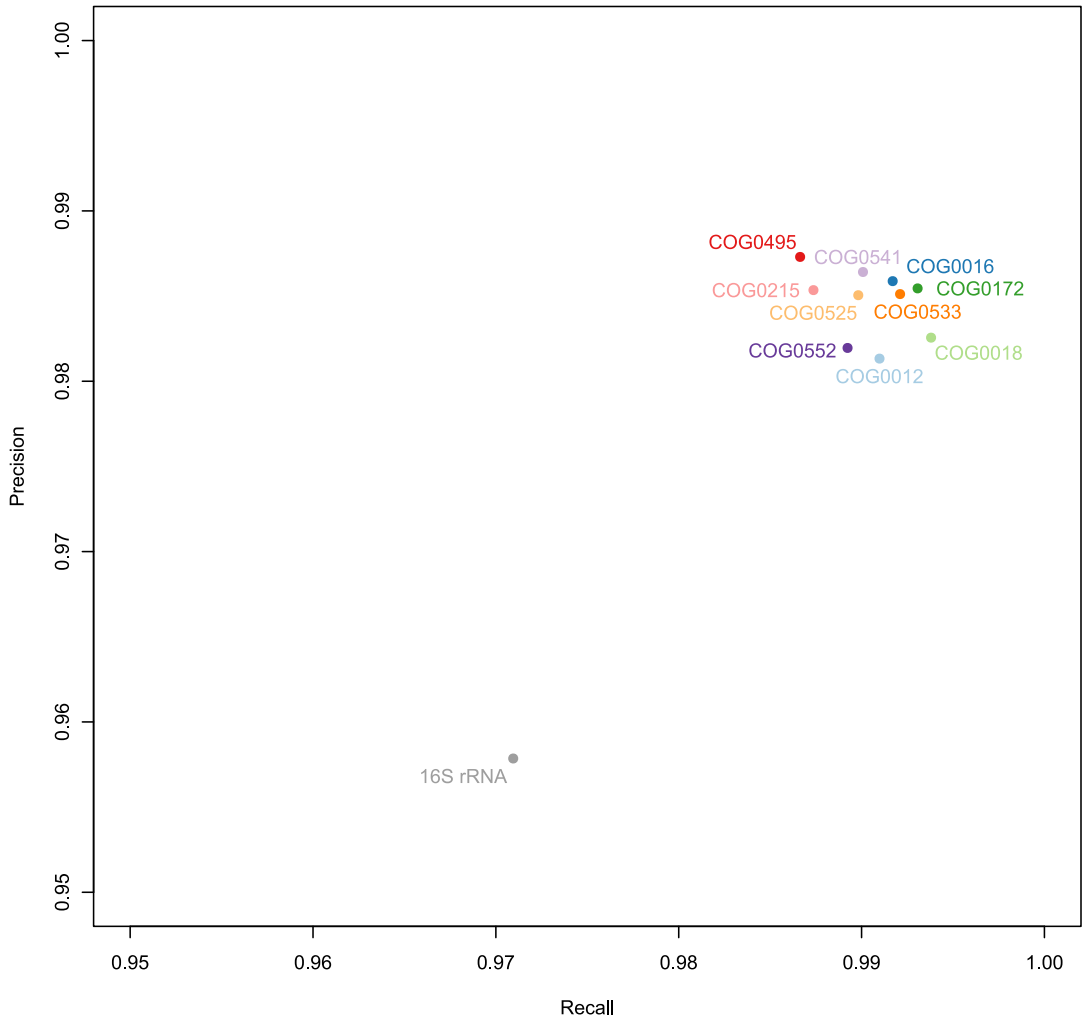
Metagenomic species profiling using universal phylogenetic marker genes

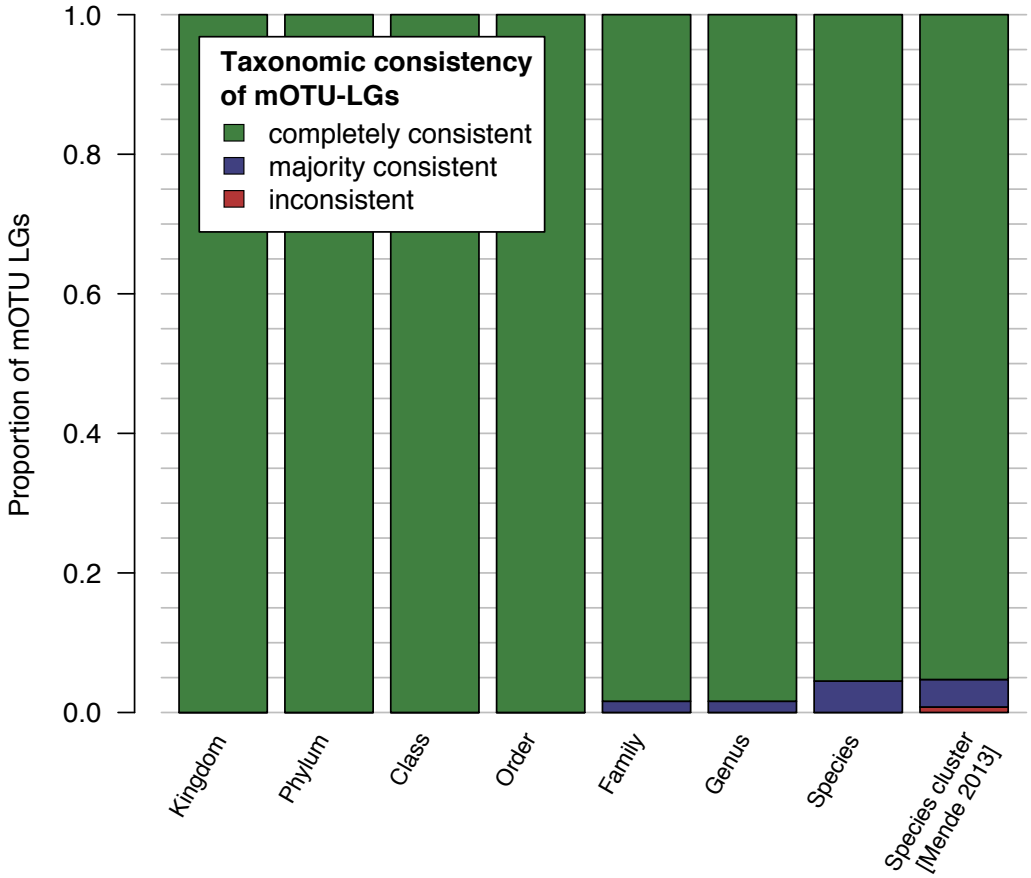
Shinichi Sunagawa, Daniel R. Mende, Georg Zeller, Fernando Izquierdo-Carrasco, Simon A. Berger, Jens Roat Kultima, Luis Pedro Coelho, Manimozhiyan Arumugam, Julien Tap, Henrik Bjørn Nielsen, Simon Rasmussen, Søren Brunak, Oluf Pedersen, Francisco Guarner, Willem M. de Vos, Jun Wang, Junhua Li, Joël Doré, S. Dusko Ehrlich, Alexandros Stamatakis, Peer Bork

Supplementary Item	Title
Supplementary Figure 1	Accuracy of species-level mOTU clustering.
Supplementary Figure 2	Linkage of mOTUs of common origin.
Supplementary Figure 3	False-discovery rate over the process of the mOTU linkage procedure.
Supplementary Figure 4	Abundance consistency within mOTU linkage groups.
Supplementary Figure 5	GC-content consistency within mOTU linkage groups.
Supplementary Figure 6	Abundance of mOTU linkage groups across samples.
Supplementary Figure 7	Ranked abundance and prevalence of mOTU-LGs
Supplementary Figure 8	Performance of relative species abundance estimations.
Supplementary Table 4	Sequence identity cutoffs used for clustering marker genes into mOTUs and ambiguous alignment rates for mOTUs of each marker gene.
Supplementary Table 5	Phylogenomic representation of mOTUs.
Supplementary Table 7	Summary of using different distance metrics for community similarity analysis.

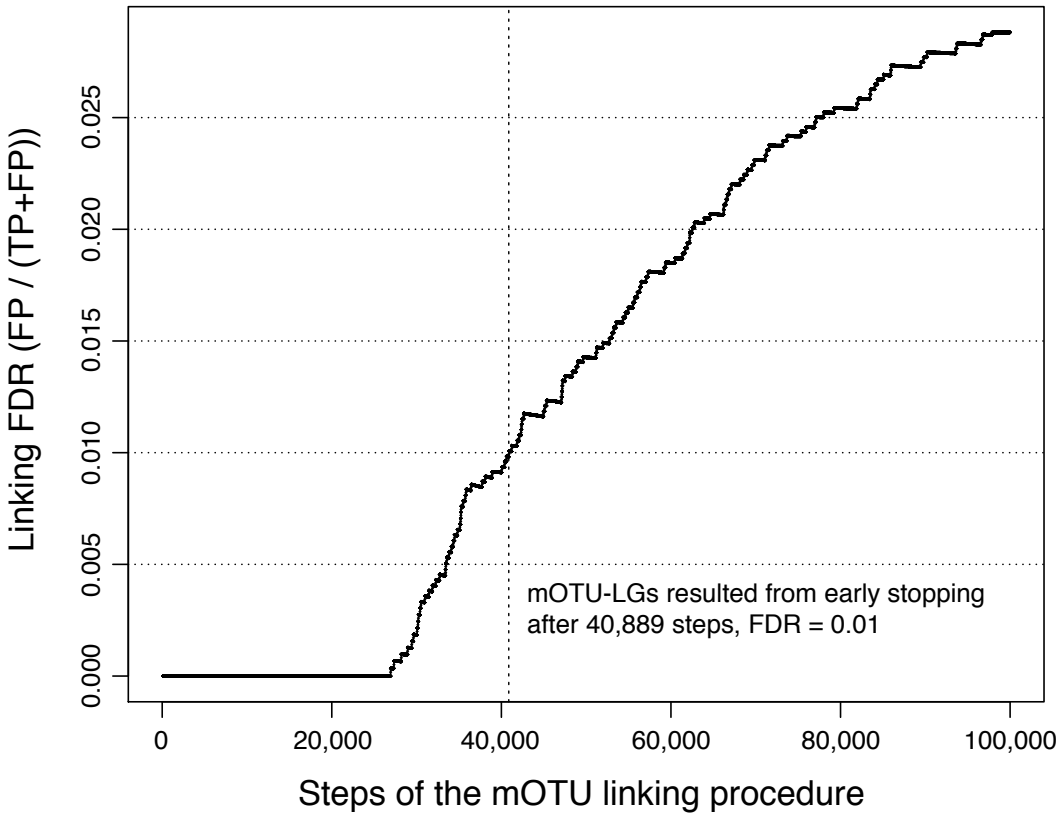
Nature Methods: doi:10.1038/nmeth.2693



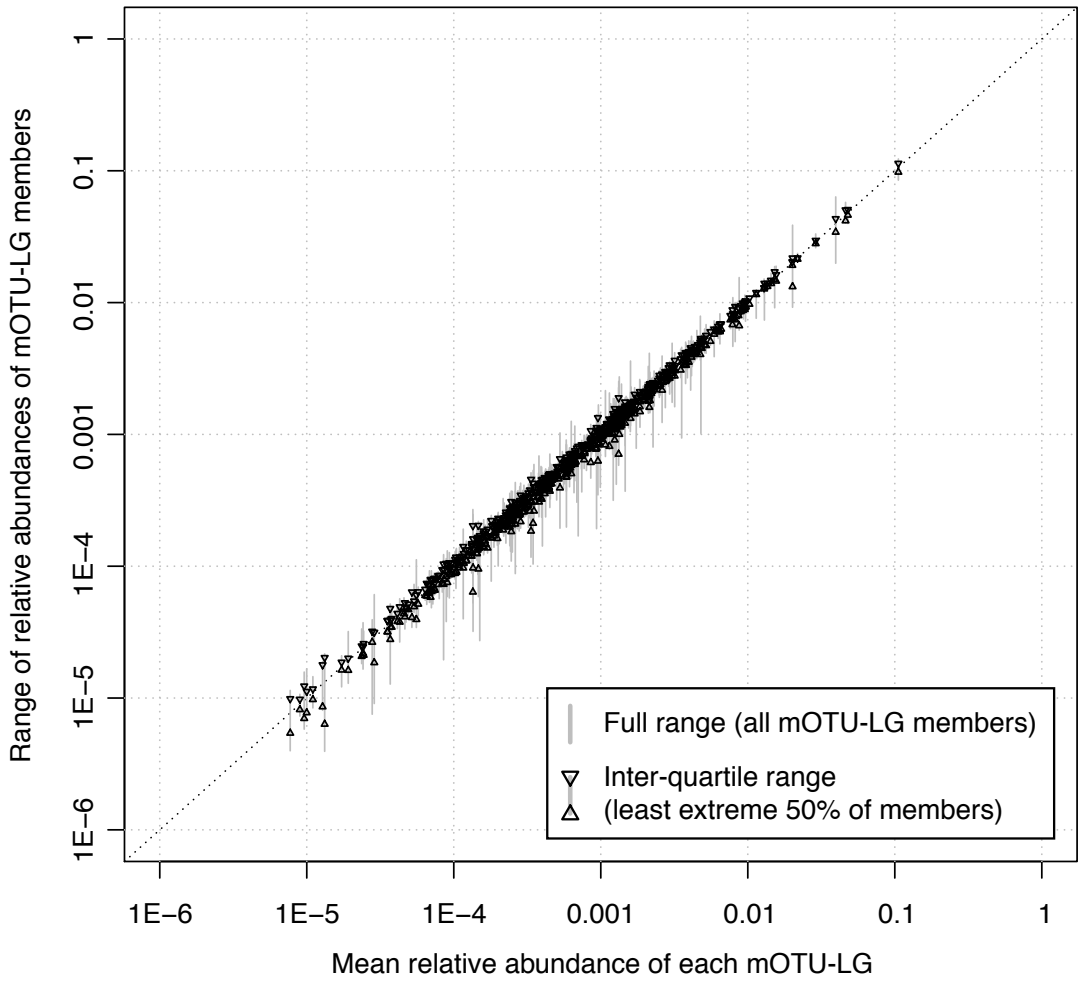
Supplementary Figure 1: Accuracy of species-level mOTU clustering. Clustering accuracy was assessed by testing whether mOTUs with at least two MGs or 16S rRNA genes that originated from a type strain reference genome were consistent regarding the taxonomic annotation of their members at the species level according to the NCBI taxonomy. According to this information, false discovery rates (FDRs) and recall values were calculated for all mOTUs. Precision (1 - FDR) and recall can take values between 0 and 1, with high values indicating a good agreement of mOTU cluster members with the NCBI taxonomy.



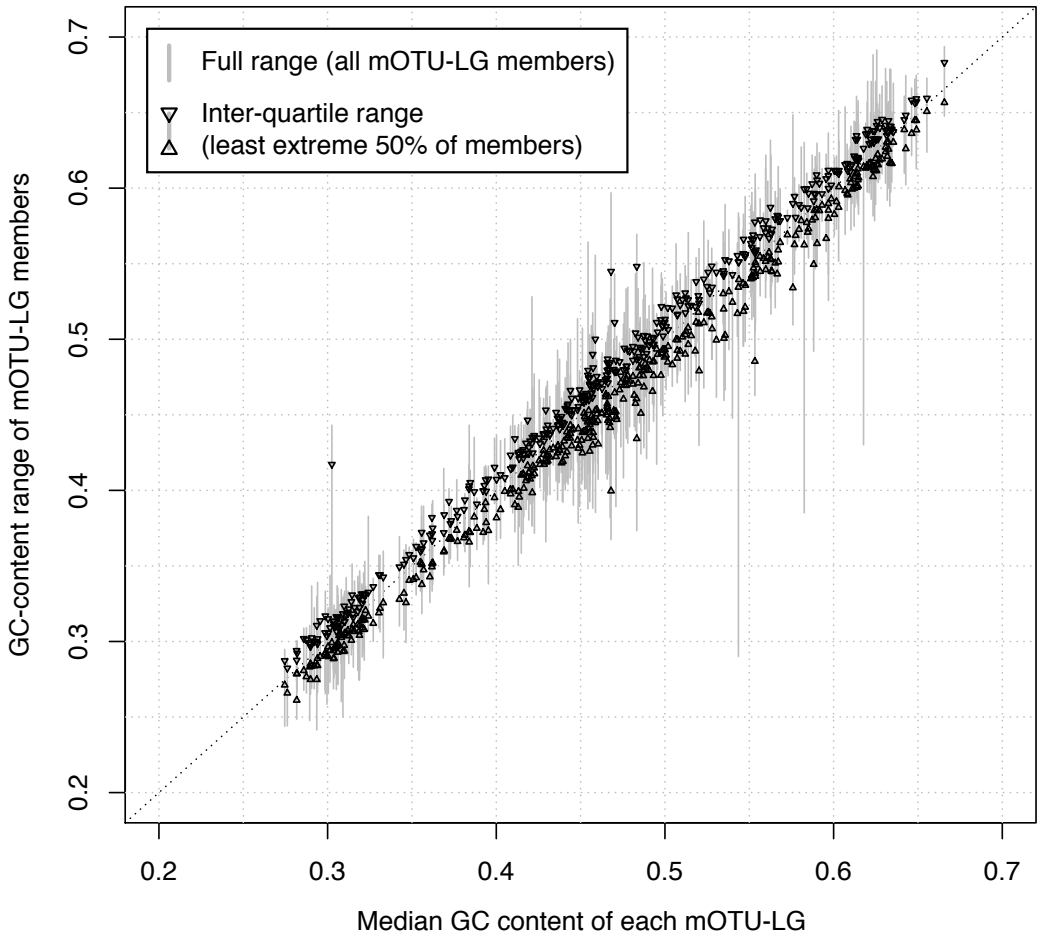
Supplementary Figure 2: Linkage of mOTUs of common origin. Accuracy of linking mOTUs to mOTU linkage groups of common origin was assessed by calculating the consistency of taxonomic information available for mOTU linkage groups (mOTU-LGs) whose members originated from annotated $mOTU_{RefMeta}$. mOTU-LGs in which more than 50% of individual mOTU taxonomic annotations agreed were labelled as "majority consistent" (see legend). The last column shows results for species-levels clusters described in Mende et al., 2013.



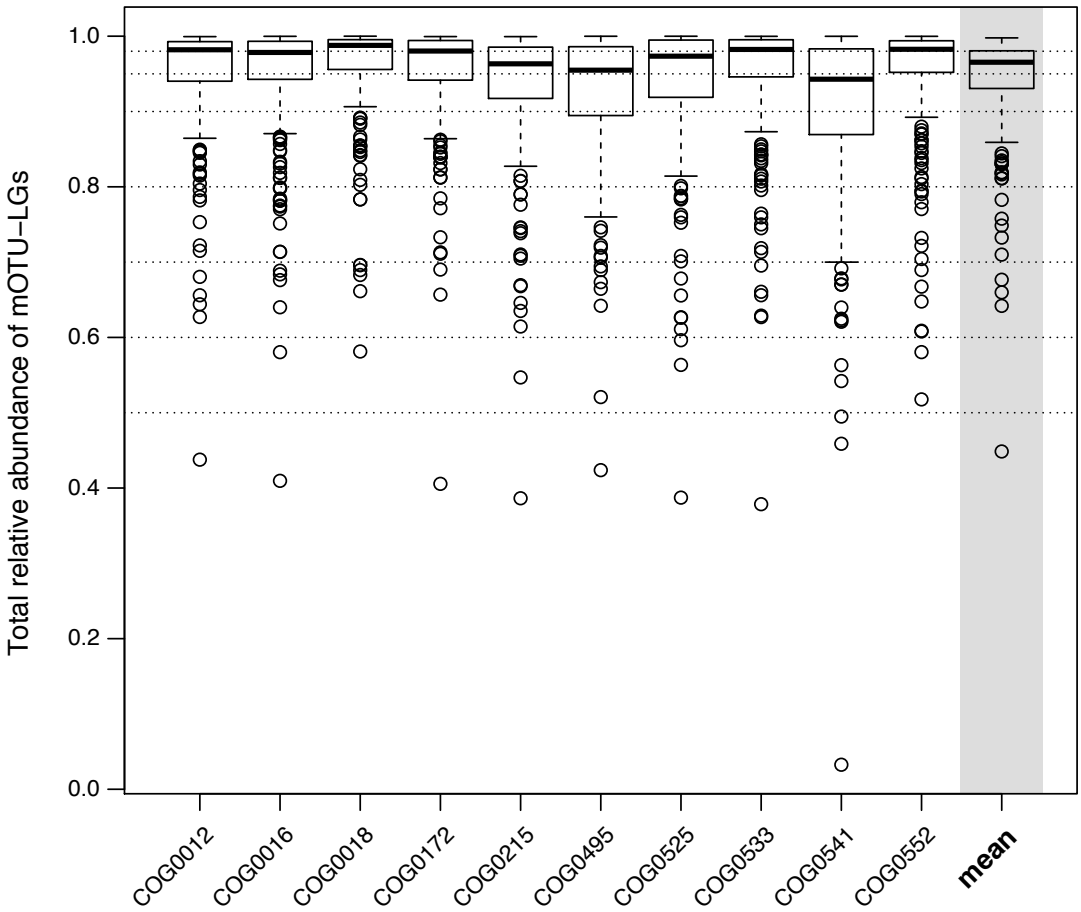
Supplementary Figure 3: False-discovery rate over the process of the mOTU linkage procedure. The false discovery rate (FDR) is defined as the proportion of false positives (FP) among all positive predictions, that is the sum of FP and true positives (TP). Whenever two mOTUs that are taxonomically annotated were linked by the algorithm, it was evaluated whether annotations agree (TP) or not (FP) and the FDR was updated accordingly. Before the FDR exceeded 0.01 (vertical line) after 40,889 linking steps, mOTU linkage was terminated.



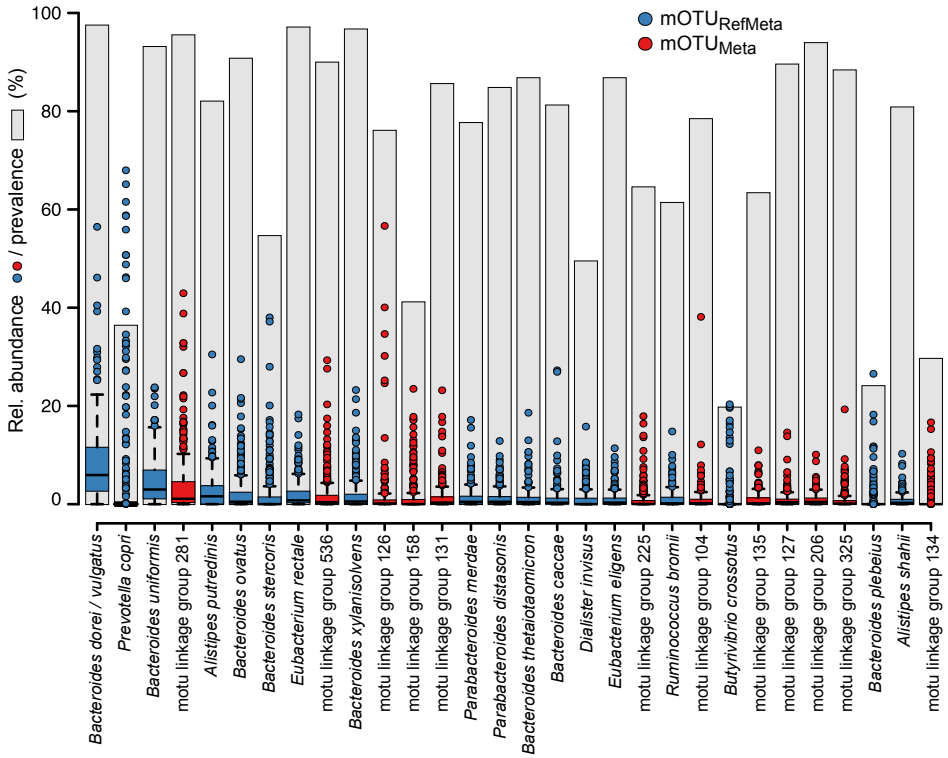
Supplementary Figure 4: Abundance consistency within mOTU linkage groups. Relative abundance estimates of individual mOTUs plotted as a function of the relative abundance of the mOTU linkage group they belong to (estimated by their median). The full range across all linkage group members as well as the inter-quartile range are shown for each mOTU linkage group (see legend). Note that in most cases the abundance estimates of individual cluster members are very close to the cluster estimate (median) indicating robustness of these estimates. Note further that mOTU abundances were not used for cluster construction, only their correlation across samples (Spearman correlation, which is transformation-invariant).



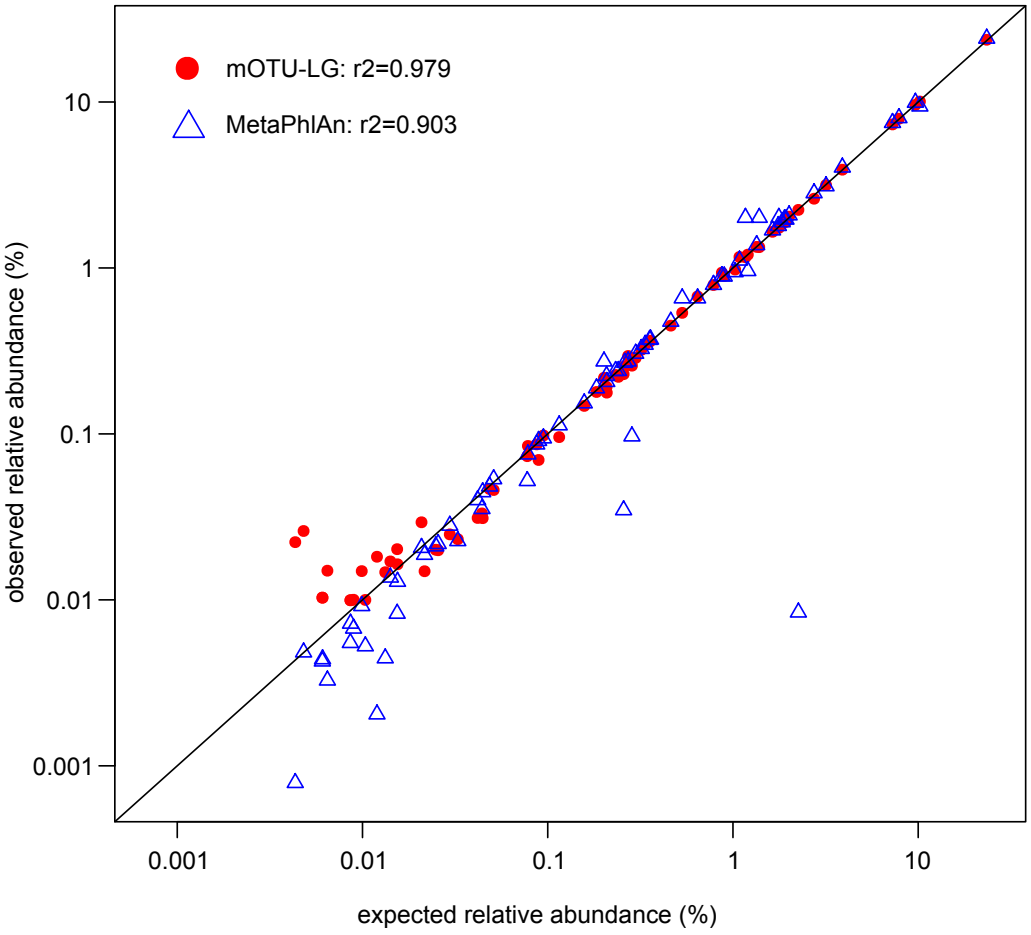
Supplementary Figure 5: GC-content consistency within mOTU linkage groups. GC-content of individual mOTUs plotted as a function of the mean GC content of the mOTU linkage group that contains them. Shown are full ranges as well as inter-quartile ranges (see legend). Homogeneity of GC content within a mOTU linkage group suggests that the contained genes likely belong to the same species, because within a genome and species, GC content generally varies much less than between different species.



Supplementary Figure 6: Abundance of mOTU linkage groups across samples. Total relative abundance of mOTUs belonging to mOTU linkage groups summarized across samples by boxplots (thick line corresponds to the median, boxes delineate the inter-quartile range (IQR) with whiskers extending to 1.5-fold IQRs, and circles indicate outliers). First ten columns show total relative abundance separately for each marker gene, while the rightmost column summarizes the distribution of per-sample averages.



Supplementary Figure 7: Ranked abundance and prevalence of mOTU-LGs. Relative abundances of the 30 most dominant (ranked by mean sample abundance), annotated and novel mOTU linkage groups (Supplementary Table 6) that represent prokaryotic species clusters are shown as blue and red boxplots, respectively. Grey bar plots show the prevalence of these species clusters in 252 human gut metagenomic samples.Performance of relative species abundance estimations.



Supplementary Figure 8: Performance of relative species abundance estimations. Observed relative species abundances using mOTU-LGs and MetaPhlAn (species-level) were benchmarked against expected values based on a simulated human faecal metagenome. Pearson correlations are shown for log-transformed relative abundances.

Supplementary Table 4: Sequence identity cutoffs used for clustering marker genes into mOTUs and ambiguous alignment rates for mOTUs of each marker gene. For each of 40 universal single-copy marker genes (MGs), sequence identity cutoffs for clustering into mOTUs are shown. Based on mOTU abundance profiles of 252 metagenomic samples (Supplementary Table 2) we calculated for each MG the fraction of reads that were mapped to more than one mOTU (ambiguous alignment rates). MGs were selected for inclusion in this study if the ambiguous alignment rate was below 7% and the false discovery rate (Supplementary Table 3) below 10%. FDR - false discovery rate.

Summary				
COG	**Mean concatenated length in 3496 genomes**	**Mean identification FDR (%)**	**Mean ambiguous alignment rate (median of 252 samples)**	
10 selected MGs	15529	1.42	3.5%	
Selected genes				
Marker Gene	**Mean length in 3496 genomes**	**Calibrated clustering cutoffs (%)**	**Identification FDR (%)**	**Ambiguous alignment rate (median of 252 samples)**
COG0012	1099	94.8	1.06	1.0%
COG0016	1058	95.8	0.14	3.4%
COG0018	1721	94.2	3.22	1.5%
COG0172	1285	94.4	3.79	3.6%
COG0215	1415	95.4	2.74	6.4%
COG0495	2571	96.4	1.87	5.5%
COG0525	2722	95.3	0.72	5.2%
COG0533	1054	93.1	0.43	0.9%
COG0541	1415	96.1	0.12	5.4%
COG0552	1189	94.5	0.12	2.5%
Excluded genes				
Marker Gene	**Mean length in 3496 genomes**	**Calibrated clustering cutoffs (%)**	**Identification FDR (%)**	**Ambiguous alignment rate (median of 252 samples)**
COG0048	391	98.4	0.18	30.1%
COG0049	476	98.7	0.15	18.5%
COG0052	782	97.2	0.20	10.1%
COG0080	433	98.6	0.43	26.8%
COG0081	696	98	0.23	15.7%
COG0085	3828	97	48.00	6.1%
COG0087	662	99	0.23	31.8%
COG0088	639	99	0.20	41.4%
COG0090	825	98.8	0.17	29.6%
COG0091	370	99.2	0.58	42.3%
COG0092	712	99.2	0.23	47.0%
COG0093	370	99	0.29	42.8%
COG0094	547	99	0.34	42.0%
COG0096	397	98.6	0.09	29.6%
COG0097	538	98.4	0.14	23.3%
COG0098	538	98.7	0.14	33.2%
COG0099	372	98.9	0.12	27.8%
COG0100	393	99	0.12	46.5%
COG0102	444	99.1	1.63	46.0%
COG0103	415	98.4	0.32	45.2%
COG0124	1306	94.5	9.67	3.0%
COG0184	278	98.2	0.44	26.1%
COG0185	282	99.3	0.20	51.4%
COG0186	268	99.3	0.35	58.4%
COG0197	425	99.3	0.12	44.6%
COG0200	446	98.4	0.14	25.3%
COG0201	1322	97.2	1.96	7.9%
COG0202	977	98.4	1.67	10.6%
COG0256	369	99	0.12	43.3%
COG0522	610	98.6	2.55	26.5%
16S rDNA	1452	98.8	NA	41.09%

Supplementary Table 5: Phylogenomic representation of mOTUs. mOTUs were defined as mOTU(RefMeta) or mOTU(Meta) depending on whether genes within a mOTU cluster were identical or similar (below the MG-specific sequence identity cutoff shown in Supplementary Table 4) to a reference marker gene sequence or only found in a metagenome, respectively. The number of mOTUs and the relative abundance per sample was calculated and broken down by mOTU type (RefMeta or Meta).

COG	All mOTUs	detected in 252 fecal samples	mOTU (Meta) clusters	fraction of detected mOTUs (%)	relative abundance (%)	mOTU (RefMeta) clusters	fraction of detected mOTUs (%)	relative abundance (%)
COG0012	2126	739	437	59.1	42.7	302	40.9	57.3
COG0016	2152	748	433	57.9	42.9	315	42.1	57.1
COG0018	2069	610	386	63.3	40.8	224	36.7	59.2
COG0172	2098	716	417	58.2	43.3	299	41.8	56.7
COG0215	2193	757	433	57.2	45.8	324	42.8	54.2
COG0495	2192	669	382	57.1	41.1	287	42.9	58.9
COG0525	2131	670	370	55.2	42.7	300	44.8	57.3
COG0533	2087	668	392	58.7	42.5	276	41.3	57.5
COG0541	2143	715	403	56.4	43.2	312	43.6	56.8
COG0552	2151	718	412	57.4	42.6	306	42.6	57.5
mean	2134	701	407	58.0	42.7	295	42.0	57.3
stdev	41	46	24	2.2	1.3	28	2.2	1.3

Supplementary Table 7: Summary of using different distance metrics for community similarity analysis. Abundance data were normalized by total abundance and a subset was additionally log10. Six different distance metrics were used to calculate community similarities across 249 samples including 88 samples from 41 individuals that were sampled more than once. For these 88 samples, we calculated the percentage of instances when the most similar sample from one individual was collected from the same individual at a different time point.

Distance metric	Log transformed relative abundances?	mOTU-linkage groups (%)	MetaPhlAn (%)	RefMG (%)	Mean across methods (%)
Euclidean	Yes	97.7	86.4	92.0	92.0
Horn-Morisita	Yes	97.7	87.5	89.8	91.7
Bray-Curtis	Yes	98.9	81.8	85.2	88.6
Jensen-Shannon	Yes	97.7	80.7	84.1	87.5
Gower	Yes	95.5	76.1	75.0	82.2
Spearman	No	85.2	77.3	67.0	76.5
Spearman	Yes	85.2	77.3	67.0	76.5
Jensen-Shannon	No	80.7	67.0	70.5	72.7
Bray-Curtis	No	68.2	51.1	47.7	55.7
Gower	No	62.5	47.7	45.5	51.9
Horn-Morisita	No	39.8	36.4	23.9	33.3
Euclidean	No	38.6	33.0	26.1	32.6