Analysis of bronchoalveolar lavage fluid metatranscriptomes among patients with COVID-19

Michael Jochum (0000-0002-2398-356X) [1], Michael D. Lee [2], Kristen Curry [3], Viktorija Zaksas [4], Elizabeth Vitalis [5], Todd Treangen [3], Krista L. Ternus [6], Kjersti Aagaard [1]

[1] Department of Obstetrics and Gynecology, Baylor College of Medicine and Texas Children's Hospital, Houston, TX, USA; [2] Blue Marble Space Institute of Science, Seattle, WA, USA; [3] Department of Computer Science, Rice University, Houston, TX, USA; [4] Center for Translational Data Science, University of Chicago, Chicago, IL, USA; [5] Inscripta, Inc, 5500 Central Ave STE 220, Boulder, CO, USA; [6] Signature Science, LLC, 8329 North Mopac Expressway, Austin, TX, USA

ABSTRACT

To better understand the potential relationship between COVID-19 disease morbidity and microbial community dynamics/functional profiles from a hologenome standpoint, we conducted a multivariate taxonomic and functional microbiome comparison of publicly available human bronchoalveolar lavage fluid (BALF) metatranscriptome samples amongst COVID-19 (*n*=32), community acquired pneumonia (*n*=25), and uninfected samples (*n*=29), and a stratified analysis based on mortality amongst the COVID-19 cohort with known outcomes of deceased *(n*=10) versus survived *(n*=15), with the overarching hypothesis being that there is a potentially informative relationship between the BALF microbiome and the severity of COVID-19 disease onset and progression. We observed 34 functionally discriminant gene ontology (GO) terms in COVID-19 disease as compared to the CAP and uninfected cohorts, and 21 GO terms functionally discriminant to COVID-19 mortality (q < 0.05). GO terms enriched in the COVID-19 cohort included hydrolase activity, and significant GO terms under the parental terms of biological regulation, viral process, and interspecies interaction between organisms, whereas GO terms enriched in the uninfected cohort compared to the COVID-19 cohort included significant GO terms under the parental terms of cellular process, metabolic process, binding, and terms classified under catalytic activity other than hydrolase activity. Notable GO terms associated with COVID-19 mortality included nucleobase-containing compound biosynthetic process, organonitrogen compound catabolic process, pyrimidine-containing compound biosynthetic process, and DNA recombination, RNA binding, magnesium and zinc ion binding, oxidoreductase activity, and endopeptidase activity. A Dirichlet multinomial mixtures clustering analysis resulted in a best model fit using 3 distinct clusters that were significantly associated with COVID-19 disease and mortality. We additionally observed discriminant taxonomic differences associated with COVID-19 disease and mortality in the genus *Sphingomonas,* belonging to the *Sphingomonadacae* family, *Variovorax,* belonging to the *Comamonadaceae* family, and in the class Bacteroidia, belonging to the order Bacteroidales. Collectively, while this data does not speak to causality nor directionality of the association, it does demonstrate a significant relationship between the human microbiome and COVID-19 morbidity and mortality, rendering testable hypotheses that warrant further investigation.

BACKGROUND & OBJECTIVES

- For respiratory viruses like SARS-CoV-2, bronchoalveolar Lavage Fluid (BALF) derived metatranscriptomes sampled from diseased host tissues represent a unique opportunity to investigate how the microbiome is responding to host a viral mediated changes in conditions surrounding SARS-CoV2 infection.
- Early in the SARS-CoV-2 outbreak, scientists openly published metatranscriptome sequences from BALF samples of patients with COVID-19 disease, prompting us to investigate **microbially derived transcriptomic changes surrounding COVID-19 moderate to severe disease and progression**, despite limitations in experimental study design.
- In contrast to other studies that focus on characteristics of the human host response or SARS-CoV-2 lineages and viral variants, our analysis specifically evaluated focused on **microbial taxonomic** and **functional profiles** of the BALF metatranscriptomes.
- Using human BALF metatranscriptomes sourced from **seven publications** and corresponding publicly available sequencing data repositories, we were able to conduct a **multivariate taxonomic** and **functional microbiome comparison** using metatranscriptome sequencing data derived from BALF specimens of subjects grouped into one of three classes: **1) Uninfected controls; 2) Community Acquired Pneumonia (CAP)** patients; or **3) patients with COVID-19 moderate to severe disease**, additionally stratified by **mortality** (Table 1).
- The objective of the study was to compare the **COVID-19** amongst **uninfected** and **CAP** patients **BALF metatranscriptomes** and: **1)** identify changes in **microbial** derived **community dynamics / gene ontologies** associated with **COVID-19.** and **2)** Predict **outcomes** amongst **COVID-19** based on metatranscriptomes profiling, with the overall hypothesis being that there is a **potential informative and discernably significant relationship between the BALF microbiome and the severity of COVID-19 disease.**

Table 1. Overview of Meta-analysis Dataset Clinical Characteristics (*n*=86)

Variable	Uninfected	Community Acquired Pneumonia	COVID-19
Cohort	29 (33.72%)	25 (29.07%)	32(37.21%)
Sex			
female	4 (18.18%)	8 (36.36%)	10 (45.45%)
male	5 (13.15%)	11 (28.94%)	22 (57.89%)
unspecified	20 (76.92%)	6(23.07%)	0 (0%)
Publication			
Chen	0 (0%)	0 (0%)	2 (100%)
Ren	9 (100%)	0 (0%)	0 (0%)
Shen	20 (32.79%)	25 (40.98%)	16 (40.98%)
Wu	0 (0%)	0 (0%)	1 (100%)
Xiong	0 (0%)	0 (0%)	4 (100%)
Zhou	0 (0%)	0 (0%)	9 (100%)
Numeric variables			
(mean ± SD)			
Age	53.2 ± 13.3 (*n*=9)	51.2 ± 19.8 *(n*=17)	47.3 ± 11.5 *(n*=32)
Temp. °C	-	38.4 ± 0.91 *(n*=15)	38.4 ± 0.715 *(n*=8)
days after onset	-	9.07 ± 3.17 *(n*=14)	12.05 ± 6.5 *(n*=41)

Table 2. Overview of COVID-19 Sample Characteristics (*n*=32)

Variable	COVID-19 Cohort
Outcome	***n* =32**
Deceased	10 (31.25%)
Survived	15 (46.87%)
Unspecified	7 (21.88%)
Cough	
aggravated	1 (2.13%)
expectoration	3 (9.38%)
intermittent	2 (6.25%)
yes	6 (18.75%)
Unspecified	20 (62.5%)
days delayed hospitalization	
mean ± SD (n)	5.27 ± 3.29 (n=11)

METHODS

Bioinformatic Processing. Illumina derived metatranscriptome sequencing reads were obtained from the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) or the China National Center for Bioinformation (CNCB) National Genomics Data Center (NGDC), trimmed, and assessed for quality before and after using FastQC and Trimmomatic. To control for different sequencing approaches by dataset (e.g., datasets being paired, or single-end reads), all paired-end reads were merged with FLASH (11) and concatenated with unmerged reads into one fastq file per sample, and low-complexity sequences were removed with fastp. Taxonomic classification was subsequently performed with Kraken2 utilizing the standard database + SARS CoV2. Classified fastq datasets were filtered of any human and phix classified reads and analyzed with SeqScreen to obtain a list of leaf-node molecular function and biological process Gene Ontology (GO) terms present within each of the samples. The CoV-IRT-Micro conda package (https://github.com/AstrobioMike/CoV-IRT-Micro) was used to propagate parent GO terms, parse SeqScreen outputs by taxonomic domain, and summarize Kraken2 taxonomic results and SeqScreen-reported protein identifiers.

Metransciptome Analysis. Parent-propagated GO term counts, sample read counts, taxonomic classifications, and curated clinical metadata were imported into a working phyloseq object and removed of potential contamination whenever negative controls were present using the package decontam when possible. After read-filtering and batch-effect sample removal, sample cohorts of *n*=29 uninfected samples from 29 subjects, *n*=25 CAP samples from 25 subjects, and *n*=32 COVID-19 samples from 18 subjects were available for comparison (total, *n*=86 BALF samples from *n*=72 subjects). Amongst the COVID-19 cohort at the time of the index study publication, *n*=10 samples were from 5 known-deceased subjects, *n*=15 samples were from 9 known-survived subjects, and *n*=7 from 4 subjects of the total 32 COVID-19 samples in this meta-analysis with unknown / unpublished survival outcomes. GO term abundances from the remaining subjects' specimens were compositionally transformed, center log ratio (CLR) normalized, and independently compared by case type (COVID-19 vs CAP and Uninfected) and survival outcome (COVID-19 only deceased vs survived) using MaAsLin2 with minimum abundance, prevalence, and significance cutoffs of 0.01, 0.1, and q < 0.05 (Benjamini-Hochberg multiple test correction). Additionally, GO term counts subjected to and unsupervised clustering community typing with Dirichlet Multinomial Mixtures (DMM) using square root scaled counts, followed statistical comparisons using analysis of variance (ANOVA) with metadata categories case type and survival outcome. Statistically significant GO terms results derived from the MaAsLin2 analysis were thereafter ordered by parental lineage and visualized alongside consensus DMM clusters and metadata categories publication, case type, and survival outcome using the package pheatmap (v1.0.12). Taxonomic comparisons were analyzed by case type and survival outcome with heat tree visualizations using log2 median ratio differences using metacoder (v0.34). An overview of the processing workflow as well as additional supplementary information about the clinical metadata, intermediate processing commands, and parameters used in the bioinformatic pipeline, can be found at the COV-IRT microbial github (https://github.com/COV-IRT/microbial) and open science framework (OSF) project (https://osf.io/7nrd3/) websites.

RESULTS

COVID-19 vs Uninfected & viral pneumonia

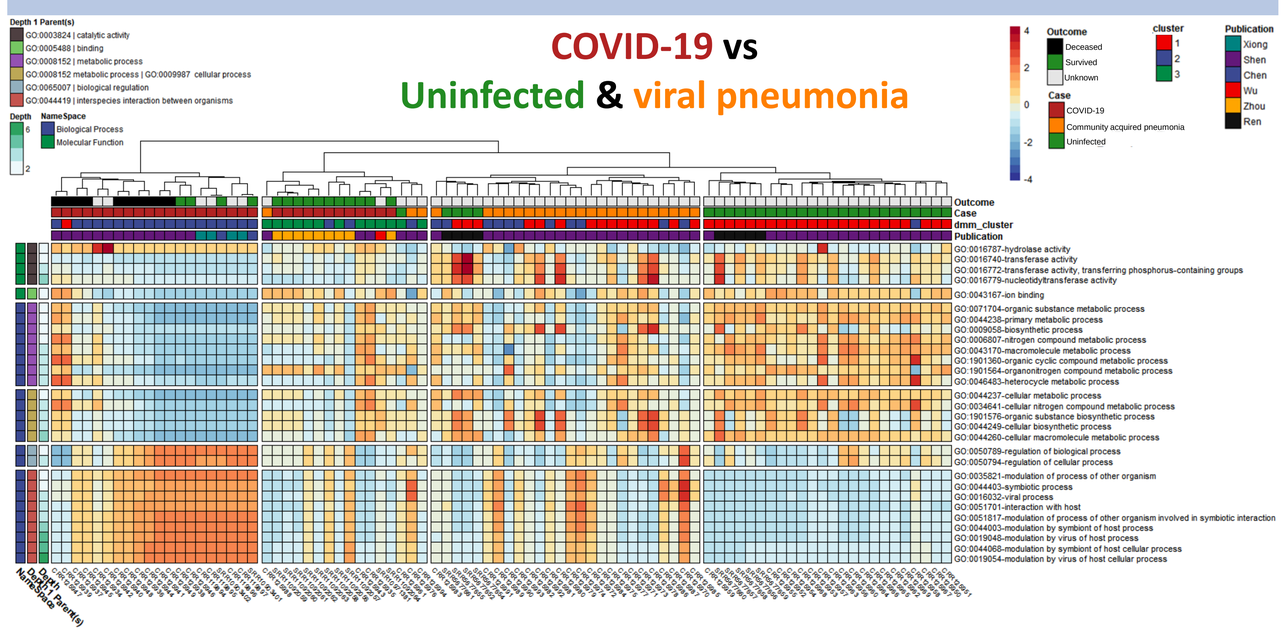

Figure 1. Heatmap with Notable Microbially-Derived Gene Ontology Functional Annotations Associated with COVID-19 (n=32), as Compared to Community Acquired Pneumonia (n=29) & Uninfected (n=25) Cohorts. Rows are sorted by parental GO terms (depth=1), and columns are clustered by Euclidean distance using ward D2 clustering. Comparisons were conducted using MaAsLin2, controlling for random effects of publication and patient and correcting for multiple tests (q<0.05).

RESULTS

Taxonomic comparison of COVID-19 BALF metatranscriptome Deceased vs Survived

Table 3. Significant taxa with Log2 median ratio counts differentially associated with COVID-19 mortality when comparing deceased (*n*=10) versus survived (*n*=15).

log_2 median ratio	Median diff	Mean diff	p value	q value	Taxonomy
2.25	0.361	0.371	0.00017	0.00691	*Comamonadaceae*
5.21	0.405	0.377	0.00017	0.00691	*Variovorax*
-5.13	-0.103	-0.104	0.0308	0.199	*Bacteroidia*
-5.18	-0.099	-0.102	0.00962	0.124	*Bacteroidales*

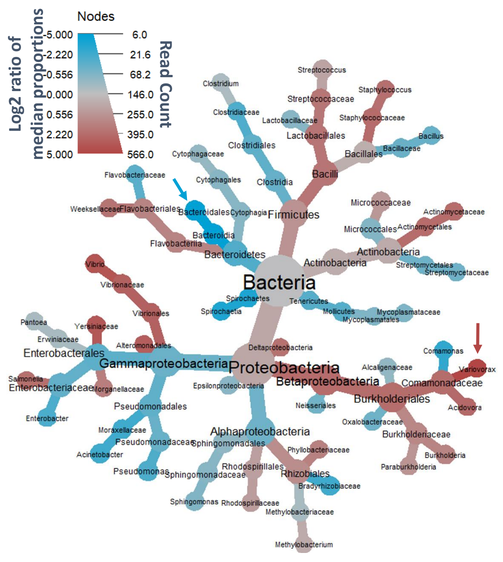

Figure 3. Taxonomic Heat Tree comparisons of BALF metatransciptome profiles with COVID-19 Mortality. Notable increases were observed in the log2 median ratios in the Family *Comamonadaceae,* genus *Variovorax,* and significant decreases in the log2 median ratios of order *Bacteroidia* and class *Bacteroidales.*

Functional comparison of COVID-19 BALF metatranscriptome Deceased vs Survived

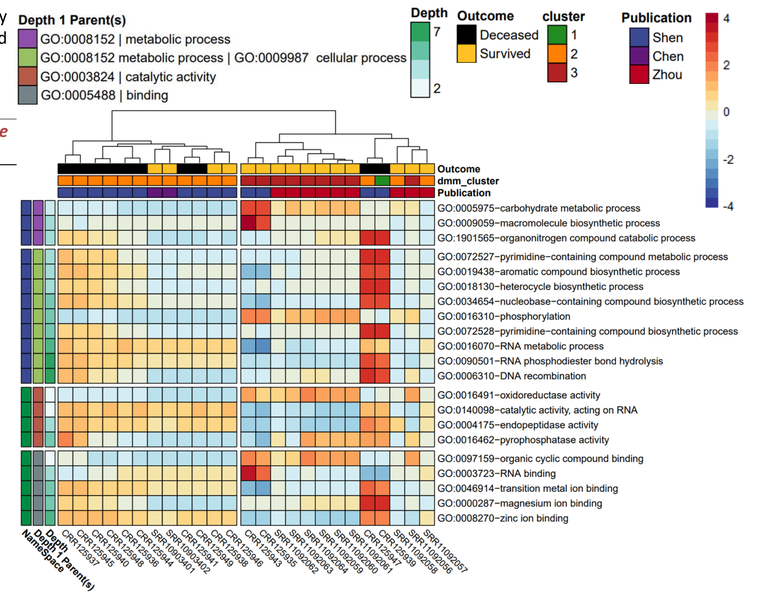

Figure 2. Heatmap of Significantly Different Gene Ontology Terms Associated with COVID-19 Mortality Comparing Deceased (n=10) versus Survived (n=15). Rows are sorted by parental GO terms (depth=1) and columns are clustered by Euclidean distance using ward D2 clustering. Comparisons were conducted using MaAsLin2, controlling for random effects of and patient and correcting for multiple tests (q<0.05).

Table 4. Predicted survival outcome using unsupervised machine learning Dirichlet Multinomial mixture clustering of BALF metatranscriptome gene ontology counts. **A)** Analysis of Variance (ANOVA) showing a statistically significant relationship between disease outcome and DMM cluster **B)** Posthoc Tukey multiple comparisons of means test using 95% CI showing significance when comparing survived vs deceased cohorts.

A)

Feature	Df	Sum Sq.	Mean Sq.	F value	Pr(>F)	
Outcome	2	3.12	1.56	8.42	0.001	**
Residuals	29	5.36	0.19			

B)

comparison	diff	lwr	upr	p adj	
NA-Deceased	0.24	-0.28	0.76	0.494	
Survived-Deceased	0.7	0.27	1.13	0.001	**
Survived-NA	0.46	-0.03	0.94	0.078	

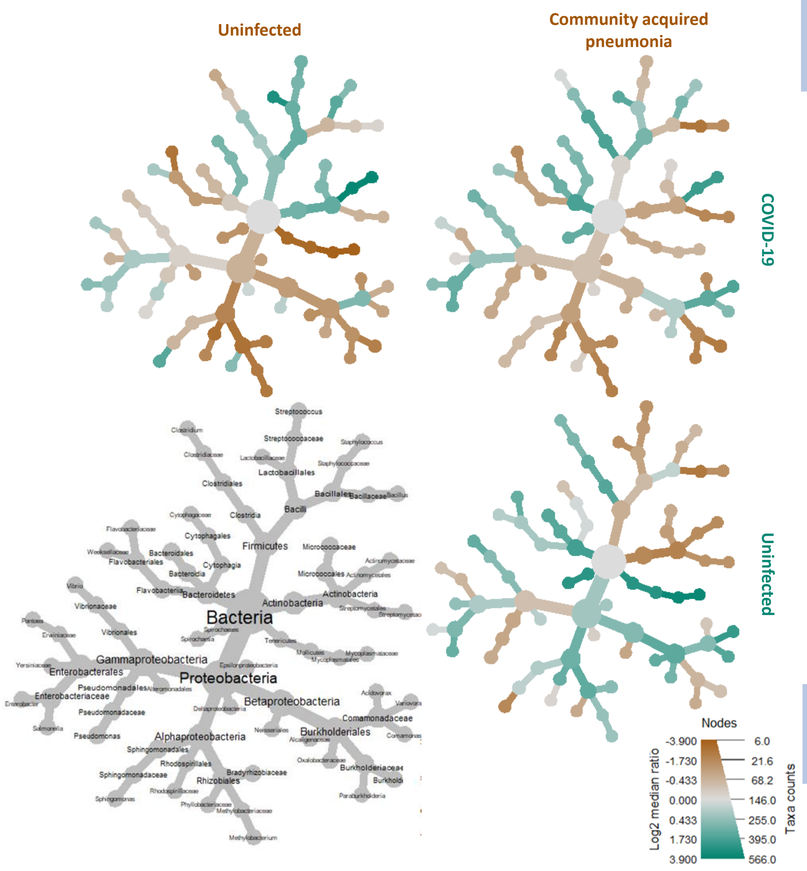

Figure 4. Heat Tree Matrix Visualizing Distinct COVID-19 vs. Uninfected & Viral Pneumonia Taxonomic Profiles. Taxonomic heat tree data matrix visualization depicting the Log 2 median ratio differences across the different cohorts. The tree depicted in grey in the lower left represents a key for the other trees. Each of smaller trees represents a comparison between the different cohorts, as labelled in the columns and rows. Taxa colored brown are more abundant among in the cohort labelled in the columns whereas taxa colored green would be more abundant in the cohort labelled in the row. For example, there were significant increases (green) identified in log2 median ratio of several species belonging to the genus *Sphingomonas* when comparing the COVID-19 vs uninfected cohort (top left) and decreases (brown) when comparing the Uninfected to the community acquired pneumonia cohort (bottom right).

CONCLUSIONS

- **There are unique and discriminant features in the BALF metatranscriptomes of COVID-19 as compared to amongst uninfected and CAP cohorts and by COVID-19 survival outcome (Survival vs deceased).**
- **Notable Gene ontologies of interest surrounding when comparing by case type and survival outcome include:**
  - **Phosphate / phosphorylation,**
  - **Metal ion binding (mg,zn,etc),**
  - **Nucleotide terms (DNA recombination / RNA binding),**
  - **Lytic activity (ie: hydrolase, endopeptidase)**
- **Taxonomic comparisons revealed significant changes in the metatranscriptomic read counts of *Comamonadaceae, Variovorax,* in the deceased cohort and *Bacteroidia, Bacteroidales* amongst those who survived.**
- **Unsupervised machine learning using Dirichlet Multinomial mixture clustering of BALF metatranscriptome gene ontology counts was capable of predicting survival outcome amongst the COVID-19-cohort.**

SIGNIFICANCE

- **This study identified significant taxonomic and functional differences in BALF metatranscriptomes associated with COVID-19 disease and death.**
- **By the nature of this analysis, while this data does not address causality nor directionality of the association, it does identify a significant relationship between the human microbiome and COVID-19 morbidity and mortality, in which the specific functions and taxa identified through this investigation warrant further study.**

ACKNOWLEDGMENTS

COVID-19 International Research Team

This project was supported by NIH NICHD : T32 HD098068 (M.J.). Additional time and support for the analysis was provided from Dr. Mike Lee at NASA / Blue Marble Space Institute of Science, Mrs. Viktoria Zaksas at the University of Chicago, Mrs. Kristen Curry and Dr. Todd Treangen Trainjen at Rice University, and by the COV-IRT microbial subgroup team leader Dr. Krista Ternus at Signature science. We would also like to thank the Texas Advanced Computing Center (TACC) at The University of Texas at Austin for providing HPC resources that have contributed to the research results reported.

REFERENCES

1) Chen L, Liu W, Zhang Q, Xu K, Ye G, Wu W, et al. RNA based mNGS approach identifies a novel human coronavirus from two individual pneumonia cases in 2019 Wuhan outbreak. Emerg Microbes Infect. 2020;9: 313–319. 2) Chen S, Zhu Q, Xiao Y, Wu C, Jiang Z, Liu L, et al. Clinical and etiological analysis of co-infections and secondary infections in COVID-19 patients: An observational study. Clin Respir J. 2021;15: 815–825. 3)Wu F, Zhao S, Yu B, Chen Y-M, Wang W, Song Z-G, et al. A new coronavirus associated with human respiratory disease in China. Nature. 2020;579: 265–269. 4) Zhou P, Yang X-L, Wang X-G, Hu B, Zhang L, Zhang W, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature. 2020;579: 270–273. 5) Xiong Y, Liu Y, Cao L, Wang D, Guo M, Jiang A, et al. Transcriptomic characteristics of bronchoalveolar lavage fluid and peripheral blood mononuclear cells in COVID-19 patients. Emerg Microbes Infect. 2020;9: 761–770. 6) Shen Z, Xiao Y, Kang L, Ma W, Shi L, Zhang L, et al. Genomic Diversity of Severe Acute Respiratory Syndrome–Coronavirus 2 in Patients with Coronavirus Disease 2019. Clinical Infectious Diseases. 2020. Doi:10.1093/cid/ciaa203. 7) Ren L, Zhang R, Rao J, Xiao Y, Zhang Z, Yang B, et al. Transcriptionally Active Lung Microbiome and Its Association with Bacterial Biomass and Host Inflammatory Status. mSystems. 2018;3. Doi:10.1128/mSystems.00199-18. 8) Sirivongrangson P, Kulvichit W, Payungporn S, Pisitkun T, Chindamporn A, Peerapornratana S, et al. Endotoxemia and circulating bacteriome in severe COVID-19 patients. Intensive Care Med Exp. 2020;8: 72. 9) Han Y, Jia Z, Shi J, Wang W, He K. The active lung microbiota landscape of COVID-19 patients. medRxiv. 2020; 2020.08.20.20144014. 10) Balaji A, Kille B, Kappell AD, Godbold GD, Diep M, Leo Elworth RA, et al. SeqScreen: Accurate and Sensitive Functional Screening of Pathogenic Sequences via Ensemble Learning. bioRxiv. 2021. p. 2021.05.02.442344. doi:10.1101/2021.05.02.442344. 11) Foster ZSL, Sharpton TJ, Grünwald NJ. Metacoder: An R package for visualization and manipulation of community taxonomic diversity data. PLoS Comput Biol. 2017;13: e1005404.