NAR Genomics and Bioinformatics, 2025, **7**, lqae191
https://doi.org/10.1093/nargab/lqae191
Advance access publication date: 8 January 2025
Pipelines and Workflows for Biological Data Analysis

OXFORD

stana: an R package for metagenotyping analysis and interactive application based on clinical data

Noriaki Sato [1], Kotoe Katayama[2], Daichi Miyaoka[3], Miho Uematsu[3,4], Ayumu Saito[1], Kosuke Fujimoto[3,4], Satoshi Uematsu[3,4] and Seiya Imoto [1,2,*]

[1]Division of Health Medical Intelligence, Human Genome Center, The Institute of Medical Science, The University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan
[2]Laboratory of Sequence Analysis, Human Genome Center, The Institute of Medical Science, The University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan
[3]Department of Immunology and Genomics, Graduate School of Medicine, Osaka Metropolitan University, 1-4-3 Asahi-machi, Abeno-ku, Osaka 545-8585, Japan
[4]Division of Metagenome Medicine, Human Genome Center, The Institute of Medical Science, The University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan

[*]To whom correspondence should be addressed. Tel: +81 3 5449 5615; Fax: +81 3 5449 5442; Email: imoto@hgc.jp

Abstract

Metagenotyping of metagenomic data has recently attracted increasing attention as it resolves intraspecies diversity by identifying single nucleotide variants. Furthermore, gene copy number analysis within species provides a deeper understanding of metabolic functions in microbial communities. However, a platform for examining metagenotyping results based on relevant grouping data is lacking. Here, we have developed the R package, stana, for the processing and analysis of metagenotyping results. The package consists of modules for preprocessing, statistical analysis, functional analysis and visualization. An interactive analysis environment for exploring the metagenotyping results was also developed and publicly released with over 1000 publicly available metagenome samples related to human diseases. Three examples exploring the relationship between the metagenotypes of the gut microbiome and human diseases are presented—end-stage renal disease, Crohn's disease and Parkinson's disease. The results suggest that stana facilitated the confirmation of the original study's findings and the generation of a new hypothesis. The GitHub repository for the package is available at https://github.com/noriakis/stana.

Introduction

With recent advancements in high-throughput sequencing technologies and the development of metagenomic analysis methodologies, comprehensive analyses of various bacterial communities in the human body and environment have become possible. Metagenomic analyses reveal mechanisms of microbial metabolism in human diseases such as neurodegenerative diseases, inflammatory bowel diseases (IBDs) and non-communicable diseases (NCDs) (1). In addition to species-level metagenomic analysis, previous studies have indicated the importance of bacterial strains existing within a species, such as those related to differing pathogenicity and drug resistance (2). Thus, attention has been focused on analyzing intraspecies diversity information contained in metagenomic data (3).

Computational pipelines have been developed to illuminate genetic diversity within species by aligning metagenomic reads to curated bacterial reference genome databases and identifying single nucleotide variants (SNVs) and gene copy number differences among species, referred to as metagenotyping (4,5). The metagenotyping data lead to unique analyses, such as source tracking using the SNV matrix and strain-level abundance estimation.

However, limited software is available for performing integrated analyses after the acquisition of metagenotyping results. This gap is particularly conspicuous in the context of group-based comparisons, which are common in clinical studies. Therefore, to address this gap, we have developed an R package, stana, to integrate and analyze data derived from metagenotyping seamlessly. This package facilitates the essential preprocessing, functional analysis, statistical examination and visualization of outcomes, with a specific emphasis on comparing the results between the groups of interest.

Also, the publicly deposited metadata coupled with metagenomic data not related to the study's main findings is useful information that needs to be examined as they could have the possibility of revealing insights not related to the study's original aims. Thus, an interactive application to analyze such intraspecies diversity across a group of disease conditions is described, with publicly available metagenomic data being profiled. The application has been published publicly to allow the interactive examination of metagenotyping analysis results from over 1000 samples of the intestinal microbiome in publicly available metagenomic datasets related to human diseases.

Received: July 18, 2024. **Revised:** October 22, 2024. **Editorial Decision:** December 16, 2024. **Accepted:** January 5, 2025

Materials and methods

Package design and functionality

The package has the core functions for loading the output of the metagenotyping software, mainly allelic counts per genome position, allele frequency tables and gene copy number variant tables. Examples are the generated data from MIDAS and MIDAS2, which include allele frequency tables and gene copy number variant profiles per sample by default and are suitable for metabolic functional analysis (4,6,7). The data generated from the other pipelines that can be imported include inStrain and metaSNV (5,8). For data produced from the other software, the functions for importing the data representing allelic counts per position, the allele frequency or the gene copy number variant matrix manually are prepared for performing the downstream analysis. The loading function for additional metagenotyping pipelines will be supported.

From the loaded profile, users can inspect basic information such as species coverage, allele frequency, major and minor allele distribution, and the number of samples profiled for the species. The typical workflow of the library is first importing metagenotyping data to the S4 class object of the R environment and then applying filtering and analytical functions, such as species- and sample-wise filters. Subsequently, using the filtered stana object, the users can search for and identify species that differ in their allele frequency or gene copy number across groups and infer the functional implications of these differences based on the gene information, whose details are described in the subsequent sections. The GitHub repository for this package is available at https://github.com/noriakis/stana with detailed documentation (https://noriakis.github.io/software/stana).

Analysis description

The package implements multiple options for investigating the intraspecies diversity. One option is inferring the consensus sequence alignment of the species within samples based on the allele frequency using an approach similar to that of MIDAS (6). The approach determines alleles for each position based on allele frequency, and the users can preset the interesting positions obtained from the other analyses in the package. This option is useful for phylogenetic tree inference based on multiple sequence alignment. Another option is using non-negative matrix factorization (NMF) applied to the data stored in the stana object. NMF approximates the indicated feature matrix, e.g. gene copy number matrix, by the product of two non-negative matrices, corresponding to feature mapping per factor and the sample profiles of factors. The rank of the factors is a parameter that should be specified, corresponding to the number of possible factors (i.e. subspecies or strain) within a species. There are options for estimating the rank, and the resulting statistics such as the estimated profiles of each factor are described when using the function, which helps to choose the rank (9–11). Also, the package statistically compares the intraspecies diversity between specified groups, e.g. clinical or environmental conditions, by performing permutational multivariate analysis of variance (PERMANOVA) on the distance matrices calculated from various imported matrices or the calculated trees (http://cran.r-project.org/package=vegan). With these functions combined, the users can identify possible candidates that differ in their intraspecies diversity between conditions.

The gene copy numbers can be compared using exact Wilcoxon rank-sum tests. Also, a function that identifies important features for distinguishing groups was prepared, which is useful for marker gene detection between groups per species based on gene copy number variants. The package supports the summarization of results of orthology or gene family assignment software, such as eggNOG-mapper v2, and functional annotation provided by the PATRIC server, which is used as the default in MIDAS, or the manual annotation data (12). This enables the functional investigation of the identified genes.

For differential set analysis (DSA), gene set enrichment analysis (GSEA) or over-representation analysis (ORA), with the function GSEA in clusterProfiler and enrichKO in MicrobiomeProfiler, were used, respectively (13,14). The input for GSEA can be chosen from various statistics, such as $\log_2$ fold changes and moderated *t*-statistics. These results can be combined with network-based analysis using the Kyoto Encyclopedia of Genes and Genomes (KEGG) PATHWAY data, helping to rank the components in the graph based on the network topology and statistics.

Visualization

The analysis results can be visualized in various ways, such as a comparison of gene copy numbers across groups, coverage of the species per sample, and heatmaps depicting gene copy numbers with functional annotation by simplifyEnrichment (15), statistics of each covariate from the calculated models per species and the functional difference table identified by DSA per species. For the KEGG ORTHOLOGY (KO) assignment, visualization of the corresponding KOs in the KEGG PATHWAY is possible using the R library ggkegg (16).

Constructing the interactive application for inspecting intraspecies diversity

Users can directly export metagenotyped datasets for inspection in the application using the package and host the application to share findings with the community, facilitating data sharing. A publicly available dataset was compiled from the gut microbiome beforehand, and the interactive application with the precalculated datasets was published at the following URL: https://metagenotype.hgc.jp. The details for the construction of the interactive application are described in Supplementary Text S1.

Statistical analysis

The *P*-values or false discovery rate-controlled *P*-values (denoted as *q*) determined by the Benjamini–Hochberg procedure below 0.05 were considered statistically significant in the analysis ($q < 0.05$). The tree visualization throughout the package and the manuscript was performed using ggtree and ggtreeExtra (17,18).

Results

An overview of the input and analysis modules of the package stana is shown in Figure 1A. The overview of the interactive application using stana is shown in Figure 1B. The publicly available datasets processed and published in the application are listed in Supplementary Table S1 along with the obtained raw sample count. The table comparing the features with the other software is presented in Supplementary Table S2. Three

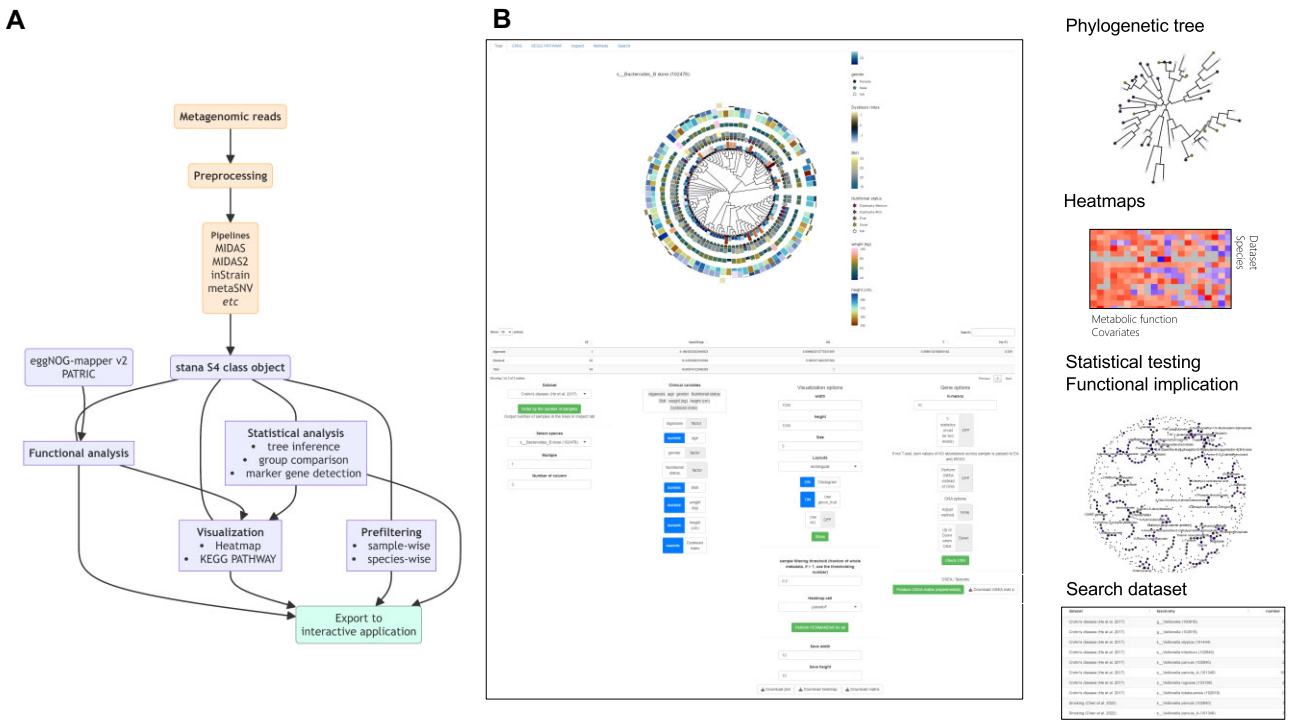

Figure 1. The overview of the analytic workflow of the R package, stana, and its interactive application. (**A**) The metagenomic reads are preprocessed and metagenotyped, and analyzed by stana using various functions, and subsequently can be exported to the interactive interface. The figure is rendered by Mermaid.js. (**B**) The overview of the interactive application using stana and publicly available datasets. The application can be directly exported from the R environment. The users can interactively perform analyses such as phylogenetic tree visualization, statistical analysis between groups, functional inference and visualization.

application examples of the assessment of intraspecies diversity in a study exploring the gut microbiome of patients with end-stage renal disease (ESRD), Crohn's disease (CD) and Parkinson's disease (PD) using a library and published interactive applications will be showcased (1,19,20).

Analysis of the ESRD dataset

First, the distances based on the phylogenetic tree inferred from the consensus multiple sequence alignment using the allele frequency table were compared between the groups of healthy controls (HC), patients with chronic kidney disease (CKD) and hemodialysis patients (HD) in the ESRD dataset for the analysis of the existence of intraspecies diversity differences. After testing for the species with a profiled sample number above 20% of the total sample number using PERMANOVA, five species, *Bacteroides_B dorei*, *Bacteroides fragilis*, *Faecalibacterium prausnitzii_G*, *Blautia_A wexlerae* and *Faecalicatena gnavus*, were statistically significant ($q < 0.05$). The cladograms of these five species with the output of the statistical values are shown in Figure 2A as the output of the interactive application.

Subsequently, the profiles of the factors calculated by the NMF are plotted between groups for *F. gnavus*, which indicates differences in the distribution of the inferred factors (Figure 2B). The differences correlated with the stage of the clinical conditions in the order of HC, CKD and HD in this analysis. The KO copy number matrix was used for NMF with a rank number of 2, which was derived from calculating the mean squared errors with cross-validation. As there is a possibility that multiple factors exist within the species, the network of cysteine and methionine metabolism pathway is plotted with the node colored by the $\log_2$ fold change between the profile of factors 2 and 1 (Figure 2C). The enzymes involved in the *S*-adenosyl-L-methionine cycle are found to be more abundant in factor 2, such as adenosylhomocysteine nucleosidase and *S*-adenosyl-L-methionine synthetase. The factor was more abundant in the CKD or HD samples over healthy control samples, which indicates that the corresponding metabolic functions could be more active in CKD or HD samples.

To further examine this observation, the relationship between tree-based distance, disease status and KO copy number across the cysteine and methionine pathway was tested using PERMANOVA, and the results of the KO copy number were statistically significant (pseudo-$F = 2.69$, $P = 0.045$). This observation could be related to the published literature that the serum *S*-adenosyl-L-methionine level is upregulated in ESRD patients (21), and possible responsibility by the intraspecies diversity of the species was suggested. The differences in intraspecies diversity and their functional implications in relation to kidney function decline were revealed using this package.

Analysis of the CD cohort

A dataset exploring the functional implications of the gut microbiome in patients with CD was investigated for the relationship between intraspecies diversity and metabolic functions. The relationship between the category of diagnosis published in the same manuscript and the intraspecies diversity of the gut microbiome was investigated and the gene copy numbers in *Bacteroides_B dorei*, which had the largest number of samples in the phylogenetic trees of the profiled species, were compared.

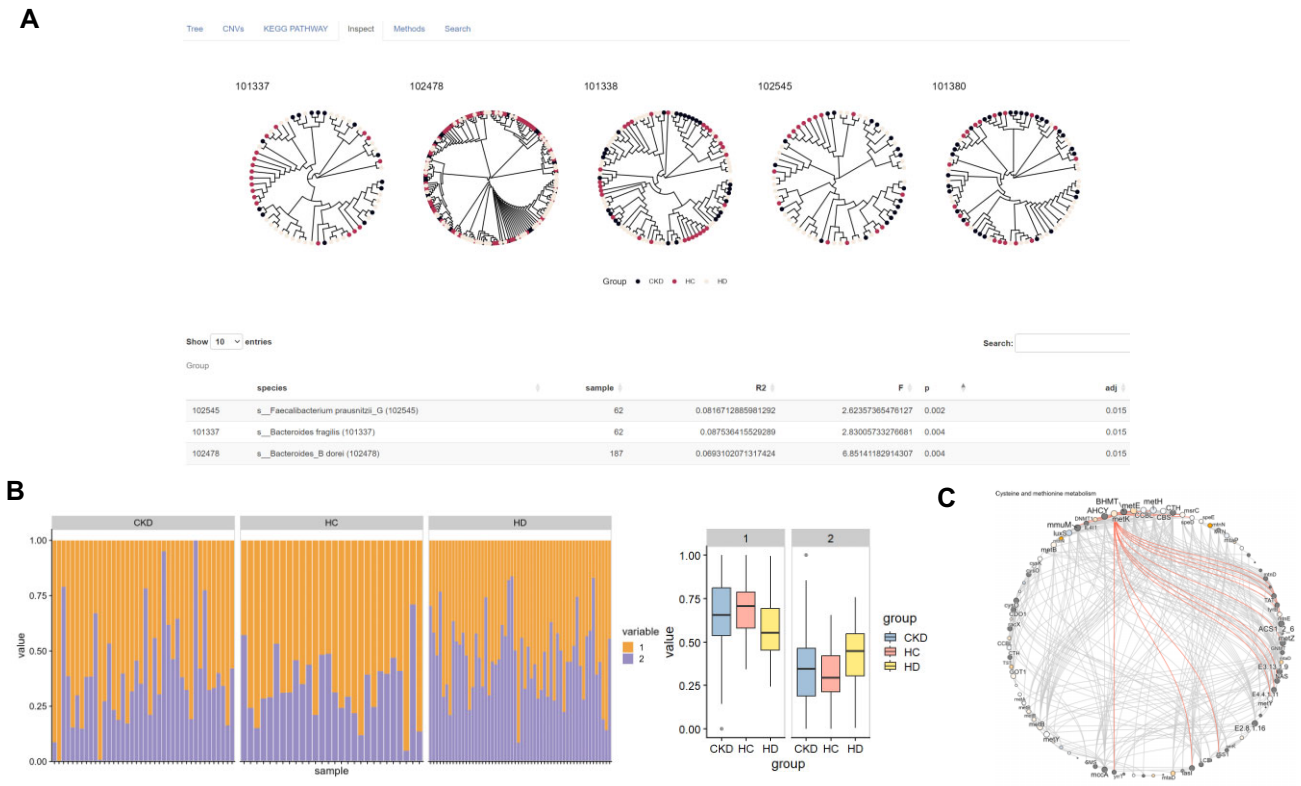

Figure 2. The analysis details of the gut microbiome of ESRD patients. An interactive application example of an ESRD dataset and the analytic example using NMF of the dataset using the package. (**A**) The significant results using the tree-based distance obtained from PERMANOVA are listed in the interactive application. (**B**) The KO copy number of the species was analyzed by the NMF and the profiles of the factors are plotted by the stacked bar plot and the box plot. (**C**) The network representation of the KEGG ORTHOLOGY involved in the cysteine and methionine metabolism. The node color indicates the $\log_2$ fold changes between the profile of factors 2 and 1. The node size indicates the degree of the nodes. The highlighted edges are those leading to metK (*S*-adenosylmethionine synthetase). KO, KEGG ORTHOLOGY; ESRD, end-stage renal disease; NMF, non-negative matrix factorization; PERMANOVA, permutational multivariate analysis of variance.

Based on the original grouping by disease diagnosis, labeled CD and Healthy, exact Wilcoxon rank-sum tests were performed on KO copy number, and the ORA on KEGG PATHWAY was performed on the gene families with higher copy numbers in CD. The lipopolysaccharide (LPS) biosynthesis was identified as statistically significant ($q < 0.05$) in the KO set. The scheme of the corresponding pathways colored by the moderated *t*-statistics between the groups is shown in Figure 3A. The *B. dorei* present in CD patients had high copy numbers of genes such as *lpxA*, which is involved in lipid A biosynthesis. Subsequently, a dendrogram was constructed using the distance matrix calculated from the gene copy number matrix subset to that related to the LPS biosynthesis pathway, as shown in Figure 3B. The trees exhibited a cluster of samples from the controls and CD patients, suggesting the existence of different functions related to the pathway. The species *B. dorei* in the gut microbiome has been reported to play a role in atherosclerosis through the reduction of LPS biosynthesis (22), and we suggested the possible role of functional differences of the corresponding species.

Analysis of the PD cohort

Intraspecies diversity of the species within the dataset exploring the gut metagenomics of patients with PD was investigated. The original manuscript reported various functional alterations in metabolic pathways.

To investigate which species might have functional differences, the unsupervised analysis can be performed by stana. The relationship between the covariates and tree-based distances across all profiled species filtered by the number of samples in the tree is evaluated. Also, GSEA was performed for all the species based on the category of the disease status, and the NES values were obtained per species. The resulting combined heatmap is shown in Figure 3C. The results suggest that in the genotype of *Streptococcus salivarius*, patients with PD had reduced gene copy numbers related to fructose and mannose metabolism pathways ($q < 0.05$), which was presumed to be related to the observation that the mannan and fructonate metabolism was reduced in patients with PD. Further, *Agathobacter rectalis*, which is related to producing short-chain fatty acids (SCFAs) (23), had the enrichment in the fatty acid metabolism and biosynthesis pathway, suggesting the possible role of intraspecies diversity of the species in PD as the SCFA level is reported to be altered in PD (24). These investigative results suggest that the package aids in the replication of the original findings in another resolution as well as generating new hypotheses.

Finally, these calculated data from multiple datasets can be merged using the common identifiers. Figure 3D shows a heatmap of the NES obtained from GSEA across two diseases, CD and PD. The gene contents and metabolic functional differences of each species across multiple datasets can be understood with plots. This enables users to inspect similarities

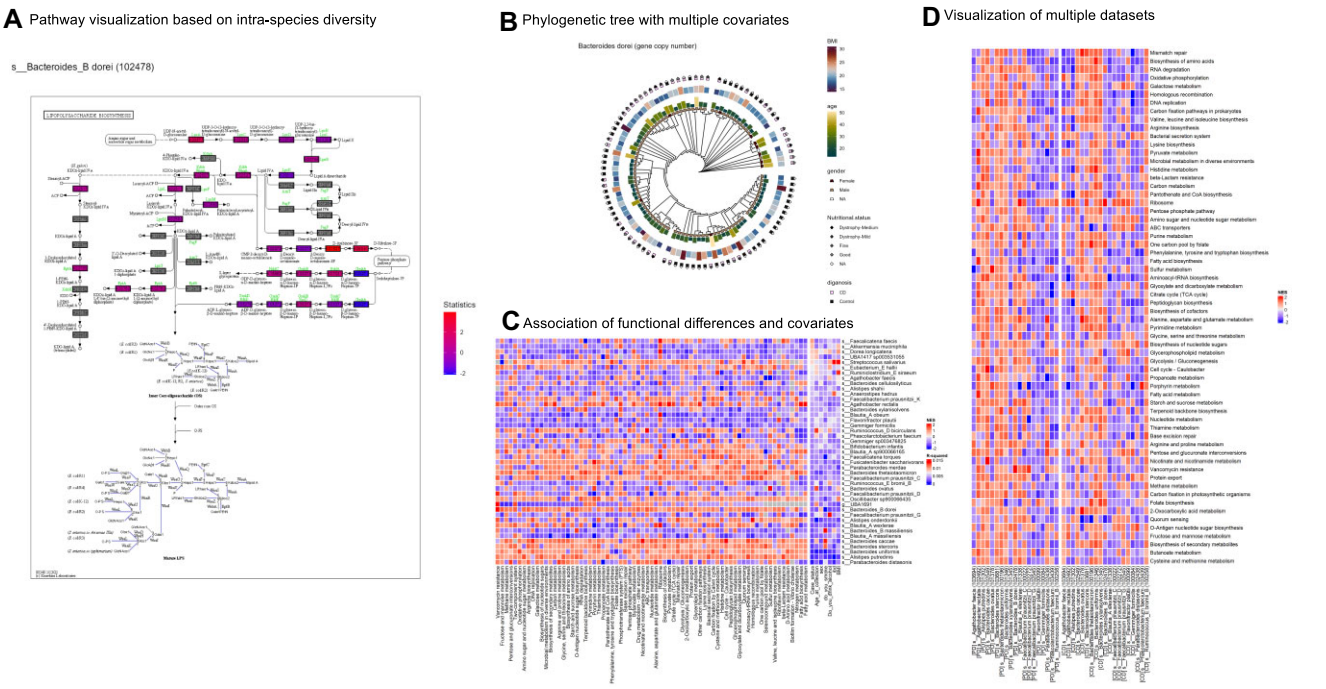

Figure 3. Analysis details of the gut microbiome of the patients with CD and PD. (**A**) The LPS biosynthesis pathway in *Bacteroides_B dorei* colored by calculated statistics between groups is shown in the CD cohort. The node indicates KO and the image was rendered internally using ggkegg. (**B**) The association between the cladogram inferred from the copy numbers of the genes related to the LPS pathway and covariate visualization for species. (**C**) Heatmap across the species demonstrating the association between intraspecies diversity and deposited metadata. The row indicates the species, the column indicates the covariates or biological pathways and the cell color indicates *R*-squared values calculated based on PERMANOVA or normalized enrichment score (NES). (**D**) NES matrix visualized across the species and the CD and PD datasets. The prefix before the species name refers to the dataset. KEGG, Kyoto Encyclopedia of Genes and Genomes; LPS, lipopolysaccharide, PD, Parkinson's disease; CD, Crohn's disease; PERMANOVA, permutational multivariate analysis of variance.

and dissimilarities of intraspecies diversity of multiple disease conditions, based on the profiled metagenotyping results.

Discussion

Here, we have described the R package for analyzing the intraspecies diversity of the microbiome and the interactive application environments for assessing the intraspecies diversity constructed using the package. The package aims to compare intraspecies diversity between groups, such as clinical conditions, and we developed an interactive application for exploring intraspecies diversity across publicly available datasets investigating human diseases and the user's datasets. This study presented three analyses of the datasets using the package and the application, highlighting that this package facilitates a confirmation and a deeper understanding of the published results of metagenomic studies, as well as new hypothesis generation.

Metagenotyping refers to genotyping across species using shotgun metagenomic sequences. Multiple computational software programs have been developed to profile metagenotypes. Intraspecies diversity has been reported to play a role in shaping clinical conditions such as pathogenicity differences. The reanalysis and re-evaluation of publicly deposited data using the proposed package could be useful for an in-depth understanding of metagenomic studies, hypothesis generation and confirmation of the results. A convenient and detailed analysis of this diversity is possible using this package and its published applications.

The definition of strain differs across studies (3); for instance, inStrain uses population average nucleotide identity thresholding to define strain, and the other software performs problem solving of strain deconvolution, like StrainFacts based on a generative model (25). Some papers assessed the distances calculated from the shared polymorphic sites or the distance based on allele frequencies to assess intraspecies diversity (26). When a PERMANOVA was used to compare the distance calculated from allele frequency to determine the difference in intraspecies diversity between groups, it may not align with the other criteria defining the strain. Although the package provides the interface for loading and analyzing all species inside the dataset, the selection of the species to be investigated could be relied upon by the ordering of statistical values, profiled sample size or prior knowledge.

Metagenotyping data can be processed in a typical computational environment when analyzing single species profiled in 100 samples using around 100 MB of the memory. However, when analyzing the data for all the profiled species in microbiome, such as summarizing the gene copy number profiles into functional gene categories, it may consume significant memory and time.

The strength lies in its design to handle and analyze SNV and gene copy number variant tables for multiple species, including the capability of performing multiple analyses on the same S4 class object, allowing for more customized comparisons between groups. There are several limitations to this library and its application. The package lacks the ability to perform the analysis like gene ortholog assignment and characterization of each SNV information. Not all pipelines

produce profiles, such as gene copy numbers, and the functions for these pipelines are limited. For complex modeling tasks often conducted in metagenomic studies, using the interactive application alone is not appropriate, and the package serves as a bridge to importing the metagenotyping data and its combination with other packages or functions for complex statistical analysis. Also, the findings of clinical diseases obtained in the study must be taken care of with caution as the effect of within-species diversity could be small, and should be supported by experimental validation. The future development goal of the package is to speed up the computation and analysis by implementing sparse class objects for handling the imported data. Also, the implementation of the other matrix deconvolution methods and comparison of metagenotyped data across different databases are deemed important.

The presented R package and the interactive analytic environment were shown to aid in understanding the role of gut microbiomes by assessing intraspecies diversity in diseases such as IBDs and NCDs. The package provides an environment to facilitate the understanding of complex intraspecies diversity information contained in metagenomic studies through the class object designed specifically for metagenotyping data and various functions for assessing diversity and functionality. Furthermore, the package and application could become important as the number of metagenomic publications increases, enabling the analyses of metagenomic data in future studies, and the published dataset can be used as a reference for studies that intend to explore intraspecies diversity in various diseases.

Data availability

All data in this study are included in this article or are accessible from the following URLs: https://github.com/noriakis/stana and https://doi.org/10.5281/zenodo.14558008. All sequencing data available in the published application were downloaded from the NCBI Sequence Read Archive or the National Microbiology Data Center, and the accession numbers are listed in Supplementary Table S1.

Supplementary data

Supplementary Data are available at NARGAB Online.

Acknowledgements

The authors thank Editage (https://editage.com) for the English language editing.

Funding

Japan Science and Technology Agency [JPMJCE1302, in part], Japan Agency for Medical Research and Development (AMED) [JP21ae0121040, in part].

Conflict of interest statement

None declared.

References

1. Wallen,Z.D., Demirkan,A., Twa,G., Cohen,G., Dean,M.N., Standaert,D.G., Sampson,T.R. and Payami,H. (2022) Metagenomics of Parkinson's disease implicates the gut microbiome in multiple disease mechanisms. *Nat. Commun.*, **13**, 6958.
2. Loman,N.J., Constantinidou,C., Christner,M., Rohde,H., Chan,J.Z.-M., Quick,J., Weir,J.C., Quince,C., Smith,G.P., Betley,J.R., *et al.* (2013) A culture-independent sequence-based metagenomics approach to the investigation of an outbreak of Shiga-toxigenic *Escherichia coli* O104:H4. *JAMA*, **309**, 1502–1510.
3. Van Rossum,T., Ferretti,P., Maistrenko,O.M. and Bork,P. (2020) Diversity within species: interpreting strains in microbiomes. *Nat. Rev. Microbiol.*, **18**, 491–506.
4. Zhao,C., Dimitrov,B., Goldman,M., Nayfach,S. and Pollard,K.S. (2023) MIDAS2: metagenomic intra-species diversity analysis system. *Bioinformatics*, **39**, btac713.
5. Olm,M.R., Crits-Christoph,A., Bouma-Gregson,K., Firek,B.A., Morowitz,M.J. and Banfield,J.F. (2021) inStrain profiles population microdiversity from metagenomic data and sensitively detects shared microbial strains. *Nat. Biotechnol.*, **39**, 727–736.
6. Nayfach,S., Rodriguez-Mueller,B., Garud,N. and Pollard,K.S. (2016) An integrated metagenomics pipeline for strain profiling reveals novel patterns of bacterial transmission and biogeography. *Genome Res.*, **26**, 1612–1625.
7. Smith,B.J., Zhao,C., Dubinkina,V., Jin,X., Moltzau-Anderson,J. and Pollard,K.S. (2024) Accurate estimation of intraspecific microbial gene content variation in metagenomic data with MIDAS v3 and StrainPGC. bioRxiv doi: https://doi.org/10.1101/2024.04.10.588779, 10 April 2024, preprint: not peer reviewed.
8. Van Rossum,T., Costea,P.I., Paoli,L., Alves,R., Thielemann,R., Sunagawa,S. and Bork,P. (2021) metaSNV v2: detection of SNVs and subspecies in prokaryotic metagenomes. *Bioinformatics*, **38**, 1162–1164.
9. Gaujoux,R. and Seoighe,C. (2010) A flexible R package for nonnegative matrix factorization. *BMC Bioinformatics*, **11**, 367.
10. Cai,Y., Gu,H. and Kenney,T. (2023) Rank selection for non-negative matrix factorization. *Stat. Med.*, **42**, 5676–5693.
11. Lin,X. and Boutros,P.C. (2020) Optimization and expansion of non-negative matrix factorization. *BMC Bioinformatics*, **21**, 7.
12. Cantalapiedra,C.P., Hernández-Plaza,A., Letunic,I., Bork,P. and Huerta-Cepas,J. (2021) eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Mol. Biol. Evol.*, **38**, 5825–5829.
13. Wu,T., Hu,E., Xu,S., Chen,M., Guo,P., Dai,Z., Feng,T., Zhou,L., Tang,W., Zhan,L., *et al.* (2021) clusterProfiler 4.0: a universal enrichment tool for interpreting omics data. *Innovation*, **2**, 100141.
14. Korotkevich,G., Sukhov,V., Budin,N., Shpak,B., Artyomov,M.N. and Sergushichev,A. (2021) Fast gene set enrichment analysis. bioRxiv doi: https://doi.org/10.1101/060012, 01 February 2021, preprint: not peer reviewed.
15. Gu,Z. and Hübschmann,D. (2023) simplifyEnrichment: a Bioconductor package for clustering and visualizing functional enrichment results. *Genomics Proteomics Bioinformatics*, **21**, 190–202.
16. Sato,N., Uematsu,M., Fujimoto,K., Uematsu,S. and Imoto,S. (2023) ggkegg: analysis and visualization of KEGG data utilizing the grammar of graphics. *Bioinformatics*, **39**, btad622.
17. Xu,S., Dai,Z., Guo,P., Fu,X., Liu,S., Zhou,L., Tang,W., Feng,T., Chen,M., Zhan,L., *et al.* (2021) ggtreeExtra: compact visualization of richly annotated phylogenetic data. *Mol. Biol. Evol.*, **38**, 4039–4042.
18. Yu,G., Smith,D.K., Zhu,H., Guan,Y. and Lam,T.T.-Y. (2017) ggtree : an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol. Evol.*, **8**, 28–36.
19. He,Q., Gao,Y., Jie,Z., Yu,X., Laursen,J.M., Xiao,L., Li,Y., Li,L., Zhang,F., Feng,Q., *et al.* (2017) Two distinct metacommunities characterize the gut microbiota in Crohn's disease patients. *Gigascience*, **6**, 1–11.

20. Zhang,P., Wang,X., Li,S., Cao,X., Zou,J., Fang,Y., Shi,Y., Xiang,F., Shen,B., Li,Y., *et al.* (2023) Metagenome-wide analysis uncovers gut microbial signatures and implicates taxon-specific functions in end-stage renal disease. *Genome Biol.*, **24**, 226.
21. Loehrer,F.M., Angst,C.P., Brunner,F.P., Haefeli,W.E. and Fowler,B. (1998) Evidence for disturbed *S*-adenosylmethionine:*S*-adenosylhomocysteine ratio in patients with end-stage renal failure: a cause for disturbed methylation reactions? *Nephrol. Dial. Transplant*, **13**, 656–661.
22. Yoshida,N., Emoto,T., Yamashita,T., Watanabe,H., Hayashi,T., Tabata,T., Hoshi,N., Hatano,N., Ozawa,G., Sasaki,N., *et al.* (2018) *Bacteroides vulgatus* and *Bacteroides dorei* reduce gut microbial lipopolysaccharide production and inhibit atherosclerosis. *Circulation*, **138**, 2486–2498.
23. Baxter,N.T., Schmidt,A.W., Venkataraman,A., Kim,K.S., Waldron,C. and Schmidt,T.M. (2019) Dynamics of human gut microbiota and short-chain fatty acids in response to dietary interventions with three fermentable fibers. *mBio*, **10**, e02566-18.
24. Chen,S.-J., Chen,C.-C., Liao,H.-Y., Lin,Y.-T., Wu,Y.-W., Liou,J.-M., Wu,M.-S., Kuo,C.-H. and Lin,C.-H. (2022) Association of fecal and plasma levels of short-chain fatty acids with gut microbiota and clinical severity in patients with Parkinson disease. *Neurology*, **98**, e848–e858.
25. Smith,B.J., Li,X., Shi,Z.J., Abate,A. and Pollard,K.S. (2022) Scalable microbial strain inference in metagenomic data using StrainFacts. *Front. Bioinform.*, **2**, 867386.
26. Baud,G.L.C., Prasad,A., Ellegaard,K.M. and Engel,P. (2023) Turnover of strain-level diversity modulates functional traits in the honeybee gut microbiome between nurses and foragers. *Genome Biol.*, **24**, 283.

Received: July 18, 2024. **Revised:** October 22, 2024. **Editorial Decision:** December 16, 2024. **Accepted:** January 5, 2025