Supplementary Materials for GMPR: A novel normalization method for microbiome sequencing data

Supplementary Note

The details of how to estimate the size factors using each normalization method are described as follows.

- GMPR (Geometric Mean of Pairwise Ratios): The size factors for all samples are calculated by `GMPR` described in the main text.
- CSS (Cumulative Sum Scaling): The size factors for all samples are calculated by applying `newMRexperiment`, `cumNorm` and `normFactors` in Bioconductor package metagenomeSeq. Normalized read counts are obtained by dividing the raw read counts by the size factors.
- RLE (Relative Log Expression): The size factors for all samples are calculated by the `calcNormFactors` with the parameter set as "RLE" in the edgeR Bioconductor package. The scaled size factors are obtained by multiplying the size factors with the total read count. Normalized read counts are obtained by dividing the raw read counts by the scaled size factors.
- RLE+ (Relative Log Expression plus pseudo-counts): The scaled size factors for all samples are calculated in the same way as RLE, except that each data entry is added with a pseudo-count 1. Normalized read counts are obtained by dividing the raw read counts by the scaled size factors.
- TMM (Trimmed Mean of M values): The size factors for all samples are calculated by the `calcNormFactors` function with the parameter set as "TMM" in the edgeR Bioconductor package. The scaled size factors are obtained by multiplying the size factors with the total read count. Normalized read counts are obtained by dividing the raw read counts by the scaled size factors.
- TMM+ (Trimmed Mean of M values plus pseudo-counts): The scaled size factors for all sample are calculated in the same way as TMM, except that each data entry is added with a pseudo-count 1. Normalized read counts are obtained by dividing the raw read counts by the scaled size factors.
- TSS (Total Sum Scaling): The size factors are taken to be the total read counts. Normalized read counts are obtained by dividing the raw read counts by the size factors.

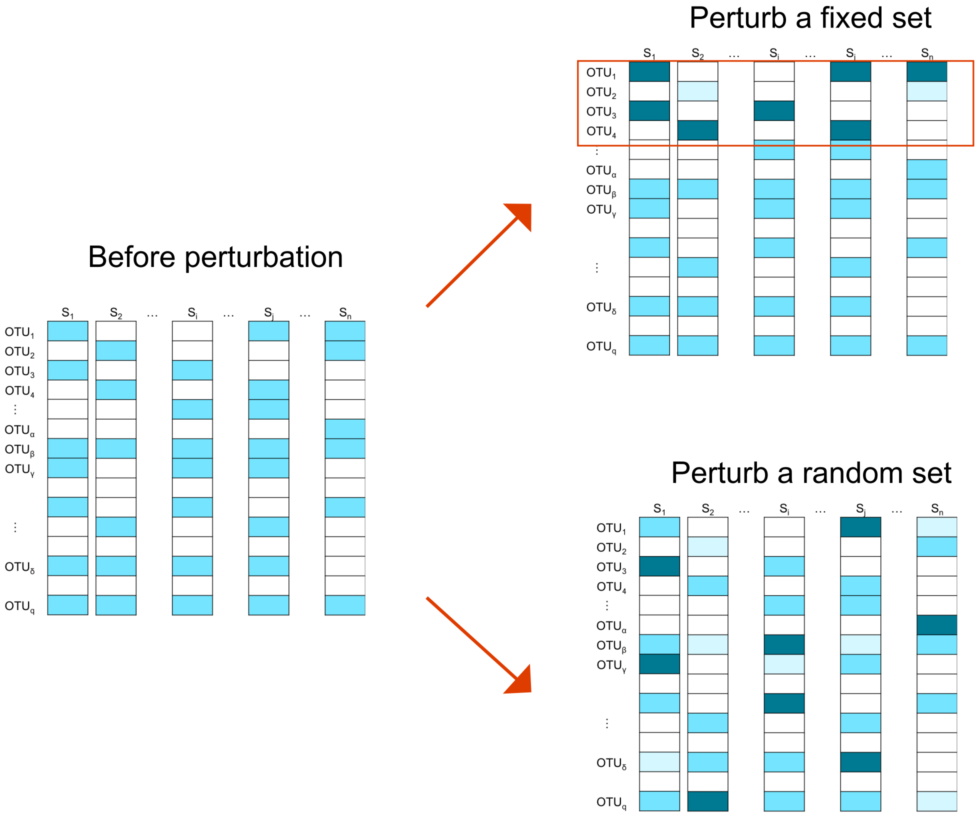

Figure S1: Illustration of the simulation strategy. In the 'fixed' perturbation approach, the same set of OTUs are decreased/increased in the same direction for all samples, reflecting differentially abundant OTUs under certain conditions such as disease state. In the 'random' perturbation approach, each sample has a random set of OTUs perturbed with a random direction, reflecting the sample-specific outliers. The darkness of the color indicates the OTU abundance.

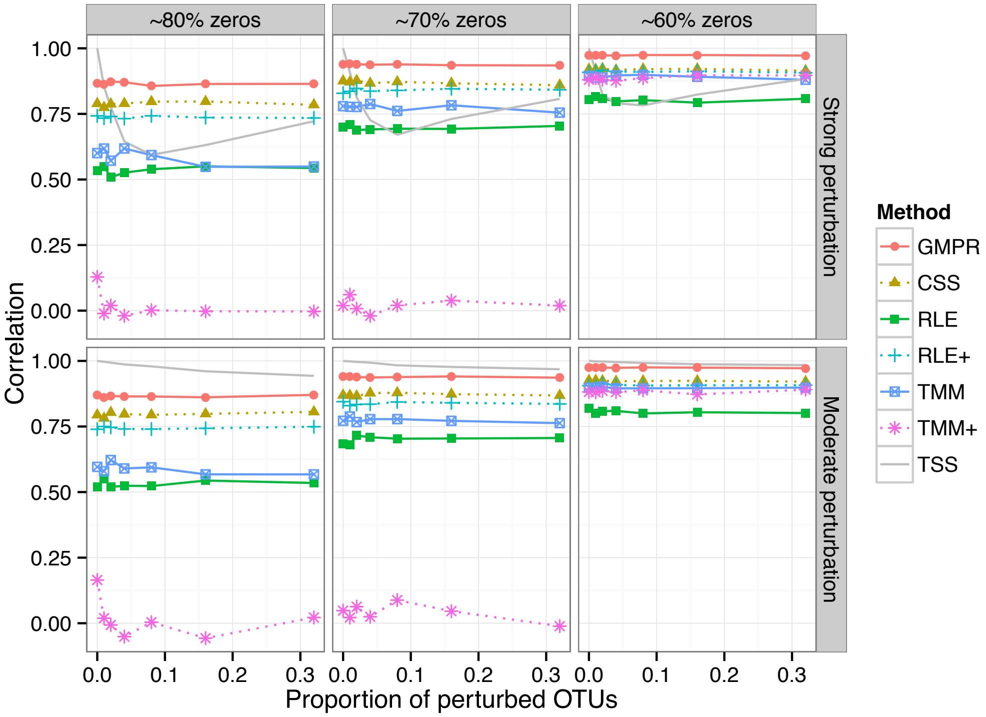

Figure S2: Spearman's correlation between the estimated size factors and the simulated 'true' library sizes when a fixed set of OTUs are perturbed. The performance of different normalization methods are compared under different level of zero inflation, percentage of perturbed OTUs and strength of perturbation.

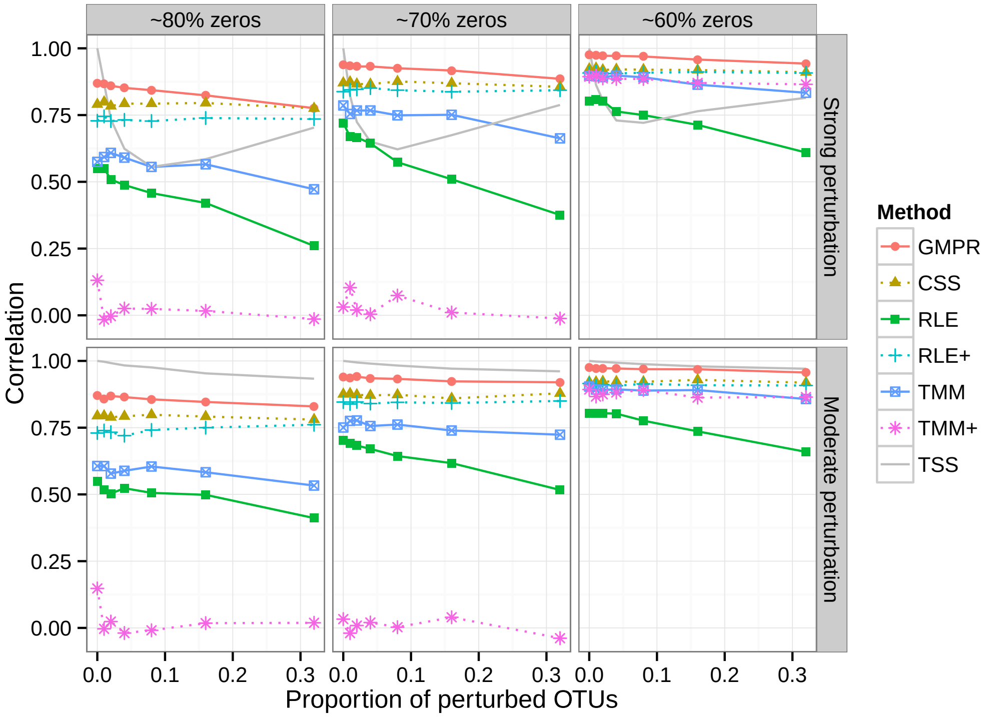

Figure S3: Spearman's correlation between the estimated size factors and the simulated 'true' library sizes when a random set of OTUs are perturbed. The performance of different normalization methods are compared under different level of zero inflation, percentage of perturbed OTUs and strength of perturbation.

Table S1: 38 gut microbiome datasets (stool samples) from qiita ($n \geq 50$)

	study.object	study.ID	sample.size
1	infant gut fecal samples	101	63
2	infant fecal samples	10293	144
3	human and canine fecal samples	10394	1535
4	mice fecal sample	10469	391
5	human fecal samples	1561	52
6	human(HIV) fecal samples	1700	58
7	Cape Buffalo fecal samples	1736	642
8	Skin, oral and fecal samples	1841	3735
9	stool New-Onset Crohns Disease	1998	284
10	TwinsUK population fecal samples	2014	1081
11	Saliva, skin and fecal samples from ICU patients	2136	554
12	human fecal samples	455	92
13	human fecal samples	457	91
14	mice fecal microbiota	654	212
15	pregnant women fecal samples	867	1007
16	human infant gut	10297	85
17	monkey gut	10315	199
18	Grant gazelle gut	10323	768
19	human gut western Oklahoma	10342	58
20	human gastrointestinal gut	1070	118
21	human gut	1189	436
22	zebrafish gut	1192	50
23	Asian primates gut	1453	318
24	cow hindgut	1621	192
25	mice gut	1634	294
26	monkey gut	1696	172
27	bat gut	1734	96
28	colobine primates gut	2182	167
29	human gut and salivary	2202	820
30	bat gut	2338	192
31	human gut and mouth, and skin	449	602
32	humann gut microbiome (mouse samples)	452	160
33	humann gut microbiome (mouse samples)	456	158
34	human gastrointestinal	492	77
35	human gut (obese and lean twins)	77	281
36	human gut	850	528
37	freshwater fish slime and gut	940	288
38	Iguanas gut	963	100

Table S2: The frequency of 1st rank in the 38 real stool microbiome data sets.

	GMPR	CSS	RLE	RLE+	TMM	TMM+	TSS	RAW
OTU(All)	22	7	0	0	0	0	8	1
OTUs(Top)	23	3	1	1	3	0	7	0
OTUs(Middle)	20	8	0	0	1	0	9	0
OTUs(Bottom)	20	8	0	0	2	2	6	0