Metagenomic species profiling using universal phylogenetic marker genes

Shinichi Sunagawa[1], Daniel R Mende[1], Georg Zeller[1], Fernando Izquierdo-Carrasco[2], Simon A Berger[2], Jens Roat Kultima[1], Luis Pedro Coelho[1], Manimozhiyan Arumugam[1,3,4], Julien Tap[1,5], Henrik Bjørn Nielsen[6,7], Simon Rasmussen[6], Søren Brunak[6,7], Oluf Pedersen[3,8–10], Francisco Guarner[11], Willem M de Vos[12,13], Jun Wang[3,4,14–16], Junhua Li[4,17,18], Joël Doré[5,19], S Dusko Ehrlich[5], Alexandros Stamatakis[2,20] & Peer Bork[1,21]

To quantify known and unknown microorganisms at species-level resolution using shotgun sequencing data, we developed a method that establishes metagenomic operational taxonomic units (mOTUs) based on single-copy phylogenetic marker genes. Applied to 252 human fecal samples, the method revealed that on average 43% of the species abundance and 58% of the richness cannot be captured by current reference genome–based methods. An implementation of the method is available at http://www.bork.embl.de/software/mOTU/.

A common approach for taxonomic profiling of microbial communities involves the sequencing and classification of amplified 16S ribosomal RNA gene (here referred to as 16S rDNA) fragments using DNA directly isolated from environmental samples. Owing to its universality in prokaryotes and the availability of large, curated reference databases, 16S rDNA is a powerful phylogenetic marker, yet its use has known problems including biases introduced by copy-number variations[1], variability in amplification efficiency[2], inconsistencies when targeting different regions of this gene[3], and problems with accurately and consistently delineating prokaryotic species[4].

In contrast to this single-gene targeted approach, shotgun sequencing of metagenomes generates millions of short reads that are randomly sampled from microbial community genomes. These reads are commonly aligned to taxonomically annotated reference genomes[5] to generate a read-abundance distribution in taxonomic bins. However, without appropriate normalization by genome size, which has to be estimated for uncharacterized species, taxonomic abundance estimates may be highly biased[5,6]. Alternatively, phylogenetic marker genes, either clade-specific[7] or universal, that are both present as single copies in most genomes[8] and rarely subject to horizontal gene transfer[9], are ideal candidates for taxonomic profiling of environmental samples[6,7,10]. Regardless of methodological differences, however, current approaches that use metagenomic shotgun sequencing data primarily for taxonomic composition analysis depend on the availability of reference (genome) sequences. Hence, they cannot resolve taxa for which no representative sequence information is available, although members of these unrepresented taxa may constitute a large fraction or even the majority of microbial communities.

To address this limitation, we developed a method based on universal, single-copy marker genes, which provide prokaryotic species boundaries at higher resolution than 16S rDNA[11], to estimate relative abundances of known and currently unknown microbial community members using metagenomics data at species-level resolution. The method clusters marker gene sequences from metagenomes and reference genomes into mOTUs. Based on covariance data across multiple samples, mOTUs of common species origin are combined into mOTU linkage groups (mOTU-LGs).

We started with 40 marker genes that previously had been used to accurately delineate prokaryotic species[11]. We calibrated and benchmarked an efficient profile hidden Markov model–based approach to identify these marker genes in reference genomes and metagenomes (**Supplementary Tables 1** and **2**, and Online Methods), and applied it to 3,496 prokaryotic reference genomes and 263 published human gut metagenomic samples[12,13] (Online Methods and **Supplementary Table 3**). Then we clustered the identified marker genes into mOTUs using marker gene–specific

[1]European Molecular Biology Laboratory, Heidelberg, Germany. [2]The Exelixis Lab, Scientific Computing Group, Heidelberg Institute for Theoretical Studies, Heidelberg, Germany. [3]The Novo Nordisk Foundation Center for Basic Metabolic Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark. [4]Beijing Genomics Institute (BGI) Shenzhen, Shenzhen, China. [5]Unité de Service 1367 Metagenopolis, Institut National de la Recherche Agronomique, Jouy en Josas, France. [6]Center for Biological Sequence Analysis, Technical University of Denmark, Kongens Lyngby, Denmark. [7]Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Kongens Lyngby, Denmark. [8]Hagedorn Research Institute, Gentofte, Denmark. [9]Institute of Biomedical Science, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark. [10]Faculty of Health Sciences, Aarhus University, Aarhus, Denmark. [11]Digestive System Research Unit, University Hospital Vall d'Hebron, Barcelona, Spain. [12]Laboratory of Microbiology, Wageningen University, Wageningen, The Netherlands. [13]Department of Bacteriology and Immunology, University of Helsinki, Helsinki, Finland. [14]King Abdulaziz University, Jeddah, Saudi Arabia. [15]Department of Biology, University of Copenhagen, Copenhagen, Denmark. [16]Macau University of Science and Technology, Macau, China. [17]BGI Hong Kong Research Institute, Hong Kong, China. [18]School of Bioscience and Biotechnology, South China University of Technology, Guangzhou, China. [19]Unité Mixte de Recherche 1319 Micalis, Institut National de la Recherche Agronomique, Jouy en Josas, France. [20]Karlsruhe Institute of Technology, Institute for Theoretical Informatics, Karlsruhe, Germany. [21]Max Delbrück Centre for Molecular Medicine, Berlin, Germany. Correspondence should be addressed to P.B. (bork@embl.de).

RECEIVED 18 JUNE; ACCEPTED 24 SEPTEMBER; PUBLISHED ONLINE 20 OCTOBER 2013; DOI:10.1038/NMETH.2693

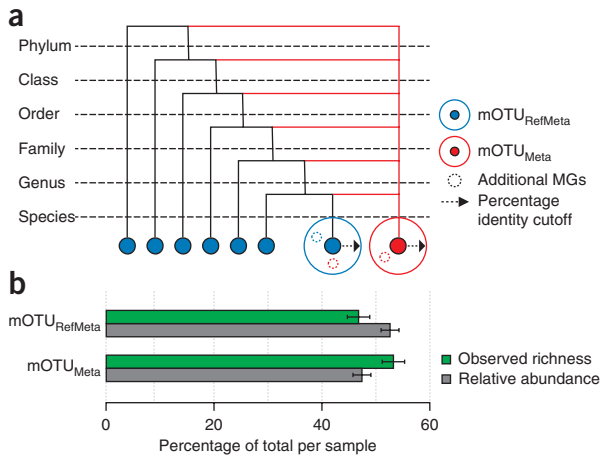

Figure 1 | Phylogenetic marker gene–based mOTUs. (**a**) Schematic showing the mOTUs that contained at least one marker gene (MG) that originated from a sequenced reference genome and at least one metagenomic MG (mOTU$_{RefMeta}$) and mOTUs that contained at least one metagenomic MG but no reference MG (mOTU$_{Meta}$). Black and red lines indicate known and unknown topologies, respectively. (**b**) Mean fractions of mOTU$_{Meta}$ and mOTU$_{RefMeta}$ of the observed mOTU richness per sample, and mean relative abundances based on mOTU abundance profiles of 252 human fecal samples.

species-level sequence identity cutoffs (**Supplementary Table 4**) to account for differences in evolutionary divergence rates across marker genes[11].

We assessed the suitability of each of the 40 marker genes for microbial composition profiling at the species level based on the false discovery rate (FDR) for their identification in reference genomes and the accuracy of their respective mOTUs in species-level profiling (**Supplementary Table 4** and Online Methods). From this analysis, we selected the ten best-performing marker genes, which had an average FDR of 1.4% (range, 0.1%–3.8%) and a mean ambiguous read alignment rate of 3.5% (range, 0.9%–6.4%; **Supplementary Table 4** and Online Methods). For comparison, the 16S rDNA, which is about ten times shorter than the 10 marker genes, had an ambiguous alignment rate of 41.1% when using species-level clustering cutoffs[11] and its accuracy in delineating prokaryotic species was lower than for any of the ten marker genes, in agreement with previous observations[11] (**Supplementary Fig. 1**).

The clustering step allowed us to estimate the fraction of gut microbial species that were not represented by currently available reference genome sequences, by classifying mOTUs based on the origin of their cluster members (i.e., whether they were similar to a reference marker gene or were only found in a metagenome; **Fig. 1a**). We defined mOTUs that recruited at least one metagenomic read and contained at least one marker gene originating from a reference genome as mOTU$_{RefMeta}$ and those that contained marker genes identified only in metagenomes as mOTU$_{Meta}$ (**Fig. 1a**). An average of 701 ± 46 (± s.d.) mOTUs per marker gene represented gut microbial species out of which as many as 58% ± 2.2% belonged to mOTU$_{Meta}$ (**Fig. 1b** and **Supplementary Table 5**). This implies that the majority of species in human gut microbial samples are not represented by current genomic resources, despite the substantial efforts that have gone into targeted genome sequencing projects with the goal of improving phylogenomic representation (ref. 14 and http://www.metahit.eu). The combined relative abundance of these species was on average 43% ± 1.3%, indicating that hitherto unsequenced gut species are likely important in the human gut ecosystem (**Fig. 1b**).

In the absence of genome sequences, it is not possible to reliably link sets of genes that originate from the same species due to the fragmented nature of metagenomic data. Such linkage groups would, however, not only increase the amount of phylogenetic

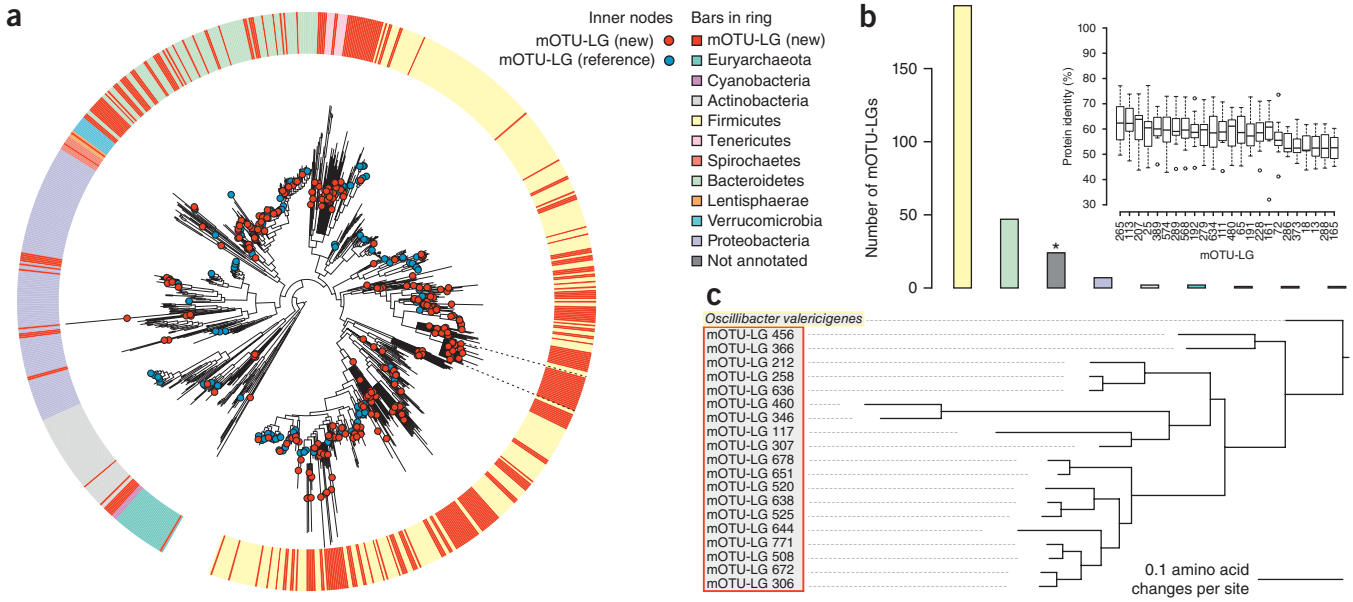

Figure 2 | Phylogenetic analysis of mOTU linkage groups. (**a**) Maximum likelihood phylogenetic tree of prokaryotic species used to infer the topology of mOTU-LGs (Online Methods). US National Center for Biotechnology Information phylum-level taxonomy is color-coded on the outer ring, and placements of mOTU-LGs are shown as circles on tree edges. Dashed lines indicate a clade of *Oscillibacter valericigenes* and related mOTU-LGs that are highlighted in **c**. (**b**) Phylum-level breakdown of new mOTU-LG (Online Methods). *, for mOTU-LGs that had no consistent annotation across their mOTU members at the phylum level, BLASTp identities of the member sequences are shown in the inset. Median protein identities (n = 6–10) with interquartile ranges (box) are shown with whiskers extending up to 1.5 times the interquartile range. (**c**) Maximum likelihood tree for a subset of mOTU-LGs that represent previously unidentified (new) species in the genus *Oscillibacter*.

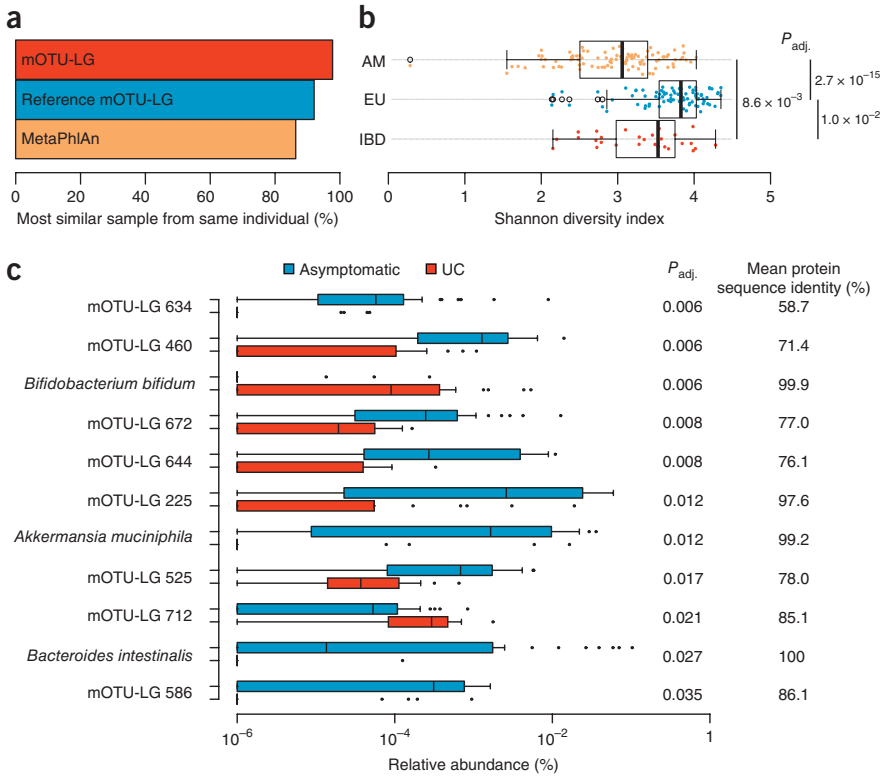

Figure 3 | Performance and application of mOTU linkage groups. (**a**) Fraction of samples originating from 43 individuals that were sampled at least twice (total 88 samples) for which the most similar sample originated from the same individual using mOTU-LG (red), a subset of mOTU-LGs that represent reference species (reference mOTU-LG) and clade-specific genes[7] at species level (MetaPhlAn). (**b**) Shannon diversity index for samples originating from US individuals (AM; $n = 97$), asymptomatic European individuals (EU; $n = 85$) and individuals diagnosed with IBD (IBD; $n = 25$). Individual samples are shown as closed circles and collective data for each group superimposed as box plots. $P_{adj.}$ denotes Bonferroni-adjusted P values of Wilcoxon's rank-sum test results. (**c**) Relative abundances of mOTU-LGs that were significantly different between fecal samples from a cohort of UC patients ($n = 21$) and matched asymptomatic individuals ($n = 35$). The mean (across mOTU-LG members) protein identity for best BLASTp hits is shown as a proxy for phylogenetic distance to the closest organism for which a reference genome sequence was available. $P_{adj.}$ values denote FDR-adjusted P values of Wilcoxon test results.

information per species, but also allow for more robust abundance estimations (by averaging across individual marker genes). Here we exploited the property that in shotgun-sequenced metagenomes genes are expected to covary in their abundance if they originate from the same species. Thus, we correlated mOTU abundances from different marker genes across 252 samples (**Supplementary Table 3**) to generate mOTU-LGs that represented 404 species (**Supplementary Figs. 2–5**), including 278 without a corresponding reference genome (**Supplementary Table 6**). On average, mOTU-LGs accounted for almost 95% of human gut microbial species (**Supplementary Fig. 6**). For 12 of the 30 most abundant species (40%), we found that representative genome sequences were lacking, highlighting the utility and importance of using reference-independent approaches to characterize microbial community structures (**Supplementary Fig. 7**).

We also tested the accuracy of taxonomic profiling using mOTU-LGs by running a benchmark against expected species abundances in a simulated gut metagenome (**Supplementary Table 1** and Online Methods). The mOTU-LG–based approach was highly accurate ($r^2 = 0.98$) and outperformed MetaPhlAn[7] ($r^2 = 0.90$), a method that uses clade-specific genes rather than universal genes for estimation of abundance (**Supplementary Fig. 8**).

We next asked whether species without a representative genome originated from only a few clades or whether they were widely distributed across the prokaryotic tree of life. To address this question, we constructed a maximum likelihood reference tree based on 1,753 species clusters[11]. Then we inferred the most likely placement of representative sequences for each mOTU-LG into this reference tree (**Fig. 2a** and Online Methods). For the 278 mOTU-LGs without species-level annotations, a phylum-level breakdown of their affiliation (**Fig. 2b**) showed that the vast majority belonged to Firmicutes (69%) and Bacteroidetes (17%). Some large clades were identified within genera for which a representative genome already existed. For example, within the genus *Oscillibacter* (phylum Firmicutes) we found 19 different mOTU-LGs, each of which suggests the existence of a species without a sequenced reference genome (**Fig. 2c**). Uncultured *Oscillibacter* have been identified as responsive to dietary change[15] and depleted in individuals with Crohn's disease[16]. The species-level resolution provided here thus might be particularly important for studying their functional role in these processes.

We also identified organisms that were highly divergent from any sequenced reference species. For example, the closest relatives for mOTU-LG 159 in the phylum Euryarchaeota were from the genera *Aciduliprofundum* and *Methanocella*, but each of them exhibited a protein sequence identity of only 54%. Similarly, a cluster of six mOTU-LGs (mOTU-LGs 13, 18, 165, 233, 286 and 373) was most closely related to Cyanobacteria, but again, protein sequence identities were as low as 53%.

Next, we evaluated how our method's ability to profile relative abundances of species for which no reference genome sequences were available affects comparative analyses of microbial community similarities. For this, we used a data set from 207 individuals: 110 individuals in Europe sampled once, and 97 individuals in the United States including 54 sampled once, 41 sampled twice and two sampled three times[13,17]. Based on evidence that variability of human gut microbial composition is smaller within individuals than between individuals[18,19], we used the subset of 43 US individuals from whom at least two fecal samples were collected 37–378 d apart, in the context of all samples, to identify samples from the same individual.

We compared our method, mOTU-LGs to two reference-based approaches; one uses the subset of mOTU-LGs that were annotated at species cluster-level (**Supplementary Table 6**) and the other one, MetaPhlAn, uses clade-specific marker genes that were

identified in reference genomes[7]. As a measure of performance, we calculated the fraction of cases in which a sample collected from one individual was more similar to another sample from the same individual than to any other sample. For mOTU-LGs, 98% of samples matched to a different sample from the same individual, compared to only 92% and 86% for annotated mOTU-LGs and MetaPhlAn, respectively (**Fig. 3a**). The higher performance of mOTU-LGs was also consistent across a broad selection of distance measures that we tested (**Supplementary Table 7**). Our results demonstrate that the increased resolution of mOTU-LGs provided by including previously unidentified species yields more accurate community similarity estimates compared to methods that solely depend on the availability of reference sequences. Furthermore, this implies that the temporal variation of microbial community structures of human gut microbiota is lower than previously assumed primarily owing to limitations of other existing methods.

Calculating species diversities can provide information about the ecology of microbial communities or may also indicate potential sampling biases. We calculated the Shannon diversity index for fecal samples from asymptomatic US individuals (here referred to as 'AM'; $n = 97$: only samples from the first time point were included) and European individuals (EU; $n = 85$), and patients with inflammatory bowel disease (IBD; $n = 25$; **Supplementary Table 3**). The differences between samples from AM, EU and IBD data sets were all significant (AM versus EU, $P = 2.7 \times 10^{-15}$; EU versus IBD, $P = 0.0086$; AM versus IBD, $P = 0.0102$; pairwise Wilcoxon test, adjusted using Bonferroni's correction; **Fig. 3b**). Samples from AM data exhibited the lowest diversity followed by those from IBD data and EU data. The reduced species diversity we detected in individuals that were diagnosed with IBD corroborated previous studies[20]. Lower species diversity in fecal samples from asymptomatic US individuals compared to those collected from IBD patients may indicate a methodological bias as the data were collected in two independent large-scale projects[12,13]. It should be noted that these projects used different methods for sample collection, DNA extraction and DNA sequencing, which highlights the need for standardizing protocols from sample collection to DNA sequencing. The International Human Microbiome Standards consortium is currently addressing this issue.

Finally, we tested for differentially abundant species between fecal samples from individuals with ulcerative colitis (UC) and healthy controls (**Supplementary Table 3**). At an FDR of 5%, 11 species were identified including *Bifidobacterium bifidum*, *Bacteroides intestinalis* and *Akkermansia muciniphila*, the latter of which has been shown to be depleted in UC patients[21,22]. In addition, we identified differentially abundant Firmicutes, putatively in the order of Clostridiales, which were highly divergent from any currently sequenced reference species (**Fig. 3c**). This result illustrates the practical utility of our method and underscores the importance of profiling currently unknown species and the need for sequencing additional genomes to better understand the functional role of these microorganisms in the human gut ecosystem.

When using reference-based methods for species-level taxonomic profiling of shotgun-sequenced metagenomic samples, data that originate from unknown species are currently ignored, summarized into a single taxonomically unassigned group or grouped at higher taxonomic levels using a last common ancestor–based approach. The main novelty of our method is to resolve this single unassigned fraction into species-level taxonomic abundances. We demonstrated the advantages of our method over reference-based methods and illustrated examples of its practical utility. In addition to gene functional[23,24] and genomic variation analyses[17], taxonomic profiling using mOTUs provides another powerful approach for the exhaustive mining of different types of information contained in metagenomic data.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank members of the European Molecular Biology Laboratory Information Technologies core facility and Y. Yuan for managing the high-performance computing resources and the members of the Bork group for fruitful discussions. This work was supported by funding from European Molecular Biology Laboratory, the European Community's Seventh Framework Programme via the MetaHIT (HEALTH-F4-2007-201052) and International Human Microbiome Standards, (HEALTH-F4-2010-261376) grants, The Novo Nordisk Foundation, The Lundbeck Foundation, institutional funding by the Heidelberg Institute for Theoretical Studies and Deutsche Forschungsgemeinschaft grants STA 860/2 and STA 860/3, the Metagenopolis ANR-11-DPBS-0001 grant, and the European Research Council Advanced Grants (MicrobesInside and CancerBiome grant agreement numbers 250172 and 268985 to W.M.d.V. and P.B., respectively).

AUTHOR CONTRIBUTIONS

P.B. and S.S. conceived the study, S.S., D.R.M., G.Z., F.I.-C., S.A.B., M.A., J.T. and A.S. designed and performed the analyses, S.S., D.R.M., G.Z., J.R.K., L.P.C. and J.L. developed and implemented the program, O.P., F.G., J.D. and J.W. provided data, S.S., D.R.M., G.Z. and P.B. wrote the manuscript, and M.A., J.T., H.B.N., S.R., O.P., F.G., W.M.d.V., S.D.E. and A.S. gave conceptual advice and revised the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Klappenbach, J.A., Saxman, P.R., Cole, J.R. & Schmidt, T.M. *Nucleic Acids Res.* **29**, 181–184 (2001).
2. Engelbrektson, A. *et al. ISME J.* **4**, 642–647 (2010).
3. Claesson, M.J. *et al. Nucleic Acids Res.* **38**, e200 (2010).
4. Gevers, D. *et al. Nat. Rev. Microbiol.* **3**, 733–739 (2005).
5. Arumugam, M. *et al. Nature* **473**, 174–180 (2011).
6. Liu, B., Gibbons, T., Ghodsi, M., Treangen, T. & Pop, M. *BMC Genomics* **12** (suppl. 2), S4 (2011).
7. Segata, N. *et al. Nat. Methods* **9**, 811–814 (2012).
8. Ciccarelli, F. *et al. Science* **311**, 1283–1287 (2006).
9. Sorek, R. *et al. Science* **318**, 1449–1452 (2007).
10. von Mering, C. *et al. Science* **315**, 1126–1130 (2007).
11. Mende, D.R., Sunagawa, S., Zeller, G. & Bork, P. *Nat. Methods* **10**, 881–884 (2013).
12. Qin, J. *et al. Nature* **464**, 59–65 (2010).
13. The Human Microbiome Project Consortium. *Nature* **486**, 215–221 (2012).
14. Nelson, K.E. *et al. Science* **328**, 994–999 (2010).
15. Walker, A.W. *et al. ISME J.* **5**, 220–230 (2011).
16. Mondot, S. *et al. Inflamm. Bowel Dis.* **17**, 185–192 (2011).
17. Schloissnig, S. *et al. Nature* **493**, 45–50 (2013).
18. Turnbaugh, P.J. *et al. Nature* **457**, 480–484 (2009).
19. Rajilic-Stojanovic, M., Heilig, H.G., Tims, S., Zoetendal, E.G. & de Vos, W.M. *Environ. Microbiol.* **15**, 1146–1159 (2012).
20. Manichanh, C., Borruel, N., Casellas, F. & Guarner, F. *Nat. Rev. Gastroenterol. Hepatol.* **9**, 599–608 (2012).
21. Rajilic-Stojanovic, M., Shanahan, F., Guarner, F. & de Vos, W.M. *Inflamm. Bowel Dis.* **19**, 481–488 (2013).
22. Png, C.W. *et al. Am. J. Gastroenterol.* **105**, 2420–2428 (2010).
23. Qin, J. *et al. Nature* **490**, 55–60 (2012).
24. Forslund, K. *et al. Genome Res.* **23**, 1163–1169 (2013).

ONLINE METHODS

Implementation of the method. Metagenomic species profiling has been implemented as a new feature in the MOCAT pipeline[25] and as a standalone tool (**Supplementary Software**). Software and scripts that were used to identify marker genes (MGs) in reference genomes and metagenomes as well as a tutorial for profiling metagenomic data sets are available at http://www.bork.embl.de/software/mOTU/.

Identification of single-copy marker genes. Profile hidden Markov models (HMMs) were generated using the hmmbuild and hmmsearch programs of HMMER[26] (v3) for 40 universal single-copy MGs[8,9] based on multiple-sequence alignments of their orthologous groups that had been previously identified in 1,497 prokaryotic genomes[17,27,28]. Prokaryotic reference genome sequences were downloaded (February 2012) from the US National Center for Biotechnology Information (NCBI) genomes database. As a quality filter, we removed complete genomes with less than 30 MGs and genomes with more than 500 contigs, yielding a set of 3,496 reference genomes (**Supplementary Table 1**). For these genomes, we used hmmsearch with a bit score cutoff of 60 to identify 138,132 MGs (39.5 per genome) in >11 million proteins by selecting the highest-scoring target sequence (best hit) for each MG in each genome (**Supplementary Table 2**). Compared to a BLAST-based annotation of the same set of proteins, the HMM–based procedure was four orders of magnitude faster in terms of computing time (134,000 versus 17.5 CPU h).

In addition to increasing the speed of MG identification, it is important to minimize the FDR when extending these search methods to metagenomic data, because selecting the best hit (as done for reference genomes) is not possible and a bit score threshold must be used instead. For example, when simply using HMMs with a cutoff of 60 bits on the set of 3,496 genomes, 15.7% more genes that are likely false positives were identified compared to the best hits only (**Supplementary Table 2**). Thus, we calibrated MG-specific bit-score cutoffs (**Supplementary Table 2**) by maximizing the accuracy (*F* score) of MG identification using a training set of 1,004 well-annotated prokaryotic genomes that are available in the eggNOG (evolutionary genealogy of genes: Nonsupervised Orthologous Groups; v3.0) database[28]. When repeating the search using these calibrated cut offs, the increase in the number of identified sequences compared to selecting the best hit only was 3.3% (expected, 39.51 MGs per genome; observed, 40.81 MGs) for all 40 MGs. This was reduced to 1.0% (expected, 38.52 per genome; observed, 38.90 MGs) when COG0085 was excluded, which alone accounted for 71% of all putative false positives (**Supplementary Table 2**).

Metagenome assembly and marker gene prediction. Raw data from 263 Illumina shotgun-sequenced gut metagenomic samples from 39 Spanish, 85 Danish and 94 US (American) individuals[12,13,17] (**Supplementary Table 3**) were quality-controlled and processed using MOCAT as previously described[25]. Briefly, raw sequencing reads were quality-trimmed with a base quality and read length cutoff of 20 and 30, respectively. High-quality reads were then assembled, assemblies were revised, and genes were predicted on contiguous sequences longer than 500 base pairs (bp) without any unknown bases ('scaftigs'). Metagenomic MGs were subsequently identified using the calibrated profile HMM bit score cutoffs (see above) in translated sequences of all full-length genes.

To test whether artificial MG sequences were expected to be commonly generated when assembling metagenomic data, we simulated a representative gut metagenome and applied the same data processing steps as described above. From the 101 most abundant genomes among 252 samples that we used for abundance profiling here and in ref. 17, we selected the overlap with the set of 3,496 reference genomes and additionally removed genomes for which full binomial species names were not available (**Supplementary Table 1**). We then calculated the expected relative abundance for the remaining 87 genomes and simulated 4.5 Gbp of paired-end 75-bp Illumina reads with modeled sequencing errors as described in ref. 29, which were assembled using the same parameters as described above. We identified a total of 184 assembled full-length MGs and aligned them to the genome sequences of origin using BLASTn. All MG sequences were aligned completely, and only a single mismatch was found among the 270,042 aligned nucleotide positions.

Clustering of marker gene sequences into mOTUs. For all MGs identified in both reference genomes and metagenomes, clustering was performed with USEARCH[30] (v4.2.66) and MG-specific species-level identify cutoffs[11] (**Supplementary Table 2**) resulting in universal single-copy MG-based mOTUs. Before clustering, we sorted the nucleotide sequences so that MGs identified in reference genome sequences were followed by those identified in metagenomes. The accuracy of clustering was evaluated by testing whether mOTUs that had at least two MGs originating from a reference genome were consistent regarding the taxonomic annotation of their members according to a curated version of the NCBI taxonomy[11]. We also compared the results to the commonly used 16S rDNA (using full-length sequences identified in 2,899 genomes listed in **Supplementary Table 1**) and found that the accuracy for delineating prokaryotic species was lower than for any of the ten MGs (**Supplementary Fig. 1**).

Abundance profiling of gut metagenomic samples using mOTUs and selection of MGs. To reduce the effect of unequal alignment probability across the length of the gene, we first dereplicated (clustered at 100% identity) marker gene sequences and extended them up to 100 bp upstream and downstream of the start and stop codon using the sequence information available on the reference genomes and assembled scaftigs of origin. Using this nonredundant, 'padded' mOTU database, abundance calculations of each mOTU were performed using MOCAT[25]. In summary, shotgun sequencing reads from 252 gut metagnomic samples (**Supplementary Table 3**) were mapped to the mOTU database using a base quality, sequence identity and alignment-length cutoff of 20, 97% and 45 bp, respectively. For each DNA fragment (insert) that mapped to one or more MGs that belonged to the same mOTU (unique mapper), we increased the count of the respective mOTU by one (note that two paired-end reads or a single read, i.e., if one of the paired-end sequence reads did not map or had previously been removed due to low quality, were counted as one insert). For each insert that mapped to MGs from *n* different mOTUs with the same alignment score (multiple mapper), the count for each of the respective mOTUs was increased by the fraction of unique mappers of these mOTUs. The insert counts were then normalized to gene length, scaled by the average gene length and rounded down to obtain integer numbers of mOTU abundances.

doi:10.1038/nmeth.2693

We used two criteria to assess the suitability of MGs given the aims of our analysis: the accuracy of their identification based on profile HMM searches and their accuracy in quantifying species abundances, the latter of which poses a challenge because of the limited length of shotgun sequencing reads typical for metagenomic data sets. We required the accuracy of MG identification to be higher than 90%, which excluded two of the 40 MGs (COG0085 and COG0124). To quantify the effect of ambiguous sequence alignment, we calculated the number of multiple mappers and found that across all samples, the mean rate of multiple mappers was 18.9%, ranging from 1.2% to 61.5% for different MGs (**Supplementary Table 4**). As a compromise between maximizing the phylogenetic signal and minimizing noise, we excluded any MG with a median multiple mapper fraction above 7% (**Supplementary Table 4**), which resulted in the ten MGs selected for this work.

Linking mOTUs of common species origin. Sequences that originate from the same prokaryotic genome are expected to covary in abundance across metagenomic samples. We thus inferred genomic linkage between mOTUs constructed from different marker genes based on their correlated abundance across metagenomes. We computed Spearman correlation between abundance estimates for each of 7,010 mOTUs across 252 metagenomic samples. To avoid spurious associations between very rare mOTUs, we excluded any mOTU with prevalence (proportion of nonzero abundances among all samples) below 2%. To link mOTUs, we employed a greedy procedure that examined the entries in the correlation matrix starting from the strongest correlations. In each step, it linked two mOTUs if the resulting group contained at most one member from each marker gene (also considering other mOTUs linked to this group in previous steps).

To estimate the accuracy of this approach, we used the taxonomic information available for MGs originating from reference genomes, i.e., those that were contained in $mOTU_{RefMeta}$ (using a curated NCBI taxonomy[11]). For this, we only considered mOTUs with a consistent taxonomic annotation, i.e., with more than half their taxonomically annotated sequences coming from the same taxon. For each step of the linking procedure we determined whether it grouped mOTUs with the same taxonomic annotation (true positives, TP) or with different annotations (false positives, FP), allowing us to monitor the FDR as TP / (TP + FP) during the linking process. Owing to the greedy nature of the algorithm, this FDR increased with the number of steps as the underlying correlation coefficients decreased (**Supplementary Fig. 3**). Based on the FDR, we terminated the linking procedure after the maximum number of steps that still maintained an FDR below 0.01 (resulting after 40,889 linking steps). Further analysis was restricted to the 404 mOTU linkage groups that contained at least six different marker genes (**Supplementary Table 6**). Together these represented an average of 94.2% of the total abundance of all detected species in our gut metagenomic data set (**Supplementary Fig. 6**). The relative abundance of each mOTU-LG in a given sample was estimated as the mean relative abundance of all of its members.

Quality assessment of mOTU linkage groups. We assessed to which extent each mOTU linkage group as a whole was consistent with respect to the taxonomic annotation of its members (**Supplementary Fig. 2**; note that these results differ from the linking FDR estimate above, which only evaluates pair-wise links between mOTUs, not the resulting group). We found that for >99% of mOTU-LGs, the majority of MGs were consistent at the species level (for >95%, all MGs were consistent); at the genus level, >98% of mOTU-LGs had completely consistent taxonomic annotations (**Supplementary Fig. 2**). We next evaluated how well the relative abundance estimates derived from individual mOTUs agreed within each cluster (**Supplementary Fig. 4**). As a final quality control step, we estimated the G+C content of the gene sequences in each mOTU and assessed the homogeneity in G+C content within each mOTU-LG, as G+C content from genes belonging to the same genome is expected to be much more homogenous than from a random sample of genes from a metagenome (**Supplementary Fig. 5**).

Taxonomic and phylogenetic diversity. Shannon diversity indices were calculated for each sample based on mOTU-LG abundances using the diversity function in the vegan R package (http://cran.r-project.org/web/packages/vegan/index.html). Differences between samples collected on the two continents and IBD patients were tested using the Wilcoxon rank-sum test, and *P* values were adjusted using Bonferroni's correction. To evaluate the phylogenetic diversity of new species, we first generated a reference tree based on the 1,753 species clusters as defined in ref. 11. For each species cluster, the genome with the highest N50 statistic (the length for which the collection of all contigs of that length or longer contains at least half of the total of the lengths of the contigs) was chosen as the representative genome, and multiple sequence alignments (MSAs) were generated for each of the 40 MG using AQUA[27] (v1.1). The 40 MSAs were concatenated into a multigene MSA comprising 50,911 protein sites. For each of the 40 MSAs, the best log-likelihood scores were calculated using RAxML[31] (v7.3.2) to determine the best-scoring empirical protein substitution model to be assigned to each MG partition of the multigene MSA.

ExaML[32] for large-scale phylogenetic inference was run to execute six independent maximum-likelihood searches for the best-scoring maximum likelihood (ML) tree on six distinct parsimony starting trees that were computed using the randomized stepwise addition order algorithm as implemented in standard RAxML (v7.3.2). Searches were conducted under the per site rate model of among-site heterogeneity. Finally, we scored the resulting six tree topologies from the ExaML searches using standard RAxML (v7.3.1) under the GAMMA model of rate heterogeneity[31]. The best-scoring tree topology was then chosen as the reference species tree for the phylogeny-aware placement of mOTU-LG sequences. PaPaRa[33] (v2) was used to align the fragments to the multigene reference MSA. Subsequently, the RAxML Evolutionary Placement Algorithm (EPA)[34] was run to place the fragments into the reference species tree. The EPA placement runs were conducted under the same partitioning scheme and protein substitution models that were used to infer the reference species tree. Although the PaPaRa and EPA algorithms were readily available, we introduced several technical and methodological improvements to increase scalability of the alignment and placement steps in order to handle large data sets. We used the heuristic tuning parameter of the EPA algorithm that relies on a fast prescoring method, which in turn reduces the number of expensive full-likelihood computations to calculate fragment insertion scores by a factor of 10 (ref. 34). The placement results were visualized using interactive Tree of Life (iTOL)[35].

Time-series benchmark. Two reference-dependent methods that calculate relative species abundances by mapping metagenomic reads to clade-specific[7] or reference mOTU-LGs (this work) were compared to our mOTU-LG–based method. Relative species abundances calculated by the MetaPhlAn pipeline were downloaded from the Human Microbiome Project Data Analysis and Coordination Center website (ftp://public-ftp.hmpdacc.org/HMSMCP/HMP.ab.txt.bz2). To assess the robustness of our results, we tested three qualitatively different distance measures (rank-based, Spearman; frequency-based, Euclidean; frequency-weighted, Jensen-Shannon Distance (JSD)) and three additional distance measures that are commonly applied in community ecological studies (Bray-Curtis, Gower and Horn-Morisita) on log-transformed (after adding a pseudocount of 10^{-6}) and non-transformed data. For each of the 88 samples originating from either one of the 41 individuals that were sampled twice or one of the two individuals that were sampled three times, we calculated the fraction of samples for which the most similar sample originated from the same individual.

Differential abundance estimation in IBD. We examined 56 gut metagenomes (**Supplementary Table 3**), 21 of which originated from patients diagnosed with UC, a subset of samples used previously for a matched analysis of IBD versus non-IBD samples[23]. Significant differences in species abundance were determined using the Wilcoxon test with a maximum q value (FDR correction) of 0.05.

25. Kultima, J.R. *et al. PLoS ONE* **7**, e47656 (2012).
26. Eddy, S.R. *PLoS Comput. Biol.* **7**, e1002195 (2011).
27. Muller, J., Creevey, C.J., Thompson, J.D., Arendt, D. & Bork, P. *Bioinformatics* **26**, 263–265 (2010).
28. Powell, S. *et al. Nucleic Acids Res.* **40**, D284–D289 (2012).
29. Mende, D.R. *et al. PLoS ONE* **7**, e31386 (2012).
30. Edgar, R.C. *Bioinformatics* **26**, 2460–2461 (2010).
31. Stamatakis, A. *Bioinformatics* **22**, 2688–2690 (2006).
32. Stamatakis, A. & Aberer, A. in *Proceedings of the 2013 IEEE 27th International Symposium on Parallel and Distributed Processing* 1195–1204 (2013).
33. Berger, S.A. & Stamatakis, A. *Bioinformatics* **27**, 2068–2075 (2011).
34. Berger, S.A., Krompass, D. & Stamatakis, A. *Syst. Biol.* **60**, 291–302 (2011).
35. Letunic, I. & Bork, P. *Bioinformatics* **23**, 127–128 (2007).

doi:10.1038/nmeth.2693