Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Development of A Method for the Detection of Antagonism in the Microbial Community and A Toolbox for Identical Sequence-Level Analysis

Sunguk Shin
Yonsei University
Jihyun F. Kim (jfk1@yonsei.ac.kr)
Yonsei University

Research Article

Keywords: Software, antagonism, 16S rDNA amplicon sequencing, microbiome, Crohn's disease

Posted Date: February 3rd, 2021

DOI: https://doi.org/10.21203/rs.3.rs-155732/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Abstract

Competition or antagonism is common in microbial interactions. Therefore, computationally identifying those relationships in the microbial community is in demand. Here, we defined exclusive correlation value (ECV) to better detect mutually exclusive relationships during microbiota analysis. We present METATEA (https://sourceforge.net/projects/metatea/) as a handy toolbox for the calculation of ECV and the analysis of community members with identical 16S rDNA sequences to define intraspecies groups. We analyzed community data related to inflammatory bowel disease to validate the utility of our ECV and METATEA. ECV identified strains antagonistic to disease-associated bacteria, thus providing their potential as probiotics in precision medicine. *Klebsiella pneumoniae* showed exclusive relationships with various bacteria, and its proportion was associated with microbial diversity and Crohn's disease. Further, many putative pathobionts were detected using METATEA, some of which were reported to be isolated from other body sites and cause illness to the host.

Introduction

Complicated and intricate networks of interactions govern microbial communities in an ecosystem. They are more often competitive or antagonistic than mutualistic. Understanding these relationships is important not only for basic understanding of the inner workings of a given microbial ecosystem, but for biotechnological application to medicine, agriculture[1], and environments. Various computational and statistical methods have been developed to assess microbial relationships. For example, Pearson correlation coefficient (PCC)[2], Spearman's rank-order correlation coefficient (SCC)[3], or maximal information coefficient (MIC)[4] had been commonly adopted in most of the previous studies to analyze relationships across microbes. However, the methods are optimized for synergistic relationships between microbes and variables with normal distribution. PCC is not optimized for non-linear relationships. In SCC, weak mutually exclusive relationships, in which the quantity of one variable negatively affects the other variable, can be detected, but non-coexisting relationships, in which the existence of one variable affects the other variable, cannot be distinguished. Although a variety of correlations can be detected using MIC, scientists have not been able to selectively distinguish exclusive relationships from others. New approaches to better detect strong mutually exclusive (almost non-coexisting) relationships across microbial members are desirable. Further, exclusive relationships may not be fully indicated by rank-based measures due to exponential growth of microbes. A convenient method to investigate exclusive relationships could enhance the development of probiotic products. For example, certain probiotic bacteria are considered to reduce intestinal colonization of pathogens[5,6]. Additionally, detection of exclusive relationships with harmful microbes will lead us to precision medicine and reduced adverse effects of probiotic agents.

Several tools such as mothur[7] and QIIME 2[8] are available for scrutinizing microbial communities in environment. Microbial members in the community are most often analyzed at the operational taxonomic unit (OTU) level that is roughly equivalent to the genus level. However, microbes in the same OTU may

have different ecological interactions and cellular functions[9]. For example, while *Escherichia coli* is an important member of the normal gut microbiota, some serotypes are pathogenic[10]. Therefore, 16S analysis below the species levels can aid better understanding of complex diseases, such as inflammatory bowel disease (IBD) that can be divided into two major types, Crohn's disease (CD) and ulcerative colitis (UC), and other forms including indeterminate colitis (IC), and help detect specific bacterial groups associated with the disease type. Thus, assignment of amplicon sequence variants instead of OTUs has been implemented in QIIME 2[8].

In this study, we developed METATEA (METAgenomics Toolbox for Efficient Analysis of microbial communities), a toolbox that enables microbiota analyses at the identical sequence level and evaluates mutually exclusive relationships between microbial members by calculating exclusive correlation value (ECV). METATEA supports various functions, such as basic quality control processes, calculation of sequence proportions, calculation of ECV, statistical analyses, etc. Output file formats were designed to easily perform network analyses and statistical tests using other tools, such as R or MATLAB. In this study, 428 IBD samples, derived from a previous microbiota study[11], were analyzed for performance assessment of METATEA and validation of ECV (Fig. 1). We presented the differences between analysis at sequence level and that at higher taxonomic level, bacterial sequences associated with disease status, and the relationship between *Klebsiella pneumoniae* and bacterial diversity.

Results

Introducing METATEA. METATEA aims to be a toolbox for easy analyses of the microbial community using amplicon sequence data (and whole genome sequence data in the future) at the sequence level and for the detection of antagonistic relationships. Since intraspecies variation in even one gene can change the virulence of a pathogen[12], we developed pipelines for analysis at the sequence level (Fig. 2). Analyzing 428 IBD samples as a case, we had challenges from sequencing errors. Sequencing errors may increase the number of unique sequences, which are weakly proportional to the total numbers of reads (Supplementary Table S1); we assumed such erroneous reads to generally occur once or twice in a sequence file. Therefore, we developed filtering step and normalization step for the removal of rare sequences and the adjustment. Also, various quality control steps were developed to accurately handle reads with Ns and short/long reads at the sequence level. For example, most ambiguous bases (Ns) from the pyrosequencing platform can be corrected[13] using METATEA. For the detection of microbial antagonism, METATEA implements the calculation of ECV (see Methods). Being more user-friendly, METATEA supports matrix output format. The format was designed to easily perform further analyses using other statistical tools like R. METATEA is written in Perl, and we support both Windows-compatible and Linux-compatible executables. METATEA Linux version was compiled from C code for relatively fast analyses of large data sets. METATEA is free, and its detailed manual is available from the project website (https://sourceforge.net/projects/metatea/).

Comparison between taxon-based and sequence-based analyses. We determined the proportional threshold (≥ 0.1%) and extracted 3,202 unique sequences and00 calculated their proportions. To validate our method, the microbial community structure was compared to that in the previous study[11], from which our data set was derived. Although the previous study had used different methods and samples relative to our current one, the regression line in Supplementary Fig. S1 showed our present results to be in accordance with the previous one's with respect to community composition (r^2 = 0.989 in HC and 0.973 in CD).

Generally, analysis results at the sequence level agreed with, and even outperformed, the results at family/genus/OTU levels. First, increase/decrease of bacteria, according to the disease subtypes, at the sequence level generally matched the results of a previous study[11] at the genus or family level (Supplementary Fig. S2). For example, five bacterial sequences in the *Pasteurellaceae* family increased in CD, as reported in a previous study[11]. However, *Bacteroidales* is comprised of various bacteria, whose increase was offset by its decrease. Second, the risk of CD was predicted using random forests at both sequence and OTU (1% and 3% dissimilarity) levels. Prediction at the sequence level outperformed that at the OTU level in terms of the area under the receiver operating characteristic (ROC) curve (AUC) shown in Fig. 3. Third, the co-occurrence of CD-associated organisms in the previous study generally coincided with the correlation network in this study (Supplementary Fig. S3), despite the use of different samples and methods. For example, both sequence identifiers S0276 (*Fusobacterium periodonticum*) and S0345 (*Veillonella dispar*) were increased in CD and showed possible agreement (Supplementary Table S2). However, we could also discover a few exceptions, like S0066 (*Clostridium symbiosum*; *Lachnospiraceae*) and S0346, which were increased in CD and showed possible agreement, although *Lachnospiraceae* was co-excluded in CD in the previous study[11].

Analyses of the microbial community composition. The most frequent sequence identifiers in our IBD data set were S0078 (*Bacteroides vulgatus*), S0052 (*Bacteroides fragilis*), S0029 (*E. coli*), S0051 (*Faecalibacterium prausnitzii*-like), and S0152 (*Bacteroides dorei*) in descending order (Supplementary Table S2). The microbial sequences were the main coefficients in our PCA analysis (Supplementary Fig. S4a). Generally, HC samples clustered in the center, slightly toward sequence S0078, in the PCA analysis. PCA results coincided with the heat map results (Supplementary Fig. S4b). Single bacterial sequences did not show high proportions in HC samples (≥ 50%), whereas certain CD and UC samples did show high proportions (≥ 60%) of sequence S0029 or S0046. Bacterial community structures in CD (pMANOVA: $P < 0.001$) and UC (pMANOVA: $P < 0.001$) were significantly different from that in HC. Several intraspecies variations in disease subtypes were identified, and their average proportions changed differently according to disease subtypes (Supplementary Table S2). Particularly, intraspecies variation of *B. vulgatus*, *E. coli*, *F. prausnitzii*, *Parabacteroides distasonis*, and *Kineothrix alysoides* in average proportion, according to disease subtypes, were prevalent, whereas variation of *B. dorei*, *Haemophilus parainfluenzae*, *Prevotella copri*, and *Fusobacterium nucleatum* were less obvious (Supplementary Fig. S5).

Detection of disease-associated community members. Multiple Mann-Whitney *U* tests were performed using METATEA to identify sequences associated with disease subtypes (Supplementary Table S3). Generally, *H. parainfluenzae*, *Dialister pneumosintes*, *F. periodonticum*, and many other sequences were significantly increased in CD. *H. parainfluenzae*-like, *F. prausnitzii*-like, *Sutterella massiliensis*-like, and many others were significantly increased in UC. We newly discovered *Veillonella atypica*, *Pseudomonas lini*, and *Hyphomicrobium zavarzinii* to possibly be related to CD, IC, and UC, respectively, probably due to high-resolution analyses at the sequence level. *Lachnospiraceae* showed variations in disease subtypes (Supplementary Fig. S6). Intraspecies variations in Z scores were apparent, particularly among sequences of low proportions (Supplementary Table S2). For example, although *F. prausnitzii* (S0376) and some *F. prausnitzii*-like sequences (S0051) were increased in HC, other *F. prausnitzii*-like sequences, such as S0397 and S2816, were increased in CD and UC, respectively (Supplementary Table S3). Generally, sequences with high Z scores (≥ 2.3264) in CD and UC compared to those in HC might be those of bacteria associated with various sites of the body or opportunistic environmental bacteria, rather than members of the normal microbiota in the gut (Supplementary Table S4). Some of these aberrant bacteria of putatively external origins are reportedly related to various diseases like endocarditis[14], carcinoma[15], and abscess[16]. For more accurate prediction of CD risk, using more biomarkers at the identical sequence level may be an appropriate approach rather than using a small number of powerful biomarkers. For example, sequences like S0362 (*Roseburia inulinivorans*) can be the most powerful biomarkers to predict CD risk (Fig. 3b). However, even the highest proportion of S0362 cannot confirm HC status (Supplementary Fig. S7).

Comparison between ECV, PCC, and SCC. ECV was designed to selectively identify highly exclusive relationships among all microbial pairs (Fig. 4). This ability is very important in terms of microbial data because most microbial pairs can show widely scattered inversely proportional correlations as shown in Fig. 5 around the median values of ECV, PCC, and SCC. ECV successfully identified strong antagonistic (almost non-coexisting) relationships between *K. pneumoniae* (S0390) and *F. prausnitzii*-like (S0764, S1650, and S25900) sequences around the maximum values (Fig. 5a). Although the negative relationships between *K. pneumoniae* and *F. prausnitzii*-like sequences are to be experimentally verified, *K. pneumoniae* increased in disease samples, whereas *F. prausnitzii* increased in control samples in previous studies[17,18].

We then compared ECV, PCC, and SCC for their capabilities to identify antagonism. Widely used PCC is one of the robust means to detect positive and negative linear correlations. Although PCC finds well co-occurring microbial members having positive correlations at the maximum values (~ 1), strong negative relationships (~-1) were undetectable using PCC (Fig. 5b). Among the microbial pairs (*B. vulgatus* and *B. dorei* as well as closely related sequences) with extremely low correlation coefficients, none had the coefficients lower than − 0.3 (Supplementary Table S5). SCC assesses monotonic relationships using a rank-based measure (whether linear or not), while PCC assesses linear correlations. SCC identified microbial pairs with inversely proportional correlations better around the minimum values than PCC (Fig. 5c). However, strong antagonistic relationships could not be identified even with SCC. For example,

S0535 (*Haemophilus pittmaniae*) and S1732 (*E. coli*-like) showed strong antagonistic relationships, and their SCC was 0.039. Pairs (*Lachnospiraceae* and *Oscillibacter ruminantium*-like; *Blautia wexlerae* and *Escherichia coli*-like; *Bacteroides uniformis* and *Escherichia coli*-like) with correlation coefficients from the minimum value were − 0.367, -0.346, and − 0.336.

Relationships between community members. Relationships across different sequences of the same species may be associated with disease subtypes and be useful for predicting CD risk. For example, IBD samples were categorized into two groups, as shown in Supplementary Fig. 8a, and high ratio of S0103 (*B. vulgatus*-like) to S0092 (*B. vulgatus*-like) might be related to CD. We successfully detected mutually exclusive relationships using ECV at three different OTU levels (0, 1, and 3% dissimilarity); S0390 (*K. pneumoniae*), 1%OTU281, and 3%OTU010 generally had significantly exclusive relationships with a variety of bacteria (Supplementary Table S6-S8; Supplementary Fig. S9) and increased in CD than in HC (Fig. 6). Particularly, high proportions of S0390 might be related to reduced bacterial diversity (Supplementary Fig. S10). Different microbes showed different ECV values with certain disease-associated microbes. We regarded 15 sequences (average proportion ≥ 0.1%) as putative probiotics that occurred at least three times in HC than in CD, and 8 sequences were removed by the ECV formula (see Methods), The 7 putative probiotic sequences, S0342 (*F. prausnitzii*), S0362 (*R. inulinivorans*), S0463 (*Blautia luti*), S0464 (*Clostridium amygdalinum*), S0661 (*Lachnospiracea_incertae_sedis*), S0706 (*Lachnospiracea_incertae_sedis*), and S0707 (*unclassified_Lachnospiraceae*) in general showed exclusive relationships with disease-associated sequences (Supplementary Table S9). Particularly S0342 showed strong mutually exclusive relationships with *E. coli* whereas S0464 did with *F. periodonticum* (Supplementary Table S10). Additionally, ECV was designed to be less affected by sample sizes, and pairs of S0390 and other sequences showed higher ECV values in HC than in CD (Supplementary Table S11). For example, ECV value of S0089 (*Bacteroides koreensis*-like) and S0390 in HC was higher than in CD, suggesting putative synergistic effects of S0390 whether it is direct or not (Supplementary Fig. S8b). Additional studies using more clinical data would validate our observation.

Discussion

This study focused on the performance analyses of METATEA with identical sequence matching, and on ECV to identify strong mutually exclusive relationships between microbes. The high resolution of sequence-based analyses enabled us to detect missing disease-associated microbes within the same taxonomic unit. ECV identified strong mutually exclusive relationships between microbes, whereas PCC and SCC could not distinguish strong mutually exclusive relationships from weak ones in microbial data sets. Although intraspecies variations can be detected using QIIME 2[8], we newly developed METATEA for the convenient calculation of ECV from microbial proportions. Also, METATEA can be easily installed regardless of the limitation of environment managers. The identical sequence matching process using METATEA can be performed without traditional time-consuming jobs, such as distance calculation and clustering compared to OTU-based tools. The efficient process will enable accurate diagnosis of diseases.

Analyzing microbial communities at the sequence level can have more merits than that at high taxonomic levels. We identified missing disease-associated microbes by the offset from the reduced strains within the same group. The high-resolution of bacterial populations at the sequence level enabled us to newly detect many disease-associated microbes, such as *V. atypica*, *P. lini*, and *H. zavarzinii*. It has been reported that IBD is linked to various diseases, such as carcinoma, endocarditis, and abscess[19–21]. Beyond our expectations, many of the specific strains of disease-associated taxa were related to various sites of the body and lethal diseases, not only due to immune dysfunction[22], but also due to various disease-associated microbes and probably due to evolutionary advantage for movement between body sites. Prediction at the sequence level outperformed that at the OTU level in terms of AUC. Also, novel relationships between strains within the same species could be identified.

The merits of analyzing microbial communities at the sequence level might stem from intraspecies variations. We identified intraspecies variation of *B. vulgatus*, *E. coli*, *F. prausnitzii*, *P. distasonis*, and *K. alysoides*. Especially, *F. prausnitzii* is reported as a potential novel probiotic bacterium for IBD although the beneficial mechanisms are still unclear[23]; however, certain *F. prausnitzii*-like sequences, such as S0397 and S2814, were increased in CD and UC, respectively. Particularly, probiotics should be carefully studied at the sequence level to reduce possible side effects due to intraspecies variations. Strains within *B. vulgatus* species might have novel relationships: *B. vulgatus*-like sequences (S0092 and S0103) showed two different ratios. The high ratio of S0103 might activate NF-κB in human gut epithelial cells[24], or inflammation of the gastrointestinal tract might result in the high ratio of S0103; more experimental studies would be required in this direction.

PCC, SCC, and ECV all efficiently detected different relationships between microbial members. Although PCC could detect positive linear correlations that correspond to co-occurrence, mutually exclusive relationships that correspond to reciprocal exclusion are poorly identified. SCC could flexibly identify negative or positive correlations, but the strongest exclusive relationships could be missed. ECV could not exactly identify positive linear correlations, but ECV could detect the strongest exclusive relationships between microbial members. ECV can be useful for the development of precision probiotics, since probiotics may have antagonistic effects of different strengths on specific pathogens. Custom probiotic supplements might be needed so as to prevent undesirable side effects. Additionally, ECV can be useful for understanding reduced microbial diversity and CD. Specific sequence identifiers have strong mutually exclusive relationships with a variety of bacterial sequences and increase in CD. For example, high proportions of S0390, a sequence of *K. pneumoniae*, are significantly associated with decreased microbial diversity. *K. pneumoniae* is a lactose-fermenting bacterium that causes infections especially in alcoholic patients[25], and ectopic colonization of *Klebsiella* spp. in the intestine drives T helper 1 cell induction and inflammation[26]. Intestinal dysbiosis and decreased microbial diversity caused by drinking alcohol are related to increased *K. pneumonia*[27]. Moreover, high-alcohol-producing *K. pneumoniae* can be associated endogenous alcohol production[28]. Whatever the exact reasons for reduced bacterial diversity and CD should be, *K. pneumoniae* may be associated with reduced microbial diversity and CD. Also, *K.*

pneumoniae might have synergistic effects with other disease-associated microbes according to differences in ECV in disease subtypes.

The efficient analyses using METATEA and ECV will enable accurate risk prediction of complex diseases, identification of more disease-associated microbes and antagonistic relationships between microbes, in-depth understanding of reduced microbial diversity, and development of precision probiotics to reduce undesirable side effects.

Materials And Methods

Definition of ECV. We defined ECV to identify strong mutually exclusive relationships. ECV was designed to be useful for the practical detection of mutually exclusive relationships between bacteria, as much as possible, so as to exclude the effects of group size. First, we generated candidate formulae to investigate mutually exclusive relationships. A single basic formula was selected that would be less affected by group size due to the division by multiplication of microbial proportion. Then, we analyzed the data distribution and performed data visualization using public data sets and those produced in house to empirically modify the formula using the numbers of zero proportions. Let *a* and *b* be the proportions of two bacterial sequences, and n be the number of observed proportions. To remove rare bacterial sequences, we counted the number of proportions consisting of zeros, since rare bacterial sequences can be shown to be exclusive by chance because rare bacterial sequences sparsely coexist and the relationships of the sequences can be easily misunderstood as exclusive relationships. Let *N(a)*, *N(b)*, and *N(ab)* be the number of zeros in *a*, *b*, and both *a* and *b*. In this study, too rare sequences were removed by the left conditions mentioned below. However, METATEA users should apply right conditions for small sample sizes (< 20):

$$N(a) \geq \frac{2n}{3}, N(b) \geq \frac{2n}{3} \text{ or } N(a) \geq \frac{n}{2}, N(b) \geq \frac{n}{2}$$

Finally, ECV was defined as

$$\mathrm{ECV} = \frac{\sum_{k=1}^{n}(a_k + b_k)^2 \times (N(a)+1) \times (N(b)+1)}{(N(ab)+1) \times \sum_{k=1}^{n}(a_k \times b_k)}$$

$\sum_{k=1}^{n}(a_k + b_k)^2 \ / \ \sum_{k=1}^{n}(a_k \times b_k)$ is the main part to detect mutually exclusive relationships (Fig. 4a). (*N(ab)*+1) is supplementary part to filter out rare sequences, thereby removing false positives incurred by skewed distributions and outliers (Fig. 4b). (*N(a)*+1) and (*N(b)*+1) are supplementary parts to accurately estimate antagonism which were obtained after observing additional gastric microbiota and soil microbiota data sets (Fig. 4c).

Public data sets. METATEA is applicable to any 16S amplicon sequence data, and ECV is applicable to various fields of natural science. Particularly, we expect that our efficient and simple analysis pipeline,

high-resolution detection of disease-associated microbes, and ECV for detection of antagonism against disease-associated microbes are suitable for analyzing gut microbiome data for clinicians. To convey the potential applications of METATEA to human gut microbiota data, we describe here the public data sets that were analyzed using METATEA. We selected 140 healthy control (HC), 214 CD, 20 IC, and 54 UC data sets in BioProject: PRJEB13679 that were relevant to the previous study[11] on IBD in order to develop sequence-supervised processes, assess the performance of METATEA, and validate ECV. Although the symptoms of IBD are heterogeneous[29], the samples were classified into four subtypes based on clinical metadata: HC, CD, IC, and UC. Only the data sets of terminal ileum biopsy (Supplementary Table S1) were downloaded using SRA Toolkit (v.2.9.0), since the terminal ileum represents a transition zone to detect both the aerobic flora and anaerobic microbes[30]. Moreover, in CD risk prediction, terminal ileum outperformed other biopsy locations, such as rectum and stool[12]; CD has been shown to frequently occur in the terminal ileum[31].

Main processing. For objective analyses at the sequence level, abnormally short or long reads (< 175 bp or >175 bp) were removed. After BLAST analysis, 11 sequences were identified as human DNA sequences occurring frequently, and were used for the decontamination step using METATEA. After quality control processes, the sum of proportions of microbial members was adjusted to one. Heat map, box plots, and PCA analysis were prepared using MATLAB (R2020a). Mann-Whitney *U* tests were performed using METATEA to compare CD/IC/UC with HC, and the significantly increased disease-associated sequences were identified. To test for differences in microbial community structures between disease subtypes, pMANOVA tests were performed using the function 'adonis2' of the vegan package in R (v4.0.0). Weighted correlation network analysis and CD risk prediction were performed using R. To predict CD risk, 80% and 20% of data sets were randomly used for training and testing, respectively. Square roots of the sequence/OTU numbers were used to tune mtry parameters. We calculated OTU (1% and 3% dissimilarity) using Mothur (v.1.44.1) with SILVA recreated SEED database (v.138). Taxonomic classification was mainly performed using RDP classifier at family level. Taxonomic classification at species level was performed using web BLAST with 16S ribosomal RNA sequence database, and the BLAST results were confirmed by RDP classifier at genus/family level. If the BLAST results were inconsistent with RDP classifier results, we considered RDP results at genus level. In case of recent re-classification, we considered literature search results. Literature was surveyed to identify the relationships between IBD and other diseases, since oral bacteria are known to be related to IBD and various diseases[32] and IBD is related to various, often lethal, diseases such as cancer[33], endocarditis[34], and neurological disorders[35]. The maximum likelihood method was selected to construct a phylogenetic tree using MEGA-X (v10.0.5). The analysis pipeline for unsupervised analysis using a few samples is described in Fig. 2.

Comparison with a previous study and determination of a proportional threshold. To compare the results in this study to that in the previous study[11], taxonomic analysis was performed using data from terminal ileum in Supplemental Table 3 of the previous study[11]. After quality control processes using METATEA, we assumed the sequencing errors to have increased the number of unique sequences, since the number

of reads was almost proportional to the number of unique sequences (Supplementary Table S1). We found putative erroneous reads, which generally occur only once in a single file, and the number of reads with proportions ≥ 0.1% were not proportional to the number of reads in each file. Therefore, we determined the proportional threshold (≥ 0.1%) for the development of sequence-supervised processes and extracted 3,202 unique sequences (Supplementary Table S12). For METATEA users, we recommend that the threshold should be determined according to the number of reads in the file since putative erroneous reads might generally occur once in a single file in this study. Generally, threshold can be determined as

$$\text{Threshold} = \frac{1}{the\ \ smallest\ \ number\ \ of\ \ reads\ \ among\ \ your\ \ files - 1}$$

But one may have to use greater proportions depending on further analyses since too big data often leads to errors when you use statistical tools.

Declarations

Acknowledgements

We thank all the members of our laboratory, including Ju Yeon Song, Ji-Won Huh, Jae Kyung Yoon, and Jongseok Kim, for useful discussion and testing ECV.

Author contributions

S.S. developed METATEA and performed the analyses. J.F.K. contributed to designing the analysis pipeline and supervised the study.

Funding

This work has been supported by the National Research Foundation (NRF-2014M3C9A3068822) funded by the Ministry of Science and ICT, Republic of Korea.

Competing Interests

The authors declare that there is no conflict of interest regarding the publication of this paper.

ORCID

Sunguk Shin https://orcid.org/0000-0002-3812-8320
Jihyun F. Kim https://orcid.org/0000-0001-7715-6992

References

1. Kwak, M. J. *et al.* Rhizosphere microbiome structure alters to enable wilt resistance in tomato. *Nat Biotechnol*, doi:10.1038/nbt.4232 (2018).
2. Sedgwick, P. Spearman's rank correlation coefficient. *BMJ.* **349**, g7327 https://doi.org/10.1136/bmj.g7327 (2014).
3. Reshef, D. N. *et al.* Detecting novel associations in large data sets. *Science.* **334**, 1518–1524 https://doi.org/10.1126/science.1205438 (2011).
4. Piewngam, P. *et al.* Pathogen elimination by probiotic Bacillus via signalling interference. *Nature.* **562**, 532–537 https://doi.org/10.1038/s41586-018-0616-y (2018).
5. Spinler, J. K. *et al.* Human-derived probiotic Lactobacillus reuteri demonstrate antimicrobial activities targeting diverse enteric bacterial pathogens. *Anaerobe.* **14**, 166–171 https://doi.org/10.1016/j.anaerobe.2008.02.001 (2008).
6. Schloss, P. D. *et al.* Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol.* **75**, 7537–7541 https://doi.org/10.1128/AEM.01541-09 (2009).
7. Bolyen, E. *et al.* Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat Biotechnol.* **37**, 852–857 https://doi.org/10.1038/s41587-019-0209-9 (2019).
8. Greenblum, S., Carr, R. & Borenstein, E. Extensive strain-level copy-number variation across human gut microbiome species. *Cell.* **160**, 583–594 https://doi.org/10.1016/j.cell.2014.12.038 (2015).
9. Feng, P. Escherichia coli serotype O157:H7: novel vehicles of infection and emergence of phenotypic variants. *Emerg Infect Dis.* **1**, 47–52 https://doi.org/10.3201/eid0102.950202 (1995).
10. Gevers, D. *et al.* The treatment-naive microbiome in new-onset Crohn's disease. *Cell Host Microbe.* **15**, 382–392 https://doi.org/10.1016/j.chom.2014.02.005 (2014).
11. Hirakawa, M. P. *et al.* Genetic and phenotypic intra-species variation in Candida albicans. *Genome Res.* **25**, 413–425 https://doi.org/10.1101/gr.174623.114 (2015).
12. Shin, S. & Park, J. Correction of sequence-dependent ambiguous bases (Ns) from the 454 pyrosequencing system. *Nucleic Acids Res.* **42**, e51 https://doi.org/10.1093/nar/gku070 (2014).
13. Chunn, C. J., Jones, S. R., McCutchan, J. A., Young, E. J. & Gilbert, D. N. Haemophilus parainfluenzae infective endocarditis. *Med. (Baltim).* **56**, 99–113 https://doi.org/10.1097/00005792-197703000-00002 (1977).
14. Castellarin, M. *et al.* Fusobacterium nucleatum infection is prevalent in human colorectal carcinoma. *Genome Res.* **22**, 299–306 https://doi.org/10.1101/gr.126516.111 (2012).
15. Rousee, J. M. *et al.* Dialister pneumosintes associated with human brain abscesses. *J Clin Microbiol.* **40**, 3871–3873 https://doi.org/10.1128/jcm.40.10.3871-3873.2002 (2002).
16. Forbes, J. D., Van Domselaar, G. & Bernstein, C. N. The Gut Microbiota in Immune-Mediated Inflammatory Diseases. *Front Microbiol.* **7**, 1081 https://doi.org/10.3389/fmicb.2016.01081 (2016).
17. Yan, Q. *et al.* Alterations of the Gut Microbiome in Hypertension. *Front Cell Infect Microbiol.* **7**, 381 https://doi.org/10.3389/fcimb.2017.00381 (2017).

18. Axelrad, J. E., Lichtiger, S. & Yajnik, V. Inflammatory bowel disease and cancer: The role of inflammation, immunosuppression, and cancer treatment. *World J Gastroenterol.* **22**, 4794–4801 https://doi.org/10.3748/wjg.v22.i20.4794 (2016).
19. Kreuzpaintner, G., Horstkotte, D., Heyll, A., Losse, B. & Strohmeyer, G. Increased risk of bacterial endocarditis in inflammatory bowel disease. *Am J Med.* **92**, 391–395 https://doi.org/10.1016/0002-9343(92)90269-h (1992).
20. Casella, G. *et al.* Neurological disorders and inflammatory bowel diseases. *World J Gastroenterol.* **20**, 8764–8782 https://doi.org/10.3748/wjg.v20.i27.8764 (2014).
21. Neuman, M. G. Immune dysfunction in inflammatory bowel disease. *Transl Res.* **149**, 173–186 https://doi.org/10.1016/j.trsl.2006.11.009 (2007).
22. Miquel, S. *et al.* Faecalibacterium prausnitzii and human intestinal health. *Curr Opin Microbiol.* **16**, 255–261 https://doi.org/10.1016/j.mib.2013.06.003 (2013).
23. P, O. C. *et al.* The gut bacterium and pathobiont Bacteroides vulgatus activates NF-κB in a human gut epithelial cell line in a strain and growth phase dependent manner. *Anaerobe.* **47**, 209–217 https://doi.org/10.1016/j.anaerobe.2017.06.002 (2017).
24. Prince, S. E., Dominger, K. A., Cunha, B. A. & Klein, N. C. Klebsiella pneumoniae pneumonia. *Heart Lung.* **26**, 413–417 https://doi.org/10.1016/s0147-9563(97)90028-5 (1997).
25. Atarashi, K. *et al.* Ectopic colonization of oral bacteria in the intestine drives TH1 cell induction and inflammation. *Science.* **358**, 359–365 https://doi.org/10.1126/science.aan4526 (2017).
26. Samuelson, D. R. *et al.* Alcohol-associated intestinal dysbiosis impairs pulmonary host defense against Klebsiella pneumoniae. *PLoS Pathog.* **13**, e1006426 https://doi.org/10.1371/journal.ppat.1006426 (2017).
27. Yuan, J. *et al.* Fatty Liver Disease Caused by High-Alcohol-Producing *Klebsiella pneumoniae*. *Cell Metab* **30**, 675–688 e677, doi:10.1016/j.cmet.2019.08.018 (2019).
28. Lamb, C. A. *et al.* British Society of Gastroenterology consensus guidelines on the management of inflammatory bowel disease in adults. *Gut.* **68**, s1–s106 https://doi.org/10.1136/gutjnl-2019-318484 (2019).
29. Quigley, E. M. & Abu-Shanab, A. Small intestinal bacterial overgrowth. *Infect Dis Clin North Am* **24**, 943–959, viii-ix, doi:10.1016/j.idc.2010.07.007 (2010).
30. Caprilli, R. Why does Crohn's disease usually occur in terminal ileum? *J Crohns Colitis.* **2**, 352–356 https://doi.org/10.1016/j.crohns.2008.06.001 (2008).
31. Han, Y. W. & Wang, X. Mobile microbiome: oral bacteria in extra-oral infections and inflammation. *J Dent Res.* **92**, 485–491 https://doi.org/10.1177/0022034513487559 (2013).
32. Axelrad, J. E., Lichtiger, S. & Yajnik, V. Inflammatory bowel disease and cancer: The role of inflammation, immunosuppression, and cancer treatment. *World J Gastroenterol.* **22**, 4794–4801 https://doi.org/10.3748/wjg.v22.i20.4794 (2016).

33. Kreuzpaintner, G., Horstkotte, D., Heyll, A., Losse, B. & Strohmeyer, G. Increased risk of bacterial endocarditis in inflammatory bowel disease. *Am J Med.* **92**, 391–395 https://doi.org/10.1016/0002-9343(92)90269-h (1992).
34. Casella, G. *et al.* Neurological disorders and inflammatory bowel diseases. *World J Gastroenterol.* **20**, 8764–8782 https://doi.org/10.3748/wjg.v20.i27.8764 (2014).
35. Casella, G. *et al.* Neurological disorders and inflammatory bowel diseases. *World J Gastroenterol* **20**, 8764–8782, doi:10.3748/wjg.v20.i27.8764 (2014).

Figures

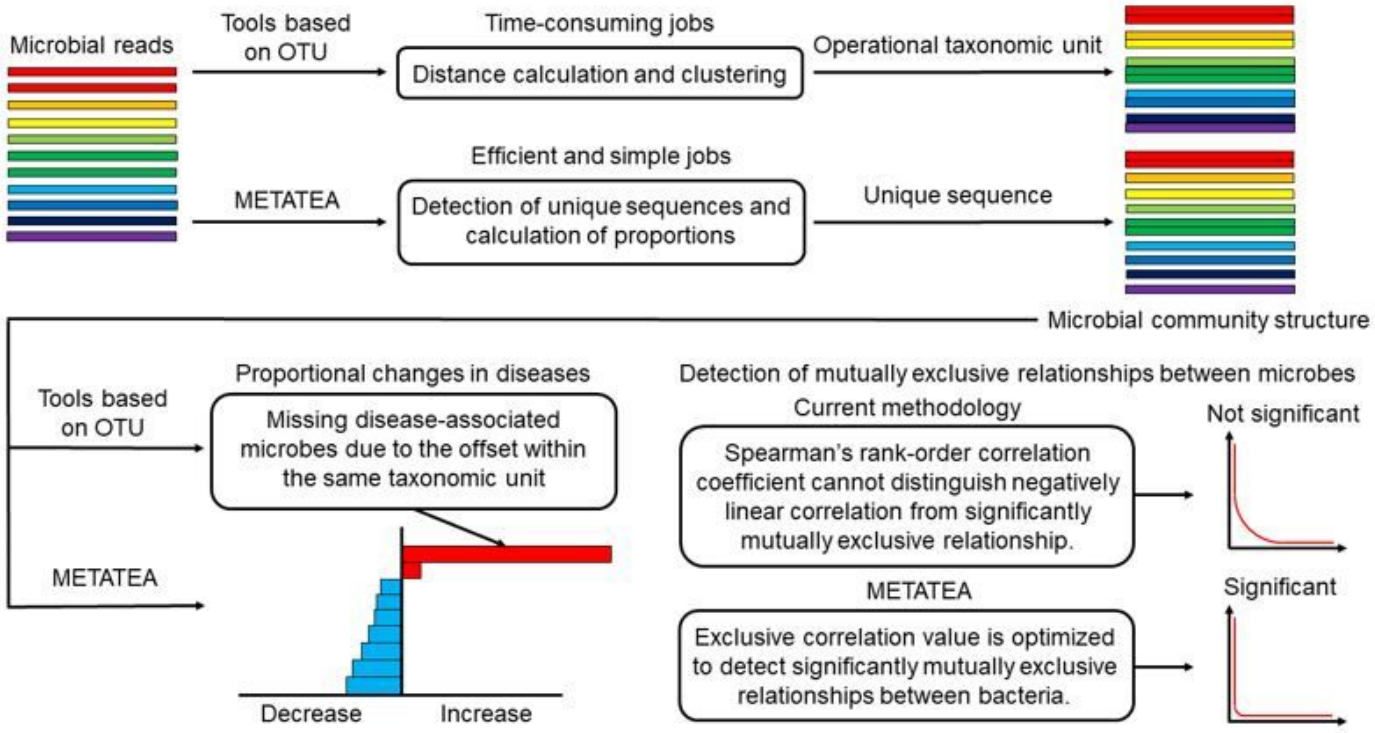

Figure 1

Comparison of METATEA and tools based on OTU.

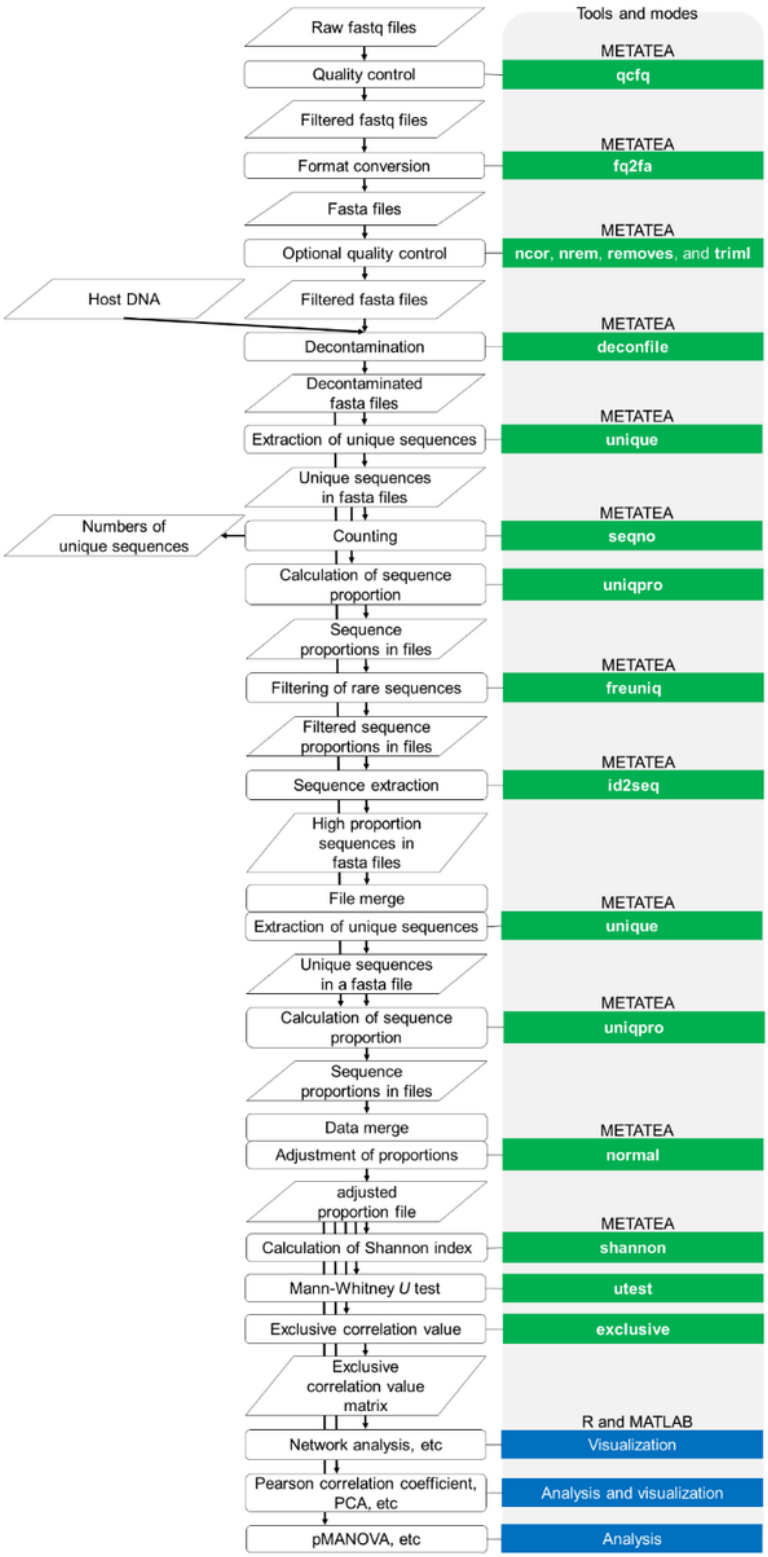

Figure 2

Schematic of the steps in unsupervised analysis using multiple samples.

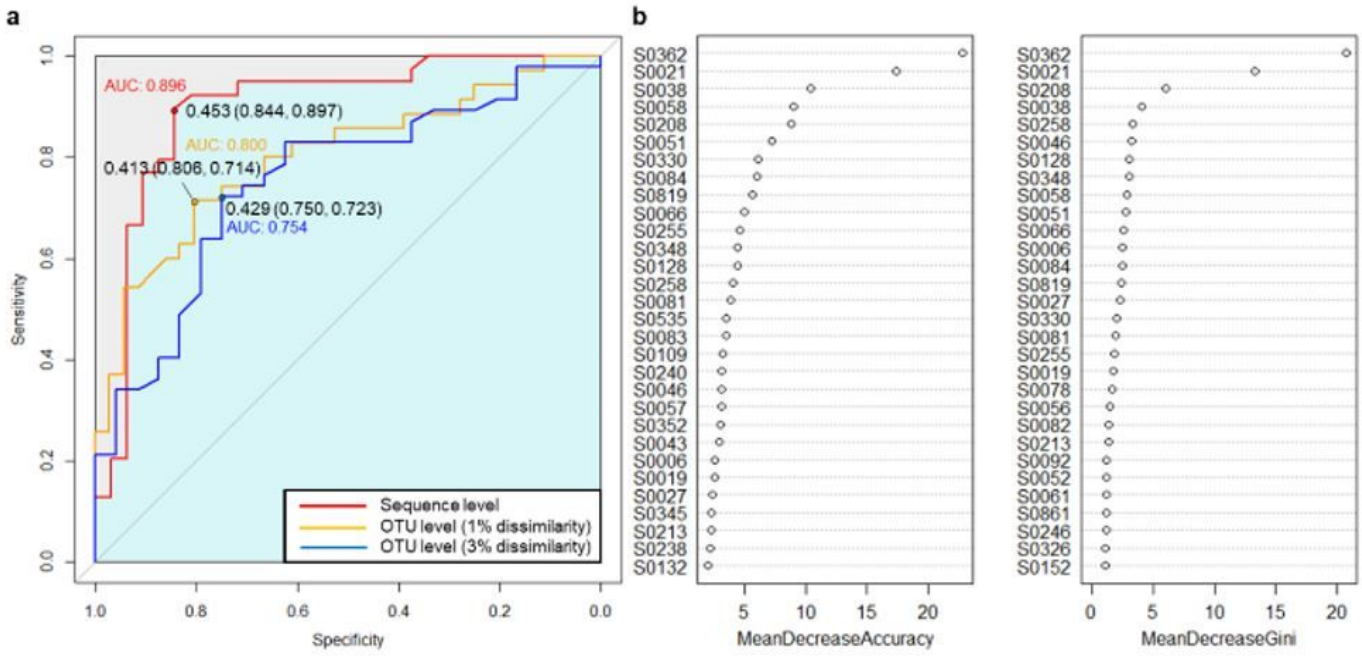

Figure 3

ROC curve for the risk of Crohn's disease and variable importance plot. a, This ROC curve compares the performance of prediction at sequence and OTU (1% and 3% dissimilarity) levels. b, Variable importance plot shows the most important microbial sequences.

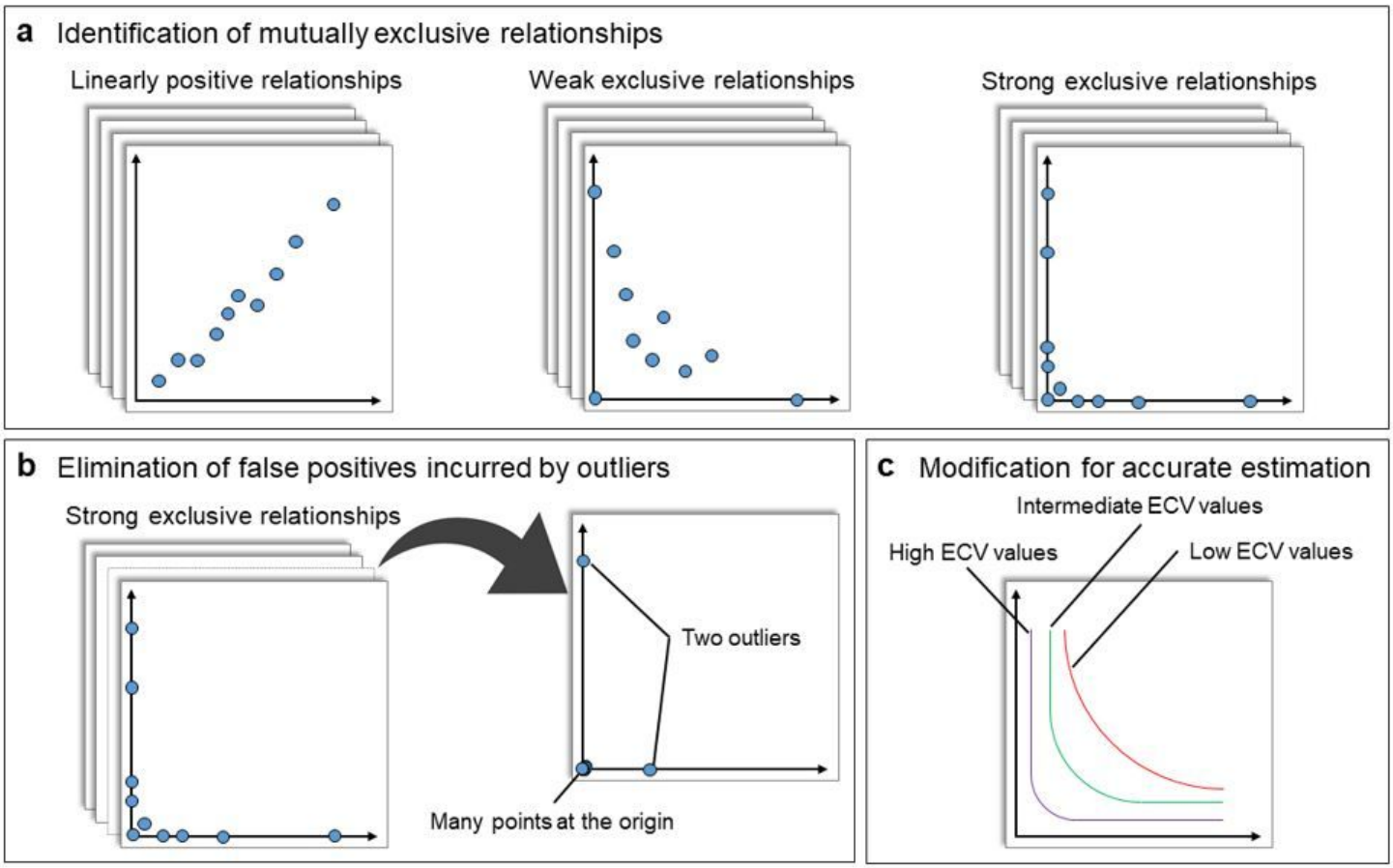

Figure 4

Examples and explanation of the ECV formula.

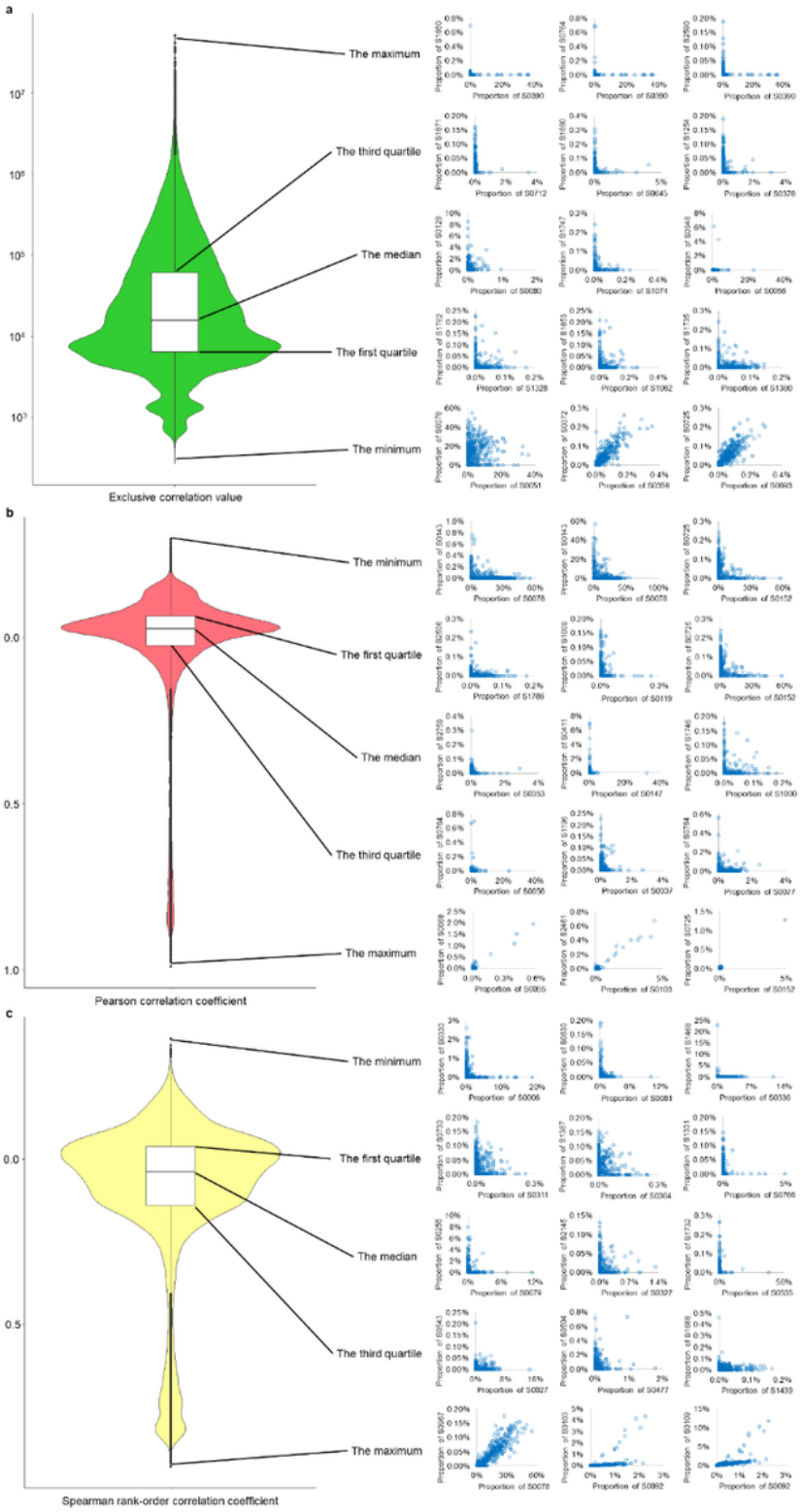

Figure 5

Plots and examples of exclusive correlation values, Pearson correlation coefficient, and Spearman's rank-order correlation coefficient. a, exclusive correlation values. b, Pearson correlation coefficient. c, Spearman's rank-order correlation coefficient. Rare bacterial sequences were removed in the same way for exclusive correlation value, Pearson correlation coefficient, and Spearman's rank-order correlation coefficient (see Definition of ECV).

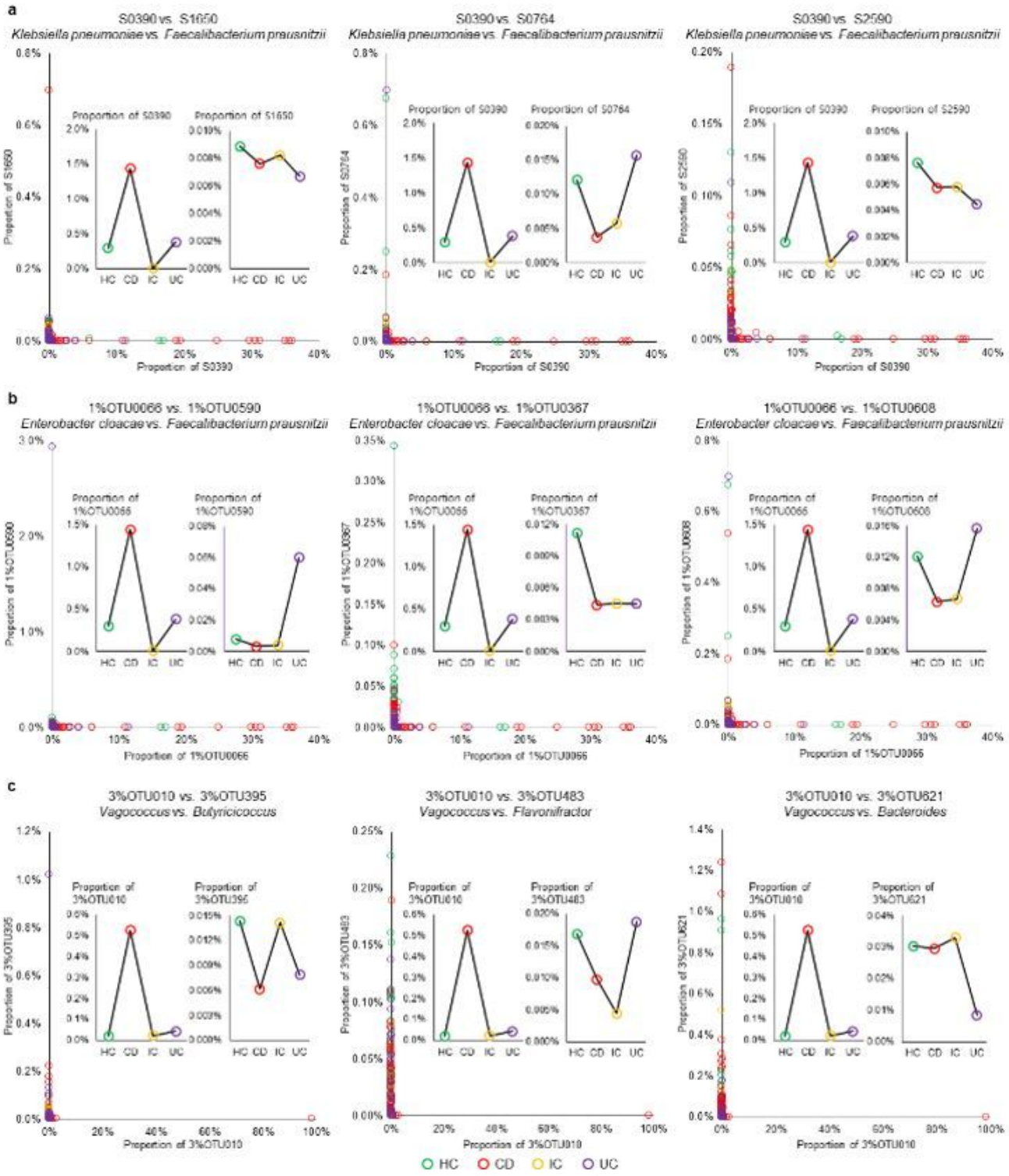

Figure 6

Examples of highly exclusive relationships. a, examples of highly exclusive relationships at sequence level. b, examples of highly exclusive relationships at 1% dissimilarity level. c, examples of highly exclusive relationships at 3% dissimilarity level. Green, red, orange, and purple circles denote HC, CD, IC, and UC samples, respectively. CD, Crohn's disease; HC, Healthy control; IC, Indeterminate colitis; UC, Ulcerative colitis.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- METATEASRSupplementaryfile1v2.47.pdf
- METATEASRSupplementaryfile2Table12v2.44.xlsx