

# AMDB: a database of animal gut microbial communities with manually curated metadata

Junwon Yang<sup>1,2,3</sup>, Jonghyun Park<sup>1,2,3</sup>, Yeonjae Jung<sup>1</sup> and Jongsik Chun<sup>1,2,3,\*</sup>

<sup>1</sup>Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul 08826, Korea, <sup>2</sup>Institute of Molecular Biology and Genetics, Seoul National University, Seoul 08826, Korea and <sup>3</sup>Department of Biological Sciences, Seoul National University, Seoul 08826, Korea

Received August 15, 2021; Revised October 11, 2021; Editorial Decision October 11, 2021; Accepted November 01, 2021

## ABSTRACT

**Variations in gut microbiota can be explained by animal host characteristics, including host phylogeny and diet. However, there are currently no databases that allow for easy exploration of the relationship between gut microbiota and diverse animal hosts. The Animal Microbiome Database (AMDB) is the first database to provide taxonomic profiles of the gut microbiota in various animal species. AMDB contains 2530 amplicon data from 34 projects with manually curated metadata. The total data represent 467 animal species and contain 10 478 bacterial taxa. This novel database provides information regarding gut microbiota structures and the distribution of gut bacteria in animals, with an easy-to-use interface. Interactive visualizations are also available, enabling effective investigation of the relationship between the gut microbiota and animal hosts. AMDB will contribute to a better understanding of the gut microbiota of animals. AMDB is publicly available without login requirements at <http://leb.snu.ac.kr/amdb>.**

## INTRODUCTION

Animal gut microbiota is a diverse microbial community that lives in the intestine of the host and consists of predominantly bacteria, as well as some archaea, fungi, protozoa and viruses (1). The gut microbiota has received widespread attention due to its potential to influence host physiology (2), immunity (3) and development (4). The gut microbiota has also been hypothesized to contribute to host evolution (5).

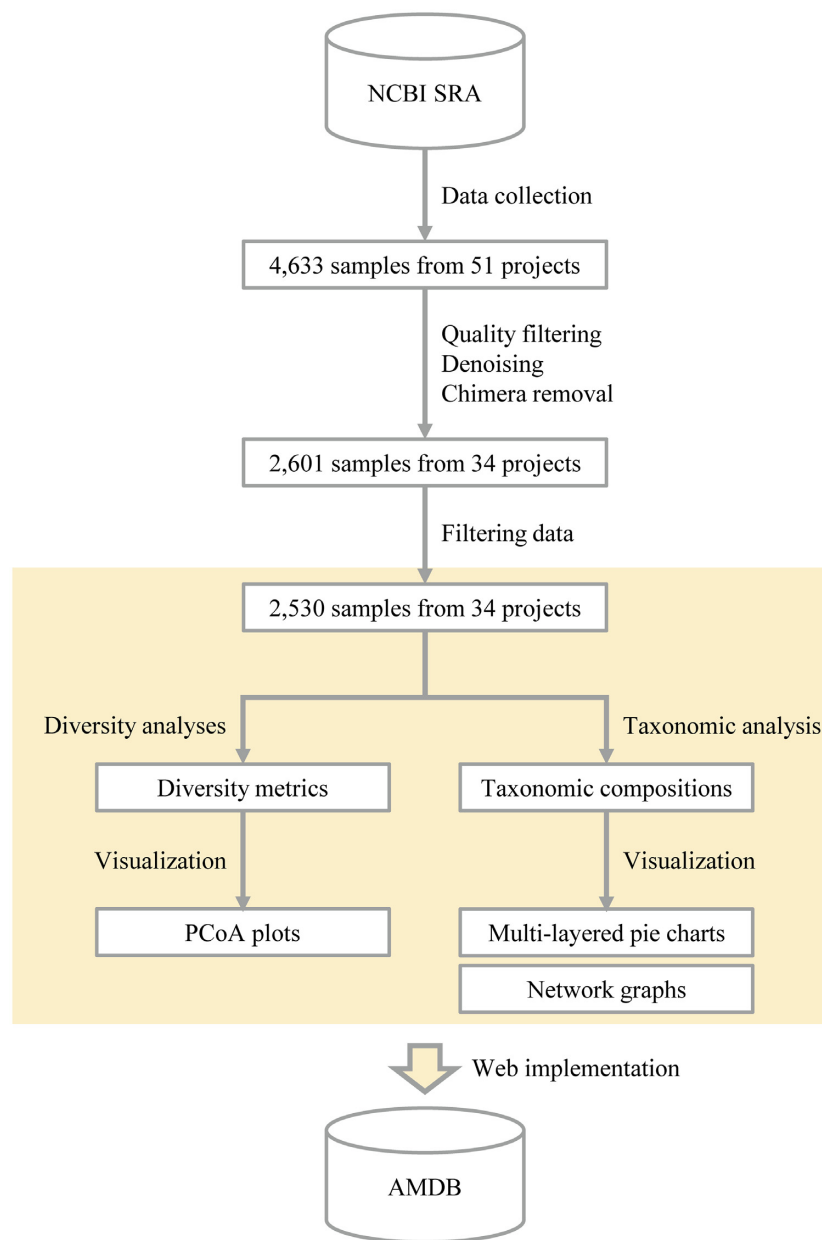
The gut microbiota and the host display a bidirectional interaction. Various studies have shown that variations in gut microbiota can be explained by differences in host characteristics (6–8). In particular, host phylogeny and diet largely account for the gut microbiota variations (7).

A recent analysis of samples from wild baboons found widespread gut microbiome heritability (9). This vertical transmission may be one of the drivers of phyllosymbiosis (10). Phyllosymbiosis is defined as ‘microbial community relationships that recapitulate the phylogeny of their host’ (11). Patterns of phyllosymbiosis have been reported in many studies (12–14). Additionally, the host diet may also affect the gut microbiota, with several studies reporting that host diet can lead to the convergence of gut microbes in the host species (10,15–18).

Despite the importance of the relationship between gut microbiota and host characteristics, specifically host phylogeny and diet, there is currently no database available that enables easy exploration of the gut microbiota of various animal hosts. Most curated databases focus only on humans (GIMICA (19), GMrepo (20) and HPMCD (21)) and mice (MMDB (22)). There are several databases that contain microbiota data from various animal hosts, including IMNGS (23), MGnify (24), MG-RAST (25) and Qiita (26). However, these databases contain data from various sources other than solely from the animal hosts, making it difficult to identify the relationship between the gut microbiota and animal hosts.

Here, we present Animal Microbiome Database (AMDB) that overcomes these limitations. AMDB includes bacterial 16S ribosomal RNA (rRNA) gene profiles from various animal species to enable the assessment of the relationship between gut microbiota and animal hosts. AMDB currently incorporates 10 478 bacterial taxa and 2530 samples from 34 projects, representing 467 animal species with manually curated metadata. This novel database (i) supports searches by the bacterial taxon of interest, (ii) provides a taxonomic composition of each sample, (iii) incorporates summary information for each project and host and (iv) includes interactive visualizations. Therefore, AMDB will help scientists to quickly access animal gut microbiota data through a user-friendly interface.

\*To whom correspondence should be addressed. Tel: +82 2 880 8153; Email: [jchun@snu.ac.kr](mailto:jchun@snu.ac.kr)  
Present address: Jongsik Chun, Department of Biological Sciences, Seoul National University, Seoul 08826, Korea.



**Figure 1.** Schematic diagram of AMDB construction. The process name for each step is displayed next to the arrow. The contents contained in AMDB is highlighted in yellow.

## MATERIALS AND METHODS

### Data collection and curation process

We manually selected candidate data for AMDB from the NCBI Sequence Read Archive (SRA) based on the following criteria: (i) samples included fecal or intestinal contents from individual healthy animal hosts, (ii) the PCR primers had to target the V4 hypervariable region of the 16S rRNA gene, (iii) amplicons had to be sequenced on Illumina instruments, (iv) samples had to be linked to research articles. For longitudinal data, only one sample was selected as follows; only one adult sample was included when the samples were from multiple life stages, and one sample from an earlier time point was selected for a given life stage. Samples that were duplicates of those previously included in the

AMDB were not included. Amplicon data from different hypervariable regions of the 16S rRNA gene cannot be directly compared due to differences in binding affinity and resolution (27,28). We only used amplicon data from the V4 hypervariable region to ensure comparability. Illumina data was used because we used the Deblur for data processing, which was designed for Illumina data (29). To ensure that samples were of high-quality, we only selected samples linked to research articles. We checked the suitability of samples by reading the publication materials and methods, and we collected metadata, including the accession numbers and the host information. We extracted information on host diets from the MammalDIET (30) and the EltonTraits database (31). A total of 4633 samples were obtained from 51 projects (Figure 1).

## Data processing

Figure 1 summarizes all of the data processing steps. The entire analysis was performed using QIIME 2 (Version 2021.2) (32). Paired-end reads were merged using VSEARCH with default parameters (33). The total number of sequencing reads was 434 900 445. The sequencing reads were quality filtered as follows; reads were truncated at any site containing >3 consecutive low-quality base (Phred score < 4), and the minimum fraction of consecutive high-quality bases to be retained was set to 75% of the length of the input sequence with no uncalled bases (Ns) (34). The total number of sequencing reads after the quality filtering was 432 039 098. The Deblur was used for denoising and chimera removal to obtain amplicon sequence variants (ASVs) using a trim length of 250 bases (29,35). The resulting ASVs from all samples were combined into a BIOM table (36). After using the Deblur, a total of 81 701 877 reads were obtained from 2601 samples (34 projects), with an average of 31 412 reads per sample (a minimum of 2 reads and maximum of 205 611 reads). Samples with a minimum of 1000 reads were included after denoising and chimera removal, and a total of 2530 samples from 34 projects were available (the total number of sequencing reads was 81 669 682).

For diversity analyses, we normalized sequencing reads to 1000 reads by rarefying (37) and scaling with ranked subsampling (SRS) (38). Alpha diversity indices, including the observed ASVs and the Shannon index (39), were calculated from samples before normalization, after rarefying and after SRS (40). All ASVs were aligned with MAFFT (41) and were used to construct a phylogenetic tree with FastTree 2 (40,42). Using the phylogenetic tree, we calculated unweighted and weighted UniFrac distances to measure beta diversity after rarefying samples to 1000 reads (37,43–47). Principal coordinate analysis (PCoA) was performed based on the unweighted and weighted UniFrac distances (48,49), and PCoA plots were visualized with Emperor (50,51).

For taxonomic analysis, taxonomy was assigned to ASVs using the q2-feature-classifier classify-consensus-vsearch (33,40,52) against the EzBioCloud (53). All matches with an identity percentage of 0.97 or higher were kept. We only used bacterial 16S rRNA gene sequences from the EzBioCloud. Multi-layered pie charts representing the taxonomic composition were visualized with Krona (54), and network graphs representing the associations between bacteria and hosts were visualized with Flourish (<https://flourish.studio/>).

## Web implementation

Data were stored in a MySQL database (<https://www.mysql.com/>). The backend was implemented in Java using Spring Boot (<https://spring.io/projects/spring-boot>). The user interface was implemented using CSS, HTML and JavaScript with Bootstrap (<https://getbootstrap.com/>), jQuery (<https://jquery.com/>) and Thymeleaf (<https://www.thymeleaf.org/>). DataTables (<https://datatables.net/>) and Plotly.js (<https://plotly.com/javascript/>) were used for data visualization. The 'Visualization' page was made by referring to the format of *Peryton* (55). We tested AMDB on Google Chrome, Microsoft Edge and Mozilla Firefox to provide a robust ser-

**Table 1.** Data summary of AMDB

Variable	N
<b>Samples</b>	2530
<b>Projects</b>	34
<b>Features (ASVs)</b>	139 375
<b>Sequencing reads</b>	81 669 682
<b>Bacterial taxa</b>	
Phylum	44
Class	115
Order	280
Family	687
Genus	2828
Species	6524
<b>Total</b>	10 478
<b>Host taxonomy</b>	
Class	9
Order	63
Family	180
Genus	369
Species	467
<b>Host diet types</b>	4

ASV: amplicon sequence variant

vice. In addition, we made AMDB accessible and legible on phone and tablet screens.

## RESULTS

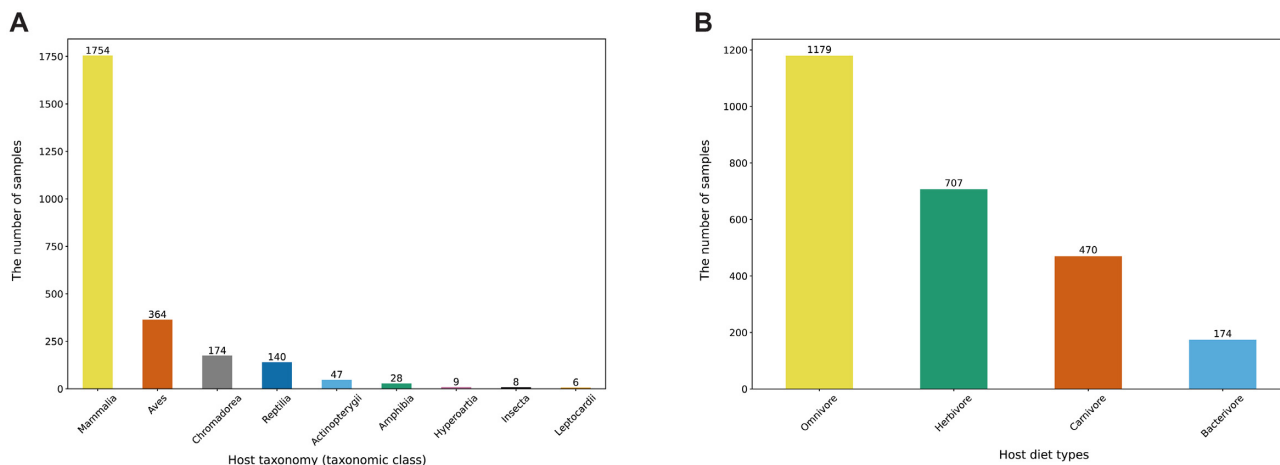
### Database statistics

Table 1 summarizes the statistics of AMDB. AMDB contains 2530 samples from 34 projects. A total of 139 375 ASVs were identified, corresponding to 81 669 682 reads. In the taxonomic analysis, 84.94% (69 367 504) reads were assigned to bacterial taxa, covering a total of 10 478 taxa. The total number of hosts in AMDB was 467 animal species, representing nine taxonomic classes (namely, 'Mammalia', 'Aves', 'Chromadorea', 'Reptilia', 'Actinopterygii', 'Amphibia', 'Hyperoartia', 'Insecta' and 'Leptocardii') and four trophic groups (namely, 'Omnivore', 'Herbivore', 'Carnivore' and 'Bacterivore'). The most abundant host taxonomic class was 'Mammalia,' which represented 69.33% (1754) of the samples, followed by 'Aves' and 'Chromadorea' which represented 14.39% (364) and 6.88% (174), respectively (Figure 2A). The most abundant trophic group was 'Omnivore,' which represented 46.60% (1179) of the samples, followed by 'Herbivore' and 'Carnivore' which represented 27.94% (707) and 18.58% (470) respectively (Figure 2B).

### Database content and usage

AMDB can be divided into four main parts, namely 'Taxa', 'Samples', 'Projects/Hosts' and 'Visualization'. 'Taxa' shows samples enriched with the bacterial taxon of interest. 'Samples' provides the gut microbiota composition of the sample of interest. 'Projects' and 'Hosts' give users summary information on the project and the host, respectively. 'Visualization' visually presents valuable information related to the relationship between the host and the gut microbiota.

'Taxa' allows users to search for the taxon of interest (Supplementary Figure S1). 'Taxa' provides taxon informa-



**Figure 2.** Bar plots showing the distribution of samples. (A) The number of samples is represented according to the host taxonomic classes. (B) The number of samples is represented according to the host diet types. Values are shown on the head of each bar. The bars are sorted in descending order by the values.

tion, including taxonomy and taxonomic rank. It also supports linking with the EzBioCloud (53), allowing for direct access to relevant information about the taxon. A list of the samples is provided, sorted according to the relative abundance of the taxon occurring within each sample. The AMDB allows the user to determine the relative abundance of the taxon based on host taxonomic ranks and host diet types using box plots. Each group name in the plot is followed by a frequency of occurrence, calculated as the number of samples containing the taxon divided by the total number of samples in the group. The viewing area of the plot can be adjusted by zooming in (dragging) or zooming out (double-clicking on the plot). Users can also click the camera icon to download the plot as a portable network graphics (.png) file.

‘Samples’ allows users to determine the gut microbiota composition of the sample of interest (Supplementary Figure S2). ‘Samples’ provides sample information, including the sample name, sampling site, accessions in the NCBI and information about the host and the respective analysis. Calculated alpha diversity indices are also displayed. A list of ASVs that make up the sample is provided. Taxonomic composition is displayed in a table and a multi-layered pie chart. The multi-layered pie chart has zooming capabilities, providing a more detailed view (double clicking on the node allows for zooming in, while zooming out is achieved by clicking on the summary pie charts present on the right-hand side of the chart). The ASV list and the taxonomic composition can be downloaded as comma-separated values (.csv) files.

Users are able to get summary information about a project and a host using ‘Projects’ (Supplementary Figure S3A) and ‘Hosts’ (Supplementary Figure S3B), respectively. ‘Projects’ provides information about the related paper, and ‘Host’ displays the taxonomy and diet type of the host. A complete list of samples related to both the project and the host is provided. Additionally, the alpha diversity indices from the samples are displayed as box plots. The average taxonomic composition from the samples is displayed in a table and a multi-layered pie chart. The sample list and

the average taxonomic composition can be downloaded as comma-separated values (.csv) files.

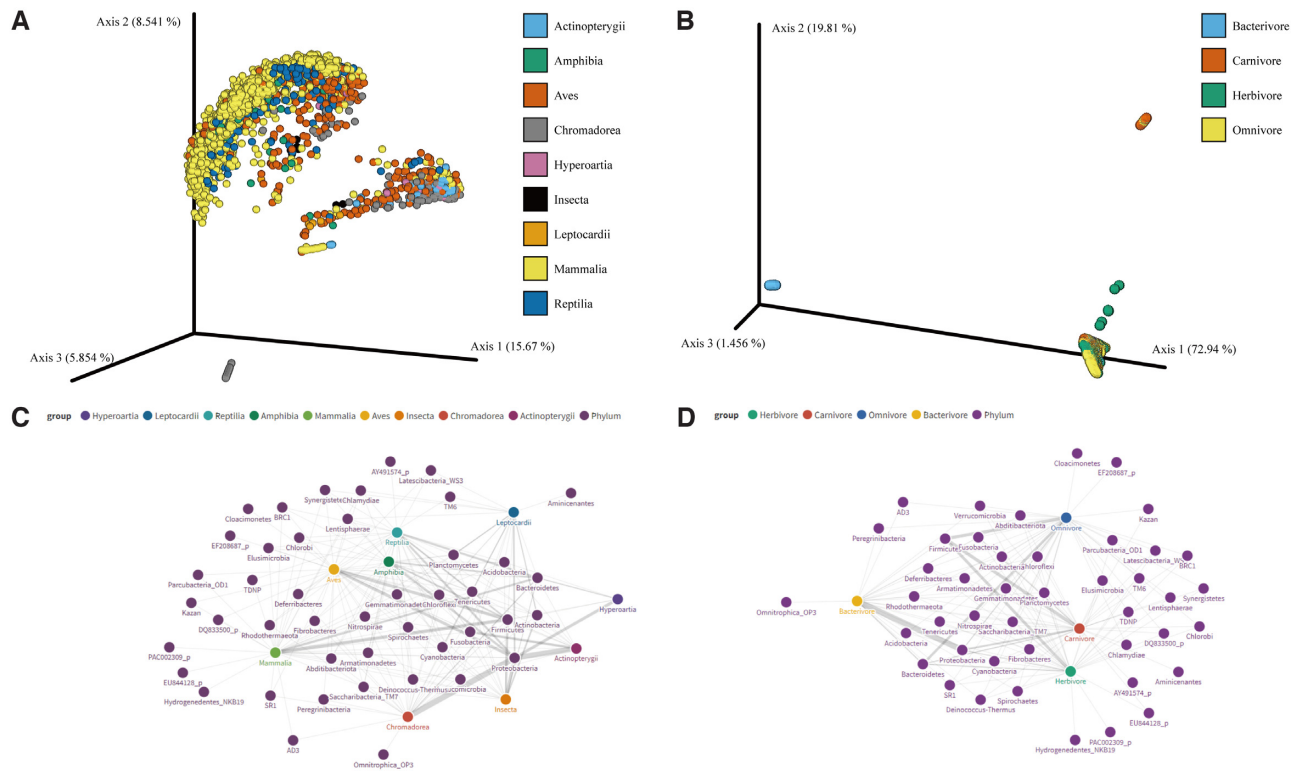
AMDB provides interactive visualizations, including PCoA plots and network graphs, to effectively describe the relationship between the samples and the respective gut microbiota on the ‘Visualization’ page. Via PCoA plots, users can assess the variations in phylogenetic structure among the samples based on unweighted UniFrac distances (Figure 3A) and weighted UniFrac distances (Figure 3B). Each point in the plot represents one sample and can be color-coded depending on the users’ choice. The plot is a draggable and zoomable 3D object, allowing users to view the plot from different perspectives. The plots can be downloaded as portable network graphics (.png) files, scalable vector graphics (.svg) files or QIIME 2 visualizations (.qzv) files (32). Network graphs show associations between the gut microbiota and host characteristics, including taxonomy (Figure 3C) and diet types (Figure 3D). Nodes in the network graphs represent taxa and host characteristics. The nodes can be moved and filtered out or in according to the users’ choice. Clicking on each node brings up a pop-up box with a link to ‘Taxa’ in a new tab. The line width of the network graph is relative to the average relative abundance of each taxon.

### Other functionalities

To better guide users, the ‘Help’ page provides an overview of AMDB with simple examples. Users can also propose candidate data for AMDB using the submission form on the ‘Contact’ page. Our team will manually check new user-submitted information, and AMDB will be updated on an ongoing basis.

## DISCUSSION

AMDB is a database for exploring the gut microbiota of various animal species. AMDB provides a search capability for the various components related to gut microbiota. For example, one may be interested in *Bilophila wadsworthia*,



**Figure 3.** AMDB visualizations. (A) PCoA plot shows variation in phylogenetic structure among 2530 samples based on unweighted UniFrac distances. Each point represents an individual sample, and samples are colored by the host taxonomic class. (B) PCoA plot shows variations in phylogenetic structure among 2530 samples based on weighted UniFrac distances. Each point represents an individual sample, and samples are colored by the host diet. (C) Associations between bacteria and host taxonomic classes are shown as a network graph. Each node represents an individual taxon (the phylum level) or a host taxonomic class. (D) Associations between bacteria and host diet types are shown as a network graph. Each node represents an individual taxon (the phylum level) or a host diet type. For network graphs, the line width is proportional to the average relative abundance of each taxon.

which is known to be related to animal-based diets in humans (56). The samples rich in this taxon can be identified in the search result of AMDB. Additionally, AMDB allows users to search for the sample based on metadata, including host taxonomy and diet types. The work from Youngblut *et al.* identified that hosts from the same species showed similar relative abundances of microbial phyla (7). This can be confirmed by comparing the microbial taxonomic compositions of samples taken from the same species. In addition, AMDB provides summary information about related projects and hosts. Users can thus compare the mouse information held within AMDB to the core microbiota of the mouse gut identified in multiple studies (57–59). Interactive visualizations are also available in AMDB. Host phylogeny and diet can explain variations in the gut microbiota (7), which can be confirmed using a PCoA plot within AMDB. The phylum Proteobacteria was identified as the dominant phylum in the samples from Actinopterygii (60), which can be identified using the network graph.

The number of available amplicon data in the NCBI SRA is continually increasing. AMDB will also be continuously updated to add additional data related to new and existing animal species. We will include new data collected by our team, as well as data based on the user-submitted information after manual curation.

Investigations into the relationship between gut microbiota and the host is a rapidly growing area of research

(61). AMDB is the first database enabling easier exploration of this relationship. AMDB comprehensively addresses the taxonomic composition of animal gut microbiota with manually curated metadata, thus assisting in providing a better understanding of the gut microbiota of animals.

## DATA AVAILABILITY

AMDB is freely available at <http://leb.snu.ac.kr/amdb>.

## SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

## ACKNOWLEDGEMENTS

We would like to thank the NCBI and all available projects for their valuable resources for AMDB.

## FUNDING

Korea Institute of Planning and Evaluation for Technology in Food, Agriculture, Forestry and Fisheries (iPET), funded by the Ministry of Agriculture, Food and Rural Affairs (MAFRA) [918013-04-4-SB010]; National Research Foundation of Korea [NRF-2014M3C9A3063541]. Funding for open access charge: Korea Institute of Planning and

Evaluation for Technology in Food, Agriculture, Forestry and Fisheries [918013-04-4-SB010].

*Conflict of interest statement.* None declared.

## REFERENCES

- Sommer, F. and Backhed, F. (2013) The gut microbiota—masters of host development and physiology. *Nat. Rev. Microbiol.*, **11**, 227–238.
- Krishnan, S., Alden, N. and Lee, K. (2015) Pathways and functions of gut microbiota metabolism impacting host physiology. *Curr. Opin. Biotechnol.*, **36**, 137–145.
- Colombo, B.M., Scalvenzi, T., Benlamara, S. and Pollet, N. (2015) Microbiota and mucosal immunity in amphibians. *Front. Immunol.*, **6**, 111.
- Fraune, S. and Bosch, T.C. (2010) Why bacteria matter in animal development and evolution. *Bioessays*, **32**, 571–580.
- Theis, K.R., Dheilly, N.M., Klassen, J.L., Brucker, R.M., Baines, J.F., Bosch, T.C., Cryan, J.F., Gilbert, S.F., Goodnight, C.J., Lloyd, E.A. *et al.* (2016) Getting the hologenome concept right: an eco-evolutionary framework for hosts and their microbiomes. *mSystems*, **1**, e00028-16.
- Ley, R.E., Hamady, M., Lozupone, C., Turnbaugh, P.J., Ramey, R.R., Bircher, J.S., Schlegel, M.L., Tucker, T.A., Schrenzel, M.D., Knight, R. *et al.* (2008) Evolution of mammals and their gut microbes. *Science*, **320**, 1647–1651.
- Youngblut, N.D., Reischer, G.H., Walters, W., Schuster, N., Walzer, C., Stalder, G., Ley, R.E. and Farnleitner, A.H. (2019) Host diet and evolutionary history explain different aspects of gut microbiome diversity among vertebrate clades. *Nat. Commun.*, **10**, 2200.
- Song, S.J., Sanders, J.G., Delsuc, F., Metcalf, J., Amato, K., Taylor, M.W., Mazel, F., Lutz, H.L., Winker, K., Graves, G.R. *et al.* (2020) Comparative analyses of vertebrate gut microbiomes reveal convergence between birds and bats. *mBio*, **11**, e02901-19.
- Grieneisen, L., Dasari, M., Gould, T.J., Bjork, J.R., Grenier, J.C., Yotova, V., Jansen, D., Gittel, N., Gordon, J.B., Learn, N.H. *et al.* (2021) Gut microbiome heritability is nearly universal but environmentally contingent. *Science*, **373**, 181–186.
- Mallott, E.K. and Amato, K.R. (2021) Host specificity of the gut microbiome. *Nat. Rev. Microbiol.*, **19**, 639–653.
- Brucker, R.M. and Bordenstein, S.R. (2013) The hologenomic basis of speciation: gut bacteria cause hybrid lethality in the genus *Nasonia*. *Science*, **341**, 667–669.
- Ochman, H., Worobey, M., Kuo, C.H., Ndjango, J.B., Peeters, M., Hahn, B.H. and Hugenholtz, P. (2010) Evolutionary relationships of wild hominids recapitulated by gut microbial communities. *PLoS Biol.*, **8**, e1000546.
- Brooks, A.W., Kohl, K.D., Brucker, R.M., van Opstal, E.J. and Bordenstein, S.R. (2016) Phylosymbiosis: relationships and functional effects of microbial communities across host evolutionary history. *PLoS Biol.*, **14**, e2000225.
- Amato, K.R., Sanders, J.G., Song, S.J., Nute, M., Metcalf, J.L., Thompson, L.R., Morton, J.T., Amir, A., McKenzie, V.J., Humphrey, G. *et al.* (2019) Evolutionary trends in host physiology outweigh dietary niche in structuring primate gut microbiomes. *ISME J*, **13**, 576–587.
- Muegge, B.D., Kuczynski, J., Knights, D., Clemente, J.C., Gonzalez, A., Fontana, L., Henrissat, B., Knight, R. and Gordon, J.I. (2011) Diet drives convergence in gut microbiome functions across mammalian phylogeny and within humans. *Science*, **332**, 970–974.
- Delsuc, F., Metcalf, J.L., Wegener Parfrey, L., Song, S.J., Gonzalez, A. and Knight, R. (2014) Convergence of gut microbiomes in myrmecophilous mammals. *Mol. Ecol.*, **23**, 1301–1317.
- McKenney, E.A., Maslanka, M., Rodrigo, A. and Yoder, A.D. (2018) Bamboo specialists from two mammalian orders (primates, carnivora) share a high number of low-abundance gut microbes. *Microb. Ecol.*, **76**, 272–284.
- Song, S.J., Sanders, J.G., Baldassarre, D.T., Chaves, J.A., Johnson, N.S., Piaggio, A.J., Stuckey, M.J., Novakova, E., Metcalf, J.L., Chomel, B.B. *et al.* (2019) Is there convergence of gut microbes in blood-feeding vertebrates? *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, **374**, 20180249.
- Tang, J., Wu, X., Mou, M., Wang, C., Wang, L., Li, F., Guo, M., Yin, J., Xie, W., Wang, X. *et al.* (2021) GIMICA: host genetic and immune factors shaping human microbiota. *Nucleic Acids Res.*, **49**, D715–D722.
- Wu, S., Sun, C., Li, Y., Wang, T., Jia, L., Lai, S., Yang, Y., Luo, P., Dai, D., Yang, Y.Q. *et al.* (2020) GMrepo: a database of curated and consistently annotated human gut metagenomes. *Nucleic Acids Res.*, **48**, D545–D553.
- Forster, S.C., Browne, H.P., Kumar, N., Hunt, M., Denise, H., Mitchell, A., Finn, R.D. and Lawley, T.D. (2016) HPMCD: the database of human microbial communities from metagenomic datasets and microbial reference genomes. *Nucleic Acids Res.*, **44**, D604–D609.
- Yang, J., Park, J., Park, S., Baek, I. and Chun, J. (2019) Introducing murine microbiome database (MMDB): a curated database with taxonomic profiling of the healthy mouse gastrointestinal microbiome. *Microorganisms*, **7**, 480.
- Lagkouvardos, I., Joseph, D., Kapfhammer, M., Giritli, S., Horn, M., Haller, D. and Clavel, T. (2016) IMNGS: a comprehensive open resource of processed 16S rRNA microbial profiles for ecology and diversity studies. *Sci. Rep.*, **6**, 33721.
- Mitchell, A.L., Almeida, A., Beracochea, M., Boland, M., Burgin, J., Cochrane, G., Crusoe, M.R., Kale, V., Potter, S.C., Richardson, L.J. *et al.* (2020) MGnify: the microbiome analysis resource in 2020. *Nucleic Acids Res.*, **48**, D570–D578.
- Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E.M., Kubal, M., Paczian, T., Rodriguez, A., Stevens, R., Wilke, A. *et al.* (2008) The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics*, **9**, 386.
- Gonzalez, A., Navas-Molina, J.A., Kosciok, T., McDonald, D., Vazquez-Baeza, Y., Ackermann, G., DeReus, J., Janssen, S., Swafford, A.D., Orchanian, S.B. *et al.* (2018) Qiita: rapid, web-enabled microbiome meta-analysis. *Nat. Methods*, **15**, 796–798.
- Jovel, J., Patterson, J., Wang, W., Hotte, N., O'Keefe, S., Mitchell, T., Perry, T., Kao, D., Mason, A.L., Madsen, K.L. *et al.* (2016) Characterization of the gut microbiome using 16S or shotgun metagenomics. *Front. Microbiol.*, **7**, 459.
- Soergel, D.A., Dey, N., Knight, R. and Brenner, S.E. (2012) Selection of primers for optimal taxonomic classification of environmental 16S rRNA gene sequences. *ISME J*, **6**, 1440–1444.
- Amir, A., McDonald, D., Navas-Molina, J.A., Kopylova, E., Morton, J.T., Zech Xu, Z., Kightley, E.P., Thompson, L.R., Hyde, E.R., Gonzalez, A. *et al.* (2017) Deblur rapidly resolves single-nucleotide community sequence patterns. *mSystems*, **2**, e00191-16.
- Kissling, W.D., Dalby, L., Flojgaard, C., Lenoir, J., Sandel, B., Sandom, C., Trojelsgaard, K. and Svenning, J.C. (2014) Establishing macroecological trait datasets: digitalization, extrapolation, and validation of diet preferences in terrestrial mammals worldwide. *Ecol. Evol.*, **4**, 2913–2930.
- Wilman, H., Belmaker, J., Simpson, J., de la Rosa, C., Rivadeneira, M.M. and Jetz, W. (2014) EltonTraits 1.0: species-level foraging attributes of the world's birds and mammals: Ecological Archives E095-178. *Ecology*, **95**, 2027–2027.
- Bolyen, E., Rideout, J.R., Dillon, M.R., Bokulich, N.A., Abnet, C.C., Al-Ghalith, G.A., Alexander, H., Alm, E.J., Arumugam, M., Asnicar, F. *et al.* (2019) Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat. Biotechnol.*, **37**, 852–857.
- Rognes, T., Flouri, T., Nichols, B., Quince, C. and Mahe, F. (2016) VSEARCH: a versatile open source tool for metagenomics. *PeerJ*, **4**, e2584.
- Bokulich, N.A., Subramanian, S., Faith, J.J., Gevers, D., Gordon, J.I., Knight, R., Mills, D.A. and Caporaso, J.G. (2013) Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. *Nat. Methods*, **10**, 57–59.
- Callahan, B.J., McMurdie, P.J. and Holmes, S.P. (2017) Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *ISME J*, **11**, 2639–2643.
- McDonald, D., Clemente, J.C., Kuczynski, J., Rideout, J.R., Stombaugh, J., Wendel, D., Wilke, A., Huse, S., Hufnagle, J., Meyer, F. *et al.* (2012) The Biological Observation Matrix (BIOM) format or: how I learned to stop worrying and love the ome-ome. *Gigascience*, **1**, 7.
- Weiss, S., Xu, Z.Z., Peddada, S., Amir, A., Bittinger, K., Gonzalez, A., Lozupone, C., Zaneveld, J.R., Vazquez-Baeza, Y., Birmingham, A. *et al.* (2017) Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome*, **5**, 27.

38. Beule, L. and Karlovsky, P. (2020) Improved normalization of species count data in ecology by scaling with ranked subsampling (SRS): application to microbial communities. *PeerJ*, **8**, e9593.
39. Shannon, C.E. (1948) A mathematical theory of communication. *Bell Syst. Tech. J.*, **27**, 379–423.
40. McKinney, W. (2010) In: *Proceedings of the 9th Python in Science Conference*. Austin, TX, Vol. **445**, pp. 51–56.
41. Katoh, K. and Standley, D.M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.*, **30**, 772–780.
42. Price, M.N., Dehal, P.S. and Arkin, A.P. (2010) FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS One*, **5**, e9490.
43. Lozupone, C. and Knight, R. (2005) UniFrac: a new phylogenetic method for comparing microbial communities. *Appl. Environ. Microbiol.*, **71**, 8228–8235.
44. Lozupone, C.A., Hamady, M., Kelley, S.T. and Knight, R. (2007) Quantitative and qualitative beta diversity measures lead to different insights into factors that structure microbial communities. *Appl. Environ. Microbiol.*, **73**, 1576–1585.
45. Hamady, M., Lozupone, C. and Knight, R. (2010) Fast UniFrac: facilitating high-throughput phylogenetic analyses of microbial communities including analysis of pyrosequencing and PhyloChip data. *ISME J*, **4**, 17–27.
46. Lozupone, C., Lladser, M.E., Knights, D., Stombaugh, J. and Knight, R. (2011) UniFrac: an effective distance metric for microbial community comparison. *ISME J*, **5**, 169–172.
47. McDonald, D., Vázquez-Baeza, Y., Koslicki, D., McClelland, J., Reeve, N., Xu, Z., Gonzalez, A. and Knight, R. (2018) Striped UniFrac: enabling microbiome analysis at unprecedented scale. *Nat. Methods*, **15**, 847–848.
48. Legendre, P. and Legendre, L. (2012) In: *Numerical Ecology*. 3rd edn. Elsevier.
49. Halko, N., Martinsson, P.-G., Shkolnisky, Y. and Tygert, M. (2011) An algorithm for the principal component analysis of large data sets. *SIAM J. Sci. Comput.*, **33**, 2580–2594.
50. Vázquez-Baeza, Y., Pirrung, M., Gonzalez, A. and Knight, R. (2013) EMPeror: a tool for visualizing high-throughput microbial community data. *Gigascience*, **2**, 16.
51. Vázquez-Baeza, Y., Gonzalez, A., Smarr, L., McDonald, D., Morton, J.T., Navas-Molina, J.A. and Knight, R. (2017) Bringing the dynamic microbiome to life with animations. *Cell Host Microbe*, **21**, 7–10.
52. Bokulich, N.A., Kaehler, B.D., Rideout, J.R., Dillon, M., Bolyen, E., Knight, R., Huttley, G.A. and Gregory Caporaso, J. (2018) Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. *Microbiome*, **6**, 90.
53. Yoon, S.H., Ha, S.M., Kwon, S., Lim, J., Kim, Y., Seo, H. and Chun, J. (2017) Introducing EzBioCloud: a taxonomically united database of 16S rRNA gene sequences and whole-genome assemblies. *Int. J. Syst. Evol. Microbiol.*, **67**, 1613–1617.
54. Ondov, B.D., Bergman, N.H. and Phillippy, A.M. (2011) Interactive metagenomic visualization in a Web browser. *BMC Bioinformatics*, **12**, 385.
55. Skoufos, G., Kardaras, F.S., Alexiou, A., Kavakiotis, I., Lambropoulou, A., Kotsira, V., Tastsoglou, S. and Hatzigeorgiou, A.G. (2021) Peryton: a manual collection of experimentally supported microbe-disease associations. *Nucleic Acids Res.*, **49**, D1328–D1333.
56. David, L.A., Maurice, C.F., Carmody, R.N., Gootenberg, D.B., Button, J.E., Wolfe, B.E., Ling, A.V., Devlin, A.S., Varma, Y., Fischbach, M.A. et al. (2014) Diet rapidly and reproducibly alters the human gut microbiome. *Nature*, **505**, 559–563.
57. Xiao, L., Feng, Q., Liang, S., Sonne, S.B., Xia, Z., Qiu, X., Li, X., Long, H., Zhang, J., Zhang, D. et al. (2015) A catalog of the mouse gut metagenome. *Nat. Biotechnol.*, **33**, 1103–1108.
58. Wang, J., Lang, T., Shen, J., Dai, J., Tian, L. and Wang, X. (2019) Core gut bacteria analysis of healthy mice. *Front. Microbiol.*, **10**, 887.
59. Yang, J. and Chun, J. (2021) Taxonomic composition and variation in the gut microbiota of laboratory mice. *Mamm. Genome*, **32**, 297–310.
60. Sullam, K.E., Essinger, S.D., Lozupone, C.A., O'Connor, M.P., Rosen, G.L., Knight, R., Kilham, S.S. and Russell, J.A. (2012) Environmental and ecological factors that shape the gut bacterial communities of fish: a meta-analysis. *Mol. Ecol.*, **21**, 3363–3378.
61. Lim, S.J. and Bordenstein, S.R. (2020) An introduction to phyllosymbiosis. *Proc Biol Sci*, **287**, 20192900.