Title: Correcting for batch effects in case-control microbiome studies.

Authors: Sean M. Gibbons[1,2,3], Claire Duvallet[1,2], and Eric J. Alm[1,2,3]

Affiliations: [1] Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA; [2] Center for Microbiome Informatics and Therapeutics, Cambridge, MA, USA; [3] The Broad Institute of MIT and Harvard, Cambridge, MA, USA

Abstract

High-throughput data generation platforms, like mass-spectrometry, microarrays, and second-generation sequencing are susceptible to batch effects due to run-to-run variation in reagents, equipment, protocols, or personnel. Currently, batch correction methods are not commonly applied to microbiome sequencing datasets. In this paper, we compare different batch-correction methods applied to microbiome case-control studies. We introduce a model-free normalization procedure where features (i.e. bacterial taxa) in case samples are converted to percentiles of the equivalent features in control samples within a study prior to pooling data across studies. We look at how this percentile-normalization method compares to traditional meta-analysis methods for combining independent p-values and to limma and ComBat, widely used batch-correction models developed for RNA microarray data. Overall, we show that percentile-normalization is a simple, non-parametric approach for correcting batch effects and improving sensitivity in case-control meta-analyses.

Author Summary

Batch effects are obstacles to comparing results across studies. Traditional meta-analysis techniques for combining p-values from independent studies, like Fisher's method, are effective but statistically conservative. If batch-effects can be corrected, then statistical tests can be performed on data pooled across studies, increasing sensitivity to detect differences between treatment groups. Here, we show how a simple, model-free approach corrects for batch effects in case-control microbiome datasets.

Introduction

Data generated by high throughput methods like mass-spectrometry, second-generation sequencing, or microarrays are sensitive to experimental and computational processing [1]. This sensitivity gives rise to 'batch effects' between independent runs of an experiment. Even when different research groups adhere

to the same methodologies, these effects can arise due to slight differences in hardware, reagents, or personnel [2]. Thus, it is inappropriate to make direct, quantitative comparisons of uncorrected data across studies.

Several tools for reducing batch effects in RNA microarray data have been developed. For example, surrogate variable analysis (SVA) estimates a set of inferred variables (eigenvectors) that explain variance associated with putative batch effects [3]. These inferred variables are then incorporated into a linear model to correct downstream significance tests. The limma package employs a similar linear correction to account for batch effects prior to statistical analysis [4]. SVA and limma are part of a family of linear batch-correction methods that use different varieties of factor analysis, singular value decomposition, or regression [3-6]. The most relied upon method to date [7], called ComBat, uses a Bayesian approach to estimate location and scale parameters for each feature within a batch [8]. All of these models are most effective when batch effects are not conflated with the true biological effects [1]. Furthermore, most batch correction methods make certain parametric assumptions.

Unfortunately, models that often work well for many types of ‘omics data may not be appropriate for microbiome datasets. In microbiome studies, batch effects are often diffuse and conflated with biological signals [9-11]. The microbiome field has also struggled with finding appropriate parametric models for bacterial abundance distributions and for dealing with zeros. This is especially true for low-biomass samples in microbiome sequencing studies, like samples taken from the built environment [12], where populations are under-sampled, the

biological signal is relatively weak, and batch effects can be quite large [13]. One way to get around this issue is to calculate statistics within a given batch, and then compare significant features across batches using classic meta-analysis techniques for combining p-values, like Fisher's and Stouffer's methods [14, 15]. These meta-analysis techniques are robust to batch effects across independent studies, but have less statistical power and ability to detect subtle differences than directly pooling data across studies.

Here, we describe a model-free data-normalization procedure for controlling batch effects in case-control microbiome studies that enables pooling data across studies. Case-control studies include a built-in population of control samples (e.g. healthy subjects) that can be used to normalize the case samples (e.g. diseased subjects). For every feature (i.e. bacterial taxon), the case abundance distributions can be converted to percentiles of the equivalent control abundance distributions (Fig. 1). Study-specific batch effects present in the case samples will also be present in the control samples, and by converting the case data into percentiles of the control distribution these effects are mitigated. Upon conversion to percentiles of the within-study controls, percentile-normalized samples from multiple studies with similar case-control definitions can be more appropriately pooled for statistical testing. We show that this approach effectively controls batch effects in microbiome case-control studies and we compare this method to pooling ComBat- or limma-corrected data, and to Fisher's and Stouffer's methods for combining independent p-values.

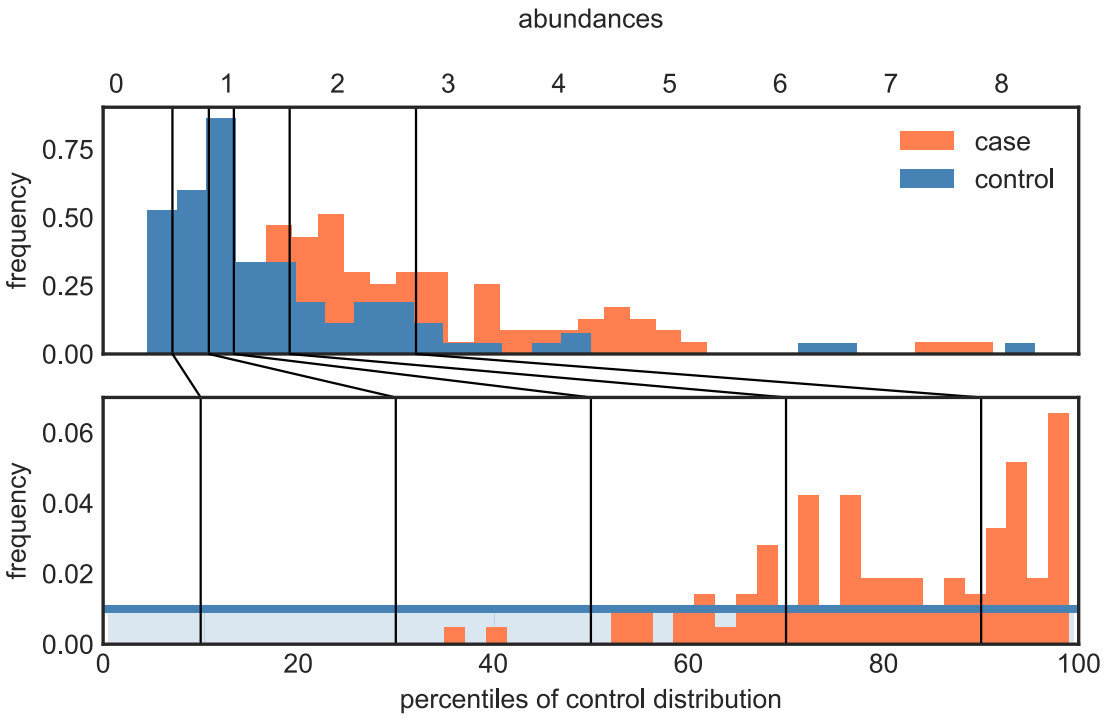


Figure 1. Percentile-normalization procedure converts case values into percentiles of the control distribution. Theoretical feature abundance distributions for the control samples (blue) and case samples (orange) are shown in the upper panel. Converting the control distribution into percentiles of itself naturally gives rise to a uniform distribution (represented by blue horizontal line in lower panel), while converting the case distribution into percentiles of the control distribution produces a non-uniform distribution when these two distributions differ (lower panel). Black lines show where control distribution percentiles lie on the original and percentile-normalized histograms (10th, 30th, 50th, 70th, and 90th percentiles). The control distribution was produced by randomly sampling 100 times from a lognormal distribution with parameters $\mu = 0.1$ and $\sigma = 0.7$. The case distribution was produced in a similar fashion, with distribution parameters $\mu = 0.8$ and $\sigma = 0.5$.

Methods

Datasets

We used a collection of case-control datasets obtained from the MicrobiomeHD database [16] to validate our batch-normalization method. We focused our analyses on studies spanning five diseases: colorectal cancer (CRC) [17-20],

107 Crohn's Disease (CD) [21-24], Ulcerative Colitis (UC) [22-24], obesity (OB) [17,
108 25-34], and *Clostridium difficile* induced diarrhea (CDI) [32, 35]. For a subset of
109 three CRC studies [17-19], we were able to obtain sequence data from the same
110 region of the 16S gene (V4) so that these data could be processed together. The
111 remaining MicrobiomeHD case-control datasets were previously processed using
112 the same pipeline (see below), and then Operational Taxonomic Units (OTUs)
113 were summarized at the genus level for comparison across studies.
114
115 *Sequence Data Processing*
116 To perform OTU-level analyses across the CRC studies, we downloaded the raw
117 data from all of the MicrobiomeHD datasets that sequenced the V4 region of the
118 16S gene. We quality filtered and length trimmed each V4 dataset as described
119 in Duvallet et al. (2017) and concatenated these raw, trimmed FASTQ files into
120 one file. We removed any unique sequences that did not appear more than 20
121 times and clustered the remaining reads with USEARCH [36] at 97% similarity.
122 We assigned these OTUs taxonomic identifiers using the RDP classifier [37] with
123 a cutoff of 0.5.
124 For genus-level analyses, OTU tables and metadata were acquired from
125 the MicrobiomeHD database (https://doi.org/10.5281/zenodo.569601). Raw data
126 were downloaded from the original studies and processed through our in-house
127 16S-processing pipeline
128 (https://github.com/thomasgurry/amplicon_sequencing_pipeline) as described
129 previously [16]. Each study's OTU table was converted to relative abundance by

130 dividing each sample by its total number of reads and collapsed to genus level by
131 summing all OTUs with the same genus-level annotation.
132 To plot data in ordination space, Bray-Curtis distances were calculated
133 from relative abundance data using Scikit-learn
134 (sklearn.metrics.pairwise.pairwise_distances; metric='braycurtis') [38]. Non-
135 metric multidimensional scaling (NMDS) coordinates were calculated for two
136 axes based on Bray-Curtis distances using Scikit-learn (sklearn.manifold.MDS;
137 n_components=2, metric=False, max_iter=500, eps=1e-12,
138 dissimilarity='precomputed').
139
140 *Percentile Normalization*
141 Relative abundance distributions were converted to percentiles using the SciPy v
142 0.19.0 [39] stats.percentileofscore method (kind='mean'). In order to avoid rank
143 pile-ups due to the presence of many zeros, we replaced zeros with pseudo
144 relative abundances drawn from a uniform distribution between 0.0 and 10^{-9} (i.e.
145 a set of random values smaller than the lowest possible relative abundance in
146 any dataset). Within each study, control distributions for each individual OTU or
147 genus were converted into percentiles of themselves and case distributions were
148 converted into percentiles of their corresponding control distribution. We have
149 written a python script that performs percentile-normalization given an OTU table,
150 a list of case sample IDs, and a list of control sample IDs as inputs
151 (https://github.com/seangibbons/percentile_normalization). A QIIME 2

152 (https://qiime2.org) plugin for running percentile-normalization is also available
153 (https://github.com/cduvallet/q2-perc-norm).
154
155 *ComBat*
156 For each disease, we applied ComBat [7] to the case-control datasets analyzed
157 in this study. Relative abundances (OTUs in the CRC analysis or OTUs
158 collapsed to the genus level in the genus-level analysis) were log-transformed
159 prior to running ComBat (default settings), adding a pseudo relative abundance
160 of half the minimal frequency (across the entire feature table) to replace zeros.
161 ComBat-corrected data were then transformed back from log-space (i.e.
162 exponential transformation) prior to downstream analyses.
163
164 *limma*
165 In addition to ComBat, we applied a linear batch correction method from the
166 limma package in R [4]. Relative abundances (zeros replaced with pseudo
167 relative abundances) were log-transformed as described above and then a linear
168 model was fit to subtract batch effects using the removeBatchEffect{} function
169 (default settings). The limma-corrected data were then transformed back from
170 log-space (i.e. exponential transformation) prior to downstream analysis.
171
172 *Statistical Analysis*
173 To calculate statistical significance, we restricted our statistical tests to
174 OTUs/genera that occurred in at least one third of control or one third of case

175 samples in order to reduce our multi-test correction penalty. We used the
176 Wilcoxon rank-sum test, as implemented in SciPy v0.19.0 (sicipy.stats.ranksums)
177 [39], to determine significant differences between independent groups of
178 samples. Wilcoxon tests were run either within or across studies. In order to
179 calculate statistics across studies, normalized case and control samples from
180 multiple studies of the same disease were combined together into the same OTU
181 table. Hereafter, combining datasets is referred to as 'pooling.' P-values were
182 multiple-test corrected using the Benjamini-Hochberg False Discovery Rate
183 (FDR) procedure, as implemented in StatsModels v 0.8.0
184 (statsmodels.sandbox.stats.multicomp.multipletests) [40]. Differences in overall
185 community structure were assessed using the Permutational Multivariate
186 Analysis of Variance (PERMANOVA) test in R's vegan package [41] as
187 implemented in scikit-bio (skbio.stats.distance.permanova). Fisher's and
188 Stouffer's methods for combining p-values were performed using SciPy v0.19.0
189 (scipy.stats.combine_pvalues; method='fisher' or method='stouffer'). For
190 Stouffer's method, weights for each study were calculated as the square root of
191 the number of cases plus the number of controls. OTUs/genera with significant
192 responses in opposing directions across studies were excluded from Fisher and
193 Stouffer analyses.
194
195 *in silico experiments*
196 We ran an *in silico* titration experiment using the OTU-level data to
197 simulate pooling of control samples from different datasets before calculating

198 significant differences. Healthy samples from one study were mixed with healthy
199 samples from another study at different proportions prior to calculating significant
200 differences in OTU frequencies between cases and controls. Case and control
201 groups were subsampled to 40 samples each. Control samples were substituted
202 by randomly selected samples from another study along a fractional gradient (0-
203 100% control samples from another study). We calculated significant differences
204 between case and control groups using the Wilcoxon rank-sum test and applied
205 an FDR correction. OTUs with q-values ≤ 0.05 were considered significant. The
206 titration experiment was rerun 20 times, and the results were averaged.
207 Similar to the titration experiment, we ran an OTU-level analysis of how
208 batch-correction methods might impact false-positive rates by randomly selecting
209 40 control samples from the Baxter et al. (2016) study as artificial 'controls' and
210 40 control samples from the Zeller et al. (2014) study as artificial 'cases' (across
211 20 iterations) for each data type (i.e. raw, percentile-normalized, limma-
212 corrected, and ComBat-corrected). We then calculated significant differences
213 between these artificial 'case' and 'control' groups as outlined above to generate
214 p-values for each OTU.
215
216 **Results**
217 *Batch effects at OTU-level resolution*
218 To minimize possible biases across data sets, we identified three colorectal
219 cancer (CRC) studies that sequenced the same region of the 16S gene (V4). We

220 reprocessed the raw sequence data from each study in the same quality filtering
221 and OTU picking pipeline to obtain bioinformatically-standardized results. OTUs
222 that occurred in at least one third of case or one third of control samples (i.e.
223 within individual studies) were retained for all downstream statistical analyses.
224 Despite standardizing the bioinformatic processing of these data, we saw
225 significant batch effects in healthy patients across studies (PERMANOVA $p <$
226 0.001; Fig. 2). The similarity between samples from the Baxter et al. (2016) and
227 Zackular et al. (2014) studies is due to the fact that they were sourced from the
228 same patient cohort (although samples were processed separately), making this
229 comparison a good pseudo-negative control for batch effects [17, 19]. There was
230 an apparent reduction in the batch effect after applying ComBat, although
231 differences between batches remained weakly significant (PERMANOVA $p =$
232 0.008, Fig. 2) [7]. Due to the non-independence between the Baxter and Zackular
233 patient cohorts, we removed the smaller of the two studies (Zackular) from all
234 downstream analyses. Out of a total of 1,021 OTUs that passed our abundance
235 filter, 681 differed significantly in uncorrected relative abundance between the
236 Baxter and Zeller healthy controls (FDR $q \leq 0.05$).

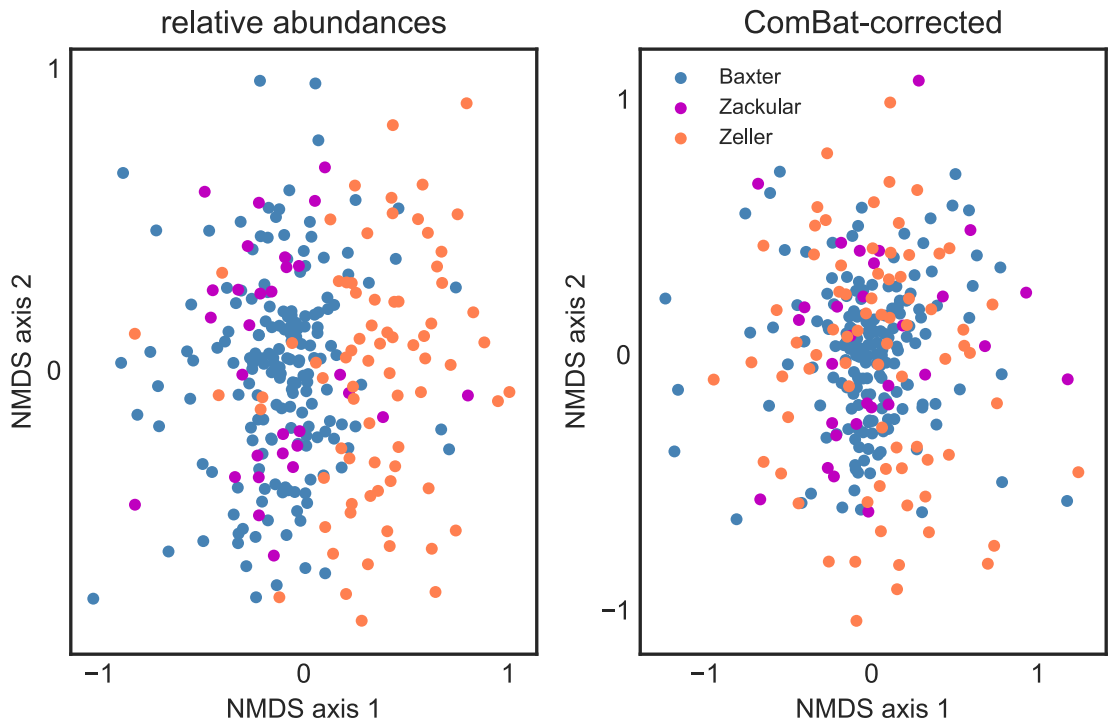


Figure 2. Batch effects between healthy controls from different studies can be reduced by ComBat. Non-metric multidimensional scaling (NMDS) plot showing the distribution of healthy controls from three colorectal cancer studies in ordination space (Bray-Curtis distances of relative abundance OTU-level data). Despite standardized bioinformatic processing, healthy patients differed significantly in their gut microbiomes across studies (PERMANOVA $p < 0.001$). Studies were still significantly different after applying ComBat, an established batch-correction method (PERMANOVA $p < 0.01$).

We ran an *in silico* titration experiment to simulate pooling of control samples from different datasets before calculating significant differences. Healthy samples from one study were mixed with healthy samples from another study at different proportions prior to calculating significant differences in OTU frequencies between cases and controls (see conceptual outline in Fig. 3). For non-normalized data, the number of significant OTUs greatly increased due to batch effects as more control samples were substituted in from another study. This result highlights the danger of pooling raw data across batches. ComBat-

and limma-corrected data performed better than uncorrected data, but still showed many spurious results as the proportion of control samples from another study increased (Fig. 3). Percentile-normalization showed no increase in spurious results over the titration gradient (Fig. 3). Although we do see several significant CRC-associated OTUs in the full dataset (see below), these were not detected in the titration experiment due to the loss of statistical power when reducing case and control groups to only 40 samples each.

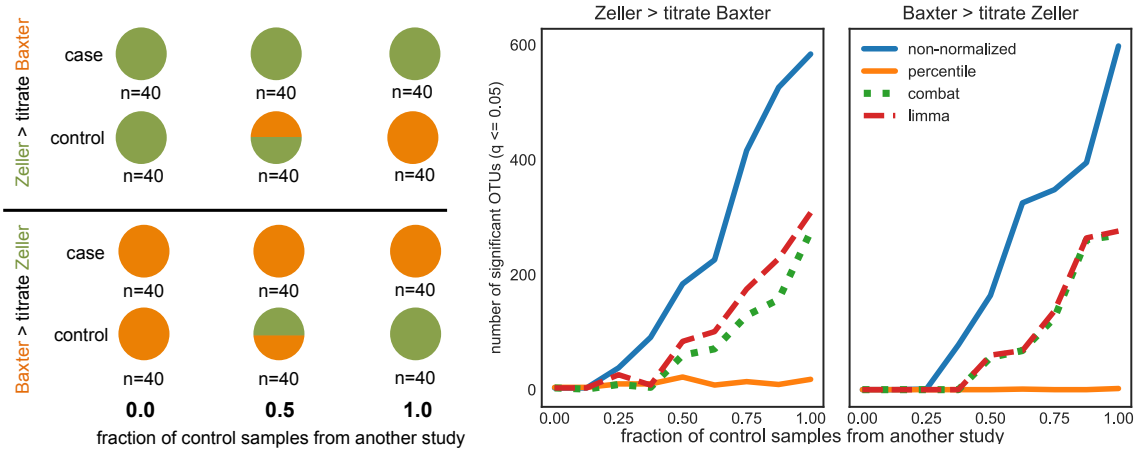
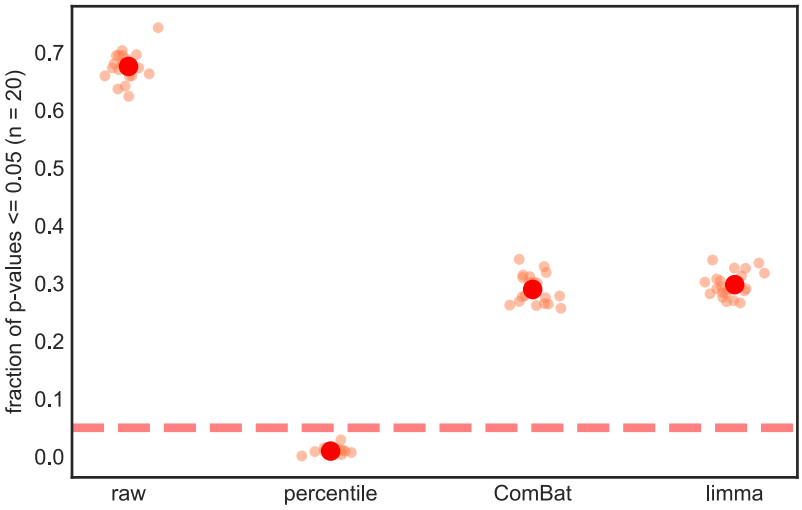


Figure 3. Pooling non-normalized samples from different studies can give rise to many spurious associations. The control group from one study is gradually substituted with randomly chosen control samples from another study (non-normalized, percentile-normalized, limma-corrected, and ComBat-corrected), keeping the total number of case and control samples fixed at n=40 (see conceptual illustration on the left). Mixing in non-normalized control samples from another study gave rise to spurious results due to batch effects (blue lines). ComBat- and limma-corrected data showed fewer spurious associations (green and red lines). Percentile-normalization showed no increase in spurious results along the titration gradient (orange lines).

We ran a second *in silico* experiment to determine whether the false-positive rate was impacted by our different batch-correction methods. We randomly selected 40 samples from the Baxter healthy controls as artificial

278 ‘controls’ and 40 samples from the Zeller healthy controls as artificial ‘cases’ for
279 each data type (i.e. non-normalized, percentile-normalized, limma-corrected, and
280 ComBat-corrected) and calculated significant OTU-level differences between
281 these groups. We repeated this process twenty times to generate a set of p-value
282 distributions. We found that the fraction of p-values ≤ 0.05 can be as high as
283 ~70% for the non-normalized data (Fig. 4). This result matches with our finding
284 that 681 out of the 1,021 OTUs in this dataset differed significantly across Baxter
285 and Zeller controls (q ≤ 0.05). Each normalization technique drastically reduced
286 the number of false positives, but percentile-normalization gave the best results
287 (Fig. 4). When low-abundance OTUs were included in the analysis, ComBat and
288 limma showed highly skewed p-value distributions, giving rise to a large number
289 of false positives than the non-normalized data (Fig. S1).



290

291 **Figure 4. False positive rates are reduced by batch-correction methods.**
292 Random sets of 40 Baxter controls and random sets of 40 Zeller controls were

293 selected for null case-control comparisons (20 iterations). Smaller points show
294 the fraction of p-values ≤ 0.05 within a given iteration, while larger dots show the
295 average value across all 20 iterations. Within each category, smaller points are
296 randomly jittered along the x-axis for better visualization. The fraction of p-values
297 ≤ 0.05 is highly inflated for non-normalized data (red dashed line shows the null-
298 expectation for p-values). Only abundant OTUs (detected in at least a third of
299 case or control samples) were included in this analysis.
300
301

302 We next assessed the performance each cross-study analysis method by
303 comparing OTU-level results across two independent CRC datasets. In the
304 Baxter study, there were 172 healthy (control) samples and 120 CRC (case)
305 samples, with 14 OTUs (from *Fusobacterium*, *Coprococcus*, *Butyricicoccus*,
306 *Gemmiger*, *Faecalibacterium*, *Roseburia*, *Parvimonas*, *Haemophilus*,
307 *Porphyromonas*, *Peptostreptococcus*, *Streptophyta*, *Bacteroides* and *Clostridium*
308 *XIVa* genera) showing significant differences in abundance between cases and
309 controls (FDR q ≤ 0.05). For Zeller, there were 71 control and 40 case samples,
310 with 18 OTUs (from *Butyricicoccus*, *Butyricimonas*, *Fusobacterium*, *Closridium*
311 *XIVa*, *Streptococcus*, *Parabacteroides*, *Alistipes*, *Anaerostipes*, *Parvimonas*,
312 *Peptostreptococcus*, *Blautia*, *Dialister,* and *Bacteroides* genera) that differed
313 significantly across cases and controls (FDR q ≤ 0.05).

314 In the absence of batch effects, pooling data across datasets of the same
315 disease should increase sensitivity to detect significant cross-study associations.
316 We pooled percentile-normalized, limma-corrected, and ComBat-corrected data,
317 respectively, across Baxter and Zeller studies to look for OTUs that differed
318 significantly across cases and controls. These pooled results were then

319 compared to classic methods for combining p-values from each dataset's
320 individual results (above). For the percentile-normalized data, we found 39 OTUs
321 that differed significantly between cases and controls (FDR q ≤ 0.05), 21 of which
322 overlapped with the within-study results. The pooled limma-corrected and
323 ComBat-corrected data resulted in 37 and 36 significant OTUs, respectively. 35
324 of the OTUs identified as significant by ComBat were also significant in the limma
325 results. 30 of the limma results and 29 of the ComBat results were also
326 significant in the percentile-normalization results, respectively. Fisher's method
327 identified seven significant OTUs from *Clostridium XIVa*, *Streptococcus*,
328 *Fusobacterium*, *Parvimonas*, *Peptostreptococcus*, and *Anaerostipes* genera,
329 which were also found in the percentile-normalized results. Stouffer's method
330 identified the same seven OTUs found using Fisher's method. Overall, the
331 pooling methods improve statistical power to detect significant OTUs over
332 traditional meta-analysis methods. For example, particular OTUs from
333 *Desulfovibrio* and *Parabacteroides* genera were identified as significantly
334 enriched in CRC patients in the pooled results (ComBat, limma, and percentile-
335 normalized), but not in the within-study results or in the Fisher and Stouffer
336 results (Fig. 5). Pooled analysis of percentile-normalized data also identified on
337 *Enterobacter* OTU enriched in caner patients, two OTUs from the
338 *Lachnospiraceae* family that were enriched in controls and one *Lachnospiraceae*
339 that was enriched in cases, which were missed by the within-study analyses. In
340 all, 18 OTUs were identified in the pooled, percentile-normalized results that
341 were missed by the within-study analyses (Fig. 5). These additional taxonomic

associations (e.g. *Desulfovibrio*, *Costridium XIVa*, and *Lachnospiraceae*) are consistent with prior meta-analyses of CRC microbiome studies [16, 42].

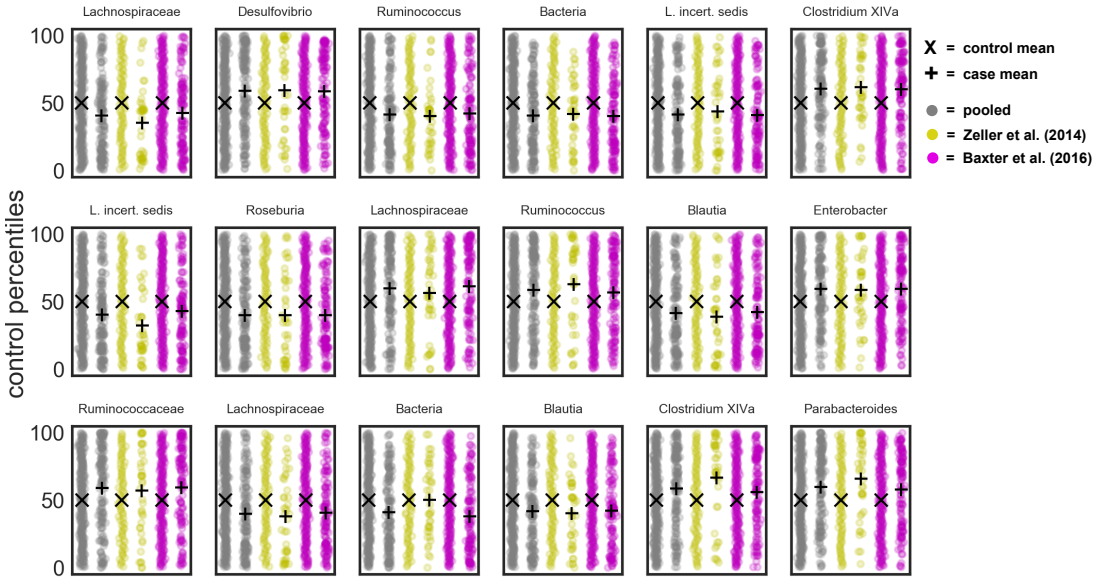


Figure 5. OTUs significant across CRC studies, but not within a given study. Pooling data provides greater statistical power to detect subtle, yet consistent differences in OTU abundances across sample groups. 18 OTUs are labeled by their most resolved taxonomic annotation. Each OTU in this plot was not found to be significant within either Baxter or Zeller studies, but became significant after pooling the percentile-normalized datasets ($q \leq 0.05$).

Batch effects at genus-level resolution across multiple diseases

In order to assess the performance of different meta-analysis techniques across a larger set of studies and diseases, we summarized OTU abundances at the genus level for five diseases across 18 studies - *Clostridium difficile induced diarrhea* (CDI), Crohn's disease (CD), ulcerative colitis (UC), obesity (OB), and CRC. There were a total of 306 unique genera detected across studies. There

360 were two CDI case-control studies: Schubert et al. (2014) had 154 control and 93
361 case samples [32]; Vincent et al. (2013) had 25 control and 25 case samples
362 [35]. There were four inflammatory bowel disease (IBD) studies that included CD
363 patients and three that also included UC patients: Papa et al. (2012) had 24 non-
364 IBD control samples, 23 CD samples, and 43 UC samples [22]; Morgan et al.
365 (2012) had 18 control, 61 CD and 47 UC samples [23]; Willing et al. (2010) had
366 35 control, 16 UC and 29 CD samples [24]; Gevers et al. (2014) had 16 non-IBD
367 control and 146 CD samples, with no UC samples [21]. There were eleven
368 studies with lean and obese (OB) cohorts: Turnbaugh et al. (2009) had 33
369 controls and 102 cases [25]; Goodrich et al. (2014) had 428 controls and 185
370 cases [29]; Escobar et al. (2014) had 10 controls and 10 cases [26]; Zhu et al.
371 (2014) had 16 controls and 25 cases [43]; Jumpertz et al. (2011) had 12 controls
372 and 9 cases [33]; Ross et al. (2015) had 26 controls and 37 cases [28]; Zupancic
373 et al. (2012) had 96 controls and 101 cases [27]; Baxter et al. (2016) had 125
374 controls and 47 OB cases [17]; Schubert et al. (2014) had 68 controls and 34 OB
375 cases [32]; Wu et al. (2011) had 59 controls and 9 cases [31]; and Zeevi et al.
376 (2015) had 567 controls and 151 cases [30]. There were four independent CRC
377 studies, including the Baxter and Zeller studies listed in the OTU-level analysis
378 (see above for sample sizes). The remaining two CRC studies are Wang et al.
379 (2012), which had 54 control and 44 case samples [44], and Chen et al. (2012),
380 which had 22 controls and 21 cases [20].

381 Most genera that differed significantly within a given study were not
382 significant in other studies of the same disease. For example, of the 36 unique

383 genera that showed significant differences within any given OB study, none were
384 found to be significant in all studies (I = 0; Table 1). Indeed, there were no genera
385 that were significant across all studies in the majority of diseases studied (I = 0;
386 Table 1). CDI only had two studies, and of the 38 significant results, only six were
387 shared across both datasets. Overall, few genera were significant within two or
388 more studies (2N ≤ 6; Table 1).

389 The number of genera that differed significantly across pooled cases and
390 controls changed depending on how the data were batch-corrected (Table 1). For
391 every disease, percentile-normalization yielded the largest number of significant
392 genera when compared to other methods. Overall, ComBat- and limma-
393 corrections resulted in many fewer significant genera, especially for CD, UC, and
394 CRC (Table 1). Half of the IBD (CD and UC) studies included non-IBD patients
395 with inflammatory symptoms as controls rather than clinically healthy patients.
396 These biologically relevant differences in inflammatory symptoms between
397 control cohorts were conflated with batches and were likely smoothed out by
398 ComBat and limma corrections. Fisher's and Stouffer's methods consistently
399 identified fewer significant associations than percentile-normalization (Table 1).
400 Pooling data prior to running a statistical test is a more sensitive technique than
401 combining independently calculated p-values [45]. Thus, percentile-normalization
402 increases the statistical power to detect differences across studies while
403 controlling for false positives and batch effects.

disease	method	number of significant genera pooled	within
CDI (N=2)	percentile	37	U = 38 | 2N = 6
	Fisher	14	
	Stouffer	14	
	ComBat	36	
	limma	36	
CD (N = 4)	percentile	19	U = 13 | I = 0 | 2N = 2
	Fisher	5	
	Stouffer	5	
	ComBat	2	
	limma	1	
UC (N = 3)	percentile	11	U = 17 | I = 0 | 2N = 1
	Fisher	3	
	Stouffer	3	
	ComBat	5	
	limma	5	
CRC (N = 4)	percentile	12	U = 20 | I = 0 | 2N = 3
	Fisher	9	
	Stouffer	7	
	ComBat	5	
	limma	5	
OB (N = 11)	percentile	22	U = 36 | I = 0 | 2N = 6
	Fisher	4	
	Stouffer	6	
	ComBat	13	
	limma	15	

Table 1. Normalization methods impact the number of significant genus-level associations between cases and controls across multiple diseases. Numbers of genera that differ significantly between cases and controls for five diseases. In the 'disease' column, CDI = *Clostridium difficile* induced diarrhea, CD = Crohn's Disease, UC = Ulcerative Colitis, CRC = Colorectal Cancer, and OB = obesity. 'N =' shows the number of studies included in each meta-analysis. The method column indicates how the data were processed prior to running significance tests (percentile-normalized, Fisher's method for combining p-values, Stouffer's method for combining p-values, ComBat-corrected, or limma-corrected). The significance threshold used was $q \leq 0.05$ (FDR). The 'pooled' column shows the total number of genera that were found to be significantly different *across* pooled studies for a given disease. The 'within' column shows the total number of unique (non-redundant) genera that were identified as significantly different *within* each study (U = union), the number of genera that were significant in all individual studies (I = intersection), and the number of significant genera that were consistently significant in at least two studies (2N; 2N == I for CDI).

To better assess how percentile normalization impacted the pooled results, we looked at genera that were significant within a single-study but not across studies after pooling. There were 11 genera that were significant within a subset of CRC studies, but not after pooling (Fig. 6). *Butyricimonas*, *Gemmiger*, *Flavonifractor*, *Bacteroides*, and *Roseburia* showed variable responses across studies, sometimes enriched in controls and other times enriched in cases. The remaining genera showed weak associations within one or two studies, but did not differ significantly across studies (i.e. $q \leq 0.05$). These genera that show weak or inconsistent responses across batches may not be reliable disease biomarkers. However, by including larger numbers of CRC studies in future meta-analyses it is likely that some of these genera could pass the significance threshold. Two genera - *Enterobacter* and *Pseudoflavonifractor* - were not significantly different between cases and controls within an individual study, but became significant after pooling (Fig. 7). These genera showed weak, but largely consistent enrichment in cancer patients and demonstrate the utility of pooling datasets to detect subtle differences.

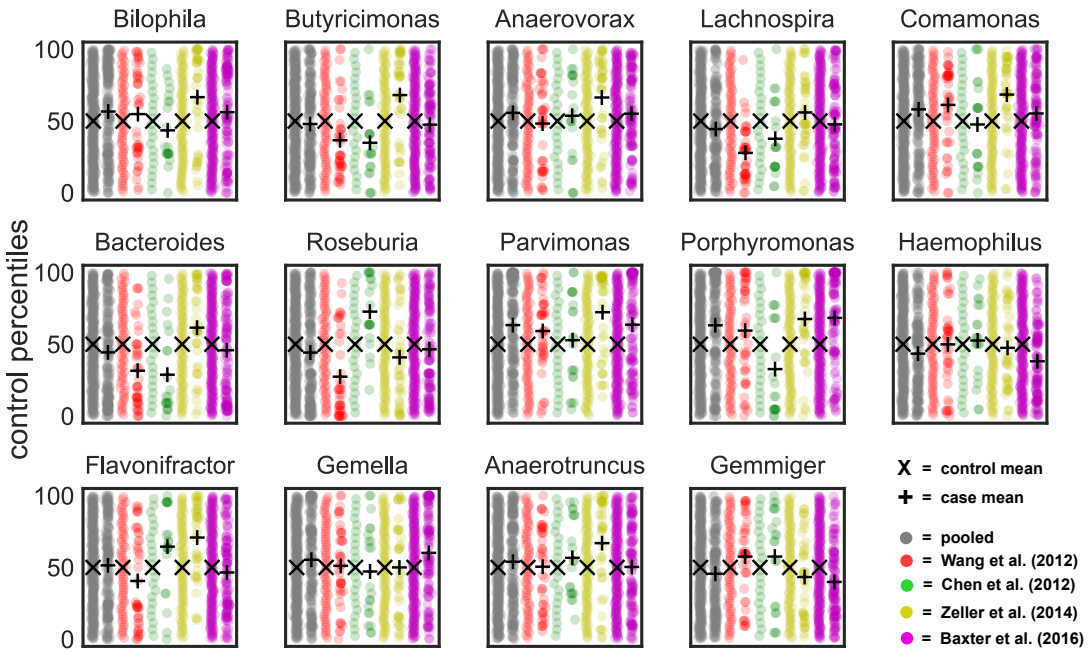


Figure 6. Genera that show a significant difference between CRC cases and controls within a given study, but not after pooling. 14 genera showed significant differences between cases and controls within a study (q ≤ 0.05), but not after pooling across CRC studies.

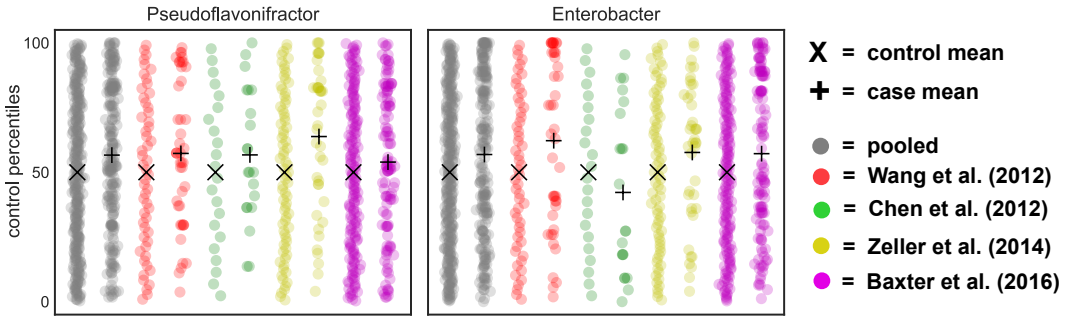


Figure 7. Genera that do not show a significant difference between CRC cases and controls within a given study, but do after pooling. Two genera did not show significant differences between cases and controls within a study, but became significant after pooling across CRC studies (q ≤ 0.05).

Discussion

450 Batch effects are unavoidable when working with high-throughput data
451 generation platforms. The RNA microarray community has been proactive in the
452 development of tools for dealing with these effects [1, 7]. However, these tools
453 are not as effective when batch effects are confounded with biological signals or
454 when parametric assumptions do not apply, which is often the case in
455 microbiome case-control studies [26, 28, 46]. Therefore, model-free methods are
456 needed for correcting batch effects across microbiome datasets. Fortunately,
457 case-control studies can be internally normalized by their own control samples.
458 Any study-specific batch effects in the case samples will be present in the control
459 samples, and by converting the case data into percentiles of the control
460 distribution these effects are attenuated without making parametric assumptions.
461 Relative abundance, limma-corrected, and ComBat-corrected data - but
462 not percentile-normalized data - quickly yielded a large number of spurious
463 results when cases from one study were tested against controls from another
464 (Fig. 3). Additionally, when control populations from different batches were
465 compared to one another, non-normalized data yielded a much larger number of
466 false positives than batch-corrected data (Fig. 4). Our percentile-normalization
467 approach was much more effective than limma and ComBat in controlling false
468 positives (Fig. 4), especially in the presence of low-abundance taxa (Fig. S1).
469 Because pooling datasets increases statistical power, it is tempting to pool
470 these data even in the absence of suitable batch-correction methods.
471 Consequently, pooling non-normalized data from different batches has been
472 common practice in the microbiome field [21, 28, 47, 48]. In this paper, we

473 demonstrate why this practice is highly inadvisable. Pooling batch-corrected data
474 from multiple studies allowed us to detect significant differences that were not
475 found within a given study (Figs. 5 and 7), while removing associations that were
476 weak or inconsistent across studies (Fig. 6). Percentile-normalized results often
477 identified significant differences between cases and controls that were missed by
478 other normalization methods (Table 1). For CDI, percentile-normalized results
479 identified about the same number of significant hits as the other batch-correction
480 methods (Table 1), which was likely due to the fact that the biological signal
481 associated with diarrhea is very strong [16]. In cases where the biological signal
482 is strong, results should be robust to the types of analyses employed. For UC
483 and CD studies (IBD), percentile-normalization identified several significant
484 genera that limma and ComBat did not (Table 1). The reduced number of
485 significant hits from limma- and ComBat-corrected data for IBD was likely due to
486 heterogeneous control cohorts across these studies (i.e. healthy patients vs. non-
487 IBD patients), which likely smoothed-out inflammation-associated signals. This
488 result highlights the importance of having consistent definitions for case and
489 control cohorts across studies.

490 We compared percentile-normalization and pooling to Fisher's and
491 Stouffer's methods for combining independent p-values. Stouffer's method is
492 similar to Fisher's, but includes weights for each p-value based on the number of
493 samples in a study. Percentile-normalization consistently identified a larger
494 number of significant hits than Stouffer's and Fisher's methods, confirming that
495 pooling data increases sensitivity.

496 In conclusion, we present a robust, model-free procedure for transforming
497 each feature in a microbiome case-control dataset into percentiles of its control
498 distribution (Fig. 1). The main conditions for applying this method are that 1) each
499 batch must have a sizeable number of control samples (i.e. the density of the
500 control distribution limits the resolution of the percentile-transformation of the
501 case samples), and 2) case and control populations should be consistently
502 defined across batches (i.e. same definition of 'healthy' or 'diseased' groups).
503 Given these caveats, percentile-normalized features can be pooled across
504 studies for univariate statistical testing (whichever test a researcher prefers -
505 ideally non-parametric), alleviating the batch effect problem. This model-free
506 procedure could also be applied to other types of 'omics datasets with
507 consistently defined internal controls. We find that this procedure allows us to
508 identify differences between cases and controls that are often missed by more
509 conservative meta-analysis techniques. Methods developed for batch-correction
510 in microarray data, like limma and ComBat, can partially reduce batch effects in
511 microbiome studies (Figs. 2-4), but appear to obscure real patterns if batch
512 effects are not independent of biological signals or if the parametric assumptions
513 of these models are not valid. We suggest that methods like limma and ComBat
514 are useful for studies lacking case and control groups. However, when studies
515 have consistently defined internal controls, percentile-normalization should be
516 the preferred batch correction approach. Future work should focus on developing
517 parametric models specifically for batch correction in microbiome datasets, which

518 could further improve sensitivity to detect subtle biological differences across

519 studies.

520

521 **Acknowledgements**

522 We thank the members of the Alm lab for helpful feedback and Marc Sze and

523 Patrick Schloss for providing online code to download the Wu et al. (2011) and

524 Zeevi et al. (2015) obesity datasets.

525

526 **References**

527 1. Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, et
528 al. Tackling the widespread and critical impact of batch effects in high-throughput
529 data. Nat Rev Genet. 2010;11(10):733-9.
530 2. Schloss PD, Gevers D, Westcott SL. Reducing the Effects of PCR
531 Amplification and Sequencing Artifacts on 16S rRNA-Based Studies. PLoS One.
532 2011;6(12):e27310. doi: 10.1371/journal.pone.0027310.
533 3. Leek JT, Storey JD. Capturing heterogeneity in gene expression studies
534 by surrogate variable analysis. PLoS Genet. 2007;3(9):e161.
535 4. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers
536 differential expression analyses for RNA-sequencing and microarray studies.
537 Nucleic Acids Res. 2015;43(7):e47-e.
538 5. Alter O, Brown PO, Botstein D. Singular value decomposition for genome-
539 wide expression data processing and modeling. Proc Natl Acad Sci USA.
540 2000;97(18):10101-6.
541 6. Benito M, Parker J, Du Q, Wu J, Xiang D, Perou CM, et al. Adjustment of
542 systematic microarray data biases. Bioinformatics. 2004;20(1):105-14.
543 7. Chen C, Grennan K, Badner J, Zhang D, Gershon E, Jin L, et al.
544 Removing Batch Effects in Analysis of Expression Microarray Data: An
545 Evaluation of Six Batch Adjustment Methods. PLoS One. 2011;6(2):e17238. doi:
546 10.1371/journal.pone.0017238.
547 8. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray
548 expression data using empirical Bayes methods. Biostatistics. 2007;8(1):118-27.
549 9. Weiss S, Amir A, Hyde ER, Metcalf JL, Song SJ, Knight R. Tracking down
550 the sources of experimental contamination in microbiome studies. Genome Biol.
551 2014;15(12):564. doi: 10.1186/s13059-014-0564-2.
552 10. Shen H, Rogelj S, Kieft TL. Sensitive, real-time PCR detects low-levels of
553 contamination by Legionella pneumophila in commercial reagents. Mol Cell
554 Probes. 2006;20. doi: 10.1016/j.mcp.2005.09.007.

555 11. Nguyen NH, Smith D, Peay K, Kennedy P. Parsing ecological signal from
556 noise in next generation amplicon sequencing. New Phytol. 2015;205(4):1389-
557 93. doi: 10.1111/nph.12923.
558 12. Gibbons SM. The Built Environment Is a Microbial Wasteland. mSystems.
559 2016;1(2):e00033-16.
560 13. Chase J, Fouquier J, Zare M, Sonderegger DL, Knight R, Kelley ST, et al.
561 Geography and Location Are the Primary Drivers of Office Microbiome
562 Composition. mSystems. 2016;1(2). doi: 10.1128/mSystems.00022-16.
563 14. Fisher RA. Statistical methods for research workers: Genesis Publishing
564 Pvt Ltd; 1925.
565 15. Stouffer SA. Adjustment during army life: Princeton University Press;
566 1949.
567 16. Duvallet C, Gibbons SM, Gurry T, Irizarry RA, Alm EJ. Meta-analysis of
568 gut microbiome studies identifies disease-specific and shared responses. Nat
569 Comm. 2017;8(1):1784.
570 17. Baxter NT, Ruffin MT, Rogers MA, Schloss PD. Microbiota-based model
571 improves the sensitivity of fecal immunochemical test for detecting colonic
572 lesions. Genome Med. 2016;8(1):37.
573 18. Zeller G, Tap J, Voigt AY, Sunagawa S, Kultima JR, Costea PI, et al.
574 Potential of fecal microbiota for early‐stage detection of colorectal cancer. Mol
575 Syst Biol. 2014;10(11):766.
576 19. Zackular JP, Rogers MA, Ruffin MT, Schloss PD. The human gut
577 microbiome as a screening tool for colorectal cancer. Cancer Prev Res.
578 2014;7(11):1112-21.
579 20. Chen W, Liu F, Ling Z, Tong X, Xiang C. Human intestinal lumen and
580 mucosa-associated microbiota in patients with colorectal cancer. PLoS One.
581 2012;7(6):e39743.
582 21. Gevers D, Kugathasan S, Denson LA, Vázquez-Baeza Y, Van Treuren W,
583 Ren B, et al. The treatment-naive microbiome in new-onset Crohn's disease. Cell
584 Host Microbe. 2014;15(3):382-92.
585 22. Papa E, Docktor M, Smillie C, Weber S, Preheim SP, Gevers D, et al.
586 Non-invasive mapping of the gastrointestinal microbiota identifies children with
587 inflammatory bowel disease. PLoS One. 2012;7(6):e39242.
588 23. Morgan XC, Tickle TL, Sokol H, Gevers D, Devaney KL, Ward DV, et al.
589 Dysfunction of the intestinal microbiome in inflammatory bowel disease and
590 treatment. Genome Biol. 2012;13(9):R79.
591 24. Willing BP, Dicksved J, Halfvarson J, Andersson AF, Lucio M, Zheng Z, et
592 al. A pyrosequencing study in twins shows that gastrointestinal microbial profiles
593 vary with inflammatory bowel disease phenotypes. Gastroenterology.
594 2010;139(6):1844-54. e1.
595 25. Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, Ley
596 RE, et al. A core gut microbiome in obese and lean twins. Nature.
597 2009;457(7228):480-4.
598 26. Escobar JS, Klotz B, Valdes BE, Agudelo GM. The gut microbiota of
599 Colombians differs from that of Americans, Europeans and Asians. BMC
600 Microbiol. 2014;14(1):311.

601 27. Zupancic ML, Cantarel BL, Liu Z, Drabek EF, Ryan KA, Cirimotich S, et al.
602 Analysis of the gut microbiota in the old order Amish and its relation to the
603 metabolic syndrome. PLoS One. 2012;7(8):e43052.
604 28. Ross MC, Muzny DM, McCormick JB, Gibbs RA, Fisher-Hoch SP,
605 Petrosino JF. 16S gut community of the Cameron County Hispanic Cohort.
606 Microbiome. 2015;3(1):7.
607 29. Goodrich JK, Waters JL, Poole AC, Sutter JL, Koren O, Blekhman R, et al.
608 Human genetics shape the gut microbiome. Cell. 2014;159(4):789-99.
609 30. Zeevi D, Korem T, Zmora N, Israeli D, Rothschild D, Weinberger A, et al.
610 Personalized nutrition by prediction of glycemic responses. Cell.
611 2015;163(5):1079-94.
612 31. Wu GD, Chen J, Hoffmann C, Bittinger K, Chen Y-Y, Keilbaugh SA, et al.
613 Linking long-term dietary patterns with gut microbial enterotypes. Science.
614 2011;334(6052):105-8.
615 32. Schubert AM, Rogers MA, Ring C, Mogle J, Petrosino JP, Young VB, et
616 al. Microbiome data distinguish patients with Clostridium difficile infection and
617 non-C. difficile-associated diarrhea from healthy controls. mBio.
618 2014;5(3):e01021-14.
619 33. Jumpertz R, Le DS, Turnbaugh PJ, Trinidad C, Bogardus C, Gordon JI, et
620 al. Energy-balance studies reveal associations between gut microbes, caloric
621 load, and nutrient absorption in humans. Am J Clin Nutr. 2011;94(1):58-65.
622 34. Sze MA, Schloss PD. Looking for a Signal in the Noise: Revisiting Obesity
623 and the Microbiome. mBio. 2016;7(4):e01018-16.
624 35. Vincent C, Stephens DA, Loo VG, Edens TJ, Behr MA, Dewar K, et al.
625 Reductions in intestinal Clostridiales precede the development of nosocomial
626 Clostridium difficile infection. Microbiome. 2013;1(1):18.
627 36. Edgar RC. Search and clustering orders of magnitude faster than BLAST.
628 Bioinformatics. 2010;26(19):2460-1.
629 37. Wang Q, Garrity GM, Tiedje JM, Cole JR. Naïve Bayesian Classifier for
630 Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy. Appl
631 Environ Microbiol. 2007;73(16):5261-7. doi: 10.1128/aem.00062-07.
632 38. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et
633 al. Scikit-learn: Machine learning in Python. J Mach Learn Res.
634 2011;12(Oct):2825-30.
635 39. Jones E, Oliphant T, Peterson P. SciPy: Open source scientific tools for
636 Python, 2001–. URL http://www.scipy org. 2007;73:86.
637 40. Seabold S, Perktold J, editors. Statsmodels: Econometric and statistical
638 modeling with python. Proceedings of the 9th Python in Science Conference;
639 2010.
640 41. Oksanen J. Multivariate analysis of ecological communities in R: vegan
641 tutorial. R package version. 2011;1(7).
642 42. Tjalsma H, Boleij A, Marchesi JR, Dutilh BE. A bacterial driver–passenger
643 model for colorectal cancer: beyond the usual suspects. Nat Rev Microbiol.
644 2012;10(8):575-82.
645 43. Zhu L, Baker RD, Baker SS. Gut microbiome and nonalcoholic fatty liver
646 diseases. Pediatric Res. 2014;77(1-2):245-51.

647 44. Wang T, Cai G, Qiu Y, Fei N, Zhang M, Pang X, et al. Structural
648 segregation of gut microbiota between colorectal cancer patients and healthy
649 volunteers. ISME J. 2012;6(2):320-9.
650 45. Glass GV. Primary, secondary, and meta-analysis of research.
651 Educational Res. 1976;5(10):3-8.
652 46. Pasolli E, Truong DT, Malik F, Waldron L, Segata N. Machine learning
653 meta-analysis of large metagenomic datasets: tools and biological insights. PLoS
654 Comput Biol. 2016;12(7):e1004977.
655 47. Dubourg G, Lagier J-C, Hüe S, Surenaud M, Bachar D, Robert C, et al.
656 Gut microbiota associated with HIV infection is significantly enriched in bacteria
657 tolerant to oxygen. BMJ Open Gastroenterology. 2016;3(1).
658 48. Oliveira FS, Brestelli J, Cade S, Zheng J, Iodice J, Fischer S, et al.
659 MicrobiomeDB: a systems biology platform for integrating, mining and analyzing
660 microbiome experiments. Nucleic Acids Res. 2017;46(D1):D684-D91.
661

662

663

664

665

666

667

668

669

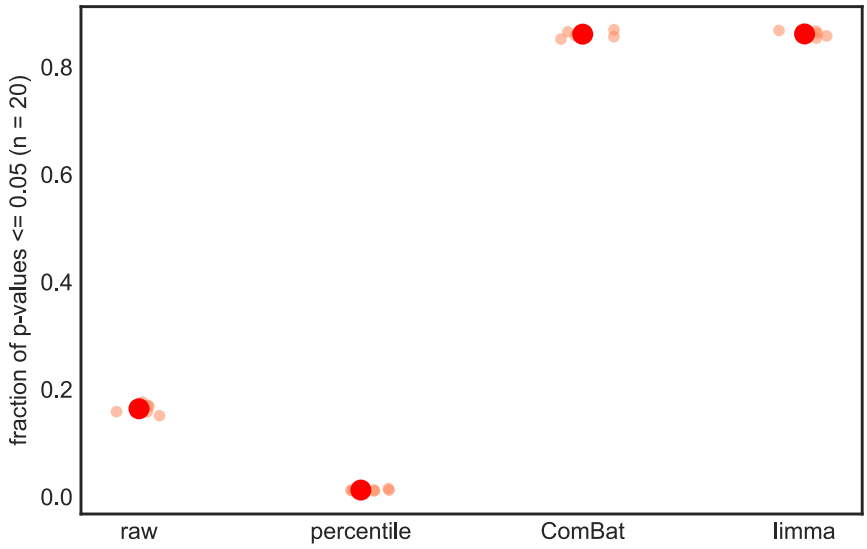
670

671

672

673

674 **Supplemental Figure**



675

676 **Figure S1. False positive rates are reduced by percentile-normalization, but**
677 **not by ComBat or limma, in the presence of low-abundance OTUs.** Random
678 sets of 40 Baxter controls and random sets of 40 Zeller controls were selected for
679 null case-control comparisons (20 iterations). Smaller points show the fraction of
680 p-values ≤ 0.05 within a given iteration, while larger dots show the average value
681 across all 20 iterations. Within each category, smaller points are randomly jittered
682 along the x-axis for better visualization. The fraction of p-values ≤ 0.05 is highly
683 inflated for all methods except percentile-normalization (red dashed line shows
684 the null-expectation for p-values). All OTUs were included in this analysis (i.e. no
685 abundance filter prior to running tests). ComBat and limma show highly skewed
686 p-value distributions when including low-abundance OTUs.
687