

Tax4Fun2: a R-based tool for the rapid prediction of habitat-specific functional profiles and functional redundancy based on 16S rRNA gene marker gene sequences

Franziska Wemheuer^{1,2}, Jessica A. Taylor³, Rolf Daniel⁴, Emma Johnston^{1,2}, Peter Meinicke⁵,
Torsten Thomas³ and Bernd Wemheuer^{3,4*}

¹Evolution and Ecology Research Centre, School of Biological, Earth and Environmental Sciences, University of New South Wales, NSW 2052, Australia

²Sydney Institute of Marine Science, Mosman, NSW 2088, Australia

³Centre for Marine Bio-Innovation, School of Biological, Earth and Environmental Sciences, University of New South Wales, NSW 2052, Australia

⁴Genomic and Applied Microbiology and Göttingen Genomics Laboratory, Institute of Microbiology and Genetics, University of Göttingen, Göttingen, Germany

⁵Department of Bioinformatics, Institute of Microbiology and Genetics, University of Göttingen, Göttingen, Germany

*** Correspondence:**

Bernd Wemheuer

bwemheu@gwdg.de

Main text

Assessing the functional capability and redundancy of a microbial community is a major challenge in environmental microbiology. To address this challenge, we developed Tax4Fun2, a R-based tool for the rapid prediction of functional profiles and functional redundancy of prokaryotic communities from 16S rRNA gene sequences. By incorporating user-defined, habitat-specific genomic information, the accuracy and robustness of predicted functional profiles can be substantially enhanced.

Microorganisms play a key role in ecosystem functioning¹. High-throughput sequencing of 16S rRNA genes is a powerful and widely used approach to study the composition and structure of microbial communities in a variety of marine²⁻⁴, terrestrial^{5,6} and host-associated^{7,8} environments. However, numerous questions in biogeochemistry and ecosystem ecology require knowledge of community functions rather than the taxonomic composition⁹. In recent years, several freely available tools such as PICRUSt¹⁰, Tax4Fun¹¹, Piphillin¹², Faprotax¹³ and paprica¹⁴ have been developed. Although these tools cannot replace the functional assessment via metagenomic shotgun sequencing, they have provided unique insights into functional capabilities of prokaryotic communities in diverse habitats, such as soil^{5,6}, marine seawater^{2,13,14}, microbial mats¹⁵ and the plant endosphere⁷.

The predictive power of these tools relies on functional information derived from genomes available in public databases. However, available genomes do not necessarily represent the total functional diversity present in the ecosystem investigated. This problem has motivated the development of predictive tools specific for the rumen microbiome¹⁶ or marine microorganisms¹³. Given the rapidly increasing number of available genomes, in particular through metagenome-assisted genome binning¹⁷, and that many research groups have access

to unpublished, habitat-specific genomic information, the incorporation of this data should enhance the accuracy of functional inferences.

To address these challenges, we developed Tax4Fun2, a novel version of Tax4Fun¹¹. Tax4Fun2 is a fast and user-friendly R package (<https://sourceforge.net/projects/tax4fun2/>) with a current default reference dataset of 275 archaeal and 12,002 bacterial genomes available through NCBI RefSeq database (assessed on 19 August 2018). A novel feature is that Tax4Fun2 can incorporate habitat-specific and user-defined data to increase the robustness and specificity of functional profiles (Fig. 1). Although Tax4Fun2 focuses on prokaryotic data, eukaryotic data can also be incorporated. Tax4Fun2 is platform-independent and highly memory-efficient, enabling researchers without extensive bioinformatics knowledge to predict functional profiles on almost every computer.

We first applied Tax4Fun2 in comparison to Tax4Fun¹¹ and PICRUSt¹⁰ using the same paired samples (16S rRNA and metagenomic data available), which were used to validate Tax4Fun¹¹ and PICRUSt¹⁰, i.e. samples derived from the human microbiome, mammalian guts, soil and from a hypersaline microbial mat (Table 1). In addition, we evaluated the predictive power (defined as high Spearman correlation coefficient) of Tax4Fun2 using ten marine seawater⁴ samples taken in the North Sea and 90 kelp-associated samples collected within the Marine Microbes Framework Data Initiative (<http://www.bioplatforms.com/marine-microbes>). Tax4Fun2 outperforms PICRUSt and Tax4Fun across all these datasets (Fig. 2a). Functional profiles predicted by Tax4Fun2 were highly correlated to functional profiles derived from the metagenomes. Although the predicted profiles for the kelp-associated communities were significantly correlated to functional profiles, the median Spearman correlation coefficient was only 0.72, indicating that a lack of suitable reference genomes limits Tax4Fun2's performance.

To address this issue, we used 68 metagenome-assembled genomes (MAGs) derived from the 90 kelp-associated metagenomes to build a kelp-specific genomic dataset. This

substantially increased the accuracy of the prediction (median Spearman correlation coefficient with user data added = 0.86) and reduced the fraction of the sequences not used in the predictions (Fig. 2b). Moreover, using the kelp-specific dataset allowed to predict functional profiles for samples, which failed when using only the default reference data because next neighbour search resulted in no close matches. These results demonstrate the benefits of incorporating habitat-specific reference databases, which distinguishes Tax4Fun2 from all other published tools.

A major question in microbial ecology is whether, and to what degree, microbial communities contain functionally redundant members, that may provide stability of ecosystem processes in the face of environmental perturbations^{18,19}. In Tax4Fun2, we introduced a functional redundancy index (FRI) with respect to single functions. The FRI is based on the proportion of species capable of performing a particular function and their phylogenetic relationship to each other. A high FRI indicates that a specific function is almost ubiquitous in all community members, whereas a low FRI suggests that the function is present in a few closely related species or has been detected in only one community member. A FRI of 0 indicates that a function is not present at all. Tax4Fun2 calculates a relative FRI (rFRI), which is normalized by the average phylogenetic distance of the community analysed in a specific survey, and the absolute FRI (aFRI), which is normalized by the average phylogenetic distance of all prokaryotes in the reference tree provided with Tax4Fun2. The rFRI can be used to compare samples within one survey, whereas the aFRI allows the comparison of functional redundancy indices across different, unrelated ecosystems.

To test the accuracy of the FRI calculation, we simulated 1,000 communities, each consisting of 100 prokaryotic genomes. We extracted the 16S rRNA gene sequences from each simulated community, clustered them at 97% similarity and calculated the FRI values. We subsequently compared these values to FRI values based on the actual genomic information of

the simulated communities. This comparison revealed that Tax4Fun2 provided a good estimate of the functional redundancy present in the microbial community (Spearman rank correlation > 90%) (Fig. 2b). We further tested this approach using the marine sea water samples⁴. Six of these samples were taken inside a phytoplankton bloom and three samples served as reference that were taken outside the bloom. Nearly 7,000 functions displayed a higher functional redundancy in the reference samples, whereas only 1,468 functions had higher redundancies inside the bloom (Fig. 2c). This indicates that the functional redundancy greatly shifts during the phytoplankton bloom. Phytoplankton blooms are usually characterized by a substrate-controlled succession, i.e. distinct bacterial clades dominate the bacterioplankton community at different stages during and shortly after the bloom²⁰. Consequently, community members involved in the turnover of certain substrates at a specific stage are predominant and thus their genomes and associated functions will be more redundant, whereas the opposite can be observed for all other community members.

Tax4Fun2 provides researchers with a unique tool to predict and investigate functional profiles of prokaryotic communities based on 16S rRNA gene data. We demonstrated the high predictive power of Tax4Fun2, which can be further enhanced by the incorporation of user-defined and habitat-specific data. Another unique feature of Tax4Fun2 is that it enables researchers to calculate the redundancy of specific functions, which is critical for the prediction how likely a specific function is lost during environmental perturbation. Tax4Fun2 with its user-friendly, simplified workflow will assist researchers considerably in the functional analysis of microbial communities.

Methods

Datasets used in this study. To compare Tax4Fun2 with Tax4Fun¹¹ and PICRUSt¹⁰, we used the same 16S rRNA datasets which were originally used to validate Tax4Fun and PICRUSt (for details see ¹¹). We further assessed the accuracy of Tax4Fun2 using 10 marine water (taken from ⁴) and 90 kelp-derived metagenomes (for details see <https://data.bioplatforms.com/organization/about/australian-microbiome>). A list with all accession number is provided in Table 1.

Processing of 16S rRNA gene data from marine water samples. Pyrosequencing data were processed using QIIME version 1.8²¹. After raw data extraction, reads shorter than 600 bp or longer than 900 bp, exhibiting low quality (<25), possessing long homopolymer stretches (>8 bp), or showing primer mismatches (>2 bp) were removed. Remaining reverse primer sequences were truncated employing cutadapt version 1.18²². Processed sequences of all samples were concatenated and denoised employing Acacia version 1.53b²³. Denoised sequences were sorted by decreasing length and clustered at 97% sequence identity in operational taxonomic units (OTUs) employing the UCLUST algorithm implemented in USEARCH version 8.1.1861²⁴. Chimeric sequences were removed using UCHIME²⁵ implemented in USEARCH in *denovo* and *reference* mode with the SILVA database (SILVA_132_SSURef_Nr99) as reference dataset²⁶.

Processing of 16S rRNA gene data derived from kelp samples. Paired reads were merged with Flash²⁷ and subsequently processed with USEARCH version 10.240²⁴. Merged reads were quality-filtered; the filtering included the removal of low-quality reads (maximum number of expected errors >2 and more than 1 ambiguous base) and those shorter than 400 bp. Processed sequences of all samples were concatenated into one file, dereplicated, and obtained unique sequences were denoised and clustered into zero-radius OTUs (zOTUs) with the *unoise3*

algorithm. A *de novo* chimera removal was included in the unnoise step. Afterwards, remaining chimeric sequences were removed using the *uchime2* algorithm²⁵ in high confidence mode with the SILVA database as reference dataset²⁶. Subsequently, processed sequences were mapped onto zOTU sequences to calculate the presence and abundance of each zOTU in every sample using the *otutab* command with *maxrejects* and *maxaccepts* options disabled.

Functional predictions based on 16S rRNA data. Functional profiles were predicted with PICRUST¹⁰, Tax4Fun¹¹ and Tax4Fun2. For PICRUST, processed sequences were clustered using QIIME^{21,28} by close reference picking against greengenes (13_5), PICRUST's default database. OTU abundances were normalized by 16S rRNA copy numbers prior to the calculation of functional profiles. For Tax4Fun, OTU sequences were aligned against the SILVA database (SILVA_123_SSURef_Nr99)²⁶ using BLAST version 2.7.1²⁹. The OTU table and the taxonomic classification were subsequently merged and used to predict functional profiles in Tax4Fun using default settings. For Tax4Fun2, functional profiles were initially aligned against the supplied 16S rRNA reference sequences by BLAST using the *runRefBlast* function. Functional predictions were subsequently calculated using the *makeFunctionalPrediction* function.

Generation of reference datasets. Tax4Fun2 is supplied with two reference datasets (Ref99NR and Ref100NR) refereeing to the similarity threshold used during 16S rRNA clustering. Each dataset consists of an association matrix with 16S rRNA reference sequences associated with functional reference profiles (number of entries in the association matrix: 4,584 and 18,479 for Ref99NR and Ref100NR, respectively).

Reference datasets were generated as follows: we downloaded all complete genomes and all genomes with the status 'chromosome' from NCBI RefSeq (assessed on 18 August

2018), resulting in 275 archaeal genomes and 12,102 bacterial genomes. Barrnap version 0.9 (<https://github.com/tseemann/barrnap>) was used to identify and extract all 16S rRNA gene sequences. All obtained sequences were subsequently concatenated into a single file, sorted by decreasing length and clustered using the UCLUST algorithm implemented in USEARCH version 10.240²⁴ at 99% and 100% sequence similarity, respectively. The longest sequence of each cluster served as 16S rRNA reference sequence.

Functional profiles were generated for each genome as follows: open-reading-frames were identified with prodigal version 2.6.3³⁰. Functional profiles were calculated based on deduced protein sequences with UProC version 1.2.0³¹ using the KEGG database for prokaryotes (July 2018 release) as reference³². To account for differences in rRNA copy numbers, functional profiles were normalized by the number of 16S rRNA genes identified in each genome. Due to the heterogeneity of 16S rRNA genes within a genome, the functional reference profile for each 16S rRNA reference sequence was generated based on the 16S rRNA clustering results: a single functional reference profile is the average normalized functional profiles of each genome with at least one 16S rRNA gene affiliated to the cluster. If more than one 16S rRNA gene sequence of a genome was assigned to the cluster, the normalized profile of the genome was multiplied by the number of 16S rRNA genes before calculating the mean profile.

The algorithm which was used to generate the reference data is implemented in the Tax4Fun2 package (function = *addUserDataByClustering*). Note that a 32-bit version of USEARCH is required to use this function. USEARCH is freely available at <https://www.drive5.com/usearch/>.

Testing the predictive power of Tax4Fun2. To test the predictive power of Tax4Fun2 compared to PICRUSt and Tax4Fun, we used the same paired samples (16S rRNA and

metagenomic data), which were originally used to validate Tax4Fun's accuracy. Functional profiles for each metagenome used in the validation process were generated as follow: protein sequences were extracted with prodigal version 2.6.3³⁰ and functional annotations were made with UProC version 1.2.0³¹ as described above for the functional genome annotation. We validated the accuracy of PICRUSt, Tax4Fun and Tax4Fun2 by comparing the functional profiles predicted to the metagenomic profile using Spearman correlation (see Figure 2). Due to several changes in the KEGG orthology since PICRUSt and Tax4Fun were developed (deprecated and new functional orthologs), a direct comparison of functional profiles predicted with all three tools is difficult. Hence, functional profiles were converted to relative abundances prior to comparison. Only functions present in the metagenomic profile and in the predictions were considered in the comparison.

Generation of metagenome-assembled genomes (MAGs). The incorporation of user-derived genomes is a key feature of Tax4Fun2, allowing users to build their own reference data. To exploit the accuracy of Tax4Fun2 with default settings (without user data) and with a user-defined reference database, we added 68 MAGs obtained from the 90 kelp metagenomes. The genomes were extracted from the metagenomes as follows: raw data were quality trimmed with Trimmomatic version 0.36³³ and subsequently assembled with metaSPAdes version 3.11.1³⁴. The coverage of each scaffold was determined by mapping the processed data on the assembled scaffolds using bowtie version 2.3.2³⁵. Scaffolds smaller than 2,500 bp were removed. After converting to bam format and sorting using samtools³⁶, the coverage was determined with the *jgi_summarize_bam_contig_depths* script. Genomes were extracted using MetaBAT version 0.32.5³⁷ and MyCC³⁸ and subsequently refined using the *binning_refiner* version 1.2³⁹. 16S rRNA gene sequences were identified using barnap version 0.9. The completeness and contamination was determined with checkM version 1.0.7⁴⁰. All genomes with more than 50%

completeness, less than 5% contamination and possessing at least one 16S rRNA gene were included as user data in the Tax4Fun2 prediction. A functional profile was generated with UProC version 1.2.0³¹ and KEGG as described above for the functional genome annotation.

Calculation of the functional redundancy index (FRI). In Tax4Fun2, we introduced the functional redundancy index (FRI). The FRI describes the redundancy of any given function in the investigated community. It incorporates the phylogenetic distribution (distance) of community members harbouring the function and their proportion in the community (see Formula 1).

$$FRI = \frac{\overline{PD}_x}{\overline{PD}} \times \frac{N_x}{N}$$

$$\frac{\overline{PD}_x}{\overline{PD}}$$

= Average phylogenetic distance of species with function x divided by the average phylogenetic distance of all species

$\frac{N_x}{N}$ = proportion of species with function x in a community with N species

To account for differences in phylogenetic distance, we introduced the absolute and relative FRI (aFRI and rFRI). The difference between them is the average phylogenetic distance used for its normalization. To calculate the aFRI, the average phylogenetic distance of all species in the full 16S rRNA reference tree is used, whereas the rFRI is normalized by the average phylogenetic distance of those species in the 16S rRNA reference tree observed in a sample. The tree for each of the two reference datasets was generated as follows: all 16S rRNA reference sequences were aligned with SINA version 1.2.11⁴¹ and the latest Silva ARB release (SILVA_132_SSUREF_NR99). The phylogenetic tree was calculated using RaxML version 8.2.11⁴² under a GTRGAMMA model and a random seed of 12345.

Testing the functional redundancy index using simulated datasets. To test the FRI accuracy, we simulated 1,000 communities each consisting of 100 genomes randomly selected from the 12,377 genomes used to generate the reference data. The genomes were selected based on random numbers generated with the *sample* function in R. To assess the phylogenetic distance between the genomes, we extracted 59 marker protein sequences based on hmm profiles derived from PFAM version 31⁴³ and TIGRFAM version 15⁴⁴. The 59 marker proteins were selected because their corresponding genes were present in 90% of all 12,377 genomes and, if present, were single-copy genes in 99% of them. These criteria were applied to archaea and bacteria independently. The extracted protein sequences of each hmm profile were aligned using mafft version 7.3.11⁴⁵. Afterwards, aligned protein sequences for each genome were concatenated. The phylogenomic tree was calculated under a WAGGAMMA model using FastTree version 2.1.10⁴⁶. An initial attempt to use RaxML⁴² failed due to the size of the alignment. Functional profiles for each genome were converted to presence absence data and the FRI was calculated for each function using the genome tree and the presence-absence data.

To calculate the FRI in Tax4Fun2, the 16S rRNA gene sequences present in the 100 genomes of each subset were clustered in operational taxonomic units (OTUs) at 97% similarity with UCLUST implemented in USEARCH version 10.240. The longest sequence of each cluster was used to represent each OTU. The FRI was subsequently calculated using Tax4Fun2. The FRIs calculated for each function by Tax4Fun2 were compared to the FRIs calculated directly from the genomes of each simulation by Spearman rank correlation in R. The OTU table necessary for the calculation contained the id and size of each OTU.

Statistical data analysis. All statistical tests were performed in R version 3.5.1⁴⁷.

Data availability. Tax4Fun2 is feely available at <https://sourceforge.net/projects/tax4fun2/>.

Acknowledgements

This research was partly funded by the German Research Foundation (DFG): research fellowship granted to B.W. and TRR51 granted to R.D. T.T. and J.A.T. were supported by Bioplatforms Australia. F.W. and E.L.J. were supported by ARC Linkage Project SHRP021212.

Author Contributions

B.W. led the project. B.W., P.M. and F.W. designed and implemented the final Tax4Fun2 algorithms, to which T.T. and R.D. made critical contributions. J.A.T. collected and analysed the kelp dataset used for the validation of Tax4Fun2. B.W., P.M. and R.D. coordinated the online implementation. F.W. and B.W. wrote the manuscript, with feedback from all other authors. All authors approved the final version of the manuscript

Competing interests

The authors declare no competing interests.

References

- 1 Falkowski, P. G., Fenchel, T. & Delong, E. F. The microbial engines that drive Earth's biogeochemical cycles. *Science* **320**, 1034-1039 (2008).
- 2 Wemheuer, B. *et al.* Linking Compositional and Functional Predictions to Decipher the Biogeochemical Significance in DFAA Turnover of Abundant Bacterioplankton Lineages in the North Sea. *Microorganisms* **5**, 68 (2017).

- 3 Sun, M. Y., Dafforn, K. A., Johnston, E. L. & Brown, M. V. Core sediment bacteria drive community response to anthropogenic contamination over multiple environmental gradients. *Environ. Microbiol.* **15**, 2517-2531 (2013).
- 4 Wemheuer, B. *et al.* The green impact: bacterioplankton response towards a phytoplankton spring bloom in the southern North Sea assessed by comparative metagenomic and metatranscriptomic approaches. *Front. Microbiol.* **6**, 805 (2015).
- 5 Kaiser, K. *et al.* Driving forces of soil bacterial community structure, diversity, and function in temperate grasslands and forests. *Sci. Rep.* **6**, 33696 (2016).
- 6 Herzog, S., Wemheuer, F., Wemheuer, B. & Daniel, R. Effects of Fertilization and Sampling Time on Composition and Diversity of Entire and Active Bacterial Communities in German Grassland Soils. *PLoS ONE* **10**, e0145575 (2015).
- 7 Wemheuer, F. *et al.* Bacterial endophyte communities of three agricultural important grass species differ in their response towards management regimes. *Sci. Rep.* **7**, 40914 (2017).
- 8 Zozaya-Valdés, E., Roth-Schulze, A. J., Egan, S. & Thomas, T. Microbial community function in the bleaching disease of the marine macroalgae *Delisea pulchra*. *Environ. Microbiol.* **19**, 3012-3024 (2017).
- 9 Louca, S. *et al.* Function and functional redundancy in microbial systems. *Nat. Ecol. Evol.* **2**, 936-943 1 (2018).
- 10 Langille, M. G. I. *et al.* Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nat. Biotech.* **31**, 814-821 (2013).
- 11 ABhauer, K. P., Wemheuer, B., Daniel, R. & Meinicke, P. Tax4Fun: predicting functional profiles from metagenomic 16S rRNA data. *Bioinformatics* **31**, 2882-2884 (2015).

- 12 Iwai, S. *et al.* Piphillin: improved prediction of metagenomic content by direct inference from human microbiomes. *PLoS ONE* **11**, e0166104 (2016).
- 13 Louca, S., Parfrey, L. W. & Doebeli, M. Decoupling function and taxonomy in the global ocean microbiome. *Science* **353**, 1272 (2016).
- 14 Bowman, J. S. & Ducklow, H. W. Microbial Communities Can Be Described by Metabolic Structure: A General Framework and Application to a Seasonally Variable, Depth-Stratified Microbial Community from the Coastal West Antarctic Peninsula. *PLoS ONE* **10**, e0135868, (2015).
- 15 Koo, H. *et al.* Comparison of two bioinformatics tools used to characterize the microbial diversity and predictive functional attributes of microbial mats from Lake Obersee, Antarctica. *Journal Microbiol. Methods* **140**, 15-22, (2017).
- 16 Wilkinson, T. J. *et al.* CowPI: A Rumen Microbiome Focussed Version of the PICRUSt Functional Inference Software. *Front. Microbiol.* **9**, 1095 (2018).
- 17 Parks, D. H. *et al.* Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat. Microbiol.* **2**, 1533-1542, (2017).
- 18 Allison, S. D. & Martiny, J. B. H. Resistance, resilience, and redundancy in microbial communities. *Proc. Natl Acad. Sci. USA* **105**, 11512 (2008).
- 19 Jurburg, S. D. & Salles, J. F. Functional Redundancy and Ecosystem Function — The Soil Microbiota as a Case Study, Biodiversity in Ecosystems, Yueh-Hsin Lo, Juan A. Blanco and Shovonlal Roy, IntechOpen, London, UK.
- 20 Teeling, H. *et al.* Substrate-Controlled Succession of Marine Bacterioplankton Populations Induced by a Phytoplankton Bloom. *Science* **336**, 608-611, (2012).
- 21 Caporaso, J. G. *et al.* QIIME allows analysis of high-throughput community sequencing data. *Nat. Meth.* **7**, 335-336 (2010).

- 22 Martin, M. *Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal* **17**, 10-12 (2011).
- 23 Bragg, L., Stone, G., Imelfort, M., Hugenholtz, P. & Tyson, G. W. Fast, accurate error-correction of amplicon pyrosequences using Acacia. *Nat. Meth.* **9**, 425-426 (2012).
- 24 Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460-2461 (2010).
- 25 Edgar, R. C., Haas, B. J., Clemente, J. C., Quince, C. & Knight, R. UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* **27** 2194-2200 (2011).
- 26 Quast, C. *et al.* The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research* **41**, D590-D596 (2013).
- 27 Magoč, T. & Salzberg, S. L. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **27**, 2957-2963 (2011).
- 28 DeSantis, T. Z. *et al.* Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl. Environ. Microbiol.* **72**, 5069-5072 (2006).
- 29 Camacho, C. *et al.* in *BMC Bioinformatics* **10**, 421 (2009).
- 30 Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010).
- 31 Meinicke, P. UProC: tools for ultra-fast protein domain classification. *Bioinformatics*, **31**, 1382-1388 (2014).
- 32 Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **28**, 27-30 (2000).
- 33 Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: A flexible trimmer for Illumina Sequence Data. *Bioinformatics* **30**, 2114-2120 (2014).
- 34 Bankevich, A. *et al.* SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *J. Comput. Biol.* **19**, 455-477 (2012).

- 35 Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Meth.* **9**, 357-359 (2012).
- 36 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).
- 37 Kang, D. D., Froula, J., Egan, R. & Wang, Z. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ* **3**, e1165 (2015).
- 38 Lin, H.-H. & Liao, Y.-C. Accurate binning of metagenomic contigs via automated clustering sequences using information of genomic signatures and marker genes. *Sci. Rep.* **6**, 24175 (2016).
- 39 Song, W.-Z. & Thomas, T. Binning_refiner: improving genome bins through the combination of different binning programs. *Bioinformatics* **33**, 1873-1875 (2017).
- 40 Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043-1055 (2015).
- 41 Pruesse, E., Peplies, J. & Glöckner, F. O. SINA: Accurate high-throughput multiple sequence alignment of ribosomal RNA genes. *Bioinformatics* **28**, 1823-1829 (2012).
- 42 Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312-1313 (2014).
- 43 Punta, M. *et al.* The Pfam protein families database. *Nucleic Acids Res.* **40**, D290-D301 (2012).
- 44 Haft, D. H., Selengut, J. D. & White, O. The TIGRFAMs database of protein families. *Nucleic Acids Res.* **31**, 371-373 (2003).
- 45 Katoh, T. Recent developments in the MAFFT multiple sequence alignment program. *Brief. Bioinformatics* **9**, 286-298 (2008).

- 46 Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLoS ONE* **5**, e9490 (2010).
- 47 R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing (2018).

Figure legends

Fig. 1 | The Tax4Fun workflow. 16S rRNA gene sequences are initially aligned against the reference sequences to identify the nearest neighbour. If user-defined data is supplied, the 16S rRNA gene sequences are additionally aligned against the sequences added by the user. The nearest neighbour in the user data is preferred if both search attempts result in significant hits. The OTU abundances for each sample are summarized based on the results from the nearest neighbour search. An association matrix (AM) containing the functional profiles of those references identified in the 16S rRNA search is generated. The summarized abundances and the functional profiles stored in the AM are merged and a metagenome is predicted for each sample. The amount of sequences/OTUs unused in the prediction is provided in a log file.

Fig. 2 | Tax4Fun2 validation. a) Correlations between functional profiles obtained from metagenomic datasets and those predicted from 16s rRNA data. Predictions were made with PICRUSt, Tax4Fun, and Tax4Fun2 using both supplied default reference datasets (Ref99NR and Ref100NR). Note that PICRUSt did not generate any prediction for the Kelp data. b) Correlations between functional profiles retrieved from 90 kelp metagenomes and those predicted with Tax4Fun2 without and with user data added and the fraction of zOTUs and sequences unused in the prediction. c) Functional redundancy indices inside and outside a phytoplankton bloom. A log ratio greater than 0 indicates that a function is more redundant

outside the bloom. All predictions were made using a 97% similarity cut off. Correlations are Spearman rank correlations based on relative abundances of KO functions. Only functions present in the metagenome and the predictions were used for comparison. Note that a direct comparison between PICRUSt, Tax4Fun and Tax4Fun2 is difficult due to changes in the KEGG database. Currently, the KEGG databases includes information for more than 10,000 protein-related KO orthologs, whereas PICRUSt and Tax4Fun only provide predictions for around 7000 KO terms.

Tables

Table 1: Accession numbers of samples/studies used to validate Tax4Fun2.

| Origin | Sample number | Accession numbers |
|---------------------|---------------|---|
| Human Microbiome | 41 | SRS011271, SRS011452, SRS011529, SRS011584, SRS011586, SRS013234, SRS013252, SRS013258, SRS013506, SRS013687, SRS013711, SRS013723, SRS014235, SRS014287, SRS014343, SRS014613, SRS014629, SRS014923, SRS015133, SRS015190, SRS015425, SRS015450, SRS015574, SRS015578, SRS015762, SRS015782, SRS015854, SRS015960, SRS016002, SRS016018, SRS016095, SRS016111, SRS016203, SRS016225, SRS016331, SRS016335, SRS016349, SRS016434, SRS016533, SRS016553, SRS016559 |
| Mammalian Gut | 56 | 4461284-301, 4461341-55, 4461357-58, 4461360-80, 4461383 (MG- RAST) |
| Microbial Mat | 10 | 4440963-71 (MG-RAST) |
| Soil | 14 | 4477803-5, 4477807, 4477872-7, 4477899, 4477902-4 (MG-RAST) |
| Water | 10 | SRA060677 |
| Kelp | 90 | 57884-936, 57938-56, 87958-74, 58019-20 (https://data.bioplatforms.com/organization/about/australian-microbiome .) |

Tax4Fun2

Input data

Reference data

Representative sequences

Nearest neighbour search by local alignment

16S rRNA reference data (Tax4Fun2)

Representative sequences

Nearest neighbour search by local alignment

16S rRNA reference data (user)

OTU table

OTU to reference profile conversion

Generation of specific association matrix

Functional reference profiles

Using matrix to merge abundance and functional data

Predicted metagenomes

