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 multiple 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 ComBat, a widely used batch-correction model developed for RNA microarray data, and traditional meta-analysis methods for combining independent p-values. Overall, we show that percentile-normalization is a simple, model-free approach for removing batch effects and improving sensitivity in case-control meta-analyses.

Author Summary

Batch effects present a significant obstacle to comparing results across independent 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 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 inadvisable to directly compare non-normalized data across studies.

Several tools for reducing batch effects in RNA expression 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. SVA is part of a family of batch-correction methods that use different varieties of factor analysis or singular value decomposition [3-5]. The most relied upon method to date [6], called ComBat, uses a Bayesian approach to estimate location and scale parameters for each feature within a batch [7]. These methods are most effective when batch effects are not conflated with the true biological effects [1]. Furthermore, these methods work best when batch effects are not diffuse and can be projected onto a low-dimensional manifold.

Unfortunately, batch effects are often diffuse and conflated with biological signals [8-10], limiting the usefulness of these methods. This is especially true for low-biomass samples in microbiome sequencing studies, like samples taken from the built environment [11], where the biological signal is relatively weak and batch effects can be quite large [12]. 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 [13, 14]. These meta-analysis techniques

are robust to batch effects across independent studies. However, in cases where pooling raw data across studies might increase statistical power to detect subtle differences or in cases where batches are not statistically independent of one another (e.g. multiple sequencing runs within the same study), these methods fall short.

Here, we describe a simple data-normalization procedure for controlling batch effects in case-control microbiome 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 (e.g. bacterial taxon) with sufficient representation in the data, 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 effectively removed. Upon conversion to percentiles of the within-study controls, percentile-normalized samples from multiple studies with similar case-control definitions can be appropriately pooled for statistical testing. We show that this approach controls batch effects in microbiome case-control studies and we compare this method to pooling non-normalized relative abundance data, pooling ComBat-corrected data, and to Fisher's and Stouffer's methods for combining p-values from unpooled analyses.

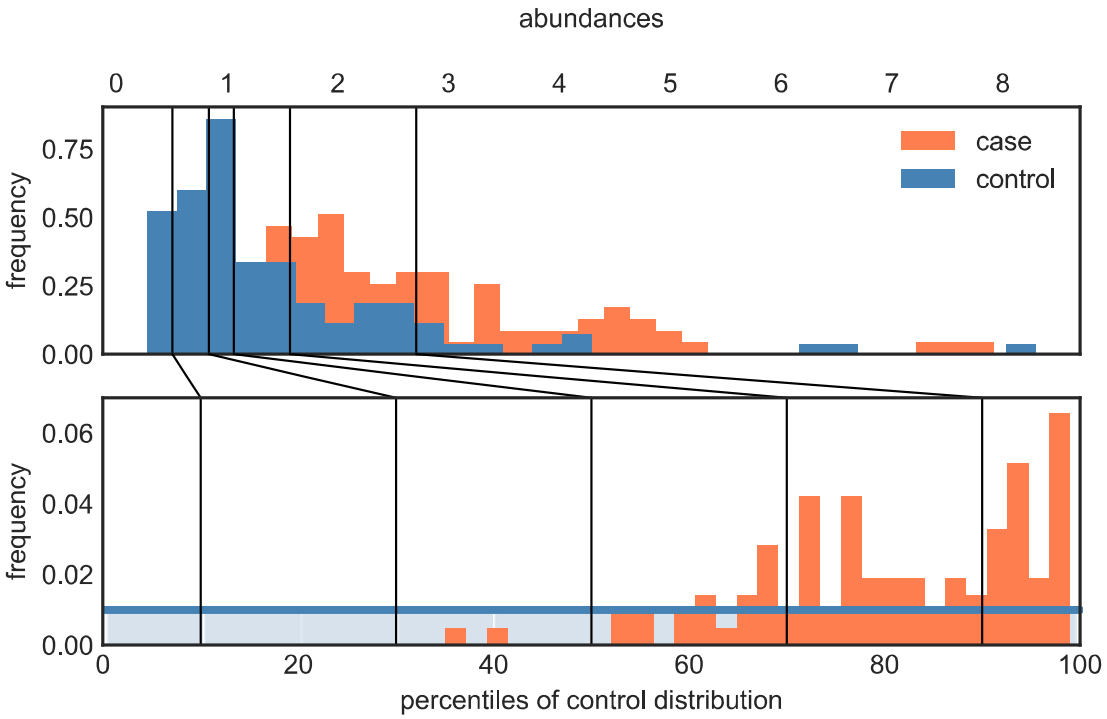

Figure 1. Theoretical feature abundance distributions for the control samples (blue) and case samples (orange) are shown in the upper panel. Converting the control distribution into percentiles of itself naturally gives rise to a uniform distribution (blue horizontal line in lower panel), while converting 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 (10^{th}, 30^{th}, 50^{th}, 70^{th}, and 90^{th} 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 [15] to validate our batch-normalization method. We focused our

analyses on studies spanning four diseases: colorectal cancer (CRC), Crohn's Disease (CD), Ulcerative Colitis (UC), and *Clostridium difficile* induced diarrhea (CDI). For a subset of three CRC studies [16-18], we were able to obtain sequence data from the same region of the 16S gene so that these data could be processed together. The remaining MicrobiomeHD case-control datasets were processed separately using the same pipeline, and then Operational Taxonomic Units (OTUs) were summarized at the genus level for comparison across studies.

Sequence Data Processing

To perform OTU-level analyses across the CRC studies, we downloaded the raw data from all of the MicrobiomeHD datasets that sequenced the V4 region of the 16S gene. We quality filtered and length trimmed each V4 dataset as described in [15] and concatenated these raw, trimmed FASTQ files into one file. We removed any unique sequences that did not appear more than 20 times and clustered the remaining reads with USEARCH [19] at 97% similarity. We assigned these OTUs taxonomic identifiers using the RDP classifier [20] with a cutoff of 0.5.

For genus-level analyses, data and metadata were acquired from the MicrobiomeHD database (https://doi.org/10.5281/zenodo.569601). Raw data were downloaded from the original studies and processed through our in-house 16S-processing pipeline (https://github.com/thomasgurry/amplicon_sequencing_pipeline) as described in [15]. Each study's OTU table was converted to relative abundance by dividing

each sample by the total number of reads and collapsed to genus level by summing all OTUs with the same genus, throwing out any OTUs which did not have a genus label.

To plot data in ordination space, Bray-Curtis distances were calculated from relative abundance data using Scikit-learn (sklearn.metrics.pairwise.pairwise_distances; metric='braycurtis) [21]. Non-metric multidimensional scaling (NMDS) coordinates were calculated for two axes based on Bray-Curtis distances using Scikit-learn (sklearn.manifold.MDS; n_components=2, metric=False, max_iter=500, eps=1e-12, dissimilarity='precomputed').

Percentile Normalization

Empirical relative abundance distributions were converted to percentiles using the SciPy v 0.19.0 [22] stats.percentileofscore method (kind='mean'). Within each study, control distributions for each individual OTU or genus were converted into percentiles of themselves and case distributions were converted into percentiles of their corresponding control distribution (Fig. 1). We restricted our analysis to OTUs that occurred in at least one third of control *or* one third of case samples in order to avoid statistical artifacts due to sampling effects. We have written a python script that performs percentile-normalization given an OTU table, a list of case sample IDs, and a list of control sample IDs as inputs (https://github.com/seangibbons/percentile_normalization)

147 *ComBat*

148 For each disease, we applied ComBat [6] to the case-control data sets analyzed
149 in this study. Relative abundances (OTUs in the CRC analysis or OTUs
150 collapsed to the genus level in the genus-level analysis) were log-transformed
151 prior to running ComBat, adding a pseudocount of 1.0 to replace zeros in the
152 OTU count matrix. ComBat-corrected data were then transformed back from log-
153 space (i.e. exponential transformation) prior to downstream analyses.

154

155 *Statistical Analysis*

156 We used the Wilcoxon rank-sum test, as implemented in SciPy v0.19.0
157 (sicipy.stats.ranksums) [22], to determine significant differences between
158 independent groups of samples. Wilcoxon tests were calculated either within or
159 across studies. In order to calculate statistics across datasets, case and control
160 samples from multiple studies of the same disease were combined together into
161 the same OTU table. Hereafter, combining datasets is referred to as 'pooling.' P-
162 values were multiple-test corrected using the Benjamini-Hochberg False
163 Discovery Rate (FDR) procedure, as implemented in StatsModels v 0.8.0
164 (statsmodels.sandbox.stats.multicomp.multipletests) [23]. Differences in overall
165 community structure were assessed using the Permutational Multivariate
166 Analysis of Variance (PERMANOVA) test in R's vegan package [24] as
167 implemented in scikit-bio (skbio.stats.distance.permanova). Fisher's and
168 Stouffer's methods for combining p-values were performed using SciPy v0.19.0
169 (scipy.stats.combine_pvalues; method='fisher' or method='stouffer'). For

Stouffer's method, weights for each study were defined as the square root of the number of cases plus the number of controls.

Results

Batch effects at OTU-level resolution

To minimize possible biases across data sets, we identified three colorectal cancer (CRC) studies that sequenced the same region of the 16S gene (V4). We reprocessed the raw sequence data from each study in the same quality filtering and OTU picking pipeline to obtain bioinformatically-standardized results. OTUs that occurred in at least a third of case or a third of control samples (i.e. either within individual studies or across studies) were retained for all downstream statistical analyses. Despite standardizing the computational processing of these data, we saw significant batch effects in healthy patients across studies (PERMANOVA $p < 0.001$; Fig. 2). The similarity between samples from the Baxter and Zackular studies is due to the fact that they were sourced from the same patient cohort, making this comparison a good pseudo-negative control for batch effects [16, 18]. There was an apparent reduction in the batch effect after applying ComBat, although differences between batches remained statistically significant (PERMANOVA $p < 0.001$; Fig. 2) [6]. Due to the non-independence between the Baxter and Zackular patient cohorts, we removed the smaller of the two studies (Zackular) from all downstream analyses. Out of a total of 5,585 OTUs found in healthy controls, 725 OTUs differed significantly in relative

abundance between the Baxter et al. (2016) and Zeller et al. (2014) controls (FDR q <= 0.05).

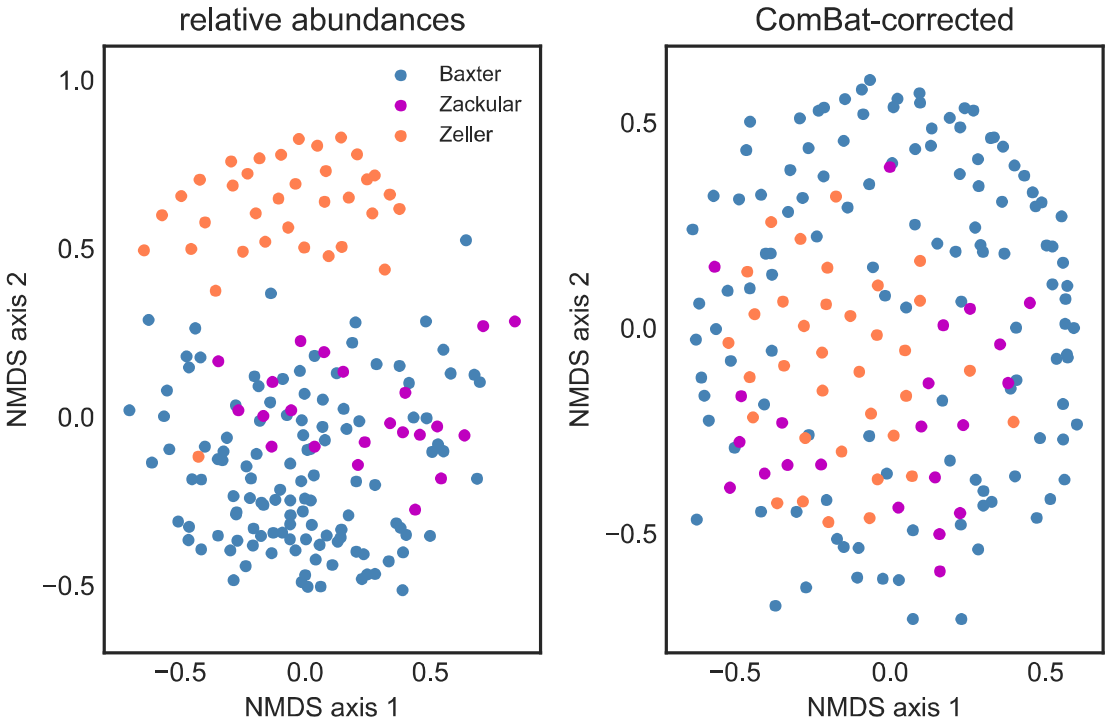

Figure 2. Non-metric multimentional scaling (NMDS) plot showing the distribution of healthy controls from three colorectal cancer studies in ordination space (Bray-Curtis distances of relative abundance data). Despite standardized bioinformatic processing, healthy patients differed significantly in their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

As expected for the Wilcoxon rank-based statistical test, within-study results at the OTU level were identical before and after percentile-normalization. In addition, these within-study results were also identical with the results from ComBat-corrected data. In the Baxter study, there were 172 healthy (control) samples and 120 CRC (case) samples, with 3 OTUs (from *Parvimonas*,

Porphyromonas, and *Peptostreptocuccus* genera) showing significant differences in abundance between cases and controls for all analyses (FDR q <= 0.05). For Zeller, there were 71 control and 40 case samples, with 4 OTUs (from *Fusobacterium*, *Closridium XIVa*, *Peptostreptococcus*, and *Dialister* genera) that differed significantly across cases and controls for all analyses (FDR q <= 0.05).

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 (Fig. 3). Case and control groups were subsampled to 30 samples each. Control samples were substituted by samples from another study along a fractional gradient (0-100% control samples from another study; see conceptual outline in Fig. 3). For the relative abundance data (non-normalized), the number of significant OTUs greatly increased due to batch effects as more control samples were substituted in from another study. However, the ComBat-corrected and percentile-normalized results were almost totally unaffected by the proportion of control samples coming from another study, indicating that batch effects were no longer driving spurious associations in the normalized data.

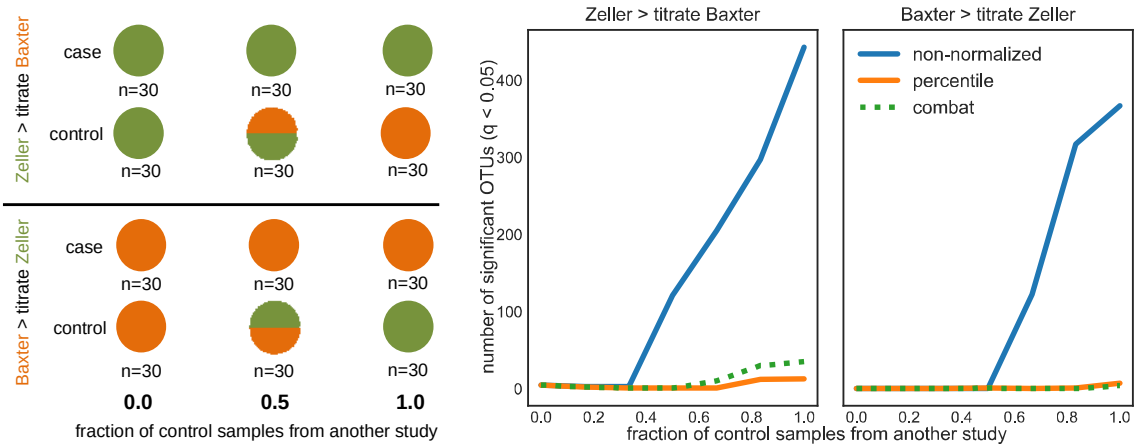

Figure 3. *In silico* titration experiment, where the control group from one study is gradually substituted with randomly chosen control samples from another study (non-normalized, percentile-normalized, and ComBat-corrected), keeping the total number of case and control samples fixed at n=30 (see conceptual illustration on the left). Mixing non-normalized data from control samples from another study often gave rise to spurious significant results due to technical differences across studies (blue lines). However, when the data were percentile-normalized or ComBat-corrected, we did not see a large increase in significant OTUs as control samples from different studies were mixed in (solid orange and green dashed lines).

In the absence of batch effects, pooling data across datasets of the same disease should increase sensitivity to detect significant associations. We pooled relative abundances, percentile-normalized abundances, and ComBat-corrected abundances, respectively, across the Baxter and Zeller studies to look for OTUs that differed significantly across cases and controls. These pooled results were then compared to classic methods for combining p-values from each dataset's individual results. For the relative abundance data, we found six OTUs (from *Porphyromonas*, *Fusobacterium*, *Clostridium XIVa*, *Peptostreptococcus*, *Dialister*, and *Parvimonas* genera) that differed significantly across cases and

controls (FDR q <= 0.05). After pooling the percentile-normalized data, we found seven OTUs that were significantly enriched in cancer patients relative to controls -- two OTUs from the *Clostridium XlVa* genus, one from *Parvimonas*, one from *Peptostreptococcus*, one from *Porphyromonas*, one from *Dialister*, and one from *Fusobacterium* (FDR q <= 0.05). The pooled ComBat-corrected results included the same significant hits identified in the percentile-normalization results. Fisher's method identified just two significant OTUs from the *Peptostreptococcus* and *Parvimonas* genera, which were also found in the pooled results. Stouffer's method identified one significant OTU from the *Peptostreptococcus* genus, which was also identified in the pooled results. Overall, the pooled methods maximize statistical power to detect significant OTUs over traditional meta-analysis methods. For example, OTUs from *Fusobacterium, Porphyromonas*, *Clostridium XIVa* and *Dialister* genera were identified as significantly enriched in CRC patients by the normalization methods but not by Fisher's or Stouffer's methods. Previous meta-analyses of CRC microbiome studies have shown these genera to be consistently associated with CRC, which supports our findings [15, 25].

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 four diseases - *Clostridium difficile induced diarrhea* (CDI), Crohn's disease (CD), ulcerative colitis (UC), and CRC - across 11 case-control studies. There were a total of 306 unique genera detected across studies. There

were two CDI case-control studies: Schubert et al. (2014) had 154 control and 93 case samples [26]; Vincent et al. (2013) had 25 control and 25 case samples [27]. There were four inflammatory bowel disease (IBD) studies that included CD patients and three that also included UC patients: Papa et al. (2012) had 24 non-IBD control sample, 23 CD samples, and 43 UC samples [28]; Morgan et al. (2012) had 18 control, 61 CD and 47 UC samples [29]; Willing et al. (2010) had 35 control, 16 UC and 29 CD samples [30]; Gevers et al. (2014) had 16 non-IBD control and 146 CD samples, with no UC samples [31]. There were four independent CRC studies, including the Baxter and Zeller studies listed in the OTU-level analysis (see above for sample sizes). The remaining two CRC studies added to the genus-level analysis are Wang et al. (2012), which had 54 control and 44 case samples [32], and Chen et al. (2012), which had 22 controls and 21 cases [33].

The number of genera that differed significantly across cases and controls changed depending on how the data were normalized, pooled, and analyzed (Table 1). Wilcoxon rank-sum tests yielded identical within-study results for non-normalized and percentile-normalized data. However, unlike the OTU-level analysis, within-study ComBat-corrected results showed fewer significant genera than the non-normalized results for unpooled, within-study tests (Table 1). Thus, in correcting for batch effects, ComBat appears to smooth out some biological signal. While pooling non-normalized data across studies is technically inappropriate, it frequently resulted in significant hits that were consistent with percentile-normalized results, suggesting that the biological signal was often

293 stronger than the batch effect. Except in the case of UC, pooling percentile-
294 normalized data consistently yielded more significant hits than pooling non-
295 normalized data (see 'across' column in Table 1). ComBat-correction generally
296 resulted in many fewer significant genera after pooling, especially for CD and
297 UC. Half of the IBD studies included non-IBD patients with inflammatory
298 symptoms as controls rather than clinically healthy patients. These biologically
299 relevant differences in inflammatory symptoms between control cohorts were
300 conflated with batches and were likely smoothed out by ComBat. In all cases,
301 Fisher's and Stouffer's methods identified fewer significant results than pooling
302 the percentile-normalized data. These results illustrate that pooling data is more
303 sensitive than classic meta-analysis techniques [34] and that percentile-
304 normalization further increases the statistical power to detect differences while
305 controlling for batch effects.
306

disease	normalization	studies	within	across	shared	Fisher	Stouffer
CDI	none	2	31	33	30	30	29
	ComBat	2	29	33	29	29	27
	percentile	2	31	33	30	30	29
CD	none	4	3	16	1	0	0
	ComBat	4	0	0	0	0	0
	percentile	4	3	18	3	0	0
UC	none	3	8	10	6	3	0
	ComBat	3	3	3	1	1	0
	percentile	3	8	9	6	0	0
CRC	none	4	7	4	2	3	0
	ComBat	4	2	4	1	0	0
	percentile	4	7	7	3	3	0

307

Table 1. Numbers of taxa that differ significantly between cases and controls for four diseases. The normalization column indicates how the data were treated prior to running significance tests (non-normalized, ComBat -corrected, or percentile-normalization). In the 'disease' column, 'CD' = Crohn's Disease, UC = Ulcerative Colitis, CRC = Colorectal Cancer, and CDI = *Clostridium difficile* induced diarrhea. The significance threshold used was q <= 0.05 (FDR). The 'within' column shows how many significant taxa were identified when running the statistics for each study independently, while the 'pooled' column shows the number of significant taxa identified when running the statistics on the combined datasets. The 'shared' column shows how many taxa overlap between the 'within' and 'pooled' columns. The 'Fisher' and 'Stouffer' columns show the number of significant taxa identified using Fisher's and Stouffer's methods for combining p-values from the independent within-study tests.

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 and also at OTUs that were significant across pooled studies but not within a given study. We ran this analysis on the CRC data, where we had the largest number of independent studies with consistently defined healthy control cohorts (n = 4). There were two genera that were significant within a subset of studies, but not across all studies after pooling (Fig. 4). *Lachnospira* was absent in three out of the five CRC studies and was enriched in controls in the two studies where it was detected. *Flavonifractor* was more abundant in cases for two studies, but this signal was not consistent across all studies. Thus, these taxa were either too rare or sensitive to different experimental and/or processing techniques to be reliable biomarkers. There were five genera that showed significant differences after pooling studies together but

were not significant in any individual study. *Escherichia/Shigella*, *Enterobacter*, and *Desulfovibrio* genera were slightly enriched in CRC patients in most studies, but did not show a statistically significant enrichment in any individual study (Fig. 5). Conversely, *Clostridium XVIII* and *Lachnospiraceae incertae sedis* genera were enriched in controls across most studies. These OTUs show small, yet consistent effect sizes across independent studies that can only be detected after pooling (Fig. 5).

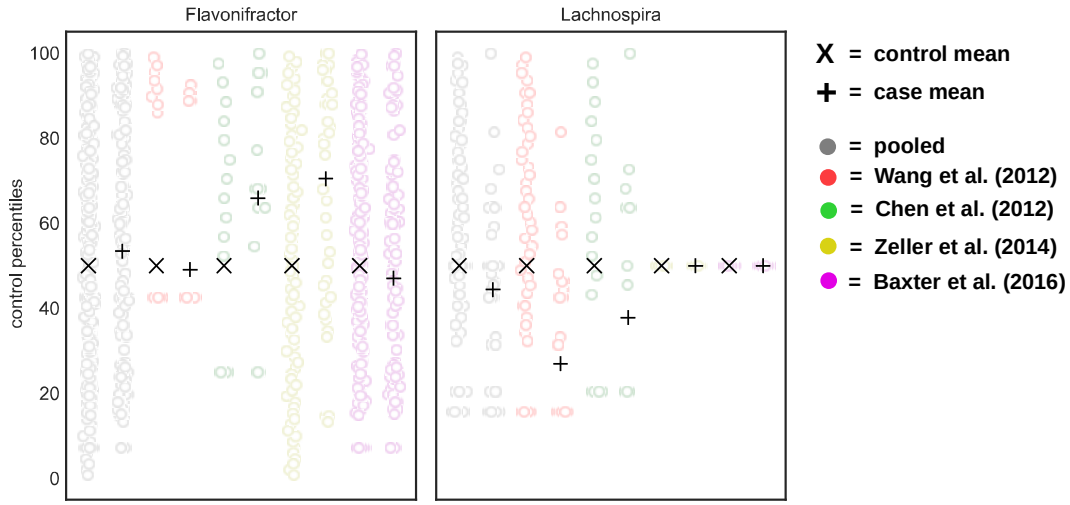

Figure 4. The *Flavonifractor* and *Lachnospira* genera showed significant differences between cases and controls within a study (FDR q <= 0.05), but not after pooling across CRC studies.

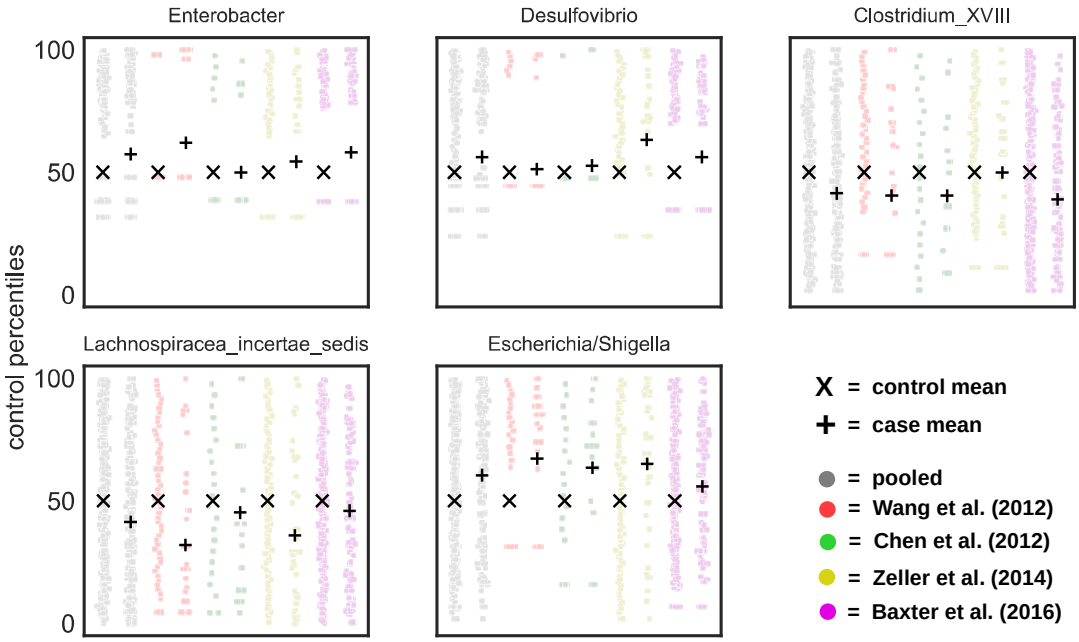

Figure 5. Five genera did not show a significant difference between cases and controls within an individual study, but were significantly different when pooling across CRC studies (FDR q <= 0.05). Scatter plots show distributions of percentile-normalized data for case and control samples across studies.

Discussion

Batch effects are unavoidable when working with high-throughput data generation platforms. The RNA microarray community has been proactive in the development of tools for dealing with these effects [1, 6]. However, these tools are not as effective when batch effects are confounded with biological signals, or when these effects cannot be projected onto a small number of dimensions, which is often the case in microbiome case-control studies [35-37]. Fortunately, case-control studies can be internally normalized by their own control samples. Any study-specific batch effects in the case samples will be present in the control

359 samples, and by converting the case data into percentiles of the control
360 distribution these effects are removed.
361 Relative abundance data -- but not the percentile-normalized or ComBat-
362 corrected data -- quickly gave spurious results when cases from one study were
363 tested against controls from another (Fig. 3). For studies with small numbers of
364 control and/or case samples, it is tempting to pool with other datasets to increase
365 statistical power. In the past, pooling of non-normalized data from different
366 studies has been done [31, 35, 38], but as demonstrated above, this is
367 inadvisable. In these scenarios, datasets can first be percentile-normalized and
368 then appropriately combined without introducing batch-related artifacts.
369 We found substantial overlap in the relative abundance and percentile-
370 normalized results when calculating significance across studies. This overlap is
371 expected when there is a strong biological signal that overrides batch effects
372 [39]. Despite the similarity between pooled relative abundance and percentile-
373 normalized results, there were several cases where the percentile-normalized
374 results identified significant differences between cases and controls that were
375 missed in the non-normalized results and there was one instance (UC) where
376 one fewer significant difference was found in the percentile-normalized results
377 (Table 1). Percentile-normalization also identified more significant hits than
378 ComBat-corrected data in the genus-level pooled analyses, especially for UC
379 and CD (IBD). The reduced number of significant hits from ComBat-corrected
380 data for IBD was likely due to heterogeneous control cohorts across these

studies (i.e. healthy patients vs. non-IBD patients), which likely smoothed-out inflammation-associated signals.

We compared normalization and pooling methods (i.e. percentile-normalization and ComBat) to Fisher's and Stouffer's methods for combining p-values. Stouffer's method is similar to Fisher's, but includes weights for each p-value based on the number of samples in a study. For all diseases, the pooling methods identified a larger number of significant hits than Stouffer's and Fisher's methods, indicating that pooling provides more sensitivity to detect differences between cases and controls. The bacterial taxa identified as significant by the percentile-normalization method were largely consistent with prior results [15].

In conclusion, we present a robust, model-free procedure for transforming each feature in a case-control dataset into percentiles of its control distribution (Fig. 1). These percentile-normalized features can be pooled across independent studies for non-parametric, univariate statistical testing, circumventing the batch effect problem. We find that this procedure allows us to identify differences between cases and controls that are often missed by more conservative meta-analysis techniques. Methods developed for batch-correction in microarray data, like ComBat, can reduce batch effects in microbiome studies (Fig. 2-3), but may also obscure real patterns if batch effects are not totally independent of biological signals. We suggest that ComBat and other similar methods are useful for studies without case and control groups. However, when studies have internal controls, percentile-normalization should be the preferred batch correction approach.

Acknowledgements

This work was supported by the Center for Microbiome Informatics and Therapeutics. We thank the members of the Alm lab for helpful feedback.

References

1. Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, et al. Tackling the widespread and critical impact of batch effects in high-throughput data. Nat Rev Genet. 2010;11(10):733-9.
2. Schloss PD, Gevers D, Westcott SL. Reducing the Effects of PCR Amplification and Sequencing Artifacts on 16S rRNA-Based Studies. PLoS One. 2011;6(12):e27310. doi: 10.1371/journal.pone.0027310.
3. Leek JT, Storey JD. Capturing heterogeneity in gene expression studies by surrogate variable analysis. PLoS Genet. 2007;3(9):e161.
4. Alter O, Brown PO, Botstein D. Singular value decomposition for genome-wide expression data processing and modeling. Proc Natl Acad Sci USA. 2000;97(18):10101-6.
5. Benito M, Parker J, Du Q, Wu J, Xiang D, Perou CM, et al. Adjustment of systematic microarray data biases. Bioinformatics. 2004;20(1):105-14.
6. Chen C, Grennan K, Badner J, Zhang D, Gershon E, Jin L, et al. Removing Batch Effects in Analysis of Expression Microarray Data: An Evaluation of Six Batch Adjustment Methods. PLoS One. 2011;6(2):e17238. doi: 10.1371/journal.pone.0017238.
7. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007;8(1):118-27.
8. Weiss S, Amir A, Hyde ER, Metcalf JL, Song SJ, Knight R. Tracking down the sources of experimental contamination in microbiome studies. Genome Biol. 2014;15(12):564. doi: 10.1186/s13059-014-0564-2.
9. Shen H, Rogelj S, Kieft TL. Sensitive, real-time PCR detects low-levels of contamination by Legionella pneumophila in commercial reagents. Mol Cell Probes. 2006;20. doi: 10.1016/j.mcp.2005.09.007.
10. Nguyen NH, Smith D, Peay K, Kennedy P. Parsing ecological signal from noise in next generation amplicon sequencing. New Phytol. 2015;205(4):1389-93. doi: 10.1111/nph.12923.
11. Gibbons SM. The Built Environment Is a Microbial Wasteland. mSystems. 2016;1(2):e00033-16.
12. Chase J, Fouquier J, Zare M, Sonderegger DL, Knight R, Kelley ST, et al. Geography and Location Are the Primary Drivers of Office Microbiome Composition. mSystems. 2016;1(2). doi: 10.1128/mSystems.00022-16.

13. Fisher RA. Statistical methods for research workers: Genesis Publishing Pvt Ltd; 1925.
14. Stouffer SA. Adjustment during army life: Princeton University Press; 1949.
15. Duvallet C, Gibbons SM, Gurry T, Irizarry RA, Alm EJ. Meta Analysis Of Microbiome Studies Identifies Shared And Disease-Specific Patterns. bioRxiv. 2017:134031.
16. Baxter NT, Ruffin MT, Rogers MA, Schloss PD. Microbiota-based model improves the sensitivity of fecal immunochemical test for detecting colonic lesions. Genome Med. 2016;8(1):37.
17. Zeller G, Tap J, Voigt AY, Sunagawa S, Kultima JR, Costea PI, et al. Potential of fecal microbiota for early□stage detection of colorectal cancer. Molec Sys Biol. 2014;10(11):766.
18. Zackular JP, Rogers MA, Ruffin MT, Schloss PD. The human gut microbiome as a screening tool for colorectal cancer. Cancer Prev Res. 2014;7(11):1112-21.
19. Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinformatics. 2010;26(19):2460-1.
20. Wang Q, Garrity GM, Tiedje JM, Cole JR. Naïve Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy. Appl Environ Microbiol. 2007;73(16):5261-7. doi: 10.1128/aem.00062-07.
21. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. J Mach Learn Res. 2011;12(Oct):2825-30.
22. Jones E, Oliphant T, Peterson P. SciPy: Open source scientific tools for Python, 2001–. URL http://www scipy org. 2007;73:86.
23. Seabold S, Perktold J, editors. Statsmodels: Econometric and statistical modeling with python. Proceedings of the 9th Python in Science Conference; 2010.
24. Oksanen J. Multivariate analysis of ecological communities in R: vegan tutorial. R package version. 2011;1(7).
25. Tjalsma H, Boleij A, Marchesi JR, Dutilh BE. A bacterial driver–passenger model for colorectal cancer: beyond the usual suspects. Nat Rev Microbiol. 2012;10(8):575-82.
26. Schubert AM, Rogers MA, Ring C, Mogle J, Petrosino JP, Young VB, et al. Microbiome data distinguish patients with Clostridium difficile infection and non-C. difficile-associated diarrhea from healthy controls. mBio. 2014;5(3):e01021-14.
27. Vincent C, Stephens DA, Loo VG, Edens TJ, Behr MA, Dewar K, et al. Reductions in intestinal Clostridiales precede the development of nosocomial Clostridium difficile infection. Microbiome. 2013;1(1):18.
28. Papa E, Docktor M, Smillie C, Weber S, Preheim SP, Gevers D, et al. Non-invasive mapping of the gastrointestinal microbiota identifies children with inflammatory bowel disease. PLoS One. 2012;7(6):e39242.

29. Morgan XC, Tickle TL, Sokol H, Gevers D, Devaney KL, Ward DV, et al. Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. Genome Biol. 2012;13(9):R79.
30. Willing BP, Dicksved J, Halfvarson J, Andersson AF, Lucio M, Zheng Z, et al. A pyrosequencing study in twins shows that gastrointestinal microbial profiles vary with inflammatory bowel disease phenotypes. Gastroenterology. 2010;139(6):1844-54. e1.
31. Gevers D, Kugathasan S, Denson LA, Vázquez-Baeza Y, Van Treuren W, Ren B, et al. The treatment-naive microbiome in new-onset Crohn's disease. Cell Host Microbe. 2014;15(3):382-92.
32. Wang T, Cai G, Qiu Y, Fei N, Zhang M, Pang X, et al. Structural segregation of gut microbiota between colorectal cancer patients and healthy volunteers. ISME J. 2012;6(2):320-9.
33. Chen W, Liu F, Ling Z, Tong X, Xiang C. Human intestinal lumen and mucosa-associated microbiota in patients with colorectal cancer. PLoS One. 2012;7(6):e39743.
34. Glass GV. Primary, secondary, and meta-analysis of research. Educ Res. 1976;5(10):3-8.
35. Ross MC, Muzny DM, McCormick JB, Gibbs RA, Fisher-Hoch SP, Petrosino JF. 16S gut community of the Cameron County Hispanic Cohort. Microbiome. 2015;3(1):7.
36. Pasolli E, Truong DT, Malik F, Waldron L, Segata N. Machine learning meta-analysis of large metagenomic datasets: tools and biological insights. PLoS Comput Biol. 2016;12(7):e1004977.
37. Escobar JS, Klotz B, Valdes BE, Agudelo GM. The gut microbiota of Colombians differs from that of Americans, Europeans and Asians. BMC Microbiol. 2014;14(1):311.
38. Dubourg G, Lagier J-C, Hüe S, Surenaud M, Bachar D, Robert C, et al. Gut microbiota associated with HIV infection is significantly enriched in bacteria tolerant to oxygen. BMJ Open Gastroenter. 2016;3(1).
39. Forslund K, Hildebrand F, Nielsen T, Falony G, Le Chatelier E, Sunagawa S, et al. Disentangling type 2 diabetes and metformin treatment signatures in the human gut microbiota. Nature. 2015;528(7581):262-6.