

Supplementary materials for “**Transformation and differential abundance analysis of microbiome data incorporating phylogeny**”

Chao Zhou^{1,3}, Hongyu Zhao^{2,3,*}, and Tao Wang^{1,3,4,*}

¹Department of Bioinformatics and Biostatistics, Shanghai Jiao Tong University

²Department of Biostatistics, Yale University

³SJTU-Yale Joint Center for Biostatistics and Data Science, Shanghai Jiao Tong University

⁴MoE Key Lab of Artificial Intelligence, Shanghai Jiao Tong University

*Corresponding author: hongyu.zhao@yale.edu; neowangtao@sjtu.edu.cn

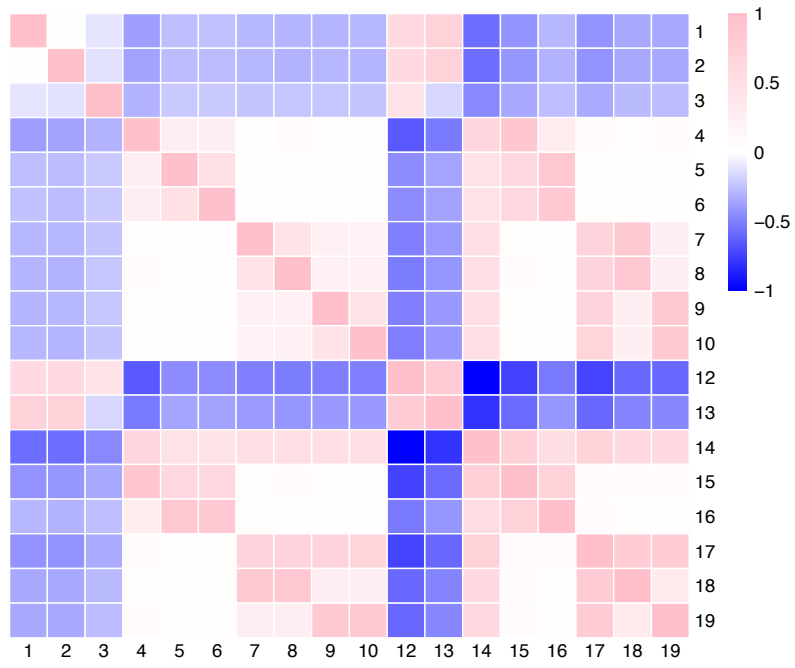


Figure S1. The sample correlations among counts on the tree in Figure 1 when the underlying distribution is DTM.

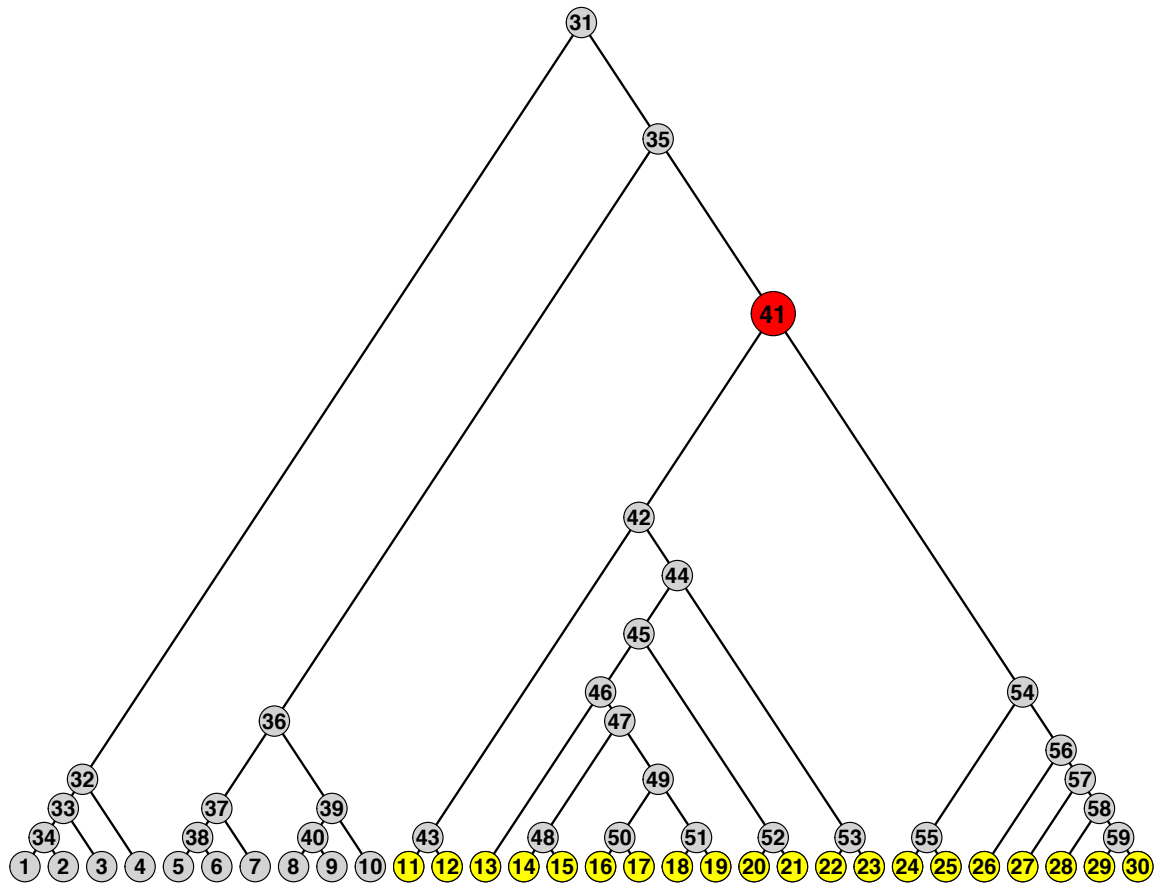


Figure S2. A binary tree with $K = 30$. The internal node 41 (red) and the corresponding leaves (yellow) were set to be differentially abundant (DA) between two groups.

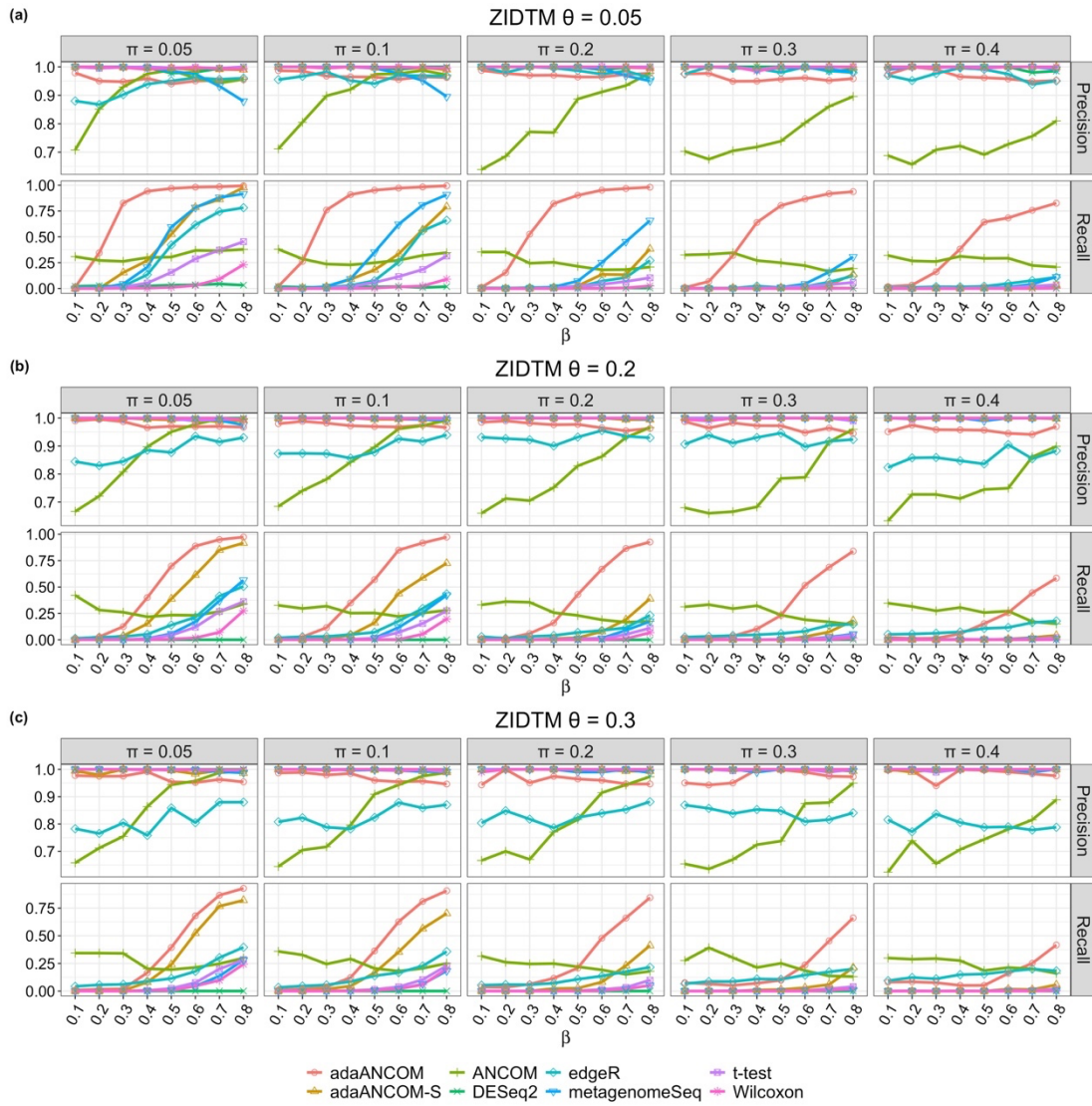


Figure S3. Precision and recall comparison of different DA testing methods. (a) Data were generated from ZIDTM with varying values of zero-inflation proportion $\pi_v = \pi$ and a fixed dispersion $\theta_v = \theta = 0.05$, with the tree and DA pattern depicted in Figure S2; (b) Data were generated in the same way as in (a), except that $\theta = 0.2$; (c) Data were generated in the same way as in (a), except that $\theta = 0.3$.

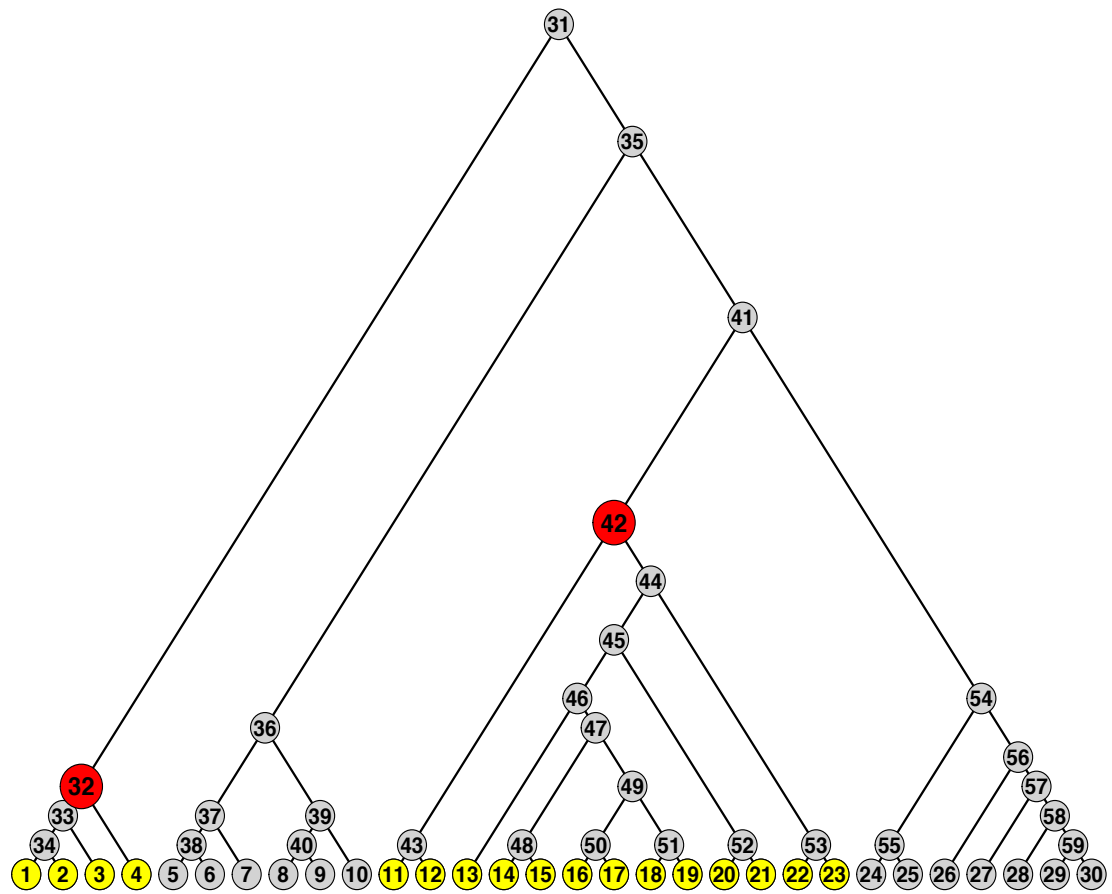


Figure S4. The same binary tree as in Figure S2, except that two internal nodes 32 and 42 (red), and the corresponding leaves (yellow) were set to be DA between two groups.

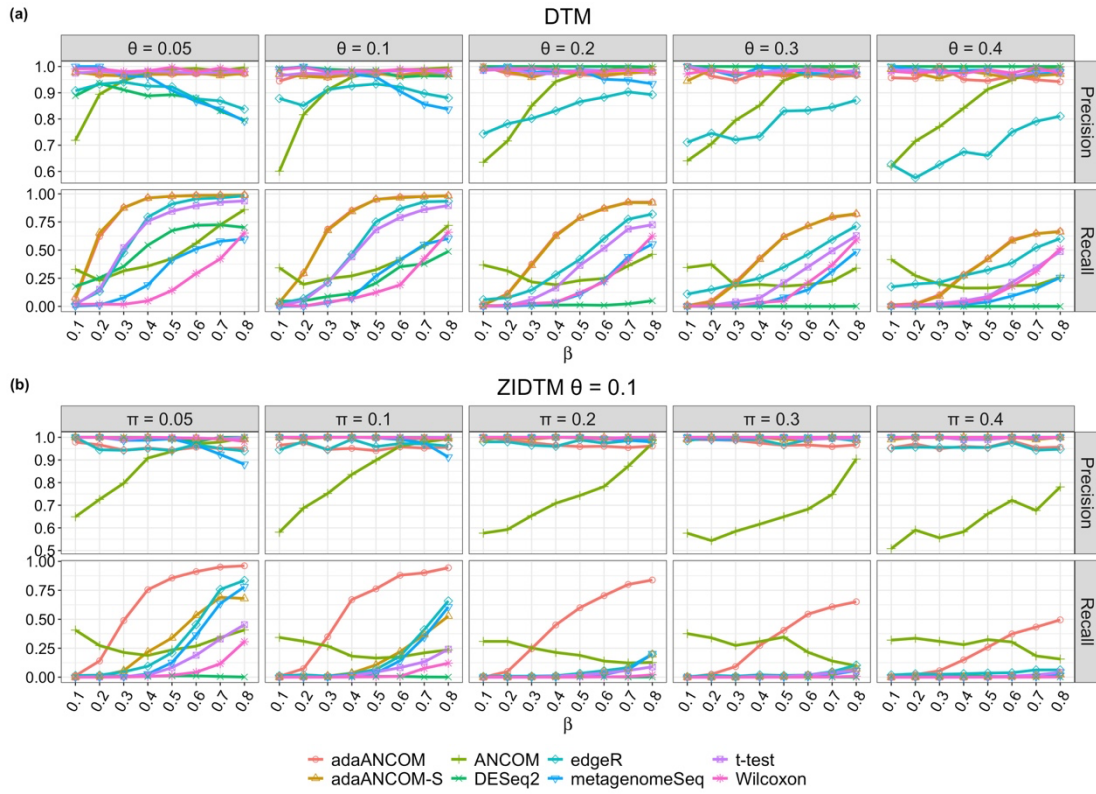


Figure S5. Precision and recall comparison of different DA testing methods. (a) Data were generated from DTM with varying values of dispersion parameter $\theta_v = \theta$ and effect size β , with the tree and DA pattern depicted in Figure S4; (b) Data were generated from ZIDTM with varying values of zero-inflation proportion $\pi_v = \pi$ and a fixed dispersion $\theta_v = \theta = 0.1$, with the tree and DA pattern depicted in Figure S4.

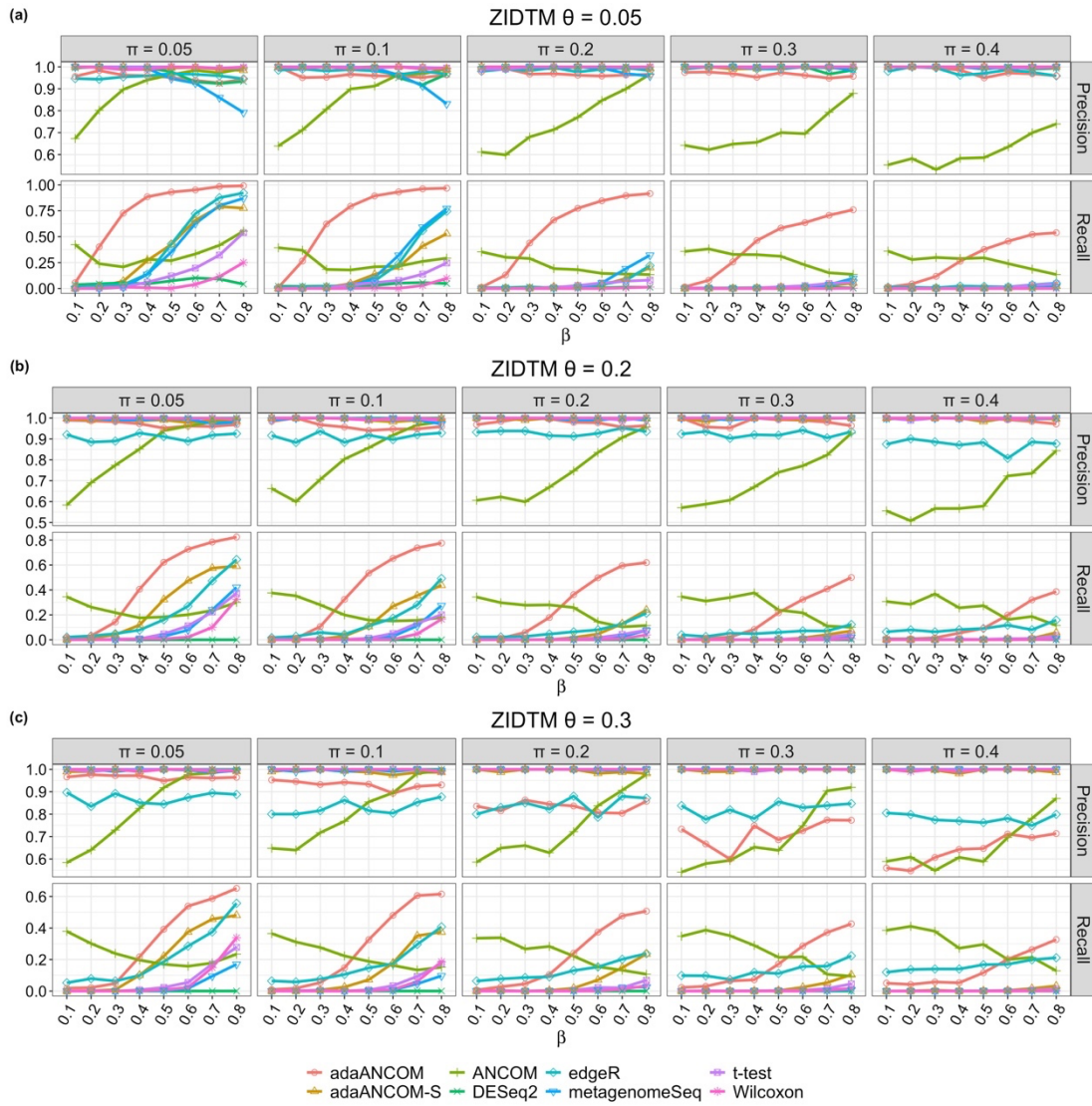


Figure S6. Precision and recall comparison of different DA testing methods. (a) Data were generated from ZIDTM with varying values of zero-inflation proportion $\pi_v = \pi$ and a fixed dispersion $\theta_v = \theta = 0.05$, with the tree and DA pattern depicted in Figure S4; (b) Data were generated in the same way as in (a), except that $\theta = 0.2$; (c) Data were generated in the same way as in (a), except that $\theta = 0.3$.

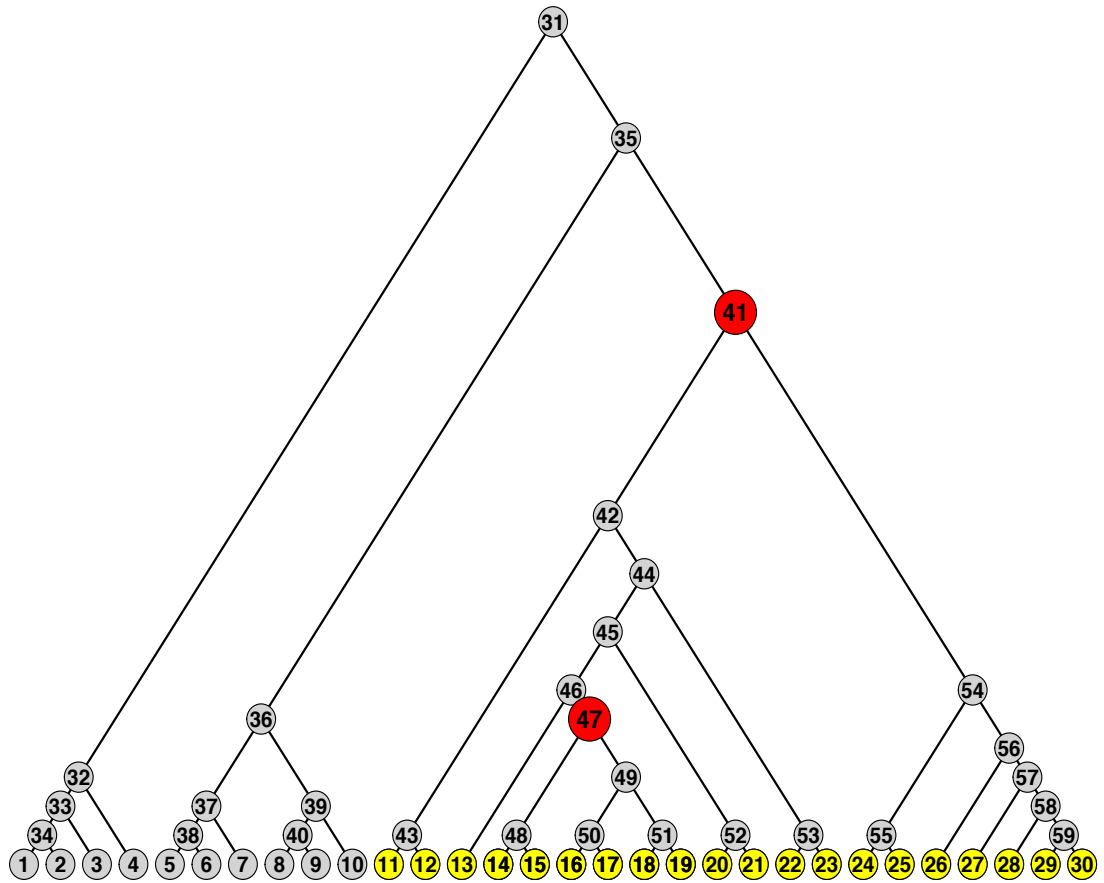


Figure S7. The same binary tree as in Figures S2 and S4, except that two nested internal nodes 41 and 47 (red), and the corresponding leaves (yellow) were set to be DA between two groups.

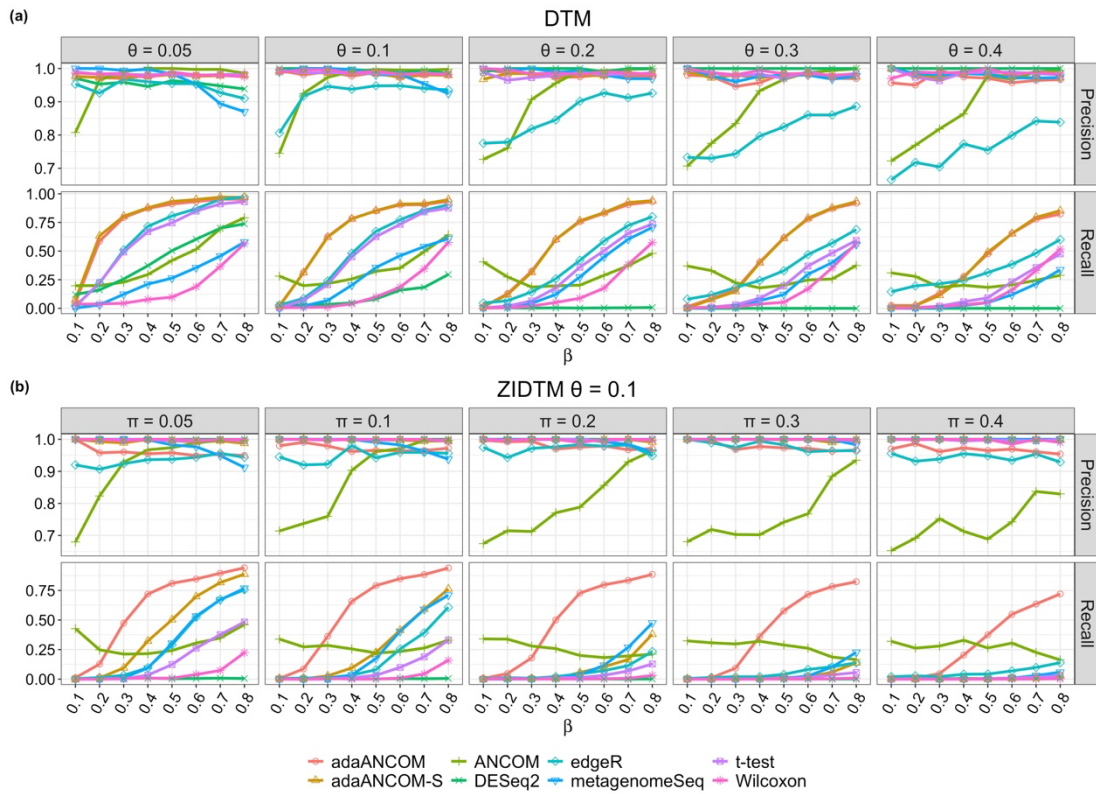


Figure S8. Precision and recall comparison of different DA testing methods. (a) Data were generated from DTM with varying values of dispersion parameter $\theta_v = \theta$ and effect size β , with the tree and DA pattern depicted in Figure S7; (b) Data were generated from ZIDTM with varying values of zero-inflation proportion $\pi_v = \pi$ and a fixed dispersion $\theta_v = \theta = 0.1$, with the tree and DA pattern depicted in Figure S7.

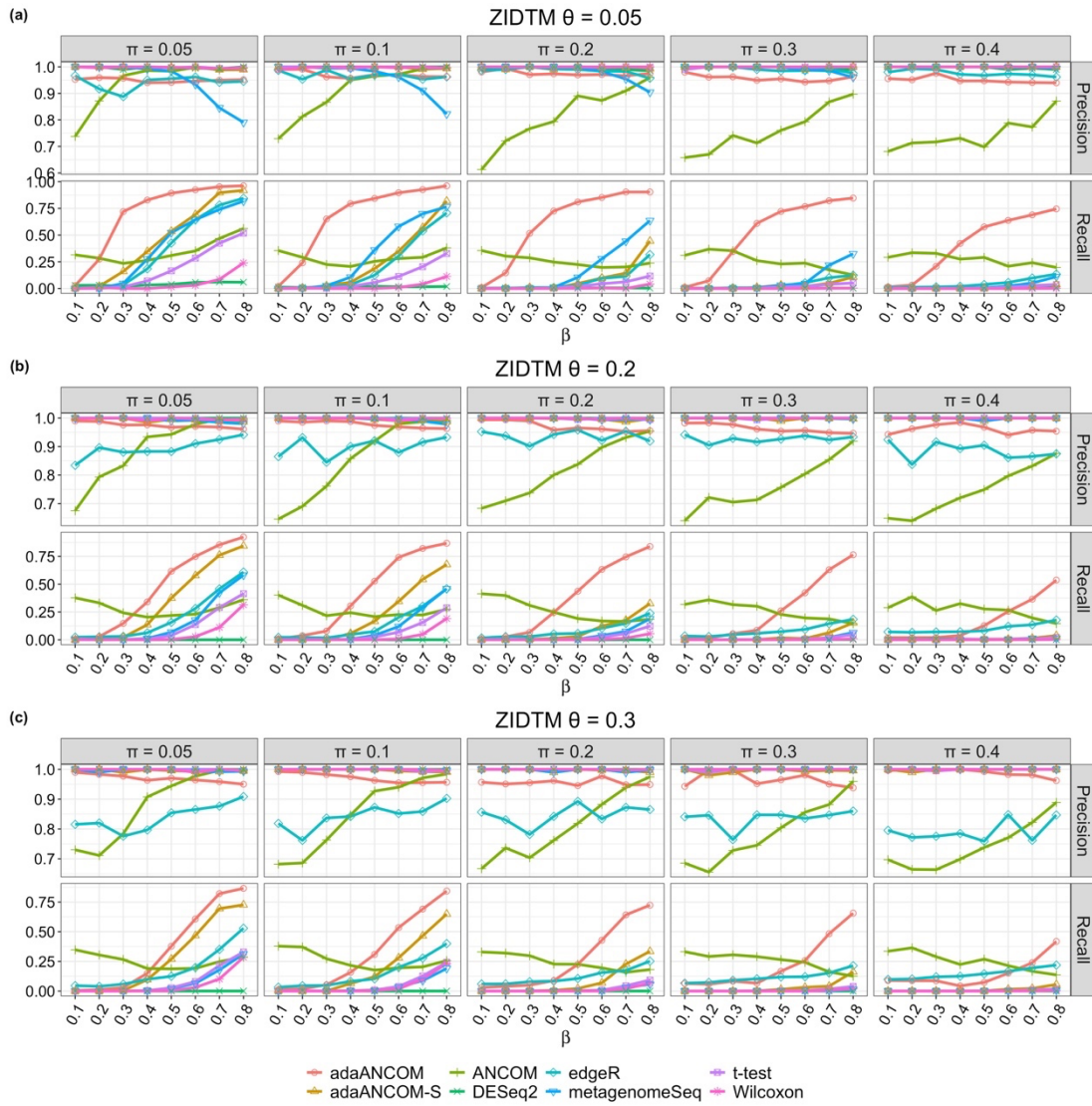


Figure S9. Precision and recall comparison of different DA testing methods. (a) Data were generated from ZIDTM with varying values of zero-inflation proportion $\pi_v = \pi$ and a fixed dispersion $\theta_v = \theta = 0.05$, with the tree and DA pattern depicted in Figure S7; (b) Data were generated in the same way as in (a), except that $\theta = 0.2$; (c) Data were generated in the same way as in (a), except that $\theta = 0.3$.

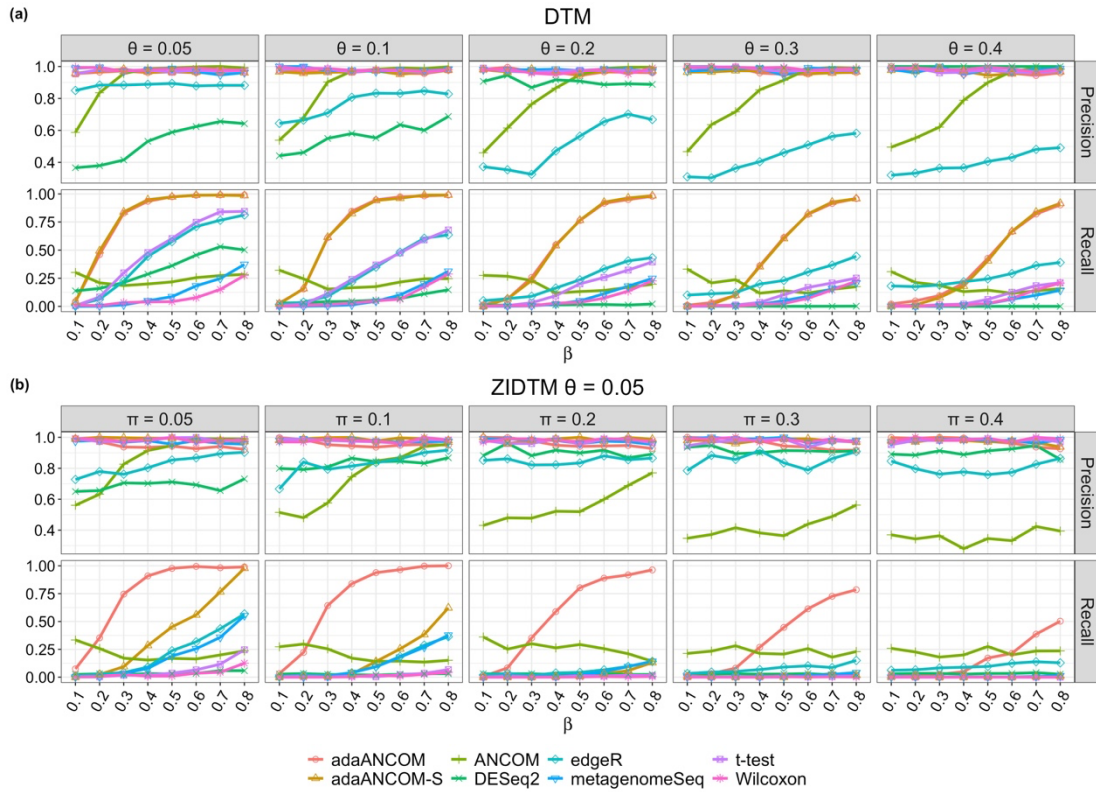


Figure S11. Precision and recall comparison of different DA testing methods. (a) Data were generated from DTM with varying values of dispersion parameter $\theta_v = \theta$ and effect size β , with the tree and DA pattern depicted in Figure S10; (b) Data were generated from ZIDTM with varying values of zero-inflation proportion $\pi_v = \pi$ and a fixed dispersion $\theta_v = \theta = 0.05$, with the tree and DA pattern depicted in Figure S10.

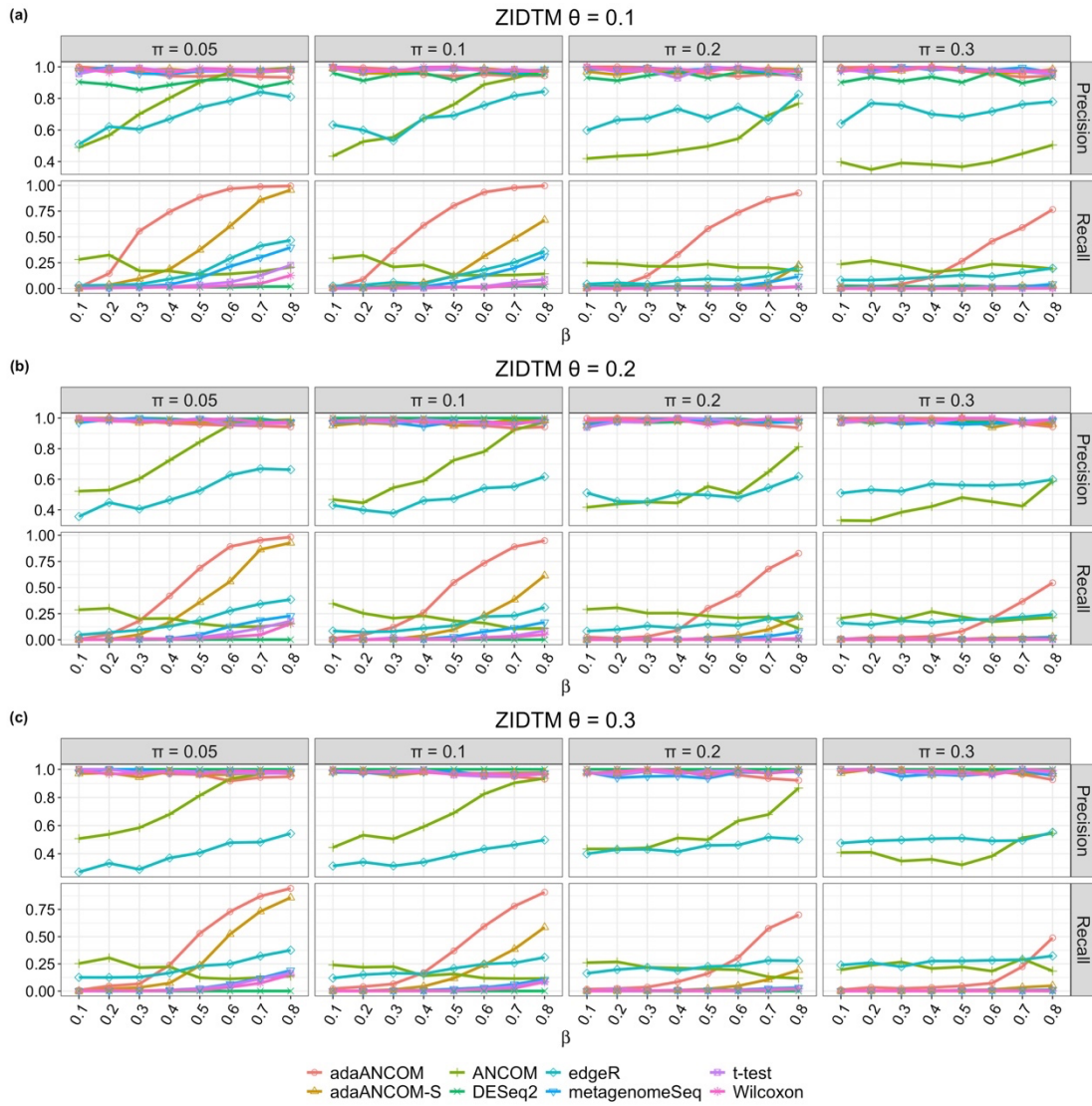


Figure S12. Precision and recall comparison of different DA testing methods. (a) Data were generated from ZIDTM with varying values of zero-inflation proportion $\pi_v = \pi$ and a fixed dispersion $\theta_v = \theta = 0.1$, with the tree and DA pattern depicted in Figure S10; (b) Data were generated in the same way as in (a), except that $\theta = 0.2$; (c) Data were generated in the same way as in (a), except that $\theta = 0.3$.

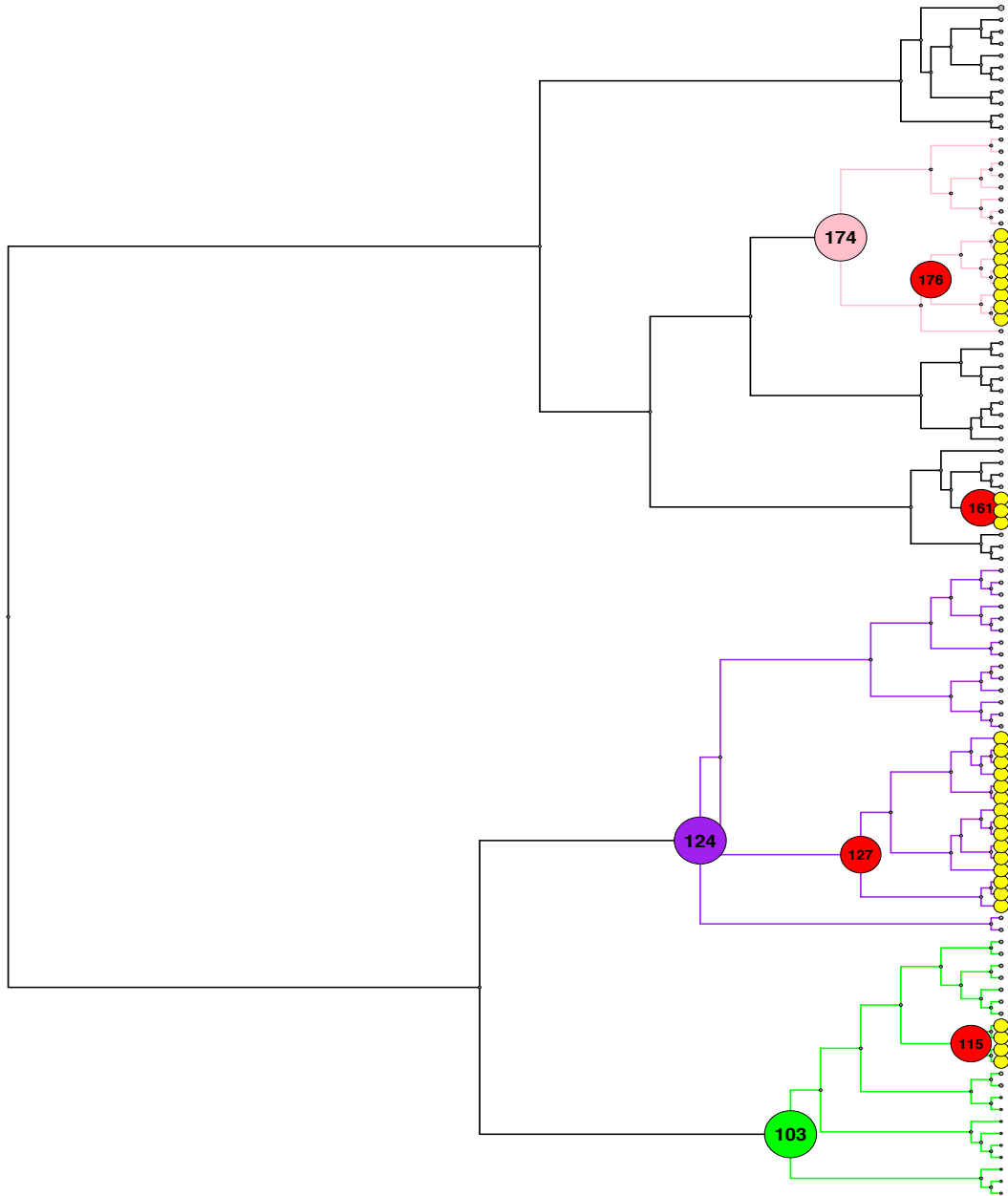


Figure S13. A binary tree with $K = 100$. DA internal nodes are shown in red and DA leaves are shown in yellow. Three subtrees, indexed by the internal nodes 103, 124, and 174, are depicted in green, blue, and pink, respectively. Data were generated from ZIDTM with parameters learned based on the HMP data, as detailed in the next figure.

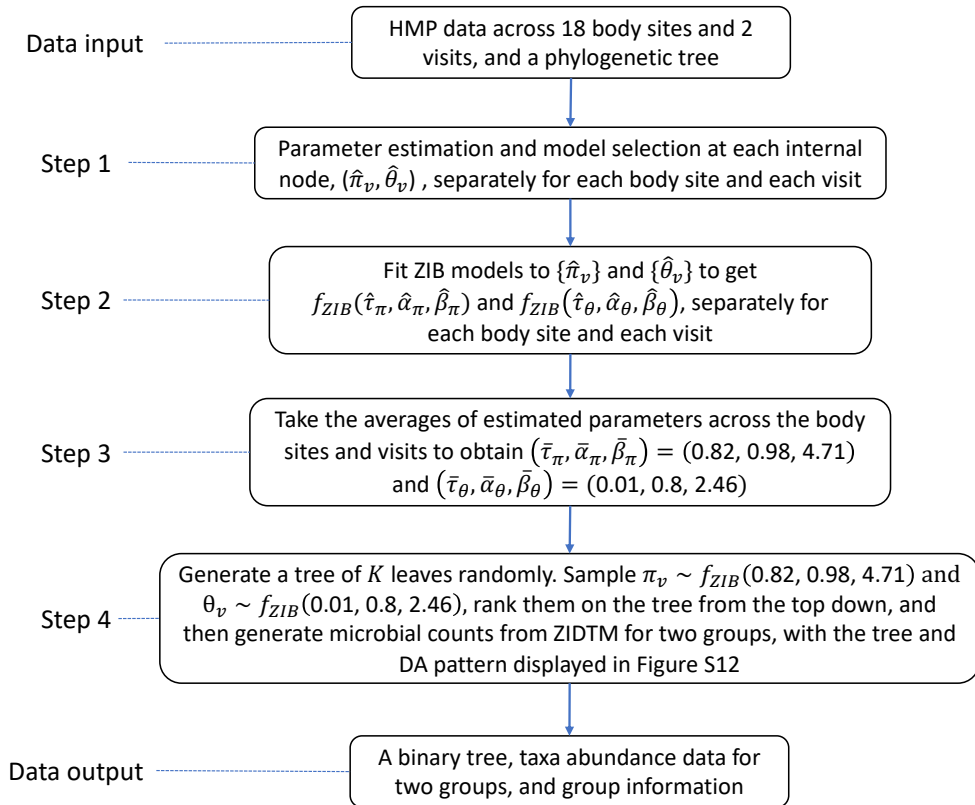


Figure S14. Data generation process. Taxa abundance data were generated from ZIDTM with parameters estimated based on HMP data. Here τ represents the probability of zero-inflation in the ZIB model. This is to be distinguished from π in zero-inflated models of the main text.

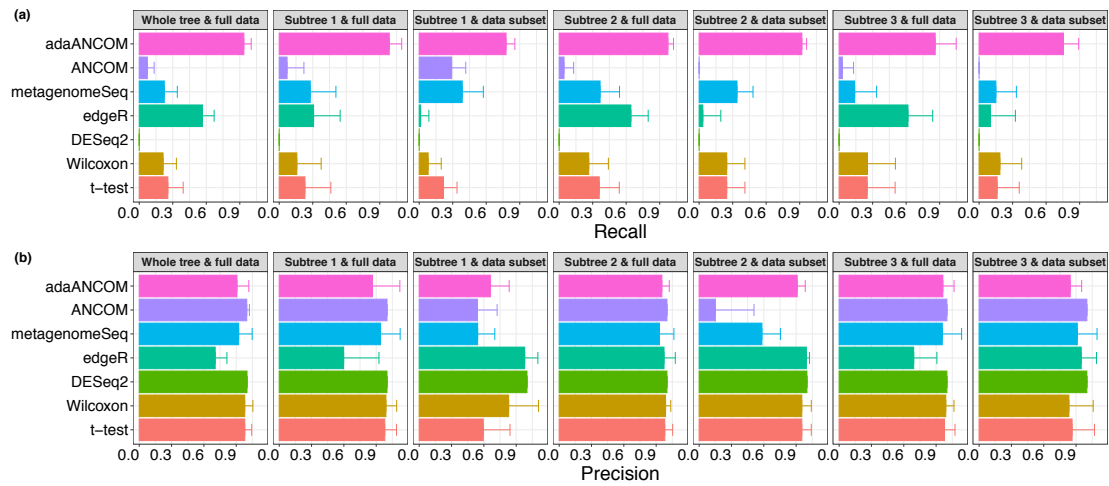


Figure S15. Recall and precision comparison of different DA testing methods. Data were generated from ZIDTM with parameters estimated based on the HMP data, and with the tree and DA pattern depicted in Figure S13. adaANCOM had the highest recall and comparable precision with other methods across all scenarios.

	Methods	DTM			ZIDTM		
		Recall	Precision	F1	Recall	Precision	F1
Setting 1	t-test	0.9460	0.6541	0.7734	0.6120	0.6263	0.6137
		0.0243	0.0065	0.0130	0.1117	0.0672	0.0670
	Wilcoxon	0.8615	0.6326	0.7295	0.8445	0.6400	0.7274
		0.0245	0.0064	0.0129	0.0594	0.0243	0.0319
	metagenomeSeq	0.7085	0.9879	0.8243	0.7990	0.6505	0.7164
		0.0356	0.0351	0.0234	0.0678	0.0278	0.0407
	ANCOM	0.7390	0.9994	0.8489	0.3571	0.8764	0.5068
		0.0430	0.0059	0.0307	0.0456	0.0609	0.0503
	DESeq2	0.8460	0.6605	0.7132	0.1295	1.0000	0.1768
		0.1845	0.0877	0.1493	0.2099	0.0000	0.2857
	edgeR	0.8915	0.6515	0.7513	0.8340	0.7490	0.7767
		0.0310	0.0528	0.0266	0.0791	0.1453	0.0581
	adaANCOM-S	0.9500	0.9595	0.9546	0.8335	0.9868	0.9017
		0.0174	0.0207	0.0138	0.0772	0.0245	0.0473
	adaANCOM	0.9505	1.0000	0.9745	0.9410	1.0000	0.9672
		0.0181	0.0000	0.0098	0.0911	0.0000	0.0522
Setting 2	t-test	0.9300	0.5489	0.6903	0.6812	0.5031	0.5775
		0.0232	0.0067	0.0116	0.0680	0.0367	0.0413
	Wilcoxon	0.8765	0.5340	0.6636	0.6935	0.4776	0.5655
		0.0177	0.0052	0.0091	0.0497	0.0201	0.0303
	metagenomeSeq	0.6265	0.8073	0.7010	0.8553	0.5762	0.6878
		0.0676	0.0817	0.0498	0.0505	0.0342	0.0342
	ANCOM	0.7291	0.7971	0.7595	0.6714	0.8044	0.7302
		0.0650	0.0410	0.0387	0.0544	0.0952	0.0565
	DESeq2	0.9106	0.5831	0.6869	0.1194	0.9293	0.1193
		0.1895	0.0764	0.0872	0.2489	0.1474	0.2482
	edgeR	0.8776	0.5425	0.6702	0.8953	0.6137	0.7247
		0.0492	0.0160	0.0237	0.0477	0.0927	0.0675
	adaANCOM-S	0.9194	0.9939	0.9541	0.5824	0.9764	0.7270
		0.0546	0.0175	0.0255	0.0717	0.0414	0.0587
	adaANCOM	0.9194	0.9983	0.9564	0.7859	0.9704	0.8662
		0.0546	0.0122	0.0290	0.0682	0.0478	0.0423

Table S1. Recall, precision, and F1 comparison of different DA testing methods. Initially, data were generated from DTM ($\theta_v = \theta = 0.1$) or ZIDTM ($\theta_v = \theta = \pi_v = \pi = 0.1$) with the same parameters for the two groups, and tree depicted in Figure S2. Then we multiplied the counts for leaf nodes in yellow depicted in Figure S2 (Setting 1) and Figure S4 (Setting 2) by some effect size for one group. The log effect size was drawn uniformly from -5 to 5 . Each method has associated with it two rows showing respectively the mean and standard error over 100 replicates.

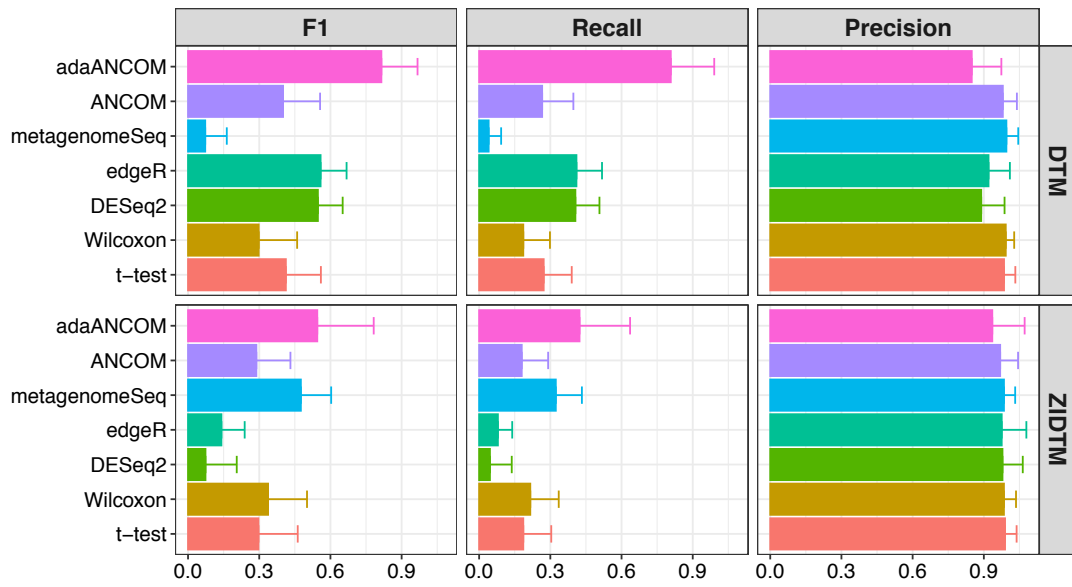


Figure S16. Recall, precision, and F1 comparison of different DA testing methods. Data were generated from SparseDOSSA 2 with parameters estimated based on synthetic DTM ($\theta_v = \theta = 0.1$) and ZIDTM ($\theta_v = \theta = \pi_v = \pi = 0.1$) data, and with the tree and DA pattern depicted in Figure S2. adaANCOM had the highest recall and comparable precision with other methods across both scenarios.

<i>Body site</i>	<i>Sample size</i>	<i>Species</i>	<i>Genus</i>	<i>Family</i>	<i>Order</i>	<i>Class</i>	<i>Phylum</i>
<i>Anterior_nares</i>	77	612	35	32	16	12	8
		0.60	0.47	0.41	0.35	0.35	0.35
<i>Attached_Keratinized_gingiva</i>	104	881	39	29	16	11	7
		0.63	0.40	0.35	0.23	0.11	0.15
<i>Buccal_mucosa</i>	104	1221	47	35	19	13	9
		0.62	0.36	0.33	0.22	0.15	0.19
<i>Hard_palate</i>	99	1244	49	35	18	14	9
		0.61	0.32	0.26	0.15	0.16	0.15
<i>Left_Antecubital_fossa</i>	60	341	33	37	16	12	6
		0.62	0.55	0.52	0.41	0.38	0.25
<i>Left_Retroauricular_crease</i>	105	489	23	25	12	8	6
		0.58	0.52	0.49	0.37	0.28	0.30
<i>Mid_vagina</i>	52	430	26	26	14	10	5
		0.65	0.57	0.53	0.43	0.33	0.16
<i>Palatine_Tonsils</i>	108	1289	56	37	21	15	10
		0.61	0.32	0.26	0.20	0.14	0.14
<i>Posterior_fornix</i>	53	388	18	21	11	8	5
		0.64	0.63	0.61	0.48	0.34	0.25
<i>Right_Antecubital_fossa</i>	63	348	35	33	16	11	6
		0.61	0.55	0.47	0.39	0.36	0.25
<i>Right_Retroauricular_crease</i>	107	440	21	24	12	8	5
		0.59	0.54	0.52	0.41	0.32	0.23
<i>Saliva</i>	79	1130	56	40	21	14	9
		0.61	0.35	0.30	0.21	0.12	0.10
<i>Stool</i>	108	1036	49	23	13	10	6
		0.65	0.38	0.36	0.44	0.32	0.15
<i>Subgingival_plaque</i>	104	1199	50	40	22	15	10
		0.63	0.32	0.33	0.27	0.18	0.20
<i>Supragingival_plaque</i>	108	1248	42	32	18	14	9
		0.61	0.29	0.28	0.21	0.20	0.19
<i>Throat</i>	90	1105	52	39	21	14	9
		0.60	0.34	0.32	0.21	0.12	0.11
<i>Tongue_dorsum</i>	112	1513	43	33	19	14	9
		0.60	0.28	0.27	0.20	0.14	0.14
<i>Vaginal_introitus</i>	52	465	30	28	14	10	5
		0.65	0.52	0.50	0.41	0.32	0.14

Table S2. Information on the preprocessed HMP data. Each body site has associated with it two rows showing respectively taxon number and zero proportion at different taxonomic levels.

<i>Body site</i>	<i>Visit</i>	<i>Species</i>	<i>Genus</i>	<i>Family</i>	<i>Order</i>	<i>Class</i>	<i>Phylum</i>
<i>Anterior_nares</i>	1	0.0736	0.2059	0.1935	0.1333	0.0909	0.2857
	2	0.0690	0.1471	0.1290	0.0667	0.0909	0.2857
<i>Attached_Keratinized_gingiva</i>	1	0.0676	0.0263	0.0357	0.0000	0.0000	0.1667
	2	0.0507	0.1053	0.1071	0.0667	0.0000	0.0000
<i>Buccal_mucosa</i>	1	0.0472	0.0870	0.0882	0.1111	0.0000	0.0000
	2	0.0552	0.0652	0.0882	0.0556	0.0833	0.1250
<i>Hard_palate</i>	1	0.0632	0.0833	0.0882	0.1176	0.1538	0.1250
	2	0.0556	0.1458	0.1176	0.1765	0.0769	0.2500
<i>Left_Antecubital_fossa</i>	1	0.0902	0.3438	0.6111	0.3333	0.4545	0.0000
	2	0.0392	0.4688	0.3333	0.3333	0.3636	0.0000
<i>Left_Retroauricular_crease</i>	1	0.0802	0.0909	0.2500	0.4545	0.4286	0.6000
	2	0.0430	0.0455	0.0417	0.0909	0.1429	0.0000
<i>Mid_vagina</i>	1	0.0427	0.0800	0.1200	0.0000	0.0000	0.0000
	2	0.0517	0.1304	0.1600	0.1538	0.1111	0.2500
<i>Palatine_Tonsils</i>	1	0.0722	0.0727	0.0278	0.1000	0.0714	0.1111
	2	0.0430	0.1091	0.1111	0.1000	0.0714	0.1111
<i>Posterior_fornix</i>	1	0.0377	0.0588	0.1000	0.0000	0.0000	0.0000
	2	0.0503	0.1176	0.1000	0.1000	0.0000	0.0000
<i>Right_Antecubital_fossa</i>	1	0.0646	0.4118	0.5000	0.6000	0.4000	0.0000
	2	0.0646	0.2941	0.3438	0.2000	0.5000	0.4000
<i>Right_Retroauricular_crease</i>	1	0.0868	0.1000	0.2174	0.1818	0.2857	0.2500
	2	0.0547	0.1000	0.1739	0.1818	0.0000	0.2500
<i>Saliva</i>	1	0.0826	0.3455	0.3846	0.4000	0.2308	0.1250
	2	0.0941	0.3455	0.3077	0.4000	0.3846	0.2500
<i>Stool</i>	1	0.1251	0.2292	0.2273	0.1667	0.0000	0.0000
	2	0.1275	0.1667	0.2727	0.2500	0.2222	0.0000
<i>Subgingival_plaque</i>	1	0.0833	0.2041	0.1538	0.1905	0.2143	0.1111
	2	0.0656	0.1429	0.1026	0.1429	0.0714	0.3333
<i>Supragingival_plaque</i>	1	0.0721	0.0976	0.1290	0.1176	0.0769	0.1250
	2	0.0669	0.1463	0.1613	0.1765	0.1538	0.1250
<i>Throat</i>	1	0.0769	0.1961	0.1842	0.1500	0.1538	0.1250
	2	0.0553	0.0784	0.1316	0.0500	0.0000	0.0000
<i>Tongue_dorsum</i>	1	0.0685	0.1667	0.1250	0.1111	0.0769	0.0000
	2	0.0526	0.1190	0.1875	0.2222	0.0769	0.1250
<i>Vaginal_introitus</i>	1	0.0421	0.0690	0.0741	0.0000	0.0000	0.0000
	2	0.0343	0.0345	0.1111	0.0000	0.0000	0.0000

Table S3. The proportions of selecting ZIBB at different taxonomic levels across 18 body sites and 2 visits for the HMP data.

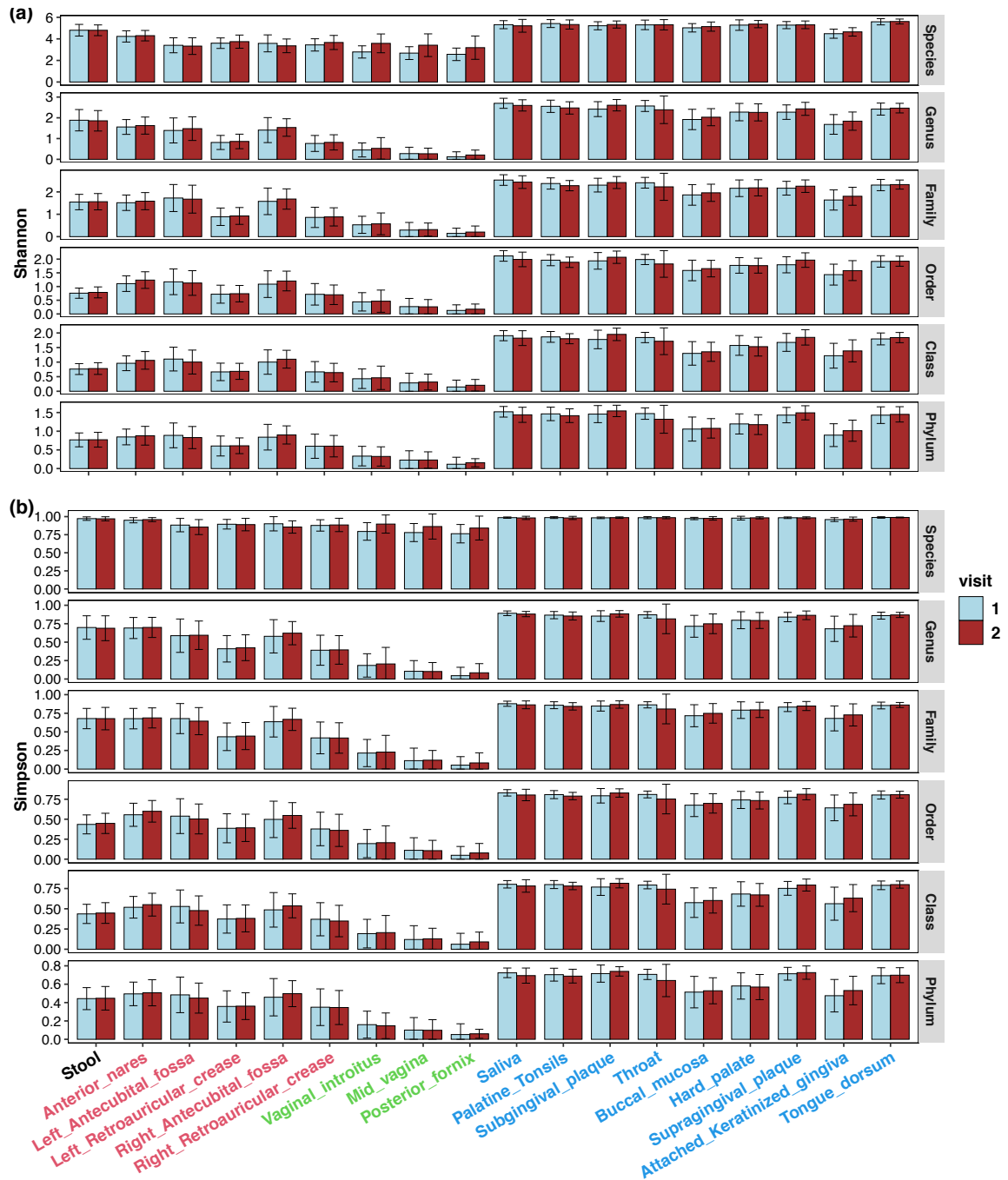


Figure S17. Comparison of Shannon's index and Simpson's index between two visits of each body site across different taxonomic levels. There is no significant difference in alpha diversity after the Bonferroni correction.

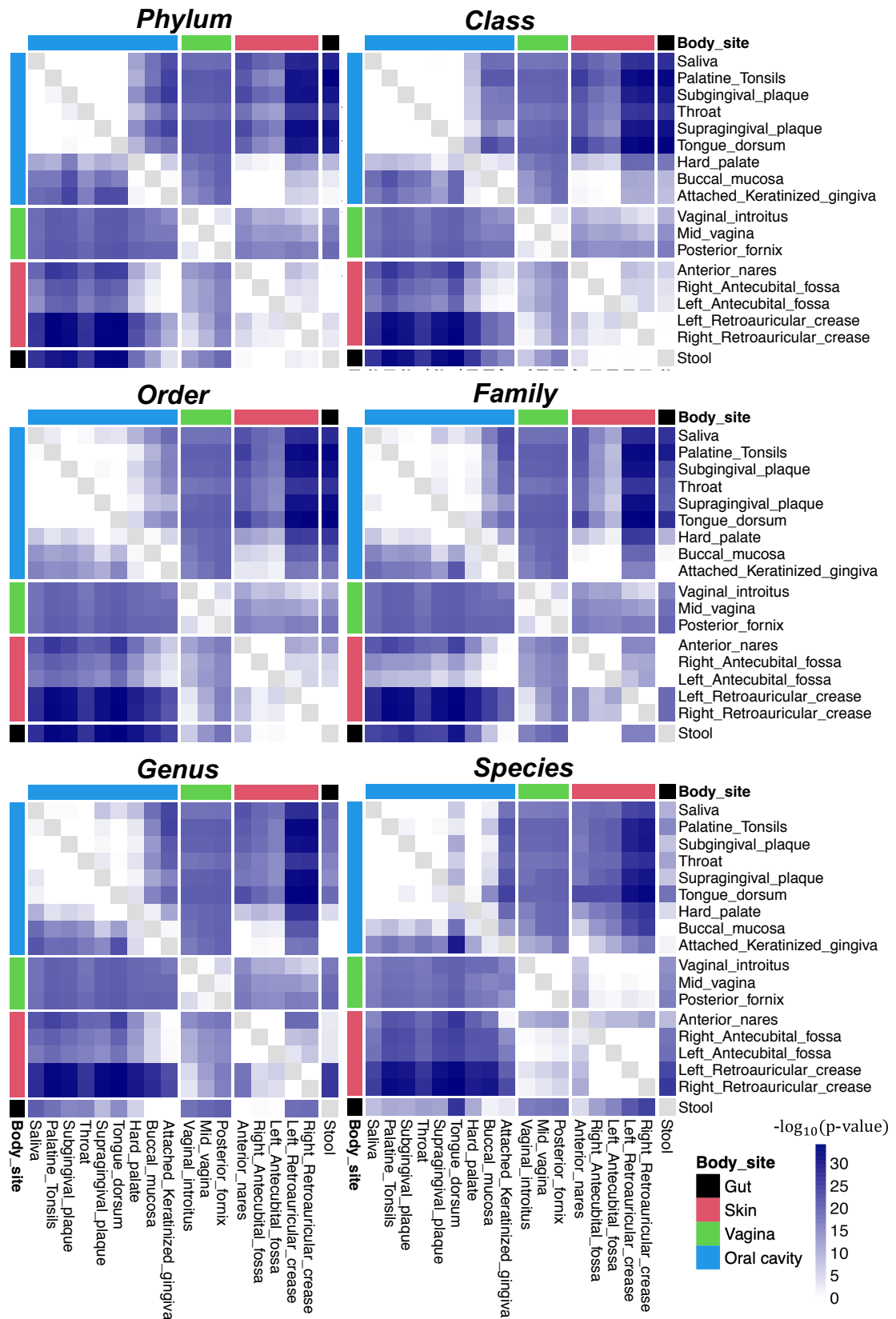


Figure S18. Comparing the alpha diversity between each pair of body sites at different levels. P-values were obtained using the Wilcoxon rank-sum test and Bonferroni correction. The upper and lower triangles show the results for the Shannon's index and the Simpson's index, respectively.

	t-test	Wilcoxon	ANCOM	metagenomeSeq	edgeR	adaANCOM	adaANCOM-S
Before	69	197	3	192	701	336	271
After	0	0	3	0	17	2	0

Table S4. Numbers of detected species by various methods, before and after multiple testing correction, for HMP stool data between two visits. There were 6965 species of 108 samples for the DA analysis and the zero proportion was about 90%. edgeR stood as an exception, and species detected by it were likely false discoveries. An error occurred for running DESeq2 to estimate the size factor.

<i>Body site</i>	Species number	t-test	Wilcoxon	ANCOM	MetagenomeSeq	edgeR	ada-ANCOM	ada-ANCOM-S
<i>Saliva</i>	839	665	611	657	603	-	97	96
<i>Anterior_nares</i>	1126	969	585	299	168	-	63	64
<i>Right_Antecubital_fossa</i>	1072	940	539	240	161	-	72	73
<i>Left_Retroauricular_crease</i>	1330	1123	860	265	105	-	78	73
<i>Left_Antecubital_fossa</i>	1209	1054	672	261	178	-	63	63
<i>Right_Retroauricular_crease</i>	1031	875	603	250	97	683	65	64
<i>Palatine_Tonsils</i>	684	444	489	496	376	-	86	86
<i>Subgingival_plaque</i>	602	219	309	277	257	405	71	63
<i>Throat</i>	930	709	724	894	714	778	95	90
<i>Vaginal_introitus</i>	500	398	306	173	131	-	47	46
<i>Buccal_mucosa</i>	617	409	446	333	308	494	77	68
<i>Hard_palate</i>	672	385	513	550	414	475	89	89
<i>Supragingival_plaque</i>	513	373	262	227	179	327	51	51
<i>Mid_vagina</i>	364	322	188	147	152	247	31	34
<i>Attached_Keratinized_gingiva</i>	622	273	399	314	383	474	55	53
<i>Posterior_fornix</i>	268	249	107	117	97	-	22	24
<i>Tongue_dorsum</i>	516	445	372	376	111	343	76	75

Table S5. DA results between stool and other body sites based on HMP data at the species level (- means an error occurred). The proportion of zeros ranged from 81% to 90%, the total number of species varied between 243 and 1123. There were some errors occurred when running DESeq2 and edgeR.



Figure S19. The phylogenetic tree of 60 bacterial taxa inferred by maximum likelihood. The text next to each OTU gives the taxonomic information and the rank of relative importance obtained by applying the random forest algorithm.

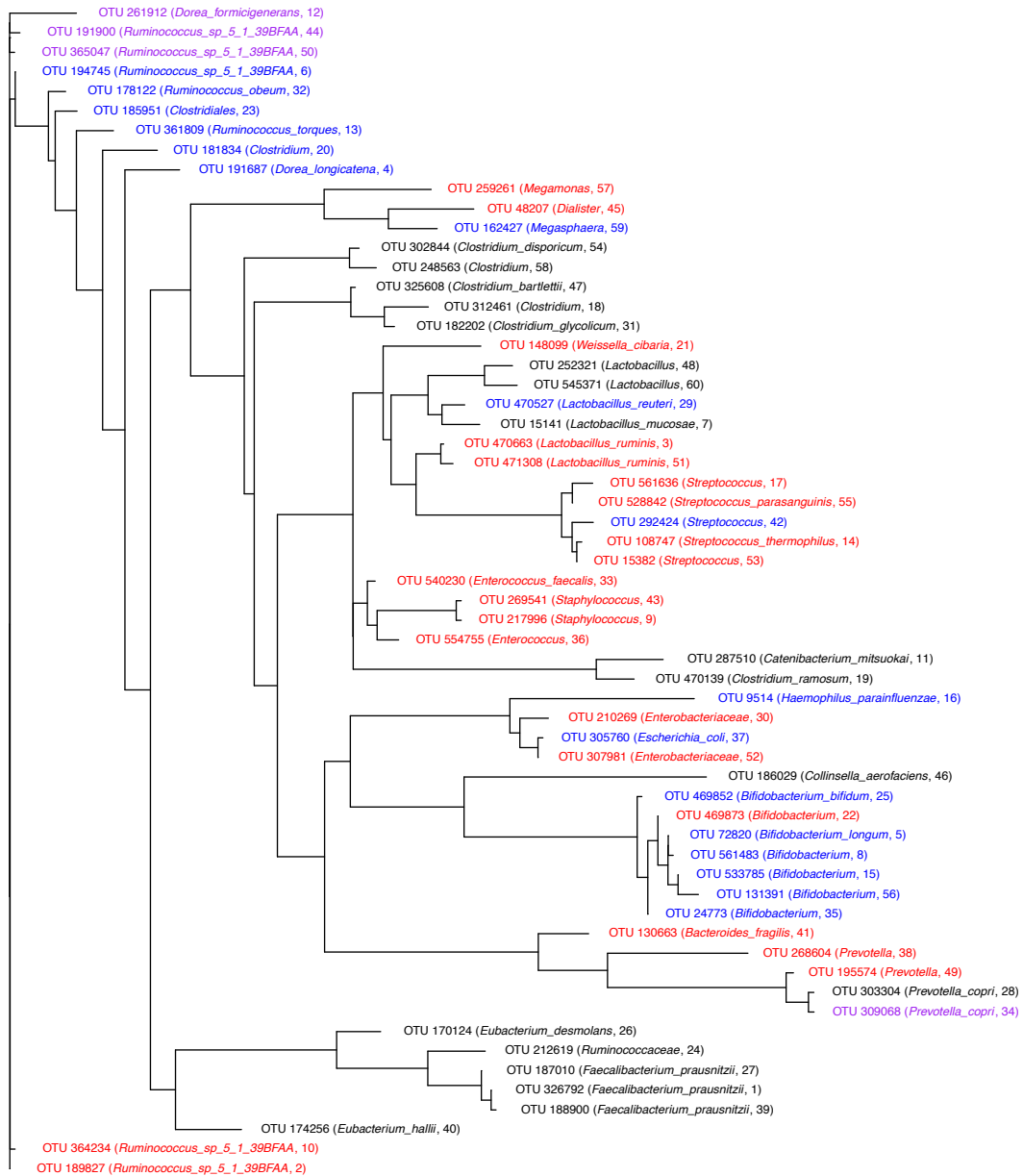


Figure S20. Tree-based visualization of DA testing results. Species detected uniquely by adaANCOM, uniquely by its competitors, and by adaANCOM and one of its competitors, are shown in red, purple, and blue, respectively.