Supplementary Information Text

Supplementary Methods

The methods described below gave detailed information regarding how the modification and expansion to the original Genome Taxonomy Database (GTDB) bacterial phylogeny was made in each step.

Determination of NCBI Taxonomy ID for Each Genome.

Tip labels in the original GTDB phylogeny represent bacterial genomes, starting with "RS_" or "GB_" and followed by the National Center for Biotechnology Information (NCBI) GenBank and RefSeq assembly accession numbers. Genome accession numbers were collected by removing the prefix of each of the tip labels in the original phylogeny. The accession numbers were then searched against the NCBI Assembly database to retrieve the most up-to-date NCBI taxonomy ID (taxID) assigned to each genome assembly. The taxIDs for bacterial genomes that were used to build the Kraken2 bacterial database were retrieved directly from the Kraken2 seqid2taxid.map file.

Full Lineage Information Extraction.

Each of the non-redundant taxIDs retrieved above was searched against the NCBI Taxonomy database to fetch its full lineage information. As for retrieved full lineage information, taxIDs that corresponded to different taxonomic ranks were collected, including superkingdom, phylum, class, order, family, genus, species group, and species. The lineage information was recorded in tip_lineage.tsv (for each of the tips present in the original phylogeny), phylo_spp_lineage.tsv (for each of the non-redundant species-level taxIDs identified in the original phylogeny), and added_spp_lineage.tsv (for each of the non-redundant species-level taxIDs identified in the Kraken2 standard bacterial reference library). All searches against NCBI databases were processed in batch using the Biopython package (1).

Adding Species to the Original Phylogenetic Tree.

For each of the unique species identified in the Kraken2 and GTDB bacterial phylogeny sources (added_spp_lineage.tsv and phylo_spp_lineage.tsv), the sequence of taxIDs at different taxonomic ranks was searched against lineage information appended to the tip labels (tip_lineage.tsv), starting from the lowest rank (species) to the highest (superkingdom). This determined the lowest taxonomic rank possible where at least one of the genomes in the original GTDB bacterial phylogeny shared with the query species. Upon determination of the taxonomic rank to map that species, all tips within the original phylogeny that shared the same taxID at the corresponding rank were extracted and their most recent common ancestor (MRCA) node was identified using the getMRCA function from the ape R package (2). A subtree rooting at the MRCA node was extracted, and the average distance of all its children tips that shared the same taxID as the query species at the predetermined taxonomic rank to the MRCA node was calculated. Then the query species was added to the original phylogeny using the add.tips function from the phangorn R package with the MRCA node being the place to bind the tip and the computed average distance being the inserted branch length (3).

Removal of Potential Outliers and Tree Pruning.

Taxonomic misclassification can result in extreme outlier tips that could cause the computed MRCA node to reside close to the base root of the phylogeny, leading to long branch length assigned to the query species. For each member in the group of reference tips that were used to map a particular species (e.g., all tips sharing genus-level taxID 226 for locating species *Alteromonas* sp. 76-1, as the exact species was not found in the original phylogeny, therefore its location was inferred using its congeneric taxa), its average distance to the remaining group members was calculated. Then the mean and standard deviation were calculated for all these average distances. Candidate outliers were defined as tips whose average distance to the remaining group member tips exceeds mean plus N times standard deviations (N = 1, 2, 3, respectively). For species mapped at different taxonomic ranks, different N values were applied where the value was determined by considering how much improvement has been made in terms

of branch length distributions and how much phylogenetic information was retained after removal of the outlier tips (Fig. S6 and Tables S3-S5).

For groups containing only two reference tips where detection of potential outlier tips based on mean and standard deviation was impractical, the fraction (distance to the MRCA node) / (distance to the base root) for the more distant tip was used to indicate if an outlier was present in the reference group (threshold used in this study: fraction ≥ 0.75 to indicate the presence of an outlier tip in the reference group). The potential outlier tip was determined based on comparing the lowest taxonomic rank these two tips shared with their corresponding neighboring tips, where 2 nodes backward were taken to extract the subtree containing the target tip and its neighboring tips. The one with a comparatively higher-level taxonomic rank shared with the neighboring tips was determined to be the outlier. In the case where the two reference tips had the same level of shared taxonomic rank, the one with a longer average distance to its neighboring tips was selected as the outlier.

The addition of species was performed again after deletion of outlier tips. The resulting phylogeny was pruned by removing all original GTDB tips where each tip label represented a single bacterial genome assembly, so that all inter-tip distances within the pruned phylogeny represented interspecific distances under NCBI taxonomy system.

Supplementary Notes

The species *Isorropodon fossajaponicum* symbiont (taxID 883811) and *Abyssogena phaseoliformis* symbiont (taxID 596095) were identified in the Kraken2 standard bacterial reference library but were not mapped to our expanded phylogeny, as they could only be mapped at the superkingdom level where the inference of their location using the entire phylogeny was computationally infeasible and biologically meaningless. Deletion of these two species should be done for the Kraken2 report before phylogeny-based diversity analysis if they are present in the community.

Not all unique species names identified in the original GTDB bacterial tree were mapped back into it, as some represent “pseudo” species-level names (e.g., Enterobacteriaceae bacterium, taxID: 1849603, which indeed represented a family but was assigned the rank species). A filter for at least having a recorded species- and genus-level taxIDs in the full lineage was applied and only the filtered list of bacterial species identified in the phylogenetic source were added back to the expanded phylogeny.

For the expanded phylogeny using scientific names as tip labels, single quotes within species scientific names were replaced with whitespaces such that these names can be displayed properly by some phylogenetic tree visualization tools. Therefore, extracting the column with taxIDs from the Kraken2 classification report and utilizing the expanded phylogeny using taxIDs as tip labels should be more robust.

Updates and merging of NCBI taxIDs are continuous, for example, during this study, taxIDs 147467 and 861208 that were retrieved from the Kraken2 seqid2taxid.map file have been merged into taxIDs 1296 and 1183401, respectively. Upon warnings about duplicated tip label names when applying the expanded phylogeny in community diversity analysis, identification and combination of merged taxonomic records must be done for the Kraken2 classification reports using basic dataframe operations, for example, in R or Python, to ensure a match between the taxonomic labels in a Kraken report (or any other classifier) and the labels present in the phylogeny.

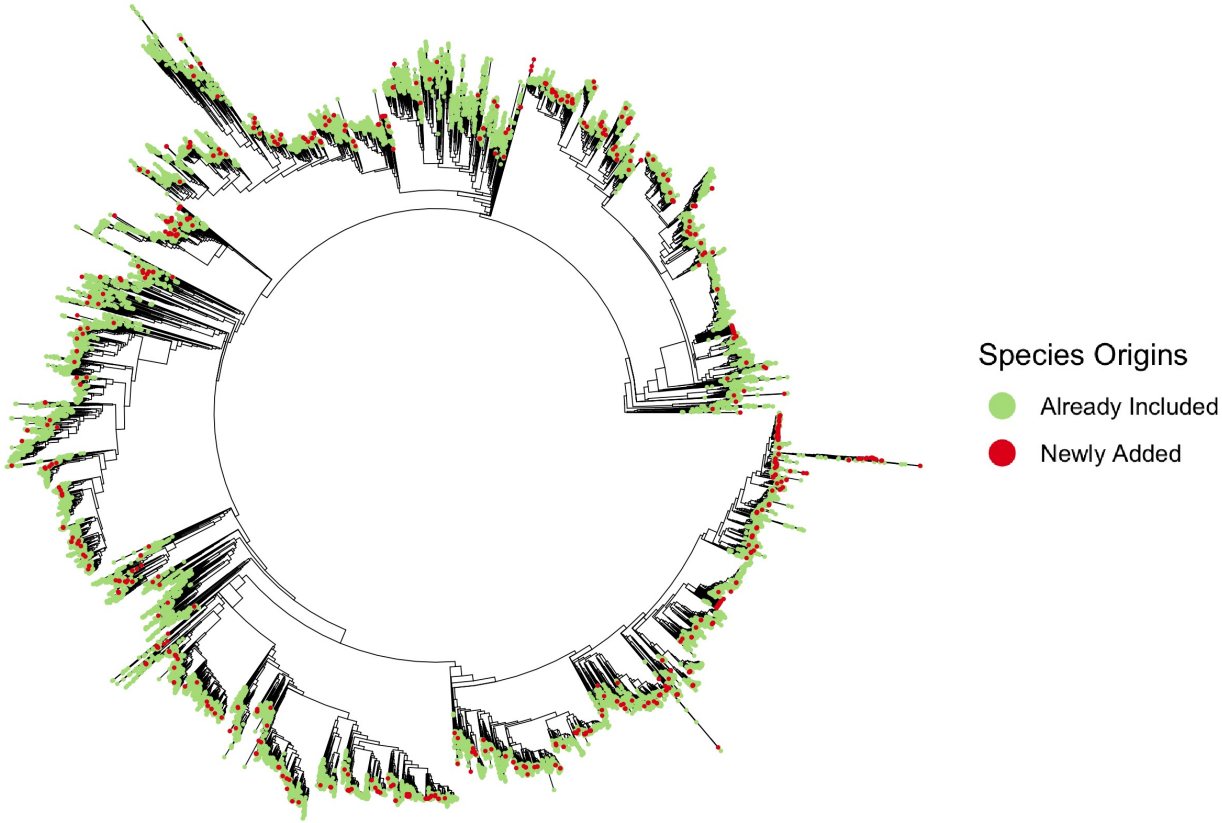

Fig. S1. Insertion places for newly added tip labels. A total of 1,411 additional species from the Kraken2 standard bacterial reference library were added to the original GTDB phylogeny. These newly added species are distributed across the entire phylogeny.

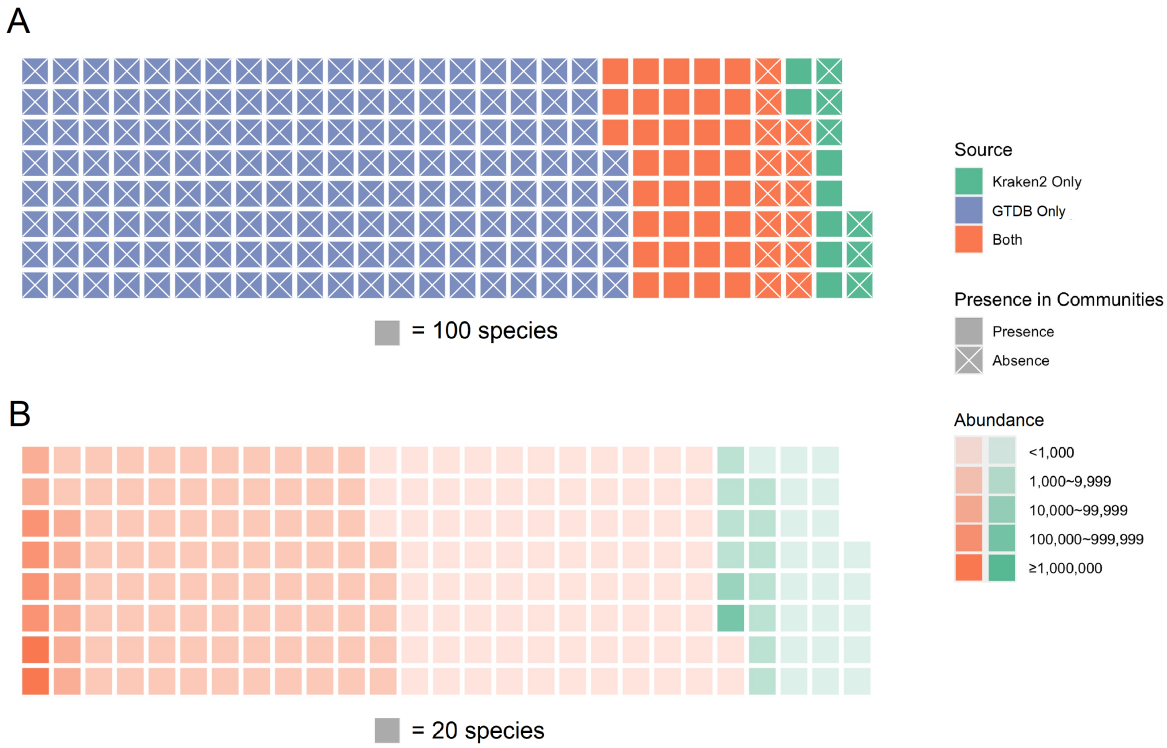

Fig. S2. Database source and community abundance of bacterial species present in the expanded phylogeny. (A) A plot to identify the source of bacterial species included in the expanded phylogeny and their presence/absence in microbiome communities used in this study: the original GTDB phylogeny significantly extended the range of bacterial taxonomy, but not all species found in the Kraken2 standard bacterial reference library can be identified in the original phylogeny (shown in green). (B) A plot for bacterial species that were present in our communities, with abundance information also presented via opacity: some species missing from the original GTDB phylogeny were found in the Kraken2 database and had high abundance, which were crucial for community diversity analysis.

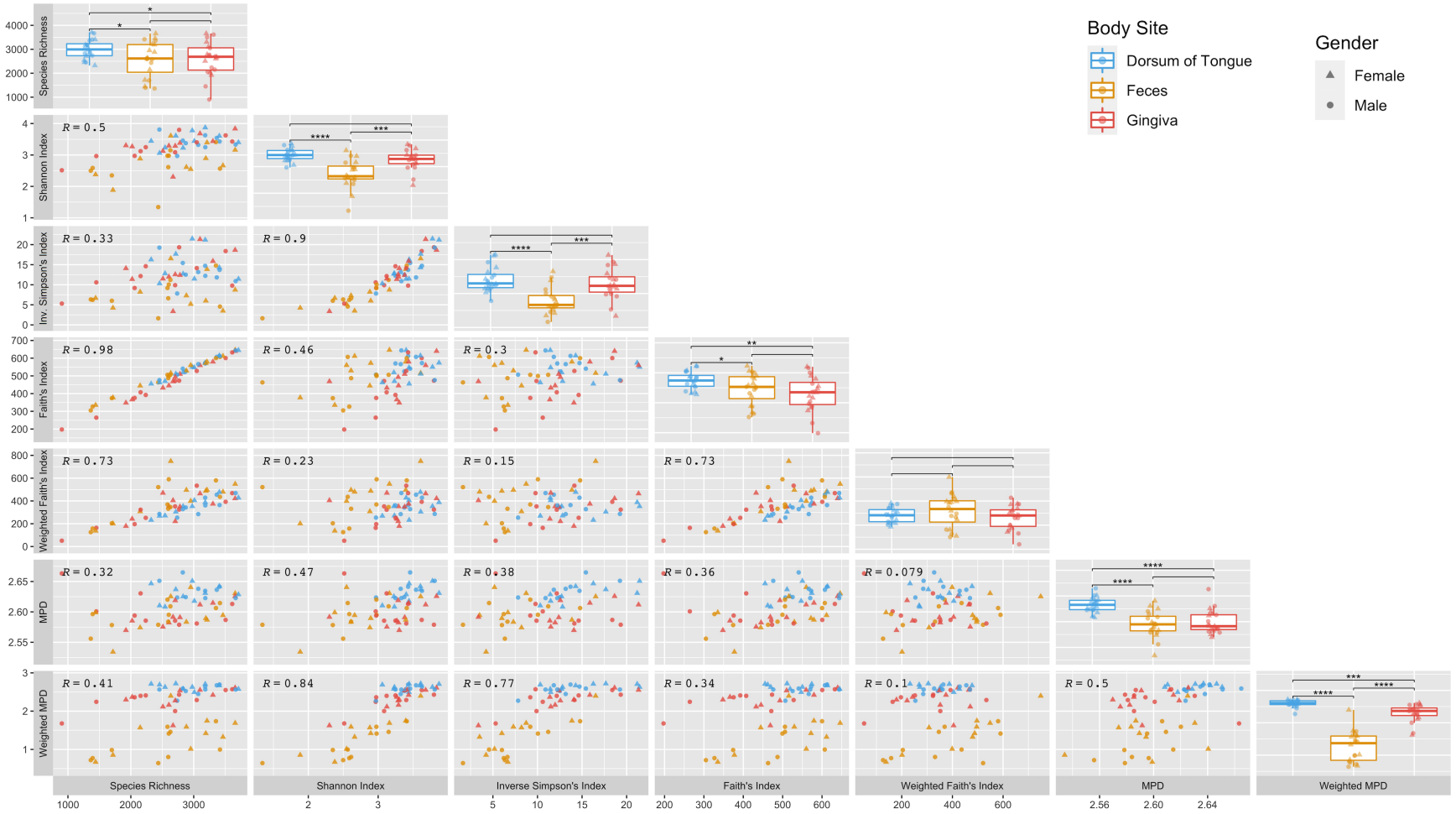

Fig. S3. Comparison of different alpha diversity metrics. Bacterial communities were grouped by different body sites. Along the diagonal are boxplots representing the distribution of each alpha diversity metric. Pairwise *t*-tests were performed to indicate if the metric was significantly different between each pair of body sites (**P* ≤ 0.05, ***P* ≤ 0.01, ****P* ≤ 0.001, *****P* ≤ 0.0001). Below the diagonal are scatterplots showing associations between each pair of these metrics and Pearson's correlation coefficients (*R*) were calculated.

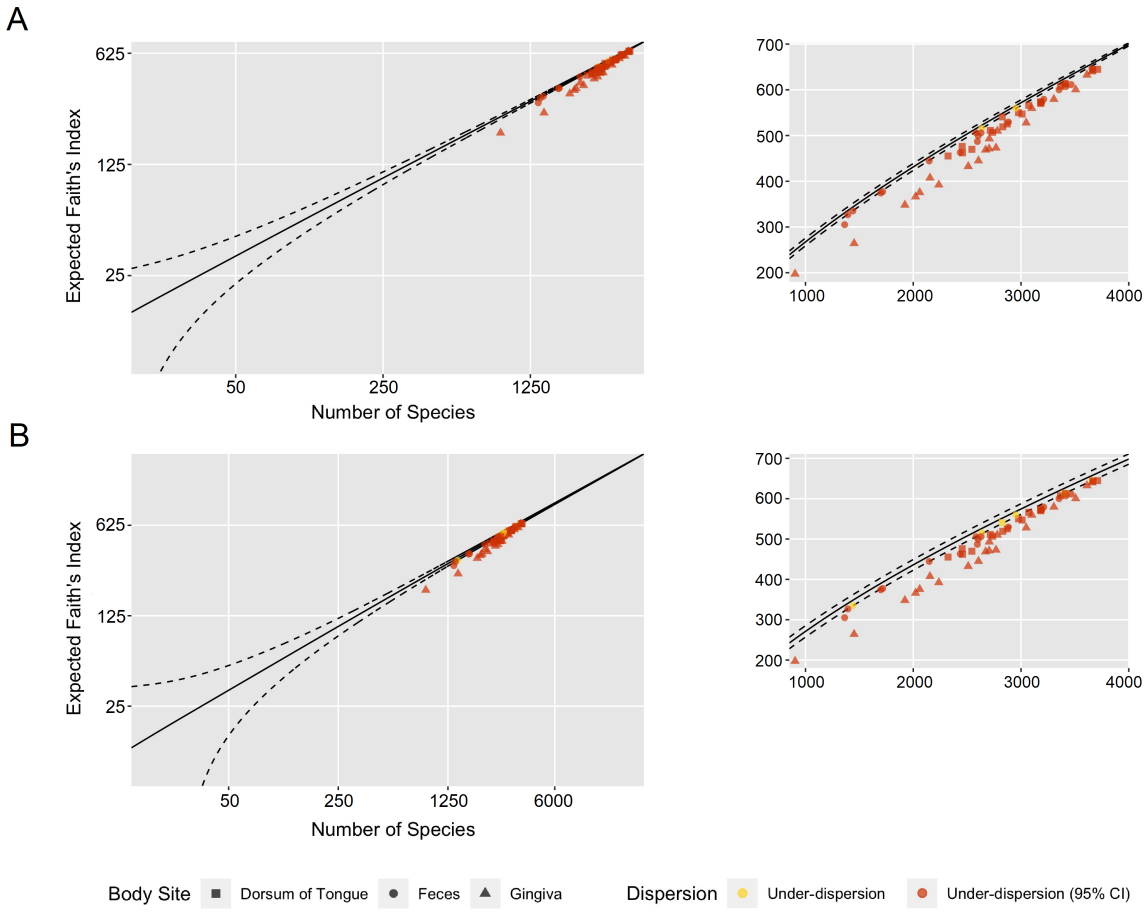

Fig. S4. Comparison of actual Faith's index of local communities with expected value under random sampling from different metacommunities. (A) the metacommunity was defined as the pool of all bacterial species identified across all communities used in this study; (B) the metacommunity was defined as the pool of all species present in our expanded phylogeny. Dashed lines represent the upper and lower bound on the 95% confidence interval, computed using the variance of sampled Faith's index. All samples had Faith's index lower than expected, suggesting phylogenetic under-dispersion, irrespective of the scale of metacommunity.

Fig. S5. Null distribution of phylogeny-based alpha diversity metrics for communities of each body sites. The generation of the null model applied random tip shuffling method to the phylogeny that either (1) only included bacterial species identified in our study (Human metacommunity); or (2) included all bacterial species present in the expanded phylogeny (Entire metacommunity). The area under the curve represents the percentile of the observed value among all values within the null distribution (1 observed value + 999 permuted values) and the two-tailed statistical significance was inferred according to the percentile.

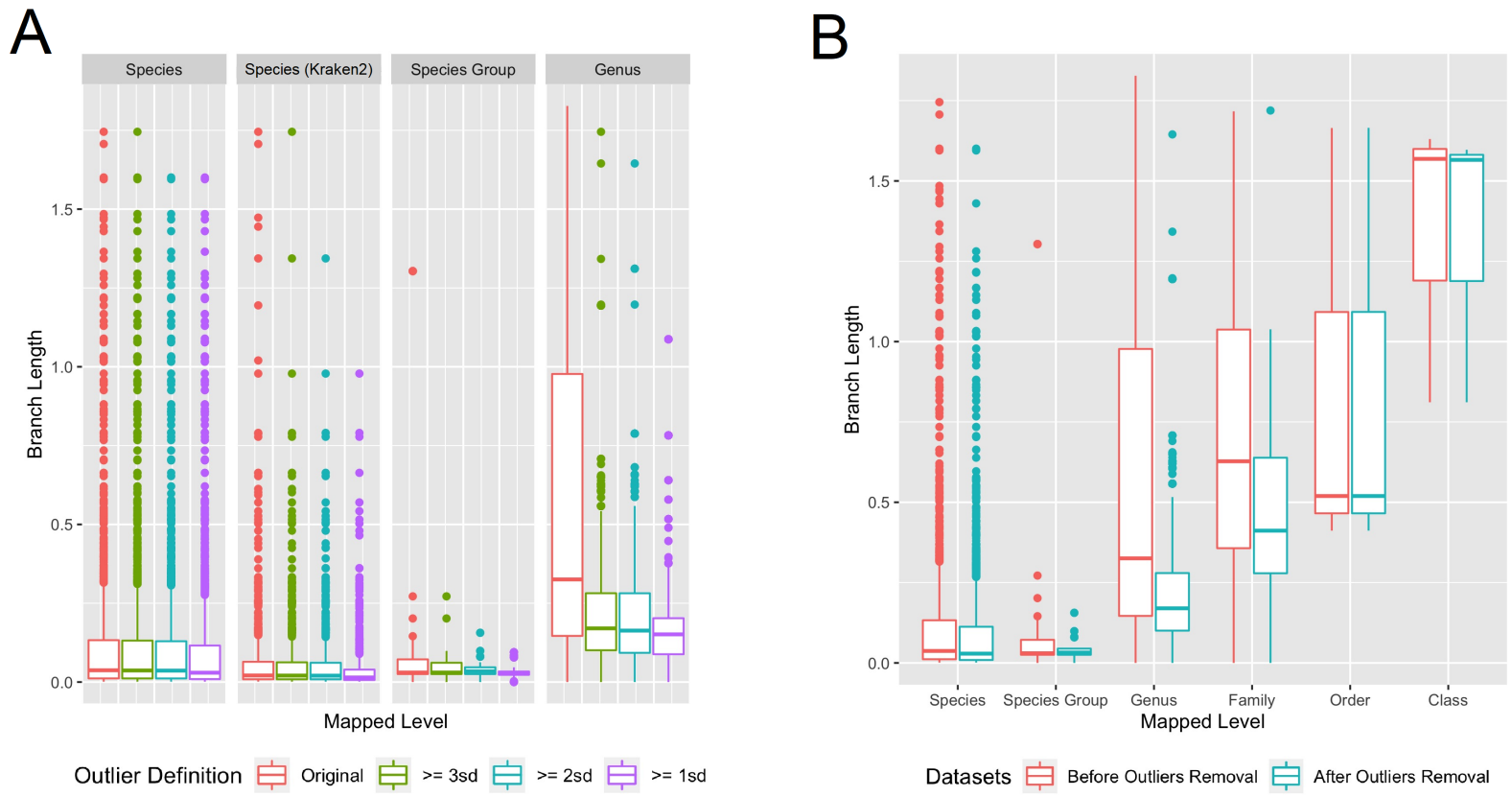

Fig. S6. Evaluation of the effect of removing potential outlier tips. (A) Changes in branch length distribution after removal of potential outlier tips defined using different thresholds. The figure is faceted by the taxonomic rank at which a query species could be mapped to the original GTDB phylogeny. For the first facet, species that were mapped with a branch length of zero (i.e., these species were found in the original phylogeny and were represented by only one tip) were excluded from the plot for better evaluation of the effects of removing outliers. The second facet only contains species present in the Kraken2 database that could be mapped at the species level. (B) Overall comparison of branch length distribution after removal of all outliers. Final thresholds used in this study are: ≥1sd for species level; ≥2sd for species group level; and ≥ 3sd for genus level.

Table S1. Effects of complete inclusion of bacterial species found in the Kraken2 database on phylogeny-based alpha diversity metrics.

Body Site	Phylogeny Used [a]	Weighted Faith's Index	Unweighted Faith's Index	Weighted MPD	Unweighted MPD
Gingiva	Incomplete	221.884 ± 85.431	385.401 ± 94.444	2.288 ± 0.290	2.606 ± 0.020
	Complete	326.897 ± 123.773	459.248 ± 116.684	2.295 ± 0.272	2.599 ± 0.023
	Difference	+ 47.32%	+ 19.16%	+ 0.31%	- 0.27%
Dorsum of Tongue	Incomplete	313.676 ± 64.675	463.536 ± 43.856	2.588 ± 0.113	2.641 ± 0.014
	Complete	352.662 ± 74.402	548.423 ± 58.199	2.597 ± 0.108	2.633 ± 0.014
	Difference	+ 12.43%	+ 18.31%	+ 0.30%	- 0.30%
Feces	Incomplete	319.820 ± 133.025	424.052 ± 81.467	1.209 ± 0.474	2.597 ± 0.027
	Complete	398.299 ± 171.364	492.259 ± 102.845	1.269 ± 0.469	2.595 ± 0.025
	Difference	+ 24.54%	+ 16.08%	+ 4.96%	- 0.08%
Overall	Incomplete	285.127 ± 106.940	424.329 ± 81.633	2.028 ± 0.679	2.615 ± 0.028
	Complete	359.286 ± 130.611	499.977 ± 101.290	2.054 ± 0.653	2.609 ± 0.027
	Difference	+ 26.01%	+ 17.83%	+ 1.28%	- 0.23%

[a] Difference is measured as the percentage of increase (+) or decrease (-) in metrics calculated based on values where all bacterial species present in the Kraken2 outputs were included (complete) compared to where only species shared by both original GTDB phylogeny and the Kraken2 sources were included (incomplete).

Table S2. Metadata of metagenomic sequencing files used in this study.

Subject ID	Body Site	Gender	Visit Number	File Name	Sample ID	File ID	Species Counts[a]
646253328	Feces	F	1	SRS1055043	596fc2de57601ec08a01fdee59051f2a	596fc2de57601ec08a01fdee59e87fbc	14201991
	Dorsum of Tongue			SRS1055046	596fc2de57601ec08a01fdee59053063	596fc2de57601ec08a01fdee59e890b0	11786671
	Gingiva			SRS1055048	596fc2de57601ec08a01fdee59053b18	596fc2de57601ec08a01fdee59e89be3	8363594
195044680	Feces	F	2	SRS1055049	596fc2de57601ec08a01fdee59068363	596fc2de57601ec08a01fdee59e8e118	14914676
	Dorsum of Tongue			SRS893324	596fc2de57601ec08a01fdee59065f61	596fc2de57601ec08a01fdee59e8bb7a	8489344
	Gingiva			SRS893333	596fc2de57601ec08a01fdee5906777e	596fc2de57601ec08a01fdee59e8d79a	17227855
765216093	Feces	M	2	SRS104036	596fc2de57601ec08a01fdee5907d2a4	596fc2de57601ec08a01fdee59e96352	12510588
	Dorsum of Tongue			SRS104039	596fc2de57601ec08a01fdee5907d7d7	596fc2de57601ec08a01fdee59e967fb	10656592
	Gingiva			SRS104045	596fc2de57601ec08a01fdee5907ed4b	596fc2de57601ec08a01fdee59e977a0	10554983
765539774	Feces	F	2	SRS104084	596fc2de57601ec08a01fdee59086a40	596fc2de57601ec08a01fdee59e9847b	19871908
	Dorsum of Tongue			SRS104087	596fc2de57601ec08a01fdee59086c5a	596fc2de57601ec08a01fdee59e9880d	9768077
	Gingiva			SRS104093	596fc2de57601ec08a01fdee59087b8a	596fc2de57601ec08a01fdee59e998a1	13142159
103092734	Feces	M	2	SRS104327	596fc2de57601ec08a01fdee590ae134	596fc2de57601ec08a01fdee59ea713b	17691361
	Dorsum of Tongue			SRS104314	596fc2de57601ec08a01fdee590ac020	596fc2de57601ec08a01fdee59ea5450	12587054
	Gingiva			SRS104320	596fc2de57601ec08a01fdee590ad2a9	596fc2de57601ec08a01fdee59ea6bc5	19450152
765377934	Feces	F	2	SRS104636	596fc2de57601ec08a01fdee590d53af	596fc2de57601ec08a01fdee59eb509f	17912802
	Dorsum of Tongue			SRS104647	596fc2de57601ec08a01fdee590d7d24	596fc2de57601ec08a01fdee59eb6d00	7529019
	Gingiva			SRS104653	596fc2de57601ec08a01fdee590d8e9b	596fc2de57601ec08a01fdee59eb74f9	13230162
765317243	Feces	F	2	SRS104693	596fc2de57601ec08a01fdee590df967	596fc2de57601ec08a01fdee59eb7fe5	13492648
	Dorsum of Tongue			SRS104704	596fc2de57601ec08a01fdee590e0645	596fc2de57601ec08a01fdee59eb92f3	11119296
	Gingiva			SRS104711	596fc2de57601ec08a01fdee590e1027	596fc2de57601ec08a01fdee59ebacfd	19313347
316129862	Feces	F	1	SRS1055056	596fc2de57601ec08a01fdee590e77b5	596fc2de57601ec08a01fdee59ebb402	5533226
	Dorsum of Tongue			SRS1055059	596fc2de57601ec08a01fdee590e87fe	596fc2de57601ec08a01fdee59ebd7ce	10517690
	Gingiva			SRS893325	596fc2de57601ec08a01fdee590e928f	596fc2de57601ec08a01fdee59ebdfa8	8058291

	Feces			SRS104975	596fc2de57601ec08a01fdee59102718	596fc2de57601ec08a01fdee59ec792b	17902835
765519544	Dorsum of Tongue	M	2	SRS104962	596fc2de57601ec08a01fdee590ffb91	596fc2de57601ec08a01fdee59ec4d1c	16660221
	Gingiva			SRS104968	596fc2de57601ec08a01fdee59100bdb	596fc2de57601ec08a01fdee59ec6786	8674705
	Feces			SRS142712	596fc2de57601ec08a01fdee5914cf11	596fc2de57601ec08a01fdee59ede00d	42364374
338793263	Dorsum of Tongue	F	2	SRS142680	596fc2de57601ec08a01fdee59144e90	596fc2de57601ec08a01fdee59edd27e	23268623
	Gingiva			SRS142664	596fc2de57601ec08a01fdee59143c4c	596fc2de57601ec08a01fdee59edcda4	15476951
	Feces			SRS143085	596fc2de57601ec08a01fdee59184991	596fc2de57601ec08a01fdee59ef04ea	22089117
355657046	Dorsum of Tongue	M	2	SRS143088	596fc2de57601ec08a01fdee59185494	596fc2de57601ec08a01fdee59ef06cc	22527320
	Gingiva			SRS143094	596fc2de57601ec08a01fdee591867b3	596fc2de57601ec08a01fdee59ef1b4f	16187838
	Feces			SRS893378	596fc2de57601ec08a01fdee5905f670	596fc2de57601ec08a01fdee59e8aeb8	18085961
366487741	Dorsum of Tongue	F	1	SRS893363	596fc2de57601ec08a01fdee5905cd25	596fc2de57601ec08a01fdee59e89fa4	8248797
	Gingiva			SRS893385	596fc2de57601ec08a01fdee5905dc94	596fc2de57601ec08a01fdee59e8a04f	21054429
	Feces			SRS893300	596fc2de57601ec08a01fdee5943fe4c	596fc2de57601ec08a01fdee59fb25e4	10165834
596625983	Dorsum of Tongue	M	3	SRS893321	596fc2de57601ec08a01fdee59440010	596fc2de57601ec08a01fdee59fb2c14	16885339
	Gingiva			SRS893328	596fc2de57601ec08a01fdee59440de8	596fc2de57601ec08a01fdee59fb372d	12414829
	Feces			SRS149181	596fc2de57601ec08a01fdee59492959	596fc2de57601ec08a01fdee59fc54f4	16453878
115629832	Dorsum of Tongue	M	3	SRS149184	596fc2de57601ec08a01fdee59492fe3	596fc2de57601ec08a01fdee59fc5a08	19880727
	Gingiva			SRS149190	596fc2de57601ec08a01fdee59493eea	596fc2de57601ec08a01fdee59fc7558	11196351
	Feces			SRS149244	596fc2de57601ec08a01fdee59499148	596fc2de57601ec08a01fdee59fc9e78	20104343
901775393	Dorsum of Tongue	M	2	SRS149231	596fc2de57601ec08a01fdee5949863f	596fc2de57601ec08a01fdee59fc8721	11646608
	Gingiva			SRS149237	596fc2de57601ec08a01fdee594987c9	596fc2de57601ec08a01fdee59fc9182	10855181
	Feces			SRS147377	596fc2de57601ec08a01fdee5938c4ec	596fc2de57601ec08a01fdee59f818a7	22129770
938202701	Dorsum of Tongue	M	2	SRS147380	596fc2de57601ec08a01fdee5938cd6c	596fc2de57601ec08a01fdee59f818bc	22776414
	Gingiva			SRS147386	596fc2de57601ec08a01fdee5938d364	596fc2de57601ec08a01fdee59f82339	22151682
486505039	Feces	M	2	SRS893288	596fc2de57601ec08a01fdee593d01dc	596fc2de57601ec08a01fdee59f90f8c	16804418
	Dorsum of Tongue			SRS893274	596fc2de57601ec08a01fdee593ce195	596fc2de57601ec08a01fdee59f8f78c	5830044

	Gingiva			SRS893282	596fc2de57601ec08a01fdee593cf4f6	596fc2de57601ec08a01fdee59f9038d	16325440
	Feces			SRS148091	596fc2de57601ec08a01fdee593f76a6	596fc2de57601ec08a01fdee59f98f83	12587705
516889361	Dorsum of Tongue	M	2	SRS148094	596fc2de57601ec08a01fdee593f83bc	596fc2de57601ec08a01fdee59f99126	14836403
	Gingiva			SRS148100	596fc2de57601ec08a01fdee593f90c1	596fc2de57601ec08a01fdee59f998ee	10664665
	Feces			SRS893285	596fc2de57601ec08a01fdee59099743	596fc2de57601ec08a01fdee59e9d5d0	11737670
188816475	Dorsum of Tongue	F	2	SRS893289	596fc2de57601ec08a01fdee5909aa37	596fc2de57601ec08a01fdee59e9ed57	8359822
	Gingiva			SRS893299	596fc2de57601ec08a01fdee5909b9a4	596fc2de57601ec08a01fdee59e9f25b	12666112
	Feces			SRS104912	596fc2de57601ec08a01fdee590f6eb8	596fc2de57601ec08a01fdee59ec2567	11485887
765661155	Dorsum of Tongue	F	2	SRS104915	596fc2de57601ec08a01fdee590f745e	596fc2de57601ec08a01fdee59ec319a	10656324
	Gingiva			SRS104921	596fc2de57601ec08a01fdee590f78a3	596fc2de57601ec08a01fdee59ec3be8	20619108

[a] Estimated using Kraken2 and Bracken, at least 1,000,000 total classified species counts was required for each metagenomic file for further analysis.

Table S3. Basic phylogenetic information retained after removal of potential outlier tips.

Defined Potential Outliers[a]	# Tips[b]	# Species[b]	# Species Group[b]	# Genus	# Family	# Order	# Class	# Phylum	# Super-kingdom
Original	45554	26811	62	3269	576	236	107	156	1
Species: 3sd	45543 (99.98%)	26811	62	3269	576	236	107	156	1
Species: 2sd	45510 (99.90%)	26811	62	3269	576	236	107	156	1
Species: 1sd	45204 (99.23%)	26811	62	3269	576	236	107	156	1
Species Group: 3sd	45545 (99.98%)	26809 (99.99%)	62	3269	576	236	107	156	1
Species Group: 2sd	45532 (99.95%)	26808 (99.99%)	62	3269	576	236	107	156	1
Species Group: 1sd	45509 (99.90%)	26794 (99.94%)	62	3269	576	236	107	156	1
Genus: 3sd	45354 (99.56%)	26667 (99.46%)	62	3269	576	236	107	156	1
Genus: 2sd	45122 (99.05%)	26488 (98.80%)	62	3269	576	236	107	156	1
Genus: 1sd	44237 (97.11%)	25751 (96.05%)	61 (98.39%)	3269	576	236	107	156	1
Final[c]	45010 (98.81%)	26662 (99.44%)	62	3269	576	236	107	156	1

[a] Only specified outliers were removed. For example, for “species: 3sd”, we only considered query species that could be mapped at the species level and removed tips within their corresponding reference groups that had an average distance of more than three standard deviation plus the mean to the remaining group member tips.
[b] Percentages were given by comparing the changed values to the corresponding values in the original phylogeny.
[c] Final thresholds used in this study are: ≥1sd for species level; ≥2sd for species group level; and ≥ 3sd for genus level.

Table S4. Summary of mapped level and average branch length of all query species.

Defined Potential Outliers[a]	Species		Species Group		Genus		Family		Order		Class		Overall	
	n	avg.	n	avg.	n	avg.	n	avg.	n	avg.	n	avg.	n	avg.
Original	20809	0.0090	59	0.1808	1299	0.5445	13	0.7104	3	0.8653	3	1.3370	22186	0.0415
Species: 3sd	20809	0.0086	59	0.1808	1299	0.4978	13	0.7105	3	0.8653	3	1.3368	22186	0.0385
Species: 2sd	20809	0.0085	59	0.1776	1299	0.4921	13	0.7105	3	0.8652	3	1.3368	22186	0.0379
Species: 1sd	20809	0.0079	59	0.0468	1299	0.4341	13	0.5333	3	0.8652	3	1.3359	22186	0.0336
Species Group: 3sd	20807	0.0087	61	0.0489	1299	0.4673	13	0.5329	3	0.8653	3	1.3368	22186	0.0362
Species Group: 2sd	20806	0.0086	62	0.0427	1299	0.4670	13	0.5331	3	0.8653	3	1.3368	22186	0.0362
Species Group: 1sd	20792	0.0086	76	0.0337	1299	0.4670	13	0.5331	3	0.8653	3	1.3368	22186	0.0362
Genus: 3sd	20665	0.0079	59	0.0513	1443	0.2321	13	0.5336	3	0.8656	3	1.3255	22186	0.0232
Genus: 2sd	20486	0.0076	60	0.0537	1621	0.2010	13	0.5339	3	0.8659	3	1.3257	22186	0.0225
Genus: 1sd	19749	0.0075	62	0.04960	2356	0.1712	13	0.4540	3	0.8654	3	1.3260	22186	0.0256
Final[b]	20660	0.0072	64	0.0399	1443	0.2302	13	0.5340	3	0.8655	3	1.3246	22186	0.0224

[a] Only specified outliers were removed. For example, for “species: 3sd”, we only considered query species that could be mapped at the species level and removed tips within their corresponding reference groups that had an average distance of more than three standard deviation plus the mean to the remaining group member tips.

[b] Final thresholds used in this study are: ≥1sd for species level; ≥2sd for species group level; and ≥ 3sd for genus level.

Table S5. Summary of mapped level and average branch length of query species present in the Kraken2 source.

Defined Potential Outliers[a]	Species		Species Group		Genus		Family	
	n	avg.	n	avg.	n	avg.	n	avg.
Original	5038	0.0115	59	0.1808	1299	0.5445	13	0.7105
Species: 3sd	5038	0.0100	59	0.1808	1299	0.4978	13	0.7105
Species: 2sd	5038	0.0092	59	0.1776	1299	0.4921	13	0.7105
Species: 1sd	5038	0.0069	59	0.0468	1299	0.4341	13	0.5333
Species Group: 3sd	5038	0.0101	59	0.0478	1299	0.4673	13	0.5329
Species Group: 2sd	5037	0.0099	60	0.0418	1299	0.4670	13	0.5331
Species Group: 1sd	5032	0.0099	65	0.0329	1299	0.4670	13	0.5331
Genus: 3sd	5017	0.0088	59	0.0513	1320	0.2185	13	0.5336
Genus: 2sd	4974	0.0083	59	0.0513	1363	0.1922	13	0.5339
Genus: 1sd	4808	0.0083	59	0.0505	1529	0.1599	13	0.4540
Final[b]	5014	0.0063	62	0.0389	1320	0.2165	13	0.5340

[a] Only specified outliers were removed. For example, for “species: 3sd”, we only considered query species that could be mapped at the species level and removed tips within their corresponding reference groups that had an average distance of more than three standard deviation plus the mean to the remaining group member tips.
[b] Final thresholds used in this study are: ≥1sd for species level; ≥2sd for species group level; and ≥ 3sd for genus level.

Supplementary Information References

1. P. J. Cock et al., Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 25, 1422-1423 (2009).
2. E. Paradis, J. Claude, K. Strimmer, APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics 20, 289-290 (2004).
3. K. P. Schliep, phangorn: phylogenetic analysis in R. Bioinformatics 27, 592-593 (2011).