Inference based PICRUSt accuracy varies across sample types and functional categories

Shan Sun[1], Roshonda B. Jones[2] and Anthony A. Fodor[1]

[1]Department of Bioinformatics and Genomics, University of North Carolina at Charlotte, Charlotte, North Carolina, USA; [2] The Saban Research Institute, Children's Hospital Los Angeles, University of Southern California, Los Angeles, California, USA

Abstract

Background: Despite recent decreases in the cost of sequencing, shotgun metagenome sequencing remains more expensive compared with 16S rRNA amplicon sequencing. Methods have been developed to predict the functional profiles of microbial communities based on their taxonomic composition, and PICRUSt is the most widely used of these techniques. In this study, we evaluated the performance of PICRUSt by comparing the significance of the differential abundance of functional gene profiles predicted with PICRUSt to those from shotgun metagenome sequencing across different environments.

Results: We selected 7 datasets of human, non-human animal and environmental (soil) samples that have publicly available 16S rRNA and shotgun metagenome sequences. As we would expect based on previous literature, strong Spearman correlations were observed between gene compositions predicted with PICRUSt and measured with shotgun metagenome sequencing. However, these strong correlations were preserved even when the sample labels were shuffled. This suggests that simple correlation coefficient is a highly unreliable measure for the performance of algorithms like PICRUSt. As an alternative, we compared the performance of PICRUSt predicted genes

to metagenome genes in inference models associated with metadata within each dataset. With this method, we found reasonable performance for human datasets, with PICRUSt performing better for inference on genes related to “house-keeping” functions. However, the performance of PICRUSt degraded sharply outside of human datasets when used for inference.

Conclusion: We conclude that the utility of PICRUSt for inference with the default database is likely limited outside of human samples and that development of tools for gene prediction specific to different non-human and environmental samples is warranted.

Key words: microbiota functional profile prediction, inference, sample type, functional category

Introduction

Recent advances in next generation sequencing are revolutionizing our understanding of complex microbial communities. Amplicon sequencing of marker genes provides information regarding the phylogenetic diversity and taxonomic composition of microorganisms present in the environment, while shotgun metagenome sequencing provides additional information on the relative abundance of functional genes. Although knowledge of taxonomy and functional genes of microorganisms are both important, functional genes are more directly related to pathways and therefore are essential for understanding the roles microorganisms play with regards to different physiological or ecological outcomes. However, the higher cost of metagenome sequencing hinders its application in studies consisting of a large number of samples, which are usually

47 necessary in order to ensure adequate statistical power for detecting true differences [1].
48 Additionally, metagenome sequencing can also be very challenging for low biomass
49 samples or samples that are dominated by non-microbial DNA [2, 3].
50
51 In order to address this problem, tools have been developed to predict microbial
52 functional genes from their taxonomic compositions inferred from more cost-effective
53 amplicon sequencing, including PICRUSt, Tax4Fun and FaproTax [4-6]. Among these
54 tools, PICRUSt is the most widely used and has been applied in hundreds of projects on
55 various environments, including human gut [7, 8], murine [9, 10], fish [11], coral [12],
56 water [13], plant [14], bioreactor [15] and soil [16]. PICRUSt predicts the genes of
57 organisms without sequenced genomes based on mapping their 16S rRNA genes to
58 homologous taxa with fully sequenced genomes. The predictions of PICRUSt are
59 therefore limited by currently available genomes, which are highly biased towards
60 microorganisms associated with human health and biotechnology use [17].
61
62 In order to gauge the reliability of PICRUSt predictions in different environments and for
63 different functional categories, we utilized human, non-human animal (gorilla, mouse and
64 chicken) and environmental (soil) datasets that were sequenced for both 16S rRNA
65 marker genes and shotgun metagenomes. We compared the predicted functional profiles
66 from PICRUSt to the functional profiles measured with shotgun metagenome sequencing.
67 We demonstrated that simple correlations such as Spearman correlation overstate the
68 accuracy of PICRUSt by not taking into account the low variance of functional profiles
69 generated from shotgun metagenome sequencing. As an alternative metric, we used

PICRUSt results for inference with simple statistical models and found reasonable performance for human datasets, which presumably reflected the better reference information we currently have for human genomes, but a sharp decrease in performance for inference in non-human samples.

Results

Spearman correlation is not a reliable measurement for the prediction accuracy of gene contents

In order to test the accuracy of PICRUSt from a range of environments, we compared PICRUSt predictions to the results of shotgun metagenome sequencing on publicly available datasets for which both metagenome and 16S rRNA sequences were available (Table S1). As we would expect from previous literature [4], gene content estimation from PICRUSt was robustly correlated with gene contents from metagenome sequencing with Spearman correlations in the range of 0.62 to 0.84 (Fig. 1). For example, in one soil sample (Fig. 1a), there is a clear correlation between the relative abundance of each gene from PICRUSt and the relative abundance from metagenome sequencing (Spearman's rho = 0.85). However, if the same comparison is made with the sample labels shuffled, the correlation that was observed is not substantially impacted (Spearman's rho = 0.84) (Fig. 1b).

The likely explanation for this observation is that across environments, there is less variation between metagenome functional profiles of samples than their taxonomic profiles (Fig. 2), an observation that has been previously made for human samples in the

Human Microbiome Project [18]. In the datasets we examined, despite sample labels being shuffled, the relative abundance of genes from PICRUSt estimates were highly correlated with the relative abundance of gene estimates from metagenome sequencing, with correlation coefficients always higher than 0.5. The PICRUSt predictions were often only marginally higher on the unpermuted data than those with sample labels shuffled, with perhaps the gorilla dataset as an exception (Fig. 1c). However, even in the gorilla samples, the difference between Spearman coefficients for permuted and unpermuted samples was only 0.12. For the 2 soil datasets, the Spearman coefficients for the unpermuted samples were not significantly different from those for the permuted samples ($P = 0.508$ and 0.794).

PICRUSt results for inference showed higher consistency with metagenome sequences in human samples than non-human samples

As an alternative evaluation to Spearman correlation, we compared how PICRUSt's gene predictions performed in a t-test to genes detected with shotgun metagenome sequencing in each of our datasets. For this purpose, we formed a null hypothesis for each gene in each dataset that there is no difference in the mean of that gene's distribution of relative abundance between the two groups in the dataset. For example, for each of the 5,574 genes detected by both PICRUSt and metagenome sequencing in the Human_KW dataset, we used a t-test to generate a P-value for the difference in gene composition between rural and urban samples. Across all the genes, there was a reasonable correlation (rho=0.49) of P-values from t-tests run on metagenome sequencing data and those on PICRUSt data (Fig. 3; top left panel). Unlike our results for Spearman correlation, this

116 rho value is sensitive to sample permutation, as when we repeated this procedure on
117 permuted data, the correlation between P-values generated from metagenome sequencing
118 and those from PICRUSt approached zero (Fig. 4). We saw a similarly robust correlation
119 (rho=0.554) for our Human_TY dataset evaluating a null hypothesis comparing US and
120 non-US samples. However, when we extended this analysis to non-human datasets (using
121 the null hypotheses for each study listed in Table S1), the inference produced by
122 PICRUSt showed a much lower similarity to inference produced by metagenome
123 sequencing (Fig. 3).
124
125 In order to ensure that the differences in performance were not due to differences in
126 sample sizes, we randomly sub-sampled each larger dataset (without replacement) to 10
127 samples (5 per group) and re-calculated the comparison of P-values between PICRUSt
128 and metagenome sequencing (Fig. S1). Even at a smaller size, data from the human
129 studies showed greater concordance than those from other environments. We conclude
130 that the difference in sample sizes between datasets does not explain the variability of
131 PICRUSt accuracy between different samples types in our study. Likewise, the effect
132 sizes of the associations with metadata, measured as R^2 in a PERMANOVA test, were
133 not substantially higher in human samples (Table S1). It therefore also seems unlikely
134 that effect size alone can explain the better concordance we observed between inference
135 results from PICRUSt and metagenome sequencing for human samples.
136
137 In order to further investigate the consistency of PICRUSt and metagenome sequencing,
138 we next asked how many genes were detected by both methods or by each method

individually. For some datasets, such as the Human_KW dataset, PICRUSt failed to predict many genes that were detected by metagenome sequencing. For other datasets, such as the soil datasets, many genes that PICRUSt predicted were not detected in metagenome sequencing and there were also many genes seen in metagenome sequencing but not in PICRUSt (Fig. 3). For the chicken dataset with an average metagenome sequencing depth of 31 million reads/sample and the gorilla dataset of 27 million reads/sample, 23.4% and 26.1% of PICRUSt predicted genes could not be detected by metagenome sequencing (Fig. 3). In addition, the metagenome sequencing of the Human_KW dataset with an average sequencing depth of 10 million reads/sample detected 13,880 genes and PICRUSt missed 59.8% of them (Fig. 3).

PICRUSt performs differently for different functional categories

We next investigated the discrepancy between PICRUSt and metagenome sequencing for inference in different functional categories (Fig. 5). When comparing the P-values from PICRUSt data to P-values from metagenome sequencing, there was a lack of significant correlations in the human gut samples for some categories including Signaling molecules and interaction, Immune system, Endocrine system, Cell growth and death, Nervous system, Digestive system, Environmental adaptation, Cardiovascular diseases, Immune diseases, Substance dependence, Circulatory system, Excretory system and Viral protein family. PICRUSt appeared to have a closer match to the metagenome sequencing data for categories such as Folding, sorting and degradation, Translation, Transcription, Replication and repair, Nucleotide metabolism, Glycan biosynthesis and metabolism, Drug resistance, Neurodegenerative diseases, Endocrine and metabolic diseases and

Aging. For the genes only detected by one method, most of the genes missed by PICRUSt belong to Environmental Information Processing, Organismal Systems and Human Diseases, while those associated with Metabolism were more likely predicted (Fig. S2). Among the genes predicted by PICRUSt but not detected by metagenome sequencing, most of them belong to Cellular processing and signaling, Signaling molecules and interaction, Genetic information processing, Substance Dependence and Viral protein family (Fig. S3).

Discussion

Microbial community functional profiles are typically of much lower variance compared to their taxonomic profiles likely because of the large proportions of “core” or “housekeeping” functions [18-20]. In this study, we showed that this lack of variance in functional profiles between samples leads to a strong correlation between functional profiles from metagenome sequencing and those estimated from references with PICRUSt, even when the sample labels are randomly shuffled (Fig. 1c). We argue that this result shows that metrics commonly used to measure gene prediction performance, such as Spearman correlation between gene composition estimated with PICRUSt and metagenome sequencing, do not give a satisfactory measure of overall accuracy. As an alternative, we evaluated the performance of PICRUSt at a community level based on inference from simple statistical models testing the association between genes and metadata. Unlike simple Spearman correlations, evaluation with inference methods are highly sensitive to shuffling sample labels (Fig. 4), which indicated that inference methods are much less affected by the relatively low variance of functional profiles. The

185 inference-based approach also has the advantage of reflecting the common use of
186 PICRUSt to reveal predicted functional profiles associated with different metadata
187 categories [11, 12, 21-23]. Incorrect estimation of differential abundance could lead to
188 false discovery of signature genes, and this concern motivated our approach to determine
189 the reliability of PICRUSt produced inference in different systems.
190
191 In this study, we selected 7 datasets from different environments which include human,
192 non-human animal and environmental (soil) samples. With inference methods, we found
193 that PICRUSt and metagenome sequencing had more consistent assessment from human
194 datasets than non-human animals or environmental datasets. It is likely that these
195 differences reflected the bias of genome databases towards human-related
196 microorganisms. However, PICRUSt still missed a large percentage of genes that were
197 detected with metagenome sequencing in human samples, and an increase in metagenome
198 sequencing depth could presumably increase the number of genes that are potentially not
199 detected by PICRUSt (Fig. 3; Table S1). Likewise, PICRUSt sometimes predicted many
200 genes not found in metagenome sequencing even in samples with presumably adequate
201 sequencing depth of millions of reads per sample, which suggested that these additional
202 genes are likely false predictions.
203
204 As a meta-analysis across multiple studies, there are systemic factors that may influence
205 the results of this study, including different sample sizes, sequencing designs and effect
206 sizes of associations with the metadata. We repeated our analysis on subsampled datasets
207 that were rarified to the number of samples in the smallest dataset that we examined and

208 observed a similar pattern of results with inference more consistent between PICRUSt
209 and metagenome sequencing for human studies (Fig. S1). This result suggests that
210 difference in sample size does not explain the better inference performance of PICRUSt
211 for the human studies. While differences in effect size and experimental design are harder
212 to control, the human studies did not have an obviously higher effect size than the non-
213 human studies as measured with a PERMANOVA test (Table S1). It therefore also seems
214 unlikely that differences in effect sizes of associations with the metadata can explain our
215 results.
216
217 Our study also examined the performance of PICRUSt for different functional categories.
218 This approach was motivated by the presumed bias in current genome databases toward
219 culturable microorganisms [24]. We reasoned that the unculturable state of
220 microorganisms could be caused by specific requirements for nutrients, temperature, pH,
221 beneficial interactions with other microbes or extremely slow growth rates [25], which in
222 turn could lead to bias in gene families in different microorganisms. Likewise, different
223 microorganisms and genes also have different rates of horizontal gene transfer and the
224 accuracy of gene content estimation may therefore vary depending on the type of the
225 genes and microbial groups [26]. We found that PICRUSt performed best for “house-
226 keeping” functions such as transcription and translation while the accuracy of functions
227 related to environmental information processing was generally much lower (Fig. 5).
228 Future algorithms for gene prediction could explicitly incorporate this performance
229 variance into a confidence score that could give users estimated error rates for prediction
230 of a given gene family.

231

232 Our analysis suggests that in order to better predict microbial functional profiles in
233 certain environments, it will be of utility to develop tools specific to that environment.
234 There have been encouraging examples in the literature of efforts to make environmental
235 specific databases such as CowPI, a functional inference tool specific to the rumen
236 microbiome, which had better estimates than PICRUSt when used for predicting
237 functional profiles in the bovine environment [27]. We can look forward to similar future
238 refinements in the next generation of these algorithms that will use appropriate reference
239 databases for an environment and analyze individual functional categories to yield
240 confidence scores for each prediction.

241

242 # Methods

243 The datasets used in this study include 2 human datasets (named as Human_KW [28] and
244 Human_TY [29] in our study after the initials of their first authors), 1 gorilla [30], 1
245 mouse [31], 1 chicken [32] and 2 soil datasets Soil_LWM [33] and Soil_AAN [34]. Each
246 dataset has publicly available 16S rRNA and metagenome sequences and is associated
247 with a two-level categorical metadata. The Human_KW study compared urban and rural
248 subjects in China, while US and non-US subjects were compared for the Human_TY
249 study. In the gorilla study, the dry and wet seasons were compared while the mouse study
250 compared community composition of two enterotypes. Lean and fat broiler chicken lines
251 were compared for the chicken study. For the Soil_LWM study, Amazon dark earth and
252 agricultural soil were compared, while forested and deforested soils were compared for
253 the Soil_AAN study. Information regarding data locations, sequencing depth, sample

254 sizes and effect sizes (measured as R^2 in the PERMANOVA test with the function
255 'adonis' in the R package 'vegan') for each study are listed in Table S1.
256
257 The PICRUSt predictions of the 16S rRNA sequences in the datasets followed the
258 developer's instructions [4]. The authors' metagenome analysis results were used when
259 available [29, 31, 33, 34], otherwise the raw sequences were analyzed with humann2
260 following the developer's instructions [35]. In each dataset, all PICRUSt-predicted gene
261 families and pathways were compared to those from metagenome sequencing to
262 determine their discrepancy in the number and types of genes revealed. For genes
263 detected by both PICRUSt and metagenome sequencing, we used two sets of methods to
264 evaluate their consistency. In a first set of methods, we analyzed the Spearman
265 correlation between PICRUSt-predicted gene composition and those from metagenome
266 sequencing. As a control, we permuted the sample labels 100 times and re-calculated
267 Spearman correlation between PICRUSt estimates and metagenome sequencing estimates
268 with sample labels shuffled.
269
270 In a second set of methods, we analyzed the consistency of PICRUSt and metagenome
271 sequencing in the P-values they generated for null hypotheses of no association with
272 metadata. For this purpose, P-values were produced with a t-test of the 2 distinguishable
273 groups in each dataset (Table S1). P-values from the t-test were log10 transformed and
274 multiplied by either 1 or -1 to include the direction of change as indicated below:
275 $$P\_t=\log10(P)* t /abs(t)$$

P_t is the transformed P-value, P is the P-value from t test, and t is the statistic of t-test. We then estimated the consistency of the log-transformed P-values from PICRUSt and metagenome sequencing with Spearman's correlation. To determine whether this method is affected by the low variance of functional profiles, we permuted sample labels of the metagenome sequencing gene compositions 100 times and re-calculated the log-transformed P-values and their correlation with the PICRUSt results.

In order to correct for differences in sample size, each dataset was also subsampled to 5 samples per group to ensure that the different sample sizes of datasets were not unduly influencing our results. The PICRUSt predictions and metagenome sequencing were also compared in each of the 48 level 2 functional categories.

Declarations

Availability of data and material

The datasets analyzed in this study are publicly available with repositories and accession numbers listed in Table S1. R scripts used in this study are available at Github (https://github.com/ssun6/Inference_picrust). Additional requests and questions can be addressed to S.S.

Competing interests

The authors declare that they have no competing interests.

References

1. Button KS, Ioannidis JP, Mokrysz C, Nosek BA, Flint J, Robinson ES, Munafò MR: Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience* 2013, 14:365.
2. Jervis-Bardy J, Leong LE, Marri S, Smith RJ, Choo JM, Smith-Vaughan HC, Nosworthy E, Morris PS, O'Leary S, Rogers GB: Deriving accurate microbiota profiles from human samples with low bacterial content through post-sequencing processing of Illumina MiSeq data. *Microbiome* 2015, 3:19.
3. Minich JJ, Zhu Q, Janssen S, Hendrickson R, Amir A, Vetter R, Hyde J, Doty MM, Stillwell K, Benardini J: KatharoSeq enables high-throughput microbiome analysis from low-biomass samples. *MSystems* 2018, 3:e00218-00217.
4. Langille MGI, Zaneveld J, Caporaso JG, McDonald D, Knights D, Reyes JA, Clemente JC, Burkepile DE, Vega Thurber RL, Knight R, et al: Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nat Biotech* 2013, 31:814-821.
5. Aßhauer KP, Wemheuer B, Daniel R, Meinicke P: Tax4Fun: predicting functional profiles from metagenomic 16S rRNA data. *Bioinformatics* 2015, 31:2882-2884.
6. Louca S, Parfrey LW, Doebeli M: Decoupling function and taxonomy in the global ocean microbiome. *Science* 2016, 353:1272-1277.
7. Gevers D, Kugathasan S, Denson LA, Vázquez-Baeza Y, Van Treuren W, Ren B, Schwager E, Knights D, Song SJ, Yassour M: The treatment-naive microbiome in new-onset Crohn's disease. *Cell host & microbe* 2014, 15:382-392.
8. Goodrich JK, Waters JL, Poole AC, Sutter JL, Koren O, Blekhman R, Beaumont M, Van Treuren W, Knight R, Bell JT: Human genetics shape the gut microbiome. *Cell* 2014, 159:789-799.
9. Carmody RN, Gerber GK, Luevano Jr JM, Gatti DM, Somes L, Svenson KL, Turnbaugh PJ: Diet dominates host genotype in shaping the murine gut microbiota. *Cell host & microbe* 2015, 17:72-84.
10. De Palma G, Blennerhassett P, Lu J, Deng Y, Park A, Green W, Denou E, Silva M, Santacruz A, Sanz Y: Microbiota and host determinants of behavioural phenotype in maternally separated mice. *Nature communications* 2015, 6:7735.
11. Bolnick DI, Snowberg LK, Hirsch PE, Lauber CL, Org E, Parks B, Lusis AJ, Knight R, Caporaso JG, Svanbäck R: Individual diet has sex-dependent effects on vertebrate gut microbiota. *Nature communications* 2014, 5:ncomms5500.
12. Ainsworth TD, Krause L, Bridge T, Torda G, Raina J-B, Zakrzewski M, Gates RD, Padilla-Gamiño JL, Spalding HL, Smith C: The coral core microbiome identifies rare bacterial taxa as ubiquitous endosymbionts. *The ISME journal* 2015, 9:2261.
13. Revetta R, Gomez-Alvarez V, Gerke T, Santo Domingo J, Ashbolt N: Changes in bacterial composition of biofilm in a metropolitan drinking water distribution system. *Journal of applied microbiology* 2016, 121:294-305.

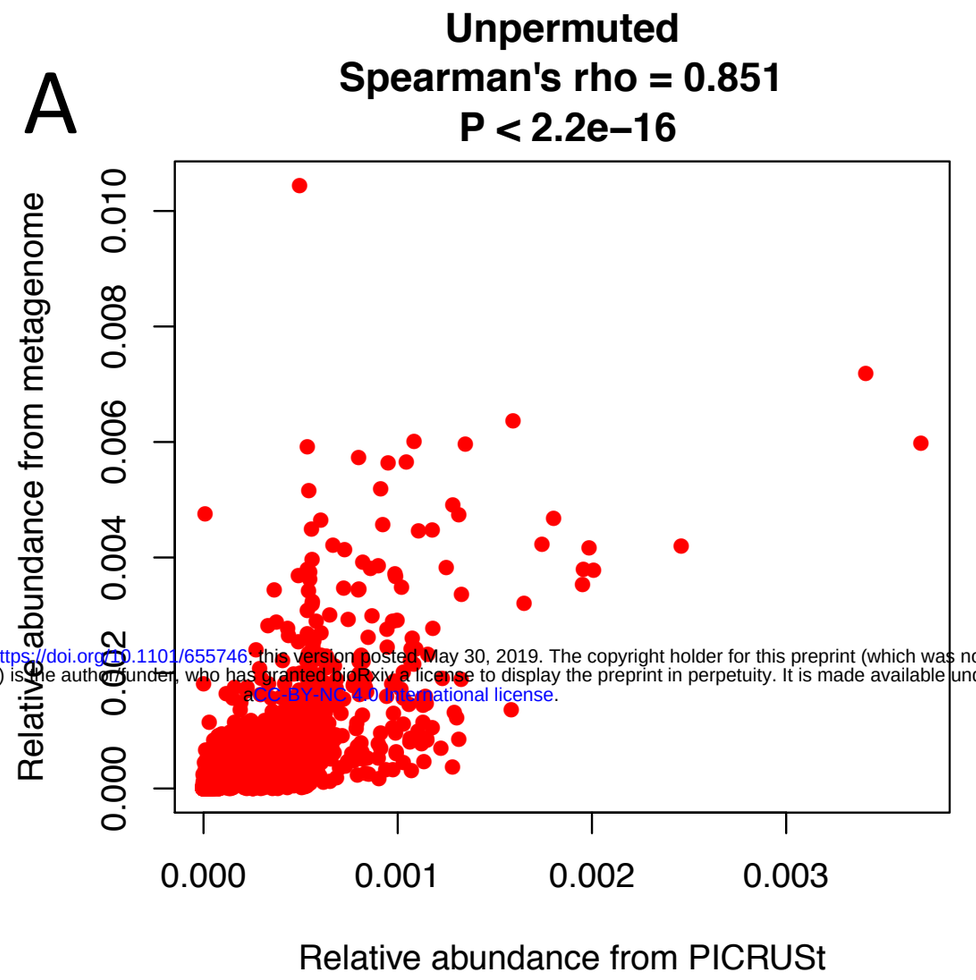
14. Zarraonaindia I, Owens SM, Weisenhorn P, West K, Hampton-Marcell J, Lax S, Bokulich NA, Mills DA, Martin G, Taghavi S: The soil microbiome influences grapevine-associated microbiota. *MBio* 2015, 6:e02527-02514.
15. Li X, Sun S, Yuan H, Badgley BD, He Z: Mainstream upflow nitritation-anammox system with hybrid anaerobic pretreatment: Long-term performance and microbial community dynamics. *Water research* 2017, 125:298-308.
16. Ling N, Zhu C, Xue C, Chen H, Duan Y, Peng C, Guo S, Shen Q: Insight into how organic amendments can shape the soil microbiome in long-term field experiments as revealed by network analysis. *Soil Biology and Biochemistry* 2016, 99:137-149.
17. Choi J, Yang F, Stepanauskas R, Cardenas E, Garoutte A, Williams R, Flater J, Tiedje JM, Hofmockel KS, Gelder B: Strategies to improve reference databases for soil microbiomes. *The ISME journal* 2017, 11:829.
18. Huttenhower C, Gevers D, Knight R, Abubucker S, Badger JH, Chinwalla AT, Creasy HH, Earl AM, FitzGerald MG, Fulton RS: Structure, function and diversity of the healthy human microbiome. *nature* 2012, 486:207.
19. Franzosa EA, Hsu T, Sirota-Madi A, Shafquat A, Abu-Ali G, Morgan XC, Huttenhower C: Sequencing and beyond: integrating molecular'omics' for microbial community profiling. *Nature Reviews Microbiology* 2015, 13:360.
20. Ortiz-Álvarez R, Fierer N, de los Ríos A, Casamayor EO, Barberán A: Consistent changes in the taxonomic structure and functional attributes of bacterial communities during primary succession. *The ISME journal* 2018, 12:1658.
21. David LA, Maurice CF, Carmody RN, Gootenberg DB, Button JE, Wolfe BE, Ling AV, Devlin AS, Varma Y, Fischbach MA: Diet rapidly and reproducibly alters the human gut microbiome. *Nature* 2014, 505:559.
22. Mao S, Zhang M, Liu J, Zhu W: Characterising the bacterial microbiota across the gastrointestinal tracts of dairy cattle: membership and potential function. *Scientific reports* 2015, 5:16116.
23. Loudon AH, Woodhams DC, Parfrey LW, Archer H, Knight R, McKenzie V, Harris RN: Microbial community dynamics and effect of environmental microbial reservoirs on red-backed salamanders (Plethodon cinereus). *The ISME journal* 2014, 8:830.
24. Grice EA, Segre JA: The human microbiome: our second genome. *Annual review of genomics and human genetics* 2012, 13:151-170.
25. Vartoukian SR, Palmer RM, Wade WG: Strategies for culture of 'unculturable'bacteria. *FEMS microbiology letters* 2010, 309:1-7.
26. Gogarten JP, Townsend JP: Horizontal gene transfer, genome innovation and evolution. *Nature Reviews Microbiology* 2005, 3:679.
27. Wilkinson TJ, Huws SA, Edwards JE, Kingston-Smith AH, Siu-Ting K, Hughes M, Rubino F, Friedersdorff M, Creevey CJ: CowPI: A Rumen Microbiome Focussed Version of the PICRUSt Functional Inference Software. *Front Microbiol* 2018, 9:1095.
28. Winglee K, Howard AG, Sha W, Gharaibeh RZ, Liu J, Jin D, Fodor AA, Gordon-Larsen P: Recent urbanization in China is correlated with a Westernized

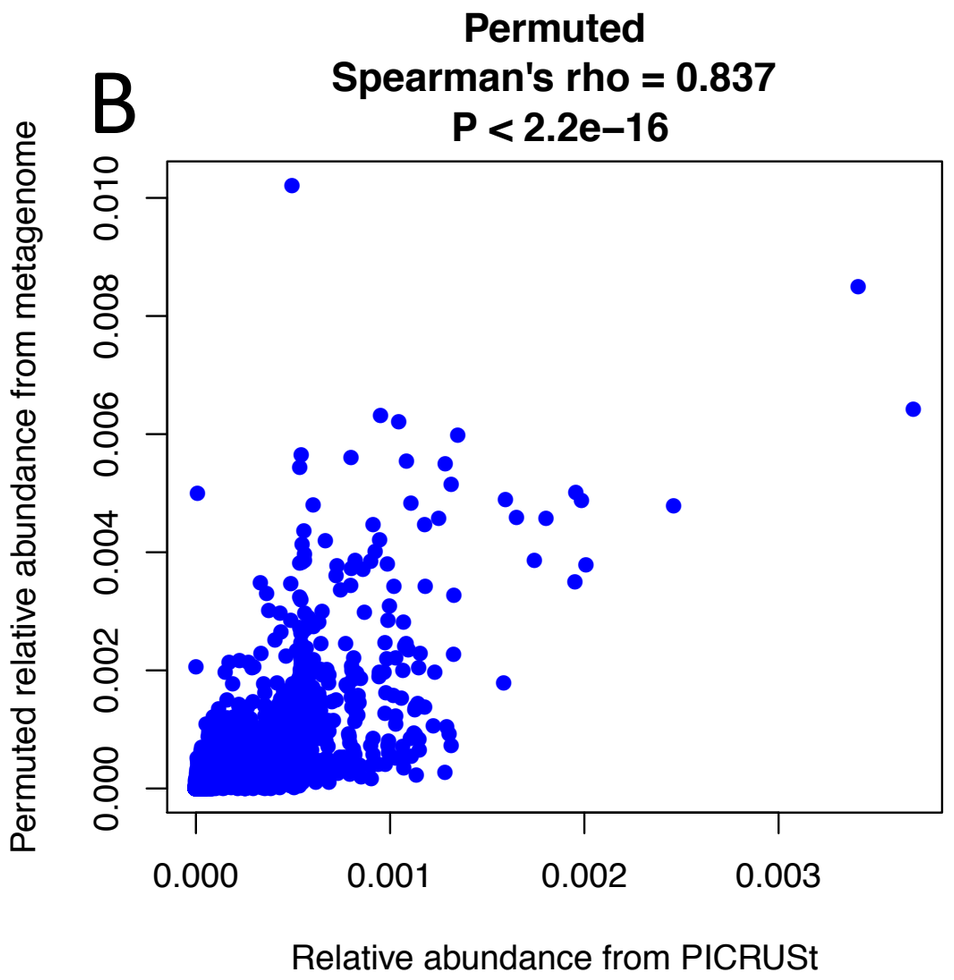
388 microbiome encoding increased virulence and antibiotic resistance genes.
389 *Microbiome* 2017, 5:121.
390 29. Yatsunenko T, Rey FE, Manary MJ, Trehan I, Dominguez-Bello MG, Contreras
391 M, Magris M, Hidalgo G, Baldassano RN, Anokhin AP: Human gut microbiome
392 viewed across age and geography. *nature* 2012, 486:222.
393 30. Hicks AL, Lee KJ, Couto-Rodriguez M, Patel J, Sinha R, Guo C, Olson SH,
394 Seimon A, Seimon TA, Ondzie AU: Gut microbiomes of wild great apes
395 fluctuate seasonally in response to diet. *Nature communications* 2018, 9:1786.
396 31. Wang J, Linnenbrink M, Künzel S, Fernandes R, Nadeau M-J, Rosenstiel P,
397 Baines JF: Dietary history contributes to enterotype-like clustering and
398 functional metagenomic content in the intestinal microbiome of wild mice.
399 *Proceedings of the National Academy of Sciences* 2014:201402342.
400 32. Hou Q, Kwok L-Y, Zheng Y, Wang L, Guo Z, Zhang J, Huang W, Wang Y, Leng L,
401 Li H: Differential fecal microbiota are retained in broiler chicken lines
402 divergently selected for fatness traits. *Scientific reports* 2016, 6:37376.
403 33. Mendes LW, Raaijmakers JM, de Hollander M, Mendes R, Tsai SM: Influence of
404 resistance breeding in common bean on rhizosphere microbiome
405 composition and function. *The ISME journal* 2017, 12:212.
406 34. Navarrete AA, Tsai SM, Mendes LW, Faust K, de Hollander M, Cassman NA,
407 Raes J, van Veen JA, Kuramae EE: Soil microbiome responses to the
408 short-term effects of Amazonian deforestation. *Molecular ecology* 2015,
409 24:2433-2448.
410 35. Franzosa EA, McIver LJ, Rahnavard G, Thompson LR, Schirmer M, Weingart G,
411 Lipson KS, Knight R, Caporaso JG, Segata N: Species-level functional profiling
412 of metagenomes and metatranscriptomes. *Nature methods* 2018, 15:962.
413

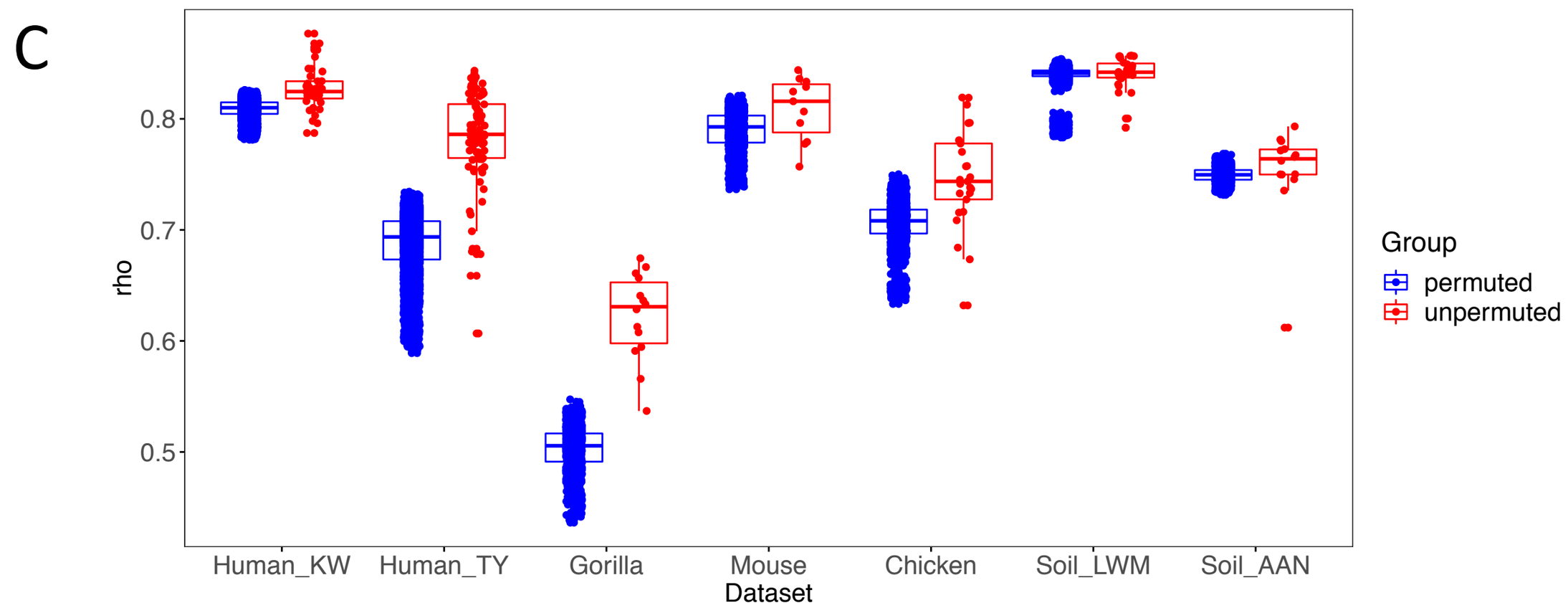
414 Fig. 1. Spearman correlations between PICRUSt and shotgun metagenome sequencing in
415 unpermuted and permuted datasets. **A** and **B:** An example showing the correlations
416 between genes relative abundances estimated by PICRUSt and metagenome sequencing
417 in a soil sample (sample BulkAG3 in soil_AAN dataset) for unpermuted (**A**) and
418 permuted (**B**) sample labels. **C:** Comparison of Spearman's rho between PICRUSt and
419 metagenome sequencing in unpermuted (red) and permuted data (blue) in all datasets. In
420 each of the 100 permutations, all sample labels were shuffled for every gene.
421

422 Fig. 2. Taxonomic (**A)** and functional profiles (**B)** of the 7 datasets in our study. The
423 taxonomic profiles were plotted at the class level and the functional profiles were plotted
424 at the broadest functional category of the KEGG database for visualization.
425
426 Fig. 3. Comparison of inferences based on gene composition estimated with PICRUSt
427 and metagenome sequencing in each of the 7 datasets. The Venn diagrams (left in each
428 panel) show the number of genes detected by both PICRUSt and metagenome sequencing
429 or by only one method, with those detected by PICRUSt in pink and those by
430 metagenome sequencing in blue. The plots (right in each panel) show log-transformed P-
431 values of the t-test evaluating the null hypothesis for each dataset (see methods and Table
432 S1) from metagenome sequencing (x-axis) and PICRUSt (y-axis) for genes in common
433 between the two methods. The sign of the log-transformed P-values reflect the direction
434 of change (see methods). For example, in the Human_KW dataset, genes higher in urban
435 subjects are in the upper-right hand quadrant and genes lower in urban are in the lower-
436 right hand quadrant. The table (inset lower right corner) shows the percentage of genes
437 called by only one method for each dataset and the coefficients of Spearman's
438 correlations between log-transformed P-values based on gene composition estimated by
439 PICRUSt and metagenome sequencing.
440
441 Fig. 4. The results of inference methods in unpermuted and permuted datasets. The red
442 points are the Spearman's rho of log-transformed P-values from PICRUSt and
443 unpermuted metagenome sequencing data for each dataset. The boxplots of blue points
444 show the Spearman's rho of log-tranformed P-values from PICRUSt and permuted

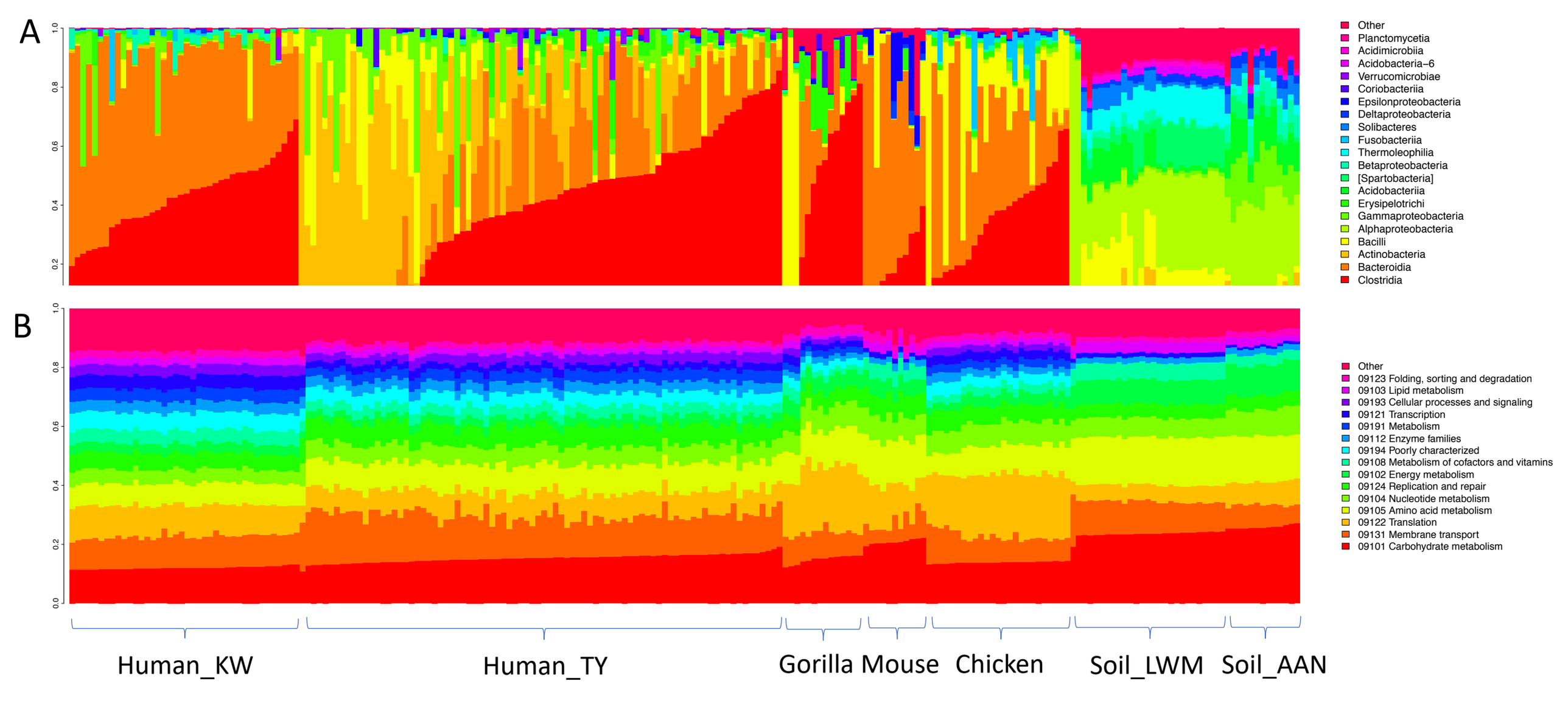
445 metagenome sequencing data in 100 permutations. In each of the 100 permutations, all
446 sample labels were shuffled for every gene.
447
448 Fig. 5. Spearman's rho of the correlations between log-transformed P-values from
449 PICRUSt and metagenome sequencing in 48 KEGG functional categories at the second
450 hierarchy level with the bar colors indicating the functional categories at the first
451 hierarchy level.



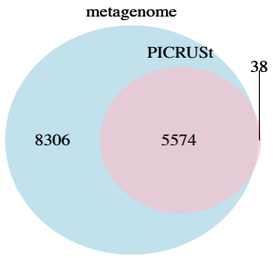


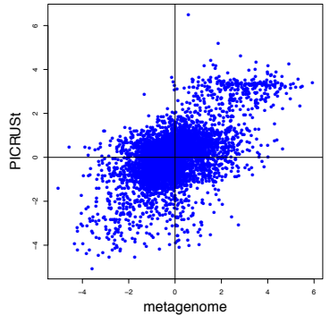


	Human_KW	Human_TY	Gorilla	Mouse	Chicken	Soil_LWM	Soil_AAN
Mean of the Spearman's rho (unpermuted)	0.827	0.782	0.622	0.809	0.747	0.840	0.753
Mean of the Spearman's rho (permuted)	0.809	0.688	0.503	0.789	0.706	0.838	0.750
P-value of the difference between the 2 sets of rho	4.41E-07	< 2.2e-16	4.6E-08	0.0435	0.000142	0.505	0.794

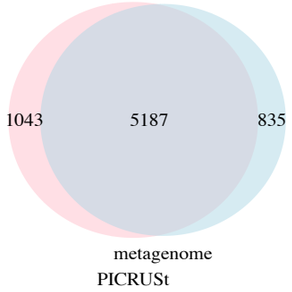

A
B
1.0
0.8
0.6
0.4
0.2
0.0
Other
Planctomycetia
Acidimicrobiia
Acidobacteria−6
Verrucomicrobiae
Coriobacteriia
Epsilonproteobacteria
Deltaproteobacteria
Solibacteres
Fusobacteriia
Thermoleophilia
Betaproteobacteria
[Spartobacteria]
Acidobacteriia
Erysipelotrichi
Gammaproteobacteria
Alphaproteobacteria
Bacilli
Actinobacteria
Bacteroidia
Clostridia
Other
09123 Folding, sorting and degradation
09103 Lipid metabolism
09193 Cellular processes and signaling
09121 Transcription
09191 Metabolism
09112 Enzyme families
09194 Poorly characterized
09108 Metabolism of cofactors and vitamins
09102 Energy metabolism
09124 Replication and repair
09104 Nucleotide metabolism
09105 Amino acid metabolism
09122 Translation
09131 Membrane transport
09101 Carbohydrate metabolism
Human_KW
Human_TY
Gorilla
Mouse
Chicken
Soil_LWM
Soil_AAN

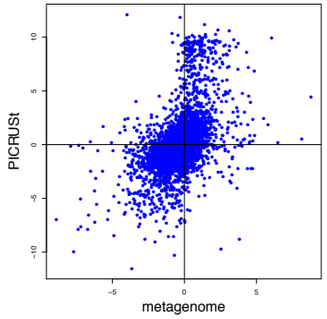
Human_KW



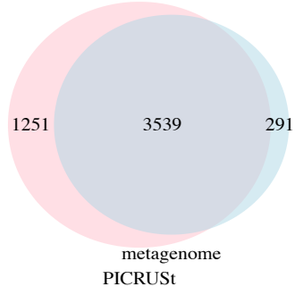


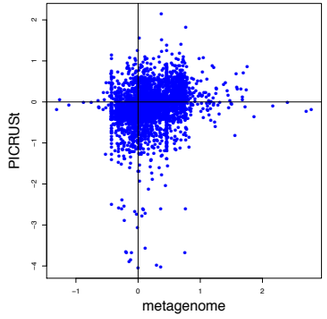
Human_TY



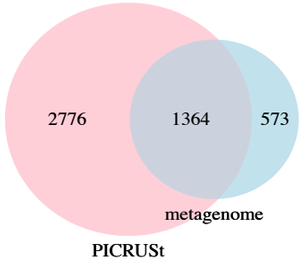


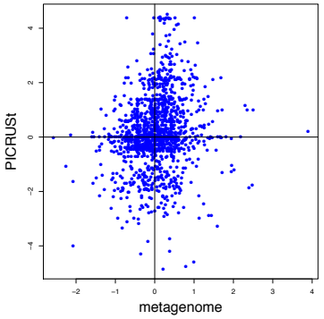
Gorilla



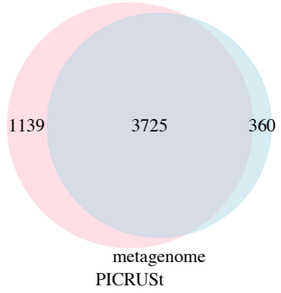


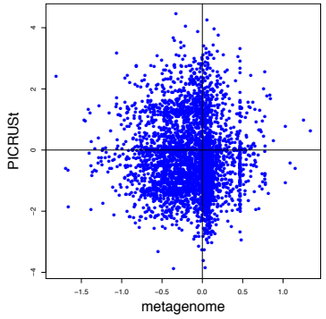
Mouse





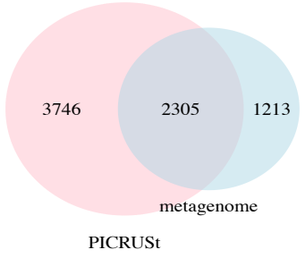
Chicken

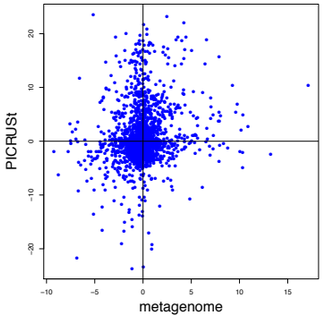




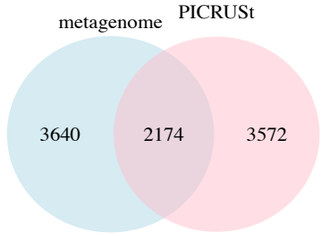
PICRUSt metagenome

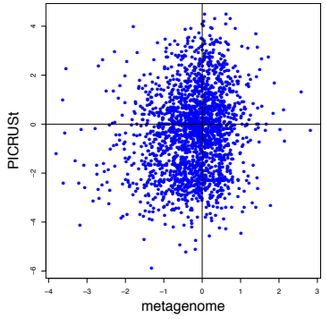
Soil _LWM



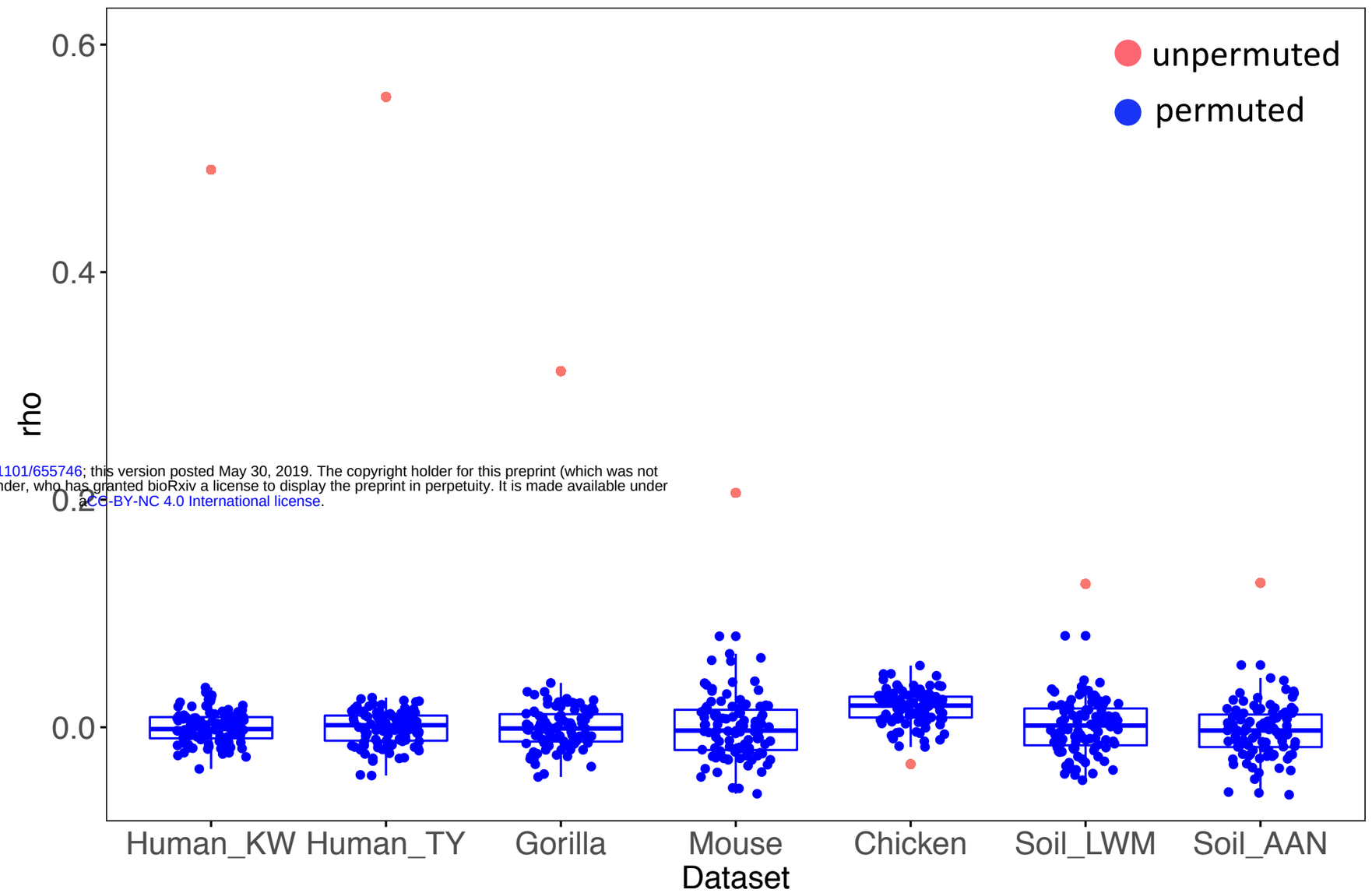


Soil _AAN





Dataset	Percentage of genes not predicted by PICRUSt	Percentage of genes predicted but not in metagenome	Spearman's rho of transformed P-values
Human_KW	59.8%	0.68%	0.490 *
Human_TY	13.9%	16.7%	0.554 *
Gorilla	7.60%	26.1%	0.313 *
Mouse	29.6%	67.1%	0.206 *
Chicken	8.81%	23.4%	-0.0324
Soil_LWM	34.5%	61.9%	0.125 *
Soil_AAN	62.6%	62.2%	0.127 *



Mean of Spearman's rho (unpermuted)	0.490	0.554	0.313	0.206	-0.0324	0.125	0.127
Mean of Spearman's rho (permuted)	-0.000486	-0.000833	-0.00103	-0.000602	0.0178	0.000146	-0.00364


Human_KW
Human_TY
Gorilla
Mouse
Chicken
Soil _LWM
Soil_AAN
09132 Signal transduction
09131 Membrane transport
09133 Signaling molecules and interaction
09167 Infectious diseases
09123 Folding, sorting and degradation
09161 Cancers
09122 Translation
09112 Enzyme families
09121 Transcription
09124 Replication and repair
09141 Transport and catabolism
09151 Immune system
09191 Metabolism
09152 Endocrine system
09101 Carbohydrate metabolism
09107 Glycan biosynthesis and metabolism
09105 Amino acid metabolism
09109 Metabolism of terpenoids and polyketides
09143 Cell growth and death
09102 Energy metabolism
09103 Lipid metabolism
09156 Nervous system
09142 Cell motility
09111 Xenobiotics biodegradation and metabolism
09168 Drug resistance
09108 Metabolism of cofactors and vitamins
09110 Biosynthesis of other secondary metabolites
09193 Cellular processes and signaling
09194 Poorly characterized
09145 Cellular community – prokaryotes
09163 Neurodegenerative diseases
09166 Endocrine and metabolic diseases
09144 Cellular community – eukaryotes
09104 Nucleotide metabolism
09154 Digestive system
09159 Environmental adaptation
09165 Cardiovascular diseases
09125 RNA family
09162 Immune diseases
09164 Substance dependence
09153 Circulatory system
09158 Development
09192 Genetic information processing
09106 Metabolism of other amino acids
09149 Aging
09157 Sensory system
09155 Excretory system
09196 Viral protein family
0.0 0.2 0.4 0.6 0.8 1.0
Spearman’s rho
Metabolism
Genetic Information Processing
Environmental Information Processing
Cellular Processes
Organismal Systems
Human Diseases
Unclassified