

Supplementary Information

for

Gut Microbiome Wellness Index 2 Enhances
Health Status Prediction from Gut Microbiome
Taxonomic Profiles

Chang and Gupta *et al.*

TABLE OF CONTENTS

Supplementary Figures

Supplementary Figure 1..... 3

Supplementary Figure 2..... 4

Supplementary Figure 3..... 5

Supplementary Figure 4..... 6

Supplementary Figure 5..... 7

Supplementary Tables

Supplementary Table 1..... 8

Supplementary Table 2..... 9

Supplementary Table 3..... 10

Supplementary Table 4..... 11

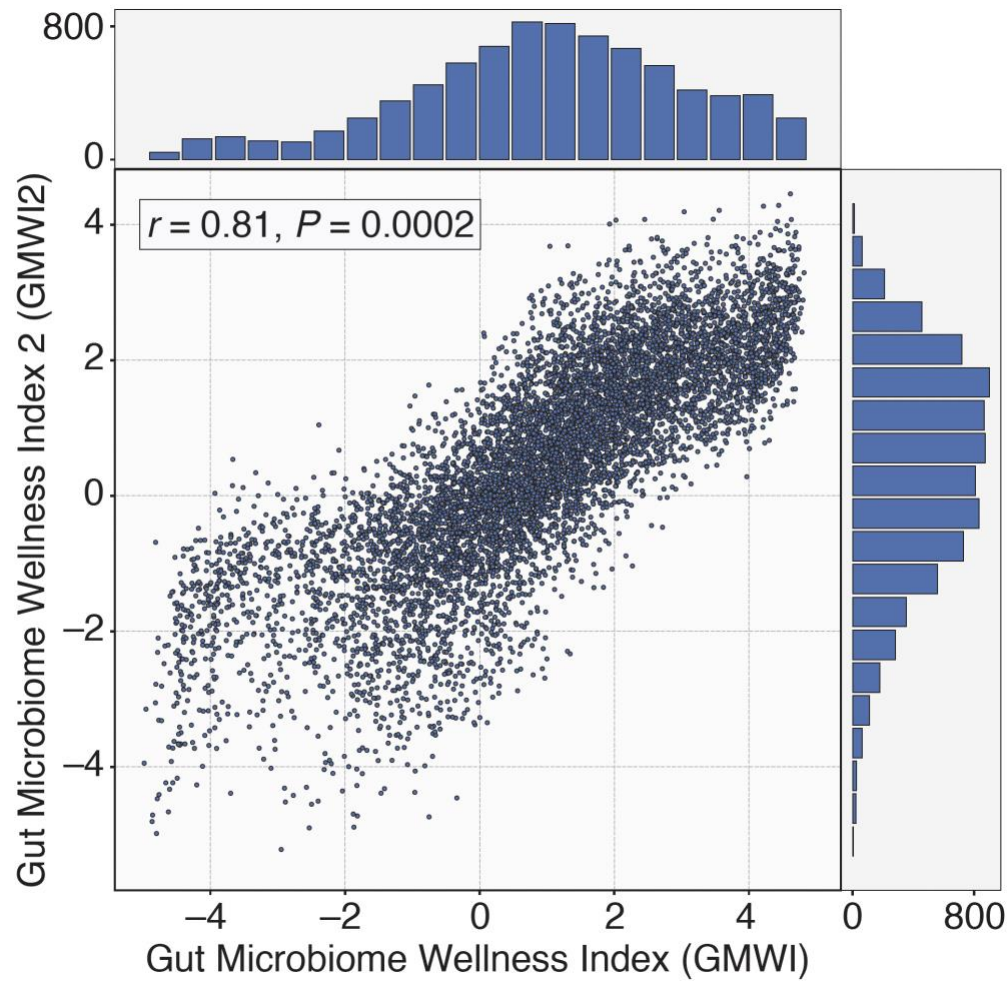
Supplementary Table 5..... 13

Supplementary Table 6..... 15

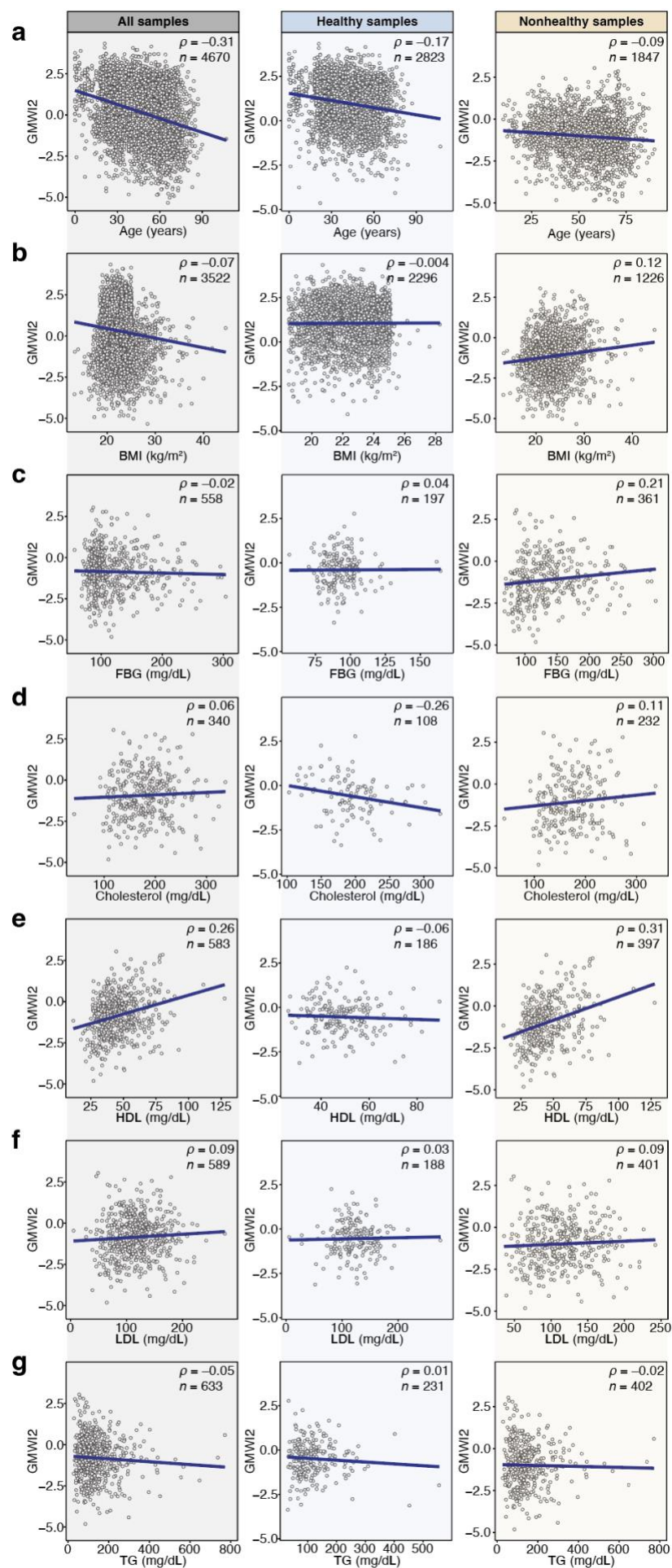
Supplementary Table 7..... 16

Supplementary Table 8..... 17

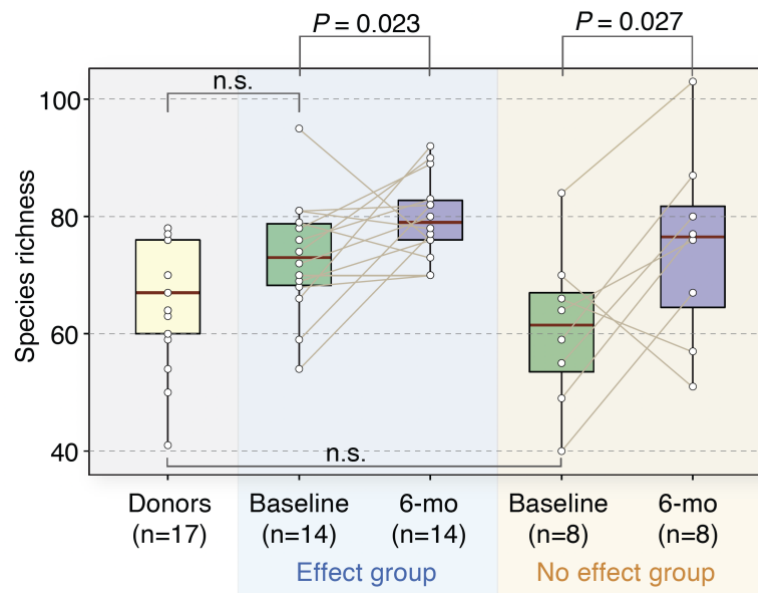
Supplementary Table 9..... 18



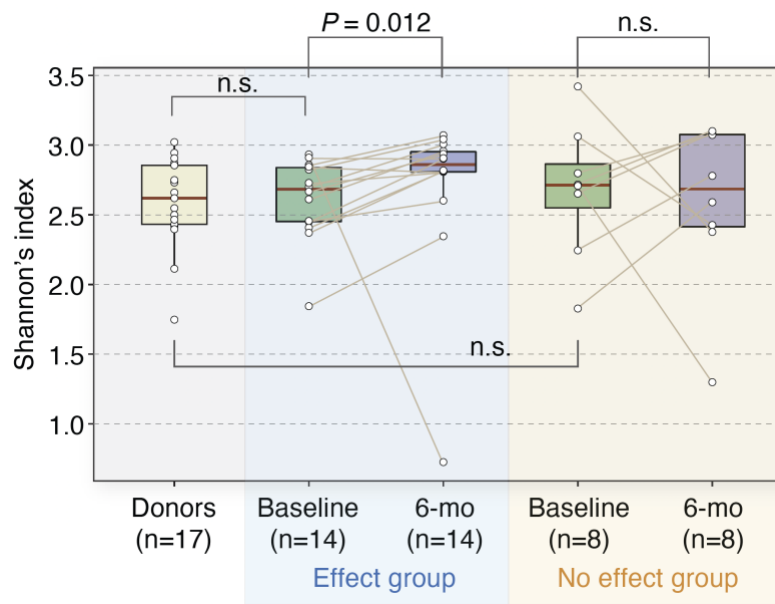
Supplementary Figure 1. Correlation between GMWI and GMWI2 scores. Scatter-plot showing the relationship between GMWI (x-axis) and GMWI2 (y-axis) calculated for each stool shotgun metagenome sample ($n = 8069$, points in plot). A strong correlation (Pearson's $r = 0.81$, $P = 0.0002$) between both versions of GMWIs was observed. Horizontal and vertical bars represent the number of scatter-plot points (i.e., samples) in each bin.



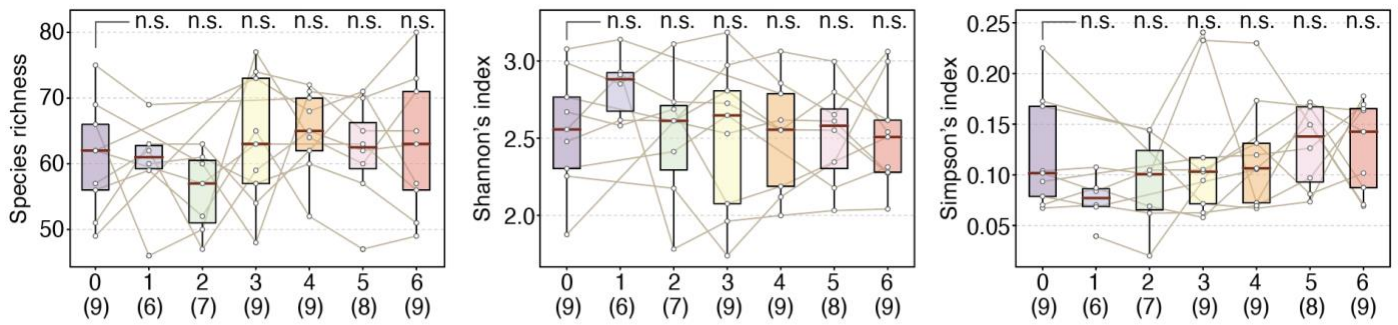
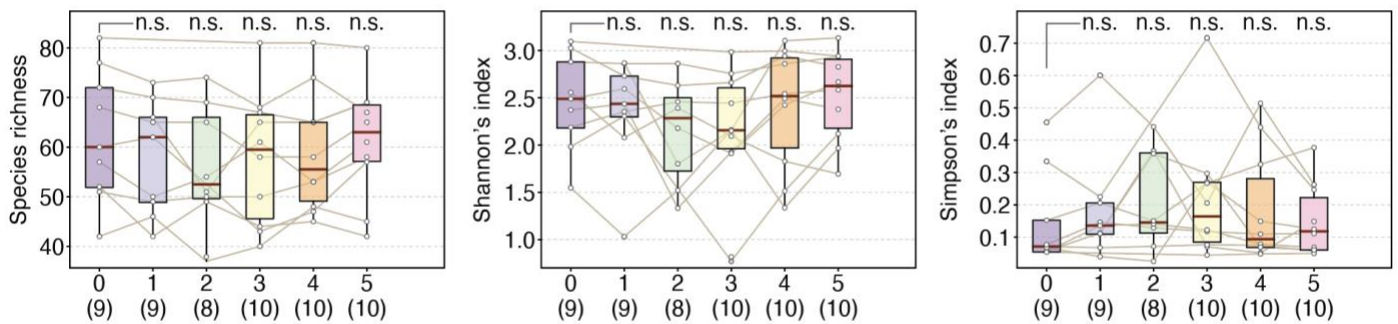
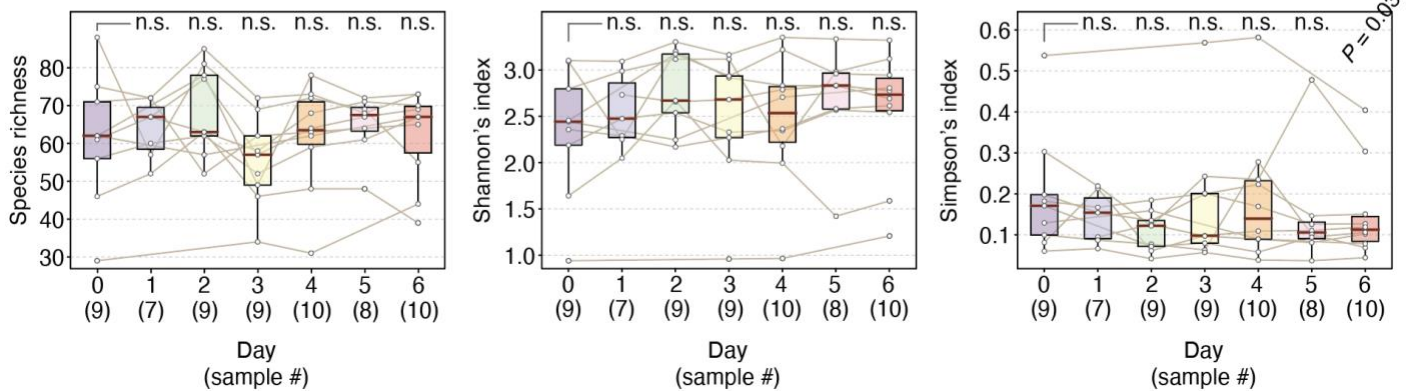
Supplementary Figure 2. Correlations between GMWI2 and clinical/demographic characteristics. Correlations are shown for **(a)** Age, **(b)** Body Mass Index (or BMI), **(c)** Fasting blood glucose (or FBG), **(d)** Cholesterol, **(e)** High-density lipoprotein (or HDL), **(f)** Low-density lipoprotein (or LDL), and **(g)** Triglycerides (or TG). All variables show weak correlations with GMWI2. ρ , Spearman's correlation coefficient; n = number of study participants (points in plot) with available clinical or demographic information.



Supplementary Figure 3. Species richness in gut microbiomes of patients with irritable bowel syndrome (IBS) observed before (Baseline) and six months after (6-mo) receiving fecal microbiota transplantation (FMT) therapy. According to the original study (Goll, R. *et al.* Effects of Fecal Microbiota Transplantation in Subjects with Irritable Bowel Syndrome are Mirrored by Changes in Gut Microbiome. *Gut Microbes* 12, 1794263 (2020)), the “Effect group” subjects were those who experienced relief in IBS symptoms, whereas the “No effect group” subjects were those who did not. Significant increase in species richness was observed in both clinical groups ($P < 0.05$). P -values were calculated using the one-sided Wilcoxon signed-rank test. n.s., non-significant ($P \geq 0.05$). n, number of FMT donor samples or FMT recipients. Standard box-and-whisker plots (e.g., center line, median; box limits, upper and lower quartiles; whiskers, $1.5 \times$ interquartile range; points, individual richness values) are used to depict groups of numerical data.



Supplementary Figure 4. Shannon's Index in gut microbiomes of patients with irritable bowel syndrome (IBS) observed before (Baseline) and six months after (6-mo) receiving fecal microbiota transplantation (FMT) therapy. According to the original study (Goll, R. *et al.* Effects of Fecal Microbiota Transplantation in Subjects with Irritable Bowel Syndrome are Mirrored by Changes in Gut Microbiome. *Gut Microbes* 12, 1794263 (2020)), the "Effect group" subjects were those who experienced relief in IBS symptoms, whereas the "No effect group" subjects were those who did not. Significant increase in Shannon's Index was observed in the "Effect group" only ($P < 0.05$). P -values were calculated using the one-sided Wilcoxon signed-rank test. n.s., non-significant ($P \geq 0.05$). n, number of FMT donor samples or FMT recipients. Standard box-and-whisker plots (e.g., center line, median; box limits, upper and lower quartiles; whiskers, $1.5 \times$ interquartile range; points, individual Shannon index values) are used to depict groups of numerical data.

a**Exclusive enteral nutrition (EEN)****b****Vegan****c****Omnivore**

Supplementary Figure 5. α -diversity indices across three diet groups. Species richness, Shannon's Index, and Simpson's Index for the **(a)** Exclusive enteral nutrition (EEN) group, the **(b)** Vegan group, and the **(c)** Omnivore group at baseline and/or across the first five to six days of dietary intervention. For all diet groups, which are described in detail in the original study (Tanes, C. *et al.* Role of Dietary Fiber in the Recovery of the Human Gut Microbiome and its Metabolome. *Cell Host Microbe* 29, 394-407.e5 (2021)), α -diversity indices did not significantly change ($P \geq 0.05$; two-sided Wilcoxon signed-rank test) compared to baseline (Day 0) except for Simpson's Index at Day 6 in the Omnivore group ($P = 0.039$). Standard box-and-whisker plots (e.g., center line, median; box limits, upper and lower quartiles; whiskers, $1.5 \times$ interquartile range; points, individual α -diversity values) are used to depict groups of numerical data.

Supplementary Table 1. GMWI2 classification performance across different taxonomic ranks.

Taxonomic rank	Balanced accuracy^b	P-value^c	Cliff's delta^d
Phylum	0.653	7.2×10^{-237}	0.451
Class	0.684	1.1×10^{-286}	0.502
Order	0.701	$< 5.0 \times 10^{-324}$	0.542
Family	0.734	$< 5.0 \times 10^{-324}$	0.599
Genus	0.768	$< 5.0 \times 10^{-324}$	0.697
Species	0.793	$< 5.0 \times 10^{-324}$	0.745
All ^a	0.799	$< 5.0 \times 10^{-324}$	0.752

^aAll taxonomic ranks were simultaneously included in training the GMWI2 model. ^bBalanced accuracy is calculated as the average of the proportions of correctly classified healthy and non-healthy samples. ^cP-value denotes the statistical significance of the differences between healthy and non-healthy GMWI2 score distributions, assessed using the two-sided Mann–Whitney *U* test. ^dCliff's delta is the measure of the effect size used.

Supplementary Table 2. Numbers (and proportions) of healthy and non-healthy samples in GMWI2 and GMWI bins. The table accounts for a total of 8,069 stool shotgun metagenome samples.

Bins	GMWI2		GMWI	
	Number of healthy samples (Proportion of total in bin)	Number of non-healthy samples (Proportion of total in bin)	Number of healthy samples (Proportion of total in bin)	Number of non-healthy samples (Proportion of total in bin)
(−6, −5]	0 (0.0%)	4 (100.0%)	0 (−)	0 (−)
(−5, −4]	3 (6.8%)	41 (93.2%)	19 (12.4%)	134 (87.6%)
(−4, −3]	11 (6.9%)	148 (93.1%)	78 (31.1%)	173 (68.9%)
(−3, −2]	57 (12.6%)	394 (87.4%)	91 (31.8%)	195 (68.2%)
(−2, −1]	252 (27.2%)	676 (72.8%)	195 (32.1%)	413 (67.9%)
(−1, 0]	833 (52.0%)	770 (48.0%)	484 (45.7%)	575 (54.3%)
(0, 1]	1380 (78.7%)	374 (21.3%)	975 (63.1%)	571 (36.9%)
(1, 2]	1713 (94.7%)	96 (5.3%)	1268 (80.4%)	309 (19.6%)
(2, 3]	1094 (98.4%)	18 (1.6%)	1123 (90.8%)	114 (9.2%)
(3, 4]	199 (99.5%)	1 (0.5%)	756 (96.4%)	28 (3.6%)
(4, 5]	5 (100.0%)	0 (0.0%)	558 (98.2%)	10 (1.8%)

Supplementary Table 3. Classification performance on training dataset for varying GMWI2 score magnitude cutoffs.

GMWI2 cutoff (Magnitude)	Number of samples retained	Balanced accuracy [†] (%)
0.0	8069	79.89
0.1	7719	81.19
0.2	7333	82.49
0.3	7004	83.53
0.4	6691	84.65
0.5	6364	85.75
0.6	6029	87.05
0.7	5664	88.20
0.8	5365	89.18
0.9	5049	90.17
1.0	4712	90.98
1.1	4412	91.63
1.2	4087	92.59
1.3	3782	93.35
1.4	3521	93.89
1.5	3238	94.56
1.6	2943	94.72
1.7	2697	95.00
1.8	2451	95.51
1.9	2210	95.53
2.0	1975	95.84
2.1	1756	96.07
2.2	1545	95.97
2.3	1347	96.10
2.4	1168	96.21
2.5	998	96.45
2.6	843	96.73
2.7	719	96.50
2.8	589	96.84
2.9	507	96.81
3.0	412	96.53

[†]Balanced accuracy is defined as the average of the proportions of healthy and non-healthy samples that were correctly classified.

Supplementary Table 4. GMWI2 classification performance on held-out studies in inter-study validation.

Author's last name (Publication year)	Number of healthy samples	Number of non-healthy samples	Balanced accuracy [†] (%)
Smits (2017)	38	0	100
Jacobson (2021)	82	0	97.56
Lokmer (2019)	37	0	97.30
Pasolli (2019)	142	0	96.48
Ang (2021)	22	0	95.45
D'Souza (2021)	128	0	95.31
Schirmer (2016)	385	0	94.03
Zeevi (2015)	900	0	93.44
Le Chatelier (2013)	88	0	93.18
Yachida (2019)	0	217	91.71
Rettedal (2021)	35	0	91.43
Tett (2019)	110	0	90.91
Asnicar (2021)	568	0	90.67
De Filippis (2019)	99	0	89.90
Liu (2016)	110	0	86.36
Costea (2017)	169	0	84.02
Xie (2016)	130	0	82.31
Roager (2019)	50	0	82.00
Backhed (2015)	100	0	81.00
Yassour (2018)	42	0	80.95
Dhakan (2019)	47	0	78.72
Zhu (2021)	32	100	76.81
Franzosa (2018)	56	157	74.57
Gu (2017)	0	94	74.47
Ananthakrishnan (2017)	0	64	71.88
Wirbel (2019)	33	22	68.94
Kim (2021)	61	0	68.85
Huttenhower (2012) and Lloyd-Price (2017) ^a	507	0	68.05
Lloyd-Price (2019)	25	61	67.87
Feng (2015)	0	46	67.39
Qin (2014)	135	152	66.59
Zeller (2014)	45	90	66.11
Vogtmann (2016)	30	51	66.08
Schirmer (2018)	18	65	65.81
Mehta (2018)	301	0	65.12
Obregon-Tito (2015)	20	0	65.00
Yang (2020)	88	92	64.90
Nielsen (2014)	82	77	64.68
Ventura (2019)	24	24	62.50
Loomba (2017)	0	86	60.47
Sun (2021)	42	0	59.52

Weng (2019)	15	64	58.44
Yu (2015)	53	75	57.71
Qin (2012)	183	186	56.68
He (2017)	40	46	56.25
Zhang (2015)	61	102	55.79
Wen (2017)	105	95	55.56
Jie (2017)	108	214	54.61
Thomas (2019)	61	99	54.27
Yang (2021)	97	97	54.12
Qi (2019)	43	0	51.16
Davies (2020)	0	44	50.00
Gupta (2020)	0	49	48.98
Karlsson (2013)	0	53	22.64

^aSamples combined from both phases of the Human Microbiome Project (HMP1 and HMP1-II). [†]Balanced accuracy is defined as the average of the proportions of healthy and non-healthy samples that were correctly classified.

Supplementary Table 5. GMWI2 scores, species richness, and Shannon Index of stool metagenomes from Goll *et al.*

Subject ID	Time-point ^a	GMWI2	Species richness	Shannon Index	Study group [†]
3	Baseline	1.97113	70	2.72843	Effect
3	6-mo	1.91417	70	2.80772	Effect
11	Baseline	−0.08548	81	2.45591	Effect
11	6-mo	0.59796	78	2.60174	Effect
16	Baseline	0.01185	59	2.37104	Effect
16	6-mo	−0.10259	82	2.81796	Effect
19	Baseline	1.72307	54	2.40814	Effect
19	6-mo	0.17037	77	2.81474	Effect
24	Baseline	0.15272	76	2.84407	Effect
24	6-mo	2.39152	83	2.80925	Effect
25	Baseline	−0.01581	66	2.91005	Effect
25	6-mo	2.20480	92	2.90612	Effect
30	Baseline	1.87864	81	2.82448	Effect
30	6-mo	2.55779	82	3.04217	Effect
31	Baseline	2.15370	74	2.85770	Effect
31	6-mo	1.85034	90	3.07087	Effect
59	Baseline	−0.13333	79	2.66701	Effect
59	6-mo	1.35974	73	2.93682	Effect
77	Baseline	−1.45895	95	2.93271	Effect
77	6-mo	0.10312	76	0.72586	Effect
78	Baseline	1.48054	72	1.84437	Effect
78	6-mo	1.03121	80	2.34608	Effect
88	Baseline	2.24283	69	2.70017	Effect
88	6-mo	2.35412	76	3.00670	Effect
89	Baseline	−0.53592	78	2.45049	Effect
89	6-mo	0.58335	89	2.90201	Effect
90	Baseline	1.21949	68	2.61136	Effect
90	6-mo	2.36111	70	2.95703	Effect
12	Baseline	1.47657	84	3.42045	No effect
12	6-mo	0.14123	103	2.37705	No effect
18	Baseline	0.22944	70	2.70798	No effect
18	6-mo	1.00296	51	1.29897	No effect
22	Baseline	1.31860	49	1.82744	No effect
22	6-mo	1.98558	77	2.58996	No effect
32	Baseline	1.31242	40	2.24464	No effect
32	6-mo	1.80714	67	2.77963	No effect
56	Baseline	1.36916	64	2.71700	No effect
56	6-mo	0.27024	76	3.10007	No effect
64	Baseline	1.82462	55	2.79772	No effect
64	6-mo	1.00534	80	3.07260	No effect
66	Baseline	0.83050	59	2.65193	No effect

66	6-mo	1.50540	87	3.08256	No effect
75	Baseline	2.06776	66	3.06168	No effect
75	6-mo	1.31720	57	2.42652	No effect
D-1_S9	-	1.12769	76	2.85490	Healthy donor
D-10_S4	-	0.69000	78	2.90576	Healthy donor
D-12_S5	-	0.92771	70	2.61911	Healthy donor
D-13U_S1	-	0.79794	41	2.65966	Healthy donor
D-14_S8	-	0.18140	76	2.46545	Healthy donor
D-14U_S10	-	0.14199	60	2.11311	Healthy donor
D-15Fryst_S3	-	0.44763	59	2.87136	Healthy donor
D-15U_S6	-	1.58681	67	2.49936	Healthy donor
D-2_S7	-	1.08010	78	2.54714	Healthy donor
D-3_S6	-	1.22138	77	2.94429	Healthy donor
D-4_S10	-	1.14330	76	2.39656	Healthy donor
D-5Fryst_S8	-	0.90896	64	2.43220	Healthy donor
D-6Fresk_S5	-	0.86373	64	3.02126	Healthy donor
D-6Fryst_S4	-	1.10229	54	2.74909	Healthy donor
D-7Fryst_S7	-	1.35535	70	2.42049	Healthy donor
D-9Feresk_S9	-	0.24005	50	1.74761	Healthy donor
D-9Fryst_S2	-	0.65847	63	2.73193	Healthy donor

^aBefore (Baseline) and six months after (6-mo) receiving fecal microbiota transplantation (FMT) from Healthy donors (-).

[†]According to the original study, “Effect group” subjects were those who experienced relief in IBS symptoms, whereas the “No effect group” subjects were those who did not.

Supplementary Table 6. GMWI2 scores of stool metagenomes from Tanes *et al.*

Diet group	Subject ID	Time-point [†]						
		Day 0	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6
EEN ^a	9003	0.934	0.483	0.501	−0.58	−0.59	-	−0.962
	9009	0.905	-	-	−0.437	−0.722	−0.689	−0.863
	9013	1.388	-	0.256	0.253	−0.041	−1.66	−1.077
	9016	1.141	-	0.903	0.06	-	−0.466	−0.679
	9024	−0.472	−0.28	−0.577	−1.043	−1.179	−1.24	−1.701
	9025	1.461	1.008	-	-	0.682	-	-
	9029	0.264	0.075	0.009	−0.101	−0.308	−0.389	−0.339
	9032	0.525	-	-	0.228	−0.146	−0.142	−0.055
	9034	0.14	−0.014	−0.05	−0.862	−0.109	−0.125	−1.077
	9038	-	−0.308	−0.728	−0.917	−1.118	−0.961	−1
Omnivore	9005	−0.624	-	-	−0.897	−0.624	-	−0.607
	9006	2.645	0.818	2.153	1.436	2.022	1.651	1.14
	9010	0.849	0.647	0.57	−0.175	0.545	0.269	0.599
	9015	0.668	0.118	0.367	0.816	1.029	−0.2	−0.691
	9019	1.472	1.915	1.693	1.402	1.73	1.686	1.533
	9030	1.276	-	0.513	0.539	1.275	0.677	0.048
	9033	−0.636	−0.929	−0.899	−0.362	−0.864	−0.589	−0.589
	9036	0.452	-	0.444	-	0.526	-	0.029
	9037	2.205	1.86	1.68	1.552	1.044	2.212	2.082
	9040	-	1.225	1.4	1.331	1.295	1.708	1.662
Vegan	9002	0.57	0.396	0.423	0.921	0.658	0.337	-
	9004	0.914	1.492	1.345	1.393	1.619	0.96	-
	9008	1.602	1.451	1.77	1.549	1.77	1.605	-
	9014	1.468	1.569	−0.509	1.405	0.884	1.849	-
	9017	1.008	0.161	1.054	0.726	0.498	0.89	-
	9022	-	1.742	1.537	1.887	2.018	1.943	-
	9026	2.342	2.128	-	1.462	1.634	1.76	-
	9027	2.92	-	-	2.486	2.433	2.627	-
	9031	1.458	1.756	1.663	1.609	1.574	1.886	-
	9035	2.892	3.045	2.796	2.621	2.371	2.277	-

^aExclusive Enteral Nutrition (EEN), which is a synthetic diet devoid of fiber. [†]Baseline and/or across the first five to six days of dietary intervention. ‘-’ denotes a time-point where a stool metagenome sample was unavailable.

Supplementary Table 7. GMWI2 scores, species richness, and Shannon Index of stool metagenomes from Palleja *et al.*

Index	Subject ID	Time-point [†]				
		Day 0	Day 4	Day 8	Day 42	Day 180
GMWI2	ERAS1	2.589	-	-0.506	-0.022	0.058
	ERAS2	1.083	0.145	-0.994	-0.123	-0.577
	ERAS3	0.889	0.526	-1.141	-0.45	-0.663
	ERAS4	1.442	2.103	-1.152	-1.126	-0.167
	ERAS5	0.98	-0.67	-0.399	0.525	0.53
	ERAS6	2.929	0.758	0.765	0.678	1.598
	ERAS7	1.936	0.161	-0.87	-0.662	0.301
	ERAS8	0.985	-	0.184	1.164	0.567
	ERAS9	2.101	1.734	-0.346	0.38	1.048
	ERAS10	2.518	-	-1.006	-0.258	-1.046
	ERAS11	2.455	1.269	-0.33	0.027	1.197
	ERAS12	1.589	0.344	-0.075	0.218	0.288
Species richness	ERAS1	74	-	59	76	81
	ERAS2	89	18	53	72	78
	ERAS3	103	96	28	48	82
	ERAS4	111	98	75	81	110
	ERAS5	95	60	47	77	72
	ERAS6	82	23	39	87	97
	ERAS7	93	80	70	80	83
	ERAS8	84	-	61	90	95
	ERAS9	92	71	50	86	95
	ERAS10	93	-	49	59	70
	ERAS11	106	46	45	95	99
	ERAS12	78	35	57	79	61
Shannon Index	ERAS1	2.886	-	1.644	2.785	2.534
	ERAS2	3.035	0.77	1.622	2.619	2.562
	ERAS3	3.035	1.636	1.428	2.223	2.545
	ERAS4	2.996	2.897	2.338	2.895	2.995
	ERAS5	2.083	0.915	1.584	2.543	2.77
	ERAS6	2.013	0.152	1.394	3.066	3.264
	ERAS7	2.909	1.553	2.389	2.372	1.504
	ERAS8	2.075	-	2.091	1.337	2.079
	ERAS9	2.934	2.812	1.745	2.952	2.779
	ERAS10	2.668	-	1.834	1.47	1.813
	ERAS11	3.101	2.012	2.239	2.822	3.139
	ERAS12	1.806	0.651	2.224	3.078	2.826

[†]Before (0) and at 4, 8, 42, and 180 days post-intervention with a cocktail of three last-resort antibiotics (meropenem, gentamicin, and vancomycin). ‘-’ denotes a time-point where a stool metagenome sample was unavailable.

Supplementary Table 8. GMWI2 scores of pooled[†] fecal samples from Lee *et al.*

Study groups ^a	Replicate 1	Replicate 2	Replicate 3
NS0	−0.117	−0.22	0.038
NS24	−0.177	−0.177	−0.340
FS24	0.065	0.350	0.350
IN24	0.725	0.556	0.713
GS24	0.611	0.608	0.478
XS24	0.192	0.303	0.100
FL24	−0.607	−0.400	−0.568

[†]19 healthy adult volunteers (14 males and 5 females) provided fecal samples, which were then pooled. ^aControl groups: NS0, no substrate addition at 0 h; NS24, no substrate for 24 h. Prebiotic groups in a 24 h *in vitro* anaerobic batch fecal fermentation system: FS24, fructooligosaccharide; IN24, inulin; GS24, galactooligosaccharide; XS24, xylooligosaccharide; FL24, 2'-fucosyllactose.

Supplementary Table 9. Nested cross-validation^δ inner loop results on each outer loop training fold, using inter-study validation (ISV) for both the inner and outer loop.

Held-out study [†]	ISV performance [¶] on the outer loop training fold for a given value of the inverse regularization strength (C)							
	0.001	0.003	0.01	0.03	0.1	0.3	1	3
Obregon-Tito (2015)	0.3779	0.6888	0.7182	0.7349	0.7210	0.7070	0.6997	0.6862
Feng (2015)	0.3590	0.6947	0.7173	0.7344	0.7143	0.7045	0.6983	0.6854
Schirmer (2016)	0.3779	0.6863	0.7092	0.7269	0.7102	0.6996	0.6945	0.6819
Huttenhower (2012)	0.3779	0.6868	0.7155	0.7384	0.7331	0.7164	0.7070	0.6945
Zeevi (2015)	0.3779	0.6650	0.7167	0.7296	0.7143	0.7044	0.7025	0.6795
Zhang (2015)	0.3661	0.6931	0.7250	0.7324	0.7219	0.7117	0.7028	0.6925
Jie (2017)	0.3653	0.6762	0.7175	0.7303	0.7142	0.7036	0.6971	0.6855
Vogtmann (2016)	0.3660	0.6882	0.7153	0.7353	0.7234	0.7058	0.7044	0.6923
Backhed (2015)	0.3779	0.6882	0.7139	0.7283	0.7155	0.7059	0.7004	0.6830
Le Chatelier (2013)	0.3779	0.6850	0.7131	0.7290	0.7125	0.7024	0.6966	0.6844
Yu (2015)	0.3668	0.6899	0.7134	0.7320	0.7186	0.7041	0.6982	0.6836
Schirmer (2018)	0.3631	0.6906	0.7194	0.7367	0.7193	0.7113	0.7042	0.6890
Zeller (2014)	0.3653	0.6936	0.7143	0.7307	0.7191	0.7076	0.7001	0.6871
He (2017)	0.3678	0.6903	0.7158	0.7331	0.7188	0.7086	0.6998	0.6911
Liu (2016)	0.3779	0.6843	0.7110	0.7283	0.7117	0.7037	0.6979	0.6821
Qin (2014)	0.3679	0.6815	0.7191	0.7315	0.7142	0.7036	0.6940	0.6833
Qin (2012)	0.3684	0.6857	0.7151	0.7336	0.7137	0.7072	0.6997	0.6925
Karlsson (2013)	0.3590	0.6977	0.7242	0.7495	0.7320	0.7185	0.7182	0.7006
Nielsen (2014)	0.3687	0.6911	0.7139	0.7344	0.7210	0.7116	0.7038	0.6952
Dhakan (2019)	0.3779	0.6881	0.7126	0.7317	0.7164	0.7066	0.6978	0.6856
Loomba (2017)	0.3590	0.6936	0.7196	0.7397	0.7253	0.7144	0.7079	0.6944
Thomas (2019)	0.3662	0.6935	0.7160	0.7333	0.7200	0.7096	0.6990	0.6892
Wirbel (2019)	0.3703	0.6901	0.7191	0.7317	0.7208	0.7069	0.6981	0.6901
Wen (2017)	0.3689	0.6900	0.7291	0.7377	0.7220	0.7096	0.7060	0.6924
Xie (2016)	0.3779	0.6839	0.7139	0.7292	0.7181	0.7041	0.7021	0.6881
Davies (2020)	0.3590	0.6914	0.7222	0.7372	0.7211	0.7082	0.7051	0.6950
Qi (2019)	0.3779	0.6893	0.7191	0.7352	0.7250	0.7075	0.7015	0.6880
Weng (2019)	0.3626	0.6853	0.7134	0.7309	0.7160	0.7037	0.6980	0.6837
Lloyd-Price (2019)	0.3645	0.6894	0.7156	0.7363	0.7190	0.7077	0.7009	0.6897
Tett (2019)	0.3779	0.6870	0.7119	0.7299	0.7147	0.7020	0.6972	0.6843
Costea (2017)	0.3779	0.6875	0.7148	0.7326	0.7156	0.7046	0.6982	0.6875
Gu (2017)	0.3590	0.6913	0.7189	0.7318	0.7256	0.7163	0.7100	0.6929
Roager (2019)	0.3779	0.6908	0.7191	0.7301	0.7186	0.7051	0.6974	0.6874

Gupta (2020)	0.3590	0.6877	0.7191	0.7381	0.7218	0.7165	0.7090	0.6943
Zhu (2021)	0.3636	0.6836	0.7153	0.7296	0.7170	0.7052	0.7044	0.6903
Yachida (2019)	0.3590	0.6852	0.7124	0.7251	0.7109	0.7077	0.7042	0.6902
De Filippis (2019)	0.3779	0.6823	0.7115	0.7297	0.7150	0.7006	0.6967	0.6855
Franzosa (2018)	0.3640	0.6870	0.7201	0.7349	0.7173	0.7071	0.7001	0.6923
Ananthakrishnan (2017)	0.3590	0.6876	0.7141	0.7356	0.7207	0.7063	0.7025	0.6904
Mehta (2018)	0.3779	0.6904	0.7224	0.7362	0.7214	0.7108	0.7099	0.6946
Asnicar (2021)	0.3779	0.6728	0.7142	0.7252	0.7149	0.7022	0.6996	0.6870
Lokmer (2019)	0.3779	0.6829	0.7120	0.7300	0.7129	0.7005	0.6941	0.6833
Pasolli (2019)	0.3779	0.6857	0.7120	0.7280	0.7123	0.7004	0.6983	0.6848
Yassour (2018)	0.3779	0.6931	0.7150	0.7304	0.7174	0.7030	0.6971	0.6841
Sun (2021)	0.3779	0.6919	0.7185	0.7355	0.7226	0.7099	0.7002	0.6894
Jacobson (2021)	0.3779	0.6841	0.7123	0.7298	0.7128	0.7022	0.6967	0.6806
D’Souza (2021)	0.3779	0.6854	0.7130	0.7296	0.7128	0.6998	0.6964	0.6823
Smits (2017)	0.3779	0.6825	0.7127	0.7297	0.7127	0.7017	0.6931	0.6804
Rettedal (2021)	0.3779	0.6866	0.7155	0.7312	0.7159	0.7020	0.6986	0.6855
Ang (2021)	0.3779	0.6903	0.7156	0.7301	0.7191	0.7088	0.7035	0.6893
Kim (2021)	0.3779	0.6884	0.7138	0.7288	0.7171	0.7056	0.7025	0.6930
Ventura (2019)	0.3684	0.6886	0.7213	0.7366	0.7217	0.7076	0.7026	0.6940
Yang (2020)	0.3682	0.6871	0.7148	0.7318	0.7205	0.7028	0.6949	0.6818
Yang (2021)	0.3684	0.6875	0.7173	0.7383	0.7257	0.7064	0.6945	0.6806

[§]The nested cross-validation protocol was conducted as follows:

1. Outer loop: Using ISV, establish the “leave-one-study-out” cross-validation structure to create multiple train-test splits.
2. Inner loop: For each outer loop train-test split, perform hyperparameter optimization within the training fold (or training set) by also using ISV.
3. For each outer loop train-test split, train the model using the optimized hyperparameters obtained from the inner loop on the training fold, and assess performance on the outer loop’s corresponding test fold (or test set).

[†]Each row represents an iteration of the outer loop, where the outer loop training fold is specified by the held-out study.

[¶]Average inner loop ISV balanced accuracies (on each outer loop training fold) are displayed for each value of C tested. The best-performing C value in each outer loop iteration is highlighted.