Strain-level analysis of microbial community and detection of mutually exclusive relationships uncover complexity of inflammatory bowel disease

Sunguk Shin
Yonsei University
Jihyun F. Kim (jfk1@yonsei.ac.kr)
Yonsei University https://orcid.org/0000-0001-7715-6992

Research

Keywords: Microbiome, microbiota, microbial diversity, 16S rDNA sequence, Crohn's disease, ulcerative colitis, Bacteroides vulgatus, Klebsiella pneumonia

Posted Date: February 24th, 2020

DOI: https://doi.org/10.21203/rs.2.24332/v1

Abstract

Background Despite intraspecies variation in ecological interaction and cellular function, most analytical methods for microbial communities that use the 16S rRNA gene are employed at the taxonomic level. Further, methods that detect positive or negative relationships between microbes such as Pearson correlation coefficient are generally applicable for linear correlations. Results We present METATEA for the analysis of community members with identical 16S rDNA sequences to define intraspecies groups, and for the calculation of exclusive correlation coefficient (ECC) to detect mutually exclusive relationships. Proportional variation of identical sequence groups in each disease subtype was revealed using a 16S rDNA data set of inflammatory bowel disease (IBD) samples. Conclusions Results at the identical sequence level complied with and outperformed those at the species or genus levels. Intraspecies variation was prevalent within Faecalibacterium prausnitzii , suggesting that some strains might be associated with diseases although it is known to be abundant in the human gut microbiota of healthy adults. IBD samples were categorized into two groups based on the ratios of certain Bacteroides vulgatus -like sequences. ECC identified strains antagonistic to disease-associated bacteria, thus proving their potential as probiotics in precision medicine. Klebsiella pneumoniae showed exclusive relationships with various bacteria, and its proportion was associated with bacterial diversity and Crohn’s disease. We expect METATEA to allow high resolution analysis of microbial communities and easy identification of pathogen-antagonistic probiotic microbes.

Background

Amplicon sequencing of the 16S rRNA gene has been one of the most popular methods to study microbial communities. Despite technological advancement, 16S rDNA (16S) sequences obtained from a next-generation sequencing platform have traditionally been analyzed at taxonomic ranks higher than genus, owing to the availability of limited nucleotide information due to the short lengths of conserved sequences, limited numbers of taxa and identified 16S sequences, and ambiguous taxonomic classification. At present, taxonomic classification that uses partial 16S sequences is mostly limited to the level of operational taxonomic units (OTUs), and most toolboxes do not support optimized analysis pipelines at the sequence level. The basic assumption underlying the existing 16S analyses at each taxonomic level is that bacteria or archaea in the same taxonomic group should have the same basic features and functions. However, this assumption is yet to be justified, particularly in microbiome studies. Microbes in the same OTU can have different ecological interactions and cellular functions [1]. For example, while Escherichia coli is an important member of the normal gut microbiota, some serotypes are pathogenic [2]. 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.

In addition, new approaches to better detect mutually exclusive relationships across microbial members are desirable. Mechanisms of exclusive relationships are known to include competition for nutrients,

production of inhibiting factors, direct antagonism by bacteriocins, competitive exclusion for binding sites, improved barrier functions, alteration of intestinal properties by increasing inflammation, and stimulation of the immune system [3]. However, mutually exclusive relationships are seldom detected due to the lack of an appropriate methodology. Pearson correlation coefficient (PCC) or maximal information coefficient (MIC) [4] had been commonly adopted in most of the previous studies to analyze relationships across microbes. However, mutually exclusive relationships do not belong to linear relationships denoted by PCC. Although a variety of correlations can be detected using MIC, scientists have not been able to selectively distinguish exclusive relationships from others. 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].

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 coefficient (ECC). METATEA supports various functions, such as basic quality control processes, calculation of sequence proportions, calculation of ECC, 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 [7], were analyzed for performance assessment of METATEA and validation of ECC. We present 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

Development of METATEA and definition of exclusive correlation coefficient

METATEA is a toolbox for 16S analyses at the sequence level to assess intraspecies variation and for detection of mutually exclusive relationships, like antagonism between pathogens and probiotics. In brief, METATEA first extracts unique 16S sequences and calculates their proportions (Figure 1). At first, quality control is performed before matching the sequence reads exactly. For example, ambiguous bases (Ns) in reads can be corrected [8], or reads with Ns can be removed. Too short or too long reads can be removed or optionally trimmed. Unique sequences can then be extracted, and their proportions calculated. Users can optionally determine proportional thresholds, since the number of reads can be almost proportional to the number of unique sequences (Additional file 1: Table S1), probably due to sequencing errors. Statistical analysis and visualization are performed using METATEA and other toolboxes like R.

We also defined ECC to identify mutually exclusive relationships. Although negative linear correlation can be detected using PCC, mutually exclusive relationships in bacterial proportions appear to be inversely proportional, rather than being negatively linear. Unlike PCC, ECC is optimal to detect non-linear mutually exclusive relationships (Additional file 1: Figure S1). For example, among three examples around the PCC value of zero, only the relationship between sequence identifiers S0245 (*Bacteroides vulgatus*-like) and

S1754 (*B. vulgatus*-like) appeared to be mutually exclusive. However, all three examples with the highest ECC values showed mutually exclusive relationships.

Comparison between taxon-based and sequence-based analyses

To analyze 428 IBD samples, sequence-supervised processes were developed. Sequencing errors may increase the number of unique sequences (Additional file 1: Table S1); we assumed such erroneous reads to generally occur only once in a sequence file. Therefore, we determined the proportional threshold (≥ 0.1%) for the development of sequence-supervised processes in order to confirm that the extracted unique reads exist at least twice in a sequence file. We extracted 3,204 unique sequences and calculated their proportions. To validate our method, the microbial community structure was compared to that in the previous study [7], 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 Additional file 1: Figure S2 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 0.977 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 [7] at the genus or family level (Additional file 1: Figure S3). For example, five bacterial sequences in the *Pasteurellaceae* family increased in CD, as reported in a previous study [7]. However, *Bacteroidales* comprises of various bacteria, whose increase was offset by their 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 Figure 2a. Third, the co-occurrence of CD-associated organisms in the previous study generally coincided with the correlation network in this study (Additional file 1: Figure S4), despite the use of different samples and methods. For example, both sequence identifiers S0276 (*Fusobacterium periodonticum*) and S0346 (*Veillonella parvula*) were increased in CD and showed possible agreement (Additional file 1: 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 the previous study [7].

Analyses of the microbial community composition

The most frequent sequence identifiers in our data set were S0078 (*B. vulgatus*), S0052 (*Bacteroides fragilis*), S0029 (*E. coli*), S0051 (*Faecalibacterium prausnitzii*), and S0152 (*Bacteroides dorei*) in descending order (Additional file 1: Table S2). S0078, S0052, S0029, and S0152 were the main coefficients in our PCA analysis (Figure 3a), while S0051 was missing due to low standard deviation. Generally, HC samples clustered in the center, slightly toward sequence S0078, in the PCA analysis. PCA results coincided with the heat map results (Figure 3b). Single bacterial sequences did not show high proportions in HC samples (≥ 50%), whereas certain CD and UC samples did show high proportions (≥

50%) of sequence S0029 or S0046. For comparison across disease subtypes, pMANOVA tests were performed. Bacterial community structures in CD (pMANOVA: $P < 0.001$) and UC (pMANOVA: $P < 0.001$) were significantly different from that in HC. Several intraspecies 16S sequences were identified, and their average proportions changed differently according to disease subtypes (Additional file 1: Table S2). Particularly, intraspecies variation of *B. vulgatus*, *F. prausnitzii*, *E. coli*, *Parabacteroides distasonis*, and *Kineothrix alysoides* in average proportion, according to disease subtypes, were prevalent, whereas variation of *B. dorei*, *Haemophilus parainfluenzae*, and *Prevotella copri* were less obvious (Figure 4).

Detection of disease-associated community members

Multiple Mann-Whitney *U* tests were performed using METATEA to identify sequences associated with disease subtypes (Additional file 1: Table S3). Generally, *H. parainfluenzae*, *Dialister pneumosintes*, *F. periodonticum*, and many other sequences were significantly increased in CD. *H. parainfluenzae*-like, *F. prausnitzii*-like, *S. 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. In the phylogenetic tree of sequences with average proportion $\geq 0.1\%$, closely related sequences tended to show similar Z scores (Additional file 1: Figure S5), although it was not so for all similar sequences (Additional file 1: 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 (Additional file 1: Table S3). Some of these aberrant bacteria of putatively external origins are reportedly related to various lethal diseases like endocarditis [9], carcinoma [10], and brain abscess [11]. (Additional file 1: Table S4).

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 (Figure 2b). However, even the highest proportion of S0362 cannot confirm HC status (Additional file 1: Figure S6a), although both maximum and average proportions increased/decreased roughly together in UC (Additional file 1: Figure S6c). 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 Additional file 1: Figure S7a, and high ratio of S0103 (*B. vulgatus*-like) to S0092 (*B. vulgatus*-like) might be related to CD. More studies would warrant the relationship between S0103 and S0092.

Exclusive correlation coefficient and its application

Although PCC can be one of the robust means of finding co-occurring bacteria in terms of positive linear correlation, we found negative linear correlation to not be appropriate for finding mutually exclusive relationships between bacteria, since negative linear correlation is different from mutually exclusive relationship in bacterial proportions (Additional file 1: Table S5). Moreover, zero PCC could not selectively detect mutually exclusive relationships (Additional file 1: Figure S1). Thus, we defined ECC to directly

identify exclusive relationships between bacteria (see the “Methods” section). We successfully detected significantly exclusive relationships 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 (Additional file 1: Table S6,7,8; Additional file 1: Figure S8) and increased in CD than in HC (Figure 5). Particularly, high proportions of S0390 might be related to reduced bacterial diversity (Additional file 1: Figure S9). Relationships between bacterial sequences associated with disease status were explored using ECC and Mann-Whitney *U* test. We regarded five sequences (average proportion ≥ 0.1%) as putative probiotics that occurred at least three times in HC than in CD. Among the five sequences, two were removed by the ECC formula (see Methods), and the relationships between disease-associated sequences and the three sequences, S0208 (*Eubacterium rectale*), S0362 (*R. inulinivorans*), and S0464 (*Clostridium amygdalinum*), were explored using ECC. Generally, the three sequences showed strong exclusive relationships with disease-associated sequences (Additional file 1: Table S9). However, they showed different ECC values depending on the disease-associated sequences (Additional file 1: Table S10). Additionally, certain sequences showed higher ECC values in HC than in CD. For example, ECC value of S0089 (*Bacteroides koreensis*-like) and S0390 in HC was higher than in CD, suggesting putative synergistic effects of disease-associated sequences (Additional file 1: Figure S7b). Additional studies using more clinical data would validate our observation.

Discussion

This study focused on the intraspecies variation in disease risk and relationships between microbes at the strain level. Disease-associated bacteria and their relationships with other members of the community have been identified for decades at species or higher taxonomic levels. PCC and MIC are not optimized for detecting mutually exclusive relationships across the members. For these reasons, we developed METATEA that would enable the analysis of 16S sequences at the identical sequence level and detect mutually exclusive relationships. METATEA has three major advantages: (i) analysis at the identical sequence level outperforms that at higher taxonomic levels; (ii) analysis using ECC easily detects putative probiotic microbes against disease-associated microbes and helps better understanding of the relationships between bacterial diversity and certain bacteria like K. pneumoniae; and (iii) METATEA supports various functions, and its output formats are designed to easily perform further analyses using other statistical tools.

Analyzing microbial communities at the sequence level can have more merits than that at high taxonomic levels. For example, a single taxonomic group is comprised of a variety of strains, and increased strains in certain disease subtypes can be offset by the reduced strains within the same group (Additional file 1: Figure S3). The high-resolution of bacterial populations at the sequence level enabled us to detect many disease-associated microbes, such as V. atypica, P. lini, and H. zavarzinii; IBD could be related to various diseases beyond our expectations (Additional file 1: Table S4), such as endocarditis, carcinoma, and abscess [12–14], not only due to immune dysfunction [15], but also due to various disease-associated microbes. Moreover, disease risk prediction at sequence level outperformed that at OTU (1% and 3% dissimilarity) levels in Fig. 2, probably due to intraspecies variation in the risk of CD

(Fig. 4). Especially, F. prausnitzii is known as a potential novel probiotic bacterium for IBD [16]; however, certain F. prausnitzii-like sequences, such as S0397 and S2816, were increased in CD and UC, respectively (Additional file 1: Table S3). Additionally, the ratios of different strains in the same species might be associated with diseases. In this study, the abnormal ratio between S0092 (B. vulgatus-like) and S0193 (B. vulgatus-like) might be associated with CD (Additional file 1: Figure S7a). The imbalanced proportions of B. vulgatus-like bacteria might activate NF-κB in human gut epithelial cells [17], or inflammation of the gastrointestinal tract might result in imbalance of B. vulgatus-like bacteria; more experimental studies would be required in this direction.

We successfully detected mutually exclusive relationships between bacteria using ECC. ECC can be useful for the development of precision probiotics, since probiotics may have antagonistic effects of different strengths on pathogens. Also, custom probiotic supplements might be needed so as to prevent undesirable side effects. For example, high proportions of S0352 (P. copri-like) and S0362 do not always suggest being healthy (Additional file 1: Figure S6a). However, S0362 might efficiently decrease Haemophilus haemolyticus, while S0464 might efficiently decrease H. parainfluenzae and F. periodonticum (Additional file 1: Table S10). ECC 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. pneumonia, are significantly associated with decreased microbial diversity (Additional file 1: Figure S9), and ectopic colonization of Klebsiella spp. in the intestine drives T helper 1 cell induction and inflammation [18].

K. pneumoniae is a lactose-fermenting bacterium that causes infections especially in alcoholic patients [19]. Moreover, high-alcohol-producing K. pneumoniae is associated with up to 60% of individuals with nonalcoholic fatty liver disease in a Chinese cohort [20]. Either K. pneumoniae is strongly associated with endo-alcohol production and decreased microbial diversity, or intestinal dysbiosis and decreased microbial diversity caused by drinking alcohol are related to increased K. pneumoniae [21]. Studying gut dysbiosis using ECC would help in understanding complex diseases, such as diabetes, depression, and IBD in more details. Additionally, ECC was designed to be less affected by sample sizes. Therefore, ECC differences according to disease subtypes enabled us to detect putative synergistic effects of S0089 (B. koreensis-like) and S0390, although this could also be caused by drinking alcohol.

Conclusions

METATEA is a toolbox optimized for 16S microbiome studies at the sequence level and for the detection of mutually exclusive relationships between bacteria. METATEA does not require time-consuming processes, such as alignment and clustering (Additional file 1: Table S11), since it identifies each unique sequence. Although METATEA does not yet support visualization and a suite of statistical analyses, we designed its output formats to enable easy performance of further analyses using statistical tools. One example is the format of ECC matrix from METATEA. The format was designed to easily perform network analyses using R (Additional file 1: Figure S8). We envision that the combination of METATEA and other

tools would help understanding complex microbial systems, increasing the accuracy of disease risk prediction, and identifying more antagonistic relationships.

Methods

Development of METATEA

We developed METATEA for both 16S analyses at the identical sequence level and detection of mutually exclusive relationships between bacteria. We devised a detailed strategy for 16S analyses at the sequence level without time-consuming processes of traditional taxonomic classification, such as alignment and clustering (Additional file 1: Table S11). METATEA supports various functions, such as basic quality control processes, file format conversion, N correction, N filtering, short read removal, long read trimming, sequence contaminant removal, frequent sequence extraction, unique sequence extraction, sequence proportion calculation, ECC calculation, and statistical analyses, as presented in Figure 1. Various quality control steps were developed to accurately handle reads with Ns and short/long reads at the sequence level. Most ambiguous bases (Ns) from the pyrosequencing platform could be corrected [8] using METATEA. METATEA and 16S sequences used in this study are available at https://sourceforge.net/projects/metatea/.

Public data sets and literature search

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 [7] on IBD in order to develop sequence-supervised processes, assess the performance of METATEA, and validate ECC. Although the symptoms of IBD are heterogeneous [22], the samples were classified into four subtypes based on clinical metadata: HC, CD, IC, and UC. Only the data sets of terminal ileum biopsy (see Additional file 2) 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 [23]. Moreover, in CD risk prediction, terminal ileum outperformed other biopsy locations, such as rectum and stool [7]; CD has been shown to frequently occur in the terminal ileum [24]. To compare the results in this study to that in the previous study [7], taxonomic analysis was performed using data from terminal ileum in Table S2 of the previous study [7]. 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 [25] and IBD is related to various, often lethal, diseases such as cancer [12], endocarditis [13], and neurological disorders [14]. 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 (Additional file 1: Table S1). We found putative erroneous reads to generally occur only once in a single file, and the number of reads with proportions ≥ 0.1% to not be 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 3,204 unique sequences.

Establishment of exclusive correlation coefficient

ECC 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 formulae to investigate mutually exclusive relationships. A single basic formula was selected that would be less affected by group size. Then, we analyzed the data distribution and performed data visualization to modify the formula. 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. Let *N(a)*, *N(b)*, and *N(ab)* be the number of zeros in *a*, *b*, and both *a* and *b*. Too rare sequences were removed by the conditions mentioned below:

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

Finally, ECC was defined as

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

Analyses of 16S rRNA genes

For objective analyses at the sequence level, abnormally short or long reads (< 175 bp or >175 bp) were removed. After BLAST analysis, 9 sequences were identified as human DNA sequences occurring frequently, and were used for the decontamination step using METATEA. After quality control processes, heat map, box plots, and PCA analysis were prepared using MATLAB (v9.4.0.813654; R2018a). 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 (v3.5.1). 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. Only proportions of 86 sequences, 90 OTUs, and 74 OTUs (average proportion ≥ 0.1%) were used for prediction at sequence, 99% similarity, and 97% similarity levels, respectively, and square roots of the sequence/OTU numbers were used to tune mtry parameters. We calculated OTU (1% and 3% dissimilarity) using Mothur (v.1.42.0) with SILVA re-created SEED database (v.132). 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. The maximum likelihood method was selected to construct a phylogenetic tree using MEGA-X (v10.0.5).

Supplemental Information Note

Additional file 1: information

Additional file 1: information accompanies this paper at https://

Additional file 1: Figures S1-9 and Table S1-11. Additional file 1: information. Additional tables and figures referenced for this study.

Additional file 2: Table S12. Sequence/OTU identifiers and their proportions for this study.

Declarations

Acknowledgements

We thank all the members of our laboratory, including Ju Yeon Song, Jae Kyung Yoon, and Jongseok Kim, for critical and useful discussion.

Authors' contributions

SS developed METATEA and performed the analyses. JFK contributed to designing the analysis pipeline and supervised the study. All authors read and approved the final version of the manuscript.

Funding

Funding for this work was provided through the National Research Foundation (NRF-2014M3C9A3068822) funded by the Ministry of Science and ICT, Republic of Korea. Publication was supported in part by the Brain Korea 21 PLUS program, Ministry of Education, Republic of Korea.

Availability of data and materials

METATEA and 3,204 unique sequences for this study are available in the SourceForge repository (https://sourceforge.net/projects/metatea). Our data sets are shown in Additional Information and were obtained from BioProject (https://www.ncbi.nlm.nih.gov/bioproject/319632).

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1] Department of Systems Biology, Division of Life Sciences, and Institute for Life Science and Biotechnology, Yonsei University, 50 Yonsei-ro, Seodaemun-gu, Seoul 03722, Republic of Korea. [2] Strategic Initiative for Microbiomes in Agriculture and Food, Yonsei University, Seoul 03722, Republic of Korea.

References

1. Greenblum S, Carr R, Borenstein E. Extensive strain-level copy-number variation across human gut microbiome species. Cell. 2015;160:583-94.
2. Feng P. *Escherichia coli* serotype O157:H7: novel vehicles of infection and emergence of phenotypic variants. Emerg Infect Dis. 1995;1:47-52.
3. O'Toole PW, Cooney JC. Probiotic bacteria influence the composition and function of the intestinal microbiota. Interdiscip Perspect Infect Dis. 2008;2008:175285.
4. Reshef DN, Reshef YA, Finucane HK, Grossman SR, McVean G, Turnbaugh PJ, et al. Detecting novel associations in large data sets. Science. 2011;334:1518-24.
5. Piewngam P, Zheng Y, Nguyen TH, Dickey SW, Joo HS, Villaruz AE, et al. Pathogen elimination by probiotic *Bacillus* via signalling interference. Nature. 2018;562:532-7.
6. Spinler JK, Taweechotipatr M, Rognerud CL, Ou CN, Tumwasorn S, Versalovic J. Human-derived probiotic *Lactobacillus reuteri* demonstrate antimicrobial activities targeting diverse enteric bacterial pathogens. Anaerobe. 2008;14:166-71.
7. Gevers D, Kugathasan S, Denson LA, Vazquez-Baeza Y, Van Treuren W, Ren B, et al. The treatment-naive microbiome in new-onset Crohn's disease. Cell Host Microbe. 2014;15:382-92.
8. Shin S, Park J. Correction of sequence-dependent ambiguous bases (Ns) from the 454 pyrosequencing system. Nucleic Acids Res. 2014;42:e51.
9. Chunn CJ, Jones SR, McCutchan JA, Young EJ, Gilbert DN. *Haemophilus parainfluenzae* infective endocarditis. Medicine (Baltimore). 1977;56:99-113.
10. Castellarin M, Warren RL, Freeman JD, Dreolini L, Krzywinski M, Strauss J, et al. *Fusobacterium nucleatum* infection is prevalent in human colorectal carcinoma. Genome Res. 2012;22:299-306.
11. Rousee JM, Bermond D, Piemont Y, Tournoud C, Heller R, Kehrli P, et al. *Dialister pneumosintes* associated with human brain abscesses. J Clin Microbiol. 2002;40:3871-3.
12. Axelrad JE, Lichtiger S, Yajnik V. Inflammatory bowel disease and cancer: The role of inflammation, immunosuppression, and cancer treatment. World J Gastroenterol. 2016;22:4794-801.
13. Kreuzpaintner G, Horstkotte D, Heyll A, Losse B, Strohmeyer G: Increased risk of bacterial endocarditis in inflammatory bowel disease. Am J Med. 1992;92:391-5.
14. Casella G, Tontini GE, Bassotti G, Pastorelli L, Villanacci V, Spina L, et al. Neurological disorders and inflammatory bowel diseases. World J Gastroenterol. 2014;20:8764-82.
15. Neuman MG. Immune dysfunction in inflammatory bowel disease. Transl Res. 2007;149:173-86.

16. Miquel S, Martin R, Rossi O, Bermudez-Humaran LG, Chatel JM, Sokol H, et al. *Faecalibacterium prausnitzii* and human intestinal health. Curr Opin Microbiol. 2013;16:255-61.
17. P OC, de Wouters T, Giri R, Mondot S, Smith WJ, Blottiere HM, 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. 2017;47:209-17.
18. Atarashi K, Suda W, Luo C, Kawaguchi T, Motoo I, Narushima S, et al. Ectopic colonization of oral bacteria in the intestine drives TH1 cell induction and inflammation. Science. 2017;358:359-65.
19. Prince SE, Dominger KA, Cunha BA, Klein NC. *Klebsiella pneumoniae* pneumonia. Heart Lung. 1997;26:413-7.
20. Yuan J, Chen C, Cui J, Lu J, Yan C, Wei X, et al. Fatty Liver Disease Caused by High-Alcohol-Producing *Klebsiella pneumoniae*. Cell Metab. 2019;30:675-88.
21. Samuelson DR, Shellito JE, Maffei VJ, Tague ED, Campagna SR, Blanchard EE, et al. Alcohol-associated intestinal dysbiosis impairs pulmonary host defense against *Klebsiella pneumoniae*. PLoS Pathog. 2017;13:e1006426.
22. Lamb CA, Kennedy NA, Raine T, Hendy PA, Smith PJ, Limdi JK, et al. British Society of Gastroenterology consensus guidelines on the management of inflammatory bowel disease in adults. Gut. 2019;68:s1-106.
23. Quigley EM, Abu-Shanab A. Small intestinal bacterial overgrowth. Infect Dis Clin North Am. 2010;24:943-59.
24. Caprilli R. Why does Crohn's disease usually occur in terminal ileum? J Crohns Colitis. 2008;2:352-6.
25. Han YW, Wang X. Mobile microbiome: oral bacteria in extra-oral infections and inflammation. J Dent Res. 2013;92:485-91.

Figures

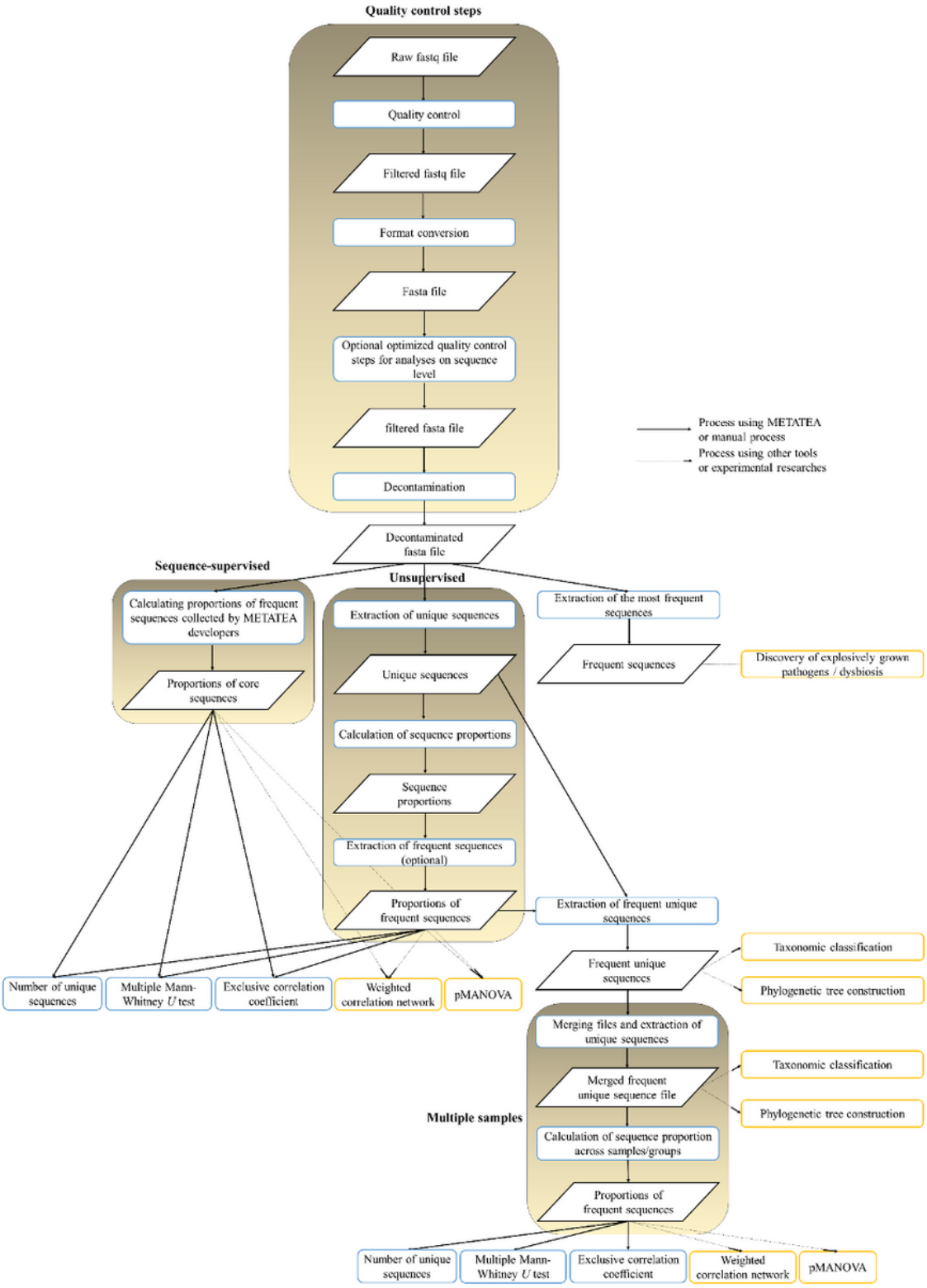

Figure 2

Simplified standard workflow diagram of METATEA. Blue rounded rectangles denote processes using METATEA. Yellow rounded rectangles denote processes using other tools.

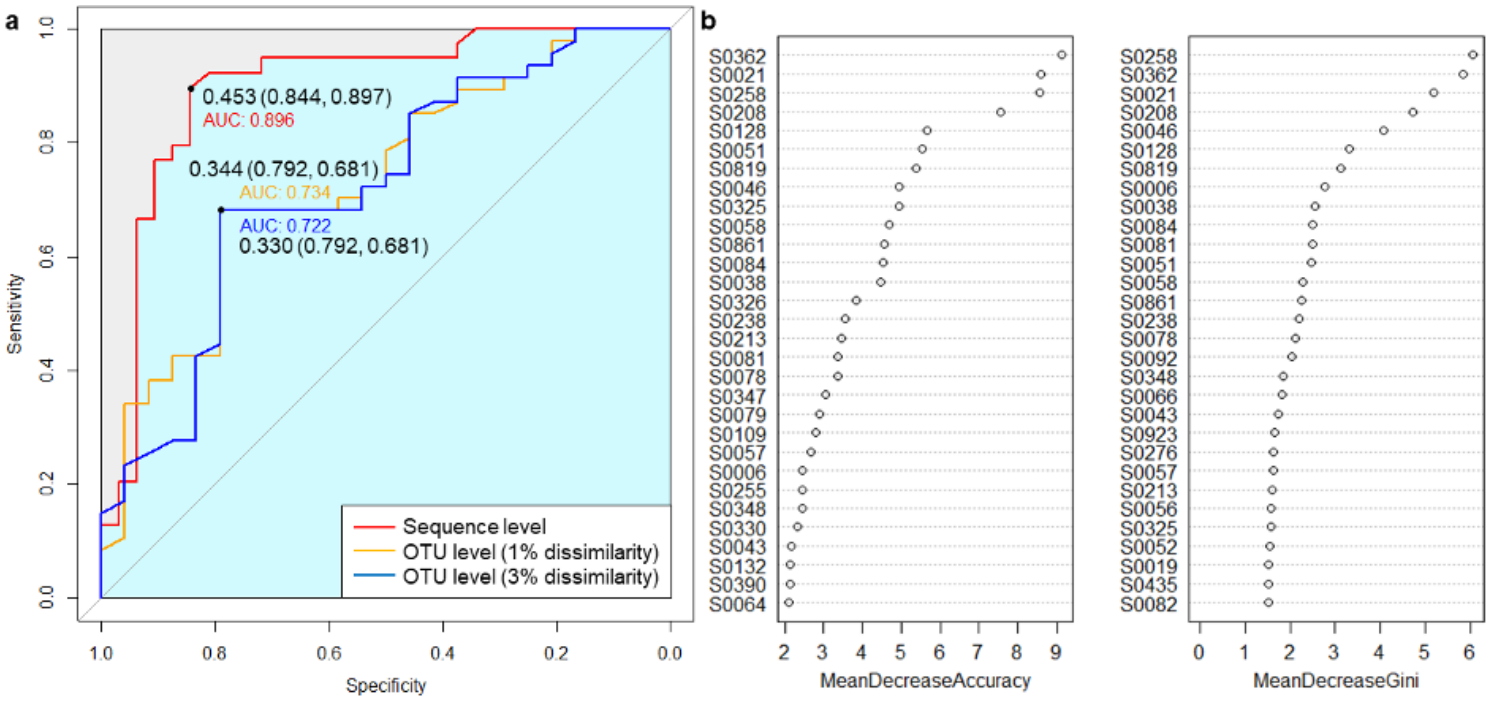

Figure 4

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, The variable importance plot shows the most important microbial sequences.

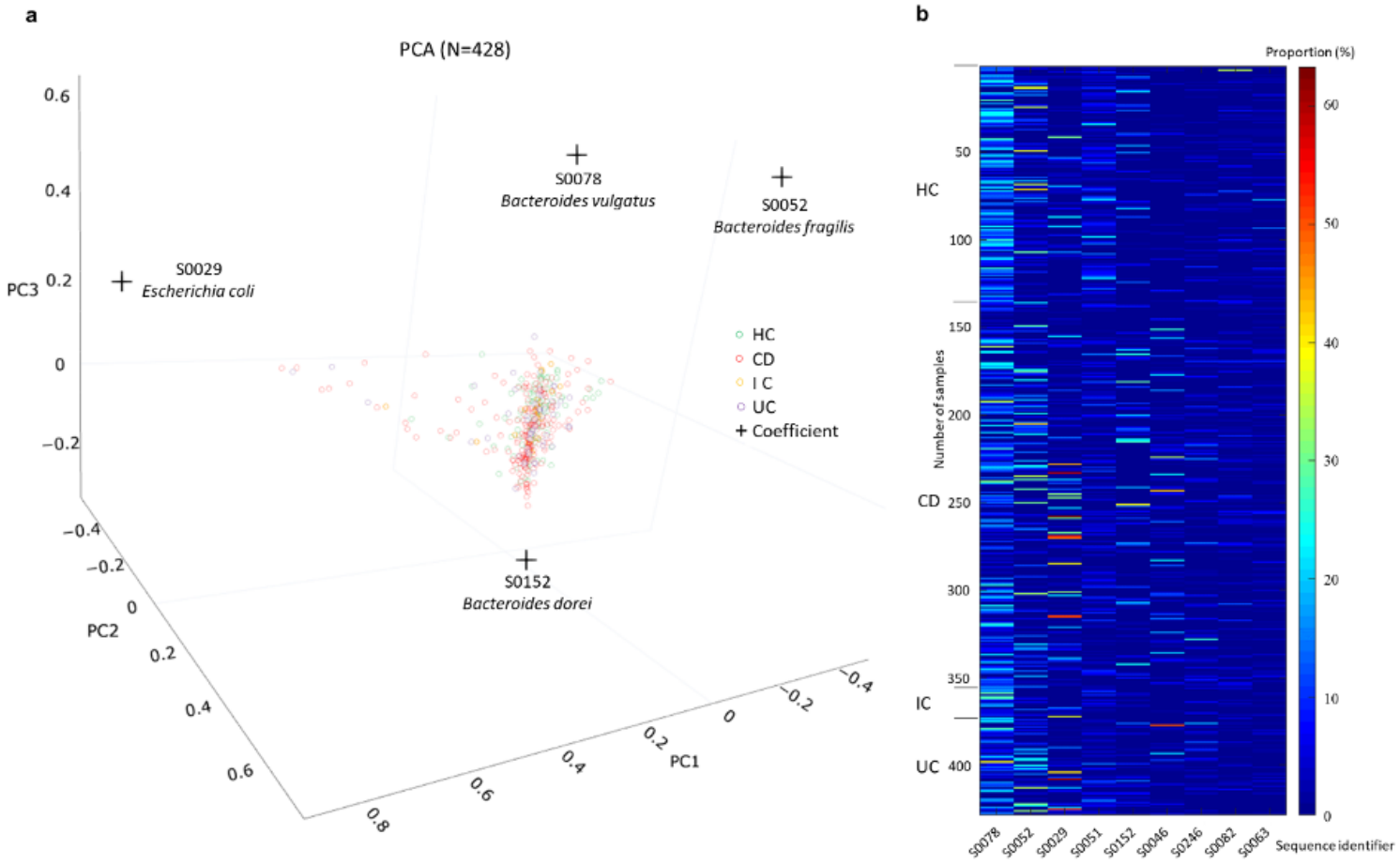

Figure 6

PCA and heat map of 428 samples. a, PCA. Four most highly correlated sequences (coefficients) with their principal components are shown. b, Heat map. Nine sequences of high average proportions (≥ 1%) are selectively shown. CD, Crohn's disease; HC, Healthy control; IC, Indeterminate colitis; UC, Ulcerative colitis.

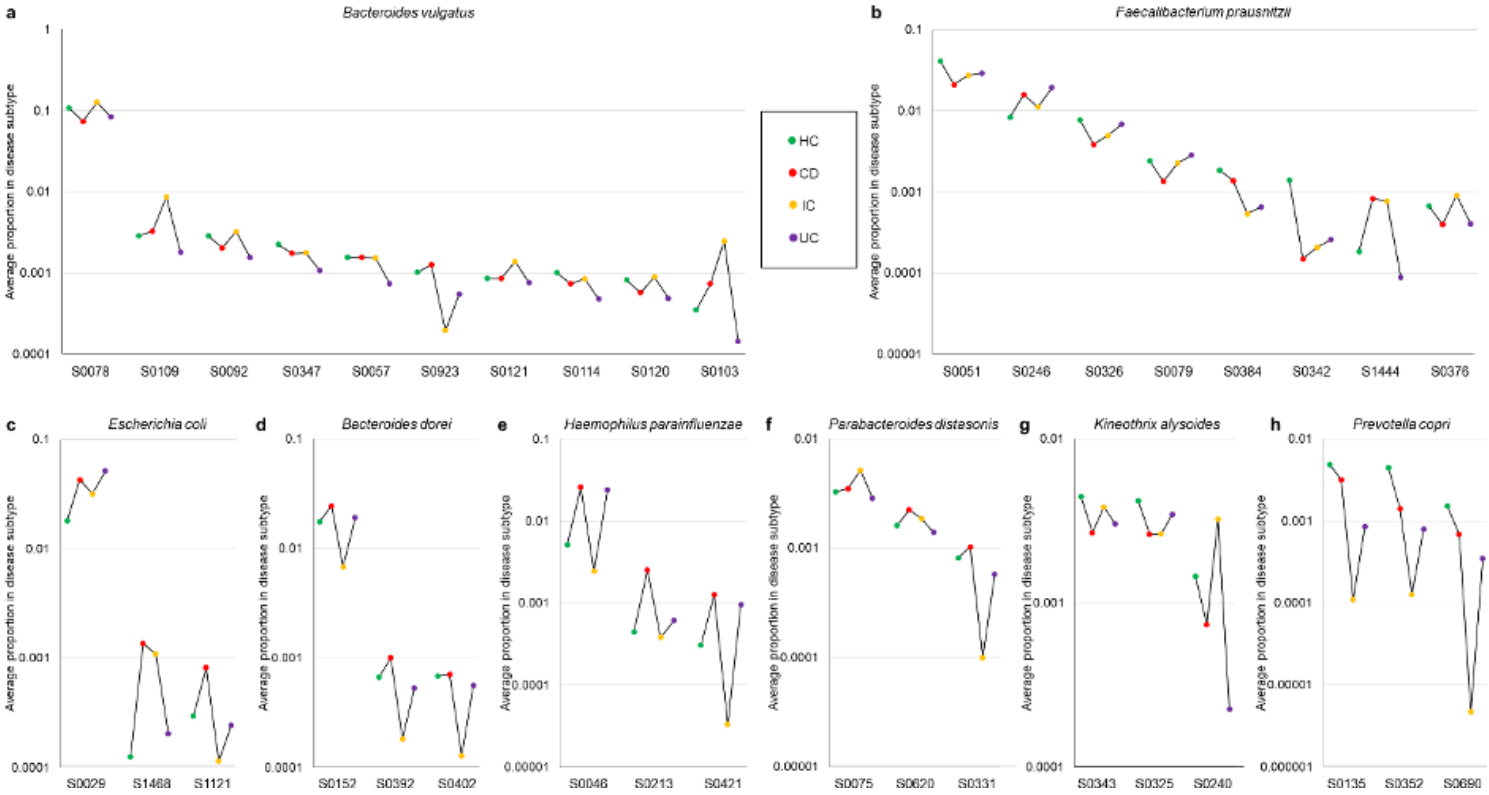

Figure 8

Examples of intraspecies variation. Average proportions of microbial sequences in disease subtypes (average proportion ≥ 0.05%). a, Bacteroides vulgatus and vulgatus-like sequences. b, Faecalibacterium prausnitzii and prausnitzii-like sequences. c, Escherichia coli and coli-like sequences. d, Bacteroides dorei and dorei-like sequences. e, Haemophilus parainfluenzae and parainfluenzae-like sequences. f, Parabacteroides distasonis and distasonis-like sequences. g, Kineothrix alysoides and alysoides-like sequences. h, Prevotella copri and copri-like sequences. Green, red, orange, and purple circles designate average proportions in HC, CD, IC, and UC, respectively. CD, Crohn's disease; HC, Healthy control; IC, Indeterminate colitis; UC, Ulcerative colitis.

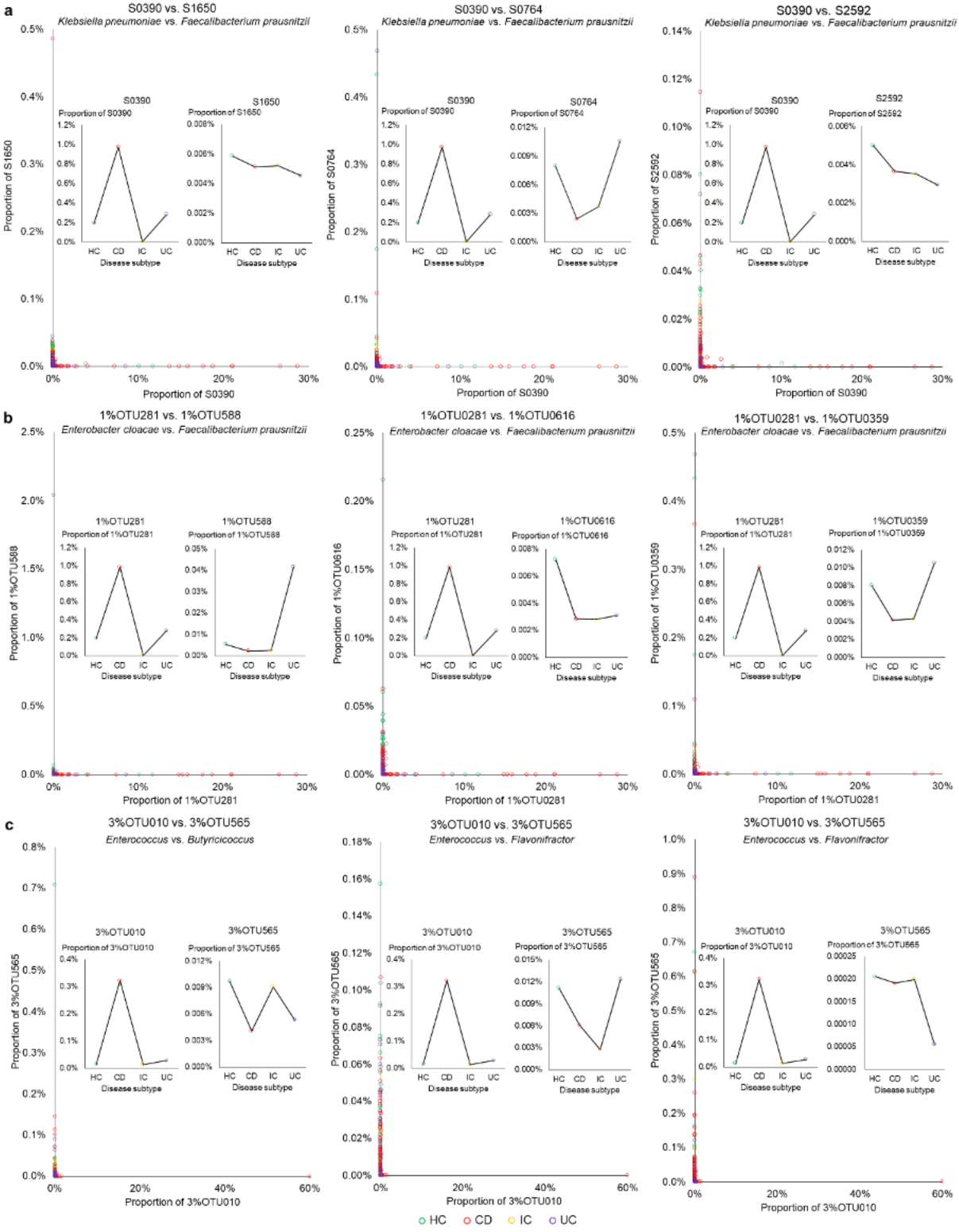

Figure 10

Examples of highly exclusive relationship. 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.

- Additionalfile11.85.docx
- Additionalfile11.85.docx
- Additionalfile2.xlsx
- Additionalfile2.xlsx