Integrative analysis of microbial 16S gene and shotgun metagenomic sequencing data improves statistical efficiency

Ye Yue
Department of Biostatistics and Bioinformatics, Emory University

Timothy Read
Department of Medicine, Division of Infectious Diseases, Emory University School of Medicine

Veronika Fedirko
Department of Epidemiology, University of Texas MD Anderson Cancer Center

Glen Satten
Department of Gynecology and Obstetrics, Emory University School of Medicine

Yi-Juan Hu (✉ yijuan.hu@emory.edu)
Department of Biostatistics and Bioinformatics, Emory University

Method Article

Keywords: microbiome, experimental bias, LOCOM, differential abundance, partial overlap, global test

Posted Date: October 3rd, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3376801/v1

Additional Declarations: No competing interests reported.

Integrative analysis of microbial 16S gene and shotgun metagenomic sequencing data improves statistical efficiency

Ye Yue[1], **Timothy D. Read**[2], **Veronika Fedirko**[3,4], **Glen A. Satten**[5],
and Yi-Juan Hu[1*]

[1]Department of Biostatistics and Bioinformatics, Emory University, Atlanta, GA, 30322, USA

[2]Department of Medicine, Division of Infectious Diseases, Emory University School of Medicine, Atlanta, GA, 30322, USA

[3]Department of Epidemiology, University of Texas MD Anderson Cancer Center, Houston, TX, 77030, USA

[4]Department of Epidemiology, Emory University, Atlanta, GA, 30322, USA

[5]Department of Gynecology and Obstetrics, Emory University School of Medicine, Atlanta, GA, 30322, USA

*Corresponding author email: yijuan.hu@emory.edu

Keywords: microbiome, experimental bias, LOCOM, differential abundance, partial overlap, global test

Abstract

Background: The most widely used technologies for profiling microbial communities are 16S marker-gene sequencing and shotgun metagenomic sequencing. Interestingly, many microbiome studies have performed both sequencing experiments on the same cohort of samples. The two sequencing datasets often reveal consistent patterns of microbial signatures, highlighting the potential for an integrative analysis to improve power of testing these signatures. However, differential experimental biases, partially overlapping samples, and differential library sizes pose tremendous challenges when combining the two datasets. Currently, researchers either discard one dataset entirely or use different datasets for different objectives.

Methods: In this article, we introduce the first method of this kind, named Com-2seq, that combines the two sequencing datasets for testing differential abundance at the genus and community levels while overcoming these difficulties. The new method is based on our LOCOM model (Hu et al., 2022), which employs logistic regression for testing taxon differential abundance while remaining robust to experimental bias. To benchmark the performance of Com-2seq, we introduce two *ad hoc* approaches: applying LOCOM to pooled taxa count data and combining LOCOM p-values from analyzing each dataset separately.

Results: Our simulation studies indicate that Com-2seq substantially improves statistical efficiency over analysis of either dataset alone and works better than the two *ad hoc* approaches. An application of Com-2seq to two real microbiome studies uncovered scientifically plausible findings that would have been missed by analyzing individual datasets.

Conclusions: Com-2seq performs integrative analysis of 16S and metagenomic sequencing data, which improves statistical efficiency and has the potential to accelerate the search of microbial communities and taxa that are involved in human health and diseases.

Background

Thanks to technological advances in high-throughput sequencing, microbiome research has proliferated in the past decade and revealed relationships between human microbiome and many diseases and conditions such as inflammatory bowl diseases [1], obesity and type II diabetes [2], and even cancers [3]. The most widely used technologies for profiling microbial communities are 16S marker-gene sequencing (16S) and shotgun metagenomic sequencing (SMS) [4]. The 16S method employs primers that target a highly variable region of the 16S ribosomal RNA gene, which is then PCR amplified and sequenced. This approach is well-tested, fast and cost effective, and provides a low-resolution view of a microbial community, typically at the genus level. On the other hand, SMS extracts all microbial genomes within a sample, which are then fragmented and sequenced. This technique offers detailed genomic information, including higher taxonomic resolution and additional functional capability, but it is 10 to 30 times more expensive to perform sequencing experiment and much more challenging to conduct bioinformatics analysis.

Both the 16S and SMS methods are subject to *experimental bias* at every step of the experiment (i.e., DNA extraction, PCR amplification, amplicon or metagenomic sequencing, and bioinformatics processing), as each step favors certain taxa over others [5]. The bias systematically distorts the measured taxon abundances from their actual values. Moreover, the bias differs significantly between 16S and SMS, as they comprise different steps (e.g., PCR amplification vs. no PCR) and even for the same step they may adopt different protocols [6]. As a result, each sequencing method leads to some taxa being underrepresented or even entirely missed [7]. As these taxa may be complementary, combining data from the two sequencing experiments could enhance the profiling of complex microbial communities.

Interestingly, many microbiome studies have performed both 16S and SMS on the same cohort of samples. In Qiita, which is currently the largest open-source microbiome study management platform (https://qiita.ucsd.edu) [8], 26 out of all 698 studies (as of June 2023) have both datasets. In fact, more studies may have both datasets, as some studies may only deposit the dataset used in the final publication. There are at least two scenarios in which this occurs. In one scenario, a study initially performs 16S and then follows up with SMS to investigate higher-resolution taxa or biological functions. In another scenario, a study initially performs SMS but adds 16S due to its low cost. The 16S data are routinely processed into a taxa count table by the popular analysis platform QIIME2 [9]. Although there is less consensus, the SMS data can also be classified into taxonomies using tools such as MetaPhlAn [10–12], Kraken [13, 14], or more recently Woltka [15]. For each of the 26 studies found in Qiita, we summarized the two taxa count tables, as well as their overlapping samples and taxa at the genus level, in Table S1. For most studies, both the overlap of samples and the overlap of genera are incomplete yet substantial, which data structure is schematically depicted in Figure 1 (left). The library sizes (i.e., depths) in the two tables may differ by orders of magnitude, ranging from 1.4 to 1500 fold. Despite their differences, the 16S and SMS taxonomic profiles for the same cohort of samples have frequently been found to yield consistent patterns of microbial signatures [7, 16–22]. This further underscores the need for an integrative analysis of the two taxonomic profiles to improve power of testing these signatures.

Currently, researchers either discard one dataset entirely or use different datasets for different objectives. To the best of our knowledge, no statistical method is available for performing integrative analysis of the two sequencing datasets. Therefore, we propose a method, named Com-2seq, to fill this gap. Our method is based on same model that underlies the LOCOM package we developed recently [23], which employs logistic regression for testing taxon differential abundance while remaining robust to experimental bias. LOCOM is a compositionally aware method that does not require pseudocounts. Our new method combines data from 16S and SMS for testing differential abundance at the genus and community levels, while overcoming the challenges including differential experimental biases, partially overlapping samples, and differential library sizes. We compare our new method to two *ad hoc* procedures: one first pools the taxon counts of the two datasets and then applies LOCOM to the pooled data, referred to as Com-count; the other applies LOCOM to the two datasets separately and then combines the two p-values at each taxon using the Cauchy p-value combination method, referred to as Com-p; see Methods for more details.

Results

Differences in experimental bias. We compared the observed relative abundances between 16S and SMS for each of the overlapping samples at each of the overlapping genera, using the ORIGINS and Dietary study data from Qiita; descriptions about the two studies are provided below. If the biases in 16S and SMS data were the same, we would expect data from each genus to fit the 45° line. The departure from this behavior seen in Figure 2 for the ORIGINS study demonstrates that there are systematic differences in observed relative abundances at many genera, even for the second most abundant genus *Haemophilus*. Further, some genera appear to be mostly absent in one dataset but reasonably abundant in the other, such as *Gemella* from 16S and *Geobacillus* from SMS. These findings confirm that 16S and SMS data have different experimental biases. Similar patterns, although with higher overdispersion (i.e., deviation of data from the fitted line), are observed in Figure 3 for the Dietary study. Despite the differences, there is clear agreement between 16S and SMS data, particularly at the most abundant genera from the ORIGINS study, motivating the use of an integrative approach that properly accounts for the differences in experimental bias.

Analysis of ORIGINS data. We analyzed data generated from the Oral Infections, Glucose Intolerance, and Insulin Resistance Study (ORIGINS) [24] to investigate the association between periodontal bacteria and prediabetes status (yes or no) among diabetes-free adults, without adjusting for any other risk factors. One subgingival plaque sample was collected from each participant and sequenced by either 16S, SMS, or both. We downloaded the 16S and SMS taxa count data with study ID 11808 from Qiita. The 16S data include 271 samples linked with meta data and having adequate ($>$ 5,000) library sizes, and 234 genera after quality control (QC, which excluded genera found in fewer than 5 samples). The SMS data include 183 samples linked with meta data and having adequate library size, and 756 genera after QC. In total, there are 302 distinct samples (56 cases and 246 controls) and 864 distinct genera, among which 152 samples (47 cases and 105 controls) have both 16S and SMS data for 125 genera. The mean library sizes of the 16S and SMS data are 26,950 and 176,321, respectively, resulting in a depth ratio of 1:6.5.

We began by analyzing data for the overlapping samples and genera, with a main purpose of method comparison. We applied Com-2seq, Com-count, and Com-p for integrative analysis of the 16S and SMS data, as well as LOCOM to analyze each dataset separately, and summarized their global test p-values and detected differentially abundant genera (at the nominal FDR level of 10%) in Table 1. Com-2seq and Com-count yielded significant global p-values (at the nominal level of 0.05) and detected 7 and 3 genera, respectively, whereas the analysis of either dataset alone and the integrative analysis by combining p-values (Com-p) produced non-significant global p-values and detected zero or at most one genus. For more information on the detected genera, see Table S2 and Figure S1.

We proceeded to analyze the full data for scientific discovery and also summarized the main results in Table 1. In this case, Com-count is invalid (as demonstrated in simulation results below) and was thus not applied. Once again, Com-2seq yielded a significant global p-value, while the p-values of the remaining methods failed to reach the significance level. Com-2seq detected three genera, *Butyrivibrio*, *Gemella*, and *Ignavigranum*, whereas the analysis of the 16S data alone detected a different genus *Chelonobacter* and the analysis of the SMS data alone and the integrative analysis by Com-p failed to detect any genus. More information on the detected genera can be found in Figure 4(a) and Table S3. Specifically, *Butyrivibrio* was captured by both sequencing techniques and assigned non-significant yet small p-values in both analyses of individual datasets. As its abundance appeared consistently lower in prediabetic participants across both datasets, this trend was deemed significant overall by Com-2seq. Indeed, *Butyrivibrio* is known to produce short-chain fatty acids such as butyrate, which has been found to improve insulin sensitivity in mice [25]. *Gemella* was only identified in a few samples by 16S sequencing and thus excluded from the analysis of 16S data alone by the LOCOM filter. On the other hand, it was efficiently captured in all samples by SMS (∼1.5% relative abundance) and revealed to be notably more abundant in prediabetic participants, resulting in its detection by Com-2seq (which analysis salvaged the 16S data on this genus that were discarded in the analysis of 16S data alone). This finding is also plausible, as *Gemella* has been linked to various infections, including those affecting heart valves [26], brain membranes [27], and bloodstreams [28]. *Ignavigranum* was completely missed by 16S sequencing and its detection by Com-2seq was solely driven by its differential abundance revealed by SMS. *Chelonobacter* exhibited significant differential abundance using the 16S data, but this signal was not replicated using the SMS data, leading to a non-significant overall result by Com-2seq.

Analysis of Dietary data. We also analyzed data from Amato et al. [29] to test the influence of host dietary niche (folivore vs. non-folivore) on the gut microbiome of wild non-human primates, while controlling for host phylogeny (categorized as apes, lemurs, new world monkeys, and old world monkeys). One fecal sample was collected for each animal and sequenced by either 16S, SMS, or both. We downloaded the 16S and SMS taxa count data with study ID 11212 from Qiita, the statistics of which are listed in Table S1. In total, there are 172 distinct samples (94 folivore and 78 non-folivore) and 2062 distinct genera, among which 76 samples (40 folivore and 36 non-folivore) have both 16S and SMS data for 236 genera. The ratio of 16S to SMS mean library sizes is 1:9.8.

Table 1 (lower panel) shows that all analysis methods yielded highly significant global p-values and detected a large number of differentially abundant genera at the nominal FDR level of 10%, in both the analysis of data for the overlapping samples and genera and the analysis of the full data. In both of these analyses, Com-2seq detected the most genera, far exceeding the number of detections by analyzing the 16S or SMS data alone or by employing Com-p. Figure 4(b) displays the Venn diagram of the detected genera by different methods when analyzing the full data. Com-2seq detected 185 novel genera that were missed by analyzing the 16S and SMS data separately, whereas Com-p detected only 8 such genera.

Simulation results. To develop confidence in our new method, we used simulated data to confirm the results we observed in the analysis of the ORIGINS and Dietary data; the details of the simulation studies are presented in Methods. In brief, we considered three causal mechanisms M1, M2, and M3, which assume the causal (i.e., trait-associated) taxa to be moderately abundant, very abundant, and rare, respectively. We varied the overdispersion parameter τ, which controls the deviation of 16S and SMS relative abundances from the fitted line (e.g., Figures 2 and 3), between 0.01 and 0.001, and we also varied the ratio of 16S to SMS mean library sizes between 1:1 and 1:10.

The results for the complete-overlap case, in which all samples have both 16S and SMS data, based on each of the four combinations of the overdispersion and depth ratio values are shown in Figures 5 and S2–S4, respectively. All global tests controlled the type I error at the nominal level and all taxon-level tests controlled the FDR at the nominal level across all scenarios. All integrative analyses significantly improved statistical efficiency over the analyses of individual datasets, at both the taxon and global levels. The increase in sensitivity of detecting causal taxa is substantial, as one sequencing experiment failed to capture certain causal taxa and the combination of two experiments provided better coverage. The boost in global power is less pronounced, because the Harmonic Mean and Cauchy statistics underlying the global tests were dominated by a few smallest individual p-values and less influenced by the total number of causal taxa. Among the three integrative approaches, Com-2seq always yielded the highest or nearly highest power and sensitivity.

Similar patterns of results are observed in the partial-overlap case (Figure 6), in which part of samples have both 16S and SMS data and the remaining samples have data from either 16S or SMS. One difference from the complete-overlap case is that, Com-count failed to control the type I error, as its zero-filling strategy tended to create spurious associations. Another difference pertains to the permutation scheme in Com-2seq, which was based on three strata of samples in the partial-overlap case, as illustrated in Figure 1 (right). It is the only one that led to correct type I error, compared to two alternative schemes: one pools samples having data from either of the two experiments into one stratum (a total of two strata), and the other pools all samples together (one stratum). Finally, the same patterns of results persisted when a confounder was simulated (Figure S5). We confirmed that the confounding effect was substantial, leading to highly inflated type I error if it were not controlled for. All methods yielded proper type I error after adjusting for the confounder.

Discussion

We have presented a new method, Com-2seq, that is the first method for integrative analysis of 16S and SMS data. Com-2seq is specifically designed to combine these datasets in a way that properly accounts for their differences in experimental bias, sequenced samples, and library sizes. We have demonstrated that Com-2seq substantially improves statistical efficiency over analysis of either of the two datasets and works better than the taxa-count-pooling and *p*-value-combination approaches. The inference of Com-2seq is based on a permutation procedure that preserves any sample structure and thus valid for partially overlapping samples as well as small sample sizes. Com-2seq allows testing of a trait that is binary, continuous, or multivariate (e.g., a categorical trait with more than two levels or a set of multiple measurements), and supports adjustment of confounding covariates.

We have implemented Com-2seq as a new function in our R package LOCOM, which is available on GitHub at https://github.com/yijuanhu/LOCOM. Com-2seq performs best when given count data. However, some bioinformatics programs for processing shotgun metagenomic data, such as Kraken, output relative abundance data only. In this case, Com-2seq can still be used, although the performance may be slightly worse than if count data were available. Com-2seq automatically adjusts to handle only relative abundance data.

Methods

Com-2seq. Let $Y_{ik,j}$ be the read count of taxon j ($j = 1, \ldots, J$) in sample i ($i = 1, \ldots, n$) obtained from data source k ($k = 1, 2$, 16S or SMS), and let Z_i be a vector of covariates excluding the intercept. To fit the LOCOM model simultaneously to 16S and SMS data, we choose one taxon (without loss of generality, taxon J) that has the largest mean relative abundance across both data sources to be the reference taxon (which does not need to be a null taxon), and compare each of the remaining taxa to the reference taxon using logistic regressions. Because the most abundant taxa can usually be effectively captured by both 16S and SMS, our selection criterion results in a reference taxon that is among the top abundant taxa in each data source. We define $p_{ik,j}$ to be the expected value of the relative abundance in data source k and relate this relative abundance to the true relative abundance $\pi_{i,j}$, which is unrelated to any data source, by $\log(p_{ik,j}) = \mathbb{I}(k = 1)\gamma_{1,j} + \mathbb{I}(k = 2)\gamma_{2,j} + \log(\pi_{i,j}) + \alpha_{ik}$, where $\gamma_{1,j}$ and γ_{2j} are source-specific bias factors and $\alpha_{i,k}$ is the normalization factor that ensures $\sum_{j=1}^{J} p_{ik,j} = 1$. Following [30], we assume $\log(\pi_{i,j}) = \log(\pi_j^0) + Z_i^{\mathrm{T}}\beta_j$, where β_j contains the effect sizes and π_j^0 contains the baseline relative abundances. When comparing taxon j to the reference taxon J, we define $\mu_{ik,j} = p_{ik,j}/(p_{ik,j} + p_{ik,J})$ to obtain a standard logistic regression

$$\log\{\mu_{ik,j}/(1 - \mu_{ik,j})\} = \eta_{k,j} + Z_i^{\mathrm{T}}\beta_j$$

for each j ($j = 1, 2, \ldots, J - 1$). The intercept $\eta_{k,j}$ includes the taxon- and source-specific bias information, and for this reason is considered as a free parameter for each j, k.

To combine data from both sources for inference on the β_js, we solve the estimating equation (EE)

$$\mathbb{U}_j(\eta_{1,j}, \eta_{2,j}, \beta_j) = \sum_{i=1}^{n}\sum_{k=1}^{2} \omega_{ik,j} \left(\frac{Y_{ik,j}}{Y_{ik,j} + Y_{ik,J}} - \mu_{ik,j} \right) \begin{bmatrix} \mathbb{I}(k = 1) \\ \mathbb{I}(k = 2) \\ Z_i \end{bmatrix} = 0 \qquad (1)$$

or the bias-corrected estimating equation (derived in Supplementary Materials) when taxon j has sparse count data, where $\omega_{ik,j}$ is a weight for the measurement of sample i from data source k. The EE of LOCOM is a special case of Equation (1) when $\omega_{ik,j}$ is set to $Y_{ik,j} + Y_{ik,J}$ for one data source and $\omega_{ik,j} = 0$ for the other.

When $\omega_{ik,j} = Y_{ik,j} + Y_{ik,J}$, the summation in Equation (1) is at the "count" level, and measurements with small $Y_{ik,j}$ and $Y_{ik,J}$ tend to contribute less. These weights are optimal when there is minimal overdispersion in the count data (as the equation becomes a score equation). Additionally, these weights are reasonable for a taxon when the count data from one data source are very sparse (i.e., mostly zeros) while the data from the other source are non-sparse. However, as taxa count data are frequently overdispersed, the amount of information quickly reaches a plateau as long as there is moderate coverage of reads. In this case, it is sensible to use $\omega_{ik,j} = 1$, which puts the summation in Equation (1) at the "relative abundance" level and treats data from the two data sources equally. These weights are particularly important when the library sizes of SMS are orders of magnitude (e.g., 10 times) higher than that of 16S (Table S1), to prevent the SMS data from dominating the results. Since it is unknown a priori as for which weighting scheme works better for which taxa, Com-2seq fits (1) using both weights and then forms an omnibus test that combines both results. When a taxon has only been observed in one data source, Com-2seq only uses weights $\omega_{ik,j} = Y_{ik,j} + Y_{ik,J}$ and Equation (1) reduces to the LOCOM EE (Figure 1, right). When only relative abundance data are available, Com-2seq only uses weights $\omega_{ik,j} = 1$.

Because there is no a priori knowledge about whether the reference taxon is null or causal (i.e., associated with the trait), Com-2seq follows LOCOM to reduce the dependence on the choice of reference taxon by testing each β_j against their median, which represents a null reference when assuming that more than half of the taxa are null taxa. Like LOCOM, Com-2seq bases inference on permutation using Potter's method [31]. Care should be taken to preserve the sample structure when generating permutation replicates. The permutation procedure should be stratified to restrict the shuffling of trait residuals in one stratum of samples having data from both sources, one stratum of samples having data from 16S only, and one stratum of samples having data from SMS only (Figure 1, right). Finally, for testing the community-level (i.e., global) hypothesis that there are no differentially abundant taxa in the community, the Harmonic Mean (HM) of the taxon-specific p-values is used as the test statistic and assessed for significance using the permutation replicates generated above.

Filtering out rare taxa in Com-2seq. We develop a filter to exclude very rare taxa that can jeopardize the validity of Com-2seq. Recall that the filter in LOCOM removes rare taxa present in fewer than 20% of samples. In Com-2seq, a sample with a non-zero count from either data source contributes to the 3rd equation of Equation (1), making the taxon analyzable if more than 20% of such samples exist. In addition, we keep the taxon as long as it passes the LOCOM filter applied to either 16S or SMS data, since the LOCOM filter works well for the "worse" case when the data from one source consist entirely of zeros.

Alternative approaches Com-count and Com-p. Com-count first creates an "augmented" table consisting of the union of samples and the union of taxa across the two data sources, then pools (i.e., sums up) the read counts for cells that have both 16S and SMS data, copies the read counts for cells that have one data source, and fills in zeros for cells that have no data source. Then, LOCOM is applied to the augmented table. The zero-filling strategy may create spurious associations, when samples filled with zeros have a different trait distribution from the others. Moreover, the results of Com-count tend to be dominated by data from the source that has much larger library sizes than the other source. Lastly, the LOCOM filter applied to the "augmented" table is more stringent than the Com-2seq filter above.

Com-p applies LOCOM separately to the two taxa count tables, and then combines the two p-values at each overlapping taxon into one p-value using the Cauchy p-value combination method [32]. The resulting p-values, along with p-values from LOCOM for non-overlapping taxa, are used to detect differentially abundant taxa after multiple test correction by the Benjamini-Hochberg method [33] and also combined into one p-value for testing the global hypothesis (also by the Cauchy method). At each taxon, the two p-values are expected to be correlated since they are based on at least some overlapping samples; at the global level, the p-values across interacting taxa may also be correlated. This motivated us to adopt the Cauchy [32] or HM [34] combination method, which accounts for such correlations; we chose Cauchy over HM because HM led to inflated error rate in our simulations (results not shown). Note that the HM method here produces p-values based on asymptotic theories, whereas Com-2seq assesses the significance of the HM statistics via permutation. Com-p is inherently unable to produce an overall p-value that is more significant than the most significant individual p-value, while Com-2seq can achieve this by combining the two datasets at the more appropriate taxon-count and relative-abundance levels. Additionally, Com-p combines p-values without considering directions of association in the two data sources, while Com-2seq is able to strengthen the signals if they are consistent across the two data sources and disregard the ones that are contradictory. Furthermore, taxa that fail the LOCOM filter applied to each taxa count table are completely missed by Com-p, but they still have a good chance of passing the Com-2seq filter (which is essentially based on the pooled data), especially when the sample overlap is substantial.

Simulation settings. Our simulations are based on data on 856 taxa of the upper-respiratory-tract (URT) microbiome by Charlson et al. [35]. We fixed the sample size to 100 unless otherwise specified and considered a binary trait T_i throughout the simulations. In some cases, we also simulated a continuous confounder C_i by drawing values from $U[-1, 1]$ for samples with $T_i = 0$ and from $U[0, 2]$ for those with $T_i = 1$. We used the two sets of causal taxa employed in [23], M1 and M2, which are a random sample of 20 taxa with mean relative abundances greater than 0.005 and the five most abundant taxa, respectively. In addition, we considered a set of rare causal taxa by randomly sampling 50 taxa with mean relative abundances between 0.0005 and 0.001; we refer to this set as M3. When a confounder was present, we randomly sampled 5 taxa with mean relative abundances greater than 0.005 to be associated with the confounder.

We assumed that the 856 taxa form the complete set of underlying taxa in the community and generated bias factors $\gamma_{1,j}$ and $\gamma_{2,j}$ for taxon j from 16S and SMS, respectively. We set $\gamma_{1,j}$ and $\gamma_{2,j}$ to a very small value of -5 to create missingness for specific taxa in each data source. Specifically, in M1 (M2 and M3), we selected two sets of five (two and five) non-overlapping causal taxa to be missing in 16S and SMS, respectively. Additionally, we sampled 20% of non-causal taxa to be missing in each data source. We set $\gamma_{1,j}$ and $\gamma_{2,j}$ for the most abundant taxon to 1, reflecting the efficiency of both sequencing methods for capturing this taxon. For all other taxa, we independently drew $\gamma_{1,j}$ and $\gamma_{2,j}$ from $N(0, 0.5^2)$.

We then simulated read count data for the 856 taxa in the two data sources, taking into account the effects of the trait and confounder as well as the influences of bias factors. First, we drew the baseline relative abundances $(\pi_{i,1}^{(0)}, \pi_{i,2}^{(0)}, \ldots, \pi_{i,J}^{(0)})$ of all taxa for each sample from $Dirichlet(\bar{\pi}, \theta)$, where $\bar{\pi}$ contains the mean relative abundances and θ is the overdispersion parameter estimated from fitting the Dirichlet-Multinomial (DM) model to the URT data. The overdispersion parameter θ controls the sample heterogeneity in baseline relative abundances, not including the variability in read count data. Thus, we set θ to 0.01, which is half of the overall overdispersion (0.02) estimated from the URT data. Then, we formed the expected value of the observed relative abundances obtained from the kth data source, $p_{ik,j}$, by spiking in the causal taxa and confounder-associated taxa, then imposing bias factors on all taxa, and finally normalizing the relative abundances to have a sum of 1, resulting in the following equation:

$$p_{ik,j} = \frac{\exp\left[\mathbb{I}(k=1)\gamma_{1,j} + \mathbb{I}(k=2)\gamma_{2,j} + \beta_{1,j}T_i + \beta_{2,j}C_i\right]\pi_{ij}^{(0)}}{\sum_{j'=1}^{J}\exp\left[\mathbb{I}(k=1)\gamma_{1,j'} + \mathbb{I}(k=2)\gamma_{2,j'} + \beta_{1,j'}T_i + \beta_{2,j'}C_i\right]\pi_{ij'}^{(0)}}. \tag{2}$$

Here, $\beta_{1,j} = 0$ for null taxa, and $\beta_{2,j} = 0$ for confounder-independent taxa. For simplicity, we set $\beta_{1,j} = \beta$ for all causal taxa, which is referred to as the *effect size*, and we fixed $\beta_{2,j} = \log(1.5)$ for all confounder-associated taxa. Subsequently, we generated the read count data for sample i in data source k from the DM distribution with mean $(p_{ik,1}, p_{ik,2}, \ldots, p_{ik,J})$, overdispersion parameter τ_k, and library size drawn from $N(\nu_1, (\nu_1/3)^2)$ and $N(\nu_2, (\nu_2/3)^2)$ for

16S and SMS, respectively, with left truncation at 2,000. We fixed $\nu_1 = 10,000$ and varied ν_2 to achieve the depth ratio of ν_1:ν_2 = 1:1 or 1:10. The overdispersion parameter τ_k controls the deviation of the observed relative abundances from their expected values $p_{ik,j}$ in the kth data source. We set $\tau_1 = \tau_2 = \tau$ without loss of generality and varied τ between 0.01 and 0.001, which correspond to the large and small deviation as seen in the Dietary and ORIGINS data, respectively.

In the complete-overlap case, we generated taxa count data from both sources for all 100 samples. In the partial-overlap case, we collected data from both sources for 40 samples (15 cases and 25 controls), from 16S only for 40 samples (30 cases and 10 controls), and from SMS only for 20 samples (5 cases and 15 controls), which resulted in a total of 100 samples, 50 cases and 50 controls, and varying case-control ratios across the three strata of samples.

We applied Com-2seq, Com-count, and Com-p to perform integrative analysis of 16S and SMS data, using the LOCOM results from analyzing individual datasets as benchmarks. We evaluated the sensitivity and empirical FDR of each method for testing individual taxa at the nominal FDR level of 20%, and the type I error and power of each method for testing the global association at the nominal level of 0.05. The type I error results were based on 10,000 replicates of simulated data, while all other results were derived from 1,000 replicates.

To confirm that the simulated data captured the important features of the ORIGINS data shown in Figure 2 and the Dietary data shown in Figure 3, we present scatter plots for the simulated data based on the sample size ($n = 152$) of the ORIGINS study and $\tau = 0.001$ in Figure S6, and scatter plots for the simulated data based on the sample size ($n = 76$) of the Dietary study and $\tau = 0.01$ in Figure S7. We observe that the range of observed relative abundances at each taxon as governed by the overdispersion parameter θ, the deviation of the fitted line from the 45° reference line as determined by the bias factors $\gamma_{1,j}$ and $\gamma_{2,j}$, and the deviation of individual data points from the fitted line as controlled by the overdispersion parameter τ, all resemble those of the real data.

References

1. Gevers D, Kugathasan S, Denson LA, Vázquez-Baeza Y, Van Treuren W, Ren B, et al. The treatment-naive microbiome in new-onset Crohn's disease. Cell host & microbe. 2014;15(3):382–392.
2. Hartstra AV, Bouter KE, Bäckhed F, Nieuwdorp M. Insights into the role of the microbiome in obesity and type 2 diabetes. Diabetes care. 2015;38(1):159–165.
3. Marchesi JR, Dutilh BE, Hall N, Peters WH, Roelofs R, Boleij A, et al. Towards the human colorectal cancer microbiome. PloS one. 2011;6(5):e20447.
4. Knight R, Vrbanac A, Taylor BC, Aksenov A, Callewaert C, Debelius J, et al. Best practices for analysing microbiomes. Nature Reviews Microbiology. 2018;16(7):410–422.
5. McLaren MR, Willis AD, Callahan BJ. Consistent and correctable bias in metagenomic sequencing experiments. Elife. 2019;8:e46923.
6. Nearing JT, Comeau AM, Langille MG. Identifying biases and their potential solutions in human microbiome studies. Microbiome. 2021;9(1):1–22.
7. Peterson D, Bonham KS, Rowland S, Pattanayak CW, Consortium R, Klepac-Ceraj V. Comparative analysis of 16S rRNA gene and metagenome sequencing in pediatric gut microbiomes. Frontiers in microbiology. 2021;12:670336.
8. Gonzalez A, Navas-Molina JA, Kosciolek T, McDonald D, Vázquez-Baeza Y, Ackermann G, et al. Qiita: rapid, web-enabled microbiome meta-analysis. Nature methods. 2018;15(10):796–798.
9. Bokulich NA, Kaehler BD, Rideout JR, Dillon M, Bolyen E, Knight R, et al. Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. Microbiome. 2018;6(1):1–17.
10. Segata N, Waldron L, Ballarini A, Narasimhan V, Jousson O, Huttenhower C. Metagenomic microbial community profiling using unique clade-specific marker genes. Nature

methods. 2012;9(8):811–814.

11. Truong DT, Franzosa EA, Tickle TL, Scholz M, Weingart G, Pasolli E, et al. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. Nature methods. 2015;12(10):902–903.

12. Beghini F, McIver LJ, Blanco-Míguez A, Dubois L, Asnicar F, Maharjan S, et al. Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. elife. 2021;10:e65088.

13. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome biology. 2014;15(3):1–12.

14. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. Genome biology. 2019;20:1–13.

15. Zhu Q, Huang S, Gonzalez A, McGrath I, McDonald D, Haiminen N, et al. Phylogeny-aware analysis of metagenome community ecology based on matched reference genomes while bypassing taxonomy. Msystems. 2022;7(2):e00167–22.

16. Clooney AG, Fouhy F, Sleator RD, O'Driscoll A, Stanton C, Cotter PD, et al. Comparing apples and oranges?: next generation sequencing and its impact on microbiome analysis. PloS one. 2016;11(2):e0148028.

17. Hillmann B, Al-Ghalith GA, Shields-Cutler RR, Zhu Q, Gohl DM, Beckman KB, et al. Evaluating the information content of shallow shotgun metagenomics. Msystems. 2018;3(6):e00069–18.

18. Mas-Lloret J, Obón-Santacana M, Ibáñez-Sanz G, Guinó E, Pato ML, Rodriguez-Moranta F, et al. Gut microbiome diversity detected by high-coverage 16S and shotgun sequencing of paired stool and colon sample. Scientific data. 2020;7(1):92.

19. Durazzi F, Sala C, Castellani G, Manfreda G, Remondini D, De Cesare A. Comparison between 16S rRNA and shotgun sequencing data for the taxonomic characterization of the gut microbiota. Scientific reports. 2021;11(1):1–10.

20. Biegert G, El Alam MB, Karpinets T, Wu X, Sims TT, Yoshida-Court K, et al. Diversity and composition of gut microbiome of cervical cancer patients: Do results of 16S rRNA sequencing and whole genome sequencing approaches align? Journal of microbiological methods. 2021;185:106213.

21. Zuo W, Wang B, Bai X, Luan Y, Fan Y, Michail S, et al. 16S rRNA and metagenomic shotgun sequencing data revealed consistent patterns of gut microbiome signature in pediatric ulcerative colitis. Scientific Reports. 2022;12(1):6421.

22. de Vries J, Saleem F, Li E, Chan AWY, Naphtali J, Naphtali P, et al. Comparative Analysis of Metagenomic (Amplicon and Shotgun) DNA Sequencing to Characterize Microbial Communities in Household On-Site Wastewater Treatment Systems. Water. 2023;15(2):271.

23. Hu Y, Satten GA, Hu YJ. LOCOM: A logistic regression model for testing differential abundance in compositional microbiome data with false discovery rate control. Proceedings of the National Academy of Sciences. 2022;119(30):e2122788119.

24. Demmer R, Jacobs Jr D, Singh R, Zuk A, Rosenbaum M, Papapanou P, et al. Periodontal bacteria and prediabetes prevalence in ORIGINS: the oral infections, glucose intolerance, and insulin resistance study. Journal of dental research. 2015;94(9_suppl):201S–211S.

25. Gao Z, Yin J, Zhang J, Ward RE, Martin RJ, Lefevre M, et al. Butyrate improves insulin sensitivity and increases energy expenditure in mice. Diabetes. 2009;58(7):1509–1517.

26. La Scola B, Raoult D. Molecular identification of Gemella species from three patients with endocarditis. Journal of clinical microbiology. 1998;36(4):866–871.

27. Ruoff KL. Miscellaneous catalase-negative, gram-positive cocci: emerging opportunists. Journal of Clinical Microbiology. 2002;40(4):1129–1133.

28. Woo P, Lau S, Fung A, Chiu S, Yung R, Yuen K. Gemella bacteraemia characterised by 16S ribosomal RNA gene sequencing. Journal of clinical pathology. 2003;56(9):690–693.

29. Amato KR, G Sanders J, Song SJ, Nute M, Metcalf JL, Thompson LR, et al. Evolutionary trends in host physiology outweigh dietary niche in structuring primate gut microbiomes. The ISME journal. 2019;13(3):576–587.

30. Zhao N, Satten GA. A log-linear model for inference on bias in microbiome studies. Statistical Analysis of Microbiome Data. 2021;p. 221–246.

31. Potter DM. A permutation test for inference in logistic regression with small-and moderate-sized data sets. Statistics in medicine. 2005;24(5):693–708.

32. Liu Y, Xie J. Cauchy combination test: a powerful test with analytic p-value calculation under arbitrary dependency structures. Journal of the American Statistical Association. 2020;115(529):393–402.

33. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the royal statistical society Series B (Methodological). 1995;p. 289–300.

34. Wilson DJ. The harmonic mean p-value for combining dependent tests. Proceedings of the National Academy of Sciences. 2019;116(4):1195–1200.

35. Charlson ES, Chen J, Custers-Allen R, Bittinger K, Li H, Sinha R, et al. Disordered microbial communities in the upper respiratory tract of cigarette smokers. PloS one. 2010;5(12):e15216.

36. Firth D. Bias reduction of maximum likelihood estimates. Biometrika. 1993;80(1):27–38.

Funding

This research was supported by the National Institutes of Health award R01GM141074 (Hu, Satten), and the Cancer Prevention and Research Institute of Texas (CPRIT) Rising Stars Award RR200056 (Fedirko).

Acknowledgement

Not applicable.

Availability of data and materials

The R package `LOCOM` and simulation code are available on GitHub at `https://github.com/yijuanhu/LOCOM` The ORIGINS and dietary data can be found in Qiita: ORIGINS https://qiita.ucsd.edu/study/description/11808 and dietary https://qiita.ucsd.edu/study/description/11212.

Authors' contributions

Y.J.H. conceived this work, developed methodologies, conducted numerical studies, and wrote the manuscript; Y.Y. developed methodologies and conducted numerical studies; T.D.R, V.F., and G.A.S. interpreted results and revised the manuscript.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

No consent for publication was required for this study; all microbiome datasets used here are publicly available.

Ethics approval and consent to participate

This study only involved secondary analyses of existing, de-identified datasets; as such it does not require separate IRB consent.

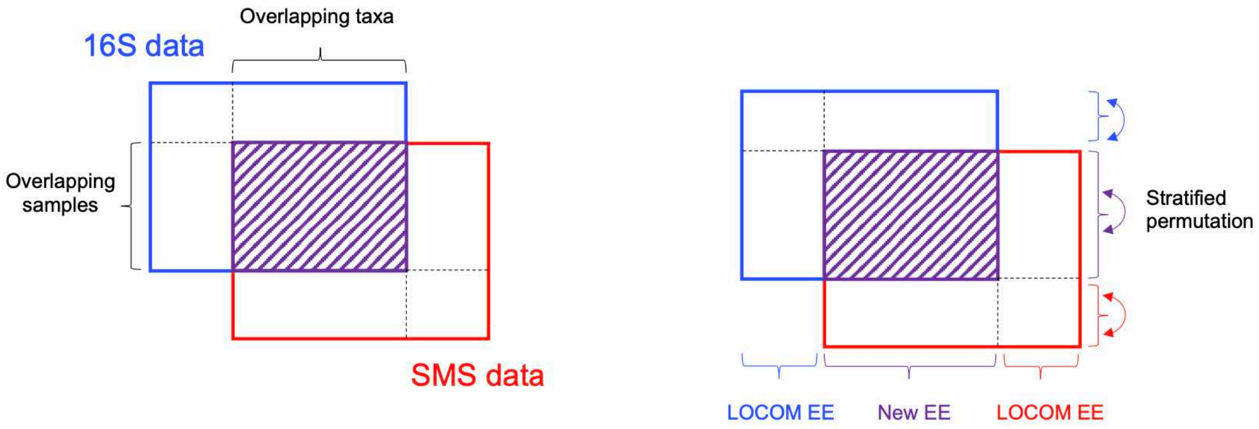

Figure 1: Illustration of the data structure (left) and strategies for analyzing the data (right). The shaded area indicates the data for the overlapping samples and taxa between the 16S and SMS taxa count data. New EE is Equation (1) or bias-corrected Equation (S3) in this paper. LOCOM EE is the estimating equation in the LOCOM paper [23].

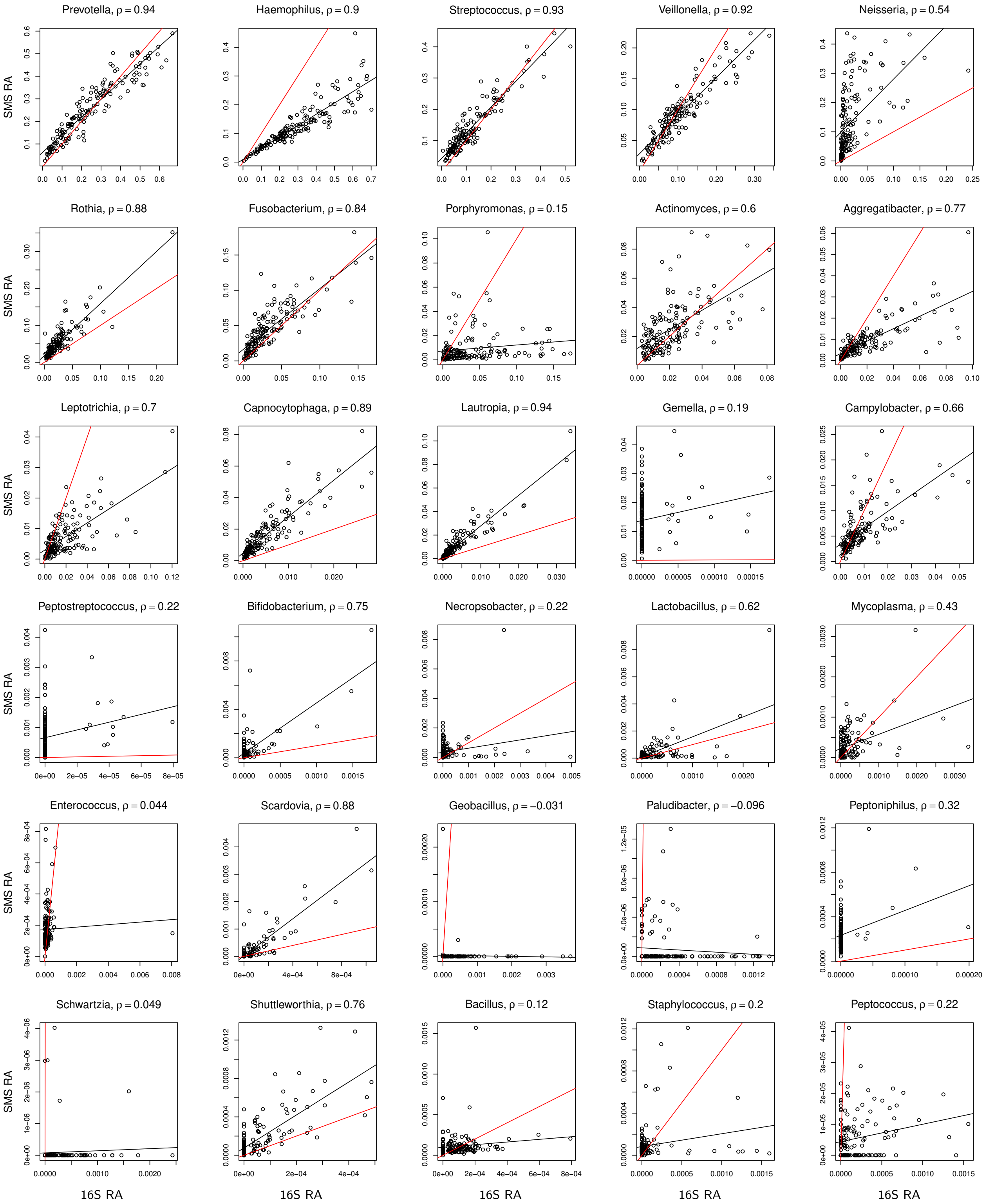

Figure 2: Scatter plot of observed relative abundances (RA) between 16S and SMS for the top 1–15 and 36–50 most abundant genera (ordered by decreasing abundance) in the ORIGINS study. The ρ value is the Pearson correlation coefficient. The red line is the 45° reference line. The black line depicts a fitted linear regression. The observed relative abundances were calculated based on the 125 overlapping genera.

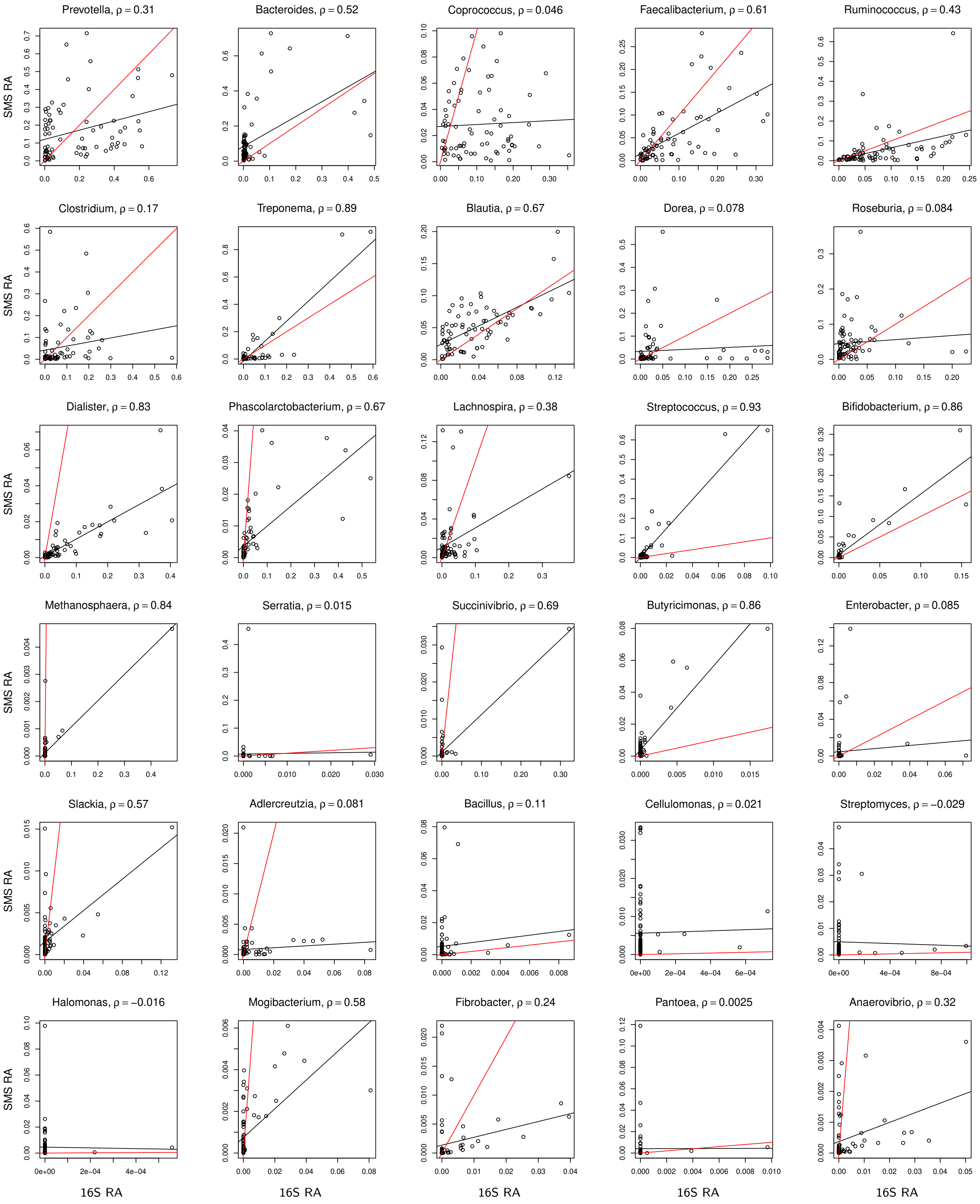

Figure 3: Scatter plot of observed relative abundances between 16S and SMS for the top 1–15 and 36–50 most abundant genera in the Dietary study. See more information in the caption of Figure 2.

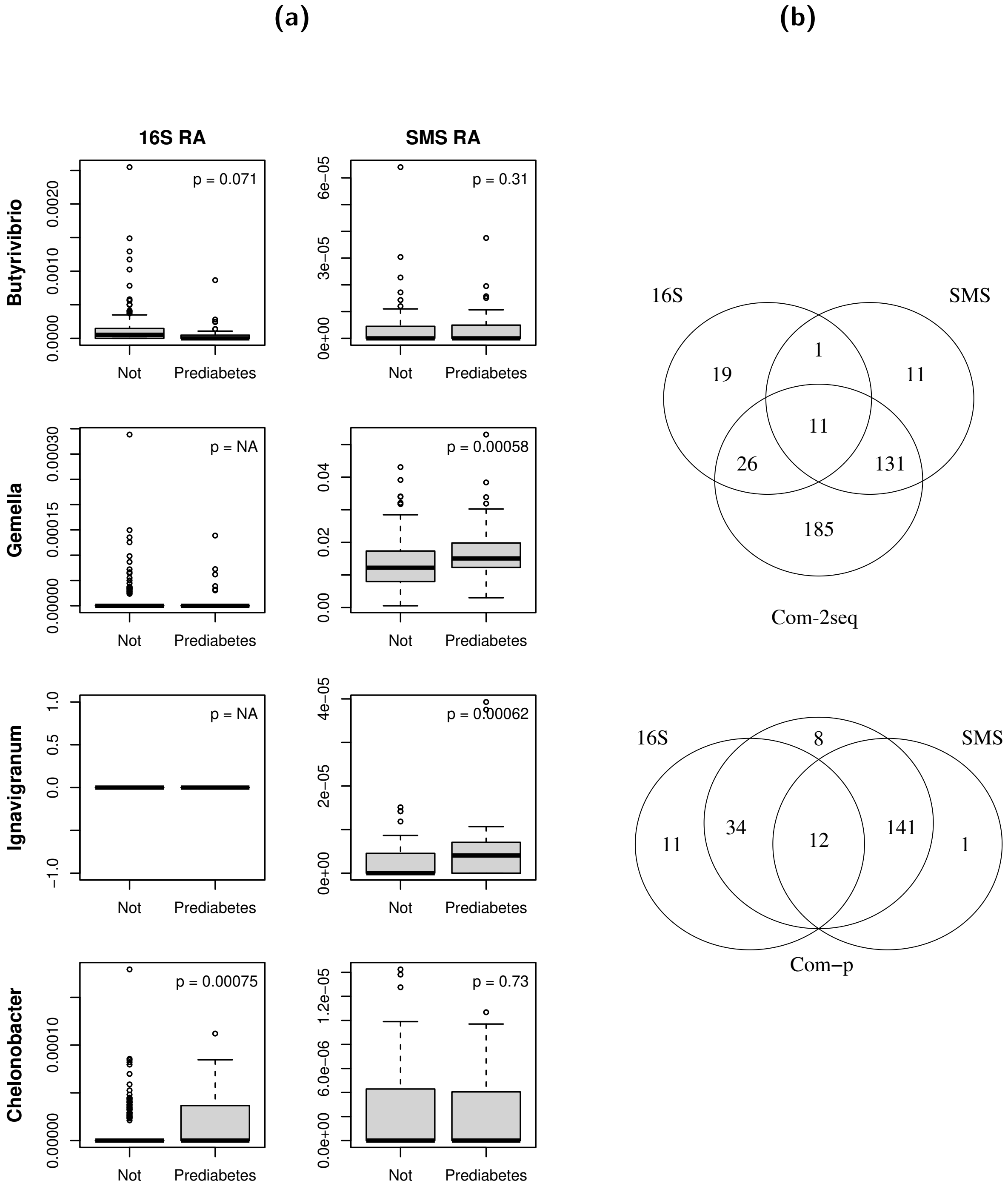

Figure 4: More information on the detected genera in the analysis of the real datasets. (a) Distribution of observed relative abundances (RA) by prediabetes status for the detected genera in the analysis of the full ORIGINS data. The observed relative abundances were calculated based on all genera for all samples in the 16S or SMS data. The *p*-values are from the analysis of individual 16S and SMS datasets, as also listed in Table S3. (b) Venn diagram for the detected genera in the analysis of the full Dietary data.

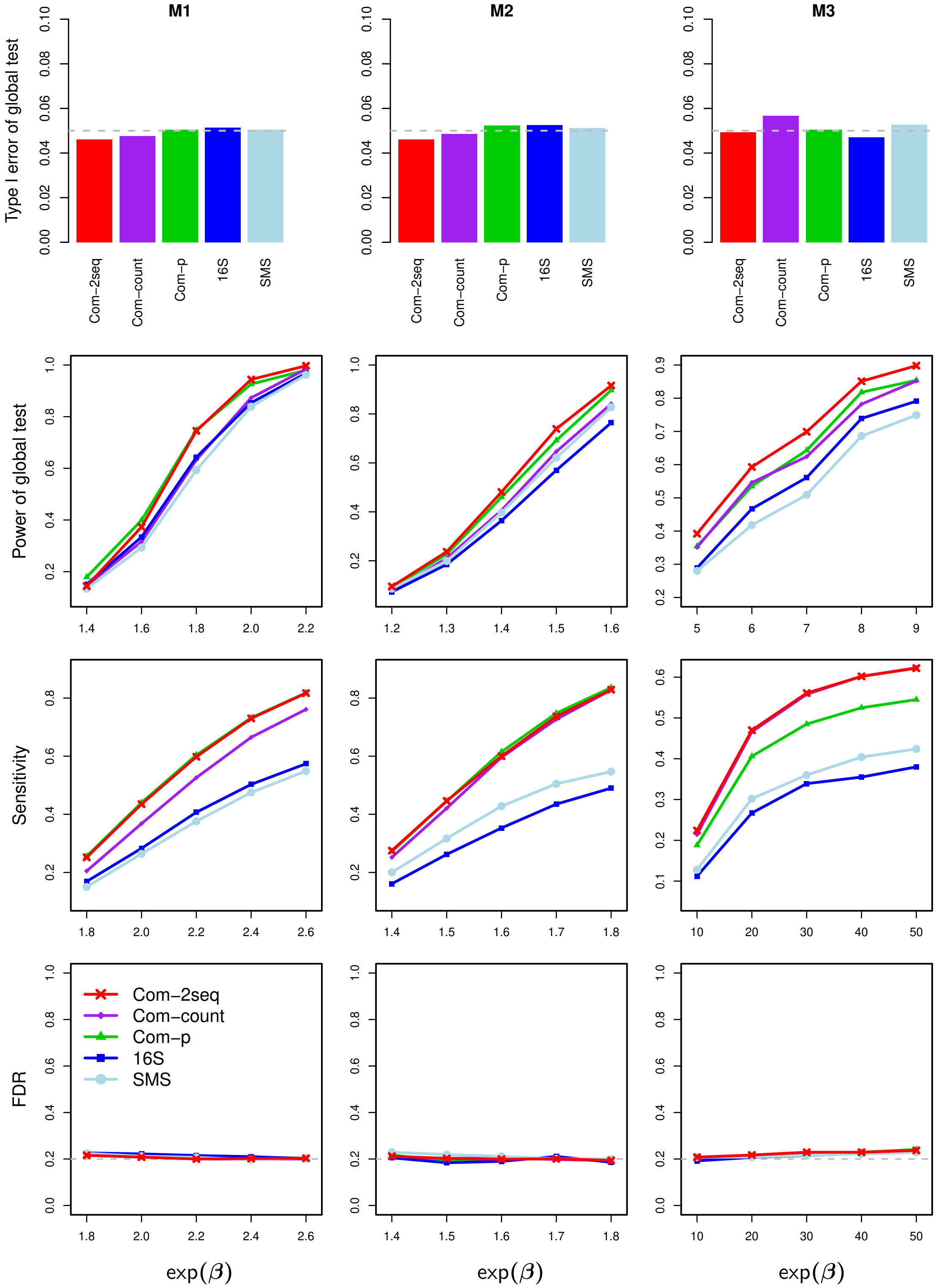

Figure 5: Type I error rate (first row, when $\beta = 0$) and power (second row) for testing the global association at the nominal level of 0.05 (gray dashed line), and sensitivity (third row) and empirical FDR (fourth row) for testing individual taxa at the nominal FDR level of 0.2 (gray dashed line), based on data simulated with completely overlapping samples, overdispersion of $\tau = 0.01$, and depth ratio of 1:10.

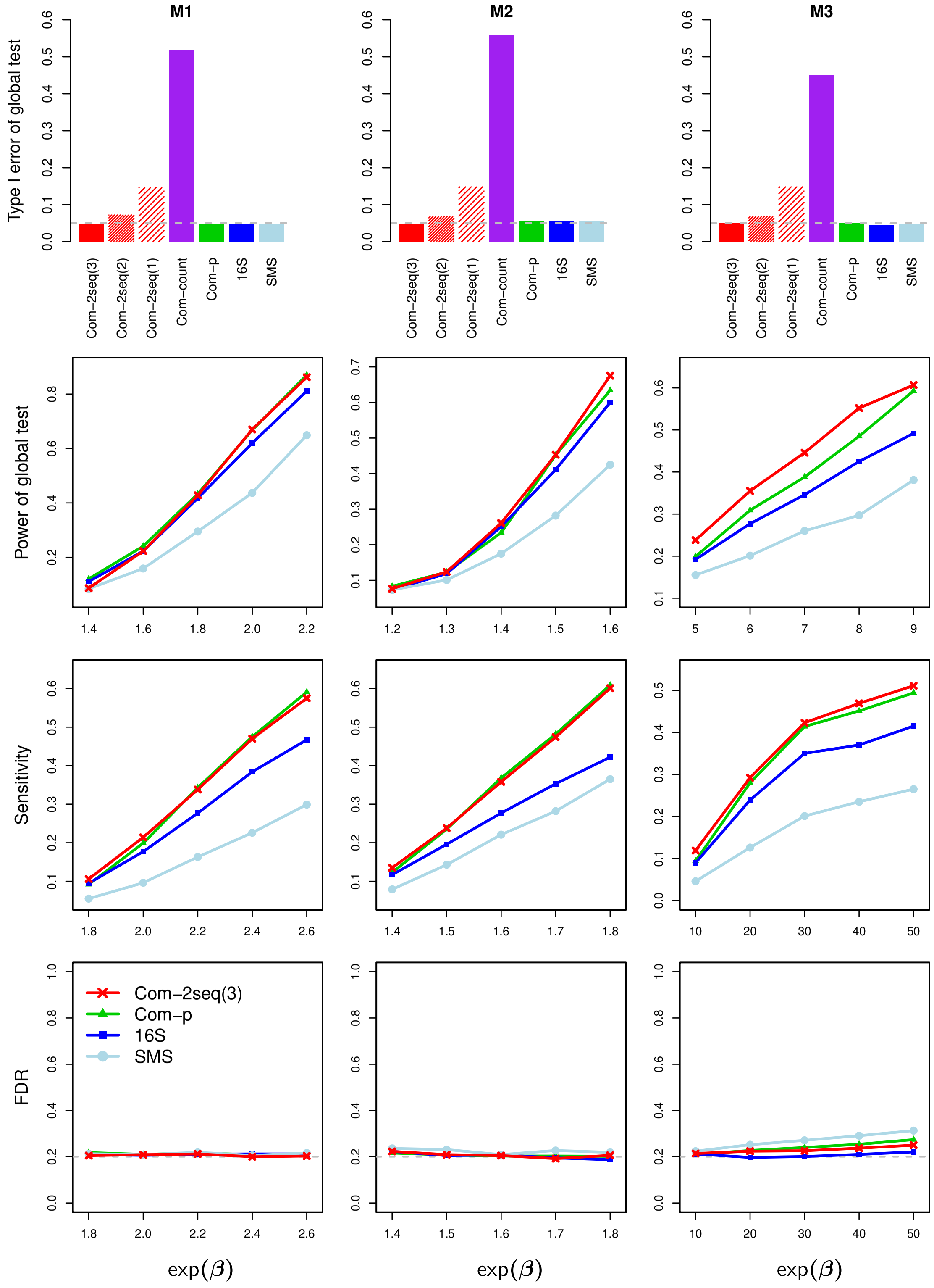

Figure 6: Results for data simulated with partially overlapping samples. Com-2seq annotated with “(3)”, “(2)” and “(1)” used permutation schemes that were based on three strata (the proposed one), two strata, and one stratum, respectively. See more information in the caption of Figure 5.

Table 1: Global test p-value and detected differentially abundant genera in the analysis of the real datasets

Dataset	Method	Global p-value	Detected genera
ORIGIN data for overlapping samples and genera	Com-2seq	0.0345	*Pseudoalteromonas*, *Actinomyces* *Capnocytophaga*, *Lonepinella* *Campylobacter*, *Gemella*, *Kocuria*
	Com-count	0.0226	*Campylobacter*,*Gemella*, *Butyrivibrio*
	Com-p	0.106	None
	16S	0.0689	*Campylobacter*
	SMS	0.170	None
Full ORIGIN data	Com-2seq	0.0340	*Ignavigranum*, *Gemella*, *Butyrivibrio*
	Com-p	0.245	None
	16S	0.0682	*Chelonobacter*
	SMS	0.0866	None
Dietary data for overlapping samples and genera	Com-2seq	0.0001	54
	Com-count	0.0006	27
	Com-p	0.00213	36
	16S	0.0001	24
	SMS	0.0002	23
Full Dietary data	Com-2seq	0.0003	353
	Com-p	0.00196	195
	16S	0.0001	57
	SMS	0.0003	154

Note: Com-count is invalid for analyzing data from partially overlapping samples and was thus not applied to the analysis of the full data. The nominal FDR level is 10%.

Supplementary Files

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

- 2023092216Sshotgunsup.pdf