أكاديمية كاوست
KAUST ACADEMY

جامعة الملك عبدالله للعلوم والتقنية
King Abdullah University of Science and Technology

Counting Studies vs. Combining Them: Re-Analyzing 28 Gut-Microbiome Studies

Raiyan Subedar, Hugo Tavares

rayan.subedarLive@gmail.com, KAUST Academy Bioinformatics Program

Abstract

Duvallet et al. (2017) pooled 28 case–control gut microbiome studies and reported 51 bacterial genera forming a shared cross-disease response. That conclusion rests on *vote counting*: tallying how many studies cross a significance threshold. Vote counting discards effect magnitude and cannot distinguish studies that agree from studies that contradict one another. We first reproduce the original analysis exactly—all 30 case–control comparisons match the published q-values to 10^{-6}, and both published genus lists reproduce at 100%—then re-analyse the same data using random-effects meta-analysis of standardised effect sizes. Pooling identifies 43 shared genera against the paper's 51, and reports a quantity the original method cannot: between-study heterogeneity. Median I^2 is 73% across all genera and 84% among those the paper designates as shared, meaning 49 of its 51 shared genera rest on studies that substantially disagree. A sensitivity analysis shows only 15 of our 43 survive once two datasets sharing control samples with two others are removed, a loss far larger than removing any random pair. In parallel, a global cross-cohort classifier is weak (mean AUC 0.599) and essentially blind to disease-specific information (removing an entire disease changes it by −0.017), yet where a real signal exists a specialist recovers it: a CRC model reaches 0.763 on a cohort no model has seen, against 0.653 for the general model. HIV transfers below chance because its cohort effect sizes are negatively correlated ($r = -0.42$); we trace this to a lifestyle confounder, with *Bacteroides* differing by risk group at $g = -1.84$ among HIV-negative participants—larger than, and opposite to, the apparent HIV effect. From healthy control samples alone, a classifier identifies the 16S region used with 76% accuracy against a 44% baseline. We also document an unreported edge case in the original Stouffer implementation that inverts one genus's direction. The published findings are not wrong, but the evidence supporting them is weaker than a count of significant studies implies.

Date: August 19, 2026

Code: https://github.com/YOURUSERNAME/microbiome-meta-reanalysis

Contents

1 Introduction **2**

1.1 Motivation . . . 2
1.2 Problem statement . . . 2
1.3 Contributions . . . 2
1.4 Assumptions and scope . . . 3

2 Related Work **3**

2.1 The original study . . . 3
2.2 Compositional data and normalisation . . . 3
2.3 Differential abundance methods disagree . . . 3
2.4 Meta-analysis methodology . . . 3

3 Methodology **4**

3.1 Data . . . 4
3.2 Reproduction pipeline . . . 4
3.3 Effect-size meta-analysis . . . 4
3.4 Classification . . . 5
3.5 Sensitivity analyses . . . 5

4 Results **5**

4.1 The reproduction is exact . . . 6
4.2 An error in the published Stouffer implementation . . . 6
4.3 Pooling effect sizes gives a different answer . . . 6
4.4 The studies substantially disagree . . . 6
4.5 Study agreement varies sharply by disease . . . 7
4.6 The HIV inversion is a lifestyle confounder, not the virus . . . 7
4.7 A general classifier exists but is weak . . . 8
4.8 Where real biology exists, a specialist finds it . . . 8
4.9 Laboratory protocol is strongly detectable . . . 9

5 Discussion **12**

5.1 What the heterogeneity result does and does not show . . . 12
5.2 Two methods, two answers, from one author . . . 12
5.3 Effect size, not consistency, predicts transferability . . . 12
5.4 The HIV inversion . . . 12
5.5 Confounding between study and protocol . . . 13
5.6 Non-independence as a shared problem . . . 13

6 Limitations and Future Work **13**

6.1 Limitations . . . 13
6.2 Future work . . . 14

7 Conclusion **14**

1 Introduction

1.1 Motivation

The human gut contains trillions of bacteria, collectively the gut microbiome, and its composition shifts in many diseases. Over the past fifteen years dozens of research groups have investigated this the same way: recruit patients with some condition, recruit healthy controls, sequence the bacterial 16S rRNA gene from stool samples, and ask which bacteria differ between the groups.

The difficulty is that individual studies are small and disagree. One cohort might enrol 40 patients in France, another 300 in China, with different diets, different DNA extraction kits, different sequencing platforms and different regions of the 16S gene amplified. The field accumulated many individual results and little consensus.

duvallet2017 addressed this by reprocessing 28 published case–control studies through a single uniform pipeline—identical OTU definitions, identical reference database—and asking whether some bacteria shift in *every* disease: a general signature of ill health rather than a disease-specific one. The paper reports 51 such genera and is widely cited.

1.2 Problem statement

The paper's conclusion depends on how the per-study results are combined. For each genus it asked, study by study, whether the difference between patients and controls reached $q < 0.05$, and then counted the studies that did. A genus significant in at least two *different diseases* was designated part of the shared response. This procedure is known as **vote counting**.

Vote counting discards information in two specific ways.

First, it cannot detect consistent weak evidence. Four studies each finding the same modest effect, each individually short of significance, cast zero votes between them and the genus never appears in the results. Combined properly the same data can be clearly significant. The bias runs against small cohorts, which is most of this dataset.

Second, and more importantly for a claim about a *shared* response, it cannot distinguish agreement from contradiction. Two genera both "significant in two studies" are treated identically whether their studies agree closely or point in opposite directions. Whether the underlying studies agree is precisely what determines how much a shared-response claim is worth.

1.3 Contributions

- **An exact reproduction** of the original analysis. The published code targets Python 2.7 with dependencies that no longer install, so the pipeline was reimplemented from source as a specification. All 30 case–control comparisons match the published q-values, and both published genus lists reproduce at 100%.
- **A random-effects re-analysis** that pools standardised effect sizes rather than counting significant studies, and reports between-study heterogeneity (I^2)—a quantity absent from every published analysis of this dataset.
- **A demonstrated non-independence problem** in the source data: two datasets share control samples with two others, and removing them costs far more significant genera than removing any random pair.
- **Cross-study classification**, testing whether a model trained on some cohorts generalises to an unseen one—a question the original work never asked—together with a study-of-origin control that quantifies how strongly laboratory protocol is detectable in healthy samples.
- **An unreported implementation error** in the original Stouffer code, which converts a

zero-effect q-value into a p-value of exactly zero and inverts at least one genus's reported direction.

1.4 Assumptions and scope

This is a re-analysis of existing data, not a new method and not new data collection. It inherits the MicrobiomeHD curators' processing decisions: OTU clustering, taxonomic assignment and label harmonisation are taken as given. Analysis is restricted to the 28 folders and 30 case–control comparisons in the original analysis set, so that differences in conclusion are attributable to the combining method rather than to a different choice of studies.

2 Related Work

2.1 The original study

duvallet2017 assembled the MicrobiomeHD database and analysed 28 case–control 16S studies spanning ten diseases. Its central methodological choice is to analyse each study separately and combine afterwards. Per genus per study it applies a Kruskal–Wallis test with Benjamini–Hochberg correction within that study, then designates genera disease-specific when significant in ≥ 2 studies of one disease, and shared when significant in ≥ 2 different diseases.

Less prominently, the accompanying code also implements Stouffer's weighted z-score method, combining per-study p-values weighted by $\sqrt{n}$. This is a stronger combining rule than counting and is reproduced here as the comparison baseline, since a re-analysis that only outperformed vote counting would be answering the weaker challenge. Its output was never published as a supplementary file.

Neither method uses effect magnitude, and neither yields a heterogeneity estimate.

2.2 Compositional data and normalisation

16S abundances are compositional: they are proportions constrained to sum to one, so an increase in one taxon forces apparent decreases elsewhere. gloor2017 argue this invalidates analyses treating relative abundances as independent quantities, and recommend log-ratio transformations. The centered log-ratio (CLR) used here follows that recommendation. The original paper used relative abundance without rarefaction; we reproduce it on that scale and depart deliberately for the effect-size estimation, reporting the consequence.

2.3 Differential abundance methods disagree

nearing2022 compared many differential abundance methods across 38 datasets and found they frequently identify different taxa from identical input. That result is the present study's concern raised one level: if method choice changes which taxa are called significant within a study, it should also change which are called shared across studies. This work tests that for one influential published claim.

2.4 Meta-analysis methodology

Combining effect sizes rather than significance is standard in clinical evidence synthesis. Under a random-effects model each study has its own true effect drawn from a distribution, with between-study variance τ^2; I^2 expresses the share of observed variation attributable to real disagreement rather than sampling error higgins2002. veroniki2016 review estimators for τ^2 and recommend Paule–Mandel or REML over the older DerSimonian–Laird, which underestimates

heterogeneity when few studies are available—the regime here, with three to five cohorts per disease.

Microbiome-specific tools have since appeared, notably MMUPHin ma2022mmuphin, which performs batch adjustment and meta-analysis jointly. It was not run here; see Section 6.

3 Methodology

3.1 Data

The MicrobiomeHD database supplies 33 study folders; the original analysis set comprises 28. Two folders each yield two case–control comparisons—Schubert's cohort contributes both *C. difficile* and non-*C. difficile* diarrhoea against the same controls, and Zhu's contributes both NASH and obesity—giving 30 comparisons in total. Both figures are correct and are reported together throughout.

After quality filtering and label assignment, 4,026 samples remain (2,210 cases, 1,816 controls), with cohorts ranging from 23 to 613 samples. Diseases number 11 rather than 13, because the original code folds both diarrhoeal datasets into a single `cdi` group before analysis.

3.2 Reproduction pipeline

The published code targets Python 2.7 with pinned versions of pandas, scipy, statsmodels and scikit-bio that cannot be installed on any current Python distribution. It was therefore read as a written specification and reimplemented in Python 3.10, with the published outputs as the check.

Per study, in order:

1. load the RDP-assigned OTU table and metadata; transpose to samples-in-rows;
2. apply the per-dataset sample filter given in the analysis manifest;
3. retain samples possessing both 16S data and metadata;
4. drop samples with ≤ 100 total reads;
5. drop OTUs with < 10 reads in total;
6. drop OTUs present in $\leq 1\%$ of samples;
7. re-check sample read depth;
8. convert to relative abundance (no rarefaction appears anywhere in the original code);
9. collapse to genus, discarding taxa without a genus assignment;
10. assign case/control status from the harmonised `DiseaseState` field;
11. Kruskal–Wallis test per genus;
12. Benjamini–Hochberg correction within the study;
13. sign the q-value by mean(cases) − mean(controls).

Filtering occurs at OTU level *before* the genus collapse; reordering these steps changes the results. Two undocumented details were required: one dataset stores numeric sample identifiers, which pandas reads as integers against a string OTU index, silently discarding every sample; and three datasets require label rewriting before case assignment, which is the mechanism separating the two comparisons drawn from each shared folder.

3.3 Effect-size meta-analysis

The re-analysis branches from step 9. Genus-level *counts* receive a pseudocount of 0.5 and a centered log-ratio transform; counts rather than relative abundance are used because CLR is scale-invariant whereas an additive pseudocount is not. For each genus in each study we record the mean,

standard deviation and sample size in both groups, from which `statsmodels.effectsize_smd` yields Hedges' g and its variance. Genera are pooled where measured in at least three studies of the relevant group, a threshold fixed in advance.

Pooling uses a random-effects model with the Paule–Mandel estimator, chosen over DerSimonian–Laird for its behaviour with few studies veroniki2016; DerSimonian–Laird is reported alongside as a robustness check. Both I^2 and τ^2 are truncated at zero as convention requires. Pooling is performed twice: within each disease, answering which genera associate with a specific condition; and across all diseases, which is the comparison against the original shared-response claim.

3.4 Classification

Genus-level relative abundances form the feature matrix. The per-study baseline reproduces the original protocol: a random forest under five-fold stratified cross-validation, scored by AUC.

All settings for the transfer analysis are fixed in a single configuration file so that every figure and table draws on identical choices; earlier versions kept separate dataset lists that drifted apart and reported different AUCs for the same disease. Five datasets are excluded and three reserved. Two studies (`ibd_papa`, `ibd_gevers`) have `nonIBD` rather than healthy controls; one (`cdi_youngster`) has only four controls, which returned a spurious AUC of 1.000; two (`noncdi_schubert`, `ob_zhu`) share control samples with another dataset, so holding one out still leaves its controls in training. Three cohorts present in MicrobiomeHD but never analysed by the original paper (`crc_zackular`, `ob_jumpertz`, `ob_escobar`) are held out untouched as a validation set. This leaves 25 training datasets and 3,467 samples (1,865 sick, 1,602 healthy).

Rather than one model per disease, a single *global* sick-versus-healthy random forest (300 trees, fixed seed) is trained across all diseases. A genus enters the feature set only if it is measured in at least 80% of training datasets, leaving 39 genera; keeping only genera common to every study leaves too few, while zero-filling the union lets a model identify a sample's source study from which genera are missing. Three evaluations follow. *Leave one dataset out* trains on all datasets but one and tests on the held-out one, repeated for each. *Leave one disease out* removes every cohort of a disease at once, testing whether the model relies on disease-specific information. *Held-out validation* trains on the four analysed CRC cohorts and tests a CRC-specialist model against the general model on the reserved `crc_zackular` cohort, which no model has seen. Random forest is compared against L_2-regularised logistic regression and gradient boosting.

A study-of-origin classifier provides the control. Trained on *control samples only*—so that disease cannot serve as a proxy—it predicts which sequencing platform, 16S region and extraction protocol a sample came from.

3.5 Sensitivity analyses

Four checks were run. Zero replacement was varied between an additive pseudocount and multiplicative replacement. The τ^2 estimator was varied between Paule–Mandel and DerSimonian–Laird. The two datasets sharing control samples with other datasets were removed and the analysis repeated. Because removing data reduces power regardless of the reason, that last check is accompanied by a control: fifteen trials removing two *randomly chosen* datasets, establishing what an ordinary loss of two datasets costs.

4 Results

4.1 The reproduction is exact

Reproduction quality sets the ceiling on everything that follows: a difference in conclusions is only interpretable if the baseline is known to be correct.

All 30 case–control comparisons reproduce the published per-study q-values, agreeing to a relative tolerance of 10^{-6} with a maximum absolute difference of 5.0×10^{-13}—the precision at which the published file was written, not a computational disagreement. Applying the original counting rules to those q-values reproduces both published genus lists exactly, including the headline 24 health-associated, 20 disease-associated and 7 mixed genera.

Table 1 Reproduction of the published analysis. Exact agreement across every check.

Check	Result	Target
Per-study q-values	30 / 30 datasets	exact
Agreement at 10^{-6}	100%	—
Disease-specific genera	100%	published list
Shared genera	100%	24 / 20 / 7

4.2 An error in the published Stouffer implementation

The original Stouffer script converts signed two-sided q-values into one-sided p-values. When a study's effect is exactly zero the stored q-value is exactly zero, and the conversion returns a one-sided p-value of zero, giving an infinite z-score that dominates the weighted sum. Eleven cells across ten genera are affected. At least one genus, *Porphyromonas*, is reported in the opposite direction as a consequence: five studies place it higher in disease, but a single zero-effect cell drives the combined result to health-associated. The corresponding vote-counting code guards against the same case explicitly; the Stouffer script does not.

4.3 Pooling effect sizes gives a different answer

Figure 1 illustrates the mechanism. Random-effects pooling identifies 43 shared genera against the original 51, with substantial but incomplete overlap: 21 genera are called shared by both methods, 30 by counting alone and 22 by pooling alone.

Table 2 Genera identified as a shared cross-disease response, by combining method. All three operate on identical per-study results; the differences are entirely methodological.

Method	Shared genera	Breakdown
Vote counting (original)	51	24 health, 20 disease, 7 mixed
Stouffer (original)	111	—
Random-effects pooling (ours)	43	12 health, 31 disease

4.4 The studies substantially disagree

The quantity unavailable to either original method is heterogeneity. Median I^2 is 73.4% across all pooled genera, indicating that most of the variation between studies is real disagreement rather than sampling noise.

Critically, heterogeneity is *higher* among the genera the original paper designates as shared: median I^2 of 84.0% versus 68.2% for the remainder, with 49 of the 51 exceeding 50% (Figure 2).

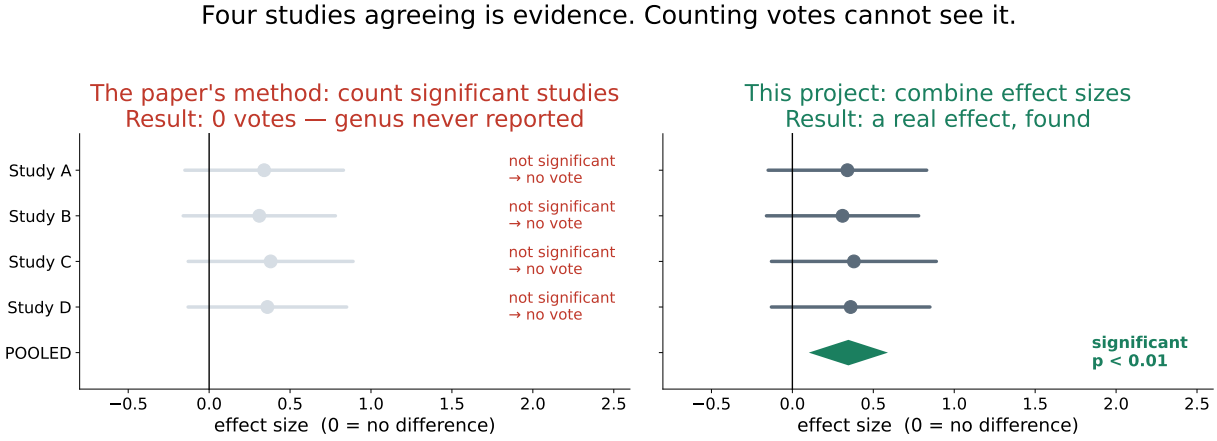


Figure 1 Why counting significant studies loses information. Four studies find the same modest effect; each is individually non-significant, so vote counting records zero votes and the genus is never reported. Pooling the same effect sizes yields a clearly significant result.

This follows from the counting rule itself—a genus qualifies by reaching significance in two different diseases, which is most easily achieved when results are inconsistent.

High heterogeneity is not an artefact of small study counts: the correlation between I^2 and the number of contributing studies is 0.081.

4.5 Study agreement varies sharply by disease

Comparing effect sizes between cohorts of the same disease (Figure 3) shows CDI, CRC and obesity cohorts positively correlated (r = +0.39, +0.30, +0.27), while HIV cohorts are *negatively* correlated ($r = -0.19$ averaged pairwise, reaching −0.42 between the two largest). The HIV studies find opposite effects for the same bacteria.

Figure 4 shows the HIV case in detail. Were the two cohorts in agreement, points would lie along the rising diagonal; instead 65% of genera are assigned opposite directions by the two studies.

4.6 The HIV inversion is a lifestyle confounder, not the virus

The genera that flip direction between HIV cohorts—*Prevotella*, *Alloprevotella*, *Catenibacterium*—are markers of an enterotype that tracks diet and sexual-behaviour risk group rather than HIV itself. Testing this directly, among HIV-*negative* participants alone (on the original paper's sample set), *Bacteroides* differs between risk groups at Hedges' $g = -1.84$ ($p = 0.0001$) and *Prevotella* at $g = +1.13$ ($p < 0.001$). The apparent effect of HIV status on the same genera is smaller and mostly non-significant—*Prevotella* $g = -0.43$ ($p = 0.06$)—and, for *Bacteroides*, points the opposite way ($g = +0.77$). Because cohorts enrolled different proportions of these risk groups, the "HIV signal" points in different directions in different studies. What the classifiers detect is the confounder, which is why a model trained on two cohorts reverses on the third.

Two datasets share control samples with two others: the Schubert pair share 154 controls and the Zhu pair share 16. Meta-analysis assumes independence.

Removing the duplicates reduces significant genera from 43 to 15, while effect sizes themselves barely change (correlation 0.983)—the signature of duplicated evidence inflating apparent precision rather than altering estimates. Fifteen control trials removing two randomly chosen datasets instead yielded a median of 41 significant genera (range 21–47), and no trial fell to 15 or below. The loss is specific to the non-independence, not to the reduction in data.

Table 3 Sensitivity to non-independent datasets, with a random-removal control establishing the ordinary cost of losing two datasets.

Analysis	Datasets	Significant genera
Main analysis	27	43
Duplicated pair removed	25	15
Random pair removed (median of 15)	25	41

4.7 A general classifier exists but is weak

The per-study random forest baseline reproduces the published AUCs with correlation 0.933 and mean absolute difference 0.038. Exact agreement is unattainable for a stochastic method across a decade of library versions; disagreements concentrate in the smallest cohorts and where the published AUC was already near chance.

The single global sick-versus-healthy model achieves a mean AUC of only **0.599** across the 25 held-out datasets. Removing an entire disease from training changes performance by only −0.017 on average (Table 4)—not statistically distinguishable from zero—so the model was never relying on disease-specific information: it works where any disrupted gut is easy to spot and fails otherwise. *C. difficile* infection, the most disturbed community in the data, is the clear exception, transferring at AUC 0.73–0.87; HIV falls below chance.

Table 4 Global-model transfer by disease. *Leave-one-dataset-out* tests on an unseen cohort of the same disease; *leave-one-disease-out* removes every cohort of that disease from training. A value of 0.5 is chance.

Disease	Leave-dataset-out	Leave-disease-out	Change
CDI	0.759	0.733	−0.026
IBD	0.662	0.708	+0.046
Obesity	0.574	0.608	+0.035
CRC	0.535	0.505	−0.030
HIV	0.609	0.489	−0.120
Mean (all)	**0.599**		−0.017

No model family rescues performance: random forest averages 0.599, gradient boosting 0.566 and logistic regression 0.550. Substantially different algorithms occupy a narrow band, indicating the limiting factor is the data rather than the model, and restricting the features to the eleven meta-analysis genera does not help (0.583).

4.8 Where real biology exists, a specialist finds it

The weakness of the general model is not the whole story. On the reserved `crc_zackular` cohort—60 people from a study the original paper never analysed and no model here has seen—a CRC-specialist random forest reaches AUC **0.763** (95% CI 0.63–0.88), while the general model manages only 0.653 on the same people. Colorectal cancer carries a genuine, transferable signature, driven by the oral-associated genera raised in tumour tissue, and a model built for it recovers that signature where a one-size model dilutes it. The two clouds of predicted scores overlap substantially: this is a population-level association, not a diagnostic test.

4.9 Laboratory protocol is strongly detectable

Trained on healthy control samples alone, where no disease is present to serve as a proxy, a classifier identifies the 16S region used with 76.3% accuracy across nine classes against a 43.8% baseline, and the sequencing platform with 81.3% against 53.6%. Technical provenance is at least as detectable in these data as disease status, which bounds how much of any cross-study signal could be biological.

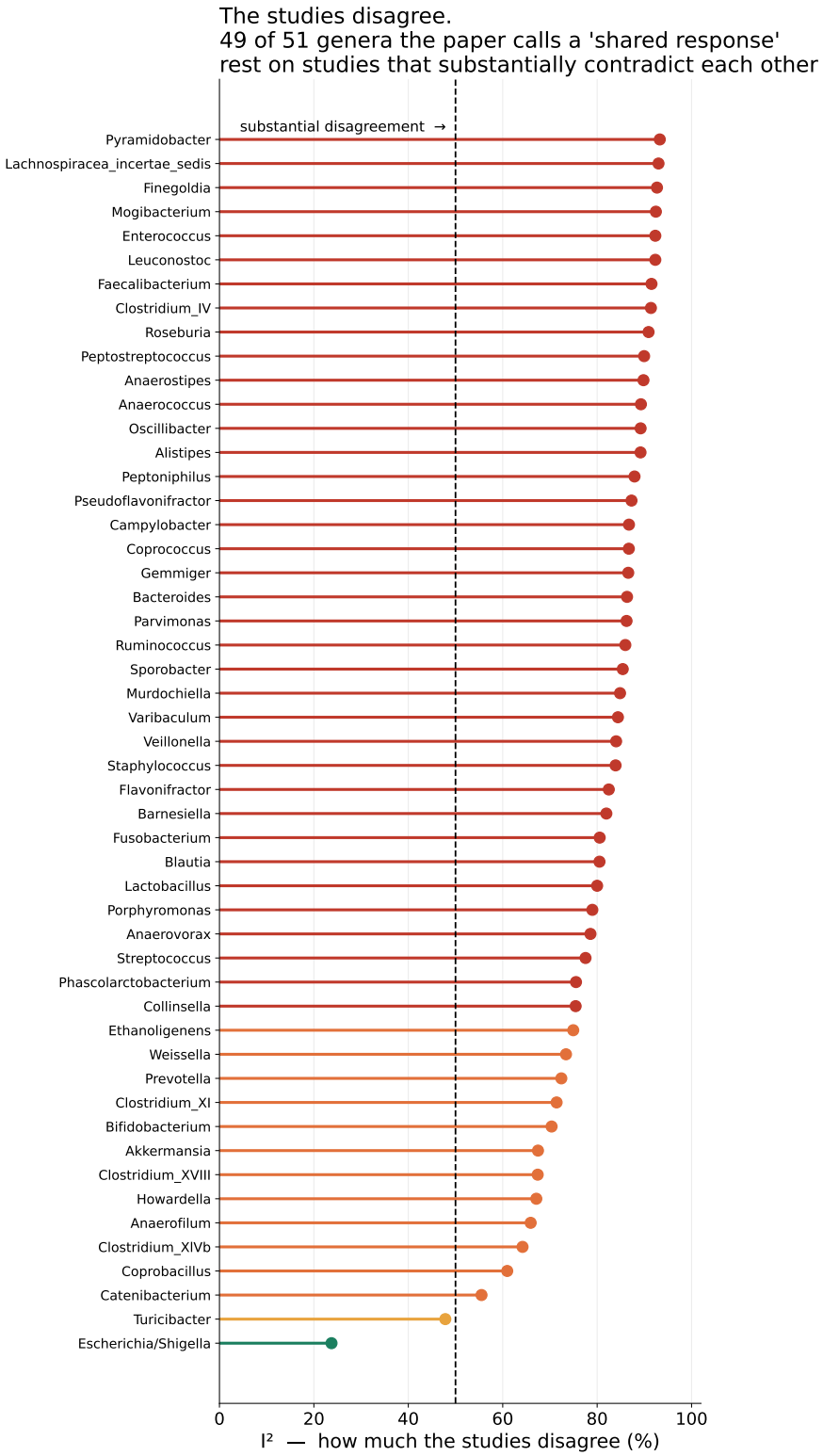


Figure 2 Every genus the original paper designates as a shared cross-disease response, ranked by between-study heterogeneity. Only *Escherichia/Shigella* shows low disagreement.

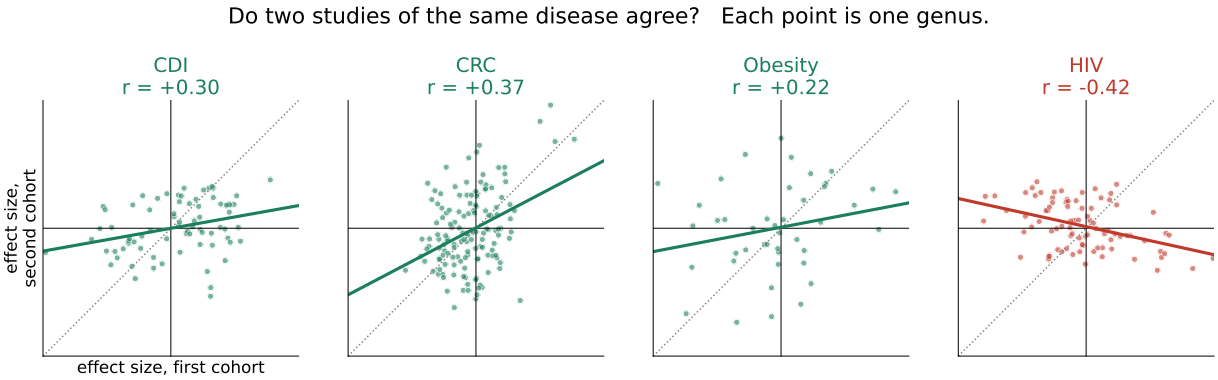


Figure 3 Effect-size agreement between two cohorts of the same disease, one point per genus. Three diseases show positive agreement; HIV is inverted.

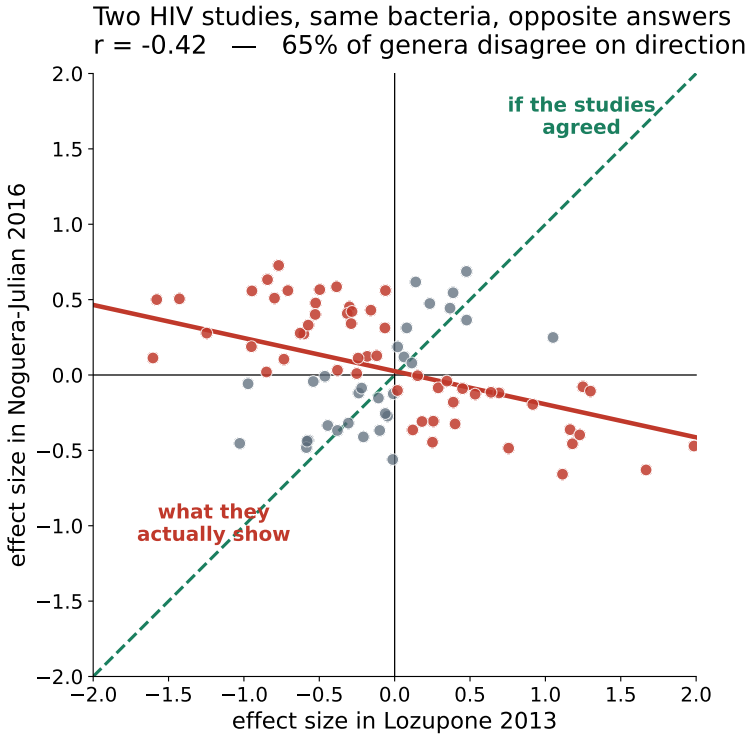


Figure 4 Effect sizes for the same genera in the two largest HIV cohorts. The dashed line marks where points would fall if the studies agreed; the solid line is the observed trend.

5 Discussion

5.1 What the heterogeneity result does and does not show

Median I^2 of 73% across all pooled genera means high heterogeneity is a property of this dataset generally, not a defect peculiar to the original paper's conclusions. It would be wrong to claim its findings are uniquely fragile.

The defensible claim is narrower and more interesting. The genera designated as a shared cross-disease response are the *most* heterogeneous subset (84.0% against 68.2%). That is a predictable consequence of the selection rule rather than a coincidence: requiring significance in two different diseases preferentially selects genera behaving inconsistently, because inconsistent behaviour across diverse conditions is what makes a genus significant in several of them for unrelated reasons.

High heterogeneity does not make the reported genera false. Populations, diets, and laboratory protocols genuinely differ, and real biological effects can vary in magnitude between settings. What it establishes is that a list assembled by counting significant studies conveys more confidence than the underlying evidence supports, and that the original method could not have revealed this.

5.2 Two methods, two answers, from one author

The comparison that requires least interpretation is internal to the original work. Vote counting and Stouffer's method are both implemented in the published code, both operate on identical per-study results, and they identify 51 and 111 genera respectively—disagreeing on direction for seven. Before effect sizes enter the discussion at all, the choice of combining rule already changes the answer.

This mirrors, one level up, the finding of nearing2022 that differential abundance methods disagree across datasets. Here the disagreement is between methods for *combining* studies rather than for analysing one.

5.3 Effect size, not consistency, predicts transferability

The classification results complicate the expectation drawn from Phase 4. Heterogeneity above 70% suggested models should transfer badly, and the general model is indeed weak—mean AUC 0.599, and removing an entire disease barely moves it (−0.017). But the pattern across diseases does not follow I^2. CDI has the highest heterogeneity and the best transfer (0.73–0.87); HIV has lower heterogeneity and fails below chance. And where a genuine disease-specific signal exists, a specialist recovers it: the CRC-specialist model reaches 0.763 on a wholly unseen cohort against the general model's 0.653.

What separates the diseases is effect magnitude. CDI yielded 69 significant genera under pooling; HIV yielded none. *C. difficile* infection is an acute condition that disrupts the gut community severely, so its signal survives a change of cohort even when inconsistent in detail. A classifier can also aggregate many weak features jointly, whereas meta-analysis evaluates each genus separately—so the two halves of this study measure genuinely different properties, and should not be expected to agree.

5.4 The HIV inversion

HIV is the clearest single result. Cross-study AUC falls below chance across cohorts and model families, which cannot be attributed to noise or to a particular algorithm. The explanation is visible directly in the effect sizes: the two largest HIV cohorts are negatively correlated ($r = -0.42$),

with 65% of genera disagreeing on direction. A model trained on two cohorts therefore learns a pattern that reverses in the third.

The cause is identifiable. The genera that flip—*Prevotella* and its enterotype partners—track diet and sexual-behaviour risk group, not the virus. Among HIV-negative participants alone, *Bacteroides* differs between risk groups at $g = -1.84$ ($p = 0.0001$), larger than and opposite to its apparent HIV effect ($g = +0.77$). Cohorts that enrolled different proportions of these groups therefore report opposite "HIV" signals, and the classifiers are detecting the confounder. Three independent lines of evidence agree that no virus-specific signature survives: the original paper's own supplementary tables list no disease-specific genera for HIV; our pooling finds none significant; and classification is anti-predictive. For HIV, the absence of a reproducible microbiome signature is a consistent finding rather than a null result.

5.5 Confounding between study and protocol

That a classifier identifies the 16S region from healthy samples at 76% accuracy against a 43.8% baseline demonstrates that laboratory protocol leaves a large, learnable signature in exactly the samples where no disease can explain it.

This bounds rather than decomposes the problem. Each study contributes one disease, so study identity and disease are confounded by construction and cannot be separated with this design. The result establishes a ceiling on how much of any observed cross-study signal could be technical; it does not establish how much actually is. Meta-regression with sequencing platform, 16S region and extraction protocol as moderators would move from bounding to attribution, and is the natural next step.

5.6 Non-independence as a shared problem

The finding that removing two datasets sharing control samples costs 28 of 43 significant genera, against a random-removal median of 41, is a caveat on this analysis as much as on the original. The same duplication inflates vote counting, since those controls contribute to two datasets there too. It is inherited from a design decision in the source database—treating one cohort compared against two different patient groups as two independent datasets—and any re-analysis of MicrobiomeHD should account for it.

6 Limitations and Future Work

6.1 Limitations

One disease dominates.. Of 91 significant genera identified within diseases, 69 come from CDI alone; obesity contributed one and HIV none. Statements about "disease in general" are therefore substantially statements about *C. difficile* infection, an acute condition with unusually large effects.

Study and disease are confounded by construction.. Each study contributes a single disease, so "this genus responds to disease" cannot be fully separated from "this laboratory's protocol produces this reading." The study-of-origin classifier bounds this contribution but cannot resolve it.

Transform dependence.. I^2 is computed on CLR-transformed abundances. Linear rescaling cannot affect it—Hedges' g is already standardised, and z-scoring within studies leaves both g and I^2 unchanged to floating-point precision. A non-linear transform can, however: for *Fusobacterium* in colorectal cancer, I^2 is 80.9% on the CLR scale and 36.0% on relative abundance. A systematic comparison across all genera was not completed and is the most significant outstanding check.

Few studies per genus.. Most genera identified as both significant and consistent rest on the minimum three studies. An I^2 of zero at $k = 3$ more often reflects insufficient power to detect disagreement than genuine agreement. Only *Escherichia/Shigella* ($k = 18$) is supported by enough cohorts for its low heterogeneity to be persuasive.

Conservative feature selection in classification.. Only genera present in at least 80% of training datasets enter the global model—39 of roughly 290, and as few as 26 for the held-out validation cohorts. This avoids a model identifying its training cohorts from missing values, but discards genera that may be informative where measured.

Fixed hyperparameters.. Model hyperparameters were left at defaults rather than tuned within each training fold. Nested cross-validation would be more rigorous, though with three to five cohorts per disease the inner folds become very small.

Checks that came back negative.. Two sensitivity analyses found no effect and are reported for completeness. Varying zero replacement between an additive pseudocount and multiplicative replacement left Hedges' g correlated at 0.9924 with a median absolute difference of 0.018. Substituting DerSimonian–Laird for Paule–Mandel did not change any conclusion.

6.2 Future work

Transform sensitivity.. Recompute the cross-disease pooling on relative abundance and compare the I^2 distributions directly. This is the one outstanding check capable of qualifying the central result, and should precede any further extension.

Meta-regression.. Adding sequencing platform, 16S region and extraction protocol as moderators would convert the heterogeneity finding from a bound into an attribution—stating what fraction of between-study variation is explained by named technical variables rather than only that technical variables are detectable.

An established-tool comparison.. MMUPHin ma2022mmuphin performs batch adjustment and meta-analysis jointly and would provide independent validation of a pipeline implemented here from first principles. Its batch-adjustment function is arguably the more relevant of the two, given the confounding discussed above.

Revisiting excluded samples.. The original label lists discard 359 overweight participants as neither case nor control. Obesity has the most cohorts and thus the greatest influence on pooled cross-disease estimates, making this the one excluded group whose inclusion could plausibly alter the result. Any such analysis must apply the widened definition to every obesity cohort simultaneously.

7 Conclusion

We reproduced duvallet2017 exactly—30 of 30 case–control comparisons matching the published q-values, and both published genus lists at 100%—and then re-analysed the same data by pooling standardised effect sizes rather than counting significant studies.

Pooling identifies 43 shared cross-disease genera against the original 51, overlapping on 21. More consequentially, it supplies a quantity neither original method could: between-study heterogeneity. Median I^2 is 73% across all genera and 84% among those designated shared, so 49 of the original 51 rest on studies that substantially disagree with one another. A sensitivity analysis reduces our own 43 to 15 once two datasets sharing control samples with two others are removed—a loss no random pair of removals approached.

Classification tells a consistent story from an independent direction. A general cross-cohort model is weak (mean AUC 0.599) and barely uses disease-specific information (removing an entire disease changes it by −0.017); no alternative model family recovers the loss. Yet where a genuine signal exists a specialist finds it—a CRC model scores 0.763 on a wholly unseen cohort against the general model's 0.653—and where the label is confounded, specialising hurts: HIV inverts entirely, performing below chance, traceable to a lifestyle confounder in which *Bacteroides* differs by risk group at $g = -1.84$ among HIV-negative participants, larger than and opposite to the apparent HIV effect. From healthy samples alone a classifier identifies the 16S region used with 76% accuracy against a 44% baseline, showing that laboratory protocol is at least as detectable in these data as disease.

We also document an unreported edge case in the original Stouffer implementation, which converts a zero-effect q-value into a p-value of exactly zero and thereby inverts the reported direction of at least one genus.

The original work is not wrong. Its analysis was standard for its time, its numbers reproduce exactly, and the uniform reprocessing it contributed is what made this re-analysis feasible at all. The conclusion is narrower and more useful than a refutation: vote counting cannot distinguish studies that agree from studies that contradict each other, and for a claim about a *shared* response across diseases, that distinction is the entire question. Where the studies do agree—most clearly *Escherichia/Shigella*, supported by 18 cohorts with low heterogeneity and a tight confidence interval—pooling states the case more strongly than counting ever could.