Metagenomic microbial community profiling using unique clade-specific marker genes

Article *in* Nature Methods · June 2012

DOI: 10.1038/nmeth.2066 · Source: PubMed

6 authors, including:

Nicola Segata
Università degli Studi di Trento
226 PUBLICATIONS **7,163** CITATIONS
SEE PROFILE

Annalisa Ballarini
Università degli Studi di Trento
22 PUBLICATIONS **707** CITATIONS
SEE PROFILE

Olivier Jousson
Università degli Studi di Trento
190 PUBLICATIONS **2,614** CITATIONS
SEE PROFILE

Curtis Huttenhower
Harvard University
411 PUBLICATIONS **9,548** CITATIONS
SEE PROFILE

Metagenomic microbial community profiling using unique clade-specific marker genes

Nicola Segata[1], Levi Waldron[1], Annalisa Ballarini[2], Vagheesh Narasimhan[1], Olivier Jousson[2] & Curtis Huttenhower[1]

Metagenomic shotgun sequencing data can identify microbes populating a microbial community and their proportions, but existing taxonomic profiling methods are inefficient for increasingly large data sets. We present an approach that uses clade-specific marker genes to unambiguously assign reads to microbial clades more accurately and >50× faster than current approaches. We validated our metagenomic phylogenetic analysis tool, MetaPhlAn, on terabases of short reads and provide the largest metagenomic profiling to date of the human gut. It can be accessed at http://huttenhower.sph.harvard.edu/metaphlan/.

Microbial communities are responsible for a broad spectrum of biological activities carried out in virtually all natural environments including oceans[1], soil[2] and human-associated habitats[3–5]. Profiling the taxonomic and phylogenetic compositions of such communities is critical for understanding their biology and characterizing complex disorders such as inflammatory bowel diseases[4,6] and obesity[7] that do not appear to be associated with any individual microbes.

Metagenomic shotgun sequencing provides a uniquely rich profile of microbial communities, with each data set yielding billions of short reads sampled from the DNA in the community. A community's taxonomic composition can be estimated from such data by assigning each read to the most plausible microbial lineage, often with a taxonomic resolution not achievable by profiling the universal 16S ribosomal RNA marker gene alone. Both alignment- and composition-based approaches have been developed for this task, and the two approaches have also been integrated in hybrid methods (see Online Methods). However, none of these methods has simultaneously achieved both the efficiency and the species-level accuracy required by current high-complexity data sets because of computational limitations, untenable accuracy for short (<400 nucleotide) reads and the need to normalize read counts into clade-specific relative abundances (**Supplementary Note 1**).

We thus present MetaPhlAn (Metagenomic Phylogenetic Analysis), a tool that accurately profiles microbial communities and requires only minutes to process millions of metagenomic reads. MetaPhlAn estimates the relative abundance of microbial cells by mapping reads against a reduced set of clade-specific marker sequences that are computationally preselected from coding sequences that unequivocally identify specific microbial clades at the species level or higher taxonomic levels and cover all of the main functional categories

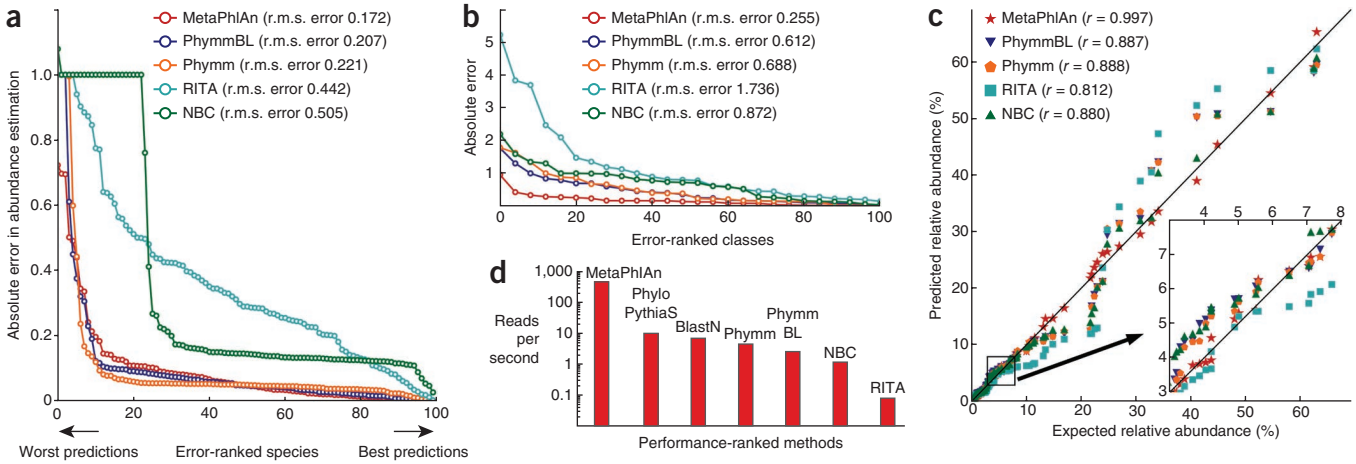

Figure 1 | Comparison of MetaPhlAn to existing methods. We used ten total synthetic metagenomes to compare MetaPhlAn to PhymmBL[12], BLAST[13], the Rapid Identification of Taxonomic Assignments (RITA) pipeline[14] and the naive Bayes classifier (NBC)[15]. (**a**,**b**) Absolute and r.m.s. errors with respect to 100 total organisms in one synthetic metagenome at the species (**a**) and class (**b**) level. (**c**) Correlations of inferred and true species abundances for eight non–evenly distributed synthetic metagenomes. (**d**) Read rates for the tested methods on single CPUs.

[1]Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts, USA. [2]Centre for Integrative Biology, University of Trento, Trento, Italy. Correspondence should be addressed to C.H. (chuttenh@hsph.harvard.edu).

RECEIVED 28 NOVEMBER 2011; ACCEPTED 7 MAY 2012; PUBLISHED ONLINE 10 JUNE 2012; DOI:10.1038/NMETH.2066

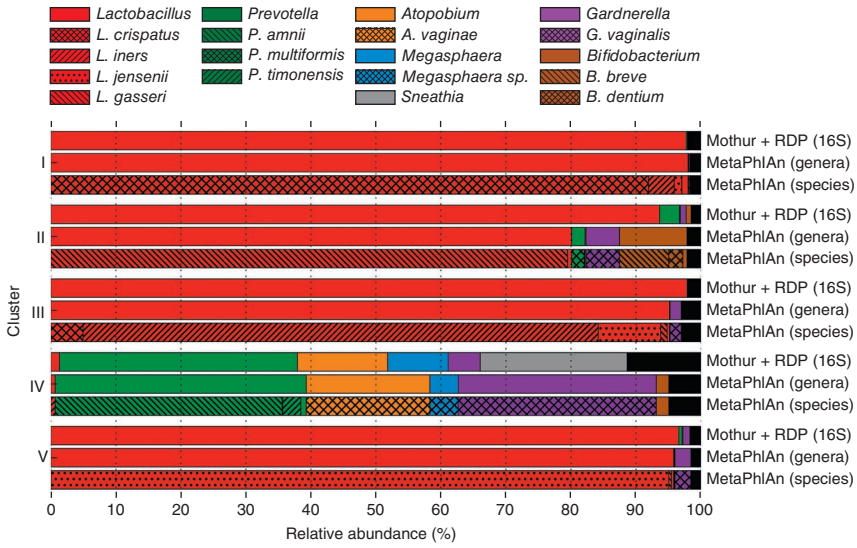

Figure 2 | Composition of healthy vaginal microbiota. MetaPhlAn species and genus abundances and 16S phylotype abundances for 51 healthy vaginal microbiomes from the Human Microbiome Project are shown. Samples were naively grouped by assigning each based on its dominant (>50%) *Lactobacillus* species or by the absence (<2%) of any *Lactobacillus*. For each cluster (named from I to V[5]) we report averages across samples for all genera and species as inferred by MetaPhlAn and, for genera, as estimated by the combination of the mothur 16S rRNA processing pipeline and the Ribosomal Database Project (RDP) classifier (see Online Methods) applied to 16S rRNA gene sequences from the same specimen.

(**Supplementary Fig. 1**). Starting from the 2,887 genomes available from the Integrated Microbial Genomes (IMG) system (2011 July)[8], we identified more than 2 million potential markers from which we selected a subset of 400,141 genes most representative of each taxonomic unit (Online Methods). The resulting catalog spans 1,221 species with 231 (s.d. 107) markers per species and >115,000 markers at higher taxonomic levels (available at http://huttenhower.sph.harvard.edu/metaphlan/).

The MetaPhlAn classifier compares each metagenomic read from a sample to this marker catalog to identify high-confidence matches. This can be done very efficiently, as the catalog contains only ~4% of sequenced microbial genes, and each read of interest has at most one match due to the markers' uniqueness. Because spurious reads are unlikely to have significant matches with a marker sequence, no preprocessing of metagenomic DNA (such as error detection, assembly or gene annotation) is required. The classifier normalizes the total number of reads in each clade by the nucleotide length of its markers and provides the relative abundance of each taxonomic unit, taking into account any markers specific to subclades. Microbial reads belonging to clades with no available sequenced genomes are reported as an 'unclassified' subclade of the closest ancestor for which there is available sequence data.

We first evaluated MetaPhlAn's performance in estimating microbial community composition using synthetic data. We constructed ten data sets comprising 4 million noisy reads from 300 organisms. MetaPhlAn mapped a number of reads consistent with the fraction of lineage-specific genomic regions (7.7% and 8.4%), correctly identified all 200 organisms in the two high-complexity data sets and accurately estimated their species relative abundances (r.m.s. error of 0.17 and 0.14), with 75% of species within 10% deviation from expected value (**Fig. 1a** and **Supplementary Fig. 2**). Similar performance was observed for higher taxonomic ranks (**Fig. 1b** and **Supplementary Figs. 2** and **3**) and for the eight non–evenly distributed data sets that better mimic the abundance distributions of real communities (Pearson $r > 0.991$, species-level Pearson $P < 2 \times 10^{-21}$) (**Fig.1c** and **Supplementary Fig. 4**).

As compared to all prior methods, which are instead optimized for read-based statistics, we considered our microbial-cell relative-abundance estimation more biologically informative. Microbial clade abundances were thus estimated by normalizing read-based counts by the average genome size of each clade. MetaPhlAn compares favorably to existing methods on all tested synthetic metagenomes (**Fig. 1** and **Supplementary Figs. 2–4**), with PhymmBL being the closest (but substantially slower) alternative. This also held true for the more challenging scenario in which metagenomes contained microbes without reference genomes (**Supplementary Tables 1** and **2**).

Notably, MetaPhlAn achieved a classification rate of about 450 reads per second on standard single-processor systems, thus greatly outperforming all existing methods (**Fig. 1d**; note that PhyloPythiaS provides only genus-level predictions). This allowed us to provide what is, to our knowledge, the first practical high-throughput assessments of several real-world metagenomes at the species level, as detailed below.

We first characterized the vaginal microbiota of asymptomatic premenopausal adult women enrolled in the Human Microbiome Project (HMP)[3], analyzing 51 metagenomes sampled from the posterior fornix. MetaPhlAn detected 98 clades with abundances >0.5% in at least one sample (32 species from 17 genera). *Lactobacillus* was consistently the most abundant genus; it represented >50% of the bacterial community in 49 of the 51 samples and >98% in half of the samples, thus confirming its well-established role in healthy vaginal microbiomes[5,9]. Sequencing of the 16S rRNA gene in an independent cohort[5] previously identified five distinct community types, each characterized by a specific *Lactobacillus* species or by the absence of any of them; in MetaPhlAn's results, these five groups are easily identifiable (**Supplementary Fig. 5**) and cluster naturally by species (**Fig. 2**). Although the characterization of *Lactobacillus* species-level operational taxonomic units in 16S pyrosequencing data is sensitive to the region being sequenced, we performed a direct comparison with 16S data from the same HMP specimens. Despite extensive technical differences between 16S pyrosequencing and shotgun sequencing, the estimated relative abundances were similar in all clusters. Moreover, MetaPhlAn's native coverage of all species with sequenced members further details the structure of these clusters. For example, *Lactobacillus* is not the only genus with species-specific differences among these microbiome types; *Bifidobacterium* species are present in cluster II as *B. breve* and *B. dentium,* whereas in cluster IV we identified an unclassified member distinct from all sequenced species. Similarly, *Prevotella* is represented by *P. multiformis* in cluster II in contrast to *P. amnii* and *P. timonensis* in cluster IV.

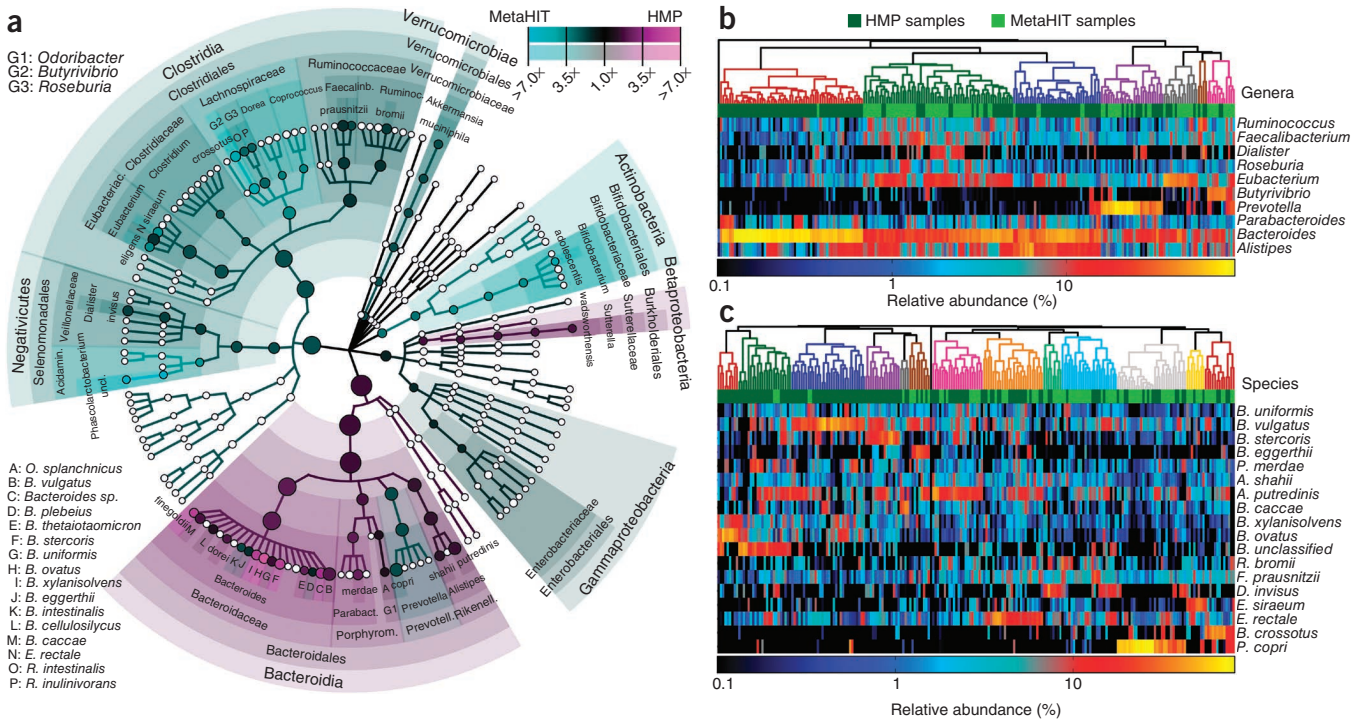

Figure 3 | The gut microbiota in asymptomatic Western populations as inferred by MetaPhlAn on 224 samples combining the HMP and MetaHIT cohorts. (**a**) Taxonomic cladogram reporting all clades present in one or both cohorts (≥0.5% abundance in ≥1 sample). Circle size is proportional to the log of average abundance; color represents relative enrichment of the most abundant taxa (≥1% average in ≥1 cohort) between the HMP (n = 139) and MetaHIT (n = 85, healthy only) populations. (**b**,**c**) Genus- (**b**) and species-level (**c**) taxonomic profiles of the most abundant clades, hierarchically clustered (average linkage) with the Bray-Curtis similarity, reveal sets of samples with similar microbial community compositions. With the exception of the cluster dominated by genus *Bacteroides* (*B. vulgatus* and *B. ovatus* in particular), samples from both studies are present in all groups, thus confirming substantial consistency of the gut microbiota characterized by independent and geographically distant Western-diet asymptomatic cohorts. Only species and genera with at least 7.5% abundances at the 95th percentile of their distribution are reported.

MetaPhlAn's methodology is restricted neither to bacteria alone nor to human-associated microbiomes, and it allowed us to investigate the microbial flora collected from oxygen minimum zones at intermediate depths in the eastern tropical South Pacific[10]. This marine ecosystem proved to include a substantial fraction of archaea, but Proteobacteria (mainly Alphaproteobacteria) was the most abundant phylum, representing approximately half of the community. Depth-associated shifts were observable for Bacteroidia, Chlamydiae, and Gammaproteobacteria, whereas the *Cenarchaea* dropped off specifically within the deepest sample. MetaPhlAn's relative abundances (**Supplementary Fig. 6**) were consistent with BLAST-based approaches[10] and confirmed its applicability for communities with limited coverage of reference genomes and its suitability for environmental metagenomic samples. Moreover, as the number of sequenced organisms continues to increase, MetaPhlAn's species specificity in such environments will automatically improve without computational drawbacks.

MetaPhlAn's read-based estimation of relative abundances enabled straightforward integration of multiple cohorts sequenced with different technologies and depths; specifically, to comprehensively characterize the asymptomatic human gut microbiota, we combined 224 fecal samples (>17 million reads) from the HMP[3] and the Metagenomics of the Human Intestinal Tract (MetaHIT) project[4], the two largest gut metagenomic collections available. MetaPhlAn detected 102 species present at least once at >0.5% abundance (**Fig. 3a**) consistently among different markers for the same clade (**Supplementary Fig. 7**). The MetaHIT project has previously characterized gut microbiomes as arising from three distinct and stable microbiome types (enterotypes[11]), and we investigated this hypothesis by hierarchically clustering the 224 samples separately at the genus and species levels (**Fig. 3b**,**c**). In some cases, enterotype-like discrete prevalence patterns were readily apparent, with the genus *Prevotella* being the most striking example and *Butyrivibrio* showing similar behavior but for fewer samples. Conversely, many samples were characterized by high fractions of *Bacteroides* resembling enterotype 1 (ref. 11), but this genus' overall relative abundance formed a continuum across samples, as did those of several other genera including *Eubacterium* and *Alistipes*.

MetaPhlAn's estimates of species-level abundance allowed us to refine this investigation (**Fig. 3c**). Although enterotype 2 remained clearly identifiable, the *Bacteroides* were diversified in a manner quite similar to *Lactobacilli* in the vaginal microbiota, albeit with more species and less exclusive dominance. This suggests the existence of more complex community patterns than those captured by the proposed genus-level enterotypes (**Supplementary Note 2**). The integrated data set also furnished the opportunity to investigate differences between independent and geographically unrelated healthy Western-diet populations (**Fig. 3** and **Supplementary Note 3**). Overall, this analysis showed that MetaPhlAn is effective in processing very large metagenomic data sets with different short read lengths at high taxonomic resolution, enabling meta-analyses difficult to achieve using other technologies.

Shotgun metagenomic data are rapidly decreasing in cost to a per-sample level comparable to that of 16S gene surveys. Community-wide sequence reads already provide unique insights into gene function, metabolism and polymorphisms that are unavailable from individual marker genes. By enabling efficient, high-resolution taxonomic profiling in such data, MetaPhlAn provides a further advantage over 16S rRNA–based investigations, which can be difficult to extend past a genus level of resolution. Metagenomic sequencing also provides better statistical support (~10^8 reads per sample) than 16S pyrosequencing approaches (typically <10^4 reads per sample), and the sequencing protocols do not require potentially biased amplification steps. Finally, the MetaPhlAn database of clade-specific markers is constructed by a fully automated computational pipeline, which will allow improved accuracy as additional microbial genomes become available and will improve support for gene markers' intraclade universality and interclade uniqueness.

METHODS

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

Note: Supplementary information is available in the online version of the paper.

ACKNOWLEDGMENTS

We would like to thank F. Stewart and E. DeLong for their helpful input during this study; D. Gevers, S. Sykes and K. Huang for their feedback on the methodology and J. Reyes and G. Weingart for their assistance with the implementation. This work was supported by US National Institutes of Health grant 1R01HG005969 and National Science Foundation grant DBI-1053486 to C.H.

AUTHOR CONTRIBUTIONS

N.S., A.B., O.J. and C.H. conceived the method; N.S. implemented the software; N.S. and C.H. performed the experiments; N.S., L.W., V.N. and C.H. analyzed the data; and N.S. and C.H. wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/doifinder/10.1038/nmeth.2066. Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. DeLong, E.F. *Nat. Rev. Microbiol.* **3**, 459–469 (2005).
2. Daniel, R. *Nat. Rev. Microbiol.* **3**, 470–478 (2005).
3. The Human Microbiome Project Consortium. *Nature* advance online publication, doi:10.1038/nature11209 (14 June 2012).
4. Qin, J. *et al. Nature* **464**, 59–65 (2010).
5. Ravel, J. *et al. Proc. Natl. Acad. Sci. USA* **108**, 4680–4687 (2011).
6. Veiga, P. *et al. Proc. Natl. Acad. Sci. USA* **107**, 18132–18137 (2010).
7. Turnbaugh, P.J. *et al. Nature* **457**, 480–484 (2009).
8. Markowitz, V.M. *et al. Nucleic Acids Res.* **38**, D382–D390 (2010).
9. Fredricks, D.N., Fiedler, T.L. & Marrazzo, J.M. *N. Engl. J. Med.* **353**, 1899–1911 (2005).
10. Stewart, F.J., Ulloa, O. & DeLong, E.F. *Environ. Microbiol.* **14**, 23–40 (2012).
11. Arumugam, M. *et al. Nature* **473**, 174–180 (2011).
12. Brady, A. & Salzberg, S. *Nat. Methods* **8**, 367 (2011).
13. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. *J. Mol. Biol.* **215**, 403–410 (1990).
14. Parks, D.H., MacDonald, N. & Beiko, R. *BMC Bioinformatics* **12**, 328 (2011).
15. Rosen, G.L., Reichenberger, E.R. & Rosenfeld, A.M. *Bioinformatics* **27**, 127–129 (2011).

ONLINE METHODS

Implementation and availability. An open-source implementation of MetaPhlAn, the corresponding core-gene catalog and species-level results for the analyzed gut and vaginal metagenomes are available online at http://huttenhower.sph.harvard.edu/metaphlan/. The taxonomic profiles of the 691 HMP shotgun samples from 18 different body habitats can be accessed at http://hmpdacc.org/HMSMCP/.

MetaPhlAn pipeline, database and classifier overview. MetaPhlAn estimates microbial relative abundances by mapping metagenomic reads against a catalog of clade-specific marker sequences currently spanning the bacterial and archaeal phylogenies. Clades are groups of genomes (organisms) that can be as specific as species or as broad as phyla. Clade-specific markers are coding sequences (CDSs) that satisfy the conditions of (i) being strongly conserved within the clade's genomes and (ii) not possessing substantial local similarity with any sequence outside the clade. The definition of such markers is somewhat sensitive to the availability of sequenced genomes, especially point (i), because a gene can be present in all available sequenced genomes in a clade but missing from some yet-to-be-sequenced strains. However, we did demonstrate in several experiments that the definition is effective even with only partial coverage of reference genomes (**Supplementary Tables 1** and **2**).

Thus, starting from the 2,887 genomes currently available from IMG (2011 July)[8], we identified more than 2 million potential markers meeting this level of stringency and allowing for sequencing and annotation errors. We then selected the subset of 400,141 genes most representative of each taxonomic unit. This selection employed a two-step computational pipeline that performed an intraclade CDS clustering and then an extraclade sequence uniqueness assessment; the method was based loosely on our previous system for detecting core genes[16]. The resulting marker catalog spans 1,221 species with an average of 231 (s.d. 107) markers per species and has only 12 species (<1% of the total) with fewer than 15 markers. Nine of these are species in the *Brucella* genus, which is known to have high genotypic homogeneity and unclear taxonomy[17]. Even these rare cases in which species are not completely characterized by specific markers, however, are themselves well covered at higher taxonomic levels. For example, the *Brucella* genus (as opposed to its constituent species) has 346 markers covering almost 300,000 nucleotides (nt); in total, 375 of 652 genera had >250 markers, as did 80 of 278 families and 22 of 130 orders. These markers were in addition to those for corresponding subclades, thus allowing MetaPhlAn to recover relative abundances within broader clades even in the absence of sequenced genomes for all organisms in a community. The generation of this catalog of marker genes is an offline procedure that we perform regularly as a relevant set of newly sequenced microbial genomes is available, and the catalog is downloaded automatically with the associated classifier.

The MetaPhlAn classifier compares metagenomic reads against this precomputed marker catalog using nucleotide BLAST searches[13] in order to provide clade abundances for one or more sequenced metagenomes. This achieves an approximately two-order-of-magnitude speedup compared to applying BLAST to the full catalog of microbial genomes, primarily because of the reduced size of our reference catalog. To estimate each clade's relative abundance in terms of cell counts (that is, whole-genome counts, rather than single-read counts), the MetaPhlAn classifier normalizes the total number of reads in each clade by the nucleotide length of its markers.

Genomic data sources and preprocessing. The input genomic data for training MetaPhlAn consists of a catalog of finished and draft genomes and the corresponding functionally unannotated CDS calls. Specifically, the system automatically retrieves the nucleotide sequences of bacterial and archaeal genomes from the IMG system[8] (2,887 total for this publication, 1,449 finished and 1,438 draft). These are screened for minimum length (>50,000 nt), minimum number of CDSs (>50), minimum percentage of coding genome (>75%) and taxonomic label (at or beneath the genus level). A total of 2,834 genomes (2,727 bacteria and 107 archaea) pass this quality-control screening, and after a minimal manual curation of the corresponding taxonomy, they span 2 domains, 33 phyla, 66 classes, 130 orders, 278 families, 652 genera and 1,221 species for a total of 2,383 taxonomic clades. The raw nucleotide sequences, the CDS calls (the list of gene or hypothetical gene start and end nucleotide positions) and the taxonomic classification of the genomes are the three only inputs for our system and can be retrieved from any available genomic database. The current version of MetaPhlAn is based on the genomic data contained in IMG version 3.4 (2011 July); different and future versions of the IMG database or other genomic databases can be used with minimal configuration changes in the pipeline.

Identification of clade-specific core genes. The offline identification of clade-specific gene markers was developed from a previous method for identifying core genes at multiple taxonomic levels[16]. As a first step, each genome is independently processed and all its CDSs are clustered to obtain a bag-of-genes representation of the genome; only one representative (called the seed) for clusters with more than one CDS is retained. The second step of the procedure considers the seeds of all the genomes in a given species and identifies the seeds with sequence homology in all genomes using UCLUST[18] with a nucleotide identity threshold of 75%. Several refinements are introduced to make the core gene identification procedure robust with respect to the missing genomic region in the draft genomes and to errors and noise in the original CDS call and in the taxonomic assignment of the genomes. First, the clade-specific gene families are aligned against the raw genomes and when new high-confidence matches are found (75% identity), CDS calls are modified accordingly and the gene family is expanded. Second, we relaxed the definition of core genes: given an empirically estimated probability that a CDS is missing from a genome for experimental rather than biological reasons (called the misannotation rate and set to 0.1), we computed the posterior probability density function (using the beta distribution) to model the presence or absence of a gene family in the genomes, and we rejected a CDS as core gene if its homology pattern in the clade deviated from the expected baseline misannotation error rate with confidence greater than 95%.

The procedure is recursively applied to higher taxonomic levels (from genera to phyla) considering all seeds of CDS families in direct-descendant clades that possess a sufficiently large core confidence score to potentially satisfy the 95% confidence score of the parent clade. Notice that some genomes are taxonomically assigned

by IMG directly to genus level, and thus they are not inserted in the species-level core-gene identification process. Differently from the standard core-gene definition[16], we are considering core-gene families that are conserved within their clade but not within their parent clades, as the purpose here is to identify genes that characterize each microbial clade. These alignment parameters are tuned here to identify nucleotide representatives along all CDS lengths and thus do not necessary capture broader functional categories.

Screening of core genes for unique taxonomic marker genes. Each clade-specific core-gene family is aligned against all the raw genomes in order to identify those core genes that are not uniquely present in the clade and exclude them from the candidate marker list. In this way, the gene catalog eliminates the genes that cannot be unequivocally associated with a clade; this is the case, for example, for the 16S rRNA gene and all other genes ubiquitous within the microbial taxonomy[19,20] as well as for genes likely to have been horizontally transferred. For each core-gene family, a uniqueness index is estimated that is inversely proportional to the number of genomes containing a local homology with the seed of the core-gene family. Multiple sets of sequence-alignment parameters (*e* values from 1×10^{-6} to 1×10^{-50}, minimum alignment length 20–50 nt) have been applied to this procedure in order to rank the uniqueness of a gene with the stringency of the homology match and consequently drive the preference of inclusion in the final reduced marker catalog.

Multicopy genes (including paralogs detected above) were excluded at this point for all clades with at least 300 markers because single-copy genes provide a more direct quantitative estimate of genome abundance when available. Interestingly, the widely used 16S rRNA marker gene is itself a multicopy gene with a strain-specific number of copies and, to the best of our knowledge, is typically used for community profiling without adjusting for the number of copies. Finally, a heuristic considering the uniqueness index, the confidence of the core genes, the different stringency parameters and the length of the genes is applied to obtain a consistent maximum number of core genes (350) for each medium- to low-level taxonomic clade; it maintains, however, the same baseline cutoff thresholds in all clades. This offline automatic pipeline produced the unique taxonomic markers that populate the database of 400,141 CDSs currently used by the taxonomic classifier.

Mapping of metagenomes to the marker gene catalog and estimation of organismal relative abundances. The selection of the marker genes described above is relatively computationally intensive (typically requiring several CPU-days), but it needs to be performed only once when the set of reference genomes is modified, usually because of the addition of newly sequenced genomes. MetaPhlAn users do not need to perform this task, as we provide the most updated reference marker set. The algorithm for estimating the taxonomic composition of the sample using the marker catalog is available for download (see "Implementation and availability") and is described below.

MetaPhlAn first needs to compare each metagenomic read from a sample with the catalog of marker genes for finding sequence matches. Although the software performs this operation internally using the NCBI program BLASTn (default threshold on the e-value of 1×10^{-6}), any other sequence-matching method (including accelerated methods like MapX (Real Time Genomics), MBLASTX (Multicoreware) and USEARCH[18] can be used by providing MetaPhlAn with the generated alignment results in a tabular format. MetaPhlAn also provides an option to perform the alignment operation with multiple processors, which involves internally splitting the input reads into multiple independent sets that can be aligned using different CPUs. The classifier then assigns read counts to the microbial clades according to the alignment prediction; in the rare case of multiple matches with markers from different clades, only the best hit is considered. For each clade, read counts can be assigned directly using the clade-specific markers or indirectly by considering the reads assigned to all direct descendants. Total read counts are estimated with the direct approach if the clade has a sufficient total size of the marker set (default 10,000 nt); otherwise the indirect approach is preferred. Moreover, for all clades except the leaves of the taxonomic tree (species), the two estimations are compared, and an 'unclassified' subclade is added when the clade-specific read count is larger than the sum of descendant counts; the difference between the two estimations is assigned to the new descendant. Relative abundances are estimated by weighting read counts assigned using the direct method with the total nucleotide size of all the markers in the clade and normalizing by the sum of all directly estimated weighted read counts.

Synthetic metagenomes and comparison with existing approaches. The generation of the ten synthetic communities used in our evaluation was inspired by Mavromatis *et al.*[21] and consisted of two high-complexity evenly distributed metagenomes (HC1 and HC2) and eight low-complexity log-normally distributed metagenomes (LC1–LC8). The 100 genomes for each HC community and the 25 genomes for each LC community were randomly chosen from the Kyoto Encyclopedia of Genes and Genomes (KEGG) v54[22] genome catalog. Reads 100 nt long were sampled from the selected genomes using a MAQ[23] error model constructed using real Illumina reads. The number of reads sampled from each genome was proportional to the size of the genome, and the distribution of the genome abundances was even for HC1 and HC2 and log-normal for LC1–LC8. The resulting synthetic metagenomes are available at the MetaPhlAn website.

We also assessed the accuracy with which MetaPhlAn detects novel organisms (those without closely related sequenced genomes) as compared to existing approaches. We selected 20 genomes from the Reference Sequence database (RefSeq)[24] versions 46–51 that belong to species not represented in the IMG repository. Although these 20 genomes may be included in IMG in the near future, they were not included in MetaPhlAn's current core-gene inputs because of incomplete sequence validation, unverified gene calls and annotations or missing database depositions. Fortunately, this made them appropriate candidates for held-out test data. For each of these 20 genomes, 25,000 short reads were generated using the same strategy adopted for creating synthetic metagenomes to assess MetaPhlAn's classification of novel taxa.

We compared MetaPhlAn's performance on these data (**Supplementary Tables 1** and **2**) and on the synthetic data above (**Fig. 1** and **Supplementary Figs. 2–4**) relative to several existing classes of methods for taxonomic classification of

doi:10.1038/nmeth.2066

metagenomic reads that are based on alignment and composition. Alignment-based methods rely on comparing reads to a set of reference genomes and include NCBI BLAST[13] and other BLAST-based approaches such as MEtaGenome ANalyzer (MEGAN)[25,26], Multiple Taxonomic Ranks base clustering (MTR)[27], PArsimony-based Phylogeny-Aware short Read Alignment (PaPaRa)[28] and CARMA3 (ref. 29). Composition-based methods exploit features such as *k*-mer frequency and include PhyloPythia/PhyloPythiaS[30,31], Phymm[12,32], the naive Bayesian classifier (NBC)[15,33], RAIphy[34] and MetaCluster 3.0 (ref. 35). These two types of approaches have also been integrated in hybrid methods such as PhymmBL[12,32] and rapid identification of taxonomic assignments (RITA)[14], which generally result in improved accuracy relative to either method alone. We thus compared MetaPhlAn to the six most popular methods representative of the current state of the art: Phymm, PhymmBL and the BLAST-based component of PhymmBL have been installed locally on the same machine on which MetaPhlAn ran, whereas RITA and NBC each provides a convenient web server, as does PhyloPythiaS.

Description of novel Human Microbiome Project and MetaHIT metagenome analyses. The human metagenomic data we analyzed with MetaPhlAn are available in public repositories. The 51 vaginal and 139 fecal samples from the HMP[3] can be accessed at http://hmpdacc.org/HMASM/, from which we used the 'WGS' reads (not the 'PGA' assemblies). The 16S rRNA profiling of HMP vaginal microbiomes used for comparison (**Fig. 2**) was performed from the mothur pipeline[36] data followed by phylotype profiling with the Ribosomal Database Project (RDP) classifier[37] (http://hmpdacc.org/HMMCP/). The 85 fecal samples from MetaHIT[4] were downloaded from the European Nucleotide Archive (http://www.ebi.ac.uk/ena/, study accession number ERP000108), and the marine minimum oxygen zone samples[10] are available in the NCBI Sequence Read Archive under accession number SRA023632.

The full tables of results for the vaginal and gut metagenomes are available at http://huttenhower.sph.harvard.edu/metaphlan/.

16. Segata, N. & Huttenhower, C. *PLoS ONE* **6**, e24704 (2011).
17. Bohlin, J. *et al. BMC Evol. Biol.* **10**, 249 (2010).
18. Edgar, R.C. *Bioinformatics* **26**, 2460–2461 (2010).
19. Wu, M. & Eisen, J.A. *Genome Biol.* **9**, R151 (2008).
20. Ciccarelli, F.D. *et al. Science* **311**, 1283–1287 (2006).
21. Mavromatis, K. *et al. Nat. Methods* **4**, 495–500 (2007).
22. Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M. & Hirakawa, M. *Nucleic Acids Res.* **38**, D355–D360 (2010).
23. Li, H., Ruan, J. & Durbin, R. *Genome Res.* **18**, 1851–1858 (2008).
24. Pruitt, K.D., Tatusova, T., Klimke, W. & Maglott, D.R. *Nucleic Acids Res.* **37**, D32–D36 (2009).
25. Huson, D.H., Auch, A.F., Qi, J. & Schuster, S.C. *Genome Res.* **17**, 377–386 (2007).
26. Huson, D.H., Mitra, S., Ruscheweyh, H.J., Weber, N. & Schuster, S.C. *Genome Res.* **21**, 1552–1560 (2011).
27. Gori, F., Folino, G., Jetten, M.S.M. & Marchiori, E. *Bioinformatics* **27**, 196–203 (2011).
28. Berger, S.A. & Stamatakis, A. *Bioinformatics* **27**, 2068–2075 (2011).
29. Gerlach, W. & Stoye, J. *Nucleic Acids Res.* **39**, e91 (2011).
30. McHardy, A.C., Rigoutsos, I., Hugenholtz, P., Tsirigos, A. & Martin, H.G. *Nat. Methods* **4**, 63–72 (2007).
31. Patil, K.R. *et al. Nat. Methods* **8**, 191–192 (2011).
32. Brady, A. & Salzberg, S.L. *Nat. Methods* **6**, 673–676 (2009).
33. Rosen, G., Garbarine, E., Caseiro, D., Polikar, R. & Sokhansanj, B. *Adv. Bioinformatics* **2008**, 205969 (2008).
34. Nalbantoglu, O.U., Way, S.F., Hinrichs, S.H. & Sayood, K. *BMC Bioinformatics* **12**, 41 (2011).
35. Leung, H.C. *et al. Bioinformatics* **27**, 1489–1495 (2011).
36. Schloss, P.D. *et al. Appl. Environ. Microbiol.* **75**, 7537–7541 (2009).
37. Cole, J.R. *et al. Nucleic Acids Res.* **37**, D141–D145 (2009).