Published online 19 March 2018

Nucleic Acids Research, 2018, Vol. 46, No. 10 ***e59***
doi: 10.1093/nar/gky174

GraftM: a tool for scalable, phylogenetically informed classification of genes within metagenomes

Joel A. Boyd†, Ben J. Woodcroft† and Gene W. Tyson*

Australian Centre for Ecogenomics, School of Chemistry and Molecular Biosciences, University of Queensland, St Lucia 4072, Queensland, Australia

Received November 13, 2017; Revised February 20, 2018; Editorial Decision February 23, 2018; Accepted March 06, 2018

ABSTRACT

Large-scale metagenomic datasets enable the recovery of hundreds of population genomes from environmental samples. However, these genomes do not typically represent the full diversity of complex microbial communities. Gene-centric approaches can be used to gain a comprehensive view of diversity by examining each read independently, but traditional pairwise comparison approaches typically over-classify taxonomy and scale poorly with increasing metagenome and database sizes. Here we introduce GraftM, a tool that uses gene specific packages to rapidly identify gene families in metagenomic data using hidden Markov models (HMMs) or DIAMOND databases, and classifies these sequences using placement into pre-constructed gene trees. The speed and accuracy of GraftM was benchmarked with *in silico* and *in vitro* mock communities using taxonomic markers, and was found to have higher accuracy at the family level with a processing time 2.0–3.7× faster than currently available software. Exploration of a wetland metagenome using 16S rRNA- and methyl-coenzyme M reductase (McrA)-specific gpkgs revealed taxonomic and functional shifts across a depth gradient. Analysis of the NCBI nr database using the McrA gpkg allowed the detection of novel sequences belonging to phylum-level lineages. A growing collection of gpkgs is available online (https://github.com/geronimp/graftM_gpkgs), where curated packages can be uploaded and exchanged.

INTRODUCTION

Microorganisms play critical roles in environmental and host-associated systems, yet our understanding of their phylogenetic and functional diversity is far from complete. Exploration of microbial diversity has been greatly accelerated by metagenomic approaches that bypass traditional culture-based bottlenecks by providing direct access to the genomic information of microorganisms within an environmental sample (1). In recent years, improvements in sequencing technology and bioinformatic tools for handling metagenomic data, including novel genome binning approaches (2–4), have enabled the recovery of population genomes from a wide range of environments (5,6). However, these approaches typically only recover members of a community that have sufficient coverage, minimal population heterogeneity, few sequence repeats and distinct genome nucleotide composition (7). As a result, current genome-centric approaches do not capture the full phylogenetic and functional diversity of a microbial community, particularly in complex environments (8,9).

Gene-centric analysis, the identification of genes in unassembled metagenomic data, directly assesses diversity within a microbial community by examining each unassembled read independently, providing access to populations that are refractory to assembly. Widely used tools for gene-centric analysis such as MEGAN (10), MG-RAST (11) and RITA (12) classify sequences using pairwise comparisons against a reference database with BLAST (13) or similar software. More recently, amino acid sequence comparison tools have been developed which are several orders of magnitude faster than BLAST (14,15), making pairwise-based approaches more computationally tractable for gene-centric analyses of metagenomes (16). While these approaches can accurately identify gene sequences if close relatives are present in the database, divergent sequences from lineages without representation are typically poorly classified. Further, the speed of pairwise comparison depends on the size of the database and as these databases grow, speed will again become a limiting factor for these tools. A number of other tools that exploit tetranucleotide frequencies provide alternate methods for community profiling (7), but most these over-classify novel sequences and cannot profile a community in the context of a single gene of interest.

Hidden Markov models (HMMs) are an alternative approach to gene-centric analysis (17). HMMs statistically describe groups of related sequences, often entire gene fami-

*To whom correspondence should be addressed. Email: g.tyson@uq.edu.au
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

lies, and can sensitively detect remote homology in sequence searches. However, HMM searches do not taxonomically classify reads and are therefore often coupled to phylogenetic approaches (18–21). Phylogenetic approaches benefit from robust models of evolution, do not rely on cut-off scores such as *E*-values, and provide more accurate classifications than homology-based methods (20,22).

Phylogenetic classification can be based on trees constructed *de novo* (18,19), but this approach is computationally costly and comparing independently analysed metagenomes is difficult because tree topologies may change between samples. In contrast, maximum likelihood placement of reads into fixed trees (23–25) scales linearly with input data size and allows robust statistical comparisons to be made between samples (20,26). The gene-centric tool metAnnotate (21) pairs a HMM-based search step with classification using pplacer (23). For each sample, metAnnotate generates a *de novo* phylogenetic tree using sequences from the NCBI nr database for classification. Currently, metAnnotate is restricted to NCBI as a source of reference sequences to create gene trees and does not allow manual expert curation. After placement using pplacer, metAnnotate assigns taxonomy to each read by taking the most common classification within a sub-tree (21). PhyloSift (20) also uses pplacer to phylogenetically classify sequences, but has limited search speed as it utilizes the pairwise comparison tool LAST (27) to identify gene sequences.

Here, we present GraftM, an open source tool for phylogenetically informed classification of metagenomic sequences. GraftM uses the rapid open reading frame finder OrfM (28) to translate nucleotide sequences, HMMs or DIAMOND-based pairwise comparison to search for target gene-families, and phylogenetic placement using pplacer for classification. GraftM provides the tools required for creating new gene specific GraftM packages (gpkgs) for bacterial, archaeal and eukaryotic gene families or subfamilies, enabling reproducible phylogenetic analysis of metagenomic data. Using conserved and metabolic markers we demonstrate that GraftM outperforms similar tools in terms of runtime, search sensitivity and classification accuracy.

MATERIALS AND METHODS

Implementation and dependencies

GraftM is implemented in Python 2.7 and is dependent several Python packages, as well as on OrfM (28), pplacer (23), DIAMOND (14), HMMER (29), fxtract (github.com/ctSkennerton/fxtract), MAFFT (30), Krona (31) and FastTree (32).

Creation of GraftM packages (gpkgs) for benchmarking

16S rRNA gpkg. A 16S rRNA gpkg was created from the 2013/08 public release of the Greengenes database (33). GraftM create was run using these sequences and the taxonomy-decorated phylogenetic tree for the 97% nucleotide identity representative OTU set (ftp://greengenes.microbio.me/greengenes_release/gg_13_8_otus). *Ribosomal protein gpkgs.* Gpkgs were created for ribosomal proteins by starting with the set of HMMs included with PhyloSift (20). These HMMs were used to search with HMMER, using an E-value cutoff of 1e–40, against the set of finished and permanent draft proteomes from the IMG (34) that were >90% complete and <5% contaminated according to CheckM v1.0.5 (35). To prevent contaminated genomes introducing error into the taxonomic annotations, only those genomes where a single hit was found were utilized. To limit the effect of taxonomic bias toward lineages with a greater number of sequenced genomes, only a single protein from each species (one representative per species, using a type strain where possible and including all those without species level taxonomic classification) were used. Proteomes were searched using GraftM graft using default parameters, after which 15 ribosomal markers were determined to be single copy on the basis of their being detected as having a single hit in >5900 of the 6215 genomes. GraftM packages for the 15 protein-coding genes were generated with GraftM create using those sequences found in single copy, a previously generated HMM and the corresponding IMG taxonomy for each genome. *Functional and taxonomic McrA gpkgs*. Two gpkgs were constructed for the alpha subunit of the methyl coenzyme M reductase (*mcra*) gene. Amino acid sequences for the McrA protein family and paralogous MrtA sequences were sourced from IMG (February 2014) using the BLASTP tool provided online. Spurious hit sequences were removed by manual inspection. Genes for the Bathyarchaeotal (36) and Vertataearchaeotal (37) orthologues were sourced from NCBI. The first taxonomy-annotated gpkg was created using the default GraftM create pipeline using the sequences and their associated genome taxonomy. The second was created by re-decorating the McrA tree with functional, rather than taxonomic information. This second tree was annotated according to their substrate utilization: acetoclastic (from acetate) comprised of the order *Methanosarcinales*; hydrogenotrophic (from hydrogen, carbon dioxide and/or formate), comprised of the Methanomicrobiales, Methanocellales, Methanococcales and Methanobacteriales; methylotrophic (from methylated compounds) comprised of the Methanomassiliicoccales, Methanofastidiales and Vertataearchaeota. Lineages within the Bathyarchaeota were recently found to encode *mcra*, though their metabolism is not yet confirmed. These sequences were included in the gpkg, but left unannotated. The McrA tree was curated with these functional groupings using ARB (38), with the exception of the *Methanosarcina* which are thought to be capable of producing methane from all three substrate groups (39). The Methanosarcinaceae were annotated as a clade separate to the exclusively acetoclastic Methanosaetaceae.

Data sources, and sampling and extraction of metagenomic DNA

in vitro mock metagenomes. To validate the search and classification steps of GraftM, three biological replicates of a mock community composed of the genomic sequences of 54 microorganisms spanning 41 taxonomic families were examined (40). The expected community composition was known since construction of the mock community involved pooling DNA extracted from a separate culture of

each community member. Community profiles generated by GraftM were benchmarked by comparison with this gold standard. Details on the preparation and sequencing of the metagenome libraries can be found in *Rinke et al.* (40). *in silico mock metagenomes.* Error-free, synthetic 150 bp paired-end reads were generated for each lineage of the *in vitro* mock at 20× coverage using a synthetic paired end read simulator sammy.pl (github.com/minillinim/sammy/blob/master/sammy.pl) using default parameters. *Environmental metagenomes.* Environmental soil was sampled from a site of thawing permafrost in Abisko National Park, northern Sweden in June 2012 (41). Three subsamples were taken at different depths: shallow (0 cm), middle (6.5 cm) and deep (12.5 cm). To demonstrate the expand_search function in GraftM, a further 19 metagenomes from a Sphagnum dominated bog were selected to analyse. Details of the preparation and sequencing of the metagenome libraries are detailed in Woodcroft *et al.* 2017 (submitted). *Symbiodinium metagenomes.* Three replicate *Acropora tenuis* samples were collected from a coral reef at Orpheus Island at Pioneer Bay, Great Barrier Reef in 2016 and preserved in Glycerol-TE buffer. A single sample of *A. tenuis* sperm was taken following a spawning event in 2016. DNA from all samples were and extracted using MoBio Ultra Clean powerlyzer extraction kits, and prepared using Nextera. Sequencing was completed at the Australian Centre for Ecogenomics with an Illumina NextSeq. *NCBI nr database.* The NCBI nr database was downloaded from the NCBI FTP server (2.9×10^9 amino acids) in January, 2016.

Defining correct classification of in vitro and in silico mock metagenomes

When the *in silico* mock was created, the origin of each read pair was recorded and the correct classification of each read was defined as the gene loci from which it originated as determined using Prokka (42) version 1.11. A read was assigned to a particular gene if at least one base of the read overlapped that gene's coding region. Correct classifications for the *in vitro* mock metagenomes were determined using the same criteria after mapping the metagenome reads to the genomes of each isolate used in the mock using BWA-MEM v0.7.12 (43), taking the primary alignment only.

Analysis of *in silico*, *in vitro* mocks and environmental metagenomes with GraftM and metAnnotate

The 16S rRNA and 15 ribosomal protein gpkgs were used to search both the forward and reverse reads of each mock *in vitro* and *in silico* metagenomes using default parameters. HMMs from the 15 ribosomal protein packages were provided to metAnnotate (downloaded 19 November 2015), which was run using the 'phylogenetic' pipeline with default parameters. The translate command from biosquid 1.9g+cvs20050121 (eddylab.org/software.html) was used with the -a flag to translate the *in silico* mock due to a the extended runtime of metAnnotate when provided with nucleotide sequences. The output ORFs were provided to metAnnotate as amino acid sequence. The database used for metAnnotate was the NCBI RefSeq database (44) after filtering sequences containing B, Z, J or X amino acids. The command line version of metAnnotate is currently not designed for non-coding genes such as 16S rRNA so comparisons in this study were restricted to protein coding genes. Lineages with <2 representatives in the reference database were removed from the analysis. GraftM was run on the three environmental metagenomes using the 16S rRNA, McrA and functional McrA gpkgs with default parameters.

Application to NCBI's nr

GraftM graft was run on NCBI's nr database (downloaded January 2016) using both the default pipeline and the DIAMOND search and classify pipeline, both with an *E*-value cutoff of 1e–30. Phylogenetic analysis of unclassified McrA sequences recovered using the default pipeline was performed by aligning these sequences with those from the McrA gpkg and the novel sequences using MAFFT, and constructing a phylogenetic tree using FastTree v2.1.7, both with default parameters. The phylogenetic tree was bootstrapped using 100 replications.

RESULTS

GraftM has been developed for the rapid and accurate phylogenetic classification of genes of interest in large metagenomic datasets. The GraftM pipeline is divided into three steps, create, search and classify (Figure 1). In the create step, a GraftM package (gpkg) for a given gene is created using homologous full length nucleotide or amino acid sequences and their associated taxonomy to generate search databases and a reference phylogenetic tree. The search step employs the gpkg to identify sequences belonging to the gene family using either HMMs or DIAMOND. The classify step assigns taxonomy to these sequences by placing them into the reference phylogenetic tree.

GraftM search sensitivity and specificity

To investigate the sensitivity and specificity of the search step, *in vitro* and *in silico* benchmark datasets were created. The *in vitro* dataset was derived by sequencing DNA from 54 isolates pooled using a log-normal community structure (40), and the *in silico* mock was constructed using synthetic 150 bp paired-end reads generated from these reference genomes. Functional and taxonomic assignments for each read were defined by mapping to the reference genomes with BWA-MEM (43) for the *in vitro* mock, and using the known genomic position for the *in silico* mock.

To assess the sensitivity of GraftM nucleotide searches, implemented primarily for the detection of 16S rRNA genes, sequences from the Greengenes database (33) clustered at 97% nucleotide identity were used to create a 16S rRNA gene gpkg. Specific bacterial and archaeal search HMMs were included in the gpkg to improve its search sensitivity. When applied to the *in vitro* and *in silico* mock datasets, GraftM search identified 75.2% and 83.7% of 16S rRNA reads, respectively (Figure 2A). The majority of false negative reads were located at the start and end of the gene, where the reads did not fully overlap with the gene (Figure 2B). The false positive rate was low for both the *in vitro* and *in silico* mocks (0.02 false positives per true positive).

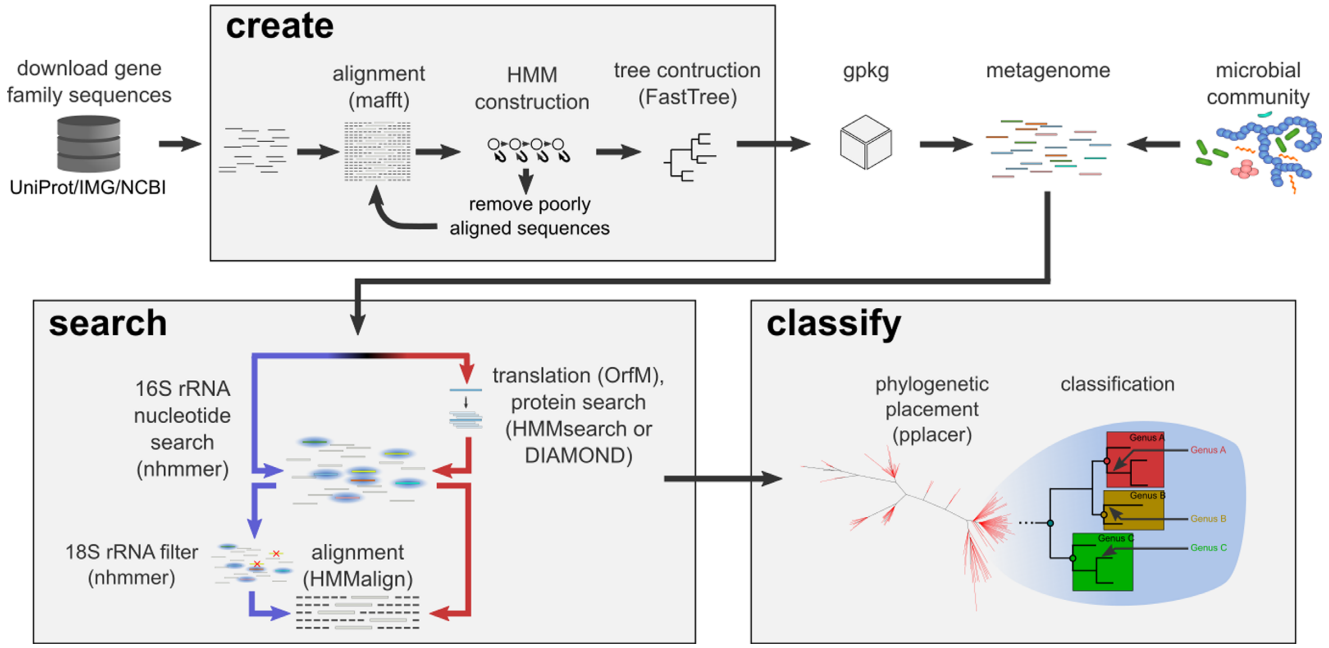

Figure 1. Schematic of the GraftM pipeline, outlining the create, search and classify stages. Within the search step, the red arrow indicates the amino acid pipeline and the blue arrow indicates the nucleic acid pipeline.

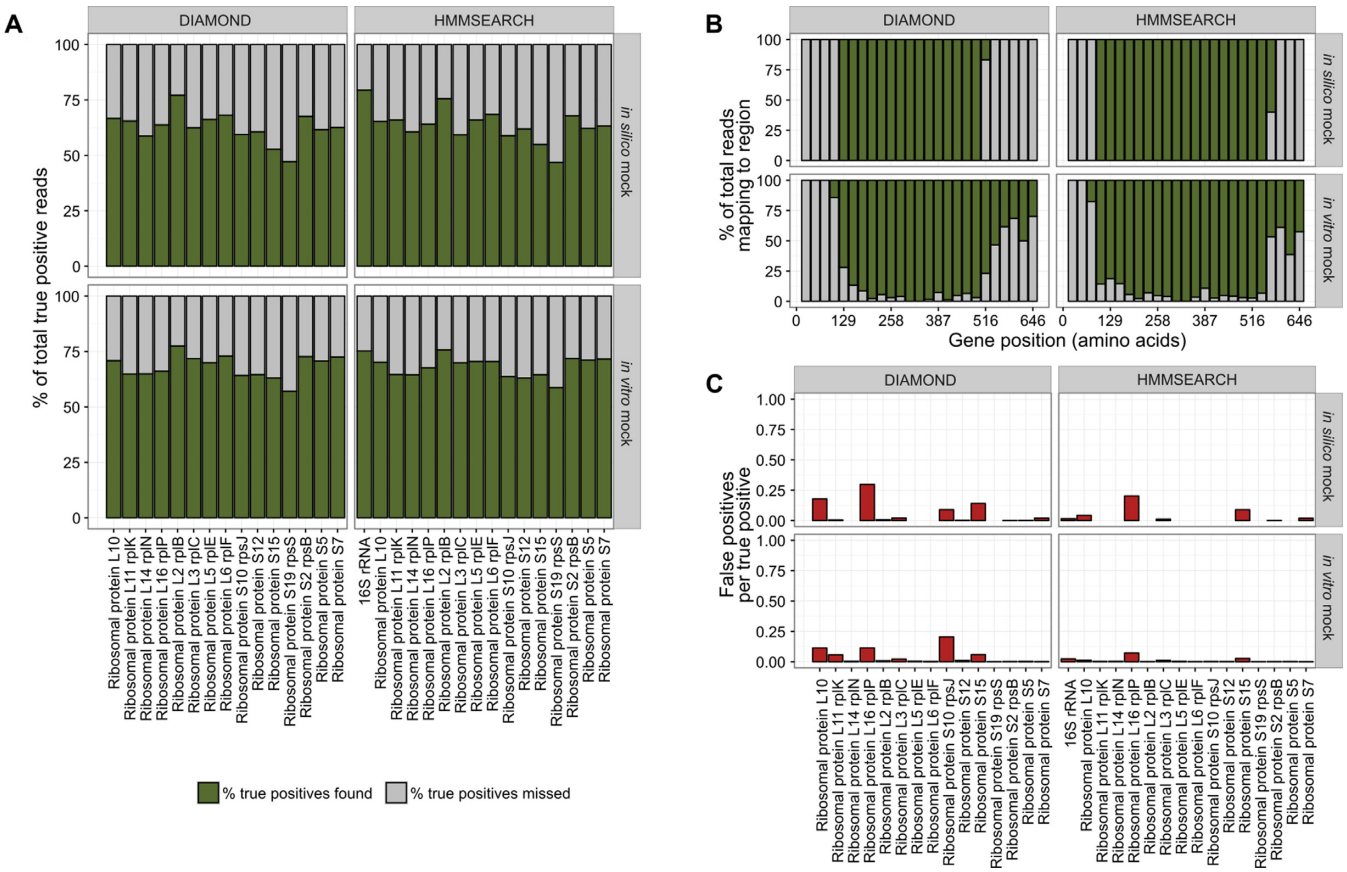

Figure 2. (**A**) Recovery of true positives by the HMMsearch and DIAMOND search methods on the *in silico* and *in vitro* mock datasets. (**B**) Distribution of true positive reads along the gene locus, using ribosomal protein S7 used as an example. (**C**) False positive rate measured as false positive per true positive for each ribosomal protein and the 16S rRNA gene.

GraftM amino acid searches for HMMSEARCH and DIAMOND were validated on the mock datasets with a set of 15 gpkgs for single copy ribosomal protein genes (Figure 2A and B). On average, HMMSEARCH had a similar true positive detection rate (*in silico* 73.3 ± 2.0%; *in vitro* 68.1 ± 1.2%; Figure 2A) to DIAMOND (*in silico* 72.8 ± 2.0%; *in vitro* 68.3 ± 2.0%; Figure 2B). However, DIAMOND returned slightly more false positives for all genes (*in silico* 0.05 ± 0.02 false positives per true positive; *in vitro* 0.04 ± 0.02; Figure 2B) than HMMSEARCH (*in silico* 0.05 ± 0.02; *in vitro* 0.02 ± 0.01; Figure 2A). For both search methods, the majority of false negative reads were located at the start and end of each gene (Figure 2C). Across the entirety of each gene, the false negative rate of GraftM was higher for the *in vitro* mock compared to the *in silico* mock due to sequencing error and read length heterogeneity (Supplementary Table S2). Given the greater false positive rate of DIAMOND, HMMSEARCH was selected as the default search method for GraftM.

GraftM classification accuracy on mock metagenomes

Reads recovered in the search step are assigned taxonomy based on their placement in a reference phylogenetic tree (Figure 1). The 16S rRNA gene gpkg accurately classified sequences at the family level (*in silico* 92.3%, *in vitro* 81.7%), but had lower sensitivity at the genus level (*in silico* 61.6%, *in vitro* 37.3%; Figure 3A). Reads that were unresolved at the genus level belonged to lineages that were paraphyletic in the reference tree (e.g. *Escherichia*, *Hydrogenobaculum*; Supplementary Table S3), or lineages with few sequenced representatives (e.g. *Methanopyrus*, *Comonomus*), and as such GraftM was unable to accurately classify these reads.

The protein classification accuracy of GraftM was compared with metAnnotate's phylogenetic pipeline (21) using 15 ribosomal proteins on the *in silico* and *in vitro* mocks. For the *in silico* mock, GraftM accurately classified a significantly higher percentage of reads at the phylum, class, order and family levels (phylum, class, order: t-test p-value <0.0001, family: <0.05). At the genus level, metAnnotate misclassified a higher proportion of reads (Figure 3), while GraftM was more conservative and did not classify these reads beyond the family level. GraftM was substantially faster than metAnnotate, running 3.6× faster (Supplementary Figure S6).

Analysis of the NCBI nr database using GraftM

To demonstrate the utility and scalability of GraftM for interrogating large datasets, a gpkg for a methanogenesis marker gene, methyl-coenzyme M reductase alpha subunit (McrA), was used to search and classify sequences from the NCBI nr database. The analysis ran in 18 minutes with the default pipeline (HMMSEARCH+pplacer; 14 888 sequences identified) and 34 minutes using the pairwise classification pipeline (DIAMOND blastp for search and classification; 15 010 sequences identified) on a 40 logical core system. The sequences identified by these approaches were highly consistent (98.9% of reads), with a small number of sequences detected by HMMSEARCH+pplacer only (20 sequences) or DIAMOND blastp only (142 sequences). Both pipelines identified a similar number of sequences belonging to the Methanomicrobiales (37%), Methanosarcinales (28%) and Methanomicrobiales (20%; Figure 4A). McrA sequences belonging to the recently discovered bathyarchaeotal and verstratearchaeotal clades were detected at low abundance using both pipelines (<2%; Figure 4A). The HMMSEARCH+pplacer pipeline identified divergent sequences including 144 sequences that were not classified at the domain level. These sequences were over-classified by the DIAMOND pipeline (Figure 4B; Supplementary Table S4). Construction of a *de novo* phylogenetic tree using these sequences revealed a novel clade basal to the Bathyarchaeota ('Red Sea' group) and two novel clades basal to the Verstratearchaeota ('Verstratearchaeota-like' groups; Figure 4C). The discovery of these novel McrA sequences demonstrates GraftM's ability to rapidly extract novel functional diversity from large public datasets.

Using GraftM on environmental metagenomes

To demonstrate GraftM's ability to assess the composition and diversity of genes of interest in complex metagenomes, the distribution and diversity of archaea and bacteria was examined in a fen soil core (shallow, middle, and deep) from Stordalen mire, Sweden (Woodcroft et al. submitted). The overall community structure, assessed using the 16S rRNA gene gpkg, showed the community was diverse with high relative abundances of Bacteroidales (6–7%), Actinomycetales (3–7%) and an unclassified order within the Chloroflexi (5–9%; Figure 5A). Deeper samples had higher abundances of Acidobacteriales, Myxococcales and the unclassified Chloroflexi order, while Actinomycetales decreased in abundance (Figure 5A).

Fen environments are a globally significant source of CH_4, a potent greenhouse gas produced by methanogens (45). The McrA gpkg was applied to assess the diversity of methanogens within the Stordalen mire fen samples. Methanogens belonging to the Methanobacteriales, Methanosarcinales, Methanomicrobiales, Methanocellales and Methanomassiliicoccales were detected across all depths. The highest number of McrA reads were detected at the mid-depth sample (0.0018% of sequenced reads), followed by the deep and shallow samples (0.0009% and 0.0001%, respectively). The relative abundance of the acetoclastic methanogens belonging to the genus *Methanosaeta* remained relatively stable throughout the soil column (17.3% - 21.0% of total McrA reads), but different hydrogenotrophic orders were dominant at the surface relative to the middle and deep samples (shallow = Methanobacteriales, 38.6%; middle, deep = Methanomicrobiales, 37.7–27.9%; Figure 5B). Alternate annotation of the McrA gpkg to link the phylogeny to acetoclastic, hydrogenotrophic and methylotrophic pathways allowed shifts in the primary mode of methanogenesis to be examined (Figure 5C). Despite the phylogenetic shifts in the methanogenic community, hydrogenotrophic methanogenesis was dominant at all depths (61.1–64.7% of McrA reads). Methylotrophic methanogenesis was elevated in the mid-depth sample (10.3%) relative to the shallow and deep samples (∼5.1%).

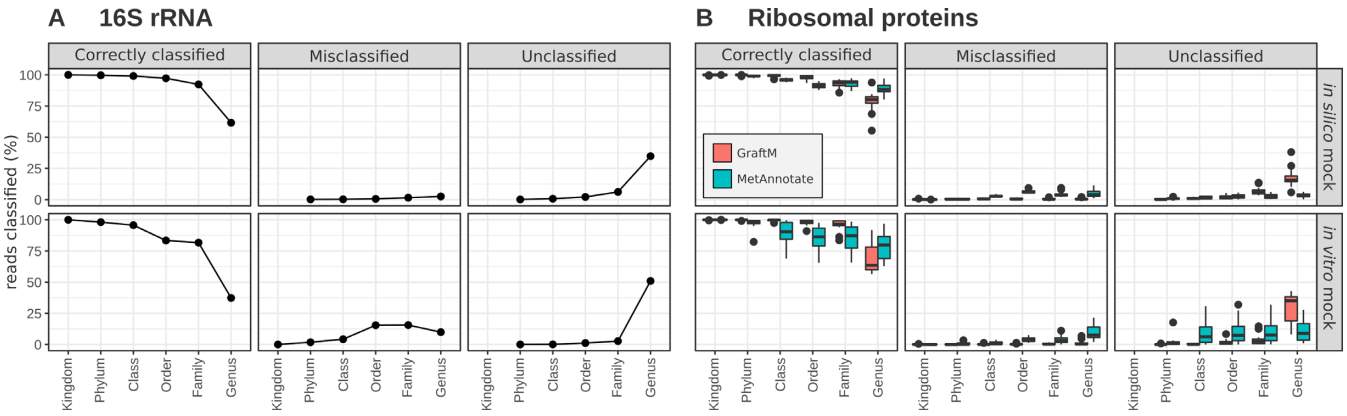

Figure 3. Classification accuracy of 150 bp reads from *in silico* and *in vitro* mock. (**A**) Accuracy of phylogenetic classification using the 16S rRNA gpkg. (**B**) Accuracy of phylogenetic classification using 15 ribosomal protein gpkgs for GraftM and metAnnotate.

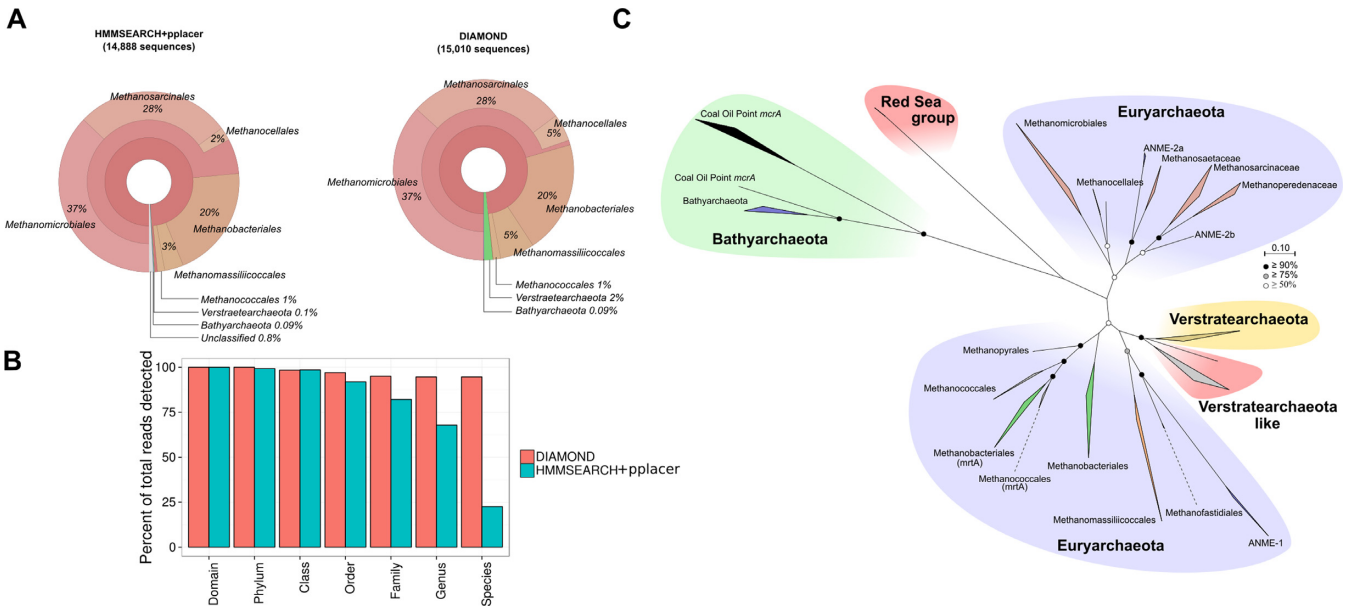

Figure 4. (**A**) Krona plots showing the order-level classification of partial and full-length McrA sequences identified by the HMMSEARCH+pplacer and DIAMOND pipelines from the NCBI's nr database. (**B**) Percentage of reads left classified at each taxonomic rank by each pipeline. (**C**) Maximum-likelihood tree of full length McrA sequences from the McrA gpkg and partial/full length McrA sequences identified in the NCBI nr database that were unclassified by GraftM. Support values from 100 bootstrap replicates of ≥50% are indicated as white circles, ≥75% as gray circles and ≥90% as black circles. Clades are labeled according to the lowest common ancestor of the lineages within. Red clades indicate lineages originating from the NCBI nr database that were not classified to a phyla by GraftM.

DISCUSSION

The size of metagenomic datasets continues to increase as a result of improvements in sequencing technology, but analysis of these large datasets is hampered by computational processing time and resource requirements. Identifying and classifying genes of interest in metagenomes is typically performed using pairwise comparison tools such as BLAST (13). These tools only provide meaningful results if the metagenome contains sequences closely related to those in a reference database. However, as the size of metagenomes and reference databases grow these approaches will likely become computationally prohibitive. HMMs are an alternative and scalable method of identifying genes in metagenomes, but require a post-processing step to assign taxonomic or functional classifications. GraftM was developed to enable fast and accurate identification and classification of genes in metagenomes using HMM searches in combination with robust phylogenetic placement methods.

GraftM uses pre-prepared databases (gpkgs) that can be built using amino acid or nucleic acid reference sequences and their annotations, but also allows custom phylogenies, alignments, search HMMs or DIAMOND databases to be used. In contrast, metAnnotate is restricted to amino acid reference sequences and their taxonomy from NCBI, limiting its speed and reproducibility when analyzing multiple metagenomes, and preventing the use of annotations from expertly curated databases (e.g. GreenGenes, GTDB) or custom functional annotation (e.g. protein superfamily phylogenies). This approach also leaves analyses vulnerable

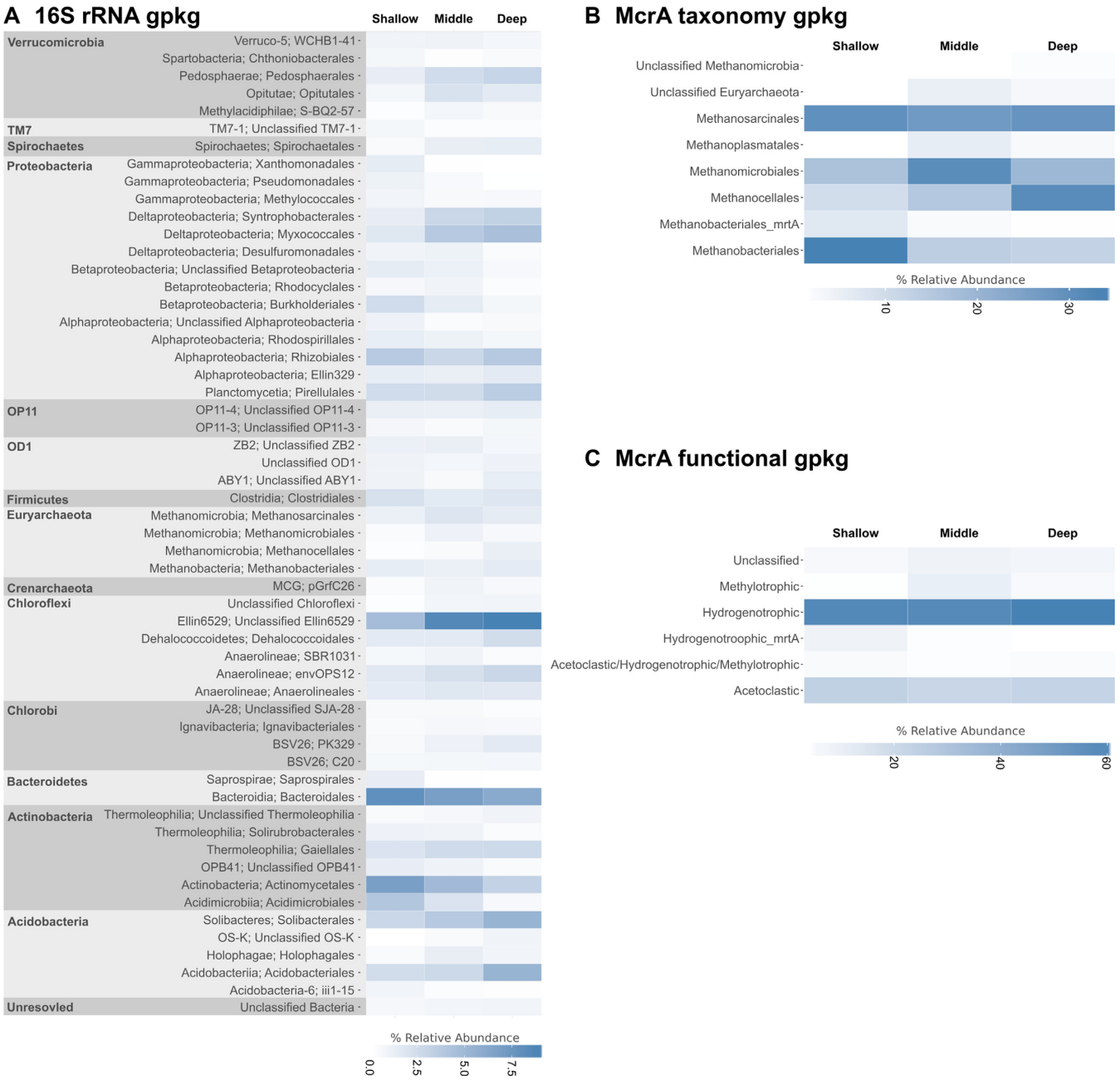

Figure 5. Shifts in microbial community structure across a permafrost active layer metagenome. (**A**) A heatmap using the log10 transformed relative abundance of the microbial community, clustered at the order level using the 16S rRNA gene gpkg. (**B**) Taxonomic composition as determined by the McrA gpkg. (**C**) Composition of the community as defined by manually refined version of the McrA tree, annotated with main methanogenic substrates used by each lineage.

to the mis-annotations that are common in public databases (46,47). The use of pre-constructed gpkg databases and a computationally efficient search step that minimises I/O overhead (see Methods) enables GraftM to run substantially faster than metAnnotate (21) and scale to the increasing size of metagenomes (Supplementary Figure S6). A growing number of gpkgs are available online (https://github.com/geronimp/graftM_gpkgs), and users are encouraged to submit new gpkgs through this portal. To facilitate the transfer of large gpkgs, GraftM provides a compression/decompression utility ('GraftM archive').

Using *in silico* and *in vitro* mocks GraftM was shown to sensitively and quickly identify metagenome reads from a gene family of interest (Figure 2). However, both pairwise (DIAMOND) and model based (HMMSEARCH) approaches for searching metagenomes have reduced sensitivity at the start and end of target genes (Figure 2C), where short reads only partially overlap with the ORF. This limitation will be alleviated as sequencing reads increase in length. Poorly conserved regions in HMMs also have reduced sensitivity when applied to environmental samples with divergent homologs (Supplementary Note 2, Supple-

mentary Figure S2), a problem partially addressed by building additional expand_search HMMs with sequences derived from assembled metagenomic data (Supplementary Note 2, Supplementary Figures S4 and S5). An additional masking step where poorly conserved regions in the HMM are removed from analysis may provide a more accurate estimation of relative abundance for a gene of interest.

Despite performing accurately overall to the family level (Figure 3), GraftM's clade-based approach to classification can result in poor resolution when paraphyletic taxonomy is used to annotate the tree (e.g. *Escherichia* and *Hydrogenobaculum*; Supplementary Table S4), an issue that will improve as the current taxonomy is replaced by a standardised genome-based taxonomy (e.g. GTDB, http://gtdb.ecogenomic.org). Poor resolution can also occur when taxonomy derived from one marker gene (e.g. 16S rRNA gene) is applied to a different gene with an incongruent phylogeny. Generally, this issue can be resolved by manually refining annotations, but in some cases gene phylogenies have minimal taxonomic or functional signal and are not well suited to a tree-based analysis. Finally, sequences belonging to lineages that are not well represented in the tree (e.g. *Leptothrix* and *Methanopyrus*) are poorly resolved, but the rapid expansion of reference databases driven by genome-centric metagenomics will enable increasingly specific annotations to be made.

The two primary applications of GraftM are to characterise the composition of a sample using taxonomic marker genes (e.g. 16S rRNA, Figure 5A), and to target specific populations or functions using marker genes (e.g. McrA, Figure 5B). GraftM can also be applied to some eukaryotic genes (e.g. 28S rRNA, Supplementary Figure S7), though its sensitivity when searching metagenomes for eukaryotic protein-coding genes may be reduced by intron/exon boundaries. Further, GraftM can be used to functionally annotate proteins by using superfamily phylogenies as reference trees (e.g. DMSO superfamily), or to distinguish substrate specificity within gene families (glycoside hydrolase subfamilies, RAS superfamily, P-type ATPase family; Figure 5C). These approaches also allow gene duplication and functional divergence to be analysed within a phylogenetic context, which are both commonly overlooked aspects of functional annotation (Supplementary Figure S8).

The speed of GraftM allows it to be used to screen large public datasets for genes of interest (e.g. McrA), which lead to the discovery of three novel McrA-containing lineages distinct from the Euryarchaeota, Bathyarchaeota (36) and Verstraetearchaeota (37); Figure 4). Given the phylogenetic novelty of these sequences, GraftM did not specify a prediction of the enzyme's substrate, where the pairwise method (DIAMOND) over-classified these sequences. This use-case illustrates how phylogenetic approaches to annotation can sensitively distinguish novel function from sequence divergence.

While pplacer has high speed and linear algorithmic complexity, high throughput placement of reads into large phylogenies (e.g. 16S rRNA) could be expedited using algorithms that decompose the tree into more computationally tractable subsets that are less memory- and CPU-intensive (25). Future development of GraftM will also focus on expanding its genome annotation capabilities, where the taxonomy of the microorganism and the annotation of multiple genes can guide our understanding of the microbial community's overall metabolism.

DATA AVAILABILITY

GraftM is part of the M-tools suite, available at ecogenomic.org/software. GraftM is licensed under the GNU General Public License version 3+ (GPLv3+), with stable versions of the code available on pip (pypi.python.org/pypi/graftm), and development code is available on GitHub (github.com/geronimp/graftM). Coral and permafrost metagenomes have been submitted to the NCBI BioProjects PRJNA386568 and PRJNA434624, respectively.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors would like to thank Mircea Podar and Stuart Denman for providing DNA used to create the mock community. The assistance in library preparation and sequencing by Margaret Butler and Serene Low is much appreciated. We thank Lauren Messer and Steven Robbins for their advice in selecting an appropriate marker gene to detect the sub-clade of *Symbiodinium*. We thank Donovan Parks for extensive advice. We thank the Australian Centre for Ecogenomics for testing and providing valuable input to GraftM.

FUNDING

Genomic Science Program of the United States Department of Energy Office of Biological and Environmental Research [DE-SC0004632, DE-SC0010580, DE-SC0016440]; Australian Research Council (ARC) Postgraduate Award (to J.A.B.); ARC Discovery Early Career Researcher Award [DE-160100248 to B.J.W.]; University of Queensland Vice Chancellor Research Focused Fellowship (to G.W.T.). Funding for open access charge: Australian Research Council [DE-160100248]; U.S. Department of Energy, Office of Science, Biological and Environmental Research [DE-SC0004632].
Conflict of interest statement. None declared.

REFERENCES

1. Tyson,G.W., Chapman,J., Hugenholtz,P., Allen,E.E., Ram,R.J., Richardson,P.M., Solovyev,V.V., Rubin,E.M., Rokhsar,D.S. and Banfield,J.F. (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*, **428**, 37–43.
2. Albertsen,M., Hugenholtz,P., Skarshewski,A., Nielsen,K.L., Tyson,G.W. and Nielsen,P.H. (2013) Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nat. Biotechnol.*, **31**, 533–538.
3. Wu,Y.-W., Tang,Y.-H., Tringe,S.G., Simmons,B.A. and Singer,S.W. (2014) MaxBin: an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm. *Microbiome*, **2**, 26.
4. Kang,D.D., Froula,J., Egan,R. and Wang,Z. (2015) MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ*, **3**, e1165.

5. Wrighton,K.C., Thomas,B.C., Sharon,I., Miller,C.S., Castelle,C.J., VerBerkmoes,N.C., Wilkins,M.J., Hettich,R.L., Lipton,M.S., Williams,K.H. *et al.* (2012) Fermentation, hydrogen, and sulfur metabolism in multiple uncultivated bacterial phyla. *Science*, **337**, 1661–1665.
6. Brown,C.T., Hug,L.A., Thomas,B.C., Sharon,I., Castelle,C.J., Singh,A., Wilkins,M.J., Wrighton,K.C., Williams,K.H. and Banfield,J.F. (2015) Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature*, **523**, 208–211.
7. Sczyrba,A., Hofmann,P., Belmann,P., Koslicki,D., Janssen,S., Dröge,J., Gregor,I., Majda,S., Fiedler,J., Dahms,E. *et al.* (2017) Critical assessment of metagenome interpretation—a benchmark of metagenomics software. *Nat. Methods*, **14**, 1063–1071.
8. Prosser,J. (2015) Dispersing misconceptions and identifying opportunities for the use of 'omics' in soil microbial ecology. *Nature*, **13**, 439–446.
9. Howe,A.C., Jansson,J.K., Malfatti,S.A., Tringe,S.G., Tiedje,J.M. and Brown,C.T. (2014) Tackling soil diversity with the assembly of large, complex metagenomes. *Proc. Natl. Acad. Sci. U.S.A.*, **111**, 4904–4909.
10. Huson,D.H., Mitra,S., Ruscheweyh,H.-J., Weber,N. and Schuster,S.C. (2011) Integrative analysis of environmental sequences using MEGAN4. *Genome Res.*, **21**, 1552–1560.
11. Glass,E.M., Wilkening,J., Wilke,A., Antonopoulos,D. and Meyer,F. (2010) Using the metagenomics RAST server (MG-RAST) for analyzing shotgun metagenomes. *Cold Spring Harbor Protocols*, **2010**, doi:10.1101/pdb.prot5368.
12. MacDonald,N.J., Parks,D.H. and Beiko,R.G. (2012) Rapid identification of high-confidence taxonomic assignments for metagenomic data. *Nucleic Acids Res.*, **40**, e111.
13. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
14. Buchfink,B., Xie,C. and Huson,D.H. (2014) Fast and sensitive protein alignment using DIAMOND. *Nat. Methods*, **12**, 59–60.
15. Mai,H., Li,D., Zhang,Y., Leung,H.C.-M., Luo,R., Ting,H.-F. and Lam,T.-W. (2016) AC-DIAMOND: accelerating protein alignment via better SIMD parallelization and space-efficient indexing. In: *Bioinformatics and Biomedical Engineering*. Lecture Notes in Computer Science. Springer, Cham, pp. 426–433.
16. de Vries,M., Schöler,A., Ertl,J., Xu,Z. and Schloter,M. (2015) Metagenomic analyses reveal no differences in genes involved in cellulose degradation under different tillage treatments. *FEMS Microbiol. Ecol.*, **91**, fiv069.
17. Krogh,A., Brown,M., Mian,I.S., Sjölander,K. and Haussler,D. (1994) Hidden Markov models in computational biology. Applications to protein modeling. *J. Mol. Biol.*, **235**, 1501–1531.
18. Krause,L., Diaz,N.N., Goesmann,A., Kelley,S., Nattkemper,T.W., Rohwer,F., Edwards,R.A. and Stoye,J. (2008) Phylogenetic classification of short environmental DNA fragments. *Nucleic Acids Res.*, **36**, 2230–2239.
19. Schreiber,F., Gumrich,P., Daniel,R. and Meinicke,P. (2010) Treephyler: fast taxonomic profiling of metagenomes. *Bioinformatics*, **26**, 960–961.
20. Darling,A.E., Jospin,G., Lowe,E., Matsen,F.A. IV, Bik,H.M. and Eisen,J.A. (2014) PhyloSift: phylogenetic analysis of genomes and metagenomes. *PeerJ*, **2**, e243.
21. Petrenko,P., Lobb,B., Kurtz,D., Neufeld,J. and Doxey,A. (2015) MetAnnotate: function-specific taxonomic profiling and comparison of metagenomes. *BMC Biol.*, **13**, 92.
22. Von Mering,C., Hugenholtz,P., Raes,J., Tringe,S., Doerks,T., Jensen,L. and Bork,P. (2007) Quantitative phylogenetic assessment of microbial communities in diverse environments. *Science*, **315**, 1126–1130.
23. Matsen,F., Kodner,R. and Armbrust,V. (2010) pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics*, **11**, 538.
24. Berger,S.A., Krompass,D. and Stamatakis,A. (2011) Performance, accuracy, and web server for evolutionary placement of short sequence reads under maximum likelihood. *Systematic Biol.*, **60**, 291–302.
25. Mirarab,S., Nguyen,N. and Warnow,T. (2012) SEPP: SATé-enabled phylogenetic placement. *Biocomputing*, **2012**, 247–258.
26. Matsen,F. and Evans,S. (2013) Edge principal components and squash clustering: using the special structure of phylogenetic placement data for sample comparison. *PLoS One*, **8**, e56859.
27. Kiełbasa,S.M., Wan,R., Sato,K., Horton,P. and Frith,M.C. (2011) Adaptive seeds tame genomic sequence comparison. *Genome Res.*, **21**, 487–493.
28. Woodcroft,B.J., Boyd,J.A. and Tyson,G.W. (2016) OrfM: a fast open reading frame predictor for metagenomic data. *Bioinformatics*, **32**, 2702–2703.
29. Eddy,S.R. (2011) Accelerated Profile HMM Searches. *PLoS Comput. Biol.*, **7**, e1002195.
30. Katoh,K. and Standley,D.M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.*, **30**, 772–780.
31. Ondov,B.D., Bergman,N.H. and Phillippy,A.M. (2011) Interactive metagenomic visualization in a Web browser. *BMC Bioinformatics*, **12**, 385.
32. Price,M., Dehal,P. and Arkin,A. (2010) FastTree 2–approximately maximum-likelihood trees for large alignments. *PLoS One*, **5**, e9490.
33. McDonald,D., Price,M.N., Goodrich,J., Nawrocki,E.P., DeSantis,T.Z., Probst,A., Andersen,G.L., Knight,R. and Hugenholtz,P. (2012) An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME J.*, **6**, 610–618.
34. Markowitz,V.M., Chen,I.-M.A., Palaniappan,K., Chu,K., Szeto,E., Grechkin,Y., Ratner,A., Jacob,B., Huang,J., Williams,P. *et al.* (2012) IMG: the Integrated Microbial Genomes database and comparative analysis system. *Nucleic Acids Res.*, **40**, D115–D122.
35. Parks,D.H., Imelfort,M., Skennerton,C.T., Hugenholtz,P. and Tyson,G.W. (2015) CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.*, **25**, 1043–1055.
36. Evans,P.N., Parks,D.H., Chadwick,G.L., Robbins,S.J., Orphan,V.J., Golding,S.D. and Tyson,G.W. (2015) Methane metabolism in the archaeal phylum Bathyarchaeota revealed by genome-centric metagenomics. *Science*, **350**, 434–438.
37. Vanwonterghem,I., Evans,P.N., Parks,D.H., Jensen,P.D., Woodcroft,B.J., Hugenholtz,P. and Tyson,G.W. (2016) Methylotrophic methanogenesis discovered in the archaeal phylum Verstraetearchaeota. *Nat. Microbiol.*, **1**, 16170.
38. Ludwig,W., Strunk,O., Westram,R., Richter,L., Meier,H., Yadhukumar, Buchner,A., Lai,T., Steppi,S., Jobb,G. *et al.* (2004) ARB: a software environment for sequence data. *Nucleic Acids Res.*, **32**, 1363–1371.
39. Galagan,J.E., Nusbaum,C., Roy,A., Endrizzi,M.G., Macdonald,P., FitzHugh,W., Calvo,S., Engels,R., Smirnov,S., Atnoor,D. *et al.* (2002) The genome of M. acetivorans reveals extensive metabolic and physiological diversity. *Genome Res.*, **12**, 532–542.
40. Rinke,C., Low,S., Woodcroft,B.J., Raina,J.-B., Skarshewski,A., Le,X.H., Butler,M.K., Stocker,R., Seymour,J., Tyson,G.W. *et al.* (2016) Validation of picogram- and femtogram-input DNA libraries for microscale metagenomics. *PeerJ*, **4**, e2486.
41. Johansson,T., Malmer,N., Crill,P.M., Friborg,T., Åkerman,J.H., Mastepanov,M. and Christensen,T.R. (2006) Decadal vegetation changes in a northern peatland, greenhouse gas fluxes and net radiative forcing. *Global Change Biol.*, **12**, 2352–2369.
42. Seemann,T. (2014) Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, **30**, 2068–2069.
43. Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.
44. Haft,D.H., DiCuccio,M., Badretdin,A., Brover,V., Chetvernin,V., O'Neill,K., Li,W., Chitsaz,F., Derbyshire,M.K., Gonzales,N.R. *et al.* (2017) RefSeq: an update on prokaryotic genome annotation and curation. *Nucleic Acids Res.*, **46**, D851–D860.
45. Schuur,E.A.G., McGuire,A.D., Schädel,C., Grosse,G., Harden,J.W., Hayes,D.J., Hugelius,G., Koven,C.D., Kuhry,P., Lawrence,D.M. *et al.* (2015) Climate change and the permafrost carbon feedback. *Nature*, **520**, 171–179.
46. Schnoes,A.M., Brown,S.D., Dodevski,I. and Babbitt,P.C. (2009) Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. *PLoS Comput. Biol.*, **5**, e1000605.
47. Kozlov,A.M., Zhang,J., Yilmaz,P., Glöckner,F.O. and Stamatakis,A. (2016) Phylogeny-aware identification and correction of taxonomically mislabeled sequences. *Nucleic Acids Res.*, **44**, 5022–5033.