1   **PICRUSt2: An improved and customizable approach for metagenome inference**

2   Gavin M. Douglas[1], Vincent J. Maffei[2], Jesse Zaneveld[3], Svetlana N. Yurgel[4], James R. Brown[5],

3   Christopher M. Taylor[2], Curtis Huttenhower[6], Morgan G. I. Langille[1,7,]*

4

5   [1]Department of Microbiology and Immunology, Dalhousie University, Halifax, NS, Canada

6   [2]Department of Microbiology, Immunology, and Parasitology, Louisiana State University Health

7   Sciences Center, New Orleans, Louisiana, USA

8   [3]University of Washington, Seattle, Washington, USA

9   [4]Department of Plant, Food, and Environmental Sciences, Dalhousie University, Truro, NS,

10   Canada

11   [5]Computational Biology, GlaxoSmithKline R&D, Collegeville, Pennsylvania, USA

12   [6]Harvard T. H. Chan School of Public Health, Boston, Massachusetts, USA

13   [7]Department of Pharmacology, Dalhousie University, Halifax, Nova Scotia, Canada

14   *Corresponding author: morgan.langille@dal.ca

15

16

17

18

19

20

21

22

23

1   One major limitation of microbial community marker gene sequencing is that it does not provide

2   direct information on the functional composition of sampled communities. Here, we present

3   PICRUSt2 (https://github.com/picrust/picrust2), which expands the capabilities of the original

4   PICRUSt method[1] to predict the functional potential of a community based on marker gene

5   sequencing profiles. This updated method and implementation includes several improvements

6   over the previous algorithm: an expanded database of gene families and reference genomes, a

7   new approach now compatible with any OTU-picking or denoising algorithm, and novel

8   phenotype predictions. Upon evaluation, PICRUSt2 was more accurate than PICRUSt1 and other

9   current approaches overall. PICRUSt2 is also now more flexible and allows the addition of

10   custom reference databases. We highlight these improvements and also important caveats

11   regarding the use of predicted metagenomes, which are related to the inherent challenges of

12   analyzing metagenome data in general.

13   The most common approach for profiling communities is to sequence the highly

14   conserved 16S rRNA gene. Functional profiles cannot be directly identified from 16S rRNA

15   gene sequence data due to strain variation and because 16S rRNA genes are not unique among

16   microbes, but several approaches have been developed to infer approximate microbial

17   community functions from taxonomic profiles (and thus amplicon sequences) alone[1–6].

18   Importantly, these methods predict functional potential, i.e. functions encoded at the level of

19   DNA. Although shotgun metagenomic sequencing (MGS) directly samples genetic functional

20   potential within microbial communities, this methodology is not without limitations. In

21   particular, functional inference from amplicon data remains important for samples with

22   substantial host contamination (e.g. biopsy samples), low biomass, and where metagenomic

23   sequencing is not economically feasible.

1   PICRUSt[1] (hereafter "PICRUSt1") was the first tool developed and the most widely used

2   for metagenome prediction, but like any inference model has several limitations. First, the

3   standard PICRUSt1 workflow requires input sequences to be operational taxonomic units

4   (OTUs) generated from closed-reference OTU picking against a compatible version of the

5   Greengenes database[7]. Due to this limitation, the default PICRUSt1 workflow is incompatible

6   with sequence denoising methods[8], which produce amplicon sequence variants (ASVs) rather

7   than OTUs. ASVs have finer resolution, allowing closely related organisms to be more readily

8   distinguished. Lastly, the prokaryotic reference databases used by PICRUSt1 have not been

9   updated since 2013 and lack many recently added gene families and pathway mappings.

10   The PICRUSt2 algorithm includes new steps that optimize genome prediction, which we

11   hypothesized would improve prediction accuracy (**Fig 1**). These are: (1) study sequences are now

12   placed into a pre-existing phylogeny rather than relying on discrete predictions limited to

13   reference OTUs (**Fig 1b**); (2) predictions are based off of a greatly increased number of

14   reference genomes and gene families (**Fig 1c**); (3) pathway abundance inference is now more

15   stringently performed (**Supp Fig 1**); (4) predictions can now be made for higher level

16   phenotypes; and (5) custom databases are easier to integrate into the prediction pipeline.

17   PICRUSt2 integrates multiple high-throughput, open-source tools to predict the genomes

18   of environmentally sampled 16S rRNA gene sequences. ASVs are placed into a reference tree,

19   which is used as the basis of functional predictions. This reference tree contains 20,000 full 16S

20   rRNA genes from prokaryotic genomes in the Integrated Microbial Genomes (IMG) database[9].

21   Phylogenetic placement in PICRUSt2 is based on running three tools: HMMER

22   (www.hmmer.org) to place ASVs, EPA-ng[10] to determine the optimal position of these placed

23   ASVs in a reference phylogeny, and GAPPA[11] to output a new tree incorporating the ASV

1   placements. This results in a phylogenetic tree containing both reference genomes and

2   environmentally sampled organisms, which is used to predict individual gene family copy

3   numbers for each ASV. This procedure is re-run for each input dataset, allowing users to utilize

4   custom reference databases as needed, including those that may be optimized for the study of

5   specific microbial niches.

6   As in PICRUSt1, hidden state prediction (HSP) approaches are used in PICRUSt2 to

7   infer the genomic content of sampled sequences. The castor R package[12], which is substantially

8   faster than the ape package[13] used previously in PICRUSt1, now performs the core HSP

9   functions. As in PICRUSt1, ASVs are corrected by their 16S rRNA gene copy number and then

10   multiplied by their functional predictions to produce a predicted metagenome. PICRUSt2 also

11   provides the ASV contribution of each predicted function allowing for taxonomy-informed

12   statistical analyses to be conducted. Lastly, pathway abundances are now inferred based on

13   structured pathway mappings, which are more conservative than the bag-of-genes approach

14   previously used in PICRUSt1.

15   The new PICRUSt2 default genome database is based on 41,926 bacterial and archaeal

16   genomes from the IMG database[9] as of November 8, 2017, which is a >20-fold increase over the

17   2,011 IMG genomes used for PICRUSt1 predictions. Many of these genomes are from strains of

18   the same species and have identical 16S rRNA genes. We de-replicated the identical 16S rRNA

19   genes across these genomes, which resulted in 20,000 final 16S rRNA gene clusters.

20   As a result of this increased database size, the taxonomic diversity of the PICRUSt2

21   reference database has markedly increased compared to PICRUSt1 (**Fig. 1c**). The clearest

22   increases in diversity have been driven by increases at the species and genus levels (5.3-fold and

1   2.2-fold increases respectively). However, all taxonomic levels exhibited increased diversity,

2   including the phylum level where the coverage increased from 39 to 64 phyla (1.6-fold increase).

3
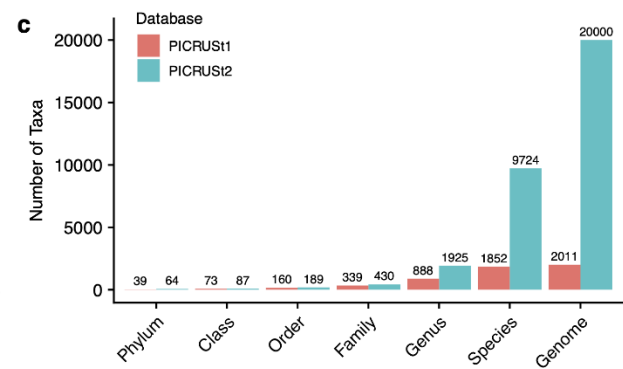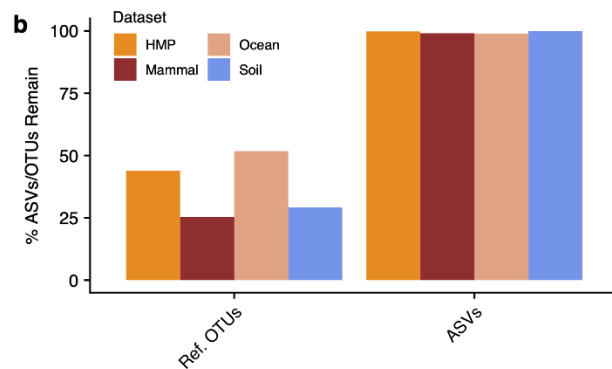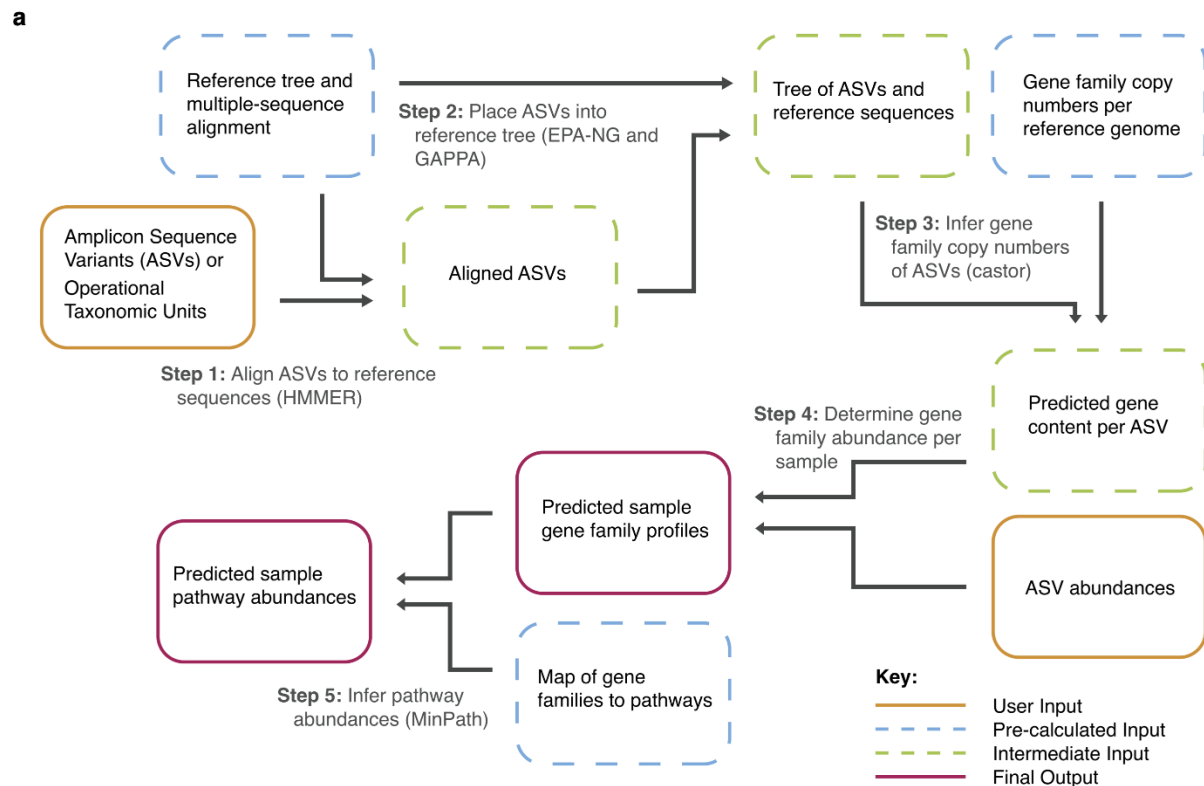
4

5   **Figure 1: PICRUSt2 algorithm and major updates.** (a) The PICRUSt2 method consists of
6   phylogenetic placement, hidden-state-prediction, and sample-wise gene and pathway abundance
7   tabulation. ASV sequences and abundances are taken as input, and gene family and pathway abundances
8   are output. All necessary reference tree and trait databases for the default workflow are included in the
9   PICRUSt2 implementation. (b) The default PICRUSt1 pipeline restricted predictions to reference
10   operational taxonomic units (Ref. OTUs) within the Greengenes database. This requirement resulted in

1 the exclusion of many study sequences across four representative 16S rRNA gene sequencing datasets. In
2 contrast, PICRUSt2 relaxes this requirement and is agnostic to whether the input sequences are within a
3 reference or not, which results in almost all of the input amplicon sequence variants (ASVs) being
4 retained in the final output. (c) A drastic increase in the taxonomic diversity within the default PICRUSt2
5 database is observed compared to PICRUSt1.
6

7 PICRUSt2 predictions based on the following gene family databases are supported by

8 default: Kyoto Encyclopedia of Genes and Genomes[14] (KEGG) orthologs (KO), Enzyme

9 Classification numbers (EC numbers), Clusters of Orthologous Genes[15] (COGs), Protein

10 families[16] (Pfam) and The Institute for Genomic Research's database of protein FAMilies[17]

11 (TIGRFAM) (**Supp Table 1**). PICRUSt2 distinctly improves on PICRUSt1 by including gene

12 families more recently added to the KEGG database. Specifically, the total number of KOs has

13 now increased from 6,911 to 10,543 (1.5-fold increase) in PICRUSt2 compared to PICRUSt1.

14 We validated PICRUSt2 metagenome predictions using samples from seven published

15 datasets that have been profiled both by 16S rRNA marker gene sequencing and shotgun

16 metagenomics sequencing (MGS). These included three human-associated microbiome datasets:

17 57 stool samples from Cameroonian individuals[18,19], 91 stool samples from Indian individuals[20],

18 and 137 samples spanning the human body (from the Human Microbiome Project[21] [HMP]).

19 These validation datasets also included non-human associated environments, including: 77 non-

20 human primate stool samples[22], eight mammalian stool samples[23], six ocean samples[24], and 22

21 bulk soil and blueberry rhizosphere samples[25]. These datasets span varying degrees of challenge

22 for accurate metagenome inference due to environmental and technical factors (**Supp Table 2**).

23 We generated PICRUSt2 KO predictions from 16S rRNA marker gene data for each

24 dataset. We compared these predictions to KO relative abundances profiled from the

25 corresponding MGS metagenomes, which served as a gold-standard to evaluate prediction

26 performance. We performed the same analysis with four alternative functional prediction

1   pipelines: PICRUSt1, Piphillin, PanFP, and Tax4Fun2. We calculated Spearman correlation

2   coefficients (hereafter "correlations") for matching samples between the predicted KO

3   abundance and MGS KO abundance tables after filtering all tables to the 6,220 KOs that could

4   be output by all tested databases (**Fig 2**). The correlation metric represents the similarity in rank

5   ordering of KO abundances between the predicted and observed data. The correlations based on

6   PICRUSt2 KO predictions ranged from a mean of 0.79 (standard deviation [sd] = 0.028; primate

7   stool) to 0.88 (sd = 0.019; Cameroonian stool dataset). For all seven datasets, PICRUSt2

8   predictions either performed best or were comparable to the best prediction method (paired-

9   sample, two-tailed Wilcoxon tests [PTW] $P < 0.05$). Correlations based on PICRUSt2

10 predictions were notably higher for non-human associated datasets. This result could indicate an

11 advantage of phylogenetic-based methods over non-phylogenetic-based methods, such as

12 Piphillin, for environments poorly represented by reference genomes.

13
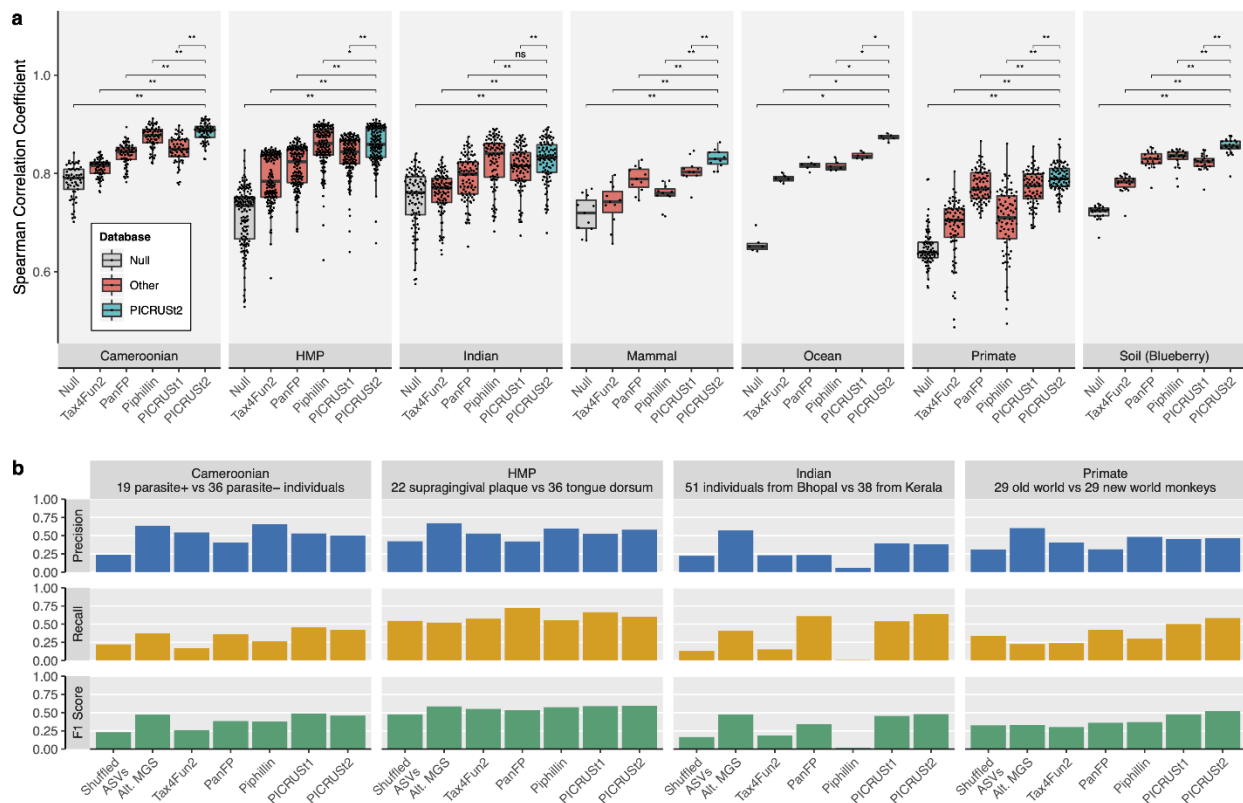
1 **Figure 2: PICRUSt2 performs best or is comparable to other tools based on Spearman correlation**
2 **coefficients and differential abundance results.** Validation results of PICRUSt2 KEGG ortholog (KO)
3 predictions comparing metagenome prediction performance against gold-standard shotgun metagenomic
4 sequencing (MGS). (a) Boxplots represent medians and interquartile ranges of Spearman correlation
5 coefficients observed in stool samples from Cameroonian individuals (n=57), the human microbiome
6 project (HMP, n=137), stool samples from Indian individuals (n=91), non-human primate stool samples
7 (n=77), mammalian stool (n=8), ocean water (n=6), and blueberry soil (n=22) datasets. The significance
8 of paired-sample, two-tailed Wilcoxon tests is indicated above each tested grouping (*, **, and ns
9 correspond to P < 0.05, P < 0.001, and not significant respectively). Note that the y-axis is truncated
10 below 0.5 rather than 0 to better visualize small differences between categories. (b) Comparison of
11 significantly differentially abundant KOs between predicted metagenomes and MGS. Precision, recall,
12 and F1 score are reported for each category compared to the MGS data. Precision corresponds to the
13 proportion of significant KOs for that category also significant in the MGS data. Recall corresponds to the
14 proportion of significant KOs in the MGS data also significant for that category. The F1 score is the
15 harmonic mean of these metrics. The subsets of the four datasets tested (which were the only ones with
16 adequate sample sizes for this analysis) and the sample groupings compared are indicated above each
17 panel. The parasite referred to for the Cameroonian dataset is *Entamoeba*. Wilcoxon tests were performed
18 on the KO relative abundances after normalizing by the median number of universal single-copy genes in
19 each sample. Significance was defined as a false discovery rate < 0.05. The "shuffled ASVs" category
20 corresponds to PICRUSt2 predictions with ASV labels shuffled across a dataset (see Supplementary
21 Text). The "Alt. MGS" category corresponds to an alternative MGS processing pipeline where reads were
22 aligned directly to the KEGG database rather than the default HUMAnN2 pipeline.
23

24 Gene families regularly co-occur within genomes, and so the use of correlations to assess

25 gene-table similarity may be limited by the lack of independence of gene families within a

26 sample (**Supp Fig 2**). To address this dependency, we compared the observed correlations

27 between paired MGS and predicted metagenomes to correlations between MGS functions and a

28 null reference genome, comprised of the mean gene family abundance across all reference

29 genomes. For all datasets, PICRUSt2 metagenome tables were more similar to MGS values than

30 the null (**Fig 2a**). However, this increase over the null expectation is predominately driven by

31 each dataset's predicted genome content (rather than that of individual samples). This is

32 demonstrated by the fact that these correlations are actually only slightly significantly higher

33 than those observed when ASV labels are shuffled within a dataset (**Supp Fig 3**). The observed

34 correlations for the shuffled ASVs ranged from a mean of 0.77 (sd = 0.196; primate stool) to

35 0.84 (sd = 0.178; blueberry rhizosphere). Biologically these results are consistent with several

1   patterns. First, gene families are correlated in copy number across diverse taxa (as captured by

2   the 'Null' dataset). Second, these correlations are stronger within than between environments (as

3   shown by the difference between the 'Null' and 'Shuffled ASV' results). Lastly, environment-to-

4   environment differences tend to be larger than sample-to-sample differences within an

5   environment (as shown by the differences between PICRUSt2 predictions and the 'Shuffled

6   ASV' results).

7   A complementary approach for validating metagenome predictions is to compare the

8   results of differential abundance tests on 16S-predicted metagenomes to MGS data. A recent

9   analysis of Piphillin suggested that this tool out-performs PICRUSt2 based on this approach[26].

10   We similarly performed this evaluation on the KO predictions for four validation datasets (**Fig**

11   **2b**; see Supplementary Text). Overall, PICRUSt2 displayed the highest F1 score, the harmonic

12   mean of precision and recall, compared to other prediction methods (ranging from 0.46-0.59;

13   mean=0.51; sd=0.06). However, all prediction tools displayed relatively low precision, the

14   proportion of significant KOs that were also significant in the MGS data. In particular, precision

15   ranged from 0.38-0.58 (mean=0.48; sd=0.08) for PICRUSt2 and 0.06-0.66 (mean=0.45; sd=0.27)

16   for Piphillin. In all cases, PICRUSt2 predictions out-performed ASV-shuffled predictions, which

17   ranged in precision from 0.22-0.42 (mean=0.30; sd=0.09). In addition, differential abundance

18   tests performed on MGS-derived KOs from an alternative MGS-processing workflow resulted in

19   only marginally higher precision (ranging from 0.57-0.67; mean=0.62; sd=0.04). Taken together,

20   these results highlight the difficulty of reproducing microbial functional biomarkers with both

21   predicted and actual metagenomics data.

22   MetaCyc pathway abundances are now the main high-level predictions output by

23   PICRUSt2 by default. The MetaCyc database is an open-source alternative to KEGG and is also

1 a major focus of the widely-used metagenomics functional profiler, HUMAnN2[27]. MetaCyc

2 pathway abundances are calculated in PICRUSt2 through structured mappings of EC gene

3 families to pathways. These pathway predictions performed better than the null distribution for

4 all metrics overall (PTW P < 0.05; **Fig 3a** and **Supp Fig 4-5**) compared to MGS-derived

5 pathways. Similar to our previous analysis, shuffled ASV predictions representing overall

6 functional structure within each dataset accounted for the majority of this signal (**Supp Fig 4**). In

7 addition, differential abundance tests on these pathways showed high variability in F1 scores

8 across datasets and statistical methods with the ASV shuffled predictions contributing the

9 majority of this signal (**Supp Fig 6**; F1 scores ranged from 0.23-0.62 (mean=0.41; sd=0.17) and

10 0.22-0.60 (mean=0.34; sd=0.18) for the observed and ASV shuffled PICRUSt2 predictions,

11 respectively). Again, these results suggest that identifying robust differentially abundant

12 metagenome-wide pathways is difficult and highlights the challenge of analyzing microbial

13 pathways in general.

14 Predictions for 41 microbial phenotypes, which are linked to IMG genomes[28], can also

15 now be generated with PICRUSt2. These represent high-level microbial metabolic activities such

16 as "Glucose utilizing" and "Denitrifier" that are annotated as present or absent within each

17 reference genome. Use of this database was motivated by the predictions made by the tools

18 FAPROTAX[29] and Bugbase[30]. We performed a hold-out validation to assess the performance of

19 PICRUSt2 phenotype predictions, which involved comparing the binary phenotype predictions

20 to the expected phenotypes for each reference genome. Based on F1 score (mean=84.8%;

21 sd=9.01%), precision (mean=86.5%; sd=6.21%), and recall (mean=83.5%; sd=11.4%), these

22 predictions performed significantly better than the null expectation (**Fig 3b**; Wilcoxon tests P <

23 0.05).

1
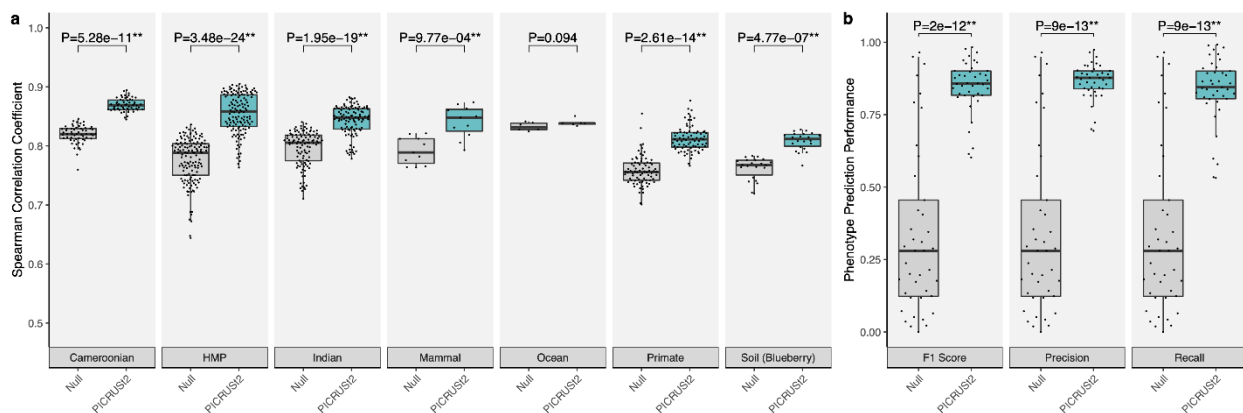
Figure 3: PICRUSt2 accurately predicts MetaCyc pathways and phenotypes for characterizing overall environments. (a) Spearman correlation coefficients between PICRUSt2 predicted pathway abundances and gold-standard metagenomic sequencing (MGS). Results are shown for each validation dataset: stool from Cameroonian individuals, The Human Microbiome Project (HMP), stool from Indian individuals, mammalian stool, ocean water, non-human primate stool, and blueberry soil. These results are limited to the 575 pathways that could potentially be identified by PICRUSt2 and HUMAnN2. (b) Performance of binary phenotype predictions based on three metrics: F1 score, precision, and recall. Each point corresponds to one of the 41 phenotypes tested. Predictions assessed here are based on holding out each genome individually, predicting the phenotypes for that holdout genome, and comparing the predicted and observed values. The null distribution in this case is based on randomizing the phenotypes across the reference genomes and comparing to the actual values, which results in the same output for all three metrics. The P-values of paired-sample, two-tailed Wilcoxon tests is indicated above each tested grouping (* and ** correspond to $P < 0.05$ and $P < 0.001$, respectively). Note that in panel a the y-axis is truncated below 0.5 rather than 0 to better visualize small differences between categories. The sample sizes in panel a are 57 (Cameroonian), 137 (HMP), 91 (Indian), 8 (mammal), 6 (ocean), 77 (primate), and 22 (soil).

There are two major criticisms of amplicon-based functional prediction. First, the predictions are biased towards existing reference genomes, which means that rare environment-specific functions are less likely to be identified. This limitation will be partially addressed as the number of high-quality available genomes continues to grow. Moreover, PICRUSt2 allows user-specified genomes to be used for generating predictions, which provides a flexible framework for studying particular environments. The second major criticism is that amplicon-based predictions cannot provide resolution to distinguish strain-specific functionality within the same species. This is an important limitation of PICRUSt2 and any amplicon-based analysis, which can only differentiate taxa to the degree they differ at the amplified marker gene sequence.

1   In summary, PICRUSt2 is a more flexible and accurate method for performing marker

2   gene metagenome inference. We have highlighted the improved performance of PICRUSt2

3   compared to other metagenome inference methods while also describing limitations with

4   identifying consistent differentially abundant functions in microbiome studies. We hope that the

5   expanded functionality of PICRUSt2 will continue to allow researchers to identify potentially

6   novel insights into functional microbial ecology from amplicon sequencing profiles.

7

8   **Code and data availability**

9   PICRUSt2 is available at: https://github.com/picrust/picrust2. The Python and R code used for

10   the analyses and database construction described in this paper are available online at

11   https://github.com/gavinmdouglas/picrust2_manuscript. This repository also includes the

12   processed datafiles that can be used to re-generate the findings in this paper. The accessions for

13   all sequencing data used in this study are listed in the supplementary information.

14

15   **Acknowledgements**

1   in part by NIH NIDDK grants U54DK102557 and R24DK110499. SYN is funded by an NSERC

2   Discovery Grant.

3

4   **<u>References</u>**

5   1.   Langille, M. G. I. *et al.* Predictive functional profiling of microbial communities using

6   16S rRNA marker gene sequences. *Nat. Biotechnol.* **31**, 814–821 (2013).

7   2.   Aßhauer, K. P., Wemheuer, B., Daniel, R. & Meinicke, P. Tax4Fun: Predicting functional

8   profiles from metagenomic 16S rRNA data. *Bioinformatics* **31**, 2882–2884 (2015).

9   3.   Jun, S. R., Robeson, M. S., Hauser, L. J., Schadt, C. W. & Gorin, A. A. PanFP:

10   Pangenome-based functional profiles for microbial communities. *BMC Res. Notes* **8**, 497

11   (2015).

12   4.   Bowman, J. S. & Ducklow, H. W. Microbial communities can be described by metabolic

13   structure: A general framework and application to a seasonally variable, depth-stratified

14   microbial community from the coastal West Antarctic Peninsula. *PLoS One* **10**, e0135868

15   (2015).

16   5.   Iwai, S. *et al.* Piphillin: Improved prediction of metagenomic content by direct inference

17   from human microbiomes. *PLoS One* **11**, e0166104 (2016).

18   6.   Wemheuer, F. *et al.* Tax4Fun2: a R-based tool for the rapid prediction of habitat-specific

19   functional profiles and functional redundancy based on 16S rRNA gene marker gene

20   sequences. *bioRxiv* (2018).

7. 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).

8. Callahan, B. J. *et al.* DADA2: High resolution sample inference from amplicon data. *Nat. Methods* **13**, 581–583 (2016).

9. Markowitz, V. M. *et al.* IMG: The integrated microbial genomes database and comparative analysis system. *Nucleic Acids Res.* **40**, 115–122 (2012).

10. Barbera, P. *et al.* EPA-ng: Massively Parallel Evolutionary Placement of Genetic Sequences. *Syst. Biol.* **68**, 365–369 (2019).

11. Czech, L. & Stamatakis, A. Scalable methods for analyzing and visualizing phylogenetic placement of metagenomic samples. *PLoS One* **14**, e0217050 (2019).

12. Louca, S. & Doebeli, M. Efficient comparative phylogenetics on large trees. *Bioinformatics* **34**, 1053–1055 (2018).

13. Paradis, E., Claude, J. & Strimmer, K. APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics* **20**, 289–290 (2004).

14. Kanehisa, M., Goto, S., Sato, Y., Furumichi, M. & Tanabe, M. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.* **40**, 109–114 (2012).

15. Tatusov, R. L., Galperin, M. Y., Natale, D. A. & Koonin, E. V. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.* **28**, 33–36 (2000).

16. Finn, R. D. *et al.* Pfam: The protein families database. *Nucleic Acids Res.* **42**, 222–230

(2014).

17. Haft, D. H., Selengut, J. D. & White, O. The TIGRFAMs database of protein families. *Nucleic Acids Res.* **31**, 371–373 (2003).

18. Morton, E. R. *et al.* Variation in Rural African Gut Microbiota Is Strongly Correlated with Colonization by *Entamoeba* and Subsistence. *PLoS Genet.* **11**, e1005658 (2015).

19. Lokmer, A. *et al.* Use of shotgun metagenomics for the identification of protozoa in the gut microbiota of healthy individuals from worldwide populations with various industrialization levels. *PLoS One* **14**, e0211139 (2019).

20. Dhakan, D. B. *et al.* The unique composition of Indian gut microbiome, gene catalogue, and associated fecal metabolome deciphered using multi-omics approaches. *Gigascience* **8**, 1–20 (2019).

21. Huttenhower, C. *et al.* Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207–214 (2012).

22. Amato, K. R. *et al.* Evolutionary trends in host physiology outweigh dietary niche in structuring primate gut microbiomes. *ISME J.* **13**, 576–587 (2019).

23. Finlayson-Trick, E. C. L. *et al.* Taxonomic differences of gut microbiomes drive cellulolytic enzymatic potential within hind-gut fermenting mammals. *PLoS One* **12**, e0189404 (2017).

24. Gillies, L. E., Thrash, J. C., deRada, S., Rabalais, N. N. & Mason, O. U. Archaeal enrichment in the hypoxic zone in the northern Gulf of Mexico. *Environ. Microbiol.* **17**,

1   3847–3856 (2015).

2   25.   Yurgel, S. N., Nearing, J. T., Douglas, G. M. & Langille, M. G. I. Metagenomic

3   Functional Shifts to Plant Induced Environmental Changes. *Front. Microbiol.* **10**, 1682

4   (2019).

5   26.   Narayan, N. R. *et al.* Piphillin predicts metagenomic composition and dynamics from

6   DADA2-corrected 16S rDNA sequences. *BMC Genomics* **21**, 56 (2020).

7   27.   Franzosa, E. A. *et al.* Species-level functional profiling of metagenomes and

8   metatranscriptomes. *Nat. Methods* **15**, 962–968 (2018).

9   28.   Chen, I. M. A. *et al.* Improving Microbial Genome Annotations in an Integrated Database

10   Context. *PLoS One* **8**, e54859 (2013).

11   29.   Louca, S., Parfrey, L. W. & Doebeli, M. Decoupling function and taxonomy in the global

12   ocean microbiome. *Science* **353**, 1272–1277 (2016).

13   30.   Ward, T. *et al.* BugBase predicts organism-level microbiome phenotypes. *bioRxiv* (2017).

14