PICRUSt2: An improved and extensible approach for metagenome inference

Gavin M. Douglas[1], Vincent J. Maffei[2], Jesse Zaneveld[3], Svetlana N. Yurgel[4], James R. Brown[5], Christopher M. Taylor[2], Curtis Huttenhower[6], Morgan G. I. Langille[1,7,*]

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

[2]Department of Microbiology, Immunology, and Parasitology, Louisiana State University Health Sciences Center, New Orleans, Louisiana, USA

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

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

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

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

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

*Corresponding author: morgan.langille@dal.ca

Abstract

One major limitation of microbial community marker gene sequencing is that it does not provide direct information on the functional composition of sampled communities. Here, we present PICRUSt2, which expands the capabilities of the original PICRUSt method to predict approximate functional potential of a community based on marker gene sequencing profiles. This updated method and implementation includes several improvements over the previous algorithm: an expanded database of gene families and reference genomes, a new approach now compatible with any OTU-picking or denoising algorithm, novel phenotype predictions, and novel fungal

reference databases that enable predictions from 18S rRNA gene and internal transcribed spacer amplicon data. Upon evaluation, PICRUSt2 was more accurate than PICRUSt1 and other current approaches and also more flexible to allow the addition of custom reference databases. Last, we demonstrate the utility of PICRUSt2 by identifying potential disease-associated microbial functional signatures based on 16S rRNA gene sequencing of ileal biopsies collected from a cohort of human subjects with inflammatory bowel disease. PICRUSt2 is freely available at: https://github.com/picrust/picrust2.

Introduction

Next-generation sequencing has enabled high-throughput profiling of microbial communities, which has improved our understanding of microbial ecology in diverse environments. The most common approach for profiling communities is to sequence the highly conserved 16S rRNA gene (16S), which acts as an identifier that may be enumerated to determine the relative abundances of prokaryotic groups present. Functional profiles cannot be directly identified from 16S sequence data due to strain variation and the lack of uniqueness of 16S rRNA gene sequences among microbes, but several approaches have been developed to infer approximate microbial community functions from taxonomic profiles (and thus amplicon sequences) alone[1–5]. Importantly, these methods predict functional potential, i.e. functions encoded at the level of DNA. Although shotgun metagenomic sequencing (MGS) directly samples genetic functional potential within microbial communities, this methodology is not without limitations. In particular, functional inference from amplicon data remains important for samples with substantial host contamination (e.g. biopsy samples), low biomass, and where metagenomic sequencing is not economically feasible.

PICRUSt[1] (hereafter "PICRUSt1") was the first tool developed and the most widely used for metagenome prediction, but like any inference model has several limitations. First, by default PICRUSt1 requires input sequences to be operational taxonomic units (OTUs) generated from closed-reference OTU picking against a specific, compatible version of the Greengenes database[6]. Due to this limitation, PICRUSt1's implementation is incompatible with sequence denoising methods[7–9], which are rapidly becoming the predominant approach as they enable sequence resolution down to the single-nucleotide level. This improved resolution allows closely related organisms to be better distinguished and thus more precise gene annotations are associated with a given 16S sequence. In addition, PICRUSt1 cannot be used with 18S rRNA gene (18S) and internal transcribed spacer (ITS) sequencing data as its database is limited to prokaryotic community predictions. Lastly, the prokaryotic reference databases used by PICRUSt1 have not been updated since 2013 and lack many recently added gene families and pathway mappings.

Since PICRUSt1 was published a number of similar metagenome inference tools have been developed. All of these approaches aim to capture the shared phylogenetic signal in the distribution of functions across taxa; however, they differ in many fundamental ways. Tax4Fun[2] and Piphillin[5] produce metagenome predictions based on the nearest-neighbour mappings of reference 16S sequences to study 16S sequences. This process is pre-computed in advance for Tax4Fun and is restricted to reference OTUs in the SILVA database[10]. In contrast, this procedure is quickly re-run by Piphillin for each dataset and either OTUs or ASVs can be input. PanFP[3] is another similar tool that bases metagenome prediction on the taxonomic classifications of 16S sequences. PICRUSt1 differs from all three of these approaches since it explicitly accounts for each unknown organism's position in a phylogenetic tree, which enables more sophisticated

methods for inferring hidden states to be employed. PAPRICA[4] is an alternative approach that does use a phylogenetic approach but has focused primarily on marine microbes and has not been tested against rigorously against a known gold-standard.

Here, we present PICRUSt2, which addresses the limitations of the above tools and results in the most accurate prediction method to date. PICRUSt2 relies on a new algorithm that leverages recently developed short-read placement tools that insert sequences into an existing phylogenetic tree, a more accurate method relative to *de novo* amplicon tree-building[11] as used in PICRUSt1. We compare PICRUSt2 hidden-state prediction (HSP) against the other HSP tools mentioned above and demonstrate that PICRUSt2 out-performs these tools in predicting gene family relative abundances as well as gene family presence and absence. Lastly, we demonstrate the utility of PICRUSt2 to infer the functional capacity of sequenced human ileal communities from an inflammatory bowel disease cohort and identify novel microbial signatures linked with host metabolome and transcriptome data from the same subjects.

Results

The PICRUSt2 algorithm includes several improved steps that optimize genome prediction and were hypothesized to improve prediction accuracy (**Fig 1**). The key improvements are: (1) study sequences are now placed into a pre-existing phylogeny rather than relying on discrete predictions limited to reference OTUs (**Fig 1B**); (2) a greatly increased number of reference genomes and gene families are used for generating predictions (**Fig 1C**); (3) pathway abundance inference is now more stringently performed (**Supp Fig 1**); (4) predictions can now be made for fungal amplicon sequences and for (5) higher level phenotypes, and (6) custom databases are

easier to integrate into the prediction pipeline. PICRUSt2 is available as a standalone command-line implementation and as a QIIME2 plugin[12] (https://github.com/picrust/picrust2).

Functional inference from variant sequences

PICRUSt2 integrates multiple high-throughput, open-source tools to predict the genomes of environmentally sampled 16S amplicon sequences. Amplicon sequence variants (ASVs) are placed into a reference tree, which is used as the basis of functional predictions. This reference tree contains 20,000 full 16S sequences from prokaryotic genomes from the Integrated Microbial Genomes database[13]. Phylogenetic placement in PICRUSt2 is based on a sequence of three steps: HMMER (www.hmmer.org) to place ASVs, EPA-NG[14] to determine the best position of these placed ASVs in a reference phylogeny, and GAPPA[15] to output a tree of the most likely ASV placements. This results in a phylogenetic tree containing both reference genomes and environmentally sampled organisms, which is used to predict individual gene family copy numbers for each ASV. Since this procedure is re-run for each input dataset, users can specify custom reference databases with improved resolution for specific environments[16].

As in PICRUSt1, HSP approaches are used in PICRUSt2 to infer the genomic content of sampled sequences. The castor R package[17], which is substantially faster than the ape package[18] used previously in PICRUSt1, now performs the core HSP functions and allows for greater flexibility and future improvements in modelling gene evolution. As in PICRUSt1, ASVs are corrected by their 16S copy number and then multiplied by their functional predictions to produce a predicted metagenome. PICRUSt2 also provides the taxonomic contribution of each predicted function allowing for taxonomy-informed statistical analyses to be conducted (see example below). Lastly, pathway abundances are now inferred based on structured pathway

mappings, which are more conservative than the bag-of-genes approach previously used in PICRUSt1 (see Online Methods).

Updated Reference Database

The new PICRUSt2 default genome database is based on 41,926 bacterial and archaeal genomes from the Integrated Microbial Genomes (IMG) database[13] as of November 8, 2017, which is a >20-fold increase over the 2,011 IMG genomes used for PICRUSt1 predictions. Many of these genomes are from strains of the same species and have identical 16S sequences. We de-replicated the identical 16S sequences across these genomes, which resulted in 20,000 final 16S clusters. A total of 3,002 of these clusters are composed of more than one identical 16S sequence that were collapsed together by clustering. The sequences contained in these 3,002 clusters make up 59.5% of the original 41,926 genomes. Among the 20,000 sequence clusters, there is a mean of 2.1 sequences per cluster (standard deviation [sd] = 562.5). The cluster with the highest sequence count of 1,379 corresponded to strains of *Staphylococcus aureus*. We observed a mean clustered-sequence length of 1489.6 base-pairs (bp; sd = 65.8) overall.

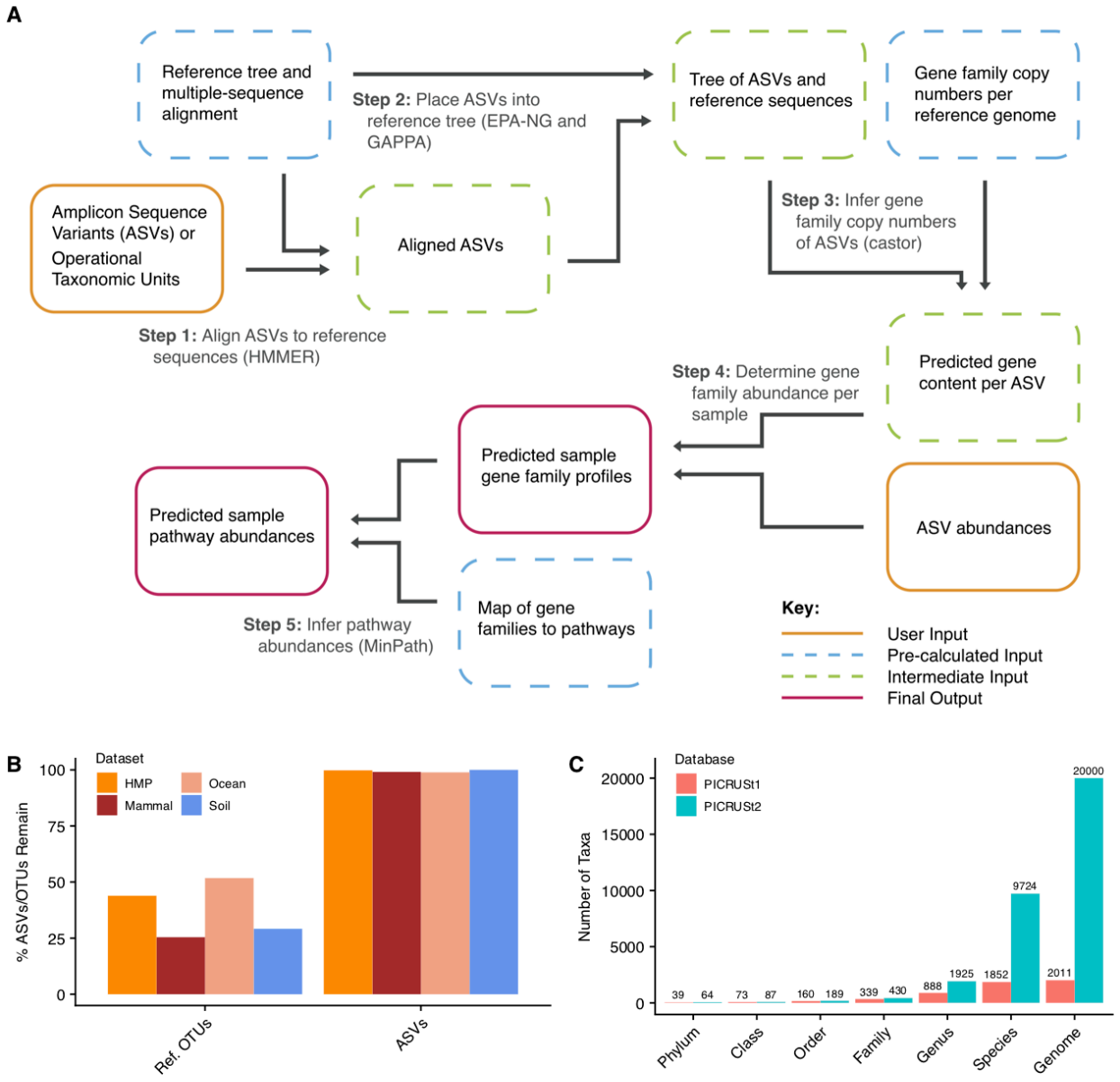

Figure 1: PICRUSt2 algorithm and major updates. (A) The PICRUSt2 method now consists of phylogenetic placement, hidden-state-prediction, and sample-wise gene abundance tabulation. ASV sequences and abundances are taken as input, and stratified gene family and pathway abundances are output. All necessary reference tree and trait databases for the default workflow are included in the PICRUSt2 implementation. (B) The PICRUSt1 pipeline restricted predictions to reference operational taxonomic units (Ref. OTUs) within the Greengenes database by default. This requirement resulted in a substantial proportion of the study sequences to be excluded from the analyses across four 16S rRNA gene sequencing datasets. In contrast, since PICRUSt2 relaxes this requirement and

is agnostic to whether the input sequences are within a reference or not, almost all of the input amplicon sequence variants (ASVs) are retained in the final output. (C) A drastic increase in the taxonomic diversity within the default PICRUSt2 database is observed compared to PICRUSt1.

As a result of this increased database size, the taxonomic diversity of the PICRUSt2 reference database has markedly increased compared to PICRUSt1 (**Fig. 1C**). The clearest increases in diversity have been driven by increases at the species and genus levels (5.3-fold and 2.2-fold increases respectively). However, all taxonomic levels exhibited increased diversity, including the phylum level where the coverage increased from 39 to 64 phyla (1.6-fold increase). PICRUSt2 predictions based on the following gene families are supported by default: Kyoto Encyclopedia of Genes and Genomes[19] (KEGG) orthologs (KO), Enzyme Classification numbers (EC numbers), Clusters of Orthologous Genes[20] (COGs), Protein families[21] (Pfam) and The Institute for Genomic Research's database of protein FAMilies[22] (TIGRFAM). While similar gene families are shared across these databases (e.g. many KOs are highly related to EC numbers), they each, nevertheless, provide novel insights (**Supp Table 1**). One key drawback of the current PICRUSt1 database is that it is based on an outdated version of the KEGG database, which is missing many recently defined gene families that are now included in the PICRUSt2 database. An increase in the total number of KOs from 6,911 to 10,543 (1.5-fold increase) was observed in the final database.

Validation of 16S rRNA gene-based metagenome predictions

To validate metagenome predictions made by PICRUSt2, we assessed four previously published datasets with paired 16S amplicon and shotgun metagenomics sequencing (MGS) data from a range of environments: (1) 116 samples spanning the human body (from the Human Microbiome

Project[23] [HMP]), (2) 8 mammalian stool samples[24], (3) 6 ocean samples[25], and (4) 22 bulk soil and blueberry rhizosphere samples[26]. The key assumption of these analyses is that the MGS data is a "gold-standard" for evaluating prediction performance. However, the true underlying microbial gene abundances of a sample are likely incompletely observed by MGS data due to inadequate sequencing depth, biases in library preparation, and the limitations of short read annotation. Nonetheless, these datasets were chosen since they represent varying degrees of challenge for accurate metagenome inference due to environmental and technical factors characteristic of microbiome studies (**Supp Table 2**). Generating ASVs for these datasets resulted in a total of 1865, 323, 1148, and 3333 ASVs for the HMP, mammalian, ocean, and soil datasets, respectively.

We previously developed the nearest sequenced taxon index (NSTI) as a metric for summarizing a microbial taxonomic profile's novelty relative to isolate genomes. This measures the abundance-weighted distances in the phylogenetic tree between taxa (ASVs) from a community and the tips of the nearest sequenced neighbours. These NSTI distributions are calculated automatically in PICRUSt2 and differed significantly among the evaluation datasets, demonstrating their range of "unusualness" relative to sequenced isolates (Kruskal-Wallis $\chi 2$=6,504.3, P= 6.7×10^{-16}; **Supp Fig 2A and B**). Datasets from more well-characterized communities have lower mean NSTI values overall as expected, ranging from 0.11 (sd: 0.49) in the HMP dataset to 0.51 (sd: 2.06) in the ocean dataset. A maximum NSTI cut-off of 2 is implemented by default in PICRUSt2, as a guideline to prevent unconsidered interpretation of overly-speculative inferences, which resulted in a mean of 0.4% of ASVs being excluded across these datasets. These excluded ASVs correspond to either eukaryotic sequences or microbial phyla with no reference genomes available.

We generated PICRUSt2 KO predictions for each dataset and compared them to KO abundances profiled from the corresponding MGS metagenomes (see Online Methods). We also predicted KO abundances with four alternative hidden-state prediction pipelines: PICRUSt1, Piphillin, PanFP, and Tax4Fun. In addition, we ran PICRUSt2 on the table and sequences prepared for input into PICRUSt1 (closed-reference OTU-picking against the Greengenes database, see Online Methods) to investigate whether different pre-processing pipelines affected the predictions. We calculated Spearman correlation coefficients (hereafter "correlations") between the predicted KO abundances and MGS KO abundance tables for each sample after filtering all tables to the 5,996 shared KOs present within each tested HSP tool database (**Fig 2**). The correlation metric represents the similarity in rank ordering of KO abundances between the predicted and observed data. For all four datasets, PICRUSt2 predictions yielded significantly higher mean correlations compared to the other tools (paired-sample, two-tailed Wilcoxon tests [PTW] $P < 0.05$) with means of 0.859 (sd = 0.036), 0.830 (sd = 0.0184), 0.874 (sd = 0.007), and 0.855 (sd = 0.017) for the HMP, mammalian, ocean, and soil datasets, respectively. Notably, the PICRUSt2 predictions based on the input files prepped with an OTU-picking pipeline resulted in significantly higher correlations compared to the ASV pipelines for the HMP and ocean datasets (PTW $P < 0.05$). We also explored how the correlation performance shifts as the NSTI cut-off is changed and found that correlations decrease as the NSTI cut-off decreases over all datasets (**Supp Fig 3**).

Gene families regularly co-occur within genomes, and so the use of correlations to assess gene-table similarity may be limited by the lack of independence of gene families within a sample. For instance, a high covariance in gene family abundances can occur simply because genomes are composed of similar gene families (**Supp Fig 4**; see Online Methods). To address

this dependency, we compared the observed correlations between paired MGS and predicted metagenomes to correlations between MGS functions and the mean gene family abundance across all reference genomes ("Null"). For all datasets and tested NSTI cut-offs, PICRUSt2 metagenome tables were more similar to MGS values than the null expectation (**Fig 2**).

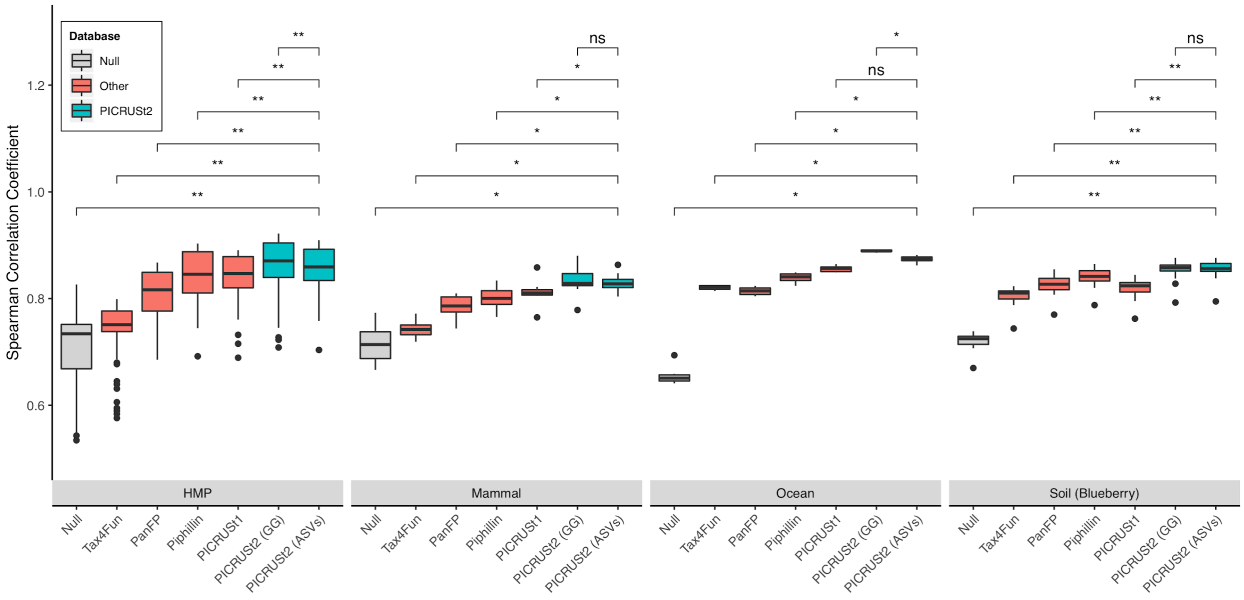

Figure 2: PICRUSt2 prediction outperforms existing HSP tools based on Spearman correlation coefficients. Validation results of PICRUSt2 comparing metagenome prediction performance against gold-standard shotgun metagenomic sequencing (MGS). Boxplots represent medians and interquartile ranges of Spearman correlation coefficients observed in the human microbiome project (HMP, n=116), mammalian stool (n=8), ocean water (n=6), and blueberry soil (n=22) datasets. PICRUSt2 predictions are shown both for when input data are restricted to Greengenes reference operational taxonomic units (GG) and when amplicon sequence variants (ASVs) are input, to differentiate algorithmic vs. database improvements. The significance of paired-sample, two-tailed Wilcoxon tests is indicated above each tested grouping (*, **, and ns correspond to $P < 0.05$, $P < 0.001$, and not significant respectively). Note that the y-axis is truncated at 0.5 rather than 0 to better visualize small differences between categories.

Next, we investigated metagenome predictions based on the presence and absence of output KOs (**Supp Fig 5**). We assessed relative differences in precision and recall (see Online Methods) to compare PICRUSt2 with alternative approaches. The resulting gene abundance tables were first rounded to the nearest integer before converting to binary, presence-absence tables of 1s and 0s. In this analysis, null prediction values indicate the baseline effect of gene family conservation on precision or recall prior to metagenome prediction. PICRUSt2 predictions made with an NSTI cut-off of 2 exhibited significantly higher precision against all other HSP approaches over all four datasets with the exception of PICRUSt1 mammalian predictions (**Supp Fig 5**; PTW $P < 0.05$; mean and sd of precision for predictions per dataset: HMP [0.857, 0.052], mammal [0.792, 0.062], ocean [0.830, 0.021], and soil [0.699, 0.047]). In contrast to the Spearman correlation coefficient analyses, PICRUSt2 predictions made on closed-reference OTUs exhibited significantly less precision than ASVs as well (PTW $P < 0.05$). In addition, precision increased with decreasing NSTI cut-off (**Supp Fig 5**). These analyses demonstrate that PICRUSt2 exhibits consistently high recall and precision in all datasets and that a clear trade-off between these two metrics was evident in each dataset. Overall, PICRUSt2 compromised the least between these two metrics in each sampling environment.

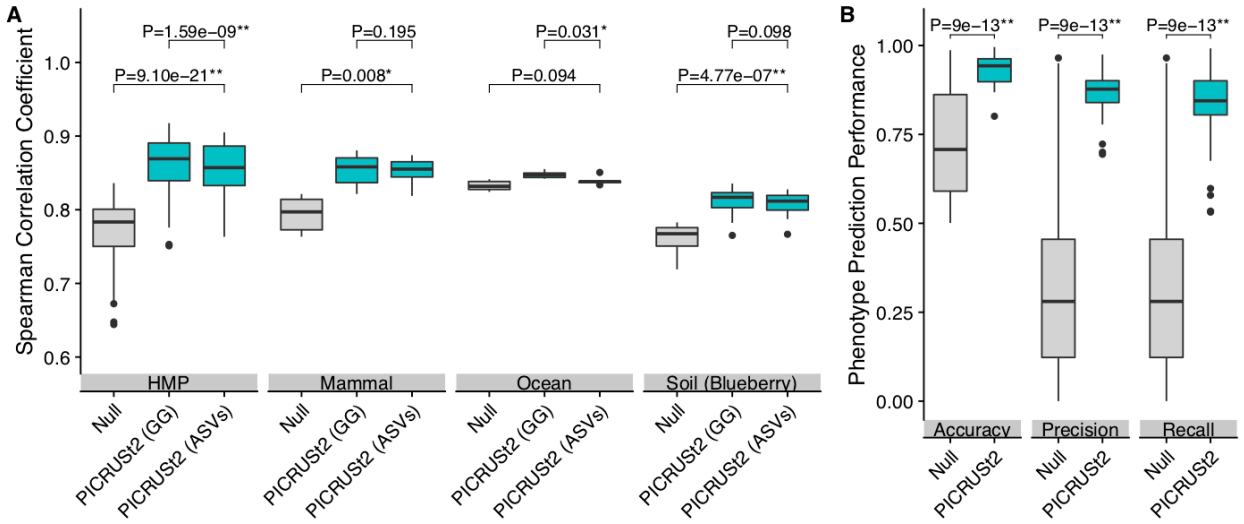

Figure 3: PICRUSt2 accurately predicts MetaCyc pathways and phenotypic data. (A) Spearman correlation coefficients between PICRUSt2 predicted pathway abundances and gold-standard metagenomic sequencing (MGS). Results are shown for each dataset: The Human Microbiome Project (HMP), mammalian stool, ocean water, and blueberry soil. The PICRUSt2 predictions are shown for when input data is restricted to Greengenes reference operational taxonomic units (GG) or when amplicon sequence variants (ASVs) are input. (B) Performance of binary phenotype predictions based on three metrics: accuracy, precision, and recall (see Online Methods). 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 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 at 0.5 rather than 0 to better visualize small differences between categories. The sample sizes are 116, 8, 6, and 22 for the HMP, mammal, ocean, and soil datasets, respectively.

We repeated these validation analyses for EC gene family predictions and MetaCyc[27] pathway abundance predictions for each dataset. We compared the performance of EC predictions to PAPRICA, which is another EC-based 16S HSP tool. EC-based PICRUSt2 prediction results were consistent with KO-based results, and correlations were significantly

higher than PAPRICA for all datasets (PTW $P < 0.05$; **Supp Fig 6**). In addition, the recall of PICRUSt2 predictions was significantly higher than PAPRICA although this came at a cost of prediction precision (PTW $P < 0.05$; **Supp Fig 7**). MetaCyc pathway abundances, calculated using structured mappings of EC gene families to pathways, also performed better than the null distribution for all metrics overall (PTW $P < 0.05$; **Fig 3A** and **Supp Fig 8-9**). One exception was the finding of no significant difference between the null (mean: 0.833; sd: 0.007) and ASV-based PICRUSt2 correlations (mean: 0.839; sd: 0.006) in the ocean dataset (PTW P=0.094; **Fig 3A**). However, the predictions for this dataset are clearly different from the null in terms of precision and recall (**Supp Fig 9**), which is reflected by a significantly higher mean F1 score of 0.881 (sd: 0.006) compared to 0.694 (sd: 0.004) in the null (PTW P=0.03).

Lastly, we validated the predictions of 41 microbial physiology phenotypes that are available for IMG genomes[28]. These represent higher-level microbial metabolic activities such as "Glucose utilizing" and "Denitrifier" that are annotated as present or absent within each reference genome (see Online Methods). This database was motivated by the predictions made by the tools FAPROTAX[29] and Bugbase[30] although the exact phenotypic predictions are not directly comparable. We performed a hold-out validation to assess the performance of PICRUSt2 phenotype predictions, which involved comparing the binary phenotype predictions to the expected phenotypes for each reference genome. Based on overall accuracy (mean=93.3%; sd=4.40%), precision (mean=86.5%; sd=6.21%), and recall (mean=83.5%; sd=11.4%), these predictions performed significantly better than the null expectation (**Fig 3B**; Wilcoxon tests $P < 0.05$).

Validation of fungal metagenome inference

We next assessed PICRUSt2's new capabilities for predicting metagenomes based on fungal amplicon sequencing of either the 18S rRNA gene (18S) or internal transcribed spacer (ITS) regions. Data used for these predictions included EC abundances from 294 fungal genomes from the 1000 Fungal Genomes Project that were publicly available as of November 16, 2018 and passed quality control criteria (see Online Methods). Unlike the prokaryotic database, only a minority of 18S and ITS sequences were redundant across genomes (7.5% and 8.5%, respectively). A total of 7 and 8 phyla as well as 183 and 209 genomes are represented in the ITS and 18S databases, respectively (see **Supp Table 3** for the database counts at all taxonomic levels).

We first evaluated the performance of PICRUSt2 18S and ITS metagenome predictions by leave-one-out cross-validation of individual genomes. Spearman correlations were calculated between the predicted EC abundance profiles in each held-out genome and the EC abundances in the known genome. For both the 18S (Spearman Rho mean=0.821; sd=0.141) and ITS databases (Spearman Rho mean=0.822; sd=0.135), the predictions were significantly better than the null expectation (Wilcoxon test $P < 0.001$; **Supp Fig 10**). Similar to the 16S-based validations, genome prediction accuracy decreased as reference genomes were artificially held out of the training dataset at increasing taxonomic scale, suggesting that overall accuracy is hampered for those lineages without comprehensive representative genomes.

Next, we evaluated the performance of fungal EC predictions on two amplicon sequencing datasets with paired MGS data using the same approach as with 16S predictions. The two validation datasets used were the same 22 blueberry soil samples described above, which also underwent 18S sequencing[26], and eight wine fermentation internal transcribed spacer (ITS1)

sequencing samples[31]. EC predictions for both of these datasets were significantly more similar to the MGS gold-standard compared to the null expectation based on correlations (**Fig 4A**; $P=9.5x10^{-7}$ and $P=7.8x10^{-3}$ for the blueberry soil and wine fermentation datasets respectively). However, the correlations observed for these datasets was substantially lower than for the 16S-based validations described above, which is to be expected as these metagenomes include a substantial amount of functions from non-fungal origins. The mean correlations in the blueberry soil dataset were 0.340 (sd: 0.016) and 0.365 (sd: 0.019) for the null expectation and PICRUSt2 predictions, respectively. There was a larger mean difference for the wine fermentation dataset where the mean correlations were 0.492 (sd: 0.004) and 0.611 (sd: 0.028) for the null expectation and PICRUSt2 predictions, respectively. Interestingly, the correlations based on predicted MetaCyc pathway abundances were slightly lower than the null values for the blueberry soil 0.461 (sd: 0.009) and wine fermentation 0.501 (sd: 0.008) datasets (**Fig 4B**).

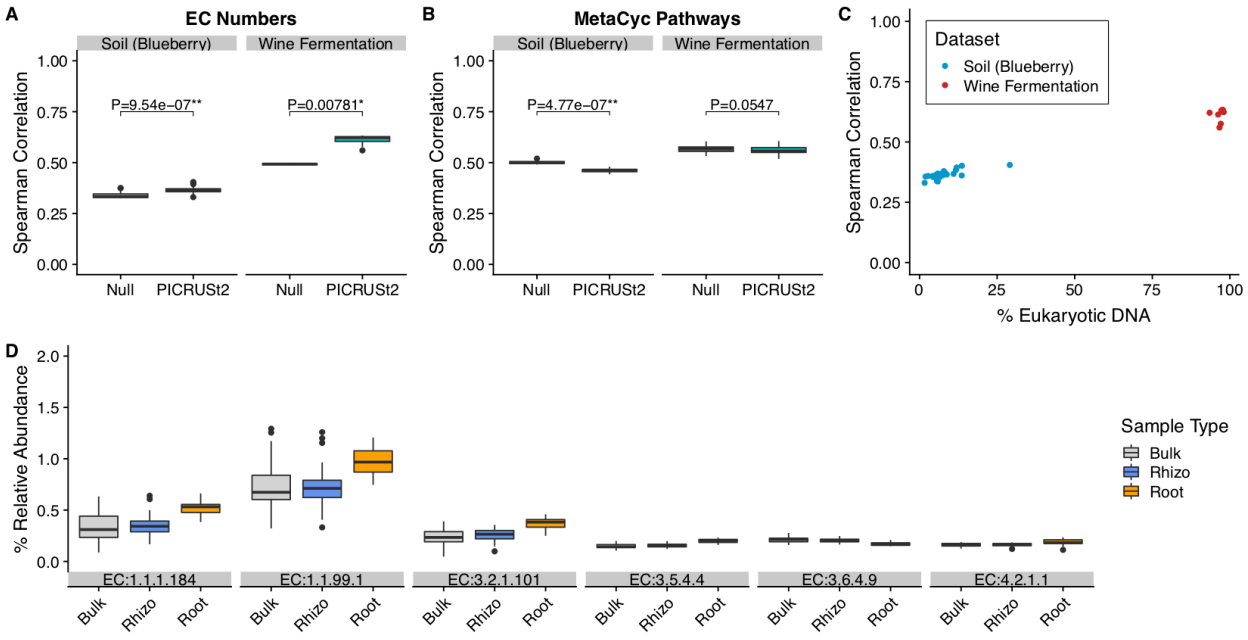

Figure 4: PICRUSt2 18S rRNA gene and internal transcribed spacer predictions exceed null prediction accuracy. (A) Spearman correlation coefficients between amplicon predicted Enzyme Classification number abundances and gold-standard shotgun metagenomic (MGS) profiles from the same biological samples. (B) Spearman correlation coefficients between amplicon-predicted MetaCyc pathway abundances and MGS on the same biological samples. For panels A and B, 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). (C) The Spearman correlation coefficients as shown in panel A re-plotted against the percent of non-animal and non-plant eukaryotic DNA within each sample. The blueberry soil dataset consists of 22 18S rRNA gene sequencing samples and the wine fermentation dataset consists of eight internal transcribed spacer region one (ITS1) sequencing samples. (D) The relative abundance of significantly informative EC numbers ($P < 0.001$) in a Random Forest model for distinguishing blueberry soil and root samples by sample type. Only significant EC numbers with a mean relative abundance greater than 0.15% are shown. The EC numbers shown correspond to carbonyl reductase (NADPH) (EC:1.1.1.184), choline dehydrogenase (EC:1.1.99.1), mannan endo-1,6-alpha-mannosidase (EC:3.2.1.101), adenosine deaminase (EC:3.5.4.4), chaperonin ATPase (EC:3.6.4.9), and carbonate dehydratase (EC:4.2.1.1). There are 26, 33, and 32 samples for the bulk, rhizosphere (rhizo), and root environments, respectively.

One potential factor affecting these results is the percent of eukaryotic DNA within the MGS data. A low percent of eukaryotic DNA would result in prokaryotes mainly contributing to gene family abundances. The percent of eukaryotic DNA (after excluding plant and animal DNA) within the MGS datasets differed dramatically between the blueberry soil (mean: 8.17%; sd: 5.82) and wine fermentation datasets (mean: 96.72%; sd: 1.44; **Fig 4C**). This low percent of eukaryotic DNA in the blueberry soil dataset could partially account for the relatively poor performance we observed (**Fig 4A**).

To nonetheless show that the PICRUSt2 predictions for the blueberry soil dataset are informative, we ran PICRUSt2 on 18S sequencing data from additional blueberry soil samples with no matching MGS data as well as blueberry root samples from the same dataset. We then

generated a Random Forest model to identify the most informative predicted EC numbers that distinguish samples by whether they were taken from a bulk soil, rhizosphere, or root environment. This model resulted in a classification accuracy of 68%, which was substantially better than the random expectation (accuracy 36%) and identified 32 significantly informative EC numbers ($P < 0.001$; see Online Methods; **Fig 4D**).

Functional profiling of inflammatory bowel disease using PICRUSt2

Finally, to demonstrate the utility of PICRUSt2 in making functional inferences in a human health context where only amplicon sequencing is feasible, we profiled 27 ileal biopsy samples from subjects with Crohn's disease (CD) and 20 control subjects (non-IBD). These data are a subset of the Inflammatory Bowel Disease Multi'Omics Database[32], which provides "multi-omics" data to identify host and microbial features associated with inflammatory bowel disease (IBD). Our analysis was based on 16S marker gene libraries collected from biopsy samples in this dataset. Importantly, MGS could not practically be performed on these (or any typical) biopsy samples due to the overwhelming predominance of human host DNA, which competes with microbial DNA for sequencing reads. As such, MGS data was produced only for subject stool samples for this study, which are analyzed here in addition to accompanying human RNA-seq transcriptional profiles (from the same biopsies) and metabolomic profiles from paired stool.

We first analyzed these data with a typical testing framework for identifying significant microbial features: testing for significantly differentially abundant ASVs clustered by taxonomy as well as inferred pathway abundances. Based on this standard approach we identified no pathways that significantly differed between CD and non-IBD subjects (FDR < 0.1; see Online Methods). However, four taxa were identified with a differential relative abundance between

non-IBD and CD subjects based on a lenient FDR q-value cut-off of 0.2 (**Supp Fig 11**). These included three taxa within the Clostridiales order, which were increased in relative abundance in control subjects, and the phylum Proteobacteria at higher relative abundance in CD subjects.

We next focused on the predicted MetaCyc pathways inferred by PICRUSt2 for ASVs underlying the significantly differentially abundant taxa: (1) 35 significant Clostridiales ASVs and (2) 192 Proteobacteria ASVs. For each predicted pathway in the community, we calculated the ratio of the abundance of that pathway contributed (i.e. potentially produced) by ASVs within the group of interest compared to the pathway's abundance contributed by all other ASVs. Based on this approach we identified 3 pathways significantly contributed by Proteobacteria (**Fig 5A**; Wilcoxon test FDR < 0.05). In addition, the relative contribution to 75 pathways by Clostridiales significantly differed between CD and non-IBD subjects (**Supp Fig 12**; Wilcoxon test FDR < 0.05). These results demonstrate how PICRUSt2 stratified outputs allow integration of functional predictions with taxonomic findings as opposed to treating the two independently as in non-contributor stratified metagenomic analyses.

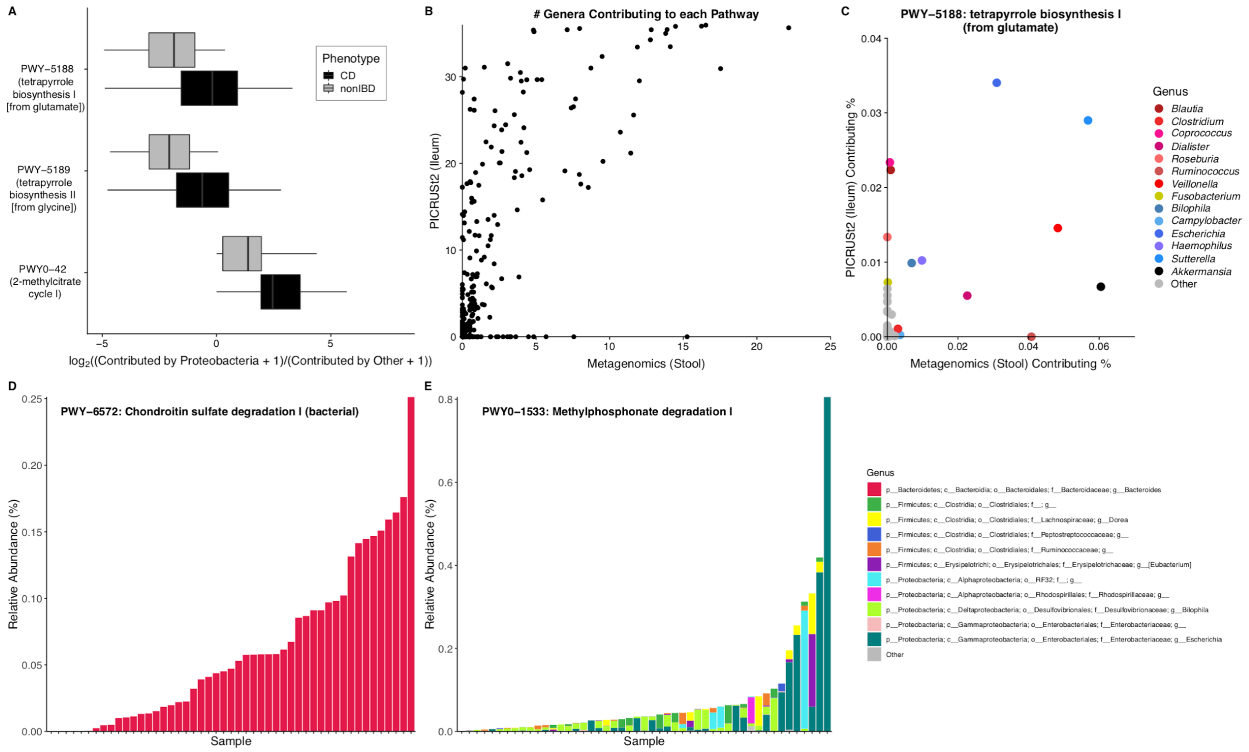

Figure 5: Applying PICRUSt2 to Crohn's disease cohort yields novel insights. (A) Predicted MetaCyc pathways significantly contributed by Proteobacteria in the ileum of subjects with Crohn's disease (CD, n=27) compared to healthy controls (non-IBD, n=20) based on PICRUSt2 inference from 16S rRNA gene (16S) sequencing. These significant pathways are: PWY-5188 (tetrapyrrole biosynthesis I [from glutamate]), PWY-5189 (tetrapyrrole biosynthesis II [from glycine]), and PWY0-42 (2-methylcitrate cycle I). (B) The mean number of classified genera contributing to each of the 313 MetaCyc pathways identified in either the ileum 16S sequencing (by PICRUSt2) or shotgun metagenomics sequencing (MGS) of the stool of the same CD subjects. (C) The top classified genera contributing to the relative abundance of PWY-5188 based on PICRUSt2 predictions of 16S sequencing in ileum tissue or MGS of the stool of the same subjects. Only the top 10 contributing genera are shown. Genera of the same phylum are shades of the same colour (Firmicutes and Proteobacteria are shades of red and blue, respectively). (D and E) Taxonomic breakdown of genera contributing to the predicted relative abundance of (D) PWY-6572 and (E) PWY0-1533 across all CD samples. Unclassified genera were included in these stacked bar charts, unlike in panel C. Only the top ten genera contributing to either PWY-6572 or PWY0-1533 are labelled.

We next investigated whether analysis of stool MGS data rather than ileal PICRUSt2 predictions resulted in substantially different conclusions, either due to methodology or body site. The number of classified genera contributing to each pathway within CD subjects differed strikingly depending on whether the contributors were identified through ileal 16S sequencing or stool MGS (mean difference: 7.3; sd: 10.2; **Fig 5B**). While a small number of pathways were uniquely identified by stool MGS, for most pathways a greater number of contributing taxa were identified by PICRUSt2. This is most likely due to the much greater taxonomic diversity accessible through amplicon databases than through reference genome isolates. This result could also be due to biological differences between the stool and ileal samples. However, we also identified a similar trend in all four paired 16S-MGS datasets, although the magnitude of difference depended greatly on sequencing technology and sampling environment (**Supp Figure 13**). Not just the number of taxa, but also their identities, differed between stool metagenomes versus biopsy inferences. For instance, for the tetrapyrrole biosynthesis I (from glutamate) pathway (PWY-5188), the top contributors differ between phylum Proteobacteria in biopsy 16S profiles, while *Akkermansia* (in phylum Verrucomicrobia) is the top contributor identified in the MGS data (**Fig 5B**).

Last, we tested whether the PICRUSt2 predictions give novel insights into CD biomarkers by associating 207 predicted pathways with both 583 metabolites from paired stool metabolomic profiles and the ileal transcription levels for six human host genes of interest (see Online Methods). We identified no significant associations between predicted pathway and stool metabolite levels, but 29 associations between the predicted pathway and ileal transcript levels (FDR < 0.1; **Supp Table 4**). Some of these significant associations are driven largely by individual taxa. For example, since the predicted relative abundance of chondroitin sulfate

degradation is entirely contributed by *Bacteroides* (**Fig 5D**), the association between this pathway and NAT8 expression (partial R=-0.58) is trivially due to the relative abundance of this genus. However, not all significant associations are driven by individual taxa. For instance, there is no single taxon driving the association of the predicted relative abundance of the methylphosphonate degradation I pathway with MMP3 expression (partial R=-0.62; **Fig 5E**). This association is an example of PICRUSt2 predictions yielding potentially novel insights beyond those of the originating amplicon-based taxonomic profiles.

Discussion

We developed the PICRUSt2 algorithm as a major update to the widely used PICRUSt1 method for functional inference from amplicon-derived taxonomic profiles. The new method introduces capabilities for 18S and ITS amplicon profiling, phylogeny-based ASV analysis, a greatly expanded reference dataset, and novel functional and phenotype predictions. PICRUSt2 significantly outperformed both PICRUSt1 and other HSP tools on all four paired 16S-MGS datasets tested. Lastly, we have applied PICRUSt2 to mucosa-associated microbial communities sequenced from ileal biopsies of human subjects with and without IBD. This identified the taxonomic contributors of inflammation-linked pathways at the mucosa, an analysis not possible from stool metagenomics or from typical biopsy-based analyses alone.

An informed understanding of amplicon-based functional inferences, their strengths and weaknesses, and the accuracies of estimated metagenomes is crucial to the application of PICRUSt2 and related methods. PICRUSt2 prediction accuracy exceeded null values, for example, in all analyses except for null precision (**Supp Fig 5A**). When evaluated using Spearman correlation coefficients, however, the non-independence of gene families from the

same organisms could greatly influence results. In other words, the copy numbers of many bacterial and archaeal gene families tend to be highly correlated across lineages. The null metagenome as a result is enriched for gene families that are shared between many reference genomes. As such, the null (when rounded to the nearest integer) should naturally observe a low false-positive incidence of gene families and a high false-negative incidence. Thus, a high null precision and low null recall were anticipated, as observed in our analyses, suggesting the need to assess these accuracy metrics simultaneously and not independently. While PICRUSt2 predictions introduced more false-positive gene families compared to the null, this is an expected side-effect of genome prediction where an inherent trade-off exists between precision and recall. Indeed, when considering the F1 score (the harmonic mean of precision and recall), KOs predicted with PICRUSt2 for the HMP dataset had a score of 0.856 compared to 0.582 for the null expectation.

We demonstrated the utility of PICRUSt2 by identifying previously identified and novel pathways and taxa linked to biomarkers of Crohn's disease within ileal biopsy samples, which are difficult to assess with MGS. The decrease and increase in the relative abundance of Clostridiales and Proteobacteria, respectively, in subjects with CD is a microbial signature of CD identified by several prior studies[33,34]. In addition, the concentration of two tetrapyrrole compounds, urobilinogen and urobilin, was recently identified to be lower in CD subjects versus controls[35]. Our analysis newly suggests this feature of CD could be related to the bloom of Proteobacteria, which warrants further investigation. More generally, we demonstrated that PICRUSt2 can access functional potential both of rare or intractable taxa not well-identified by shotgun metagenomics, and in environments such as tissue biopsies in which they are difficult to directly measure, while retaining excellent agreement with direct measurements of paired stool.

There are two natural criticisms of amplicon-based HSP. First, the predictions are biased towards existing reference genomes, which means that rare environment-specific functions are less likely to be identified. This issue will be partially addressed as the number of sequenced genomes, and especially high-quality metagenome-assembled genomes, continues to grow. PICRUSt2 allows user-generated genomes to be included in its reference database providing a flexible and extensible framework for projects focused on particular hosts or 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 HSP tools, which can only differentiate taxa to the degree they differ at the amplified marker gene sequence. However, despite this limitation, we have demonstrated that PICRUSt2 predictions have overall high accuracy and that they can yield novel biological insights. Importantly, as with PICRUSt1 and any similar methods, any insights produced by analyzing PICRUSt2 results should primarily be used for hypothesis generation; additional direct measurements are always needed to validate predictions of functional potential.

These cautions are especially true when interpreting any PICRUSt2 fungal amplicon-based predictions. Although we have shown above that the fungi EC number predictions are more accurate than random it is important to emphasize that these predictions are based on only approximately 200 fungal genomes and thus these predictions should be treated cautiously. Currently assessing fungal metagenome inference performance is challenging since the majority of paired amplicon-MGS datasets are dominated by prokaryotic DNA. Nonetheless, we have demonstrated a proof-of-concept of fungal metagenome inference that may be valuable to advanced users and will continue to improve in performance as more fungal genomes become available.

In summary, PICRUSt2 is an improved, more flexible, and accurate method for performing marker gene metagenome inference. No restrictions are imposed on input 16S amplicon sequences, which may be generated by any conventional read-processing algorithm and may now be accompanied by other amplicons such as 18S or ITS sequences. We have simplified the PICRUSt workflow for ease of use and customization, allowing genome prediction to be run on any marker gene sequencing library. Gene family predictions are made using greatly expanded reference databases as well as updated pathway mappings. These predictions can now be directly stratified by taxonomic contributors, which will improve interpretability. We hope that the substantially expanded functionality of this approach will continue to allow researchers to identify potentially novel insights into functional microbial ecology from amplicon sequencing taxonomic profiles.

Acknowledgements

We would like to thank Zhenjiang Xu for providing alternative databases of 16S sequences and bacterial gene family abundances, which were not used for this manuscript. We would also like to thank Amy Chen for helping us gain access to the IMG database files used for the PICRUSt2 default database. Lastly, we would also like to thank Heather McIntosh for her help designing the pipeline flowchart. GMD is funded by a Natural Sciences and Engineering Research Council (NSERC) Alexander Graham Bell Graduate Scholarship (Doctoral). VM is funded by an NIH/NIAAA Ruth L. Kirschstein National Research Service Award (F30 AA026527). MGIL is funded by an NSERC Discovery Grant and an NSERC Collaborative Research Development with co-funding from GlaxoSmithKline to MGIL and JB. CH is funded in part by NIH NIDDK grants U54DK102557 and R24DK110499. SYN is funded by an NSERC Discovery Grant.

Online Methods

Code availability

PICRUSt2 is available at: https://github.com/picrust/picrust2. The Python and R code used for the analyses and database construction described in this paper as well as key intermediate files are available online at https://github.com/gavinmdouglas/picrust2_manuscript.

PICRUSt pipeline updates

The analyses in this paper are based on PICRUSt2 version 2.1.0-b. In addition to the improvements reported in the Results section, several other updates have also been made to the PICRUSt pipeline. Since the HSP step is now run using the castor R package, other inference approaches like maximum parsimony (MP) may be performed in realistic time-frames besides phylogenetic independent contrasts[36], which was the default approach in PICRUSt1. The default HSP method is now MP with a parameter weighting the contribution of branch lengths set to 0.5 (*edge_exponent* option in castor package). This parameter value was chosen since setting this parameter to a non-zero value resulted in more reproducible predictions.

In addition, now that any study sequences can be input to PICRUSt, and not just Greengenes closed-reference OTUs, an NSTI screening step is recommended to eliminate sequences above a certain cut-off. The default NSTI cut-off in PICRUSt2 is 2, which was chosen as an extremely lenient cut-off intended to eliminate problematic sequences. Only one ASV in the IBD example dataset was above this cut-off, which corresponded to a mitochondrial sequence. The only sequences above this cut-off in the HMP validation dataset corresponded to two 18S ASVs that were clustered within the 16S dataset. Similarly, although 13/1148 of ASVs

in the ocean dataset were above the NSTI cut-off of 2, these ASVs corresponded to candidate taxonomic groups that have no representative reference genomes in the default PICRUSt2 database. Based on these observations, we believe this cut-off should be suitable for most scenarios; however, users can select a NSTI value that best fits their study design and environment (i.e. whether to maximize precision or recall).

Transforming gene family predictions to pathway abundances in PICRUSt1 was done by assuming that the abundance of each gene family contributed equal abundance to all pathways containing the gene family (i.e. if a gene family can be involved in 10 pathways the gene family abundance would be added equally to the abundance of all 10 pathways). Although this approach is easy to understand, it results in a high false-positive rate of identifying pathways present. To improve on this approach, we adapted the approach taken by HUMAnN2[37] v0.11.1 into the PICRUSt2 pipeline. MinPath[38] (v1.2 as modified for the HMP workflow[39]) is first run to identify the minimum pathways present given the gene families present. By default, these predictions are made based on the EC number predictions after regrouping them to MetaCyc reactions to predict MetaCyc pathway abundances. The mappings files and code for regrouping to MetaCyc reactions and mapping from reactions to structured pathways were taken from HUMAnN2. We further split the pathway mapping files into prokaryotic and fungal sets based on the taxa where these pathways have been identified (as reported in the MetaCyc online database).

16S database processing

A total of 52,217 genomes were parsed for 16S sequence and gene family counts from IMG on 8 Nov. 2017. Genomes lacking a 16S sequence length of at least 1,250 bp or that were identified as eukaryotic marker genes were removed. The gene families in these annotations corresponded to

these databases: KEGG (v77.1), Pfam (v30), TIGRFAM (v15), COG (v2014), EC numbers (as of 21 Jan 2016).

We also created an alternative trait database containing phenotypes defined by IMG[28]. These phenotypes are more directly interpretable than the gene family databases described above. Files listing IMG genome ids positives for one of 65 phenotypes were downloaded on 8 Jan 2019. The presence and absence of phenotypes was re-coded as 1 and 0. Prototrophic and auxotrophic phenotypes for the same compound were combined into a single prototrophic phenotype (auxotrophs are coded as 0 and unknown phenotypes are coded as NA). After this merging step, and removing two extremely rare phenotypes, there was a final set of 41 phenotypes remaining.

To identify low-quality, incomplete genomes, and possible misassembled assemblies we calculated the median number of single-copy KEGG orthologs (KO) as previously identified for the tool MUSiCC[40]. Since these genes are expected to be found in single-copies within each genome, we reasoned that incomplete or contaminated genomes could be identified by a median copy number less or greater than 1, respectively. Accordingly, we discarded all genomes with a median number of single-copy genes that differed from 1. In addition, we discarded genomes lacking a sufficient number of genes within any gene family. The minimum number of gene families per genome was chosen based on visualizing the distribution of gene family numbers over all genomes and choosing a cut-off that eliminated outliers (minimum cut-offs of unique gene families were 500, 250, 500, 750, and 350 for the COG, EC, KO, Pfam, and TIGRFAM databases). Importantly, this filtering means that endosymbionts and other organisms with reduced genome sizes will be underrepresented in the PICRUSt2 reference database. After these

filtering steps, a total of 10,291 genomes were discarded, producing a final set of 41,926 genomes.

Since prokaryotes often have multiple 16S rRNA gene copies, the centroid 16S sequence per genome was identified using the VSEARCH[41] (v2.4.4) *cluster-fast* command with an identity cut-off of 90%. In cases where multiple centroid sequences were found, a single centroid was chosen randomly. The final 16S sequences were used to build a multiple sequence alignment (MSA) using ssu-align (v0.1.1; http://eddylab.org/software/ssu-align/) against the bacteria alignment model. Only weak masking of this output MSA was performed (*ssu-mask* options: *--pf 0.001 --pt 0*). The custom Python script *derep_fasta.py* was used to identify sequences in this alignment after this masking step. A phylogenetic tree was built from this MSA with RAxML-ng[42] (v0.6.0) using the *GTR+G* model. The custom Python script *mean_16S_function_counts.py* was used to calculate the mean gene family abundances for all sequences within a cluster. These values were then rounded to the nearest integer. The full NCBI taxonomic lineage of all 16S clusters was called by the taxizedb R package (https://github.com/ropensci/taxizedb) using the species name provided by the IMG FASTA metadata. Gene family trait depths were calculated using the castor R package function *get_trait_depth* with default settings, which is based on the consenTRAIT metric[43].

18S and ITS database processing

A total of 574 publicly available fungi genomes were downloaded from the 1000 Fungal Genomes Database (http://1000.fungalgenomes.org) on November 16, 2018. The 18S rRNA genes were annotated using barrnap (v0.9-dev; https://github.com/tseemann/barrnap), and 18S sequences were parsed from the genomes using the custom Python script *rRNA_from_gff3.py*.

ITS sequences were identified and parsed from all genomes using ITSx[44] (v1.0.11) using the *--only_full T* and *–heuristics* options. Sequence length cut-offs for the 18S and ITS sequences were 605-3,076 bp and 146-2,570 bp, respectively. BUSCO[45] (3.0.2) was run to identify incomplete and contaminated genomes with the *fungi_odb9* database. Only the genomes with a completeness of at least 70%, and a duplicated metric (which is based on the copy number of single-copy genes) of at maximum 10% were retained. After restricting genomes to those that passed these quality cut-offs that also had at least one passing amplicon region, there were 229 genomes in the 18S database and 201 genomes in the ITS database.

The 18S and ITS sequences were then dereplicated as above for the 16S sequences. The 18S MSA was built using the ssu-align pipeline as above (using the *eukarya* alignment model) whereas the ITS MSA was built using MAFFT[46] (v7.407) with the *–genafpair* and *–maxiterate 1000* options. Phylogenetic trees for both MSAs were built with RAxML-ng (v0.8.0) as above except a guide tree enforcing a taxonomic topology was also used. EC number copy numbers per genome were downloaded for these genomes also from the 1000 Fungal Genomes Database. Mean EC number abundances were calculated for dereplicated amplicon sequences using the same the approach as above for the prokaryotic databases.

Amplicon dataset processing

An in-depth comparison of the technologies and sequencing depths for each of the four 16S validation datasets is shown in **Supplementary Table 2**. The processing pipelines and filtering criteria differed for each dataset due to technical differences between them. The key difference was that DADA2[7] was run for the HMP dataset since this was Roche 454 sequence data. Deblur[9] was run on all other validation datasets since they were Illumina sequence data (**Supp Table 2**).

The HMP raw data can be download from https://www.hmpdacc.org/HMIWGS/healthy/ (HMP). The mammalian stool sequencing data can be downloaded from the Short Read Archive (SRA) under accessions SRP115632 (MGS) and SRP115643 (16S). The ocean sample sequencing data can be downloaded from SRA project SRP056891. The blueberry soil samples are available at SRA project accessions PRJNA389786 (16S) and PRJNA391782 (18S). The blueberry root samples are available under SRA project accessions PRJNA434066 (16S) and PRJNA434067 (18S). The matching MGS data for these blueberry root and soil samples is under accession PRJNA484230. The processed ITS output files for the wine fermentation dataset are online as part of a GigaDB dataset (http://gigadb.org/dataset/100309), and the raw MGS data is available as part of SRA project accession PRJNA305659.

The HMP reads were filtered using DADA2 (v1.6.0) options. Denoising these sequences with DADA2 resulted in 1,865 ASVs after discarding ASVs with a minimum frequency of 10 and discarding 17 samples with fewer than 2,000 reads. The reverse complement of these sequences was taken before running PICRUSt2. The mammalian stool dataset was run using deblur with the default options in QIIME 2[12] (v2017.12), which resulted in 323 ASVs after discarding 2 samples with final read counts less than 3,220. The ocean dataset was also processed with deblur and no post-processing was done since all samples had high depth (minimum of 62,498 reads), which resulted in 1,148 final ASVs. The blueberry soil 16S dataset was also processed with deblur, and all 22 samples were retained since the minimum depth was 5,550 reads, which resulted in a total of 3,333 ASVs. Using the same approach except also restricting the output ASVs to those classified as fungi, the blueberry soil 18S data contained a total of 1,048 ASVs and a minimum sample depth of 1,981 reads was output. A total of 3,691 ASVs and a minimum depth of 2,091 ASVs was produced when re-running this pipeline with all

blueberry-associated 18S samples (i.e. including blueberry root and soil samples with no matching MGS data). The R package rfPermute (v2.1.6) was run with 501 trees, 1,000 replicates, and the default *mtry* setting to identify significantly different predicted pathways between the three environments based on PICRUSt2 predictions run on this full blueberry-associated dataset. Previously clustered ITS sequences and a processed abundance table were acquired for the wine fermentation dataset[31]. These files were used since no raw ITS reads could be located for this dataset.

Before running PICRUSt1 and Tax4Fun on each dataset, the 16S reads were separately processed with QIIME1 pipelines[47] (v1.9.0). Briefly, these reads were joined with PEAR[48] (v0.9.10), and chimeric reads were removed using the VSEARCH implementation of the UCHIME algorithm[49]. Open-reference OTU-picking at 97% identity was run against Greengenes v13_8 for PICRUSt1 and against SILVA[10] (v123) for Tax4Fun input. Closed-reference OTUs were retained, and the subsequent BIOM table was filtered based on the same criteria as above. The alternative HSP tools discussed in this paper were the following versions: PICRUSt version 1.1.3 with the default reference database of pre-calculated predictions, Tax4Fun version 0.3.1, PanFP from GitHub commit: 1f49bd1b7341b47d46fa7eaa45d7771044d0efde, Piphillin online interface as of 19 Sept. 2018 at http://secondgenome.com/solutions/resources/data-analysis-tools/piphillin/, and PAPRICA version 0.4.0e.

Shotgun metagenomics sequencing validation dataset processing

All four shotgun metagenomic sequencing (MGS) datasets were processed using the same pipeline, which is described below. Each dataset was filtered using kneaddata (v0.6.1; https://bitbucket.org/biobakery/kneaddata/wiki/Home) to run (1) Trimmomatic[50] (v0.36) to

exclude low-quality reads with the options *SLIDINGWINDOW:4:20* and *MINLEN:50* and (2) bowtie2[51] (v2.3.2) to exclude reads that mapped to the human and PhiX genomes with the options *–very-sensitive* and *--dovetail*. For the blueberry-associated samples we also mapped reads against the northern highbush blueberry (*Vaccinium corymbosum)* genome[52] version W8520 (downloaded from https://www.vaccinium.org on October 29, 2018) to exclude additional contaminant reads. For samples with paired-end reads the forward and reverse reads were concatenated into the same file. HUMAnN2 was then run to identify the abundances of annotated UniRef50 gene families in each sample. The abundances of gene families were regrouped into other gene family databases as indicated in the text using *humann2_regroup_table*. Pathway abundances and coverages were produced by the default HUMAnN2 mapping files. These steps were parallelized when possible with GNU Parallel[53] (v20170722). The percent of eukaryotic DNA in the MGS data was identified with Metaxa2[54] (v2.2), which parses rRNA genes from the raw reads.

16S-MGS validation analyses

The simulation approach to illustrate the issue of high null Spearman correlation coefficients was based on the following procedure. First, for each *N* in the set of integers that span 1 to 100, two subsets of *N* genomes were sampled randomly from the reference database. The abundance of each sampled genome was taken from a negative binomial distribution family implemented in the R function, *rnbinom*, with parameters *size=10* and *prob=0.7*, which aimed to simulate the over-dispersed count data frequently encountered in sequenced data sets. Gene family abundance tables were then computed for each of the two subsets of genomes based on the abundance of each genome and the abundances of gene families within each genome. Spearman correlation

coefficients were then computed between the two gene family tables. This procedure was replicated 10 times for each *N*.

This simulation inspired the use of null distributions in our validation analyses. We calculated the correlations between the MGS metagenome or gene table and a synthetic gene table comprised of the mean gene count number across all reference genomes in the database, which is referred to as the "null expectation" through-out the main-text. The null Spearman correlation coefficient distributions of pathway abundances and coverages were similarly based on the reference genome pathways inferred from the EC reference database. For the purposes of comparing HSP tools, gene family tables were filtered to only those gene families present in the databases of all tested HSP tools. Gene families absent in all samples were retained as 0s and were not removed. We converted the predictions to binary presence (positive) and absence (negative) format before calculating precision and recall. Precision and recall are defined as $TP/\,(TP\ +\ FP)$ and $TP/\,(TP+\ FN)$, respectively, where TP=number of true positives (i.e. functions correctly called as present), FP=number of false positives, TN=number of true negatives, and FN=number of false negatives. The F1 score is the harmonic mean of precision and recall.

IBD data processing and analyses

Raw 16S reads and processed human host transcriptome and metabolome tables were downloaded from https://ibdmdb.org. The 16S data was processed using deblur and QIIME2 as described above for the validation datasets. Only ASVs found in at least 2 samples and called by at least 10 reads were retained, which resulted in 1,419 final ASVs. The MGS raw reads were processed using the same workflow as the validation datasets described above. PICRUSt2 was

run with default options except for the option *--per_sequence_contrib*, which was set to get pathway abundances within each predicted genome for ASV-specific analyses. The unstratified pathway abundances were then calculated by summing over the pathways contributed by each ASV within each sample. Only features present in at least 33% of samples were retained for all analyses. In addition, any pathways described as "superpathways" or "engineered" were excluded from these analyses. ALDEx2[55] (v1.12.0) was run with default options to identify features at differential relative abundance between Crohn's disease and control subjects for taxa and pathways independently.

Partial Spearman correlations between predicted pathway abundances and both the metabolomic and transcriptome data was conducted with the R package ppcor (v1.1). Subject consent age was controlled for when calculating the partial correlations. Before calculating these correlations, pathway abundance data was first transformed by the arcsine square-root transformation, and the metabolomic and transcriptomic datasets were transformed by $\log_{10}$ after adding a pseudocount of 1. Metabolites were limited to those with non-empty compound names and the gene expression data was limited to 11 genes known to be biomarkers of CD-associated ileal inflammation[56]: DUOXA2, MMP3, AQP9, IL8, DUOX2, APOA1, NAT8, AGXT2, CUBN, FAM151A, and NOD2. Since several of these genes are highly correlated, we removed redundant genes based on hierarchical clustering of the complement of Spearman correlation coefficients between all genes. Six clear clusters of genes were then identified and the following six representative genes for each cluster were retained for further analyses (the other genes in each cluster are indicated in parentheses): DUOX2 (DUOXA2), MMP3, AQP9 (IL8), APOA1, NAT8 (AGXT2, CUBN, FAM151A), and NOD2.

References

1. Langille, M. G. I. *et al.* Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nat. Biotechnol.* **31**, 814–821 (2013).

2. Aßhauer, K. P., Wemheuer, B., Daniel, R. & Meinicke, P. Tax4Fun: Predicting functional profiles from metagenomic 16S rRNA data. *Bioinformatics* **31**, 2882–2884 (2015).

3. Jun, S. R., Robeson, M. S., Hauser, L. J., Schadt, C. W. & Gorin, A. A. PanFP: Pangenome-based functional profiles for microbial communities. *BMC Res. Notes* **8**, 497 (2015).

4. 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).

5. Iwai, S. *et al.* Piphillin: Improved prediction of metagenomic content by direct inference from human microbiomes. *PLoS One* **11**, e0166104 (2016).

6. 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).

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

8. Edgar, R. C. UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. *bioRxiv* 081257 (2016). doi:10.1101/081257

9. Amir, A. *et al.* Deblur Rapidly Resolves Single- Nucleotide Community Sequence Patterns. *mSystems* **2**, e00191-16 (2017).

10. Quast, C. *et al.* The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Res.* **41**, 590–596 (2013).

11. Janssen, S. *et al.* Phylogenetic Placement of Exact Amplicon Sequences Improves Associations with Clinical Information. *mSystems* **3**, e00021 (2018).

12. Bolyen, E. *et al.* QIIME 2: Reproducible, interactive, scalable, and extensible microbiome data science. *PeerJ Prepr.* (2018). doi:10.7287/peerj.preprints.27295

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

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

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

16. Wilkinson, T. J. *et al.* CowPI: A rumen microbiome focussed version of the PICRUSt functional inference software. *Front. Microbiol.* **9**, 1095 (2018).

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

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

19. 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).

20. 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).

21. Finn, R. D. *et al.* Pfam: The protein families database. *Nucleic Acids Res.* **42**, 222–230 (2014).

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

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

24. 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).

25. 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**, 3847–3856 (2015).

26. Yurgel, S. N. *et al.* Variation in Bacterial and Eukaryotic Communities Associated with Natural and Managed Wild Blueberry Habitats. *Phytobiomes* **1**, 102–113 (2017).

27. Caspi, R. *et al.* The MetaCyc database of metabolic pathways and enzymes and the

BioCyc collection of pathway/genome databases. *Nucleic Acids Res.* **44**, D471–D480 (2016).

28. Chen, I. M. A. *et al.* Improving Microbial Genome Annotations in an Integrated Database Context. *PLoS One* **8**, e54859 (2013).

29. Louca, S., Parfrey, L. W. & Doebeli, M. Decoupling function and taxonomy in the global ocean microbiome. *Science (80-.).* **353**, 1272–1277 (2016).

30. Ward, T. *et al.* BugBase predicts organism-level microbiome phenotypes. *bioRxiv* (2017). doi:http://dx.doi.org/10.1101/133462

31. Sternes, P. R., Lee, D., Kutyna, D. R. & Borneman, A. R. A combined meta-barcoding and shotgun metagenomic analysis of spontaneous wine fermentation. *Gigascience* **6**, 1–10 (2017).

32. Lloyd-Price, J. *et al.* Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature* **569**, 655–662 (2019).

33. Gevers, D. *et al.* The treatment-naive microbiome in new-onset Crohn's disease. *Cell Host Microbe* **15**, 382–392 (2014).

34. Douglas, G. M. *et al.* Multi-omics differentially classify disease state and treatment outcome in pediatric Crohn's disease. *Microbiome* **6**, 13 (2018).

35. Santoru, M. L. *et al.* Cross sectional evaluation of the gut-microbiome metabolome axis in an Italian cohort of IBD patients. *Sci. Rep.* **7**, 9523 (2017).

36. Felsenstein, J. Phylogenies and the Comparative Method. *Am. Nat.* **125**, 1–15 (1985).

37. Franzosa, E. A. *et al.* Species-level functional profiling of metagenomes and metatranscriptomes. *Nat. Methods* **15**, 962–968 (2018).

38. Ye, Y. & Doak, T. G. A Parsimony Approach to Biological Pathway Reconstruction/Inference for Metagenomes. *PLoS Comput. Biol.* **5**, e1000465 (2011).

39. Abubucker, S. *et al.* Metabolic reconstruction for metagenomic data and its application to the human microbiome. *PLoS Comput. Biol.* **8**, e1002358 (2012).

40. Manor, O. & Borenstein, E. MUSiCC: a marker genes based framework for metagenomic normalization and accurate profiling of gene abundances in the microbiome. *Genome Biol.* **16**, 53 (2015).

41. Rognes, T., Flouri, T., Nichols, B., Quince, C. & Mahé, F. VSEARCH: a versatile open source tool for metagenomics. *PeerJ* **4**, e2584 (2016).

42. Kozlov, A. M., Darriba, D., Flouri, T., Morel, B. & Stamatakis, A. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics* 1–3 (2019). doi:10.1093/bioinformatics/btz305

43. Martiny, A. C., Treseder, K. & Pusch, G. Phylogenetic conservatism of functional traits in microorganisms. *ISME J.* **7**, 830–838 (2013).

44. Bengtsson-Palme, J. *et al.* Improved software detection and extraction of ITS1 and ITS2 from ribosomal ITS sequences of fungi and other eukaryotes for analysis of environmental sequencing data. *Methods Ecol. Evol.* **4**, 914–919 (2013).

45. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M.

BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).

46. Katoh, K., Misawa, K., Kuma, K. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30**, 3059–3066 (2002).

47. Caporaso, J. G. *et al.* QIIME allows analysis of high-throughput community sequencing data. *Nat. Publ. Gr.* **7**, 335–336 (2010).

48. Zhang, J., Kobert, K., Flouri, T. & Stamatakis, A. PEAR: A fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics* **30**, 614–620 (2014).

49. 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).

50. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).

51. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–9 (2012).

52. Gupta, V. *et al.* RNA-Seq analysis and annotation of a draft blueberry genome assembly identifies candidate genes involved in fruit ripening, biosynthesis of bioactive compounds, and stage-specific alternative splicing. *Gigascience* **4**, 5 (2015).

53. Tange, O. GNU Parallel: the command-line power tool. ;*login USENIX Mag.* **36**, 42–47 (2011).

54. Bengtsson-Palme, J. *et al.* metaxa2: Improved identification and taxonomic classification of small and large subunit rRNA in metagenomic data. *Mol. Ecol. Resour.* **15**, 1403–1414 (2015).

55. Fernandes, A. D. *et al.* Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis. *Microbiome* **2**, 15 (2014).

56. Haberman, Y. *et al.* Pediatric Crohn disease patients exhibit specific ileal transcriptome and microbiome signature. *J. Clin. Invest.* **124**, 3617–3633 (2014).