EVE 161:
Microbial Phylogenomics

Class #16:
Predicting Metagenomes

UC Davis, Winter 2018
Instructor: Jonathan Eisen
Teaching Assistant: Cassie Ettinger

Profiling phylogenetic marker genes, such as the 16S rRNA gene, is a key tool for studies of microbial communities but does not provide direct evidence of a community's functional capabilities. Here we describe PICRUSt (phylogenetic investigation of communities by reconstruction of unobserved states), a computational approach to predict the functional composition of a metagenome using marker gene data and a database of reference genomes. PICRUSt uses an extended ancestral-state reconstruction algorithm to predict which gene families are present and then combines gene families to estimate the composite metagenome. Using 16S information, PICRUSt recaptures key findings from the Human Microbiome Project and accurately predicts the abundance of gene families in host-associated and environmental communities, with quantifiable uncertainty. Our results demonstrate that phylogeny and function are sufficiently linked that this 'predictive metagenomic' approach should provide useful insights into the thousands of uncultivated microbial communities for which only marker gene surveys are currently available.

Profiling phylogenetic marker genes, such as the 16S rRNA gene, is a key tool for studies of microbial communities but does not provide direct evidence of a community's functional capabilities. Here we describe PICRUSt (phylogenetic investigation of communities by reconstruction of unobserved states), a computational approach to predict the functional composition of a metagenome using marker gene data and a database of reference genomes. PICRUSt uses an extended ancestral-state reconstruction algorithm to predict which gene families are present and then combines gene families to estimate the composite metagenome. Using 16S information, PICRUSt recaptures key findings from the Human Microbiome Project and accurately predicts the abundance of gene families in host-associated and environmental communities, with quantifiable uncertainty. Our results demonstrate that phylogeny and function are sufficiently linked that this 'predictive metagenomic' approach should provide useful insights into the thousands of uncultivated microbial communities for which only marker gene surveys are currently available.

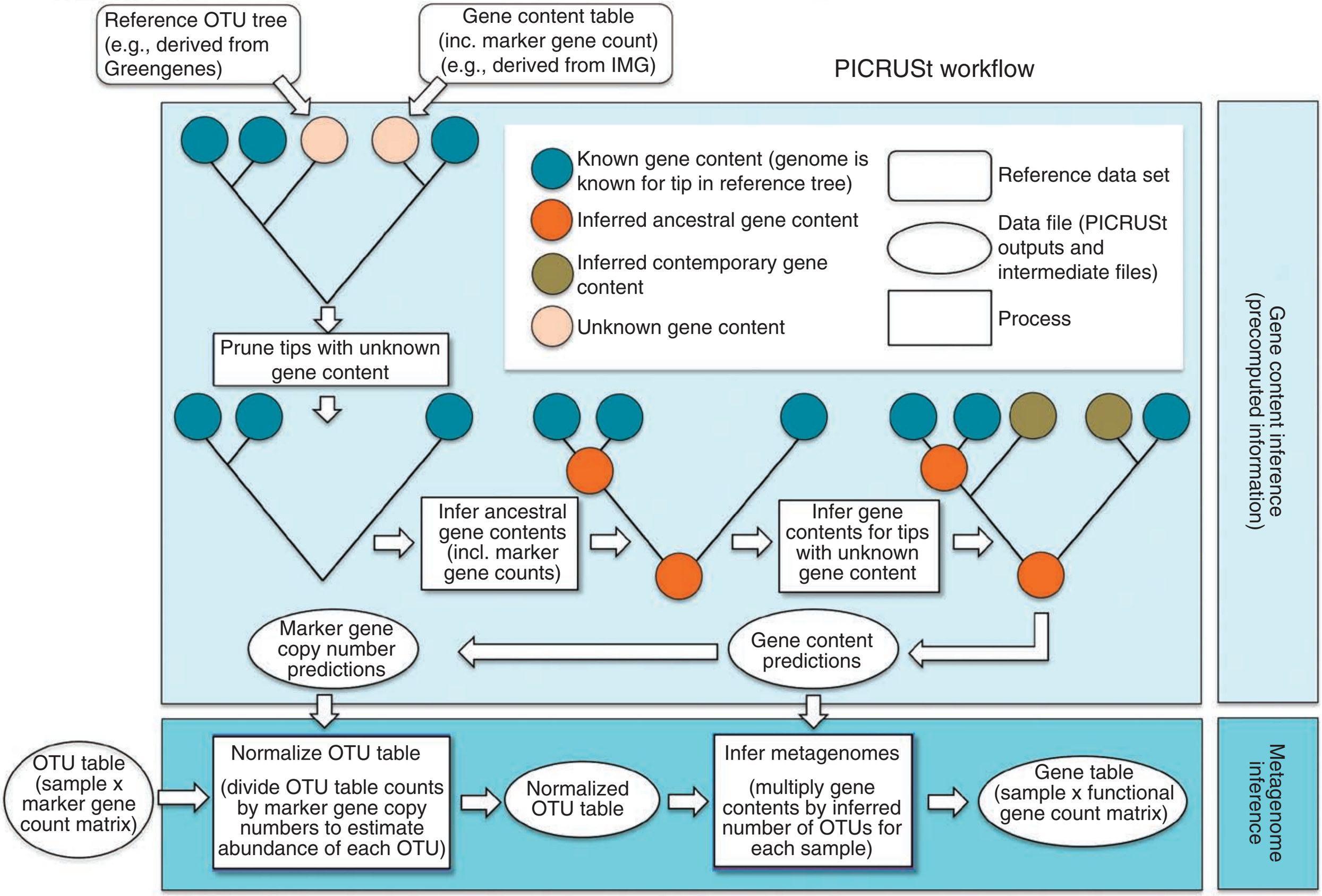

PICRUSt workflow
Reference OTU tree (e.g., derived from Greengenes)
Gene content table (inc. marker gene count) (e.g., derived from IMG)
Known gene content (genome is known for tip in reference tree)
Inferred ancestral gene content
Inferred contemporary gene content
Unknown gene content
Reference data set
Data file (PICRUSt outputs and intermediate files)
Process
Prune tips with unknown gene content
Infer ancestral gene contents (incl. marker gene counts)
Infer gene contents for tips with unknown gene content
Marker gene copy number predictions
Gene content predictions
Gene content inference (precomputed information)
OTU table (sample x marker gene count matrix)
Normalize OTU table
(divide OTU table counts by marker gene copy numbers to estimate abundance of each OTU)
Normalized OTU table
Infer metagenomes
(multiply gene contents by inferred number of OTUs for each sample)
Gene table (sample x functional gene count matrix)
Metagenome inference

Figure 1 The PICRUSt workflow. PICRUSt is composed of two high-level workflows: gene content inference (top box) and metagenome inference (bottom box). Beginning with a reference OTU tree and a gene content table (i.e., counts of genes for reference OTUs with known gene content), the gene content inference workflow predicts gene content for each OTU with unknown gene content, including predictions of marker gene copy number. This information is precomputed for 16S based on Greengenes and IMG, but all functionality is accessible in PICRUSt for use with other marker genes and reference genomes. The metagenome inference workflow takes an OTU table (i.e., counts of OTUs on a per sample basis), where OTU identifiers correspond to tips in the reference OTU tree, as well as the copy number of the marker gene in each OTU and the gene content of each OTU (as generated by the gene content inference workflow), and outputs a metagenome table (i.e., counts of gene families on a per-sample basis).

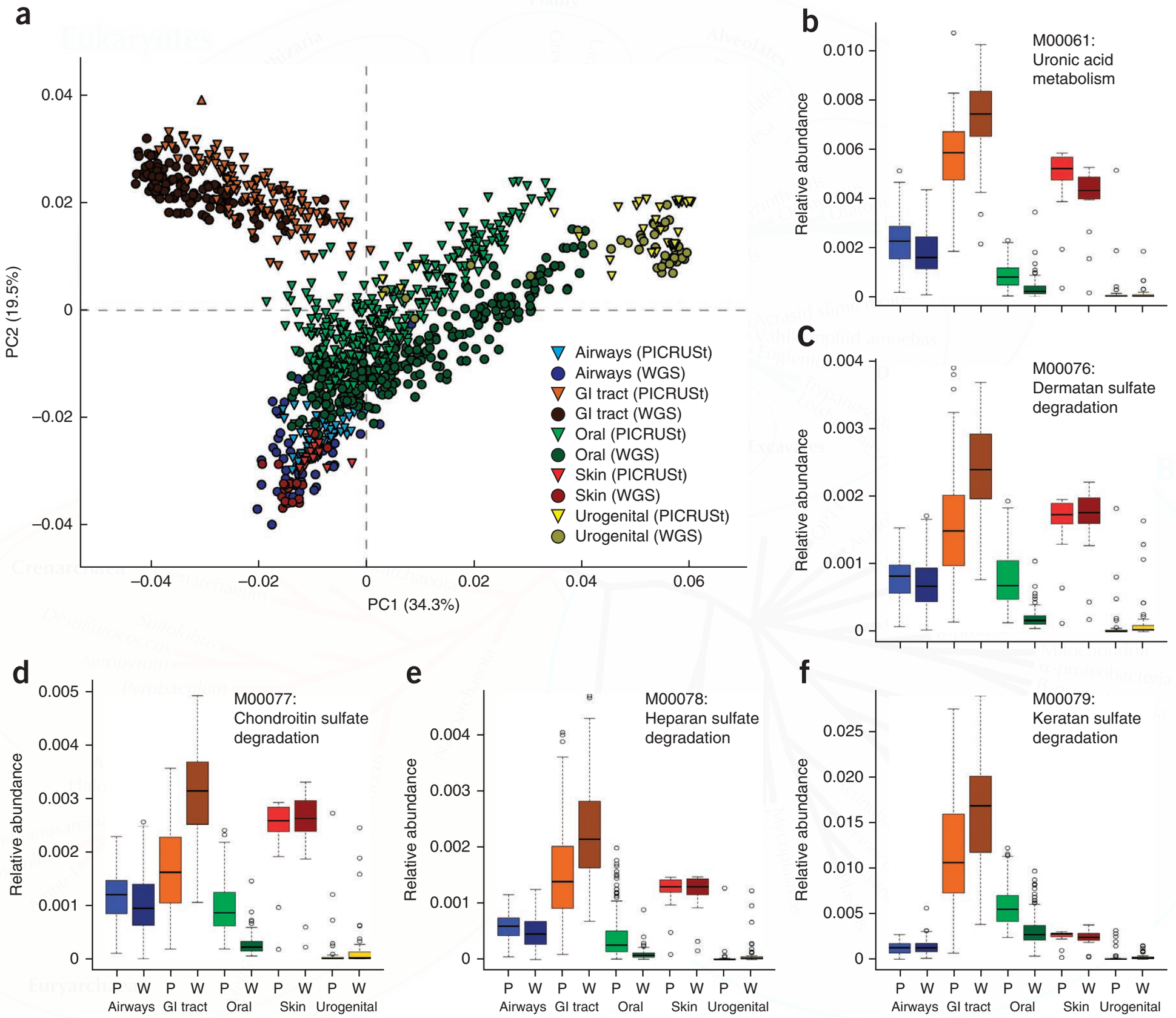

a
0.04
0.02
0
−0.02
−0.04
PC2 (19.5%)
−0.04
−0.02
0
0.02
0.04
0.06
PC1 (34.3%)
Airways (PICRUSt)
Airways (WGS)
GI tract (PICRUSt)
GI tract (WGS)
Oral (PICRUSt)
Oral (WGS)
Skin (PICRUSt)
Skin (WGS)
Urogenital (PICRUSt)
Urogenital (WGS)
b
M00061:
Uronic acid
metabolism
Relative abundance
0.010
0.008
0.006
0.004
0.002
0
c
M00076:
Dermatan sulfate
degradation
Relative abundance
0.004
0.003
0.002
0.001
0
d
M00077:
Chondroitin sulfate
degradation
Relative abundance
0.005
0.004
0.003
0.002
0.001
0
e
M00078:
Heparan sulfate
degradation
Relative abundance
0.004
0.003
0.002
0.001
0
f
M00079:
Keratan sulfate
degradation
Relative abundance
0.025
0.020
0.015
0.010
0.005
0
P W P W P W P W P W
Airways GI tract Oral Skin Urogenital

Figure 2 PICRUSt recapitulates biological findings from the Human Microbiome Project. (**a**) Principal component analysis (PCA) plot comparing KEGG module predictions using 16S data with PICRUSt (lighter colored triangles) and sequenced shotgun metagenome (darker colored circles) along with relative abundances for five specific KEGG modules: (**b**) M00061: Uronic acid metabolism. (**c**) M00076: Dermatan sulfate degradation. (**d**) M00077: Chondroitin sulfate degradation. (**e**) M00078: Heparan sulfate degradation. (**f**) M00079: Keratan sulfate degradation. All KEGG modules are involved in glycosaminosglycan degradation (KEGG pathway ko00531) using 16S with PICRUSt (P) and whole genome sequencing (W) across human body sites. Color key is the same as in **a**.

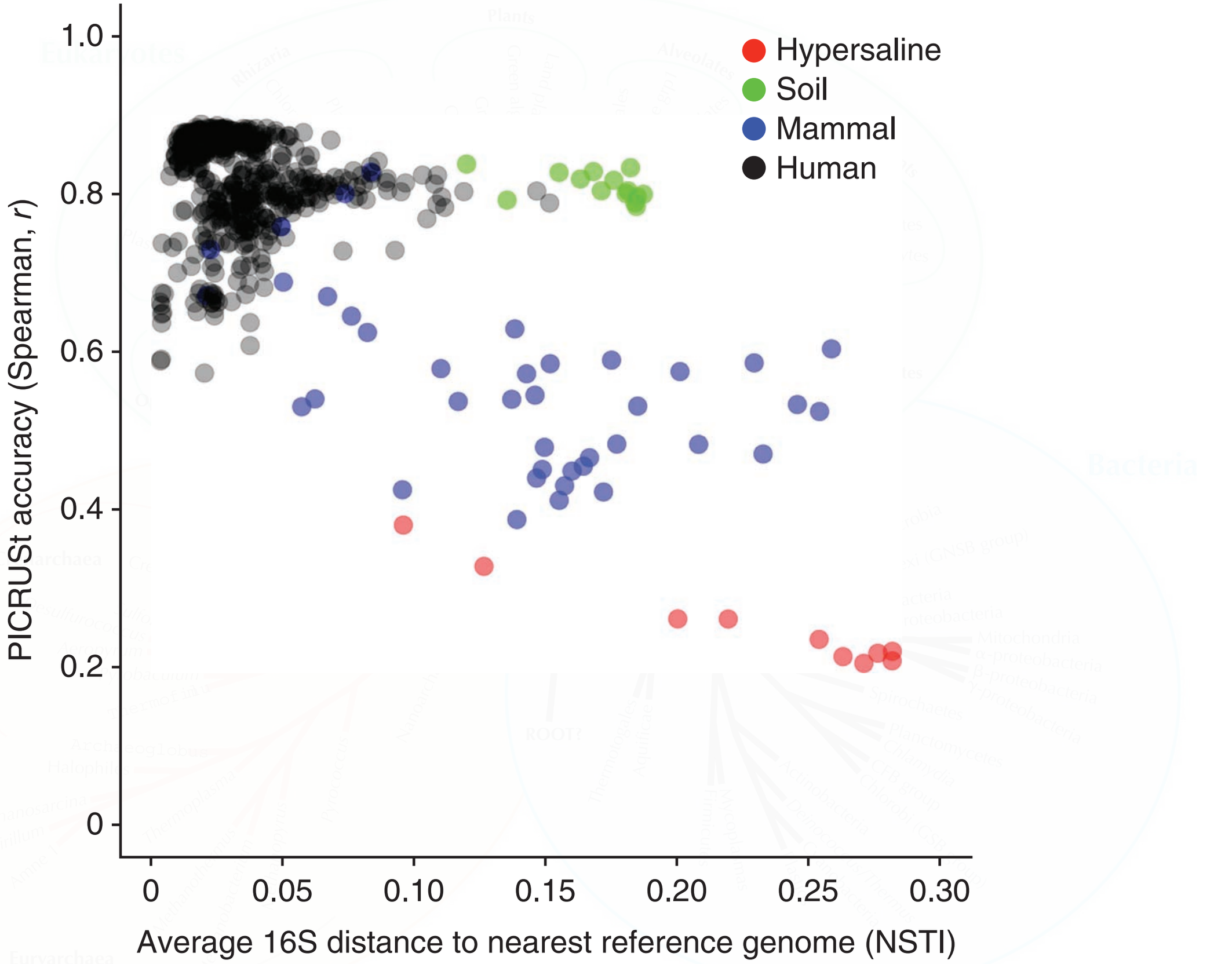

Hypersaline
Soil
Mammal
Human
PICRUSt accuracy (Spearman, r)
1.0
0.8
0.6
0.4
0.2
0
0
0.05
0.10
0.15
0.20
0.25
0.30
Average 16S distance to nearest reference genome (NSTI)

Figure 3 PICRUSt accuracy across various environmental microbiomes. Prediction accuracy for paired 16S rRNA marker gene surveys and shotgun metagenomes are plotted against the availability of reference genomes as summarized by NSTI. Accuracy is summarized using the Spearman correlation between the relative abundance of gene copy number predicted from 16S data using PICRUSt versus the relative abundance observed in the sequenced shotgun metagenome. In the absence of large differences in metagenomic sequencing depth, relatively well-characterized environments, such as the human gut, had low NSTI values and can be predicted accurately from 16S surveys. Conversely, environments containing much unexplored diversity (e.g., phyla with few or no sequenced genomes), such as the Guerrero Negro hypersaline microbial mats, tended to have high NSTI values.

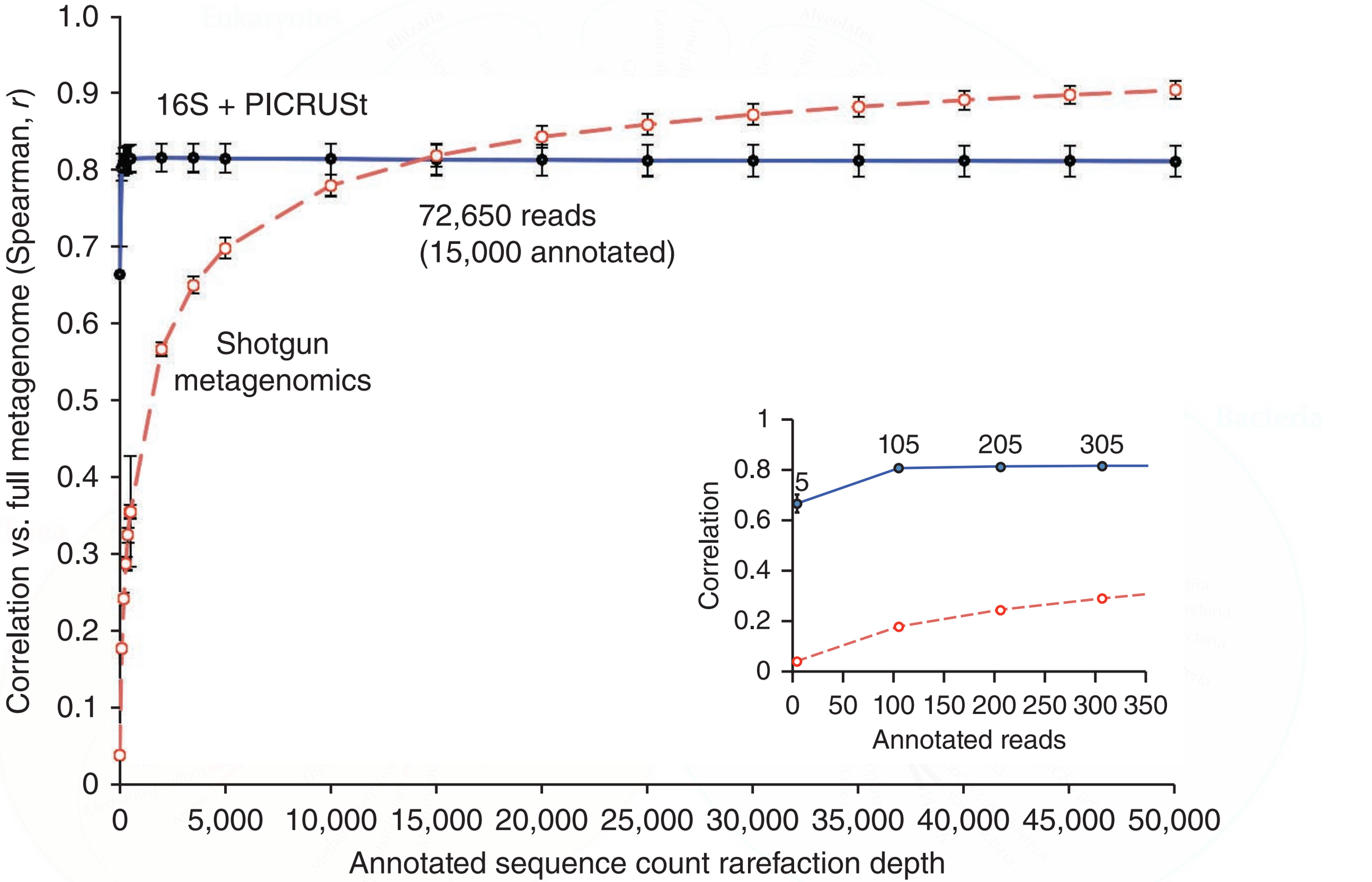
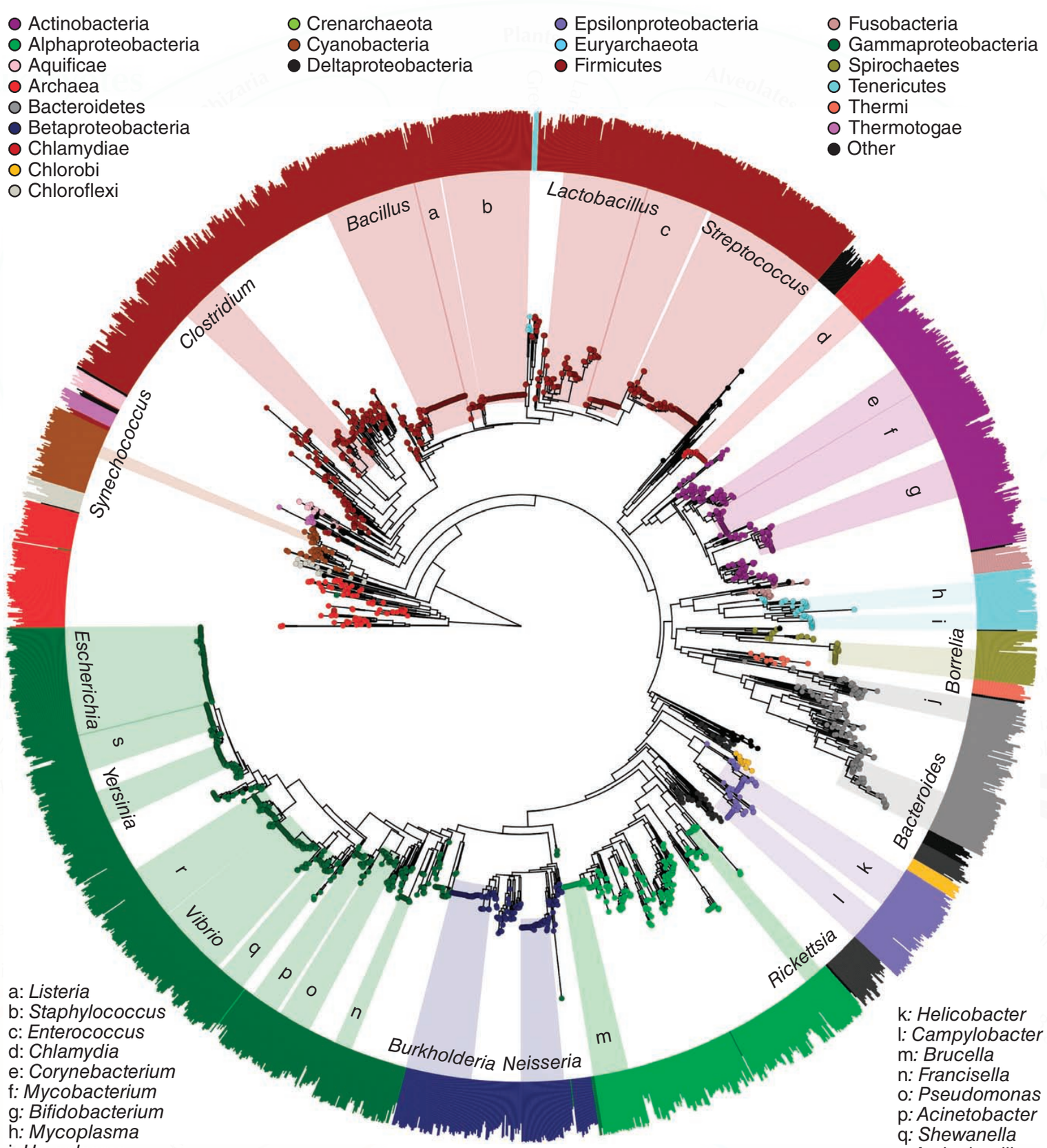

16S + PICRUSt
72,650 reads
(15,000 annotated)
Shotgun
metagenomics
Correlation vs. full metagenome (Spearman, r)
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
5,000
10,000
15,000
20,000
25,000
30,000
35,000
40,000
45,000
50,000
Annotated sequence count rarefaction depth
Correlation
1
0.8
0.6
0.4
0.2
0
105
205
305
5
0
50
100
150
200
250
300
350
Annotated reads

Figure 4 Accuracy of PICRUSt prediction compared with shotgun metagenomic sequencing at shallow sequencing depths. Spearman correlation between either PICRUSt-predicted metagenomes (blue lines) or shotgun metagenomes (dashed red lines) using 14 soil microbial communities subsampled to the specified number of annotated sequences. This rarefaction reflects random subsets of either the full 16S OTU table (blue) or the corresponding gene table for the sequenced metagenome (red). Ten randomly chosen rarefactions were performed at each depth to indicate the expected correlation obtained when assessing an underlying true metagenome using either shallow 16S rRNA gene sequencing with PICRUSt prediction or shallow shotgun metagenomic sequencing. The data label describes the number of annotated reads below which PICRUSt-prediction accuracy exceeds metagenome sequencing accuracy. Note that the plotted rarefaction depth reflects the number of 16S or metagenomic sequences remaining after standard quality control, dereplication and annotation (or OTU picking in the case of 16S sequences), not the raw number returned from the sequencing facility. The number of total metagenomic reads below which PICRUSt outperforms metagenomic sequencing (72,650) for this data set was calculated by adjusting the crossover point in annotated reads (above) using annotation rates for the soil data set (17.3%) and closed-reference OTU picking rates for the 16S rRNA data set (68.9%). The inset figure illustrates rapid convergence of PICRUSt predictions given low numbers of annotated reads (blue line).



a: *Listeria*
b: *Staphylococcus*
c: *Enterococcus*
d: *Chlamydia*
e: *Corynebacterium*
f: *Mycobacterium*
g: *Bifidobacterium*
h: *Mycoplasma*
i: *Ureaplasma*
j: *Prevotella*

k: *Helicobacter*
l: *Campylobacter*
m: *Brucella*
n: *Francisella*
o: *Pseudomonas*
p: *Acinetobacter*
q: *Shewanella*
r: *Actinobacillus*
s: *Salmonella*

Figure 5 PICRUSt prediction accuracy across the tree of bacterial and archaeal genomes. Phylogenetic tree produced by pruning the Greengenes 16S reference tree down to those tips representing sequenced genomes. Height of the bars in the outermost circle indicates the accuracy of PICRUSt for each genome (accuracy: 0.5–1.0) colored by phylum, with text labels for each genus with at least 15 strains. PICRUSt predictions were as accurate for archaeal (mean = 0.94 0.04 s.d., n = 103) as for bacterial genomes (mean = 0.95 0.05 s.d., n = 2,487).

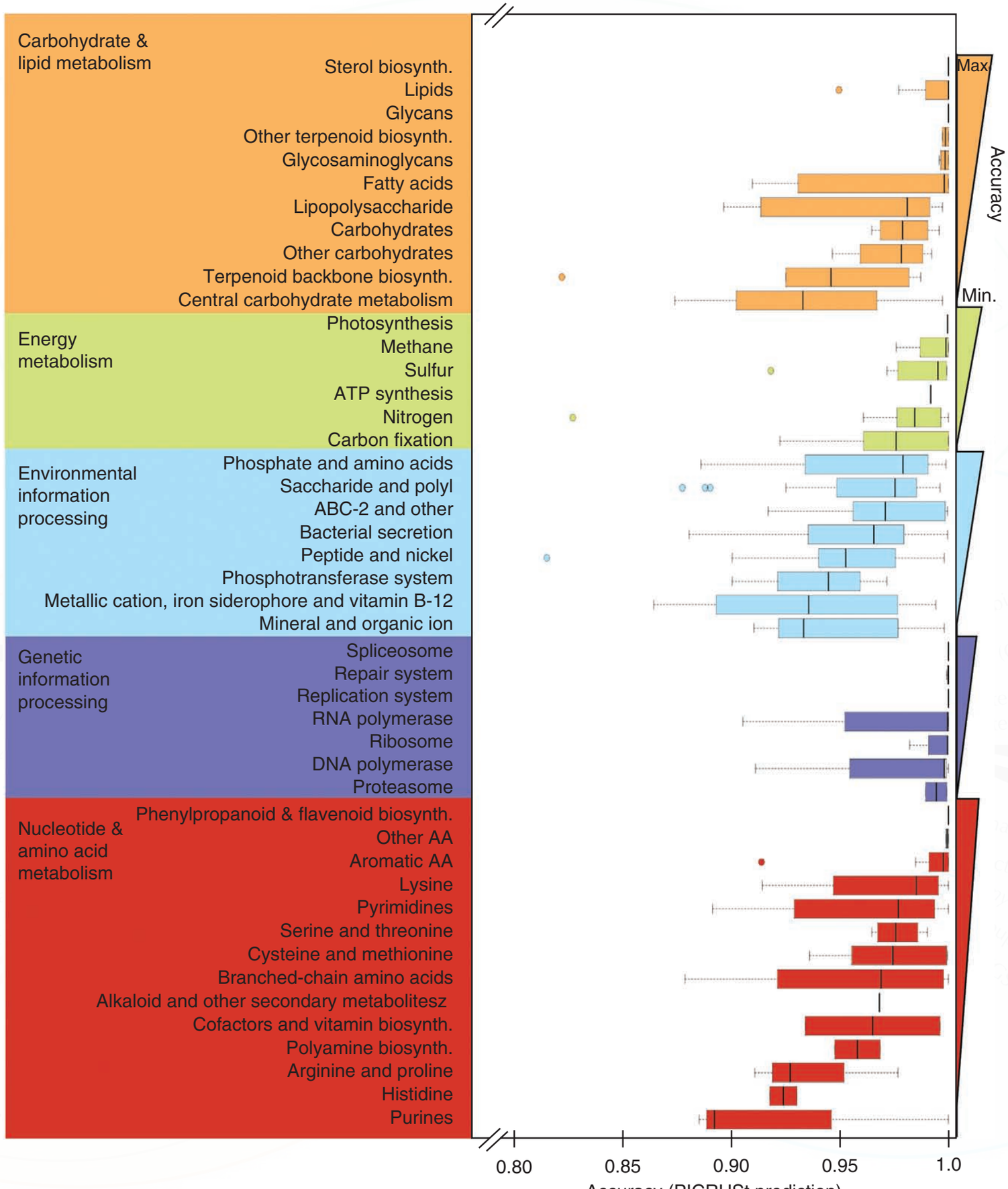

Carbohydrate & lipid metabolism
Sterol biosynth.
Lipids
Glycans
Other terpenoid biosynth.
Glycosaminoglycans
Fatty acids
Lipopolysaccharide
Carbohydrates
Other carbohydrates
Terpenoid backbone biosynth.
Central carbohydrate metabolism
Energy metabolism
Photosynthesis
Methane
Sulfur
ATP synthesis
Nitrogen
Carbon fixation
Environmental information processing
Phosphate and amino acids
Saccharide and polyl
ABC-2 and other
Bacterial secretion
Peptide and nickel
Phosphotransferase system
Metallic cation, iron siderophore and vitamin B-12
Mineral and organic ion
Genetic information processing
Spliceosome
Repair system
Replication system
RNA polymerase
Ribosome
DNA polymerase
Proteasome
Nucleotide & amino acid metabolism
Phenylpropanoid & flavenoid biosynth.
Other AA
Aromatic AA
Lysine
Pyrimidines
Serine and threonine
Cysteine and methionine
Branched-chain amino acids
Alkaloid and other secondary metabolitesz
Cofactors and vitamin biosynth.
Polyamine biosynth.
Arginine and proline
Histidine
Purines
Max.
Min.
Accuracy
0.80
0.85
0.90
0.95
1.0
Accuracy (PICRUSt prediction)

Figure 6 Variation in inference accuracy across functional modules within single genomes. Results are colored by functional category and sorted in decreasing order of accuracy within each category (indicated by triangular bars, right margin). Note that accuracy was >0.80
for all, and therefore the region 0.80–1.0 is displayed for clearer visualization of differences between modules.