Supplementary Information for Analyzing Microbiome Data with Taxonomic Misclassification using a Zero-Inflated Dirichlet-Multinomial Model

Matthew D. Koslovsky
Department of Statistics, Colorado State University

S1 MCMC Sampler Details

Jointly update α_{it} and ζ_{it} with either a Contract or Expand Step:

- Contract Step: If $\zeta_{it} = 1$, propose $\zeta'_{it} = 0$ and $\alpha'_{it} = 0$ and accept with probability

$$\min\left\{\frac{\prod_{l=1}^{\dot{z}_i} p(\boldsymbol{z}_{il}|\boldsymbol{\Theta}'_i)p(\zeta'_{it})q(\zeta_{it},\alpha_{it}|\zeta'_{it},\alpha'_{it})}{\prod_{l=1}^{\dot{z}_i} p(\boldsymbol{z}_{il}|\boldsymbol{\Theta}_i)p(\zeta_{it})p(\alpha_{it}|\zeta_{it},\gamma_{it})}, 1\right\},$$

  where $q(\cdot|\cdot) \sim \zeta_{it}\text{Gamma}(\gamma_{it}, 1) + (1-\zeta_{it})\delta_0(\cdot)$.

- Expand Step: If $\zeta_{it} = 0$, propose $\zeta'_{it} = 1$ and $\alpha'_{it} \sim \text{Gamma}(\gamma_{it}, 1)$ and accept with probability

$$\min\left\{\frac{\prod_{l=1}^{\dot{z}_i} p(\boldsymbol{z}_{il}|\boldsymbol{\Theta}'_i)p(\alpha'_{it}|\zeta'_{ij},\gamma_{ij})p(\zeta'_{it})}{\prod_{l=1}^{\dot{z}_i} p(\boldsymbol{z}_{il}|\boldsymbol{\Theta}_i)p(\zeta_{it})q(\zeta'_{it},\alpha'_{it}|\zeta_{it},\alpha_{it})}, 1\right\}.$$

Update data augmentation parameters for $\boldsymbol{\Theta}_i$:

$$p(\alpha_{it}|\cdot) \sim \text{Gamma}\left(\sum_{l=1}^{\dot{z}_i} I(z_{ilt}=1) + \gamma_{it}, 1+\mu_i\right).$$

Update auxiliary parameters for α_{it}:

$$p(\mu_i|\cdot) \sim \text{Gamma}(\dot{z}_i, \bar{\alpha}_i).$$

Update true species classification:

$$p(\boldsymbol{z}_{ic}|\cdot) \sim \text{Multinomial}(\sum_{l=1}^{\dot{z}_i} I(y_{ilc}=1), \boldsymbol{\Theta}_i \otimes \boldsymbol{\theta}_c).$$

Update regression coefficients for covariates associated with the occupancy probability:

$$p(\boldsymbol{\beta}_{\eta_t}|\cdot) \sim \text{Normal}(\tilde{\boldsymbol{\beta}}_{\eta_t}, \tilde{\Sigma}_{\eta_t}),$$

where $\tilde{\boldsymbol{\Sigma}}_{\eta_t} = (\boldsymbol{X}'_\eta \boldsymbol{\Omega}_t \boldsymbol{X}_\eta + \boldsymbol{\Sigma}_\eta^{-1})^{-1}$ and $\tilde{\boldsymbol{\beta}}_{\eta_t} = \tilde{\boldsymbol{\Sigma}}_{\eta_t}(\boldsymbol{X}'_\eta \boldsymbol{k}_t + \boldsymbol{\Sigma}_\eta^{-1}\boldsymbol{\mu}_\eta)$, with $\boldsymbol{X}_\eta$ representing the matrix of $\boldsymbol{x}_i$, $\boldsymbol{\Omega}_t$ a diagonal matrix of $\omega_{\zeta_{it}}$, $\boldsymbol{k}_t$ a vector of $(\zeta_{it} - 1/2)$, and $\boldsymbol{\Sigma}_\eta$ a $P_\eta \times P_\eta$ - dimensional diagonal matrix of σ^2_η.

Update auxiliary parameters for occupancy indicators:

$$p(\omega_{\zeta_{it}}|\cdot) \sim \text{PG}(1, \boldsymbol{x}'_i \boldsymbol{\beta}_{\eta_t}).$$

Update data augmentation parameters for classification probabilities:

$$p(a_{tc}|\cdot) \sim \text{Gamma}\left(\sum_{i=1}^{N}\sum_{l=1}^{\dot{z}_i} I(y_{ilc} = 1, z_{ilt} = 1) + \nu_{tc}, u_t + 1\right).$$

Update auxiliary parameters for classification probabilities:

$$p(u_t|\cdot) \sim \text{Gamma}\left(\sum_{i=1}^{N}\sum_{l=1}^{\dot{z}_i} I(z_{ilt} = 1), \bar{a}_t\right).$$

Update regression coefficients for relative abundances $\beta_{\gamma_{tp}}$:

- Propose $\beta'_{\gamma_{tp}} \sim \text{Normal}(\beta_{\gamma_{tp}}, 1)$.
- Accept proposal with probability

$$\min\left\{\frac{p(\alpha_t|\beta'_{\gamma_{tp}}, \boldsymbol{x}_i)p(\beta'_{\gamma_{tp}})}{p(\alpha_t|\beta_{\gamma_{tp}}, \boldsymbol{x}_i)p(\beta_{\gamma_{tp}})}, 1\right\}.$$

S2 Additional Tables and Figures

Null										
	$\beta_{\eta_{t1}} \neq 0$		$\beta_{\eta_{t1}} = 0$		$\beta_{\gamma_{t1}} \neq 0$		$\beta_{\gamma_{t1}} = 0$		$\pi_t(\boldsymbol{x}_{ip})$	
	ABS	COV	ABS	COV	ABS	COV	ABS	COV	ABS	COV
MicroMiss	0.397	0.937	0.286	0.936	0.262	0.797	0.244	0.878	0.263	0.872
ZIDM	0.479	0.942	0.407	0.925	0.453	0.586	0.254	0.894	0.349	0.781
DM	-	-	-	-	0.656	0.174	0.173	0.920	0.482	0.470
Low										
	$\beta_{\eta_{t1}} \neq 0$		$\beta_{\eta_{t1}} = 0$		$\beta_{\gamma_{t1}} \neq 0$		$\beta_{\gamma_{t1}} = 0$		$\pi_t(\boldsymbol{x}_{ip})$	
	ABS	COV	ABS	COV	ABS	COV	ABS	COV	ABS	COV
MicroMiss	0.397	0.935	0.363	0.935	0.312	0.676	0.211	0.918	0.270	0.820
ZIDM	0.597	0.842	0.495	0.910	0.876	0.015	0.126	0.968	0.601	0.312
DM	-	-	-	-	0.763	0.012	0.115	0.948	0.541	0.301
Medium										
	$\beta_{\eta_{t1}} \neq 0$		$\beta_{\eta_{t1}} = 0$		$\beta_{\gamma_{t1}} \neq 0$		$\beta_{\gamma_{t1}} = 0$		$\pi_t(\boldsymbol{x}_{ip})$	
	ABS	COV	ABS	COV	ABS	COV	ABS	COV	ABS	COV
MicroMiss	0.441	0.920	0.403	0.949	0.384	0.512	0.190	0.924	0.306	0.720
ZIDM	0.713	0.823	0.494	0.941	0.895	0.003	0.129	0.944	0.616	0.254
DM	-	-	-	-	0.774	0.005	0.103	0.952	0.547	0.265
High										
	$\beta_{\eta_{t1}} \neq 0$		$\beta_{\eta_{t1}} = 0$		$\beta_{\gamma_{t1}} \neq 0$		$\beta_{\gamma_{t1}} = 0$		$\pi_t(\boldsymbol{x}_{ip})$	
	ABS	COV	ABS	COV	ABS	COV	ABS	COV	ABS	COV
MicroMiss	0.487	0.918	0.436	0.954	0.455	0.349	0.168	0.950	0.352	0.628
ZIDM	0.850	0.876	0.390	0.978	0.880	0.014	0.159	0.801	0.610	0.224
DM	-	-	-	-	0.782	0.002	0.085	0.979	0.552	0.239

Table S1: Simulation Results: Estimation performance for $N = 50$ observations, $\dot{z}_i = 10{,}000$ reads, $T = 50$ compositional elements at varying levels of misclassification, and with overdispersion parameter $d = 0.05$. ABS - absolute value of the difference between the estimated and true parameters; COV - 0.95 coverage probabilities.

Null										
	$\beta_{\eta_{t1}} \neq 0$		$\beta_{\eta_{t1}} = 0$		$\beta_{\gamma_{t1}} \neq 0$		$\beta_{\gamma_{t1}} = 0$		$\pi_t(\boldsymbol{x}_{ip})$	
	ABS	COV	ABS	COV	ABS	COV	ABS	COV	ABS	COV
MicroMiss	0.432	0.926	0.335	0.930	0.329	0.797	0.320	0.891	0.375	0.864
ZIDM	0.570	0.920	0.510	0.918	0.425	0.699	0.267	0.942	0.350	0.849
DM	-	-	-	-	0.560	0.262	0.177	0.943	0.426	0.557
Low										
	$\beta_{\eta_{t1}} \neq 0$		$\beta_{\eta_{t1}} = 0$		$\beta_{\gamma_{t1}} \neq 0$		$\beta_{\gamma_{t1}} = 0$		$\pi_t(\boldsymbol{x}_{ip})$	
	ABS	COV	ABS	COV	ABS	COV	ABS	COV	ABS	COV
MicroMiss	0.435	0.928	0.397	0.930	0.334	0.723	0.265	0.919	0.322	0.838
ZIDM	0.572	0.840	0.474	0.903	0.774	0.058	0.138	0.982	0.544	0.454
DM	-	-	-	-	0.705	0.067	0.134	0.941	0.508	0.375
Medium										
	$\beta_{\eta_{t1}} \neq 0$		$\beta_{\eta_{t1}} = 0$		$\beta_{\gamma_{t1}} \neq 0$		$\beta_{\gamma_{t1}} = 0$		$\pi_t(\boldsymbol{x}_{ip})$	
	ABS	COV	ABS	COV	ABS	COV	ABS	COV	ABS	COV
MicroMiss	0.486	0.904	0.413	0.946	0.417	0.531	0.217	0.926	0.344	0.736
ZIDM	0.652	0.763	0.481	0.909	0.876	0.006	0.122	0.974	0.603	0.324
DM	-	-	-	-	0.761	0.022	0.113	0.950	0.541	0.302
High										
	$\beta_{\eta_{t1}} \neq 0$		$\beta_{\eta_{t1}} = 0$		$\beta_{\gamma_{t1}} \neq 0$		$\beta_{\gamma_{t1}} = 0$		$\pi_t(\boldsymbol{x}_{ip})$	
	ABS	COV	ABS	COV	ABS	COV	ABS	COV	ABS	COV
MicroMiss	0.522	0.902	0.455	0.937	0.485	0.378	0.196	0.931	0.378	0.641
ZIDM	0.776	0.739	0.484	0.926	0.908	0.002	0.121	0.975	0.623	0.265
DM	-	-	-	-	0.785	0.006	0.102	0.959	0.554	0.267

Table S2: Simulation Results: Estimation performance for $N = 50$ observations, $\dot{z}_i = 10{,}000$ reads, $T = 50$ compositional elements at varying levels of misclassification, and with overdispersion parameter $d = 0.10$. ABS - absolute value of the difference between the estimated and true parameters; COV - 0.95 coverage probabilities.

Null						
	$\beta_{\eta_{t1}}$		$\beta_{\gamma_{t1}}$		$\pi_t(\boldsymbol{x}_{ip})$	
	SENS	SPEC	SENS	SPEC	ABS	COV
MicroMissS	0.889	0.718	0.980	0.979	0.043	0.962
ZIDMbvs	0.735	0.982	1.000	0.992	0.035	0.997
DMBVS	-	-	0.036	0.977	0.206	0.535
ZINB	-	-	1.000	0.927	-	-
Low						
	$\beta_{\eta_{t1}}$		$\beta_{\gamma_{t1}}$		$\pi_t(\boldsymbol{x}_{ip})$	
	SENS	SPEC	SENS	SPEC	ABS	COV
MicroMissS	0.749	0.863	0.900	0.977	0.077	0.853
ZIDMbvs	0.000	1.000	0.141	0.996	0.195	0.359
DMBVS	-	-	0.013	0.999	0.204	0.309
ZINB	-	-	0.997	0.693	-	-
Medium						
	$\beta_{\eta_{t1}}$		$\beta_{\gamma_{t1}}$		$\pi_t(\boldsymbol{x}_{ip})$	
	SENS	SPEC	SENS	SPEC	ABS	COV
MicroMissS	0.774	0.865	0.907	0.973	0.084	0.836
ZIDMbvs	0.000	1.000	0.175	0.999	0.191	0.375
DMBVS	-	-	0.043	0.998	0.202	0.331
ZINB	-	-	0.997	0.756	-	-
High						
	$\beta_{\eta_{t1}}$		$\beta_{\gamma_{t1}}$		$\pi_t(\boldsymbol{x}_{ip})$	
	SENS	SPEC	SENS	SPEC	ABS	COV
MicroMissS	0.753	0.882	0.866	0.976	0.093	0.803
ZIDMbvs	0.000	1.000	0.247	0.996	0.186	0.366
DMBVS	-	-	0.045	1.000	0.201	0.393
ZINB	-	-	0.979	0.822	-	-

Table S3: Simulation Results: Selection performance for $N = 50$ observations, $\dot{z}_i = 10{,}000$ reads, $T = 50$ compositional elements at varying levels of misclassification, overdispersion parameter $d = 0.01$, and 50% active covariates for 10 compositional elements. SENS - sensitivity; SPEC - specificity; ABS - absolute value of the difference between the estimated and true parameters; COV - 0.95 coverage probabilities.

Null						
	$\beta_{\eta_{t1}}$		$\beta_{\gamma_{t1}}$		$\pi_t(\boldsymbol{x}_{ip})$	
	SENS	SPEC	SENS	SPEC	ABS	COV
MicroMissS	0.453	0.912	0.644	0.966	0.177	0.542
ZIDMbvs	0.652	0.945	0.903	0.961	0.065	0.797
DMBVS	-	-	0.190	0.964	0.067	0.623
ZINB	-	-	0.980	0.775	-	-
Low						
	$\beta_{\eta_{t1}}$		$\beta_{\gamma_{t1}}$		$\pi_t(\boldsymbol{x}_{ip})$	
	SENS	SPEC	SENS	SPEC	ABS	COV
MicroMissS	0.435	0.916	0.573	0.965	0.157	0.498
ZIDMbvs	0.034	0.962	0.320	0.989	0.043	0.646
DMBVS	-	-	0.172	0.991	0.046	0.590
ZINB	-	-	0.989	0.639	-	-
Medium						
	$\beta_{\eta_{t1}}$		$\beta_{\gamma_{t1}}$		$\pi_t(\boldsymbol{x}_{ip})$	
	SENS	SPEC	SENS	SPEC	ABS	COV
MicroMissS	0.455	0.913	0.619	0.959	0.162	0.521
ZIDMbvs	0.003	0.969	0.329	0.993	0.047	0.530
DMBVS	-	-	0.186	0.995	0.047	0.544
ZINB	-	-	0.986	0.643	-	-
High						
	$\beta_{\eta_{t1}}$		$\beta_{\gamma_{t1}}$		$\pi_t(\boldsymbol{x}_{ip})$	
	SENS	SPEC	SENS	SPEC	ABS	COV
MicroMissS	0.366	0.918	0.625	0.958	0.154	0.503
ZIDMbvs	0.000	0.971	0.378	0.995	0.041	0.508
DMBVS	-	-	0.322	0.996	0.041	0.531
ZINB	-	-	0.986	0.700	-	-

Table S4: Simulation Results: Selection performance for $N = 50$ observations, $\dot{z}_i = 10{,}000$ reads, $T = 250$ compositional elements at varying levels of misclassification, overdispersion parameter $d = 0.01$, and 50% active covariates for 10 compositional elements. SENS - sensitivity; SPEC - specificity; ABS - absolute value of the difference between the estimated and true parameters; COV - 0.95 coverage probabilities.

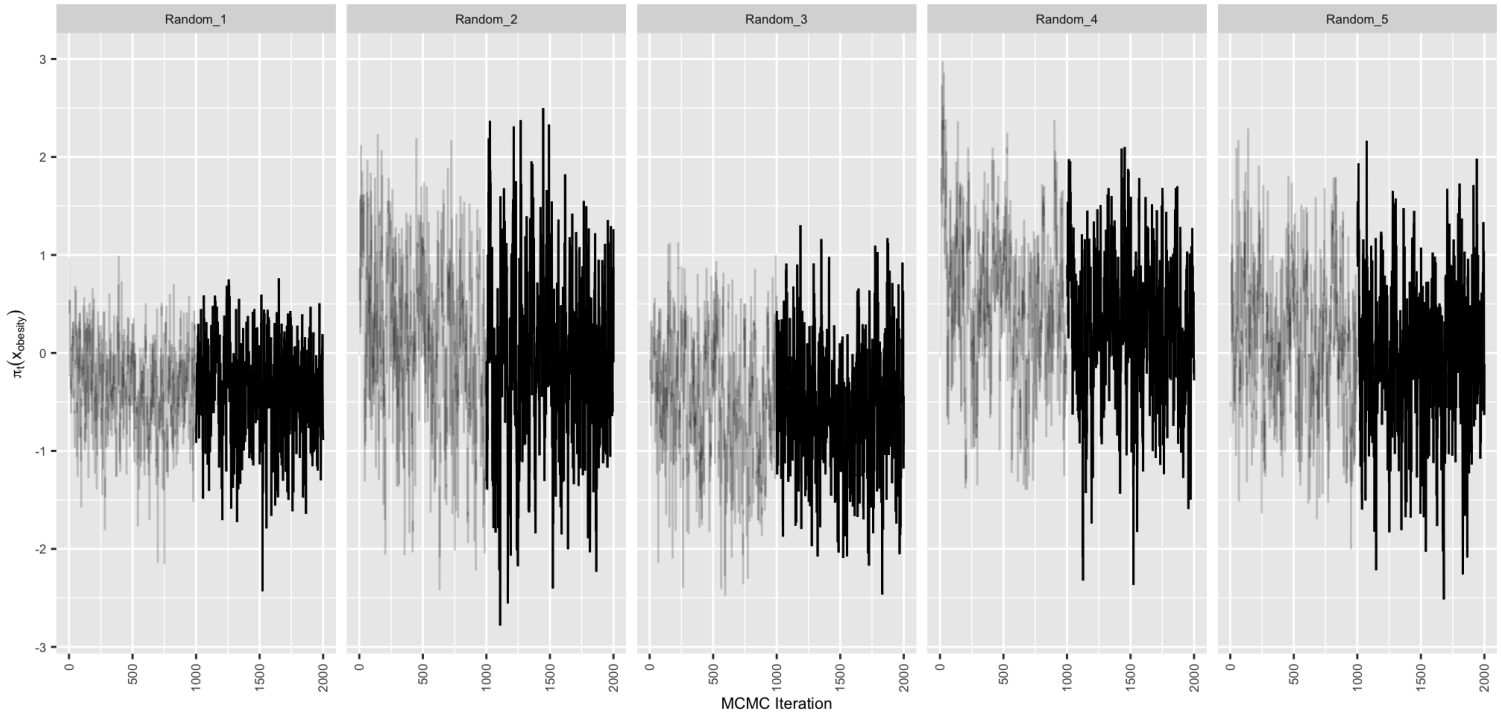

Figure S1: Posterior samples of the log multiplicative effect of being obese versus healthy for the relative abundance of 5 random taxa using the proposed MicroMiss model. The MCMC algorithm was run for 10,000 iterations, treating the first 5,000 samples as burnin and then thinning to every 5th sample for inference. Lighter grey indicates MCMC samples treated as burnin.

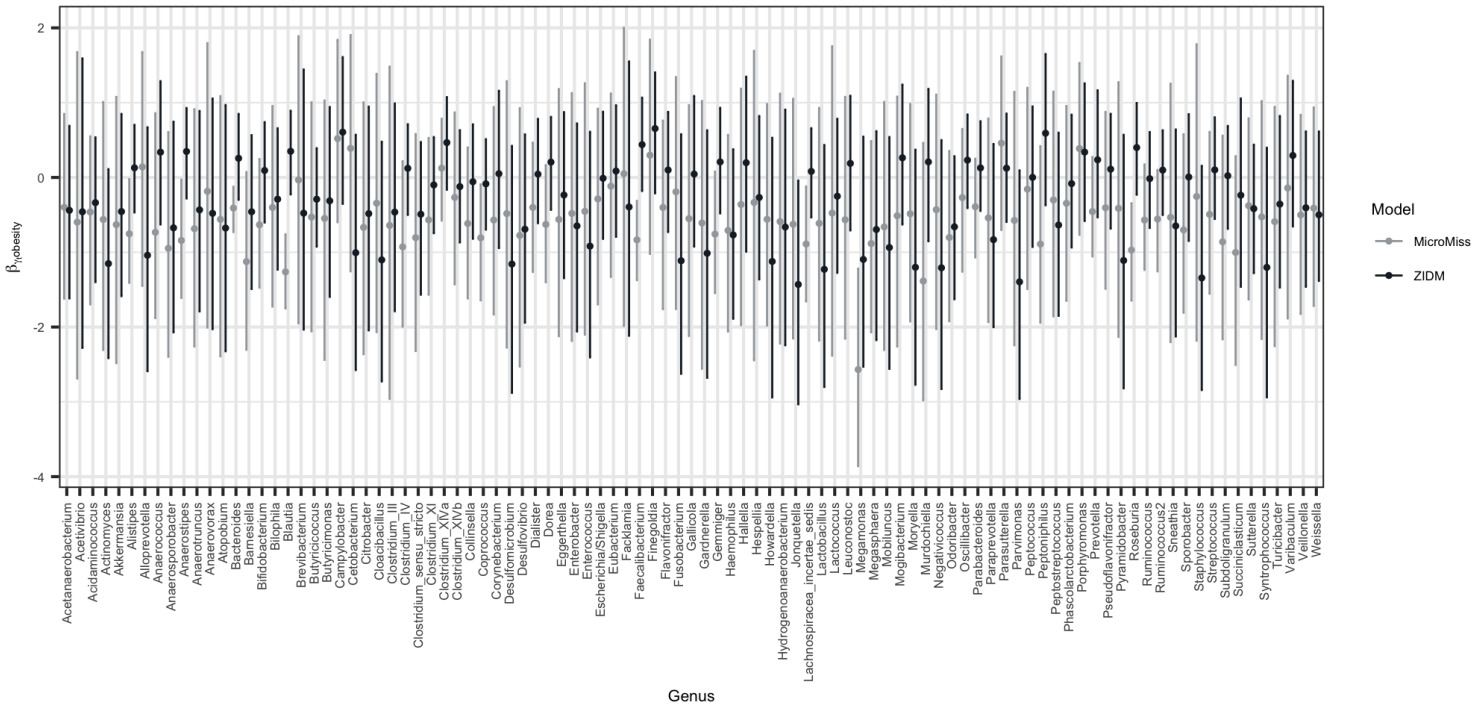

Figure S2: Posterior estimates of the log multiplicative effect of being obese versus healthy for the corresponding concentration parameters of each taxon using the proposed MicroMiss and ZIDM models. Dot represents the posterior mean with error bars capturing the 95% credible intervals.

Family	Genus	MicroMiss			ZIDM		
		Obese	Healthy	p-value	Obese	Healthy	p-value
Acidaminococcaceae	*Acidaminococcus*	0.72	1.58	0.789	0.28	2.00	0.103
	Succiniclasticum	-1.29	-2.03	0.012	-1.88	-1.61	0.454
Aerococcaceae	*Facklamia*	-6.63	0.59	< 0.001	-5.41	-2.65	< 0.001
Bacteroidaceae	*Bacteroides*	7.99	8.74	0.316	7.60	8.71	0.019
Bifidobacteriaceae	*Bifidobacterium*	3.88	5.75	0.109	3.47	6.35	0.008
Campylobacteraceae	*Campylobacter*	3.99	-0.63	< 0.001	3.58	-1.10	< 0.001
Clostridiales Incertae Sedis XI	*Anaerococcus*	4.10	-2.95	< 0.001	4.06	-1.10	< 0.001
	Finegoldia	4.37	-1.61	< 0.001	4.15	-0.80	< 0.001
	Peptoniphilus	4.68	-3.51	< 0.001	4.52	-0.75	< 0.001
Incertae Sedis XI	*Murdochiella*	2.07	-2.38	< 0.001	1.37	-1.57	0.003
Lachnospiraceae	*Anaerostipes*	3.91	5.72	0.030	3.50	6.72	< 0.001
	Blautia	5.97	8.65	< 0.001	5.29	8.79	< 0.001
	Coprococcus	3.62	5.19	0.019	4.22	5.44	0.004
	Lachnospiracea Incertae Sedis	3.83	5.82	0.005	4.27	6.63	< 0.001
	Roseburia	5.45	6.55	0.019	4.75	7.43	< 0.001
	Ruminococcus	4.59	5.92	0.033	4.61	6.67	< 0.001
Porphyromonadaceae	*Parabacteroides*	5.14	5.12	0.612	4.66	4.56	0.262
	Porphyromonas	4.25	-0.06	< 0.001	3.86	-0.02	0.001
Prevotellaceae	*Prevotella*	6.70	1.35	0.001	6.74	2.64	0.001
Rikenellaceae	*Alistipes*	4.50	5.54	0.045	3.97	5.80	0.003
Ruminococcaceae	*Clostridium IV*	1.95	4.10	0.010	2.92	5.63	< 0.001
	Faecalibacterium	6.01	8.31	0.004	5.79	8.55	< 0.001
	Ruminococcus	4.19	5.20	0.250	3.88	5.74	0.012
Veillonellaceae	*Dialister*	4.46	3.23	0.310	4.09	3.32	0.621
	Megamonas	-3.59	0.67	< 0.001	-3.66	-4.73	0.128
	Megasphaera	-3.14	-4.14	0.336	-1.92	-3.33	0.209

Table S5: Average posterior centered log ratio transformed relative abundances estimated with the proposed MicroMiss and ZIDM models for healthy and obese participants. Only genera with relative abundances above 0.01 for obese or healthy groups are presented. P-values were obtained from a rank sum test on the centered log ratio transformations of relative abundances at the genus level.

S3 Sensitivity Analysis

In this section, we evaluate the sensitivity of the application results obtained with MicroMiss to the specification of the hyperparameters for the misclassification probability, $\boldsymbol{\eta}$. Specifically, we incorporate phylogenetic information by setting $\nu_{tc} = \nu_{ct} = 100$ when genus t and c are from the same family. This assumes that genera from the same family are more likely to be misclassified to each other.

Supplementary Figure S3 presents the log multiplicative effect of being obese versus healthy for the relative abundance of each taxon using the proposed MicroMiss model with the phylogentic-tree-based prior (MicroMissPhylo) and corresponding 95% credible intervals, and Supplementary Figure S4 provides inference on the relation between obesity status and at-risk probabilities. Similar to the results obtained with MicroMiss, MicroMissPhylo also estimated a negative multiplicative effect of being obese versus healthy for *Blautia* and *Megamonus*, in addition to *Roseburia*. In addition to identifying associations between obesity and at-risk observations in *Anaerococcus*, *Finegoldia*, *Murdochiella*, *Peptoniphilus*, and *Prevotella*, similar to MicroMiss, MicroMissPhylo also found associations with *Campylobacter* and *Porphyromonas*.

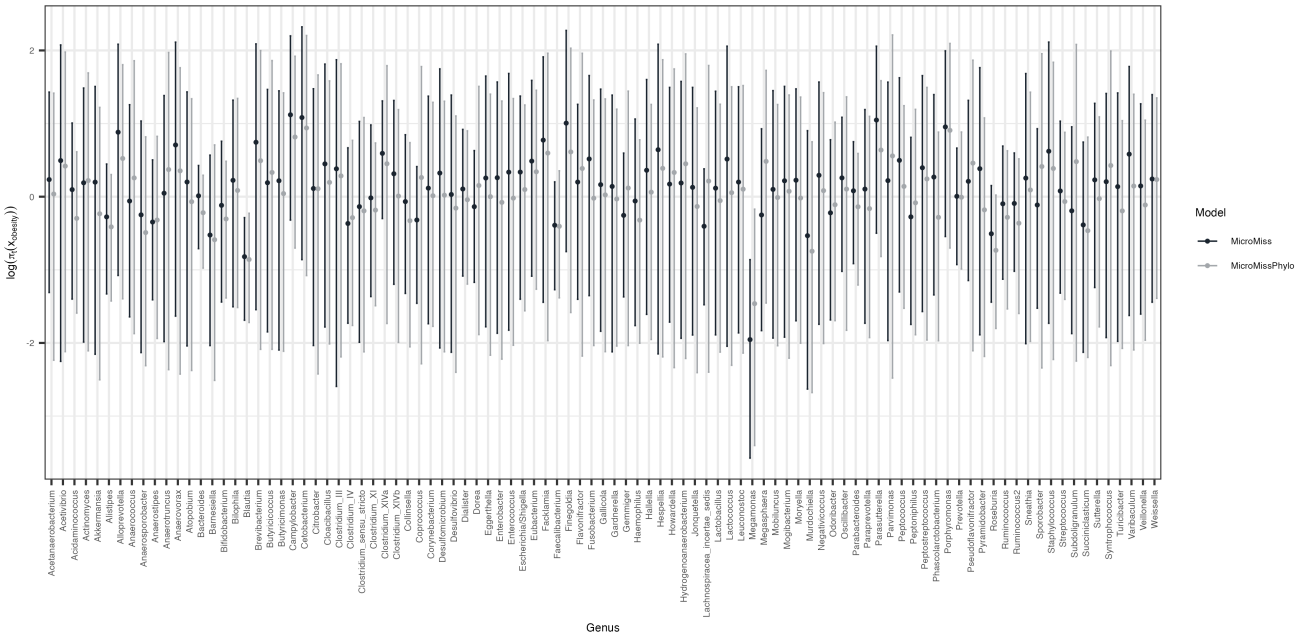

Figure S3: Posterior estimates of the log multiplicative effect of being obese versus healthy for the relative abundance of each taxon using the proposed MicroMiss and MicroMissPhylo models. Dot represents the posterior mean with error bars capturing the 95% credible intervals.

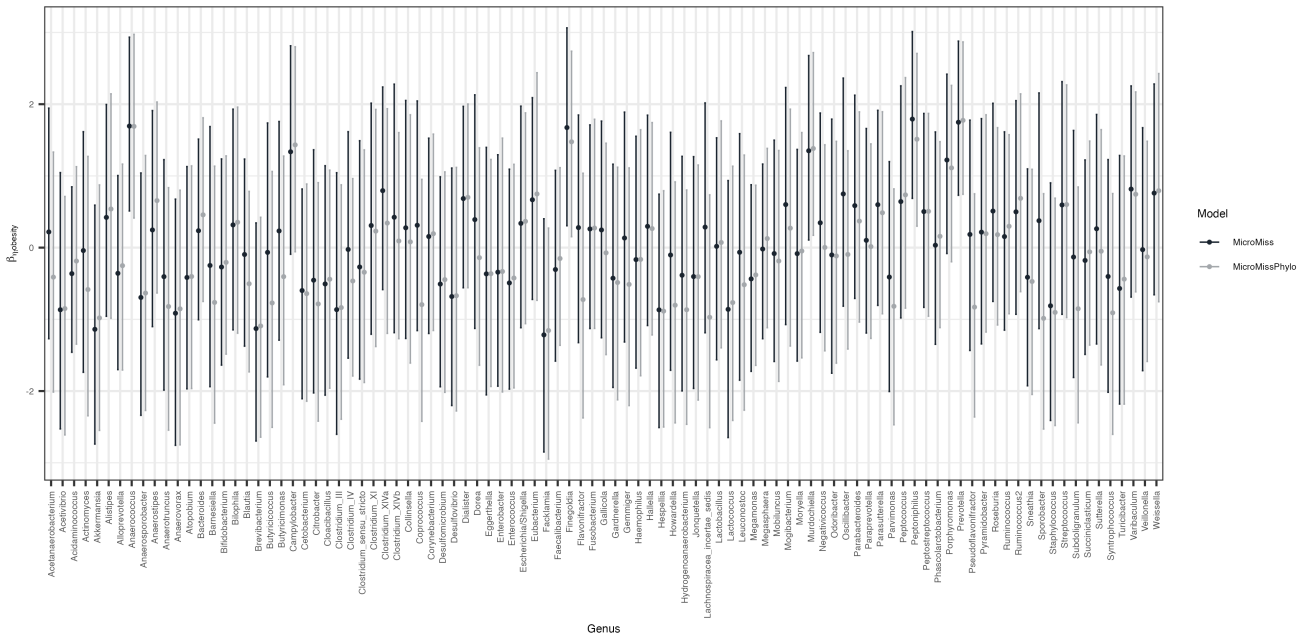

Figure S4: Posterior estimates of at-risk log odds ratios for obese versus healthy participants for each taxon using the proposed MicroMiss and MicroMissPhylo models. Dot represents the posterior mean with error bars capturing the 95% credible intervals.

Additionally, in order to assess how the results may change with higher levels of misclassification, we treated the observed classifications as the truth, and then simulated new observations with potential misclassification using a similar data generation process as the simulation study. However, instead of assuming the misclassification rates were balanced across taxa, we used phylogenetic information to generate the observations. Specifically, we set the unnormalized probability of misclassification to 1 for taxa from different families and 50 for taxa from the same family. To increase the amount of potential misclassification, we set the unnormalized probability of correct classification to 10,000, 1,000, and 100, which we refer to as *Small*, *Medium*, and *Large* amounts of misclassification, respectively. With the addition of potential misclassification in the application data, the maximum absolute difference in estimated host-level relative abundances was 0.01, 0.08, and 0.33 for the *Small*, *Medium*, and *Large* settings, respectively. The results of these analyses are plotted in Supplementary Figures S5 and S6. By introducing *potential* noise (as we may unknowingly re-classifying reads to the truth) to the observed data, we observed fairly similar results as the analysis of the raw data since a majority of the associations were insignificant. In some cases, introducing additional noise removed associations between obesity and a given taxon. For example, we observed 15 taxa from fam-

ily *Ruminococcaceae* with genus *Faecalibacterium* being the dominate taxon from this family. In the analysis of the raw data, we found a negative association between obesity status and the relative abundance for *Faecalibacterium*. However, as the amount of potential misclassification increased, the observed counts for family *Ruminococcaceae* became more evenly distributed and the association between obesity status and *Faecalibacterium* shifted towards the null. Whereas for some taxa, adding potential misclassfication biased the associations away from the null. For example, obesity status was not associated with the probability of an at-risk observation for genus *Anaerococcus* in the original analysis, but the association became more positive as the amount of misclassification increased.

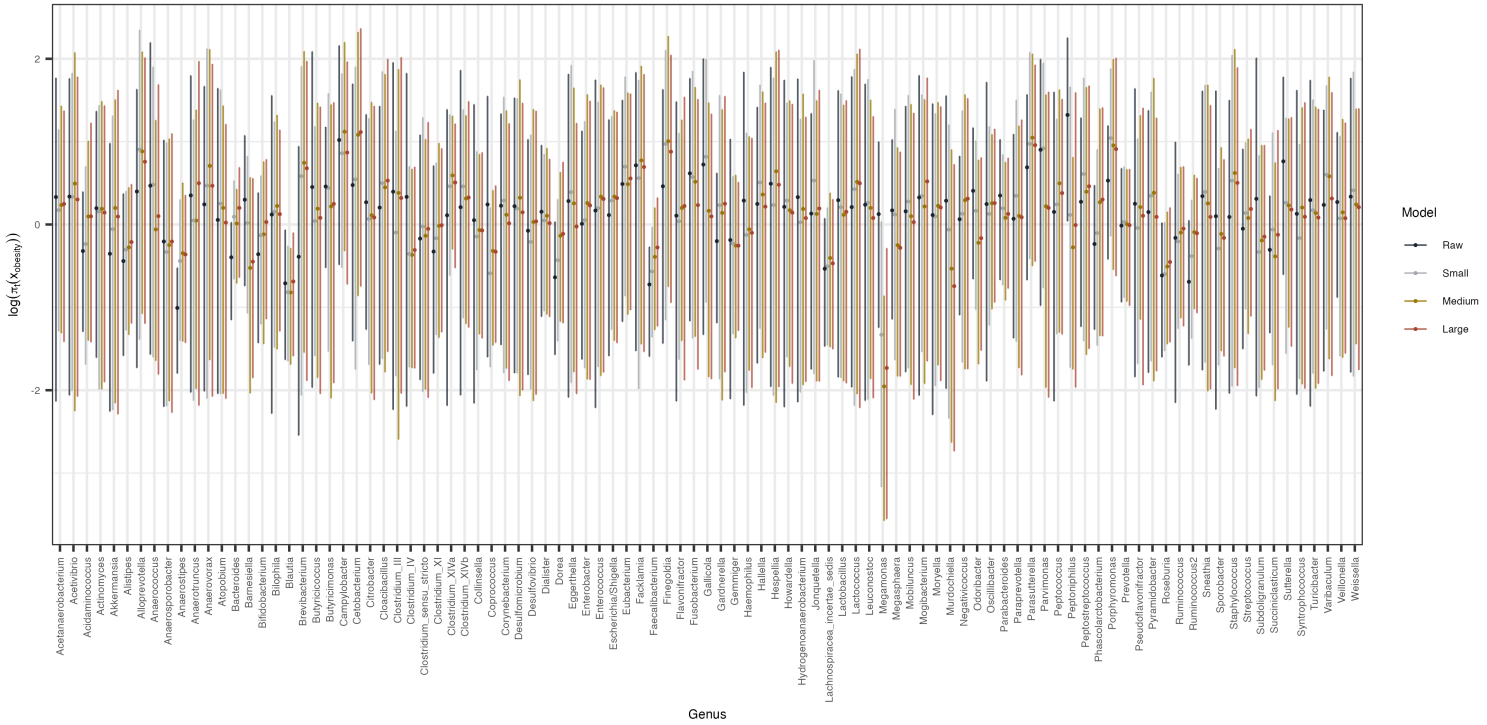

Figure S5: Posterior estimates of the log multiplicative effect of being obese versus healthy for the relative abundance of each taxon using the proposed MicroMiss model on the raw data and with small, medium, and large amounts of induced misclassification. Dot represents the posterior mean with error bars capturing the 95% credible intervals.

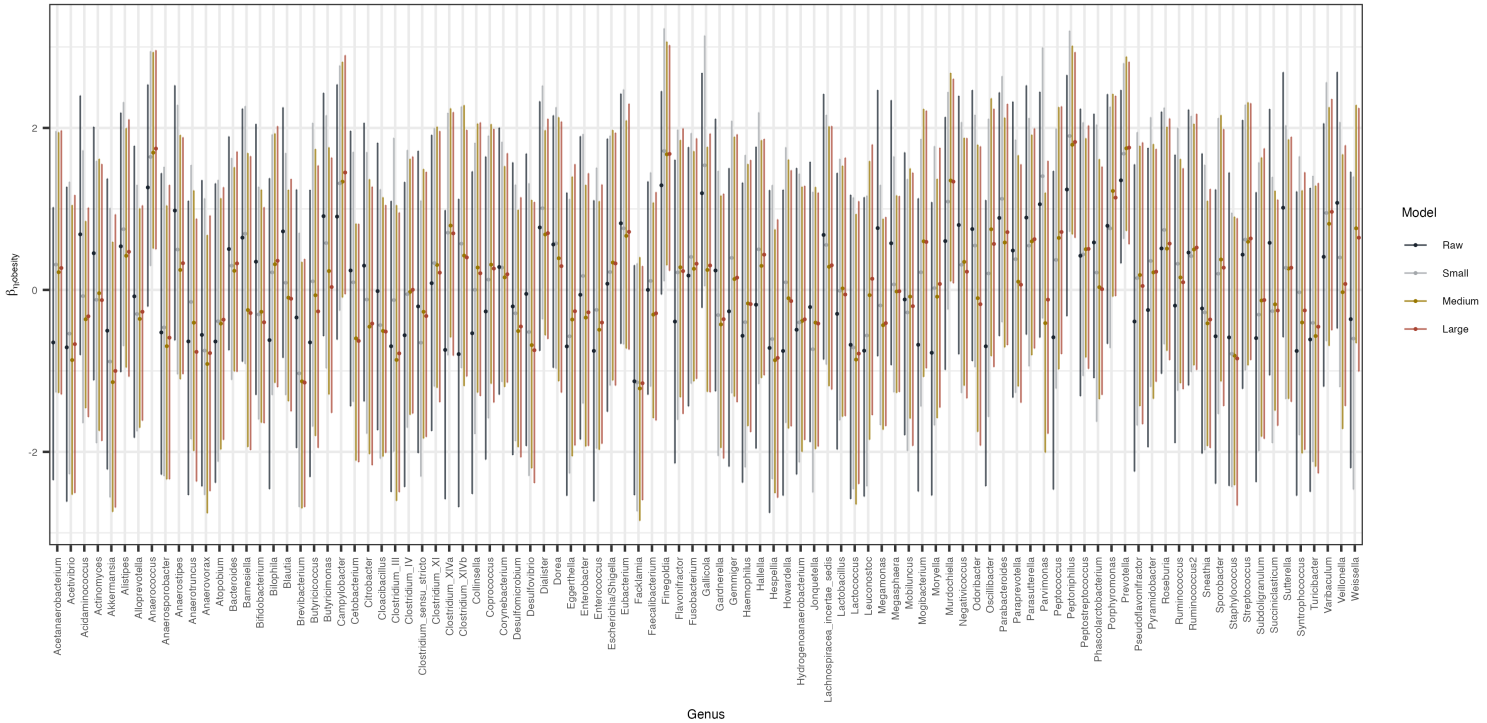

Figure S6: Posterior estimates of at-risk log odds ratios for obese versus healthy participants for each taxon using the proposed MicroMiss on the raw data and with small, medium, and large amounts of induced misclassification. Dot represents the posterior mean with error bars capturing the 95% credible intervals.