



Article

The Use of Confidence Intervals in Differential Abundance Analysis of Microbiome Data

Elizaveta Vinogradova ^{1,2,*} , Almagul Kushugulova ² , Samat Kozhakhmetov ² and Maxim Baltin ^{1,*}

¹ Center for Genetics and Life Sciences, Sirius University of Science and Technology, 1 Olympic Ave., Sirius Federal Territory, Sochi 354340, Russia

² Laboratory of Microbiome, Center for Life Sciences, National Laboratory Astana, Nazarbayev University, 53 Kabanbay Batyr Ave., Block S1, Astana Z05H0P9, Kazakhstan; akushugulova@nu.edu.kz (A.K.); skozhakhmetov.nu@gmail.com (S.K.)

* Correspondence: st.paulmississippi@gmail.com (E.V.); baltin.me@talantiuspeh.ru (M.B.)

Abstract

Differential abundance analysis (DAA) is a critical task in microbiome research aimed at identifying microbial signatures that reliably characterize groups. Research suggests that microbiome systems are relatively stable and resilient, yet even small changes under certain conditions can trigger dysbiosis. The high dimensionality of microbiome datasets exacerbates the challenge of detecting such changes by posing a multiple comparison problem that requires hypothesis filtration. Standard filtration using multiple comparison correction procedures is designed for scenarios with a high number of true positives and is often too conservative for microbiome data, where the proportion of true signals can be very low. Therefore, there is a substantial need for hypothesis filtration methods tailored to microbiome data. Confidence intervals (CIs) for between-group differences offer a powerful alternative to *p*-value filtration, as their range simultaneously conveys information about the significance, potential magnitude, and direction of the effect, as well as the certainty of the estimate itself. Microbial data can be adequately modeled using a negative binomial (NB) distribution, and its location parameter can be robustly estimated with the Hodges–Lehmann estimator (HLE). Using synthetic and experimental data, we demonstrate that hypothesis filtration based on CIs for the two-sample HLE is a robust method for comparing microbial data. Our analysis demonstrates that the HLE-CI approach provides the same level of precision as filtration using multiple-adjustment methods while achieving significantly higher recall in microbiome DAA. The results of this study suggest that HLE-CI-based filtration can be an effective step in the search for microbiome biomarkers.

Keywords: negative binomial distribution; microbiome; differential abundance analysis; confidence interval; Hodges–Lehmann estimator



Academic Editor: Kirstin Vach

Received: 10 December 2025

Revised: 27 December 2025

Accepted: 28 December 2025

Published: 2 January 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

1.1. Background

In recent years, it has become widely recognized that the human microbiome significantly influences host digestion, immunity, metabolism, and even neural function, making its balance critical for overall health [1]. A central task in microbiome research is differential abundance analysis (DAA), which aims to identify reliable marker features that distinguish groups within microbiome datasets [2–6].

Microbiome datasets are characterized by high dimensionality and often include hundreds or even thousands of features, such as bacterial taxa or genes. Biomarker discovery typically involves testing each feature for statistical associations. Because each test carries a non-zero probability of a false positive (Type I error), conducting thousands of tests creates a high risk of accidentally discovering significant results. Consequently, microbial analysis inherently presents a multiple comparison problem, highlighting the need for hypothesis filtration [7,8].

Given this context, the high dimensionality of microbiome data creates ample opportunity for false positives. Nonetheless, the number of identified microbiome markers is often strikingly small compared to the total number of taxonomic features examined [9,10], and the overwhelming majority of microbiome variation (>80%) consistently remains unaccounted for [11–13] by studied demographic, clinical, or environmental factors. This discrepancy can be explained, in considerable part, by intra-person variability and the inherent biological stability of microbiome systems.

Research indicates that human microbiome systems are generally stable and resilient [1]. For instance, longitudinal studies demonstrate high rates of re-identification (over 86%) of unique microbial species upon repeated patient examinations [14]. Similarly, core microbiome species are consistently identified across large populations and explain a substantial proportion of microbiome variation [15]. At the same time, the largest part of explained microbiome variability is attributed to intra-person differences [16], complicating pattern identification. Nonetheless, dysbiosis can arise from known subtle microbial imbalances [17]. Minor shifts in the microbiome can trigger community-wide changes in structure, metabolic activity, or both, ultimately leading to significant phenotypic effects [18–22].

Thus, one of the methodological challenges of microbial DAA is the detection of subtle changes in a microbiome that has large intra-person variation, yet which overall composition appears to be stable. The challenge of detecting these changes in microbial composition is threefold. First, increases in pathobionts or decreases in key commensals are often minor. Second, pathobionts and key commensals typically represent only a small fraction of the total microbial population. Third, the change in their relative abundance (which is already minimal) can itself be small. After conducting tests and adjusting for multiple comparisons across all taxonomic features, these subtle but biologically important changes can resurface as insignificant.

The reason for this obscuring lies in the nature of multiple comparison corrections. Although these hypothesis filtration methods are designed to control false positives (Type I errors), they inherently reduce statistical power [23]. This can be particularly problematic for high-dimensional data, such as microbiome datasets, where most features are not differentially abundant (i.e., “stable”). In this context, standard corrections become overly conservative, disproportionately penalizing the few true signals and leading to false negatives (Type II errors). Given this limitation, it can be concluded that other hypothesis filtrations are in demand in microbial DAA.

Another popular option for ranking hypotheses is filtration by effect size [2]. Most effect size measures (e.g., from the *r* and *d* families) provide information about the effect size that complements the statistical significance indicated by the *p*-value. Large effect sizes with reasonable sample sizes often reveal more about practical importance than small *p*-values. However, effect size measures present their own challenges. For example, although tools such as LEfSe (Linear discriminant analysis Effect Size) are widely popular in microbial DAA [2], they can be susceptible to generating or inflating false positives [7]. This vulnerability arises because effect sizes, whether LDA estimates from LEfSe or other common statistics such as Cliff’s δ , are single-point estimates intended to complement or

extend the underlying significance test [24]. Consequently, these point estimates remain highly dependent on sample peculiarities.

One well-known, straightforward way to address this issue is to consider confidence intervals (CIs). CIs for between-group differences can be seen as a better alternative to raw *p*-values and related point estimate statistics because they provide a range of plausible location shift values [25]. This range simultaneously conveys information about significance, the potential magnitude and direction of the effect, and the certainty of the estimate itself.

A critical question is what constitutes the optimal CI for analyzing microbiological data. Since a desirable CI roughly describes the plausible range for the true shift between the central tendencies of two samples, the first step can be to obtain robust estimates of the location parameters (the central values) of the microbial data itself.

The overdispersed, zero-inflated count data typical of microbiome samples can be adequately modeled using the negative binomial (NB) distribution [26]. This distribution models the number of failures before a specified number of successes in a series of independent Bernoulli trials. A defining feature of the NB distribution is its right-skewed, unimodal shape with a long right tail, which naturally captures the overdispersion (variance exceeding the mean) observed in microbial abundance data. The shape of the distribution, specifically its degree of skewness and the location of its peak, is governed by its parameters. As the number of required successes increases, the distribution becomes less skewed and its shape approaches that of a symmetric bell curve. Consequently, an NB distribution with substantial positive skew is most-optimal for modeling low-abundance, sparse taxa, while a distribution nearing symmetry better describes common, abundant taxa. The open question is: which measure of central tendency would most accurately represent the center of NB-distributed microbial data?

The mean is the most commonly used measure of central tendency. In the context of the NB distribution, it is directly sensitive to the parameters governing the distribution's spread and location. It also aligns well with the zero-inflated nature of microbial data, making it a theoretically useful metric for reflecting both taxonomic abundance and prevalence. However, microbiome datasets frequently contain outliers, which in practice renders the mean an unreliable statistic that is unduly influenced by extreme values. Related robust metrics, such as the trimmed and winsorized mean, offer greater resistance to extreme values. Yet these methods introduce a new challenge: determining an optimal and biologically justified threshold for trimming or winsorization [27]. Moreover, even after applying these adjustments, the reliability of the resulting estimate is not guaranteed. The median, defined as the middle value of a ranked vector, is less susceptible to outliers than the mean, providing greater robustness for sample comparisons. However, it may be the least accurate theoretical estimate of the true central parameter of a NB distribution, as it does not account for right skew. Moreover, in the context of zero-inflated microbial data, the median tends to underestimate the true central tendency and can easily yield a zero estimate for samples with low prevalence.

A robust alternative to traditional measures can be the pseudo-median, calculated via the Hodges–Lehmann estimator (HLE) [28,29]. This estimator is both sensitive to the right skew of the NB distribution without being unduly influenced by extreme outliers. In the one-sample case, the HLE is defined as the median of all pairwise averages, also known as Walsh averages. In a symmetric distribution, the HLE equals the median. However, for right-skewed NB data, the HLE lies between the median and the mean, effectively balancing the robustness of the median with the sensitivity of the mean. This balance is achieved by incorporating information about the data's spread through all pairwise averages (similar to the mean) while remaining resistant to extreme values by taking their

median. Consequently, the HLE offers a favorable compromise. The two-sample HLE is defined as the median of all pairwise differences between the observations in two samples. CIs for this estimator can be derived by inverting the Mann–Whitney U count statistic, yielding a non-parametric CI for the difference between the location parameters of two NB samples.

Additionally, unlike conventional tests that primarily yield a p -value and corresponding test statistic, the two-sample HLE estimate directly results in an intuitive and interpretable measure of the shift in central tendencies between groups (i.e., median of all observed pairwise differences in two samples). Consequently, the one-sample HLE (pseudo-median) and the two-sample HLE for differences between locations may represent a powerful, assumption-lean, and robust tool for analyzing microbial data, warranting further investigation.

1.2. The Aim and Objectives of the Study

In summary, Type II errors (false negatives) can be as critical as Type I errors (false positives) in microbial DAA. Standard multiple testing corrections tend to over-penalize comparisons in datasets with few true signals, such as those typical of many microbiome studies, often to the point of prohibiting any marker taxa detection.

The aim of this study is to demonstrate the utility of HLE-CI-based hypothesis filtration for mitigating the issue of marker taxa discovery over-penalization during hypothesis filtration using multiple testing corrections.

To this end, the study addresses two specific objectives. First, we investigate marker discovery error rates using synthetic data, comparing several approaches: raw and adjusted p -values from popular statistical tests, as well as effect-size-based filtration. Second, we perform a comparative analysis of HLE-CI filtration performance on a real-world dataset.

1.3. Summary of the Results

The results of this study demonstrate that in datasets with sparse genuine signals, conventional multiple comparison adjustments inadvertently over-penalize and obscure true biological signals. In contrast, filtration based on HLE-CIs provides significantly higher recall while maintaining superior or comparable precision. Consequently, we show that HLE-CI achieves significantly higher F1 scores, an advantage that persists even with decreasing sample sizes. The results of this study illustrate that HLE-CI-based hypothesis filtration can be a beneficial and powerful approach for biomarker discovery in microbiome data.

2. Materials and Methods

2.1. Synthetic Dataset

Simulation data were generated by random sampling from a NB distribution using the NumPy v2.0.1 package in Python v3.12. The distribution's parameters were randomly drawn for each simulation: the parameter 'n' (number of successes) was sampled from a discrete uniform interval of 1 to 10, and the probability of success 'p' was sampled from a continuous uniform interval of 0 to 1. To ensure reproducibility, the random seed was fixed for each iteration. Each simulation experiment comprised 100,000 draws to minimize stochastic effects.

2.2. Experimental Dataset

The Inflammatory Bowel Disease (IBD) dataset comprised IBD patients and their healthy relatives (HR) 16s microbiome profiling data. Patients were recruited at clinical sites in Almaty, Kazakhstan. All participants provided written informed consent. In total,

270 individuals over 18 years of age were enrolled, forming 135 matched pairs (135 IBD patients and 135 healthy relatives). This dataset collection and use was approved by the Ethics Committee of the PI “National Laboratory Astana,” Nazarbayev University (Protocol #05-2022, 21 October 2022).

Bioinformatics and Statistical Analysis of Experimental Data

Sequencing data was processed using the LotuS2 bioinformatics pipeline. Taxonomic annotation of ASVs was performed using the Lambda classifier and the SILVA v138.1 reference database. Before diversity analysis data was rarified. Biodiversity analysis, feature importance analysis and visualization were performed in Python v3.12 using the NumPy v2.0.1, SciPy v1.15.1, scikit-bio v0.6.3, scikit-learn v1.6.1, Matplotlib v3.10.0, and seaborn v0.13.2 packages. Only taxa that prevalent in 50% of samples in any group were considered for analysis. Beta diversity (between group diversity) was assessed using the Jaccard and Bray–Curtis metrics. ANOSIM tests with 999 permutations were used to assess the significance of grouping. Alpha diversity (within sample diversity) was assessed using the Observed, Pielou, and Faith indices at the ASV level.

2.3. Statistical Method Implementation

All standard statistical tests used throughout the study, including the independent *t*-test, Mann–Whitney U test, Brunner–Munzel test, and Mood’s median test, were performed using SciPy v1.15.1. Multiple comparison adjustments via the Benjamini–Hochberg procedure and the Holm–Bonferroni method were implemented using statsmodels v0.14.4. Cliff’s delta (δ) effect size was calculated directly from the Mann–Whitney U statistic. One sample and two sample Hodges–Lehmann estimators were implemented using NumPy v2.0.1 and SciPy v1.15.1 packages. For comparative analysis, the ANCOM-BC2 algorithm from the ANCOMBC v2.8.1 package in R v4.4.2 was used. All data and scripts required to reproduce the analysis presented in this work are available at <https://github.com/VeaLi/hle-ci-for-microbial-daa> (accessed on 10 December 2025).

3. Results

3.1. Empirical False Positives Rates in Negative Binomial Data

We conducted a simulation experiment to estimate the false positive rates (FPR) of several statistical methods: the independent *t*-test (TT; note that TT solution equals to OLS linear regression solution in case with one predictor, and thus omitted); the *t*-test with Welch’s correction (TTW); the *t*-test applied to log-normalized data with a pseudocount (PL-TT and PL-TTw); the Mann–Whitney U test (MWU); the Brunner–Munzel test (BMT), an alternative to the MWU that does not assume identical shapes; Mood’s median test (MMT), a more robust alternative to the MWU; Cliff’s Delta (δ), an effect size measure derived from the MWU statistic representing the degree of non-overlap; and the two-sample Hodges–Lehmann estimator (HLE)-based test (HLE-CI, which uses non-overlapping CIs for significance estimation).

We generated 100,000 random samples from a NB distribution with different location parameters. For each sample, we randomly split the data into two equal halves. As these halves are random subsets of the same large sample, no true difference should exist between them. We then applied each statistical test to compare the halves to determine if any method falsely detected a significant difference. This procedure allowed us to estimate the empirical false positive rate for each method when applied to data from a NB distribution. The results of this analysis are presented in Table 1.

Table 1. Empirical false positive error rate from 100,000 simulations drawn from a negative binomial distribution (NB) distribution.

Test/Sample Size (<i>n</i> per Group)	<i>n</i> = 15	<i>n</i> = 25	<i>n</i> = 50	<i>n</i> = 100
$\alpha = 0.05$				
TT, $p < 0.05$	4793	4869	4895	4926
TTW, $p < 0.05$	4572	4794	4884	4920
PL-TT, $p < 0.05$	4893	5059	4964	4960
PL-TTW, $p < 0.05$	4818	5032	4953	4958
MWU, $p < 0.05$	4558	4863	4922	4834
BMT, $p < 0.05$	5283	5268	5067	4903
MMT, $p < 0.05$	1792	2872	3119	3626
95% HLE-CI	1793	1458	1076	816
$\alpha = 0.01$				
TT, $p < 0.05$	925	932	900	919
TTW, $p < 0.05$	821	863	892	914
PL-TT, $p < 0.05$	994	1035	945	966
PL-TTW, $p < 0.05$	936	984	942	966
MWU, $p < 0.05$	784	908	889	932
BMT, $p < 0.05$	1370	1225	1047	1005
MMT, $p < 0.05$	278	365	571	575
99% HLE-CI	261	254	174	136
$\alpha = 0.001$				
TT, $p < 0.05$	73	80	75	99
TTW, $p < 0.05$	59	68	70	99
PL-TT, $p < 0.05$	105	106	84	82
PL-TTW, $p < 0.05$	90	98	82	81
MWU, $p < 0.05$	45	69	61	90
BMT, $p < 0.05$	270	182	119	117
MMT, $p < 0.05$	19	19	36	45
99.9% HLE-CI	14	14	8	11
Cliff's δ				
$\delta > 0.15$ (small)	45,739	33,583	17,343	5462
$\delta > 0.33$ (medium)	10,606	3606	292	2
$\delta > 0.47$ (large)	2191	285	2	0
Cliff's δ & MWU, $p < 0.05$				
$\delta > 0.15$ (small), $p < 0.05$	4558	4863	4729	4287
$\delta > 0.33$ (medium), $p < 0.05$	4443	3606	292	2
$\delta > 0.47$ (large), $p < 0.05$	2191	285	2	0

TT = independent *t*-test; TTW = independent *t*-test with Welch correction; PL = log of data with pseudo count added; MWU = Mann-Whitney-U test; BMT = Brunner-Munzel test; MMT = Mood's median test; HLE-CI = confidence interval for two-sample Hodges-Lehmann estimation of location shift.

At a significance level of $\alpha = 0.05$, all *t*-test-based methods (TT, TTW, PL-TT, PL-TTW), the MWU, and the BMT demonstrated a higher false positive rate (FPR) when comparing subsamples from the same NB distribution than the HLE-CI method. Specifically, these tests were approximately 2.5 times more likely to produce false positives for subsamples of size 15, 3.2 times for size 25, 4.5 times for size 50, and 5.6 times for size 100. The MWU performed only marginally better than the *t*-tests, while BMT performed marginally worse. MMT showed a 1.5-fold lower FPR than the other tests but remained inferior to HLE-CI. Notably, while MMT produced FPRs similar to HLE-CI for a sample size of 15, it was 1.9 times more likely to yield a false positive for size 25, 2.9 times for size 50, and 4.4 times for size 100.

As the sample size increased, the FPRs of the t -tests, MWU, BMT, and MMT also increased. This pattern is consistent with their growing statistical power, which enables the detection of increasingly negligible, biologically irrelevant differences. In contrast, the HLE-CI's FPRs decreased with larger sample sizes, indicating a corresponding increase in the accuracy of median difference estimate. Similar but more pronounced trends were observed at $\alpha = 0.01$ and $\alpha = 0.001$.

Unconditional filtration by Cliff's δ effect size (small and medium effects) resulted in the highest FPR, although this rate decreased with increasing sample size. Coupling Cliff's δ with the MWU p -value limited these spurious findings; however, the resulting performance was not substantially better than using the MWU alone without the corresponding effect size. When a large effect size threshold was used for filtration, the performance of Cliff's δ -based filtration was comparable to that of the HLE-CI method.

The results of this experiment suggest that: (1) the HLE-CI method does not produce results similar or identical to the MWU or related tests (contrary to the relationship between a t -test and its complementary CIs); (2) HLE-CI yields the lowest FPR for samples drawn from NB distribution; (3) HLE-CI demonstrates the most optimal performance even for small sample sizes ($n = 15$ – 25).

3.2. Empirical False Negative Rates in Negative Binomial Data

We conducted a further simulation experiment to compare the performance of the best performed methods: MWU and MMT with and without multiple comparisons correction (FDR-BH, denoted as q), Cliff's δ -based filtration, and the HLE-CI method in a scenario designed to detect a small number of true positive markers. Similarly to the first experiment, we generated 1000 pairs of samples from a NB distribution to simulate taxonomic feature abundances for two experimental groups. To create a dataset with a known, small fraction of true positives, we introduced a significant difference by shifting the values of 10 randomly selected features by a fixed amount (15%, 25%, or 50%) in one of the groups. This resulted in a final dataset of 1000 features containing 10 true markers. We repeated the simulation 100 times and computed the F1-score, which summarizes the balance between precision and recall. The F1-score is the harmonic mean of precision and recall, calculated as shown in Equation (1). The results are presented in Tables 2–4.

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (1)$$

At a 15% marker shift, the HLE test with 95% and 99% CIs demonstrated the best performance across all dataset sizes (Table 2), consistently outperforming the MWU and MMT at all α -levels. The use of multiple comparison adjustment resulted in suboptimal performance, effectively prohibiting marker detection in datasets with a small number of true positives at the 15% difference in marker centers. Only when the sample size was increased to 100 per group did filtration by corrected p -values show any detection capability, yielding better results than filtration by raw MWU p -values. Nevertheless, the highest F1 score was still achieved by the HLE-CI filtration.

The maximum F1 scores for largest sample size ($n = 100$) for filtration by raw p -values, adjusted p -values, and the HLE-CI results were 32.65%, 18.2%, and 50.0%, respectively. Thus, the HLE test provided a 31.8% relative improvement in the F1 score over the adjusted p -values. The highest F1 score achieved using a conditional Cliff's δ was 19.65%, which occurred at a medium effect size threshold. Furthermore, even with smaller sample sizes ($n = 15$ – 25) HLE-based filtration performed the best. Finally, the MMT did not perform better than the MWU, suggesting that the MMT's lower empirical false positive rate

(observed on Experiment 1; Table 1) was a result of its lower statistical power, rather than higher accuracy.

Table 2. F1 from 100 simulations of 1000 comparisons comparing negative binomial (NB) subsamples with 10 markers per 1000 comparisons. 15% shift between centers. Md (IQR).

Test/Sample Size (<i>n</i> per Group)	<i>n</i> = 15	<i>n</i> = 25	<i>n</i> = 50	<i>n</i> = 100
$\alpha = 0.05$				
MWU, $p < 0.05$	6.2 (3.4; 8.2)	7.9 (5.95; 11.32)	12.15 (9.95; 15.25)	17.4 (14.65; 21.2)
MMT, FDR-BH, $q < 0.05$	0.0 (0.0; 0.0)	0.0 (0.0; 0.0)	0.0 (0.0; 18.2)	18.2 (0.0; 33.3)
MMT, $p < 0.05$	5.35 (0.0; 7.7)	5.7 (4.7; 10.0)	9.3 (5.07; 14.85)	13.95 (10.3; 18.52)
MMT, FDR-BH, $q < 0.05$	0.0 (0.0; 0.0)	0.0 (0.0; 0.0)	0.0 (0.0; 0.0)	0.0 (0.0; 18.2)
95% HLE-CI	11.8 (6.8; 16.0)	17.8 (10.38; 25.8)	30.8 (23.05; 38.5)	45.5 (38.5; 52.2)
$\alpha = 0.01$				
MWU, $p < 0.01$	0.0 (0.0; 11.1)	10.5 (0.0; 19.0)	20.7 (10.0; 28.6)	32.65 (23.78; 40.35)
MMT, FDR-BH, $q < 0.01$	0.0 (0.0; 0.0)	0.0 (0.0; 0.0)	0.0 (0.0; 0.0)	18.2 (0.0; 33.3)
MMT, $p < 0.01$	0.0 (0.0; 0.0)	0.0 (0.0; 14.3)	12.5 (9.4; 18.2)	22.2 (13.3; 28.95)
MMT, FDR-BH, $q < 0.01$	0.0 (0.0; 0.0)	0.0 (0.0; 0.0)	0.0 (0.0; 0.0)	0.0 (0.0; 0.0)
99% HLE-CI	0.0 (0.0; 15.4)	15.4 (0.0; 26.7)	30.8 (16.7; 40.0)	50.0 (39.38; 57.1)
$\alpha = 0.001$				
MWU, $p < 0.001$	0.0 (0.0; 0.0)	0.0 (0.0; 16.7)	16.7 (0.0; 18.2)	30.8 (16.7; 43.72)
MMT, FDR-BH, $q < 0.001$	0.0 (0.0; 0.0)	0.0 (0.0; 0.0)	0.0 (0.0; 0.0)	0.0 (0.0; 18.2)
MMT, $p < 0.001$	0.0 (0.0; 0.0)	0.0 (0.0; 0.0)	0.0 (0.0; 15.4)	18.2 (0.0; 30.8)
MMT, FDR-BH, $q < 0.001$	0.0 (0.0; 0.0)	0.0 (0.0; 0.0)	0.0 (0.0; 0.0)	0.0 (0.0; 0.0)
99.9% HLE-CI	0.0 (0.0; 0.0)	0.0 (0.0; 18.2)	18.2 (0.0; 18.2)	33.3 (18.2; 46.2)
Cliff's δ & MWU, $p < 0.05$				
$\delta > 0.15$ (small), $p < 0.05$	6.2 (3.4; 8.2)	7.9 (5.95; 11.32)	12.5 (10.0; 15.52)	18.55 (16.08; 22.6)
$\delta > 0.33$ (medium), $p < 0.05$	6.15 (3.48; 8.2)	9.3 (4.88; 13.6)	19.65 (13.3; 29.15)	18.2 (0.0; 33.3)
$\delta > 0.47$ (large), $p < 0.05$	6.7 (3.22; 11.18)	12.5 (0.0; 15.4)	0.0 (0.0; 18.2)	0.0 (0.0; 0.0)

MWU = Mann–Whitney–U test; MMT = Mood's median test; HLE-CI = confidence interval for two-sample Hodges–Lehmann estimation of location shift.

Similarly, at a 25% marker shift, the HLE filtration with 95% and 99% CIs demonstrated the best performance across all dataset sizes (Table 3), consistently outperforming the MWU and MMT at all α -levels. The use of multiple comparison adjustment again resulted in suboptimal performance. Only when the sample size was increased to 50 per group did filtration by corrected p -values show any detection capability. Filtration by raw MWU p -values, in general, performed better than by adjusted p -values, due to higher recall. The highest F1 score was still achieved by the HLE-CI filtration.

The maximum F1 scores for largest sample size ($n = 100$) for filtration by raw p -values, adjusted p -values, and the HLE test were 57.1%, 46.2%, and 70.6%, respectively. Thus, the HLE test provided a 24.4% relative improvement in the F1 score over the adjusted p -values. The highest F1 score achieved using a conditional Cliff's δ was 46.2%, which occurred at a medium effect size threshold. For smaller sample sizes ($n = 15$ –25) HLE-based filtration again demonstrated the best performance among the methods. The MMT did not perform any better than the MWU.

Finally, at a 50% marker shift, the HLE test with 99% CIs demonstrated the best performance across all dataset sizes (Table 4), consistently outperforming the MWU and MMT at all α -levels. The use of multiple comparison adjustment resulted in suboptimal performance for smaller sample size ($n = 15$), overpenalizing marker detection. Filtration by adjusted MWU p -values, in general, performed better than by raw p -values (by increasing precision). The highest F1 score was still achieved by the HLE-CI filtration.

Table 3. F1 from 100 simulations of 1000 comparisons comparing negative binomial (NB) subsamples with 10 markers per 1000 comparisons. 25% shift between centers. Md (IQR).

Test/Sample Size (<i>n</i> per Group)	<i>n</i> = 15	<i>n</i> = 25	<i>n</i> = 50	<i>n</i> = 100
$\alpha = 0.05$, Md (IQR)				
MWU, $p < 0.05$	7.25 (4.7; 11.42)	11.5 (9.07; 14.98)	17.0 (13.68; 20.08)	22.9 (19.77; 25.4)
MMT, FDR-BH, $q < 0.05$	0.0 (0.0; 0.0)	0.0 (0.0; 0.0)	18.2 (0.0; 20.8)	46.2 (33.3; 66.7)
MMT, $p < 0.05$	6.5 (0.0; 8.07)	8.4 (4.9; 13.55)	13.8 (10.38; 18.2)	19.65 (16.23; 23.55)
MMT, FDR-BH, $q < 0.05$	0.0 (0.0; 0.0)	0.0 (0.0; 0.0)	0.0 (0.0; 0.0)	18.2 (0.0; 33.3)
95% HLE-CI	15.4 (8.9; 22.42)	24.0 (16.7; 30.8)	40.0 (32.3; 46.92)	56.7 (48.3; 62.48)
$\alpha = 0.01$, Md (IQR)				
MWU, $p < 0.01$	9.75 (0.0; 14.73)	16.35 (9.5; 22.52)	31.6 (24.62; 40.42)	48.15 (39.62; 57.35)
MMT, FDR-BH, $q < 0.01$	0.0 (0.0; 0.0)	0.0 (0.0; 0.0)	0.0 (0.0; 18.2)	46.2 (33.3; 57.1)
MMT, $p < 0.01$	0.0 (0.0; 0.0)	0.0 (0.0; 14.3)	21.1 (11.62; 28.6)	37.5 (28.6; 47.22)
MMT, FDR-BH, $q < 0.01$	0.0 (0.0; 0.0)	0.0 (0.0; 0.0)	0.0 (0.0; 0.0)	9.1 (0.0; 18.2)
99% HLE-CI	13.3 (0.0; 18.2)	19.65 (14.3; 29.15)	44.4 (33.3; 57.1)	70.6 (58.8; 78.35)
$\alpha = 0.001$, Md (IQR)				
MWU, $p < 0.001$	0.0 (0.0; 0.0)	0.0 (0.0; 18.2)	30.8 (18.2; 42.9)	57.1 (46.2; 66.7)
MMT, FDR-BH, $q < 0.001$	0.0 (0.0; 0.0)	0.0 (0.0; 0.0)	0.0 (0.0; 0.0)	18.2 (18.2; 33.3)
MMT, $p < 0.001$	0.0 (0.0; 0.0)	0.0 (0.0; 0.0)	16.7 (0.0; 18.2)	33.3 (18.2; 46.2)
MMT, FDR-BH, $q < 0.001$	0.0 (0.0; 0.0)	0.0 (0.0; 0.0)	0.0 (0.0; 0.0)	0.0 (0.0; 18.2)
99.9% HLE-CI	0.0 (0.0; 0.0)	0.0 (0.0; 18.2)	33.3 (18.2; 46.2)	59.8 (46.2; 75.0)
Cliff's δ & MWU, $p < 0.05$, Md (IQR)				
$\delta > 0.15$ (small), $p < 0.05$	7.25 (4.7; 11.42)	11.5 (9.07; 14.98)	17.4 (14.02; 20.62)	24.6 (21.42; 27.52)
$\delta > 0.33$ (medium), $p < 0.05$	7.4 (4.8; 11.5)	13.45 (8.7; 17.82)	35.3 (28.12; 50.0)	46.2 (33.3; 57.1)
$\delta > 0.47$ (large), $p < 0.05$	8.2 (5.6; 12.5)	15.4 (0.0; 25.0)	18.2 (0.0; 18.2)	0.0 (0.0; 18.2)

MWU = Mann–Whitney–U test; MMT = Mood's median test; HLE-CI = confidence interval for two-sample Hodges–Lehmann estimation of location shift.

Table 4. F1 from 100 simulations of 1000 comparisons comparing negative binomial (NB) subsamples with 10 markers per 1000 comparisons. 50% shift between centers. Md (IQR).

Test/Sample Size (<i>n</i> per Group)	<i>n</i> = 15	<i>n</i> = 25	<i>n</i> = 50	<i>n</i> = 100
$\alpha = 0.05$, Md (IQR)				
MWU, $p < 0.05$	17.0 (13.95; 20.15)	21.9 (18.48; 24.3)	25.2 (22.42; 28.6)	26.0 (23.65; 28.2)
MMT, FDR-BH, $q < 0.05$	0.0 (0.0; 0.0)	18.2 (17.82; 33.3)	66.7 (57.1; 75.0)	84.2 (77.8; 90.0)
MMT, $p < 0.05$	18.2 (12.4; 24.25)	20.0 (15.8; 25.0)	26.4 (21.7; 31.02)	26.7 (23.0; 29.6)
MMT, FDR-BH, $q < 0.05$	0.0 (0.0; 0.0)	0.0 (0.0; 18.2)	46.2 (33.3; 57.1)	70.6 (57.1; 75.0)
95% HLE-CI	31.2 (25.8; 38.7)	42.9 (36.22; 48.3)	53.55 (49.7; 61.5)	64.0 (57.1; 69.2)
$\alpha = 0.01$, Md (IQR)				
MWU, $p < 0.01$	31.6 (23.5; 40.42)	43.2 (32.0; 50.48)	57.6 (50.0; 64.0)	59.65 (53.3; 66.7)
MMT, FDR-BH, $q < 0.01$	0.0 (0.0; 0.0)	17.45 (0.0; 18.2)	57.1 (46.2; 66.7)	82.4 (75.0; 88.9)
MMT, $p < 0.01$	15.4 (0.0; 27.18)	31.2 (18.8; 40.0)	50.0 (40.0; 60.9)	58.3 (50.0; 64.08)
MMT, FDR-BH, $q < 0.01$	0.0 (0.0; 0.0)	0.0 (0.0; 0.0)	33.3 (18.2; 46.2)	66.7 (57.1; 75.0)
99% HLE-CI	40.0 (30.8; 50.0)	53.9 (42.9; 66.7)	75.0 (66.7; 80.0)	84.2 (77.8; 90.0)
$\alpha = 0.001$, Md (IQR)				
MWU, $p < 0.001$	18.2 (0.0; 30.8)	42.9 (30.8; 53.3)	73.7 (58.8; 77.8)	84.2 (77.8; 88.9)
MMT, FDR-BH, $q < 0.001$	0.0 (0.0; 0.0)	0.0 (0.0; 0.0)	46.2 (18.2; 57.1)	82.4 (72.92; 82.4)
MMT, $p < 0.001$	0.0 (0.0; 18.2)	18.2 (0.0; 33.3)	51.65 (42.18; 63.55)	70.6 (62.5; 77.8)
MMT, FDR-BH, $q < 0.001$	0.0 (0.0; 0.0)	0.0 (0.0; 0.0)	18.2 (0.0; 33.3)	57.1 (46.2; 66.7)
99.9% HLE-CI	18.2 (0.0; 33.3)	44.55 (30.8; 47.98)	72.8 (57.1; 78.95)	87.3 (82.4; 91.18)

Table 4. Cont.

Test/Sample Size (<i>n</i> per Group)	<i>n</i> = 15	<i>n</i> = 25	<i>n</i> = 50	<i>n</i> = 100
Cliff's δ & MWU, $p < 0.05$, Md (IQR)				
$\delta > 0.15$ (small), $p < 0.05$	17.0 (13.95; 20.15)	21.9 (18.48; 24.3)	25.65 (23.05; 29.0)	28.4 (25.4; 30.3)
$\delta > 0.33$ (medium), $p < 0.05$	17.2 (13.95; 20.15)	26.1 (20.78; 29.02)	70.0 (62.5; 76.2)	82.4 (75.0; 88.9)
$\delta > 0.47$ (large), $p < 0.05$	25.7 (21.1; 29.4)	48.8 (34.8; 58.8)	57.1 (46.2; 66.7)	57.1 (46.2; 66.7)

MWU = Mann–Whitney–U test; MMT = Mood's median test; HLE-CI = confidence interval for two-sample Hodges–Lehmann estimation of location shift.

The maximum F1 scores for largest sample size ($n = 100$) for filtration by raw p -values, adjusted p -values, and the HLE test were 84.2%, 84.2%, and 87.3%, respectively. Thus, the HLE test performed similarly to adjusted p -values in the task of detecting large magnitude changes (i.e., 50% shift). The highest F1 score achieved using a conditional Cliff's δ was 82.4%, which occurred at a medium effect size threshold. Again, the HLE-based filtration showed the best results for smaller sample sizes ($n = 15$ – 25), than any other method.

The results of this experiment suggest that: (1) multiple testing correction tends to overpenalize comparisons in datasets with a low proportion of true positives, such as those common in microbiome studies; (2) the HLE-CI method consistently achieves the best, or at least comparable, F1-scores across all tested sample sizes ($n = 15$ – 100) and effect sizes (15–50% marker shift). (3) HLE-CI-based filtration performs the most optimally at the 95–99% confidence levels and (4) may serve as a powerful tool for hypothesis filtration in microbial DAA, performing on par with multiple comparison correction.

3.3. Precision and Sensitivity of Hypothesis Filtration in Human Gut Microbiome Data

In the previous section, we demonstrated that spurious significant associations can be found in NB data by chance and that multiple comparison correction over-penalizes discoveries in datasets with a low proportion of true positives. Here, we assess the performance of the best-performing hypothesis filtration method from our simulation HLE-CI on real-world Inflammatory Bowel Disease (IBD) data, comparing the 16S microbiomes' profiles of IBD ($n = 135$) patients against those of their healthy relatives (HR, $n = 135$). We compare this HLE-CI-based method with ANCOM-BC2 (without correction, with default Holm–Bonferroni method and with the Benjamini–Hochberg procedure), a state-of-the-art method specifically designed for microbiome data. ANCOM-BC2 is based on a regression framework with sampling depth correction and utilizes multiple comparison correction for hypothesis filtration. Further details on ANCOM-BC2 can be found in [5]. DAA was considering only features with a prevalence of at least 50% in the full dataset (343 out of 2272 all classified taxonomic features in the dataset).

3.3.1. Preliminary Biodiversity Analysis

IBD is a significant global burden, affecting millions of people and associated with reduced quality of life and an increased risk of complications such as colorectal cancer. As shown in Figure 1, the gut microbiome structure in IBD patients differs significantly from that of healthy controls. Figure 1A shows a beta diversity analysis based on Jaccard and Bray–Curtis distances at the ASV level for the IBD and HR groups. The IBD and HR samples form significantly separate clusters (ANOSIM, $R = 0.13$, $p < 0.001$ and $R = 0.11$, $p < 0.001$), indicating distinct microbiome profiles and suggesting a dysbiotic shift in IBD patients. IBD is also characterized by chronic intestinal inflammation and a dramatic reduction in unique taxa. In this IBD patients dataset, a 2-fold reduction in alpha diversity at ASV level was observed (Figure 1B, Observed, $p = 0.001$ and Pielou, $p = 0.001$). Leave-one-out

cross-validation without any feature selection also demonstrated robust separability on genera level (Figure 1C, AUC = 0.84).

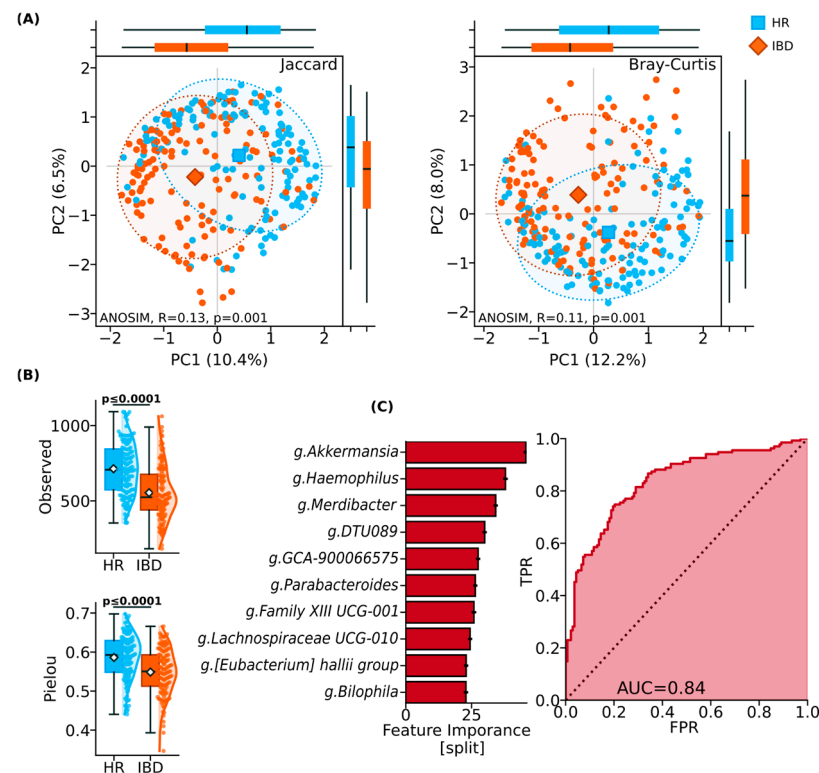


Figure 1. The metagenome of patients with inflammatory bowel disease (IBD) shows a strong structural shift and a critical reduction in biodiversity compared to their healthy relatives (HR). (A) Beta diversity (between group diversity) analysis. Jaccard and Bray–Curtis metrics. ANOSIM grouping significance test with 999 permutations. (B) Alpha diversity (within sample diversity) analysis. (C) Classification significance analysis in leave-one-out cross-validation (at the genus level). ROC analysis of classification performance. AUC values range from 0 to 1, where 1 indicates perfect classification. AUC > 0.5 indicates non-random classification. AUC = Area Under the Curve; FPR = False Positive Rate; TPR = True Positive Rate.

Overall, the diversity analysis indicates significant differences between the microbial profiles of IBD patients and HR controls, suggesting the presence of confident taxonomic markers. This and well-researched nature of IBD makes this dataset, and IBD data, in general, a popular choice [2,5] for microbiome DAA method benchmarking.

3.3.2. Differential Analysis

DAA was performed using HLE-CI filtration on rarefied, TSS-normalized relative abundance data and using ANCOM-BC2 on non-rarefied non-normalized counts. ANCOM-BC2 was run using its default settings. The ANCOM-BC2's authors recommend the Holm–Bonferroni method over the Benjamini–Hochberg procedure for multiple comparisons correction. In our experiment, we applied both the recommended Holm–Bonferroni method and the popular-choice Benjamini–Hochberg procedure. Additionally we performed an ANCOM-BC2 analysis without any correction. Features were considered significantly differentially abundant based on the following criteria: for the HLE-CI filtration, non-overlapping CIs; for ANCOM-BC2, a p/q -value < 0.05 and passing pseudo-count sensitivity analysis check (SS Filter).

To evaluate the performance of the compared DAA methods, we first analyzed their agreement on the full IBD dataset. We then assessed their performance on progressively

smaller random subsets, ranging from 100 samples per group to 15 per group. Each subset size was tested with 50 iterations to mitigate stochastic effects, and all methods were applied to the identical subsets of data in each iteration. Performance in subsets test was evaluated by calculating the percentage of markers that were re-discovered in the smaller subsets (Recall %) and the accuracy of this rediscovery (Precision %).

3.3.3. Comparison of Results Obtained on Full Data

The number of potential biomarkers identified in the full IBD dataset by two methods is presented in Table 5. ANCOM-BC2 (SS Filter) without multiple comparisons adjustment suggested the presence of 109 markers. When multiple comparison correction was applied, ANCOM-BC2 coupled with the Holm–Bonferroni method ($q < 0.05$) suggested only 47 markers, while with the Benjamini–Hochberg procedure ($q < 0.05$) it suggested 96 markers. 95% HLE-CI method proposed 134 markers, 121 markers at the 99% level, and 81 markers at the 99.9% level (Table 5).

Table 5. Number of features present at 50% prevalence in Inflammatory Bowel Disease (IBD) dataset and number of identified markers at each taxonomic level by differential abundance analysis (DAA) methods.

Taxonomic Level/DAA Method	Classified Features Present at 50% Prevalence	ANCOMB-BC2 (SS Filter), no Correction, $p < 0.05$	ANCOMB-BC2 (SS Filter), Holm–Bonferroni, $q < 0.05$	ANCOMB-BC2 (SS Filter), Benjamini–Hochberg, $q < 0.05$	95% HLE-CI	99% HLE-CI	99.9% HLE-CI
Phylum	6	3	2	2	3	3	2
Class	12	5	3	4	5	3	3
Order	25	11	6	11	11	10	5
Family	42	22	10	18	24	22	15
Genus	107	34	15	31	47	42	28
Species	70	19	7	17	20	19	13
ASV	81	15	4	13	24	22	15
Total	343	109	47	96	134	121	81

ASV = Amplicon sequence variant; HLE-CI = confidence interval for two-sample Hodges–Lehmann estimation of location shift.

Substantial overlap was observed between the methods. When comparing the 95% HLE-CI results to ANCOM-BC2 under different correction schemes, the shared markers constituted a high percentage of the ANCOM-BC2 findings: 86 markers (79%) were shared with uncorrected ANCOM-BC2; 38 markers (81%) with Holm–Bonferroni-corrected ANCOM-BC2; and 76 markers (79%) with Benjamini–Hochberg-corrected ANCOM-BC2.

When comparing the 99% HLE-CI results to ANCOM-BC2, the shared markers constituted: 83 markers (76%) were shared with uncorrected ANCOM-BC2; 37 markers (79%) with Holm–Bonferroni-corrected ANCOM-BC2; and 74 markers (77%) with Benjamini–Hochberg-corrected ANCOM-BC2.

When comparing the 99.9% HLE-CI results to ANCOM-BC2, the shared markers constituted: 62 markers (57%) were shared with uncorrected ANCOM-BC2; 36 markers (77%) with Holm–Bonferroni-corrected ANCOM-BC2; and 58 markers (60%) with Benjamini–Hochberg-corrected ANCOM-BC2.

The most conservative results were obtained by ANCOM-BC2 with the recommended Holm–Bonferroni multiple comparison scheme, which suggested the presence of only 47 markers.

These results demonstrate that: (1) the 95% HLE-CI method shows substantial (>79%) agreement with the state-of-the-art ANCOM-BC2 method; (2) on a dataset characterized by large inter-group differences and a large sample size, the number of suggested markers does not differ significantly between Benjamini–Hochberg adjusted p -values and HLE-CI filtration, as was already suggested by simulation (Experiment 2; Tables 2–4); (3) the

multiple comparison-adjusted hypothesis testing recommended by ANCOM-BC2 over-penalizes microbial marker discovery, even in data where large effects are present.

3.3.4. Comparison of Results in Experiments with Smaller Sub-Samples

Having identified the most confident markers for each DAA method using the full IBD dataset, we now empirically evaluate the certainty of these discoveries by measuring the precision and recall of their rediscovery in subsets of reduced sample size. Given the incomplete overlap between markers proposed by different DAA methods on the full dataset, this test aims to empirically estimate the credibility of the proposed markers in the first place, as true biological markers are expected to be consistently rediscovered.

The precision, recall, and F1 score obtained by each method on smaller subsets of original data relative to its own full-dataset results are shown in Table 6.

Table 6. Precision, Recall and F1 score in smaller subsamples of IBD dataset. Md (IQR).

DAA Method/Sample Size (<i>n</i> per Group)	<i>n</i> = 15	<i>n</i> = 25	<i>n</i> = 50	<i>n</i> = 100
Precision, %, Md (IQR)				
ANCOMB-BC2 (SS Filter), no correction	80.0 (71.51; 87.58)	86.69 (78.05; 93.08)	90.97 (86.91; 93.92)	96.65 (93.18; 97.87)
ANCOMB-BC2 (SS Filter), Holm–Bonferroni	0.0 (0.0; 100.0)	100.0 (0.0; 100.0)	100.0 (90.0; 100.0)	94.74 (90.91; 100.0)
ANCOMB-BC2 (SS Filter), Benjamini–Hochberg	74.17 (0.0; 100.0)	98.08 (81.44; 100.0)	96.39 (92.86; 100.0)	97.26 (95.0; 98.62)
95% HLE-CI	83.89 (70.94; 89.47)	89.33 (85.43; 93.21)	93.39 (91.07; 95.49)	95.92 (93.93; 96.71)
99% HLE-CI	93.54 (80.0; 100.0)	95.06 (90.23; 100.0)	96.75 (95.02; 98.35)	98.67 (96.97; 98.98)
99.9% HLE-CI	100.0 (0.0; 100.0)	100.0 (100.0; 100.0)	100.0 (95.03; 100.0)	96.92 (94.33; 100.0)
Recall, %, Md (IQR)				
ANCOMB-BC2 (SS Filter), no correction	16.51 (11.01; 24.54)	27.06 (19.27; 36.7)	51.38 (43.12; 58.72)	83.49 (79.13; 86.01)
ANCOMB-BC2 (SS Filter), Holm–Bonferroni	0.0 (0.0; 6.38)	6.38 (0.0; 14.89)	25.53 (15.43; 36.17)	70.21 (63.83; 76.6)
ANCOMB-BC2 (SS Filter), Benjamini–Hochberg	2.08 (0.0; 10.16)	11.46 (5.47; 20.57)	34.38 (26.3; 48.96)	76.04 (70.05; 81.25)
95% HLE-CI	20.15 (13.62; 29.1)	36.94 (30.04; 41.6)	62.69 (54.85; 67.72)	89.55 (86.01; 93.28)
99% HLE-CI	7.44 (4.13; 14.05)	17.77 (12.6; 24.38)	42.56 (35.74; 48.55)	76.86 (71.28; 79.34)
99.9% HLE-CI	1.23 (0.0; 5.86)	6.79 (3.7; 9.88)	27.78 (21.3; 36.42)	74.69 (67.9; 81.17)
F1, %, Md (IQR)				
ANCOMB-BC2 (SS Filter), no correction	27.27 (18.9; 37.97)	41.7 (31.24; 52.72)	64.74 (58.78; 71.32)	87.88 (86.29; 90.01)
ANCOMB-BC2 (SS Filter), Holm–Bonferroni	0.0 (0.0; 12.0)	12.0 (0.0; 25.45)	40.68 (26.72; 52.58)	81.01 (75.3; 84.97)
ANCOMB-BC2 (SS Filter), Benjamini–Hochberg	4.08 (0.0; 18.05)	20.47 (10.37; 32.91)	50.97 (40.89; 64.51)	85.79 (81.47; 87.64)
95% HLE-CI	32.41 (23.4; 43.88)	52.67 (43.7; 56.82)	74.94 (68.11; 77.96)	92.58 (89.85; 94.11)
99% HLE-CI	13.53 (7.94; 24.16)	29.65 (22.26; 38.62)	59.12 (52.5; 64.75)	86.3 (82.83; 87.94)
99.9% HLE-CI	2.44 (0.0; 11.07)	12.71 (7.14; 17.98)	43.47 (34.85; 52.59)	85.11 (80.07; 87.42)

Md(IQR) = median (interquartile range); HLE-CI = confidence interval for two-sample Hodges–Lehmann estimation of location shift.

The highest precision was achieved by the 99.9% HLE-CI filtration and ANCOM-BC2 (with sensitivity filter) method with the Holm–Bonferroni correction, although the latter failed to detect any markers at a sample size of 15 per group in any of the 50 iterations. The highest recall was obtained by the 95% HLE-CI method, which demonstrated substantially better performance across all sample sizes. Consequently, the highest F1-score was also achieved by the 95% HLE-CI method.

At a sample size of 15 per group, the 95% HLE-CI method showed a performance increase of 5.13% relative to uncorrected ANCOM-BC2, 32.41% relative to the Holm–Bonferroni-corrected version, and 28.33% relative to the Benjamini–Hochberg-corrected version. For a sample size of 25, the respective increases were 11.0%, 40.6%, and 32.2%. At a sample size of 50, the values were 10.2%, 34.26%, and 23.97%. Finally, for a sample size of 100 per group, the increases were 4.7%, 11.6%, and 6.78%, outperforming all other methods.

The results of this experiment suggest that: (1) HLE-CI filtration produces markers at least as robust as those identified by modern, specialized DAA tools; (2) HLE-CI achieves the highest precision and recall in marker rediscovery on increasingly smaller data subsets; (3) HLE-CI filtration outperforms other DAA methods not only by maximizing recall but also by maintaining high precision; (4) HLE-CI filtration maintains the most optimal performance even as the sample size decreases; (5) Given its optimal performance, HLE-CI may offer distinct advantages for microbial DAA due to its underlying statistical principles and the reciprocity with NB distribution properties.

Finally, an additional conclusion can be drawn from these results. The precision-recall performance of the decreasing-subsample test suggested in this study can be of interest for further research in benchmarking metagenomic tools in the absence of true data labels.

3.4. Hodges–Lehmann Estimation Significance Values and Pseudo-Median Descriptive Statistics

The results of experiments in this study demonstrated the value of HLE-CI filtration for analysis of microbiome data. Two practical questions can be additionally considered: calculating a p -value for the HLE and the ways of consistently reporting pseudo-median descriptive statistics of corresponding samples. The following two sections address these points with practical solutions.

3.4.1. p -Values Calculation for Two-Sample Hodges–Lehmann Estimation

In this analysis, we evaluated the significance of the two-sample HLE (calculated as the median of all pairwise differences between two samples) by assessing its CI. If the CI contained zero, the difference was deemed insignificant; otherwise, it was considered significant. This provides a straightforward interpretation analogous to that of a t -test CI for a difference in means. The key distinction is that the HLE-CI is not derived from or complementary to a specific parametric test. As demonstrated in Experiment 1, results from the closest corresponding Mann–Whitney U test do not converge empirically with the HLE-CI solution.

Given that p -values are the popular framework for reporting results, we propose an incremental procedure to derive a p -value equivalent from the HLE-CI analysis. Since each HLE-CI corresponds to a specific confidence level (e.g., 95%), one can directly map the CI to a p -value threshold. To achieve finer resolution, the CI can be incrementally widened to determine a more precise p -value.

For practical application, we recommend assessing significance by constructing CIs at standard alpha levels (e.g., $\alpha < 0.1$, < 0.05 , < 0.01 , < 0.001 , < 0.0001) and reporting the smallest threshold at which the CI excludes zero.

3.4.2. Box Visualization for One-Sample Hodges–Lehmann Estimation

The results of this analysis suggest that the one-sample HLE may be optimal for robustly and efficiently estimating the location parameter of a NB-distributed microbial sample. It is robust (less influenced by outliers) and efficient (closer to the true population parameter).

Since the two-sample HLE estimate represents the shift between two location parameters, reporting in a compatible format pseudo-median descriptive statistics for locations of two samples can be the most advantageous.

The pseudo-median is calculated as the median of all pairwise averages (Walsh averages) within a single sample. In a symmetric distribution, the pseudo-median equals the median, in a right-skewed NB distribution, it lies between the median and the mean.

Given that the underlying one-sample HLE is derived from a vector of pairwise averages and its center corresponds to the pseudo-median, we suggest reporting the following complementary descriptive statistics: the pseudo-median and the pseudo-interquartile range (pseudo-IQR). Similarly, a “HLE-box” plot can be constructed to visualize group distributions, adhering to the conventions of a standard boxplot but using the pseudo-median and pseudo-IQR.

This descriptive approach will align DAA HLE-CI results with statistical summaries of underlying samples, which is particularly valuable for interpreting microbial data.

4. Discussion

Microbial data present a unique analytical challenge [2–8]. Taxonomic profiles can span thousands of features, while the true signal is often extremely subtle [9–13]. The high dimensionality of these data introduces a severe multiple comparisons problem, necessitating robust hypothesis filtration [7,8]. Standard multiple comparison adjustment methods, commonly applied in microbiome studies, were designed for scenarios with a substantial proportion of true positives [23]. When applied to microbial data, these methods can over-penalize potential markers, often to the point of prohibiting any detection. This occurs because multiple testing corrections inherently reduce statistical power. Furthermore, conventional statistical tests (e.g., the T-test or Mann–Whitney U), even when applicable to NB-distributed microbial data, can produce inflated false positive rates [7,8], further destabilizing the precision-recall rates of biomarker discovery.

The results of this study demonstrate that microbial data, which can be modeled by a negative binomial (NB) distribution, can be most robustly characterized using the Hodges–Lehmann estimator (HLE). The HLE estimate lies between the median and the mean, effectively accounting for the right-hand skew of the NB distribution while mitigating the influence of extreme outliers. In both synthetic and experimental conditions, two-sample HLE-based tests demonstrated superior precision and recall rates for biomarker discovery in microbiome data. Consequently, the HLE-CI method achieved significantly higher F1 scores, an advantage that persisted even with decreasing sample sizes. These results demonstrate that HLE-CI-based hypothesis filtration can offer distinct advantages for biomarker discovery and microbiome research.

Author Contributions: Conceptualization, E.V.; methodology, E.V.; software, E.V.; validation, E.V., A.K. and S.K.; formal analysis, E.V.; investigation, E.V.; resources, M.B.; data curation, E.V.; writing—original draft preparation, E.V.; writing—review and editing, E.V., A.K. and S.K.; visualization, E.V.; supervision, M.B.; project administration, M.B.; funding acquisition, M.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Ministry of Science and Higher Education of the Russian Federation, (Agreement 075-10-2025-017 from 27 February 2025).

Institutional Review Board Statement: The data collection and study was conducted in accordance with the Declaration of Helsinki and approved by the Ethics Committee of National Laboratory Astana, Nazarbayev University (protocol No. 05-2022, date of approval 21 October 2022).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: All data and scripts required to reproduce the analysis presented in this work are available at <https://github.com/VeaLi/hle-ci-for-microbial-daa> (accessed on 10 December 2025). Raw sequencing data from this study have been deposited in the National Center

for Biotechnology Information (NCBI) BioProject Sequence Read Archive under accession number PRJNA1287050.

Acknowledgments: The first author would like to thank researcher Zharkyn Jarmukhanov of National Laboratory Astana for providing technical assistance with the raw sequencing data, and research associate Dmitry Onishchenko from Sirius University of Science and technology for valuable feedback.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CI	Confidence interval
DAA	Differential abundance analysis
HLE	Hodges–Lehmann estimator
NB	Negative binomial distribution

References

- Hou, K.; Wu, Z.-X.; Chen, X.-Y.; Wang, J.-Q.; Zhang, D.; Xiao, C.; Zhu, D.; Koya, J.B.; Wei, L.; Li, J.; et al. Microbiota in Health and Diseases. *Signal Transduct. Target. Ther.* **2022**, *7*, 135. [\[CrossRef\]](#)
- Segata, N.; Izard, J.; Waldron, L.; Gevers, D.; Miropolsky, L.; Garrett, W.S.; Huttenhower, C. Metagenomic Biomarker Discovery and Explanation. *Genome Biol.* **2011**, *12*, R60. [\[CrossRef\]](#) [\[PubMed\]](#)
- Fernandes, A.D.; Reid, J.N.; Macklaim, J.M.; McMurrough, T.A.; Edgell, D.R.; Gloor, G.B. Unifying the Analysis of High-Throughput Sequencing Datasets: Characterizing RNA-Seq, 16S rRNA Gene Sequencing and Selective Growth Experiments by Compositional Data Analysis. *Microbiome* **2014**, *2*, 15. [\[CrossRef\]](#)
- Mallick, H.; Rahnavard, A.; McIver, L.J.; Ma, S.; Zhang, Y.; Nguyen, L.H.; Tickle, T.L.; Weingart, G.; Ren, B.; Schwager, E.H.; et al. Multivariable Association Discovery in Population-Scale Meta-Omics Studies. *PLoS Comput. Biol.* **2021**, *17*, e1009442. [\[CrossRef\]](#)
- Lin, H.; Peddada, S.D. Multigroup Analysis of Compositions of Microbiomes with Covariate Adjustments and Repeated Measures. *Nat. Methods* **2024**, *21*, 83–91. [\[CrossRef\]](#) [\[PubMed\]](#)
- Nickols, W.A.; Kuntz, T.; Shen, J.; Maharjan, S.; Mallick, H.; Franzosa, E.A.; Thompson, K.N.; Nearing, J.T.; Huttenhower, C. MaAsLin 3: Refining and Extending Generalized Multivariable Linear Models for Meta-Omic Association Discovery. *bioRxiv* **2024**. [\[CrossRef\]](#)
- Nearing, J.T.; Douglas, G.M.; Hayes, M.G.; MacDonald, J.; Desai, D.K.; Allward, N.; Jones, C.M.A.; Wright, R.J.; Dhanani, A.S.; Comeau, A.M.; et al. Microbiome Differential Abundance Methods Produce Different Results across 38 Datasets. *Nat. Commun.* **2022**, *13*, 342. [\[CrossRef\]](#)
- Lin, H.; Peddada, S.D. Analysis of Microbial Compositions: A Review of Normalization and Differential Abundance Analysis. *npj Biofilm. Microbiomes* **2020**, *6*, 60. [\[CrossRef\]](#)
- Sun, W.; Zhang, Y.; Guo, R.; Sha, S.; Chen, C.; Ullah, H.; Zhang, Y.; Ma, J.; You, W.; Meng, J.; et al. A Population-Scale Analysis of 36 Gut Microbiome Studies Reveals Universal Species Signatures for Common Diseases. *npj Biofilm. Microbiomes* **2024**, *10*, 96. [\[CrossRef\]](#) [\[PubMed\]](#)
- San-Martin, M.I.; Chamizo-Ampudia, A.; Sanchiz, Á.; Ferrero, M.Á.; Martínez-Blanco, H.; Rodríguez-Aparicio, L.B.; Navasa, N. Microbiome Markers in Gastrointestinal Disorders: Inflammatory Bowel Disease, Colorectal Cancer, and Celiac Disease. *Int. J. Mol. Sci.* **2025**, *26*, 4818. [\[CrossRef\]](#)
- Clooney, A.G.; Eckenberger, J.; Laserna-Mendieta, E.; Sexton, K.A.; Bernstein, M.T.; Vagianos, K.; Sargent, M.; Ryan, F.J.; Moran, C.; Sheehan, D.; et al. Ranking Microbiome Variance in Inflammatory Bowel Disease: A Large Longitudinal Intercontinental Study. *Gut* **2021**, *70*, 499–510. [\[CrossRef\]](#) [\[PubMed\]](#)
- Scepanovic, P.; Hodel, F.; Mondot, S.; Partula, V.; Byrd, A.; Hammer, C.; Alanio, C.; Bergstedt, J.; Patin, E.; Touvier, M.; et al. A Comprehensive Assessment of Demographic, Environmental, and Host Genetic Associations with Gut Microbiome Diversity in Healthy Individuals. *Microbiome* **2019**, *7*, 130. [\[CrossRef\]](#)
- Nearing, J.T.; DeClercq, V.; Van Limbergen, J.; Langille, M.G.I. Assessing the Variation within the Oral Microbiome of Healthy Adults. *mSphere* **2020**, *5*, e00451-20. [\[CrossRef\]](#) [\[PubMed\]](#)
- Lee, S.; Meslier, V.; Bidkhori, G.; Garcia-Guevara, F.; Etienne-Mesmin, L.; Clasen, F.; Park, J.; Plaza Oñate, F.; Cai, H.; Le Chatelier, E.; et al. Transient Colonizing Microbes Promote Gut Dysbiosis and Functional Impairment. *npj Biofilm. Microbiomes* **2024**, *10*, 80. [\[CrossRef\]](#)

15. Loftus, M.; Hassounah, S.A.-D.; Yooseph, S. Bacterial Associations in the Healthy Human Gut Microbiome across Populations. *Sci. Rep.* **2021**, *11*, 2828. [[CrossRef](#)]
16. Olsson, L.M.; Boulund, F.; Nilsson, S.; Khan, M.T.; Gummesson, A.; Fagerberg, L.; Engstrand, L.; Perkins, R.; Uhlén, M.; Bergström, G.; et al. Dynamics of the Normal Gut Microbiota: A Longitudinal One-Year Population Study in Sweden. *Cell Host Microbe* **2022**, *30*, 726–739.e3. [[CrossRef](#)] [[PubMed](#)]
17. Shen, Y.; Fan, N.; Ma, S.; Cheng, X.; Yang, X.; Wang, G. Gut Microbiota Dysbiosis: Pathogenesis, Diseases, Prevention, and Therapy. *MedComm* **2025**, *6*, e70168. [[CrossRef](#)]
18. Azzolino, D.; Carnevale-Schianca, M.; Santacroce, L.; Colella, M.; Felicetti, A.; Terranova, L.; Castrejón-Pérez, R.C.; Garcia-Godoy, F.; Lucchi, T.; Passarelli, P.C. The Oral–Gut Microbiota Axis Across the Lifespan: New Insights on a Forgotten Interaction. *Nutrients* **2025**, *17*, 2538. [[CrossRef](#)]
19. Hajishengallis, G.; Liang, S.; Payne, M.A.; Hashim, A.; Jotwani, R.; Eskan, M.A.; McIntosh, M.L.; Alsam, A.; Kirkwood, K.L.; Lambris, J.D.; et al. A Low-Abundance Biofilm Species Orchestrates Inflammatory Periodontal Disease through the Commensal Microbiota and the Complement Pathway. *Cell Host Microbe* **2011**, *10*, 497–506. [[CrossRef](#)]
20. Baker, J.L.; Faustoferri, R.C.; Quivey, R.G. Acid-Adaptive Mechanisms of *Streptococcus Mutans*—the More We Know, the More We Don't. *Mol. Oral Microbiol.* **2017**, *32*, 107–117. [[CrossRef](#)]
21. Yamazaki, K.; Kamada, N. Exploring the Oral-Gut Linkage: Interrelationship Between Oral and Systemic Diseases. *Mucosal Immunol.* **2024**, *17*, 147–153. [[CrossRef](#)]
22. Li, B.; Selmi, C.; Tang, R.; Gershwin, M.E.; Ma, X. The Microbiome and Autoimmunity: A Paradigm from the Gut–Liver Axis. *Cell. Mol. Immunol.* **2018**, *15*, 595–609. [[CrossRef](#)]
23. Rothman, K.J. No Adjustments Are Needed for Multiple Comparisons. *Epidemiol. Camb. Mass* **1990**, *1*, 43–46. [[CrossRef](#)]
24. Greenland, S.; Senn, S.J.; Rothman, K.J.; Carlin, J.B.; Poole, C.; Goodman, S.N.; Altman, D.G. Statistical Tests, P Values, Confidence Intervals, and Power: A Guide to Misinterpretations. *Eur. J. Epidemiol.* **2016**, *31*, 337–350. [[CrossRef](#)]
25. Gardner, M.J.; Altman, D.G. Confidence Intervals Rather than P Values: Estimation Rather than Hypothesis Testing. *Br. Med. J. Clin. Res. Ed* **1986**, *292*, 746–750. [[CrossRef](#)] [[PubMed](#)]
26. Zhang, X.; Mallick, H.; Tang, Z.; Zhang, L.; Cui, X.; Benson, A.K.; Yi, N. Negative Binomial Mixed Models for Analyzing Microbiome Count Data. *BMC Bioinform.* **2017**, *18*, 4. [[CrossRef](#)] [[PubMed](#)]
27. Li, G.; Yang, L.; Chen, J.; Zhang, X. Robust Differential Abundance Analysis of Microbiome Sequencing Data. *Genes* **2023**, *14*, 2000. [[CrossRef](#)]
28. Hodges, J.L.; Lehmann, E.L. Estimates of Location Based on Rank Tests. *Ann. Math. Stat.* **1963**, *34*, 598–611. [[CrossRef](#)]
29. Rosenkranz, G.K. A Note on the Hodges–Lehmann Estimator. *Pharm. Stat.* **2010**, *9*, 162–167. [[CrossRef](#)]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.