessment) or different (True Case; power assessment) distributions.
From this, we report the performance of each method in terms of the Matthews correlation coefficient at $\alpha = 0.05$ (see Methods) for 4 simulation scenarios (see Methods) with both balanced and unbalanced sampling designs (Figure 3). For brevity, we shall refer to the collection of PERMANOVA, ANOSIM, and energy tests as the standard methods.

In Scenario 1, where data are simulated from a Dirichlet distribution with between-group location and dispersion effects that grow as the number of dimensions increase (see Methods), we see for the balanced design that both SelEnergyPerm and the energy test perform well over all dimensions (number of log ratios) tested. Notably, ANOSIM loses the ability to detect associations as the number of dimensions increases while PERMANOVA performs poorly over all dimensions. The poor performance of ANOSIM and PERMANOVA is directly attributable to the underlying heterogeneity of variance present in the data generated in this scenario. These limitations of PERMANOVA and ANOSIM have also been discussed [47]. The presence of dispersion effects is confirmed with the PERMDISP2 [48] method and can be observed to be steady (Figure 3 - Scenario 1) and increasing across dimensions. For the unbalanced design, SelEnergyPerm and the energy test both retain strong performance and have comparable performance over most dimensions, whereas ANOSIM completely loses the ability to detect associations under the unbalanced design and PERMANOVA continues to perform poorly across all dimensions.

For Scenario 2 (Figure 3), the data distributions for each group are simulated from two Dirichlet distributions that differ in the location of the first component and overall variance. That is, this scenario embeds a sparse signal (location shift) in the first dimension with random noise in the remaining dimensions. The results for this scenario show that for the balanced case SelEnergyPerm performs significantly better than all other methods tested. For the unbalanced case, SelEnergyPerm performs better than all other methods for smaller numbers of dimensions, however, it performs similarly to
ANOSIM as the number of dimensions increases. Notably, the performance of ANOSIM improves as the number of dimensions increases for both the balanced and unbalanced cases.

For Scenario 3 (Figure 3), the data distributions for the first class are simulated from the additive logistic normal distribution. Data for the second class are also generated from an additive logistic normal distribution with the same parameters (same covariance matrix) but with location shifts in the first 25% of the dimension. Under this scenario, we observed the performance of SelEnergyPerm to be comparable to the standard methods for the balanced case and slightly worse than the standard methods for the unbalanced case. The reduced performance in the unbalanced case is attributable to the dense signal (25% of feature) being in direct tension with the reduced feature selection subset objective of SelEnergyPerm.

Lastly, in Scenario 4 (Figure 3), a location shift only (same between class covariance structure) was embedded in the first component of two additive logistic normal distributions that increases with the number dimensions. Here, SelEnergyPerm outperformed the standard methods as the number of dimensions increased for both the balanced and unbalanced cases. While performing better overall relative to the standard methods, a notable decrease in performance was observed from the balanced to the unbalanced case for SelEnergyPerm. This decrease in performance was exacerbated among the standard methods where performance not only decreased between sampling designs but also generally declined as the number of dimensions increased in the unbalanced design.

Overall sparse association detection performance as measured by MCC, sensitivity, specificity, positive predictive value, negative predictive value, Youden index, and false-positive rate across all scenarios at an $\alpha = 0.05$ are shown in Additional File 3: Figure S3. These aggregate results demonstrate SelEnergyPerm in general outperforms the standard methods for detecting sparse associations under the synthetic data simulation scenarios presented here.
Detection of sparse associations in data simulated from real 16S and WGS datasets.

To further assess performance, we benchmarked our method against the standard methods on data simulated from properties observed in real metagenomic datasets. In this way, unique metagenomic data characteristics such as sparsity, over dispersion, and complex co-occurrence patterns are assessed synthetically. As described in the detection of sparse associations in the synthetic data section above, MCC is used to assess the ability of each method to detect associations across these settings.

In the first setting, (Figure 4 – 16S/WGS: Increasing Covariance Diff.), an increasing covariance effect with a decreasing location effect between classes was simulated using healthy subsets of 16S and WGS samples. The increasing dispersion effect is confirmed with PERMDISP2 for both sampling designs (Figure 4). For 16S and WGS data with a balanced sampling design, SelEnergyPerm outperforms the standard methods across all effect sizes and has strong performance as the number of dimensions increase. For 16S data with an unbalanced design, all methods performed poorly as the location shift effect increases. This trend is traceable to the strong embedded covariance effect between classes which is a known confounder in high-dimensional association settings [47]. Notably, only SelEnergyPerm and ANOSIM maintain positive MCCs on average, indicating these methods better control type I error (with severely limited power) under this sampling design. For WGS data with an unbalanced design, SelEnergyPerm outperformed the standard methods and had better association detection across all effect levels.

For the second simulation setting, (Figure 4 – 16S/WGS: Increasing Location Effects), we simulated large location shifts between classes by increasing the size of the association signal from 5% to 50% of all features with fixed covariance structures. These shifts were computed using synthetic subsets of WGS and 16S samples from publicly available healthy gut microbiomes. Indeed, PERMDISP2 analysis confirmed the absence of covariance effects. For both 16S and WGS data with a balanced sampling
design, SelEnergyPerm outperformed all standard methods. As expected, in both WGS and 16S data, the
performance of the standard methods increased as the association signal became less sparse. Again, for
the unbalanced design in both WGS and 16S data, SelEnergyPerm outperformed all standard methods.
Importantly, the detection ability of the standard methods improved as the association signal became
less sparse.

Finally, overall sparse association detection performance metrics are shown in Additional File 4: Figure
S4. These aggregate results demonstrate SelEnergyPerm has better overall sparse association detection
performance when compared to standard methods using data simulated from real 16S and WGS
datasets.

Microbial association between Ugandan infant’s cerebrospinal fluid microbiomes and post-infectious
hydrocephalus.

In this case-control study by Paulson et al., the cerebral spinal fluid (CSF) of Ugandan infants was
profiled using 16S rRNA amplicon sequencing to characterize microbial agents associated with Post
Infectious Hydrocephalus (PIH) following neonatal sepsis [49]. This processed gut microbiome dataset
was retrieved from microbiomeDB [50, 51] and consisted of 369 distinct taxa measured on 92 samples
(58 PIH and 34 Non-Post Infectious Hydrocephalus (NPIH) patients). Removing taxa not present in at
least 10% of samples yielded 57 total distinct taxa (i.e., 1,596 taxa log ratios). We then apply
SelEnergyPerm to determine if there was an association between the microbiome composition in the
CSF and PIH/NPIH disease status. Further, we utilize the reduced SelEnergyPerm log-ratio signature of
PIH in CSF to gain insight into specific microbiome compositional differences.

From this, we confirm with SelEnergyPerm a significant association (cF = 33.59817, empirical p = 0.007)
exists between the composition of microbes in the CSF and PIH/NPIH [49] (Figure 5A). Our method
identified a reduced log-ratio signature of 12 ratios between 13 taxa as being significantly associated
with PIH/NPIH (Figure 5B). Random forest (RF) models were used to understand the capability of the SelEnergyPerm signature for discriminating between disease statuses. Indeed, estimating the discriminatory ability with RF using 50 repeats of 10-fold cross-validation, we computed an Area Under the Receiver Operating Characteristic Curve (AUC) = 0.906 (0.879-0.935 95% CI) (Figure 5C). We emphasize, however, that the more complex RF models with all 1,596 pairwise log ratios yielded a comparable AUC = 0.892 (0.860-0.923 95% CI) (Figure 5C). In addition, microbiome analysis carried out in Paulson et al. revealed *Paenibacillus* alone to be important for predicting PIH. Here, using only the relative abundance of *Paenibacillus* with RF, we observed an AUC = 0.830 (0.792-0.867 95% CI) which was significantly lower than the SelEnergyPerm results. Combined, these results suggest the parsimonious SelEnergyPerm-derived log-ratio signature retains important disease interactions and better discriminates between PIH vs. NPIH when compared to *Paenibacillus* alone.

To understand how the log ratios in our signature work together to explain differences between the CSF microbiome of PIH vs. NPIH patients we apply principal component analysis (PCA) (Figure 5D) and analyzed the means of the log ratios. From that, examination of the distribution of samples shows the greatest separation between disease groups occurs along PC1 (Figure 5D) which explains 78.48% of the total variation. This separation indicates positive scores along PC1 are associated with non-PIH whereas negative scores are associated with PIH samples. Further, analyses of the log ratio mean between groups for each log ratio in the SelEnergyPerm signature indicate the abundance of *Paenibacillus* is significantly increased (Figure 5E) relative to taxa it’s connected to (Figure 5B). Moreover, RF variable importance indicates the log ratio between *Paenibacillus* relative to *Pseudomonas* to be most important for distinguishing between disease statuses. Indeed, analysis of PC1 loadings (Figure 5E) reveals increased abundance of *Pseudomonas* relative to *Paenibacillus* results in positive loadings (non-PIH associated) along PC1. Overall our results confirm, using pairwise log ratios derived from SelEnergyPerm, the importance of *Paenibacillus* in PIH. Additionally, we show the interaction between the abundance of
Pseudomonas relative to Paenibacillus is particularly important whereby more Pseudomonas is characteristic of non-PIH and more Paenibacillus is associated with PIH.

Association between delivery mode and the composition of infant’s gut microbiomes over the first three months of life.

Bokulich et al. monthly profiled the gut microbiome of infants with either a vaginal or cesarean delivery mode using 16S rRNA sequencing for the first two years of life [52]. The processed dataset was retrieved from the QIITA repository using study ID 10249 [53]. Specifically, we extracted samples during the first 3 months of life which totaled 230 samples from 63 infants (Cesarean = 25, Vaginal = 38). We aggregated OTU’s to the family-genus level which resulted in 140 distinct taxa (9,730 log ratios) present in at least 10% of all samples by month. Here we apply SelEnergyPerm to determine if the gut microbiomes are different between the delivery modes of infants at any of the first 4 monthly time points collected (0-3 months). Secondarily, we studied our reduced log ratio signatures to understand gut microbiome compositional differences between delivery modes at time points where significant differences were detected.

Applying SelEnergyPerm to each time point with restricted permutation testing to account for repeated host microbiomes within a collection month and correcting for multiple comparisons using the Benjamini-Hochberg (BH) procedure, we found significant differences in the composition of the gut microbiomes between delivery modes during the 0-2 month collection periods (Figure 6A). Notably, restricted permutation testing with PERMANOVA and ANOSIM using all taxa PLR failed to detect differences between the gut microbiomes at $\alpha = 0.05$. This is also clear when comparing the discriminatory ability of each set. Using Partial Least Squares Discriminate Analysis (PLS-DA) with repeated cross-validation stratified by both delivery mode and host, we observed the AUC of the SelEnergyPerm-derived signatures to be higher across all time points when compared to models trained
using all PLR (Figure 6B). We next used the reduced log-ratio signatures and their PLS-DA variable importance scores to better understand which taxa are most important for discriminating between delivery modes. Indeed, aggregating to the family level for ease of interpretation, we found during months 0 and 1 that *Bacteroidaceae* were top contributors to compositional differences (Figure 6C). This pattern changed during month 2 where *Rikenellaceae taxa* were most important for discriminating between delivery modes (Figure 6C). Finally, to understand the direction of these differences (i.e. is the numerator more abundant than denominator or vice-versa between groups for a given log ratio), we examine the log-ratio means networks using the SelEnergyPerm signature. Here, we analyze the directed log ratio means network relatively (i.e. taxa A more/less abundant than taxa B) between delivery modes (Figure 6D). Specifically, given the hub-spoke character, with a single highly connected and central node in the directed MST formed by the SelEnergyPerm signature, we can see month 0 is dominated by differences between log ratios that include *Lachnospira* and *Bacteroides*. In particular, *Lachnospira* and *Bacteroides* are more abundant relative to their network of taxa connections for infants with a vaginal delivery mode whereas the opposite is true for infants with a cesarean delivery mode. For month 1, *Bacteroides* are observed to be more abundant relative to its network of taxa connections for infants with a vaginal delivery mode. The opposite is true for infants with a Cesarean delivery mode where *Bacteroides* are less abundant within its network of taxa connections. Finally, for month 2, *Rikenellaceae* taxa can be observed to be more abundant than both *Clostradiacea* and *Proteus* taxa where the opposite is true for infants with a Cesarean delivery mode.

**Association between abnormal fecal calprotectin levels and the composition of the gut microbiome in healthy and IBD individuals.**

Here we apply SelEnergyPerm to analyze WGS microbiome data from the integrative human microbiome project (iHMP) [54], a longitudinal study designed to uncover interactions between disease and human-associated microbial communities. Specifically, using the inflammatory bowel disease (IBD)
part of the iHMP study, we tested whether there exists an association between the gut microbiome composition and abnormal levels of fecal calprotectin, a protein marker of intestinal inflammation [54]. Processed microbiome data were extracted from the Inflammatory Bowel Disease Multiomics Database [55] resulting in 399 samples (93 individuals) reporting fecal calprotectin levels that were above 120 (abnormal; n = 190) or below 50 (normal; n = 209). There were 122 species identified (7,381 log ratios) as being present in at least 10% of all samples.

Using restricted permutation testing, accounting for the order of visit and diagnosis of Ulcerative Colitis (UC), Crohn’s Disease (CD), or non-IBD, SelEnergyPerm identified a significant association (cF = 92.507, p = 0.000999, 1000 permutations) between the composition of the gut microbiome and abnormal levels of fecal calprotectin in corresponding stool samples (Figure 7A). Notably, both ANOSIM and PERMANOVA with restricted permutation designs using all log ratios detected this association. To assess whether the associated SelEnergyPerm log-ratio signature (31 log ratios between 26 species) retained enough information to adequately discriminate between levels of fecal calprotectin we estimated the discriminatory ability both using the reduced signature and using all log ratios. Using repeated cross-validation with PLS-DA we found the SelEnergyPerm signature (AUC = 0.829 (0.803 – 0.854 95%CI)) to have comparable performance to PLS-DA models trained using all log ratios (AUC = 0.833 (0.803 – 0.862 95%CI)) (Figure 7B). Examination of the latent space projection of a final PLS-DA model fit using the SelEnergyPerm signature reveals strong separation between individuals with normal vs. abnormal fecal calprotectin levels (Figure 7C). A directed log-ratio network of the SelEnergyPerm signature weighted by PLS-DA variable importance shows log ratios involving *Dialister invisus*, *Streptococcus salivarius*, *Bacteroides fragilis*, *Escherichia coli*, and *Blautia wexlerae* to be most important for discriminating between levels of fecal calprotectin (Figure 7D). Interestingly, stratifying the log-ratio signature by diagnosis reveals both shared (significant between diagnosis differences across all groups) and distinct (significant between diagnosis differences among a single group) gut microbiome
differences (Figure 7E). Particularly increased abundance of *Dialister invisus* relative to *Bacteroides ovatus*, *Intestinimonas butyriciproducens*, and *Anaerotignum lactatifermentans* was observed to be associated with abnormal fecal calprotectin independent of diagnosis.

**Association between the gut microbiomes of infants in early life and the development of allergen-specific sensitization**

In this case study, we apply SelEnergyPerm to WGS gut microbiome data from the DIABIMMUNE study [56]. The focus of this longitudinal study was to characterize interactions between the immune system and the gut microbiome in the context of autoimmunity and allergy. Specifically, the gut microbiomes of infants from Finland, Russia, and Estonia were profiled monthly during the first 3 years of life. Processed WGS data were directly accessed from [57]. Here we apply SelEnergyPerm to test if associations exist between the composition of the gut microbiome at 6-month intervals during the first 2 years of life and allergy status. Allergy status was defined as food allergy (FA) if the host reported an allergy to egg, peanuts, and/or milk at year 2 or non-FA otherwise [56]. We extracted 646 samples from 192 infants (Russia = 53, Finland = 70, Estonia = 59) across 170 unique species (14,365 log ratios).

Using restricted permutation testing to account for repeated host microbiomes and host country we applied SelEnergyPerm to each timeframe and corrected for multiple comparisons using the BH procedure. We found significant differences in the composition of the gut microbiomes between allergy status during both the first 6 months and the 6-12 month collection periods (Figure 8A). PERMANOVA and ANOSIM using all taxa PLR detected differences between the gut microbiome during the first 6 months of life but did not detect differences between the gut microbiomes during the remaining timeframes at an $\alpha = 0.05$ after correcting for multiple comparisons with the BH procedure. This is apparent when comparing the discriminatory ability between the SelEnergyPerm signature and all log ratios. Using Partial Least Squares Discriminate Analysis (PLS-DA) with repeated cross-validation
stratified by allergy status, host, and month, we observed the AUC of the SelEnergyPerm-derived signatures to be significantly higher across all time points when compared to models trained with all log ratios (Figure 8B). Using the SelEnergyPerm log ratio signatures and the corresponding PLS-DA variable important scores we next examine which taxa are important for discriminating between food allergy statuses later in life. Stratifying by month and selecting the top 5 strength (weighted degree) species from our variable importance log-ratio network, we found during the first 6 months of life *Clostridium ramosum*, *Streptococcus parasanguinis*, and *Bifidobacterium bifidum* to be major contributors of DCV between allergy status (Figure 8C). However, for the 6 – 12 month time period we found the abundance of *Clostridium hathewayi*, *Bacteroides dorei*, and *Haemophilus haemolyticus* to be major contributors of DCV (Figure 8C). A review of the log-ratio mean networks (Figure 8D) between allergy status during the first 6 months shows *Clostridium ramosum* is, in general, more abundant relative to species (node strength indicated by size) it is connected to in infants with FA vs. non-FA. Further, during the 6 – 12 month time period we see more distinct differences in the log-ratio mean networks whereby *Bacteroides dorei* can be observed to be more abundant relative to species it is connected to in FA infants. We also observe *Clostridium hathewayi* to be more abundant than the species it is connected to in infants with FA where the opposite is true in infants without FA.

**Methods and materials**

Selection-Energy-Permutation (SelEnergyPerm) for simultaneous feature selection and group association testing in sparse high-dimensional compositional data.

In this section, we explain the SelEnergyPerm framework in greater detail. First, we describe our all pairwise log ratios DCV scoring measure and then detail the construction of the weighted DCV networks. We next discuss the removal of redundant ratios using a maximum spanning tree that simultaneously maximizes log-ratio variance. After this, we introduce our network-based approach to feature selection
and the two multivariate test statistics utilized to measure the strength of the association. We then
detail our between-group association maximization algorithm with pseudocode. Finally, we describe the
approach for assessing statistical significance via permutation testing using Monte Carlo sampling.

**Differential Compositional Variation Scoring**

For a given metagenomic study, let $\mathbf{M} \in \mathbb{R}^{n \times d}$ be the taxa count table for $n$ samples and $d$ taxa. Here
we are interested in computing all $p = \frac{d(d-1)}{2}$ pairwise log ratios (PLR) of $\mathbf{M}$. However, we must first
address the problem of zero counts. While there are numerous strategies with various drawbacks to
model and impute zeros [24, 58-59] based on type/cause, there is in general no consensus of which
strategy should be used in metagenomic data. Notwithstanding, here we treat zero taxa counts as being
below the detection level, and we adopt a corresponding multiplicative replacement strategy for
imputing zeros proposed in [24] that preserves the essential log ratio and covariance structure. We
apply the closure operator to $\mathbf{M}$ to map the count data onto the unit-sum simplex where we define the
matrix $\mathbf{X}$ with elements $x_{ij}$ as

$$x_{ij} = (C[\mathbf{M}])_{ij} = \frac{m_{ij}}{\sum_{k=1}^{d} m_{ik}},$$

(1)

We then set $\delta$ to be a constant equal to the smallest nonzero value across all of $\mathbf{X}$ and impute zeros to
obtain $\mathbf{R}$ with elements

$$r_{ij} = \begin{cases} 
\delta, & x_{ij} = 0 \\
 x_{ij}(1 - \delta \sum_{k=1}^{d} \mathbf{1}_{x_{ik} = 0}), & x_{ij} > 0 \quad \text{for } (i = 1, \ldots, n).
\end{cases}$$

(2)

We then compute all PLR from $\mathbf{R}$ to obtain $\mathbf{Z} \in \mathbb{R}^{n \times p}$ where $p = \frac{d(d-1)}{2}$ indicates the number of PLRs
(up to a sign). Because feature selection is critical to maximizing power and identifying sparse signals
hidden within noisy high-dimensional data, we seek to reduce the dimensionality through feature
selection. Notably, this setting is distinct from traditional log ratio analysis [21] where dimensionality
reduction using PCA is applied to PLR/CLR transformed features to reduce dimensionality. Importantly,
the set of \( p = \frac{d(d-1)}{2} \) different PLR are not independent of one another and require careful treatment
to select ratios that are independent of each other. Here we propose Differential Compositional
Variation (DCV), a scoring measure that enables efficient screening and ranking of PLR features within
compositional data. Similar to the screening concept in [60] for ultrahigh dimensional feature spaces,
DCV is motivated by Aitchison’s compositional variation array [21] where patterns of compositional
variability for a group of data can be expressed in terms of the log ratio means \( \xi_j = E[Z_{*,j}] \) and
variances \( \tau_j = \text{var}(Z_{*,j}) \) where \( j = 1, \ldots, p \). Similarly, here we define the logratio medians to be \( \tilde{\xi}_j = \tilde{Z}_{*,j} \). Our DCV score utilizes a diverse array of 5 statistics to score the contained variation of each log
ratio. Each component of DCV provides unique insight, thereby enabling efficient screening of
uninformative log ratios for downstream multivariate analysis. Let \( y \) contain the labels for the binary
classes/groups \( c_1 \) and \( c_2 \) under consideration. In terms of \( \xi_j \) and \( \tau_j \), the first component of DCV, which
measures differences in group means, is Welch’s t-statistic defined here as

\[
\Delta^1_j = \frac{\xi_{c_1} - \xi_{c_2}}{\sqrt{\left(\frac{\tau_{c_1}}{N_x}\right) + \left(\frac{\tau_{c_2}}{N_y}\right)}}.
\]

Next, we decompose the compositional variability of each \( Z_{*,j} \) using the classical F-statistic to again
measure means of differences:

\[
\Delta^2_j = \frac{n_{c_1} \left(\xi_{c_1} - \xi_j\right)^2 + n_{c_2} \left(\xi_{c_2} - \xi_j\right)^2}{\tau_{c_1}^2 + \tau_{c_2}^2}.
\]
The third component of DCV, which measures heterogeneity of variances, is the Brown-Forsythe F-statistic. Here we define the jth log ratio to be the vector \( z = \mathbf{Z}_j \). Let \( x_{cj} \in \mathbf{z} \) where c and j denote the jth observation of the cth group. From this, let \( b_{cj} = |x_{cj} - \bar{\xi}_j| \)

\[
\Delta_j^3 = \frac{\sum_c n_c (b_{cj} - \bar{b}_j)^2}{\sum_c \sum_j (b_{cj} - b_{cj})^2 / \sum_c (n_c - 1)},
\]

where \( b_c = \sum b_{cj} / n_c \) and \( \bar{b}_j = \sum b_{cj} / \sum n_c \).

For the fourth component, we first define the empirical distribution function for each ordered log ratio \( Z_j^* \), to be

\[
F_j^c(x) = \frac{1}{n_c} \sum_{k=1}^{|c|} I_{(-\infty,x]} Z_{k,j}^c
\]

for each \( (j = 1, \ldots, p) \) where \( I_{(-\infty,x]} = \begin{cases} 0, & Z_{k,j}^c > x \\ 1, & Z_{k,j}^c \leq x \end{cases} \)

and from this, we set the fourth component equal to the Kolmogorov–Smirnov statistic by

\[
\Delta_j^4 = \sup_x \left| F_j^c(x) - F_j^c(\bar{x}) \right|.
\]

The final component of DCV measures the information gain ratio which captures the ratio between the amounts of entropy reduction when splitting the jth log ratio, as implemented using the R fselectorCpp package with default settings:

\[
\Delta_j^5 = \frac{H(\text{Group}) + H(\mathbf{z}_j) - H(\text{Group}, \mathbf{z}_j)}{H(\text{Group}, \mathbf{z}_j)}
\]

where \( H(X) = -P(x_i) \log \sum P(x_i) \) is Shannon’s entropy and \( H(X, Y) = \sum P(x, y) \log \frac{P(x, y)}{P(x)} \) is conditional Shannon’s entropy. Aggregating these different components, we define the DCV matrix

\[
\mathbf{V} = \begin{bmatrix}
\Delta_1^1 & \cdots & \Delta_1^5 \\
\vdots & \ddots & \vdots \\
\Delta_p^1 & \cdots & \Delta_p^5
\end{bmatrix}
\]

To account for differences in scale between the DCV components, we Z-score standardize each score vector (column) to define the standardized DCV matrix \( \tilde{\mathbf{V}} \) as
\[ \hat{v}_{ij} = \frac{v_{ij} - \bar{v}_{*j}}{SD(v_{*j})}. \]

The final set of DCV scores, \( \bar{V} \in \mathbb{R}^{p \times 1} \), which contains a score for each log ratio, is then defined as

\[ \bar{v}_j = \sum_{k=1}^{5} \hat{v}_{jk}, \text{ where } (j = 1, \ldots, p). \]

Construction of DCV network and conversion to maximum spanning tree

Here we leverage the inherent network structure of log ratios [27] to form our DCV network, defined as a directed graph where edges point from numerator vertices to denominator vertices. We then define \( G = (V, E, W) \) to be the DCV network where \( V \) is the set of \( d \) taxa vertices, \( E \) is the edge set formed by all \( p \) pairwise log ratios formed between taxa vertices, and edge weights \( W \) are the corresponding DCV scores in \( \bar{V} \) between classes. In the initial phase of feature selection on \( Z \), we require the log ratio subsets to meet three important properties: 1) explain maximum log ratio variance, 2) form a linearly independent set, and 3) contain maximum DCV. Notably, the column rank of \( Z \) is \((d - 1)\) and thus the initial subset need only be a connected network containing all \( d \) taxa to explain 100% of the log ratio variance contained in \( Z \). The second property requires the undirected log ratio subset to be acyclic, as may be achieved with a spanning tree; however, the number of spanning trees from \( G \) can be expressed by Cayley’s formula: \( T_{|V|} = |V|^{(|V|-2)} \). To circumvent considering this unmanageable number of spanning trees, we utilize the weights imposed from the DCV scoring to enable efficient selection of a suitable spanning tree from \( G \), as described next.

To simultaneously meet all properties described above, we compute a Maximum Spanning Tree (MST), containing all vertices without cycles and with maximum edge weight (among spanning trees). To make this computationally tractable, we first sort the log ratios of \( \bar{V} \) in descending order by DCV score to form \( \bar{V}' \) and retain the first set of \( q \) log ratios that contain all \( d \) taxa to form \( \bar{V}'' \). We then define the log ratio
network $G = (V, E, W)$ where $V$ is the set of $d$ taxa vertices, $E$ is the edge set of these $q$ pairwise log ratios, with edge weights $W$ in $\bar{V}'$. In practice, we have always found that the resulting networks at this stage are connected — in the event that the network is not a single connected component, additional log ratios from $\bar{V}'$ should be added to make it connected. Finally, from $G$ we compute the MST tree $G_{MST}$ using the R igraph package. Selecting from $Z$, we define $Z' \in \mathbb{R}^{n \times (d-1)}$ to be the subset of log ratios corresponding to the edge set of $G_{MST}$.

**Multivariate Test Statistics**

SelEnergyPerm considers two multivariate test statistics to determine the statistical significance of retained subsets of log ratios.

The first multivariate test statistic, the Distance Components F-ratio (discoF) is utilized when between-group dispersion effects are not detected in $Z'$. The discoF statistic was proposed in [62] and is similar to the traditional ANOVA ‘F’ ratio (but does not follow an F-distribution) where the total dispersion is partitioned into between- and within-group components derived from an inter-sample Euclidean distance matrix computed from $Z'$. Computation of the discoF statistic is done here using the energy R package. As described in [62] the discoF test statistic, for binary groups is of the form

$$ F_{n, \alpha} = \frac{S_{n, \alpha}}{W_{n, \alpha}/(n - 2)} $$

where $S_{n, \alpha}$ is the between-sample energy statistic, $W_{n, \alpha}$ is the within-sample dispersion statistic and $0 < \alpha \leq 2$ is the exponent on the pairwise between sample norm. See [37, 62] for specific details on computing the between- and within-group components of the discoF statistic. Here we use the energy R package to compute the discoF statistic.

The second statistic used by SelEnergyPerm is a combined-F ($cF$) statistic which is distribution-free and attempts to jointly account for differences in both location and scale between distributions.
SelEnergyPerm selects the combined-F statistic when dispersion effects between groups are detected in $Z'$. In particular, the SelEnergyPerm unscaled $cF$ statistic is the sum of F-ratios obtained from the PERMDISP2 with spatial medians [48] and PERMANOVA [35] procedures, computed using the R vegan package. In terms of the distances in $D$, we partition the variation of $Z'$ and define the combined-F statistic as

$$cF = F_{\text{location}} + F_{\text{dispersion}} = \left( \frac{SS_a}{SS_w/(n-2)} \right) + \left( \frac{SS_T}{SS_E/(n-2)} \right)$$

where $SS_a$ and $SS_T$ are the between-group sum of squares components and $SS_w$ and $SS_E$ are the within-group sum of square components of variation from the PERMANOVA ($F_{\text{location}}$) and PERMDISP2 ($F_{\text{dispersion}}$) procedures, respectively. See [35] and [48] for specific details on computing these between- and within-group components. Likewise, the scaled combined-F ($\bar{cF}$) statistic, a point estimate, is computed in the same way but with z-score standardizing relative to permutation distribution. Let $\mathbf{nF}_{\text{loc}}$ and $\mathbf{nF}_{\text{disp}}$ be $m$-dimensional vectors of null $F_{\text{loc}}$ or $F_{\text{disp}}$ statistics sampled from the permutation distribution. We then scale $\hat{F}_{\text{loc}} = \frac{F_{\text{loc}} - E[\mathbf{nF}_{\text{loc}}]}{SD(\mathbf{nF}_{\text{loc}})}$ and $\hat{F}_{\text{disp}} = \frac{F_{\text{disp}} - E[\mathbf{nF}_{\text{disp}}]}{SD(\mathbf{nF}_{\text{disp}})}$.

Finally, we define

$$\bar{cF} = \hat{F}_{\text{loc}} + \hat{F}_{\text{disp}}.$$

Of note, $\bar{cF}$ is approximate and thus the estimate has variability based on the number of samples drawn from the permutation distribution. We consider $m = 1e6$ samples here as a balance between computational cost and minimizing this variation.

**Association Maximization and Greedy Forward Selection**

In this step, we focus on the multivariate structure formed by a subset of log ratios. Specifically, we are interested in maximizing the between-group variation induced by a subset of log ratios in a low-dimensional multivariate space. To find a minimal, statistically significant subset of log ratios that
maximizes $F_{n,\alpha}$ (location effects only) or $cF$ (dispersion and location effects) between classes, we utilize a greedy forward stepwise feature selection procedure.

Algorithm 1: Association maximization with greedy forward stepwise selection

1: **Procedure** `selectionEnergy(Z', y, eps = 1e-5, patience = 25, $\alpha = 0.05$)
   
2:   if `PERMDISP2(Z', y)` < $\alpha$ then
   
3:     `testStatisticFunction = cF()`
   
4:     `Metric = 'combinedF'`
   
5:   else
   
6:     `testStatisticFunction = F_{n,\alpha}()`
   
7:     `Metric = 'discoF'`
   
8:   end if
   
9:   `X = Z'[1:3]`  \hspace{1cm}  \(\triangleright\) Start with top 3 LRs
   
10:  `maxF = testStatisticFunction(baseSet, y)`
   
11:  `improvementTime = 0`
   
12:  for $x$ in 4, ..., `numCols(Z')` do
   
13:      `X_{new} = X U Z'[x]`
   
14:     `newF = testStatisticFunction(newSet, y)`
   
15:     `diff = newF - maxF`
   
16:     if `diff >= eps` then
   
17:        `X = X_{new}`
   
18:     `maxF = newF`
   
19:   end if
   
20:  if `diff >= eps` then
improvementTime = 0
else
    improvementTime = improvementTime + 1
end if
if improvementTime > patience then
    Break
end if
end for
if Metric == ‘discoF’ then
    testStat = \( F_{n,\alpha} \) | X
else
    testStat = \( c^F \) | X
end if
return list( X, testStat )
end procedure

Association Significance testing

To assess the statistical significance of the association \( F^{obs} = \text{selectionEnergy}(\mathbf{Z}', \mathbf{y})[\text{testStat}] \) computed on the log-ratio subset \( \mathbf{S} = \text{selectionEnergy}(\mathbf{Z}', \mathbf{y})[\mathbf{X}] \) returned from Algorithm 1, we utilize permutation testing via Monte Carlo sampling [63]. Specifically, we test if the \( F^{obs} \) is more extreme than \( F \)-ratios sampled from the permutation distribution \( (F^{null}) \). In particular, log-ratio subsets and association statistics are computed using Algorithm 1 with permuted labels \( (\mathbf{y}) \). With \( k \) random samples of \( F^{null} \) the one-sided estimated exact p-value becomes
\[ \hat{p} = \frac{1 + \sum_{i=1}^{k} I(f_i^{\text{null}} > F_{\text{obs}})}{k + 1} \]

**Simulation Strategy**

We adapted several simulation settings to investigate and highlight key association detection characteristics of SelEnergyPerm when compared to ANOSIM, PERMANOVA, and the energy test. Additionally, to detect the presence of heterogeneity of multivariate dispersion between groups and understand its impact on association detection, we utilized the PERMDISP2 method as an indicator. The empirical association detection ability of each method was assessed within a binary classification framework. To do this, we measured the rate of each statistical test to correctly reject (Power) or accept (Type I Error) the null hypothesis (no difference between groups) at significance \( \alpha = 0.05 \). Further, to truly assess detection capabilities, we presented each method with binary instances drawn from either the same (Null Case) or different (True Case) distributions for each scenario using Monte Carlo simulations. The Matthews Correlation Coefficient (MCC), which effectively summarizes the binary confusion matrix, was then used to measure the overall accuracy of each method’s ability to detect associations across various simulation scenarios. MCC was computed as

\[ \text{MCC} = \frac{(TP + TN) - (FP + FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \]

where \( TP = \) true positive (reject the null hypothesis for True Case), \( TN = \) true negative (accept null hypothesis for Null Case), \( FP = \) false positive (reject the null hypothesis for Null Case), \( FN = \) false negative (accept null hypothesis for True Case). For each simulation scenario, we generated 100 simulated datasets with 40 samples in both Class 1 \( (S_1) \) and Class 2 \( (S_2) \) for the balanced binary design and 20/60 (Class 1 / Class 2) samples for the unbalanced design for each case. Given all methods rely on permutation testing for significance, we generate a common set of 150 permutations per dataset to consistently compute significance for each method across all scenarios and settings.
Simulation Scenarios (Synthetic Data)

For all synthetic data scenarios, we consider datasets with $d = 50, 150, 250$ dimensions, yielding a total of $p = 1225, 11175, 31125$ pairwise log ratios, respectively. Each of the following simulation scenarios is available in our \textit{SelEnergyPermR} R package available at \url{https://github.com/andrew84830813/selEnergyPermR} using the function \texttt{scenarioN()} where $N = [1,5]$.

In Scenario 1, for the true case, we consider both multivariate location (in all dimensions) and dispersion effects that grow with increased dimensions. The increase in dispersion with dimension is similar to settings studied in [42]. Here, data from each sample arise from the Dirichlet distribution $\text{Dir}(\alpha)$ whereby data are naturally constrained within the unit-sum simplex and are commonly used to model compositional data. Data from class 1 are simulated with $\alpha_1 = (3,3,\ldots,3)$. Data from class 2 are generated with $\alpha_2 = \left(3 \cdot \frac{\log(d)}{5}, 3 \cdot \frac{\log(d)}{5}, \ldots, 3 \cdot \frac{\log(d)}{5}\right)$ where the $\frac{\log(d)}{5}$ term shifts the overall location and increases dispersion as the dimensionality increases. For the null case, data from both classes are generated from $\text{Dir}(\alpha_1)$.

In Scenario 2, for the true case, we generate sparse count data from two Dirichlet distributions that differ in the location of the first component and overall dispersion. Library size or total counts for the $i$th sample were modeled using a negative binomial (NB) distribution and simulated as $C_i \sim \text{NB}(s \times \frac{s}{s+\mu})$ where $s = 1$ and $\mu = 1e7$. Count data for class 1 were generated by rounding:

$C_i \ast \text{Dir}(\alpha_1)$ where $\alpha_1 = \left(U_1 \sim \text{unif}(3000,5000), U_{i \in [2,10]} \sim \text{unif}(500,1500), U_{i \in [11,d]} \sim \text{unif}(1,5)\right)$

Count data for class 2 were generated as:

$C_i \ast \text{Dir}(\alpha_1)$ where $\alpha_1 = \left(U_1 \sim \text{unif}(12500,17500), U_{i \in [2,10]} \sim \text{unif}(500,1500), U_{i \in [11,d]} \sim \text{unif}(1,5)\right)$

Notably, the $U_{i \in [11,d]} \sim \text{unif}(1,5)$ terms are used to model sparsity. For the null case, data from both classes are generated from $C_i \ast \text{Dir}(\alpha_1)$ vector.
In Scenario 3, for the true case, we generate compositional data whereby there exists a large location effect that increases while the dispersion effects decrease with dimensionality. These settings are similar to settings considered for association benchmark comparisons in [42]. In particular, we simulate data from the additive logistic normal distribution on the simplex [21]. To do this we first let $S_1 = N(\mu_1, \Sigma_1)$ and $S_2 = N(\mu_2, \Sigma_2)$ be samples drawn from multivariate normal distributions. We set $\mu_1 = (0, 0, \ldots, 0)$ and $\mu_2$ to be equal to $1/\sqrt{d}$ in the first 25% of dimensions and 0 in the remaining dimensions. The covariance structure was defined in the same way as [42] where $\Sigma$ was defined with 1’s along the main diagonal and 0.2 along the two diagonals off the main. From this, $\Sigma_1 = \Sigma + \delta I_d$ and $\Sigma_2 = \Sigma + U + \delta I_d$ where $U$ is a $d \times d$ matrix with $\text{unif}(0, 32/d^2)$ entries and $\delta = |\min (\text{eigenvalues}(\Sigma), \text{eigenvalues}(\Sigma + U))| + 0.05$. Here row vectors from $S$ represent additive log-ratio (ALR) vectors and are subsequently projected onto the simplex using the inverse additive log-ratio transformation defined as $\text{ALR}^{-1} = C(\exp(s))$. For the null case, data for both classes were simulated from $N(\mu_1, \Sigma_1)$.

In Scenario 4, for the true case, we generate compositional data where there exist sparse location effects in the first dimension that grow stronger while dispersion effects grow weaker as the dimensionality increases. That is, $S_1 = N(\mu_1, \Sigma_1)$ and $S_2 = N(\mu_2, \Sigma_2)$ are defined as in scenario 3 except for $\mu_2$ being set to $\log\left(\frac{d}{3}\right)$ in the first dimensions and 0 in the remaining dimensions. The simplex projection and null case are done as described in scenario 3.

Finally, in Scenario 5 for the true case, we generate compositional data from the additive logistic normal distribution with a small location shift and large dispersion difference that increases with dimensionality. Let $S_1 = N(\mu_1, \Sigma_1)$ and $S_2 = N(\mu_2, \Sigma_2)$ be defined in as in scenario 3 except for $\mu_2$ set to $\frac{1}{\sqrt{n_1+n_2}}$ in all dimensions and $U$ now being a $d \times d$ matrix with $\text{unif}(0,32)$. The simplex projection and null case are done as described in scenario 3.
For all experimental data scenarios, we used publicly available taxa count tables where sequencing data were already pre-processed. Each of the following simulation scenarios are available in our `SelEnergyPermR` R package available at https://github.com/andrew84830813/selEnergyPermR using the functions `simFromExpData.covarianceShift()` or `simFromExpData.largeMeanShift()`.

For the 16S data characteristics, we utilized the ob_goodrich_results.tar.gz dataset from the microbiomeHD [64, 65] database. We then aggregated the taxa to the genus level (distinct genera = 247) and extracted the 428 healthy samples from the goodrich16S dataset for our 16S data simulations. For WGS data characteristics, we utilized the ZeeviD2015 [66] dataset from the curatedmetagenome [67] database. We aggregated taxa counts by species (distinct species = 1,776) and then extracted the 900 control samples for our WGS data simulations. Here we model the 16S and WGS count data using zero-inflated negative binomial (ZINB) models which have been shown to be a reasonable choice for modeling microbiome count data [68]. ZINB models were fit to the 16S and WGS dataset described above using the ZINBWAVE R package and default settings. For all experimental data scenarios, we used the fitted 16S/WGS ZINB models to simulate new samples for each dataset. That is, we first simulated 428 samples from the ZINB model for the 16S datasets or 900 samples for the WGS datasets. We then randomly select 40 samples per class ($S_1$ and $S_2$) for the balanced design and 20/60 ($S_1$ and $S_2$) samples for the unbalanced design. To reduce the presence of rare features we only retained features present in at least 15% of all samples for all datasets.

For Scenario 1, we consider for the true case in both 16S and WGS datasets, settings where the percent $P = (5, 20, 35, 50)$ of dimensions with a location shift increases while the dispersion effect between classes remains fixed. To do this, we first simulate count data $X$ from the ZINB model and map it onto
the unit-sum simplex using Eq.1 and impute zeros using Eq.2. The ALR transformation is then applied to $X$ as

$$x_{ij} = \ln \left( \frac{x_{ij}}{x_{ip}} \right), j \neq p \text{ and } p = d$$

Using the ALR-transformed data we define $Z = N(\mu, \Sigma)$ where

$$\mu = \mathbb{E}[X] = (\mathbb{E}[X_1], \mathbb{E}[X_2], ..., \mathbb{E}[X_{d-1}])^T$$

and

$$\Sigma_{ij} = \text{cov}[X_i, X_j]$$

The variance ( diag($\Sigma$) ) of each dimension is ranked in ascending order whereby $\mu$ and $\Sigma$ are reordered accordingly to form $\mu_r$ and $\Sigma_r$. Of note, this is done to ensure the location shift occurs in features with minimal variance. We then simulate $S_1$ from $N(\mu_r, \Sigma_r)$ and define $\mu_1$ and $\Sigma_1$ using Eqs.4-5. Letting $\mu_2 = \mu_1$ we then shift the first $P_i\%$ of dimensions of $\mu_2$ by a factor of 1.25. From this we simulate $S_2$ from $N(\mu_2, \Sigma_1)$. Finally, $S_1$ and $S_2$, which are in Euclidean ALR form, are mapped back to the simplex (relative abundance) using the inverse ALR transformation. For the null case data for both classes are simulated from $N(\mu_r, \Sigma_r)$.

Finally, for Scenario 2, we consider for the true case in both 16S and WGS datasets, settings with a location shift in the first 10% of dimensions confounded by increasing dispersion effects. Here we compute $S_1$ in Euclidean ALR form as described in Scenario 2 (Experimental Data) such that $S_1 \sim N(\mu_1, \Sigma_1)$. From this, $\Sigma_{s_1} = \Sigma_1 + \delta I_d$ and $\Sigma_{s_2} = \Sigma_1 + U + \delta I_d$ where $U$ is a $d \times d$ matrix with $\text{unif}(0, \beta_i)$ entries and $\delta = |\min (\text{eigenvalues}(\Sigma), \text{eigenvalues}(\Sigma + U))| + 0.05$. For 16S data $\beta = (0.10, 1.40, 2.70, 4.00)$ and for WGS data $\beta = (0.10, 4.07, 8.03, 12.00)$. Additionally, letting $\mu_2 = \mu_1$ we shift the first 10% of dimensions of $\mu_2$ by a constant factor of 1.25 for WGS data and by a factor $F = (1.20, 1.17, 1.13, 1.10)$ for 16S data. In all, the final multivariate forms are $S_1 \sim N(\mu_1, \Sigma_{s_1})$ and
$S_1 \sim N(\mu_2, \Sigma_{\Sigma_2})$. These distributions, which are in ALR form, are mapped back onto the simplex using the ALR$^{-1}$ transform. Lastly, for the null case, data for both classes are simulated from $N(\mu_1, \Sigma_{\Sigma_1})$.

**Feature Selection Benchmarks**

For the feature selection (FS) benchmark we used the Boruta R package with maxRun set to 100 and importance set to Gini for the Boruta FS. The glmnet R package was used for LASSO FS where alpha was set to 1 and lambda was optimized via cross-validation. The caret R package was used to implement RFE FS where 5-fold cross-validation was used to evaluate AUC and feature importance of sets $s = (2^1, 2^2, \ldots, 2^n)$, where $n = \text{floor}(\log_2 p)$. The FSelectorRcpp R package with default settings was used for the Information Gain Filter FS. For each Scenario (Synthetic Data) FS characteristics were evaluated on 200 synthetic datasets across feature space sizes of $p = (1225, 4950, 11175, 19900, 31125)$ log ratios. Performance characteristics considered were the number of log ratios selected, log-ratio network clustering coefficient, and the combined-F statistic. Here we use the number of log ratios selected by each method as a proxy for model complexity. Specifically, higher model complexity or the number of features retained increases the risk of overfitting and unnecessarily reduces the biological interpretation corresponding to the log ratios. Log-ratio networks were formed using the final subset selected by each method, defined as a graph where vertices represent taxa and edges connect taxa pairs to form a log ratio. Redundancy in a log-ratio network of this type can be inferred from cycles in the network. While it does not detect all cycles, the clustering coefficient can be used here to detect cycles between three nodes (closed triangles versus triplets). Computation of the global clustering coefficient was done using the R igraph package. Finally, the $cF$ statistic, which measures the strength of the overall association was computed as in Eq. 3 for each subset. All performance characteristics were evaluated in both balanced and unbalanced sampling designs. Computational time was recorded in seconds for each simulation scenario, feature space, and sample design. The recorded time represents the CPU time required by each FS method to select the final log ratio subset. All computations were run on UNC-
Chapel Hill’s Linux-based Longleaf cluster in R parallelized with 10 cores using the `foreach` R package with 5GB of RAM.

**Discussion**

We presented SelEnergyPerm, a group association testing framework for high-dimensional metagenomic data with sparse microbiome associations between groups. Our framework directly accounts for the compositional sample space imposed on these data as a result of technical variations in sample-wise library size [17, 19]. This is done by using embedded feature selection on a set of robust all pairwise log ratios to improve the detection and interpretation of sparse signals hidden in these data.

Each log ratio is first ranked using our novel DCV score followed by the application of network methods and feature selection techniques to effectively select subsets of log ratios. In tandem, these steps help to identify log ratio signatures capable of explaining microbiome-derived phenotypic differences.

Further, false discovery is properly controlled for by repeating the entire process with permuted labels in unrestricted or restricted single or restricted permutation test designs for statistical significance [64].

We assessed our method by conducting an extensive simulation study to rigorously benchmark the performance of the method relative to popular alternatives for both feature selection and association testing. Our simulation scenarios included data from both synthetic and empirical distributions where scenarios included settings with small/large location shifts embedded in sparse/dense signals and small/large location shifts with covariance differences embedded in sparse/dense association signals carried out on both balanced and unbalanced sample designs. When compared to popular alternatives, we show our SelEnergyPerm feature selection approach is, overall, able to select fewer log ratios, guarantee log-ratio subsets are independent, and better maximize between-group associations with relatively modest computational time requirements. Additionally, when compared to common association testing methods used in metagenomic studies, we show SelEnergyPerm can consistently
detect associations better than or comparable with the alternatives in nearly all simulation settings tested. The better performance of SelEnergyPerm is most notable when sparse association signals are present.

Our demonstration of how SelEnergyPerm can be used to gain robust and unique biological insight was carried out in detail with data from 4 case studies. In the first case study, SelEnergyPerm successfully detected a confirmed association between the composition of the microbiome in CSF and PIH/NPIH disease status in Ugandan infants using a reduced log ratio signature (13 of 1,596 log ratios). Further, we show, given these data that our log-ratio signature can to a greater degree discriminate between disease statuses and explain differences between infants than with a single feature or all pairwise log ratios. In our second case study, SelEnergyPerm detected an association between delivery mode and the gut microbiome composition in infants during the first 2 months of life and at the time 0 collection time. Notably, PERMANOVA and ANOSIM applied over the same time course with all log ratios failed to detect this association. In the third case study, SelEnergyPerm detected an association between the composition of the gut microbiome and abnormal fecal calprotectin levels. Here we found our fecal calprotectin associated log-ratio signature (26 log ratios) had a comparable discriminatory ability to the uninterpretable all log ratios set (7,381 log ratios), thus enabling easier biological interpretation. In the final case study, SelEnergyPerm detected and characterized associations between the microbiome composition in early life and the development of food allergy later in life.

Overall, our results demonstrate that SelEnergyPerm is a powerful framework for detecting sparse association under various scenarios. However, in the presence of heterogeneity of variance and/or unbalanced group designs, common enemies of multivariate association testing method, the power of SelEnergyPerm was reduced, albeit to a lesser degree than the standard methods tested. Therefore, caution should be used when applying SelEnergyPerm in these settings. Additionally, in some scenarios with dense association signals, the performance of SelEnergyPerm was slightly reduced when compared
to standard methods. While the power reduction was small, the enhanced interpretation from a smaller log-ratio signature may outweigh the loss of power in some settings.

Notwithstanding these limitations, SelEnergyPerm is the first method to our knowledge to fully utilize the pairwise log ratio compositional approach in a group association testing framework for metagenomic data. Importantly, given the compositional sample space imposed on these data, where features are relative, our approach enables the discovery of associations using pairwise log ratios which, by design, robustly interpret features relative to one another rather than alone. While the benefits of employing log ratios are well documented, implementing and carrying out these analyses can be challenging and time-consuming in practice. To this end, we developed an R package SelEnergyPermR with functions to perform the method presented in this paper. Additionally, our package enables rapid pre-processing of relative abundance data, calculation of all pairwise log ratios, and multiplicative zero imputation. Our package also includes functions to simulate data from all scenarios presented in this work. Lastly, our approach adds to a small list of compositional methods for testing associations [30-32] and is to our knowledge, the first compositional data method developed for sparse multivariate group association testing in metagenomic data. We also add to a small list of compositional approaches for feature selection [69] however unlike these methods, our approach directly uses pairwise log ratios which enables simple interpretation and may better elucidate taxa-taxa interactions through log ratio network analysis.

We developed SelEnergyPerm to be a powerful group association test for researchers interested in studying 16S or WGS microbiome data with a compositional approach using pairwise log ratios. While not demonstrated explicitly here, SelEnergyPerm is also compatible with multiclass group association testing. Future directions to usefully expand this methodology could focus on incorporating covariate information and extending the framework to longitudinal data.
Declarations

Ethics approval and consent to participate

All datasets used in this manuscript are publicly available. No ethics approval or consent to participate was required for this manuscript.

Consent for publication

All datasets used in this manuscript are publicly available. No ethics approval or consent to participate was required for this manuscript.

Availability of data and material

Public metagenomic data used in the analysis presented in this paper can be directly accessed for (1) Ugandan infant’s cerebrospinal fluid microbiomes and post-infectious hydrocephalus (https://microbiomedb.org/mbio/app/downloads/release-22/953b8ff2d4ba436fa16c381916a57850b43a2a58), (2) delivery mode and the composition of infant’s gut microbiomes over the first three months of life (https://qiita.ucsd.edu/; study ID = 10249), (3) adult gut microbiomes and abnormal fecal calprotectin levels (https://ibdmdb.org/tunnel/public/HMP2/WGS/1818/products; taxonomic_profiles_3.tsv.gz, HMP2 Metadata), and (4) the gut microbiome composition of infants within the first 6 months of life and future allergic sensitization to egg, milk, or peanuts (https://diabimmune.broadinstitute.org/diabimmune/; Three Study Cohort). Data used in real 16S and WGS data simulation were accessed from 16S (https://zenodo.org/record/1146764#.YOc3KehKhEY; ob_goodrich_results.tar.gz) and WGS (https://doi.org/doi:10.18129/B9.bioc.curatedMetagenomicData; ZeeviD2015).
The selEnergyPerm method has been implemented in a freely available R packaged at:

https://github.com/andrew84830813/selEnergyPermR.git. All analysis was done in R 4.0.5. Analysis R script and input (publicly accessed data) and output (processed in this study) data can be accessed from:

https://github.com/andrew84830813/SelectionEnergyPerm_Project.git

Competing interests

The authors declare that they have no competing interests.

Funding.

This research was funded by an HHMI Gilliam Award (GT11504) and a JSMF Complex Systems Scholar Award (#220020315). The content is solely the responsibility of the authors and does not necessarily represent the official views of any agency funding this research.

Authors’ contributions

ALH developed the framework methodology, performed simulations and case studies, developed the software package and wrote the manuscript. PJM oversaw the development of the framework methodology and wrote the manuscript. ALH and PJM approved the final manuscript.

Acknowledgments.

We acknowledge and are grateful to our collaborators for many conversations that helped motivate and further refine this effort, including Jeff Henderson, Elise Hickman, Ilona Jaspers, Mike Kulis, Laura Marks, and John Robinson.

References

1. Martín, R., Miquel, S., Langella, P., & Bermúdez-Humarán, L. G. (2014). The role of metagenomics in understanding the human microbiome in health and disease. Virulence, 5(3), 413–423.

https://doi.org/10.4161/viru.27864
2. Ranjan, R., Rani, A., Metwally, A., McGee, H. S., & Perkins, D. L. (2016). Analysis of the microbiome: Advantages of whole genome shotgun versus 16S amplicon sequencing. Biochemical and Biophysical Research Communications, 469(4), 967–977. https://doi.org/10.1016/j.bbrc.2015.12.083

3. Gopalakrishnan, V., Helmink, B. A., Spencer, C. N., Reuben, A., & Wargo, J. A. (2018). The Influence of the Gut Microbiome on Cancer, Immunity, and Cancer Immunotherapy. Cancer Cell, 33(4), 570–580. https://doi.org/10.1016/j.ccell.2018.03.015

4. Manichanh, C., Borruel, N., Casellas, F., & Guarner, F. (2012). The gut microbiota in IBD. Nature Reviews Gastroenterology & Hepatology, 9(10), 599–608. https://doi.org/10.1038/nrgastro.2012.152

5. Turnbaugh, P. J., Ley, R. E., Mahowald, M. A., Magrini, V., Mardis, E. R., & Gordon, J. I. (2006). An obesity-associated gut microbiome with increased capacity for energy harvest. Nature, 444(7122), 1027–1031. https://doi.org/10.1038/nature05414

6. Qin, N., Yang, F., Li, A., Prifti, E., Chen, Y., Shao, L., Guo, J., Le Chatelier, E., Yao, J., Wu, L., Zhou, J., Ni, S., Liu, L., Pons, N., Batto, J. M., Kennedy, S. P., Leonard, P., Yuan, C., Ding, W., … Li, L. (2014). Alterations of the human gut microbiome in liver cirrhosis. Nature, 513(7516), 59–64. https://doi.org/10.1038/nature13568

7. Foster, J. A., & McVey Neufeld, K.-A. (2013). Gut–brain axis: How the microbiome influences anxiety and depression. Trends in Neurosciences, 36(5), 305–312. https://doi.org/10.1016/j.tins.2013.01.005

8. Kong, H. H., Oh, J., Deming, C., Conlan, S., Grice, E. A., Beatson, M. A., Nomicos, E., Polley, E. C., Komarow, H. D., Murray, P. R., Turner, M. L., & Segre, J. A. (2012). Temporal shifts in the skin microbiome associated with disease flares and treatment in children with atopic dermatitis. Genome Research, 22(5), 850–859. https://doi.org/10.1101/gr.131029.111
9. Dewhirst, F. E., Chen, T., Izard, J., Paster, B. J., Tanner, A. C. R., Yu, W.-H., Lakshmanan, A., & Wade, W. G. (2010). The human oral microbiome. Journal of Bacteriology, 192(19), 5002–5017. https://doi.org/10.1128/JB.00542-10

10. Wilson, M. T., & Hamilos, D. L. (2014). The Nasal and Sinus Microbiome in Health and Disease. Current Allergy and Asthma Reports, 14(12), 485. https://doi.org/10.1007/s11882-014-0485-x

11. Zackular, J. P., Rogers, M. A. M., Ruffin, M. T., & Schloss, P. D. (2014). The Human Gut Microbiome as a Screening Tool for Colorectal Cancer. Cancer Prevention Research, 7(11), 1112–1121. https://doi.org/10.1158/1940-6207.CAPR-14-0129

12. Schlaberg, R. (2020). Microbiome Diagnostics. Clinical Chemistry, 66(1), 68–76. https://doi.org/10.1373/clinchem.2019.303248

13. Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., Fierer, N., Peña, A. G., Goodrich, J. K., Gordon, J. I., Huttley, G. A., Kelley, S. T., Knights, D., Koenig, J. E., Ley, R. E., Lozupone, C. A., McDonald, D., Muegge, B. D., Pirrung, M., ... Knight, R. (2010). QIIME allows analysis of high-throughput community sequencing data. Nature Methods, 7(5), 335–336. https://doi.org/10.1038/nmeth.f.303

14. Schloss Patrick D., Westcott Sarah L., Ryabin Thomas, Hall Justine R., Hartmann Martin, Hollister Emily B., Lesniewski Ryan A., Oakley Brian B., Parks Donovan H., Robinson Courtney J., Sahl Jason W., Stres Blaz, Thallinger Gerhard G., Van Horn David J., & Weber Carolyn F. (2009). Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing 15. Truong, D. T., Franzosa, E. A., Tickle, T. L., Scholz, M., Weingart, G., Pasolli, E., Tett, A., Huttenhower, C., & Segata, N. (2015). MetaPhlAn2 for enhanced metagenomic taxonomic profiling. Nature Methods, 12(10), 902–903. https://doi.org/10.1038/nmeth.3589

16. Wood, D. E., Lu, J., & Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. Genome Biology, 20(1), 257. https://doi.org/10.1186/s13059-019-1891-0
17. Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V., & Egozcue, J. J. (2017). Microbiome Datasets Are Compositional: And This Is Not Optional. Frontiers in Microbiology, 8. https://doi.org/10.3389/fmicb.2017.02224

18. Gloor, G. B., & Reid, G. (2016). Compositional analysis: A valid approach to analyze microbiome high-throughput sequencing data. Canadian Journal of Microbiology, 62(8), 692–703. https://doi.org/10.1139/cjm-2015-0821

19. Weiss, S., Xu, Z. Z., Peddada, S., Amir, A., Bittinger, K., Gonzalez, A., Lozupone, C., Zaneveld, J. R., Vázquez-Baeza, Y., Birmingham, A., Hyde, E. R., & Knight, R. (2017). Normalization and microbial differential abundance strategies depend upon data characteristics. Microbiome, 5. https://doi.org/10.1186/s40168-017-0237-y

20. Pearson, K. (1897). Mathematical contributions to the theory of evolution.—On a form of spurious correlation which may arise when indices are used in the measurement of organs. Proceedings of the Royal Society of London, 60(359–367), 489–498. https://doi.org/10.1098/rspl.1896.0076

21. Aitchison, J. (1986). The Statistical Analysis of Compositional Data. London: Chapman & Hall.

22. Greenacre, M., & Lewi, P. (2009). Distributional Equivalence and Subcompositional Coherence in the Analysis of Compositional Data, Contingency Tables and Ratio-Scale Measurements. Journal of Classification, 26(1), 29–54. https://doi.org/10.1007/s00357-009-9027-y

23. Pawlowsky-Glahn, V., & Buccianti, A. (2011). Compositional Data Analysis: Theory and Applications. John Wiley & Sons.

24. Martín-Fernández, J.-A., Hron, K., Templ, M., Filzmoser, P., & Palarea-Albaladejo, J. (2015). Bayesian-multiplicative treatment of count zeros in compositional data sets. Statistical Modelling, 15(2), 134–158. https://doi.org/10.1177/1471082X14535524
25. Martin-Fernandez, J. A., Barcelo-Vidal, C., & Pawlowsky-Glahn, V. (2003). Dealing with Zeros and Missing Values in Compositional Data Sets Using Nonparametric Imputation. Mathematical Geology, 26.

26. Egozcue, J. J., Pawlowsky-Glahn, V., Mateu-Figueras, G., & Barceló-Vidal, C. (2003). Isometric Logratio Transformations for Compositional Data Analysis. Mathematical Geology, 35(3), 279–300. https://doi.org/10.1023/A:1023818214614

27. Greenacre, M. (2019). Variable Selection in Compositional Data Analysis Using Pairwise Logratios. Mathematical Geosciences, 51(5), 649–682. https://doi.org/10.1007/s11004-018-9754-x

28. Lovell, D. R., Chua, X.-Y., & McGrath, A. (2020). Counts: An outstanding challenge for log-ratio analysis of compositional data in the molecular biosciences. NAR Genomics and Bioinformatics, 2(2). https://doi.org/10.1093/nargab/lqaa040

29. Quinn, T. P., Erb, I., Gloor, G., Notredame, C., Richardson, M. F., & Crowley, T. M. (2019). A field guide for the compositional analysis of any-omics data. GigaScience, 8(9), giz107. https://doi.org/10.1093/gigascience/giz107

30. Fernandes, A. D., Reid, J. N., Macklaim, J. M., McMurrough, T. A., Edgell, D. R., & Gloor, G. B. (2014). Unifying the analysis of high-throughput sequencing datasets: Characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis. Microbiome, 2, 15. https://doi.org/10.1186/2049-2618-2-15

31. Mandal, S., Treuren, W. V., White, R. A., Eggeshøj, M., Knight, R., & Peddada, S. D. (2015). Analysis of composition of microbiomes: A novel method for studying microbial composition. Microbial Ecology in Health and Disease, 26(1), 27663. https://doi.org/10.3402/mehd.v26.27663

32. Lin, H., & Peddada, S. D. (2020). Analysis of compositions of microbiomes with bias correction. Nature Communications, 11(1), 3514. https://doi.org/10.1038/s41467-020-17041-7

33. American Journal of Roentgenology. 2005;185: 299-309. 10.2214/ajr.185.2.01850299
34. Layeghifard, M., Hwang, D. M., & Guttman, D. S. (2017). Disentangling Interactions in the Microbiome: A Network Perspective. Trends in Microbiology, 25(3), 217–228. https://doi.org/10.1016/j.tim.2016.11.008

35. Anderson, M. J. (2017). Permutational Multivariate Analysis of Variance (PERMANOVA). In Wiley StatsRef: Statistics Reference Online (pp. 1–15). American Cancer Society. https://doi.org/10.1002/9781118445112.stat07841

36. Clarke, K. R. (1993). Non-parametric multivariate analyses of changes in community structure. Australian Journal of Ecology, 18(1), 117–143. https://doi.org/10.1111/j.1442-9933.1993.tb00438.x

37. Rizzo, M. L., & Székely, G. J. (2016). Energy distance. Wiley Interdisciplinary Reviews: Computational Statistics, 8(1), 27–38. https://doi.org/10.1002/wics.1375

38. Wu, C., Chen, J., Kim, J., & Pan, W. (2016). An adaptive association test for microbiome data. Genome Medicine, 8(1), 56. https://doi.org/10.1186/s13073-016-0302-3

39. Koh, H., & Zhao, N. (2020). A powerful microbial group association test based on the higher criticism analysis for sparse microbial association signals. Microbiome, 8(1), 63. https://doi.org/10.1186/s40168-020-00834-9

40. Baumann, K. (2003). Cross-validation as the objective function for variable-selection techniques. TrAC Trends in Analytical Chemistry, 22(6), 395–406. https://doi.org/10.1016/S0165-9936(03)00607-1

41. Lindgren, F., Hansen, B., Karcher, W., Sjöström, M., & Eriksson, L. (1996). Model validation by permutation tests: Applications to variable selection. Journal of Chemometrics, 10(5–6), 521–532. https://doi.org/10.1002/(SICI)1099-128X(199609)10:5/6<521::AID-CEM448>3.0.CO;2-J

42. Wei, S., Lee, C., Wichers, L., & Marron, J. S. (2016). Direction-Projection-Permutation for High-Dimensional Hypothesis Tests. Journal of Computational and Graphical Statistics, 25(2), 549–569. https://doi.org/10.1080/10618600.2015.1027773
43. Kursa, M. B., Jankowski, A., & Rudnicki, W. R. (2010). Boruta – A System for Feature Selection. Fundamenta Informaticae, 101(4), 271–285. https://doi.org/10.3233/FI-2010-288
44. Tibshirani, R. (2011). Regression shrinkage and selection via the lasso: A retrospective.
45. KENT, J. T. (1983). Information gain and a general measure of correlation. Biometrika, 70(1), 163–173. https://doi.org/10.1093/biomet/70.1.163
46. Granitto, P. M., Furlanello, C., Biasioli, F., & Gasperi, F. (2006). Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products. Chemometrics and Intelligent Laboratory Systems, 83(2), 83–90. https://doi.org/10.1016/j.chemolab.2006.01.007
47. Anderson, M. J., & Walsh, D. C. I. (2013). PERMANOVA, ANOSIM, and the Mantel test in the face of heterogeneous dispersions: What null hypothesis are you testing? Ecological Monographs, 83(4), 557–574. https://doi.org/10.1890/12-2010.1
48. Anderson, M. J. (2006). Distance-Based Tests for Homogeneity of Multivariate Dispersions. Biometrics, 62(1), 245–253.
49. Paulson, J. N., Williams, B. L., Hehnly, C., Mishra, N., Sinar, S. A., Zhang, L., Ssentongo, P., Mbabazi-Kabachelor, E., Wijetunge, D. S. S., von Bredow, B., Mulondo, R., Kiwanuka, J., Bajunirwe, F., Bazira, J., Bebell, L. M., Burgoine, K., Couto-Rodriguez, M., Ericson, J. E., Erickson, T., ... Schiff, S. J. (2020). Paenibacillus infection with frequent viral coinfection contributes to postinfectious hydrocephalus in Ugandan infants. Science Translational Medicine, 12(563), eaba0565. https://doi.org/10.1126/scitranslmed.aba0565
50. Oliveira, F. S., Brestelli, J., Cade, S., Zheng, J., Iodice, J., Fischer, S., Aurrecoechea, C., Kissing, J. C., Brunk, B. P., Stoeckert, C. J., Jr, Fernandes, G. R., Roos, D. S., & Beiting, D. P. (2018). MicrobiomeDB: A systems biology platform for integrating, mining and analyzing microbiome experiments. Nucleic Acids Research, 46(D1), D684–D691. https://doi.org/10.1093/nar/gkx1027
51. Paulson J et al. Study: PIH Uganda. MicrobiomeDB. 2020; release 18
   (https://microbiomedb.org/mbio/app/record/dataset/DS_953b8ff2d4)

52. Bokulich, N. A., Chung, J., Battaglia, T., Henderson, N., Jay, M., Li, H., D. Lieber, A., Wu, F., Perez-Perez, G. I., Chen, Y., Schweizer, W., Zheng, X., Contreras, M., Dominguez-Bello, M. G., & Blaser, M. J.
   (2016). Antibiotics, birth mode, and diet shape microbiome maturation during early life. Science Translational Medicine, 8(343), 343ra82-343ra82. https://doi.org/10.1126/scitranslmed.aad7121

53. Gonzalez, A., Navas-Molina, J. A., Kosciolek, T., McDonald, D., Vázquez-Baeza, Y., Ackermann, G., DeReus, J., Janssen, S., Swafford, A. D., Orchanian, S. B., Sanders, J. G., Shorenstein, J., Holste, H., Petrus, S., Robbins-Pianka, A., Brislawn, C. J., Wang, M., Rideout, J. R., Bolyen, E., ... Knight, R. (2018). Qiita: Rapid, web-enabled microbiome meta-analysis. Nature Methods, 15(10), 796–798. https://doi.org/10.1038/s41592-018-0141-9

54. Proctor, L. M., Creasy, H. H., Fettweis, J. M., Lloyd-Price, J., Mahurkar, A., Zhou, W., Buck, G. A., Snyder, M. P., Strauss, J. F., Weinstock, G. M., White, O., Huttenhower, C., & The Integrative HMP (iHMP) Research Network Consortium. (2019). The Integrative Human Microbiome Project. Nature, 569(7758), 641–648. https://doi.org/10.1038/s41586-019-1238-8

55. 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., Casero, D., Courtney, H., Gonzalez, A., Graeber, T. G., Hall, A. B., Lake, K., Landers, C. J., Mallick, H., Plichta, D. R., ... Huttenhower, C. (2019). Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. Nature, 569(7758), 655–662. https://doi.org/10.1038/s41586-019-1237-9

56. Vatanen, T., Kostic, A. D., d’Hennezel, E., Siljander, H., Franzosa, E. A., Yassour, M., Kolde, R., Vlamakis, H., Arthur, T. D., Hämäläinen, A.-M., Peet, A., Tillmann, V., Uibó, R., Mokurov, S., Dorshakova, N., Ilonen, J., Virtanen, S. M., Szabo, S. J., Porter, J. A., ... Xavier, R. J. (2016). Variation in
Microbiome LPS Immunogenicity Contributes to Autoimmunity in Humans. Cell, 165(4), 842–853. https://doi.org/10.1016/j.cell.2016.04.007

57. Broad Institute. https://diabimmune.broadinstitute.org/diabimmune/. Accessed 23 May 2019.

58. Boogaart, K. G. van den, & Tolosana-Delgado, R. (2013). Analyzing Compositional Data with R. Springer-Verlag. https://doi.org/10.1007/978-3-642-36809-7

59. Palarea-Albaladejo, J., & Martín-Fernández, J. A. (2015). zCompositions—R package for multivariate imputation of left-censored data under a compositional approach. Chemometrics and Intelligent Laboratory Systems, 143, 85–96. https://doi.org/10.1016/j.chemolab.2015.02.019

60. Fan, J., & Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 70(5), 849–911. https://doi.org/10.1111/j.1467-9868.2008.00674.x

61. Csardi G, Nepusz T (2006). “The igraph software package for complex network research.” InterJournal, Complex Systems, 1695. https://igraph.org.

62. Rizzo, M. L., & Székely, G. J. (2010). DISCO analysis: A nonparametric extension of analysis of variance. The Annals of Applied Statistics, 4(2), 1034–1055. https://doi.org/10.1214/09-AOAS245

63. Ernst, M. D. (2004). Permutation Methods: A Basis for Exact Inference. Statistical Science, 19(4), 676–685. https://doi.org/10.1214/088342304000000396

64. Goodrich, J. K., Waters, J. L., Poole, A. C., Sutter, J. L., Koren, O., Blekhman, R., Beaumont, M., Van Treuren, W., Knight, R., Bell, J. T., Spector, T. D., Clark, A. G., & Ley, R. E. (2014). Human Genetics Shape the Gut Microbiome. Cell, 159(4), 789–799. https://doi.org/10.1016/j.cell.2014.09.053

65. Duvallet, Claire, Gibbons, Sean, Gurry, Thomas, Irizarry, Rafael, & Alm, Eric. (2017). MicrobiomeHD: the human gut microbiome in health and disease [Data set]. Zenodo. http://doi.org/10.5281/zenodo.1146764
Figures

**Fig. 1:** Overview of the SelEnergyPerm framework for non-parametric group association testing in metagenomic data. a. Relative abundance/count data are transformed using all pairwise log ratios. These log ratios are subsequently scored (DCV) and used to efficiently select a subset that: (1) is independent via a maximum spanning tree, (2) maximizes the energy or association between groups via greedy optimization. The entire process is repeated using permutation testing to control false discovery and assess statistical significance. b. Detection/rejection of sparse associations hidden within high dimensional data via simultaneous feature selection and permutation testing.
Fig. 2: Comparison of SelEnergyPerm selected log ratio subsets characteristics with Boruta, Information Gain Filtering, LASSO, and RFE across five simulation scenarios with log ratio dimension = \[ \binom{50}{2}, \binom{100}{2}, \binom{150}{2}, \binom{200}{2}, \binom{250}{2} \] for the balanced sampling design. Using 200 simulations for each scenario-dimension by method we assessed: (Top Row) the clustering coefficient of log ratio networks formed by selected subsets returned from each method, (Middle Row) the magnitude of the association as measured by the cF-statistic on selected subsets returned from each method, and (Bottom Row) the number of log ratios returned by each method. Points are the mean for each experimental condition and error bars indicate 95% confidence interval.

Fig. 3: Comparison of the Matthews Correlation Coefficient measuring the ability of each method to properly detect/reject associations in data generated from synthetic distributions in both balanced and unbalanced sampling designs. For each scenario and log ratio feature space size = \[ \binom{50}{2}, \binom{100}{2}, \binom{150}{2}, \binom{250}{2} \] test datasets were simulated to include data distribution that have either true between-group differences (n=100) or no between-group difference (n=100). Results from the PERMDISP2 procedure are displayed to indicate heterogeneity of variance between groups.

Fig. 4: Comparison of the Matthews Correlation Coefficient measuring the ability of each method to properly detect/reject associations in data simulated from real 16S and WGS data distributions in both balanced and unbalanced sampling designs. For each data type and scenario, datasets were generated to include data distribution that have either true between-group differences (n=100) or no between-group difference (n=100). Results from the PERMDISP2 (dashed line) procedure are displayed to indicate heterogeneity of variance between groups.

Fig. 5: SelEnergyPerm case study examining the association between Ugandan infant’s cerebrospinal fluid microbiomes and post-infectious hydrocephalus using 16S data.

a. SelEnergyPerm permutation test results displaying the null distribution of the cF statistic (Histogram, Density, and Points) and the
empirical cF statistic (dashed red vertical line). b. Random forest (RF) importance weighted directed log ratio network (edges point from numerator to denominator) of the SelEnergyPerm selected signature (nodes = taxa, node size = weighted degree, edges = log ratio, edge width/color = RF variable importance). c. ROC comparisons of disease status discrimination using RF. Models were trained with repeated (r = 50) 10-fold cross-validation using either the SelEnergyPerm Signature, all log ratios, or *Paenibacillus* alone. d. Principal component analysis using the SelEnergyPerm signature. e. (Left) Log ratio means of each log ratio included in the SelEnergyPerm signature. (Right) Loading weights of the first principal component. Significance codes (*, **, ***, ****) indicate BH corrected p-value ≤ (0.05, 0.01, 0.001, 1e-4, 0) for between group Wilcoxon Rank Sum Test. For the log ratio means, positive values indicate numerator more abundant than the denominator and negative values indicate the denominator is more abundant numerator. Error bars indicate the 95% CI of the mean. Notably, error bars that do not span 0 indicate numerator/denominator is on average more abundant than the opposite.

**Fig. 6:** SelEnergyPerm case study examining the association between delivery mode and the gut microbiome composition of infants over the first three months of life using 16S data. a. SelEnergyPerm permutation test (permutations = 1000) results displaying the null distribution of the test statistic (violin and grey points) and the empirical test statistic (red if significant, black otherwise) with Benjamini-Hochberg corrected p values. Test statistics values were z-score scaled (by Collection Month) for ease of visualization. b. AUC comparisons of delivery mode discrimination using PLS-DA. Models were trained with repeated (r = 20) 5-fold stratified (delivery mode and host) cross-validation using either the SelEnergyPerm signature or all log ratios. Points represent the mean AUC and error bars indicate the 95% CI. c. Relative taxa strength by family measuring the importance of each taxon for discriminating between delivery modes across each collection time point. Relative strength was computed using the top 5 nodes derived from the PLS-DA variable importance weighted log ratio networks across each collection time. d. Directed (edges point from numerator to denominator) network of the
SelEnergyPerm-derived signature by month and delivery mode weighted by the absolute log ratio means (nodes = taxa, node size = mean strength, edge = log ratio, edge width = log ratio mean, red edges = negative log ratio mean (incoming node more abundant), blue edges = positive log ratio mean (outgoing node more abundant)).

**Fig. 7:** SelEnergyPerm case study examining the association between abnormal fecal calprotectin levels and the gut microbiome composition in nonIBD and IBD individuals using WGS data. 

- **a.** SelEnergyPerm permutation test results displaying the null distribution of the cF statistic (Histogram, Density, and Points) and the empirical cF statistic (dashed red vertical line).
- **b.** AUC comparisons of fecal calprotectin level (Abnormal/Normal) discrimination using PLS-DA with 2 components. Models were trained with repeated (r = 20) 10-fold cross-validation using either the SelEnergyPerm signature or all log ratios.
- **c.** Directed network (edges point from numerator to denominator) of the SelEnergyPerm-selected log-ratio signature (nodes = taxa, node size = DCV strength, edges = log ratio, edge width/color = PLS-DA Variable Importance). The top 5 taxa names by strength (PLS-DA Variable Importance) are displayed.
- **c.** PLS-DA latent space projection plot extracted from final PLS-DA model fit using the full dataset with the SelEnergyPerm signature. Points represent non-IBD or IBD samples.
- **d.** Log ratio means of each log ratio included in the SelEnergyPerm signature demonstrating shared and distinct log ratio markers across diagnosis. Positive values indicate numerator more abundant than the denominator and negative values indicate the denominator is more abundant numerator. Significance codes (ns, *, **, ***, ****) indicate BH corrected p-value ≤ (Not Significant, 0.05, 0.01, 0.001, 1e-4, 0) for between group Wilcoxon Rank Sum Test. Error bars indicate the 95% CI of the mean. Notably, error bars that do not span 0 indicate numerator/denominator is on average more abundant than the opposite.

**Fig. 8:** SelEnergyPerm case study examining the association between the gut microbiomes of infants in early life and the development of food allergy later in life. 

- **a.** SelEnergyPerm permutation test...
(permutations = 1000) results displaying the null distribution of the test statistic (violin and grey points) and the empirical test statistic (red if significant, black otherwise) with Benjamini-Hochberg corrected p values. Test statistics values were z-score scaled by collection period to improve visualization. b. AUC comparisons of future food allergy development discrimination using PLS-DA. Models were trained with repeated (r = 20) 10-fold stratified (host and food allergy development) cross-validation using either the SelEnergyPerm signature or all log ratios. Points represent the mean AUC and error bars indicate the 95% CI. c. Relative taxa strength by family measuring the importance of each taxon for discriminating between food allergy statuses later in life across each collection month. Relative strength was computed using the top 5 nodes derived from the PLS-DA variable importance weighted log ratio networks across each collection month. d. Directed (edges point out from numerator to denominator) network of the SelEnergyPerm-derived signature by collection period and food allergy development weighted by the absolute log ratio means (nodes = taxa, node size = mean strength, edge = log ratio, edge width = log ratio mean, red edges = negative log ratio mean (incoming node more abundant), blue edges = positive log ratio mean (outgoing node more abundant)).

Additional File 1: Figure S1: Feature selection computational time comparisons for balanced and unbalanced sampling designs between SelEnergyPerm, LASSO, RFE, RF, Information Gain, and Boruta across each scenario and dimension. Points are the mean for each experimental condition.

Additional File 2: Figure S2: Comparison of SelEnergyPerm selected log ratio subsets characteristics with Boruta, Information Gain Filtering, LASSO, and RFE across five simulation scenarios with log ratio dimension = \[\left\{\binom{50}{2}, \binom{100}{2}, \binom{150}{2}, \binom{200}{2}, \binom{250}{2}\right\}\] for the unbalanced sampling design. Using 200 simulations for each scenario-dimension by method we assessed: (Top Row) the clustering coefficient of log ratio networks formed by selected subsets returned from each method, (Middle Row) the magnitude of the association as measured by the cF-statistic on selected subsets returned from each method, and
The number of log ratios returned by each method. Points are the mean for each experimental condition and error bars indicate 95% confidence interval.

Additional File 3: Figure S3: Overall mean performance comparison for data generated from synthetic distributions aggregated across all scenarios and dimensions using MCC, Sensitivity, Specify, Positive predictive value (PPV), Negative predictive value (NPV), Youden Index, and False Positive Rate (FPR) metric. Error bars indicate standard error.

Additional File 4: Figure S4: Overall mean performance comparison for data generated from 16S and WGS synthetic data aggregated across all scenarios and effect levels using MCC, Sensitivity, Specify, Positive predictive value (PPV), Negative predictive value (NPV), Youden Index, and False Positive Rate (FPR) metric. Error bars indicate standard error.
Figure 1

Overview of the SelEnergyPerm framework for non-parametric group association testing in metagenomic data. 

a. Relative abundance/count data are transformed using all pairwise log ratios. These log ratios are subsequently scored (DCV) and used to efficiently select a subset that: (1) is independent via a maximum spanning tree, (2) maximizes the energy or association between groups via greedy optimization. The entire process is repeated using permutation testing to control false discovery and assess statistical significance.

b. Detection/rejection of sparse associations hidden within high dimensional data via simultaneous feature selection and permutation testing.
Figure 2

Due to technical limitations, the figure 2 caption is only available in the manuscript file.
Figure 3

Due to technical limitations, the figure 3 caption is only available in the manuscript file.
Figure 4

Comparison of the Matthews Correlation Coefficient measuring the ability of each method to properly detect/reject associations in data simulated from real 16S and WGS data distributions in both balanced and unbalanced sampling designs. For each data type and scenario, datasets were generated to include data distribution that have either true between-group differences (n=100) or no between-group difference (n=100). Results from the PERMDISP2 (dashed line) procedure are displayed to indicate heterogeneity of variance between groups.
Figure 5

SelEnergyPerm case study examining the association between Ugandan infant’s cerebrospinal fluid microbiomes and post-infectious hydrocephalus using 16S data. a. SelEnergyPerm permutation test results displaying the null distribution of the cF statistic (Histogram, Density, and Points) and the 49 empirical cF statistic (dashed red vertical line). b. Random forest (RF) importance weighted directed log ratio network (edges point from numerator to denominator) of the SelEnergyPerm selected signature (nodes = taxa, node size = weighted degree, edges = log ratio, edge width/color = RF variable importance). c. ROC comparisons of disease status discrimination using RF. Models were trained with repeated (r = 50) 10-fold cross-validation using either the SelEnergyPerm Signature, all log ratios, or Paenibacillus alone. d.
Principal component analysis using the SelEnergyPerm signature. (Left) Log ratio means of each log ratio included in the SelEnergyPerm signature. (Right) Loading weights of the first principal component. Significance codes (*, **, ***, ****) indicate BH corrected p-value < (0.05, 0.01, 0.001, 1e-4, 0) for between group Wilcoxon Rank Sum Test. For the log ratio means, positive values indicate numerator more abundant than the denominator and negative values indicate the denominator is more abundant numerator. Error bars indicate the 95% CI of the mean. Notably, error bars that do not span 0 indicate numerator/denominator is on average more abundant than the opposite.

Figure 6
SelEnergyPerm case study examining the association between delivery mode and the gut microbiome composition of infants over the first three months of life using 16S data. a. SelEnergyPerm permutation test (permutations = 1000) results displaying the null distribution of the test statistic (violin and grey points) and the empirical test statistic (red if significant, black otherwise) with Benjamini-Hochberg corrected p values. Test statistics values were z-score scaled (by Collection Month) for ease of visualization. b. AUC comparisons of delivery mode discrimination using PLS-DA. Models were trained with repeated (r = 20) 5-fold stratified (delivery mode and host) cross-validation using either the SelEnergyPerm signature or all log ratios. Points represent the mean AUC and error bars indicate the 95% CI. c. Relative taxa strength by family measuring the importance of each taxon for discriminating between delivery modes across each collection time point. Relative strength was computed using the top 5 nodes derived from the PLS-DA variable importance weighted log ratio networks across each collection time. d. Directed (edges point from numerator to denominator) network of the 50 SelEnergyPerm-derived signature by month and delivery mode weighted by the absolute log ratio means (nodes = taxa, node size = mean strength, edge = log ratio, edge width = log ratio mean, red edges = negative log ratio mean (incoming node more abundant), blue edges = positive log ratio mean (outgoing node more abundant)).
Figure 7

SelEnergyPerm case study examining the association between abnormal fecal calprotectin levels and the gut microbiome composition in nonIBD and IBD individuals using WGS data. a. SelEnergyPerm permutation test results displaying the null distribution of the cF statistic (Histogram, Density, and Points) and the empirical cF statistic (dashed red vertical line). b. AUC comparisons of fecal calprotectin level (Abnormal/Normal) discrimination using PLS-DA with 2 components. Models were trained with repeated (r = 20) 10-fold cross-validation using either the SelEnergyPerm signature or all log ratios. Points represent the mean AUC and error bars indicate the 95% CI. c. Directed network (edges point from...
numerator to denominator) of the SelEnergyPerm-selected log-ratio signature (nodes = taxa, node size = DCV strength, edges = log ratio, edge width/color = PLS-DA Variable Importance). The top 5 taxa names by strength (PLS-DA Variable Importance) are displayed. c. PLS-DA latent space projection plot extracted from final PLS-DA model fit using the full dataset with the SelEnergyPerm signature. Points represent non-IBD or IBD samples. d. Log ratio means of each log ratio included in the SelEnergyPerm signature demonstrating shared and distinct log ratio markers across diagnosis. Positive values indicate numerator more abundant than the denominator and negative values indicate the denominator is more abundant numerator. Significance codes (ns, *, **, *** , ****) indicate BH corrected p-value < (Not Significant, 0.05, 0.01, 0.001, 1e-4, 0) for between group Wilcoxon Rank Sum Test. Error bars indicate the 95% CI of the mean. Notably, error bars that do not span 0 indicate numerator/denominator is on average more abundant than the opposite.
SelEnergyPerm case study examining the association between the gut microbiomes of infants in early life and the development of food allergy later in life. a. SelEnergyPerm permutation test 5 (permutations = 1000) results displaying the null distribution of the test statistic (violin and grey points) and the empirical test statistic (red if significant, black otherwise) with Benjamini-Hochberg corrected p values. Test statistics values were z-score scaled by collection period to improve visualization. b. AUC comparisons of future food allergy development discrimination using PLS-DA. Models were trained with repeated (r = 20)
10-fold stratified (host and food allergy development) cross-validation using either the SelEnergyPerm signature or all log ratios. Points represent the mean AUC and error bars indicate the 95% CI. c. Relative taxa strength by family measuring the importance of each taxon for discriminating between food allergy statuses later in life across each collection month. Relative strength was computed using the top 5 nodes derived from the PLS-DA variable importance weighted log ratio networks across each collection month. d. Directed (edges point out from numerator to denominator) network of the SelEnergyPerm-derived signature by collection period and food allergy development weighted by the absolute log ratio means (nodes = taxa, node size = mean strength, edge = log ratio, edge width = log ratio mean, red edges = negative log ratio mean (incoming node more abundant), blue edges = positive log ratio mean (outgoing node more abundant)).

**Supplementary Files**

This is a list of supplementary files associated with this preprint. Click to download.

- AdditionalFile1.tif
- AdditionalFile2.tif
- AdditionalFile3.tif
- AdditionalFile4.tif