RESEARCH ARTICLE

Computational approach to modeling microbiome landscapes associated with chronic human disease progression

Lu Li[1], Jiho Sohn[2], Robert J. Genco[3,4], Jean Wactawski-Wende[5], Steve Goodison[6], Patricia I. Diaz[3,4]*, Yijun Sun[1,7]*

1 Department of Computer Science and Engineering, University at Buffalo, The State University of New York, Buffalo, New York, United States of America, **2** Department of Medicine, University at Buffalo, The State University of New York, Buffalo, New York, United States of America, **3** Department of Oral Biology, University at Buffalo, The State University of New York, Buffalo, New York, United States of America, **4** UB Microbiome Center, University at Buffalo, The State University of New York, Buffalo, New York, United States of America, **5** Department of Epidemiology and Environmental Health, University at Buffalo, The State University of New York, Buffalo, New York, United States of America, **6** Department of Quantitative Health Sciences, Mayo Clinic, Jacksonville, Florida, United States of America, **7** Department of Microbiology and Immunology, University at Buffalo, The State University of New York, Buffalo, New York, United States of America

* pidiazmo@buffalo.edu (PID); yijunsun@buffalo.edu (YS)

Citation: Li L, Sohn J, Genco RJ, Wactawski-Wende J, Goodison S, Diaz PI, et al. (2022) Computational approach to modeling microbiome landscapes associated with chronic human disease progression. PLoS Comput Biol 18(8): e1010373. https://doi.org/10.1371/journal.pcbi.1010373

Editor: Nicola Segata, University of Trento, ITALY

Received: January 18, 2022

Accepted: July 11, 2022

Published: August 4, 2022

Data Availability Statement: The software and preprocessed data are available at www.acsu.buffalo.edu/~yijunsun/lab/MicroDynamics.html. The software is also available via Zenodo (DOI: 10.5281/zenodo.6524575), under the terms of the MIT License. The original microbiome sequence data and patient clinical data can be obtained from the QIITA open-source microbial study management platform (https://qiita.ucsd.edu) under study ID 1629 and EBI accession no. ERP020401.

Abstract

A microbial community is a dynamic system undergoing constant change in response to internal and external stimuli. These changes can have significant implications for human health. However, due to the difficulty in obtaining longitudinal samples, the study of the dynamic relationship between the microbiome and human health remains a challenge. Here, we introduce a novel computational strategy that uses massive cross-sectional sample data to model microbiome landscapes associated with chronic disease development. The strategy is based on the rationale that each static sample provides a snapshot of the disease process, and if the number of samples is sufficiently large, the footprints of individual samples populate progression trajectories, which enables us to recover disease progression paths along a microbiome landscape by using computational approaches. To demonstrate the validity of the proposed strategy, we developed a bioinformatics pipeline and applied it to a gut microbiome dataset available from a Crohn's disease study. Our analysis resulted in one of the first working models of microbial progression for Crohn's disease. We performed a series of interrogations to validate the constructed model. Our analysis suggested that the model recapitulated the longitudinal progression of microbial dysbiosis during the known clinical trajectory of Crohn's disease. By overcoming restrictions associated with complex longitudinal sampling, the proposed strategy can provide valuable insights into the role of the microbiome in the pathogenesis of chronic disease and facilitate the shift of the field from descriptive research to mechanistic studies.

Funding: This work is supported in part by R01AI125982 (YS), R01DE024523 (JWW), and R01CA241123 (YS, SG), all funded by NIH (https://www.nih.gov/). The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Author summary

The delineation of system dynamics of a microbial community can provide a wealth of insights into the roles of the microbiome in the pathogenesis of chronic disease. However, due to the difficulty in obtaining longitudinal samples, most existing microbiome studies have been cross-sectional and largely descriptive. Here, we present a novel computational strategy that leverages massive static sample data to model microbiome landscapes associated with chronic disease development. To demonstrate the validity of the proposed strategy, we applied it to a gut microbiome dataset available from a Crohn's disease study and constructed one of the first microbial progression models of the disease. We performed a series of interrogations on the constructed model. Our analysis suggested that the constructed model recapitulated the longitudinal progression of microbial dysbiosis during the known clinical trajectory of Crohn's disease. By overcoming the sampling restrictions inherent to slowly progressive diseases, our approach is potentially widely applicable in many different studies across body sites, diseases, and other conditions.

This is a *PLOS Computational Biology* Methods paper.

Introduction

The human microbiome—trillions of microbes residing in and on human bodies—plays an essential role in many important physiological processes. Studies that are part of the Human Microbiome Project and others have significantly expanded our knowledge of the human microbiota and its implications for human health [1–5]. However, most microbiome studies performed to date have been cross-sectional, using single time-point data to examine the potential role of the microbiome in human health. While cross-sectional studies are a logical first step, these analyses are largely descriptive and provide little information about microbial community dynamics with respect to disease development. A possible way to elucidate system dynamics in this context is to assemble time-series data through repeated sampling of the same cohort of subjects across a defined disease process. This could provide a wealth of insights into pathogenesis that is unattainable through a static experimental design. However, due to economical and logistical constraints, time-course studies have generally been limited by the number of samples examined and the time period followed, and consequently data collected may only cover a partial picture of microbial dynamics [6–9]. This is particularly true when studying chronic diseases (i.e., Crohn's disease or periodontitis), the development of which can take decades. Consequently, it has been difficult to study microbial community dynamics and their possible contribution to the initiation and progression of human chronic diseases.

As cost-effective DNA sequencing technology continues to advance, large-scale epidemiological studies are providing access to data from many thousands of microbiome samples. This provides us with a unique opportunity to develop an analytical strategy that uses massive cross-sectional data, instead of time-course data, to study microbial community dynamics in disease. The strategy is based on the rationale that each static sample provides a snapshot of the disease process, and if the number of samples is sufficiently large, the footprints of individual samples populate progression trajectories, which in turn enables the recovery of microbial community dynamics by using computational approaches. To demonstrate the validity of the proposed strategy, we developed a bioinformatics pipeline and applied it to a gut microbiome dataset available from a Crohn's disease (CD) study [7]. CD is a chronic inflammatory disease characterized by discontinuous lesions that can affect the entire gastrointestinal tract. It tends to start in the teens and twenties, though it can occur at any age [10], and as there are no curative interventions currently available [11], it is considered a lifelong illness. At diagnosis, most patients present with a clinical inflammatory behavior, but stricturing or penetrating complications develop as the disease progresses [12, 13]. The extent of CD lesions also changes overtime, initially involving either the ileum or the colon and later progressing to the ileocolonic region [12]. Previous metagenomics studies suggested that CD results from aberrant immune responses to the intestinal microbiota [14, 15], but details of how microbiome shifts initiate or promote disease development, progression and symptom exacerbation are lacking. Our analysis using the developed bioinformatics pipeline revealed a double bifurcating model of microbial alterations that occur during disease development. The constructed model was validated by aligning it with clinical and molecular traits. The analysis suggested that the identified microbiome trajectories reflected changes in CD behavior, location and severity associated with disease progression. To further demonstrate the utility of the model, we projected the samples onto the identified progression paths to form pseudo-time series data and performed a series of analyses to characterize dysbiosis and microbiome functional shifts, to infer microbial interactions, and to identify key bacteria associated with CD development. By overcoming the sampling restrictions inherent to slowly progressive diseases, our approach provides a novel way to study microbial community dynamics associated with human chronic diseases.

Results

Overview of the developed bioinformatics pipeline

Fig 1 presents the flowchart of the proposed bioinformatics pipeline for microbial community dynamics analysis. Briefly, given a table of operational taxonomic units (OTUs) that summarizes the microbiome compositions of individuals, either healthy or presenting with different stages of a disease, we first perform feature selection to identify disease-related microorganisms. Then, by using the relative abundances of the selected microorganisms, we perform clustering analysis to group samples with homogenous microbial compositions and conduct embedded structure learning to construct a principal tree to mathematically describe the dynamic changes in microbial compositions associated with disease development. Finally, by using the principal tree as a backbone, we combine the principal-tree and clustering results to build a microbial progression model. See Methods for details. The software and user manual of the proposed bioinformatics pipeline are freely available at www.acsu.buffalo.edu/~yijunsun/lab/MicroDynamics.html.

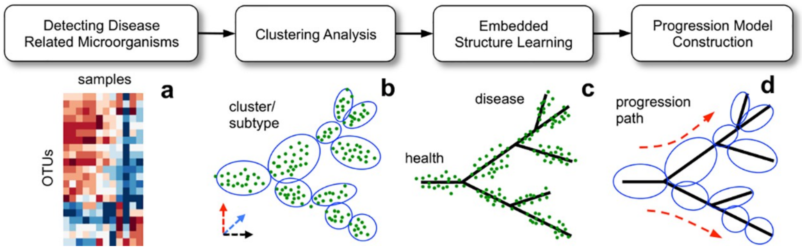
Detecting Disease Related Microorganisms
Clustering Analysis
Embedded Structure Learning
Progression Model Construction
samples
OTUs
a
cluster/ subtype
b
disease
health
c
progression path
d

Fig 1. Overview of the proposed bioinformatics pipeline for microbial community dynamics analysis. The pipeline offers an integrated suite of computational tools that allow researchers to identify disease-related microorganisms, stratify samples into clinically relevant subtypes, construct disease progression models, and delineate disease-specific community dynamics at both organism and functional levels.

https://doi.org/10.1371/journal.pcbi.1010373.g001

Constructing a microbial progression model of Crohn's disease

To demonstrate the utility of the proposed bioinformatics pipeline, we applied it to a human gut microbiome dataset obtained from a Crohn's disease study [7]. The dataset contains 312 microbiome samples collected from 49 CD patients and 9 healthy controls (HCs). The disease duration of the CD patients ranged from recent onset to 58 years. By using the Montreal classification system [16], each patient was stratified into one of three disease phenotypes: inflammatory (B1), stricturing (B2), and penetrating (B3). Patients were also classified by bowel lesion location as colonic CD (cCD), ileal CD (iCD), or ileocolonic CD (icCD), and by whether they had undergone resective surgery. See S1 Table for the summary of the study cohort and S2 Table for the detailed clinical information. For each individual, a fecal sample was collected every three months for up to two years, and the V4 hyper-variable region of the 16S rRNA gene was PCR amplified and sequenced, resulting in 90,456,980 reads with an average length of 98 bps. We used the QIIME pipeline [17] for data pre-processing and OTU table construction. Since a sample with an insufficient sequencing depth may not enable accurate estimation of microbial composition, we excluded 37 samples with less than 10^4 reads from downstream analysis (S1 Fig). By grouping the sequences into OTUs at the 3% distance level, we obtained a total of 77,286 species-level OTUs. See Methods for details.

Since only a small fraction of microorganisms are likely to be involved in disease development, the first step toward progression modeling is to identify disease-related microorganisms. We formulated it as a feature-selection problem for supervised learning and used the disease phenotypes that reflect disease severity as class labels to detect relevant microorganisms. For the purpose of the study, the LOGO algorithm [18] was employed (see Methods). This is one of the most competitive feature-selection algorithms derived to date, with excellent accuracy and computational efficiency. Since the disease progression is defined as the development of B2 or B3 in patients with B1 at diagnosis [12] and there were only 7 patients diagnosed as B3, we combined the samples in the B2 and B3 groups, forming a three-class supervised-learning problem (i.e., HC, B1, and B2/B3). The parameters of LOGO were estimated through ten-fold cross-validation (S2(A) Fig). By applying a cutoff of 0.001 to the obtained feature weights, a total of 172 OTUs were detected to be related to disease development (S2(B) Fig).

By using the relative abundance data of the identified OTUs, we next performed a clustering analysis to detect sample groups with homogenous microbial compositions. To this end, the *k*-means method [19] was employed. The number of clusters was estimated to be five by gap statistic [20] (Fig 2A). It is known that using *k*-means may result in a local optimal solution [21]. To obtain a stable and robust clustering assignment, a resampling-based consensus clustering analysis [22] was performed, where *k*-means clustering was repeated 1,000 times and in each time 80% samples were drawn randomly without replacement from the entire dataset. The results of the 1,000 runs were aggregated into a consensus matrix, providing a visual representation of the frequency of two samples being grouped into the same cluster. From the consensus matrix, we can clearly identify five blocks along the anti-diagonal line (Fig 2B), suggesting a reliable data partition. To further assess the clustering robustness, the silhouette width of each sample was calculated, which is defined as the difference between its average similarity with samples in the same cluster and the largest average similarity with samples in different clusters [23]. A cluster with an average silhouette width larger than 0 is generally considered stable. In our analysis, 255 of 275 (93%) samples had a positive silhouette width, the average silhouette widths of the five clusters ranged from 0.06 to 0.18, and the average silhouette width of all the samples equaled to 0.1, demonstrating the stability of the detected clusters (Fig 2C).

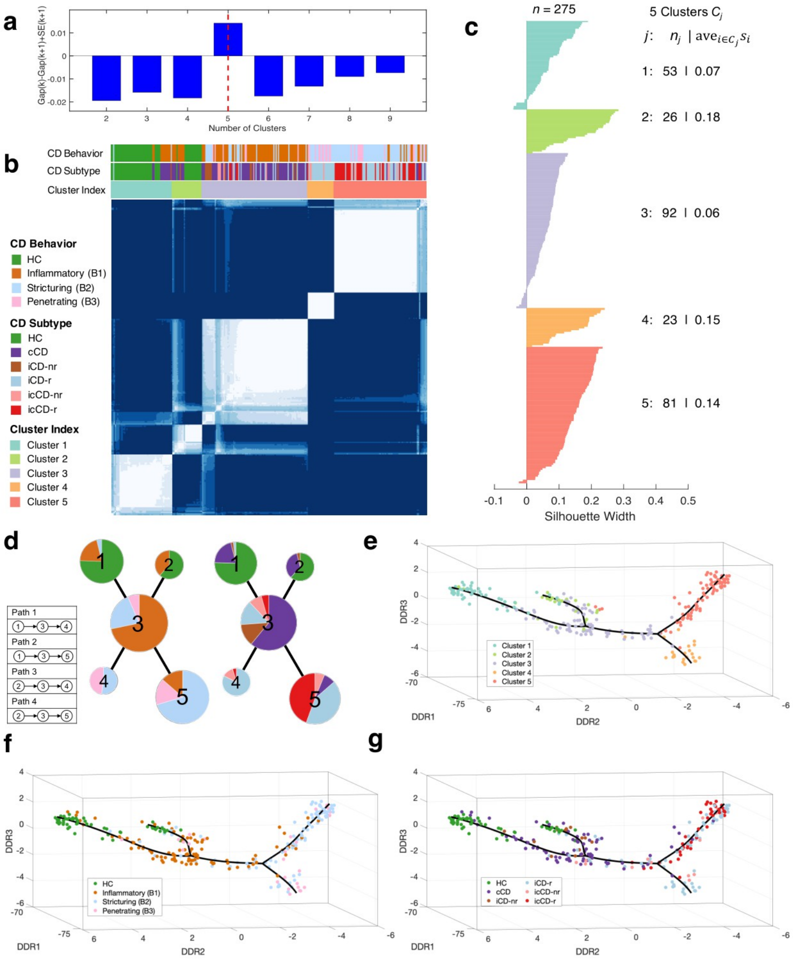

Fig 2. Microbial community dynamic analysis performed on a human gut microbiome dataset (*n* = 275) obtained from a Crohn's disease study. (**a**) The number of clusters was estimated to be five by gap statistic. (**b**) Resampling-based consensus clustering analysis identified five robust and stable clusters. (**c**) Silhouette width analysis further confirmed the robustness of clustering assignment. A total of 255 of 275 (93%) samples had a positive silhouette width, and the average was equal to 0.1. (**d**) By combining the principal-tree and clustering results, a microbial progression model of Crohn's disease was constructed, and four progression paths were identified. Each node represents an identified cluster, and the pie chart in each node depicts the percentage of the samples in the node having one of the CD behaviors (left panel) or belonging to one of the CD subtypes (right panel). (**e**-**g**) Visualization analysis provided a general view of sample distribution supported by the selected microorganisms. Each point represents a sample, which was projected onto a three-dimensional space by using the DDRTree method. Each sample was color-coded by its cluster index (**e**), CD behavior (**f**), and CD subtype (**g**), respectively. The solid line represents the constructed principal tree. HC: healthy control, cCD: colonic Crohn's disease, iCD: ileal Crohn's disease, icCD: ilealcolonic Crohn's disease, r/nr: with/without ileocaecal resection.

https://doi.org/10.1371/journal.pcbi.1010373.g002

After we grouped samples with similar microbial compositions, we next performed an embedded structure learning to construct a principal tree to mathematically describe microbial dynamics and to infer the potential progression relationships of the detected clusters. For the purpose of the study, our recently developed DDRTree algorithm [24] was employed. The basic idea is to fit a given dataset by using a minimum spanning tree with a bounded length (Fig 1C). The method can *automatically* determine the number and presence of branches and is robust against noise, rendering it particularly suitable to detect a complex tree structure hidden in a high-dimensional space. See Methods for details. We used the elbow method [25] to tune the parameters of the DDRTree method (S3 Fig).

Finally, by using the constructed principal tree as a backbone, we combined the clustering and principal-tree results to build a microbial progression model of CD (Fig 2D). We present the constructed model as an undirected graph, where each node represents a cluster, and the node size is proportional to the number of samples in the corresponding cluster. An edge connecting two nodes indicates a possible progressive relationship, and the length of the edge is proportional to the distance of the curve connecting the centers of the two nodes. The pie chart in each node depicts the percentage of the samples in the node having one of the CD behaviors or belonging to one of the CD subtypes. Our modeling analysis revealed a double bifurcating structure with four potential microbiome progression paths, starting from two distinct health-associated clusters and evolving toward two disease endpoints (Fig 2D).

We performed a series of interrogations that provided support for the constructed model. Fig 2E–2G presents the data distribution in a three-dimensional space learned by the DDRTree method. To help with visualization and to put the result into context by referring to previous studies, each sample was color-coded by its cluster index, CD behavior or CD subtype, respectively. We noticed that the overall structure of the constructed model is consistent with the data visualization result (i.e., double bifurcating), suggesting that the model faithfully reflected the data distribution. As mentioned above, changes in CD behavior are part of the natural course of CD, with the disease progression being defined as patients shifting from inflammatory (B1) to a complex behavior (either B2 or B3) [12, 13, 26]. Notably, the identified progression paths accurately reflected the changes in CD behaviors associated with disease progression. As shown in Fig 2D and 2F, the constructed model starts from two health-associated clusters, converges to Cluster 3 that is dominated by samples with an inflammatory behavior, and finally diverges to Clusters 4 and 5 that consist primarily of samples with stricturing and penetrating behaviors. It is worth noting that samples with various CD behaviors (e.g., B1) are present in nearly all the detected clusters (Fig 2D). This is possibly due to the fact that microbial compositions of CD patients are highly unstable and can be heavily influenced by various factors such as dietary changes and medications (discussed below). Changes in bowel lesion location have also been documented to occur during long-term follow-up of CD patients, with initial presentations localized to colon or ileum only and eventually involving both locations [12]. Notably, our microbial progression model captured changes in lesion location (Fig 2D and 2G). Specifically, Cluster 3 was composed mostly of patients with involvement in a single location (cCD and iCD), and Cluster 5 had a large proportion of patients with ileocolonic involvement. This suggests that there are microbial shifts associated with location changes. Our progression model also captured increased disease severity, as measured by the proportion of patients who underwent a resective procedure. Specifically, while there are only 18.8% patients in Cluster 3 having a resection, the proportions are increased to 87.0% and 88.6% for Clusters 4 and 5, respectively (Fig 2D). Taken together, the above results suggest that the constructed model recapitulates the natural history of CD and can provide a representation of the longitudinal progression of dysbiosis during the clinical trajectory of the disease.

Changes in microbiome alpha diversity along identified progression paths

Having depicted the general trend of microbiome shifts during CD development, we assessed the changes in microbiome alpha diversity along the four identified progression paths. We observed that the alpha diversity, as measured by Chao1 [27] and Shannon metrics [28], decreased significantly along each progression path (Fig 3A). We found that Cluster 3, comprised mainly of samples from B1—the early stage of the disease, showed a significantly lower alpha diversity compared with HC Cluster 1 (p-value < 0.001, ANOVA test), but not with HC Cluster 2 (Fig 3B). Furthermore, compared with the HC groups and Cluster 3, the clusters of

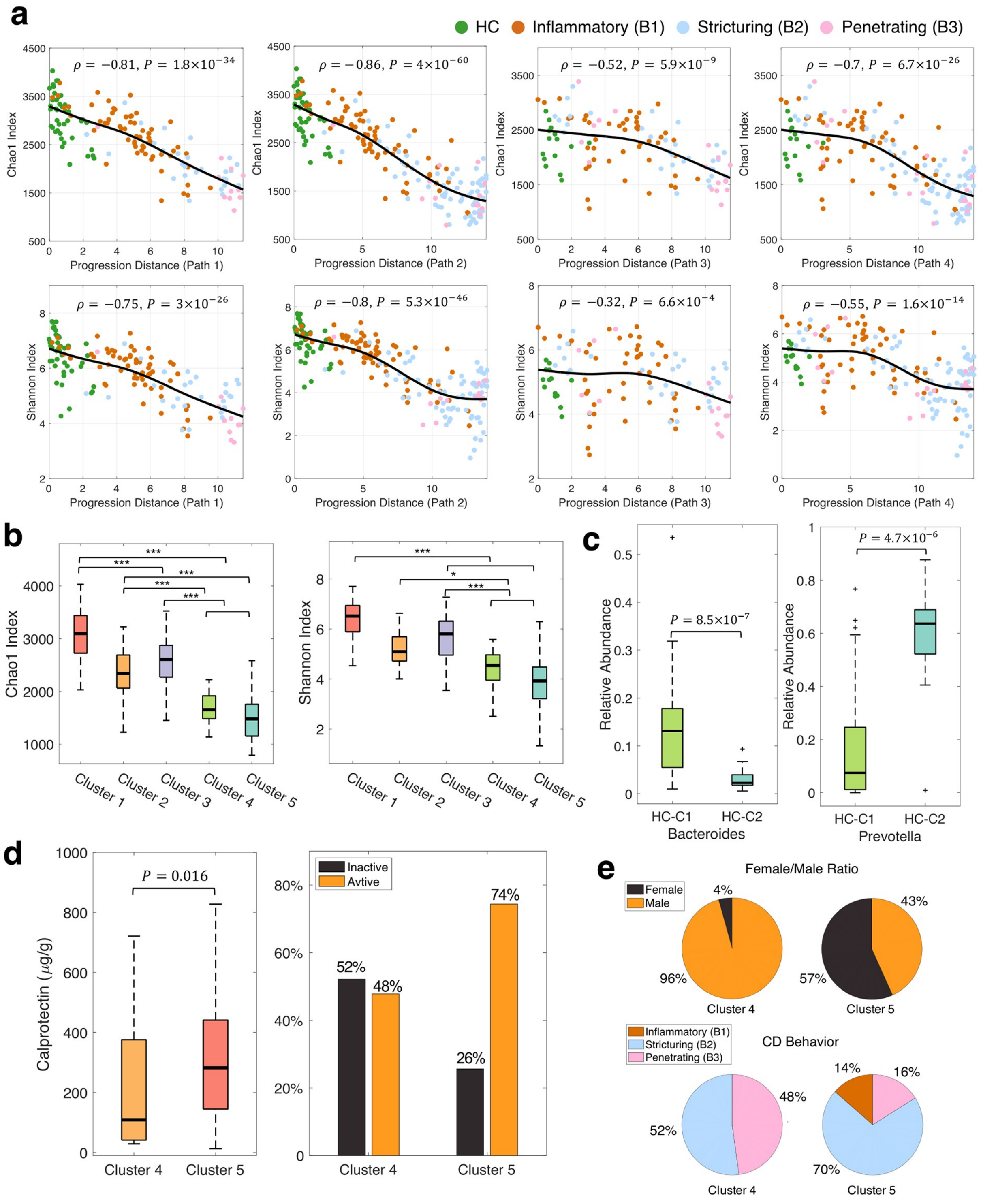

Fig 3. Changes in microbiome alpha diversity along identified progression paths and clinical characteristics of healthy states and disease endpoints. (**a**) Spearman's rank correlation analysis of alpha diversity as measured by Chao1 index and Shannon index along the four identified progression paths (see Fig 2D). To aid in visualization, each sample was annotated by its clinical behavior. (**b**) Comparison of alpha diversity of five detected clusters. The asterisks indicate the levels of significance determined by ANOVA. *: p-value < 0.05, **: p-value < 0.01, ***: p-value < 0.001. Also see S3 Table. (**c**) Enterotype analysis of the HC samples in Cluster 1 (HC-C1) and Cluster 2 (HC-C2). HC-C1 and HC-C2 correspond to the enterotypes driven by *Bacteroides* and *Prevotella*, respectively. (**d**-**e**) Comparison of clinical characteristics of patients in two disease endpoints (i.e., Clusters 4 and 5). Cluster 5 contained a significantly higher proportion of patients with active inflammation (fecal calprotectin >150 μg/g) compared with Cluster 4 (p-value = 0.016, χ^2 test). Clusters 4 and 5 exhibited significantly different female-to-male ratios and CD behavior compositions (p-value < 0.01, χ^2 test).

https://doi.org/10.1371/journal.pcbi.1010373.g003

the late-stage CD (Clusters 4 and 5) exhibited a significant reduction in alpha diversity (p-value < 0.05, ANOVA test). However, the difference in alpha diversity between Clusters 4 and 5 was not significant. The above results are consistent with the observations from prior studies [29–31] that reported a reduction in microbiome diversity with progressive CD, which provides further support for the validity of the constructed model.

Characteristics of healthy and disease endpoints

Since the modeled microbiome progression started from two distinct health-associated clusters and evolved toward two disease endpoints, we next evaluated the clinical characteristics of the starting and terminal microbial states. Several studies have demonstrated that human gut microbiota of healthy individuals can be stratified into enterotypes, which are associated with long-term diet and mainly driven by the abundances of *Bacteroides* (termed as ET-B) or *Prevotella* (termed as ET-P) [32, 33]. Notably, in our study, we observed that the HC samples were grouped into two clusters corresponding to ET-B and ET-P, respectively (Fig 3C), and the two clusters converged to the CD-associated Cluster 3 (Fig 2D). This suggests that CD may have two distinct disease origins depending on the enterotypes of individual patients. We noticed that the relative abundance of *Prevotella* in ET-P is much higher than that observed in the original enterotype paper [33]. The discrepancy may be explained by the differences in individual diets, lifestyles, and countries of origin, as reported in [34–36].

The constructed model also depicted two disease endpoints (i.e., Clusters 4 and 5) with distinct clinical characteristics. Specifically, Cluster 5 exhibited a significantly higher proportion of samples showing active inflammation compared with Cluster 4 (p-value = 0.016, χ^2 test), as measured by fecal calprotectin $> 150 \mu g/g$ [37, 38] (Fig 3D). We also observed a gender difference in the two clusters with a female-to-male ratio of 1.31 in Cluster 5 and 0.05 in Cluster 4 (p-value $\leq 8.2 \times 10^{-6}$, χ^2 test, Fig 3E). Moreover, there was a significant difference between the two clusters in terms of CD behavior (p-value $\leq 2.7 \times 10^{-3}$, χ^2 test, Fig 3E). Specifically, Cluster 5 was dominated by B2 cases—patients with stricturing CD. In contrast, patients with either stricture or penetration were dominant in Cluster 4. To rule out the possibility that the difference in inflammation levels between Clusters 4 and 5 was related to disease behavior, we compared fecal calprotectin levels of patients with the same disease behavior in the two clusters. Our analysis showed that within each CD behavior Cluster 5 always demonstrated a higher portion of samples with active inflammation (S4 Fig) and thus Cluster 5 can be considered as a more severe phenotype. In summary, our analysis suggests that there may be two forms of late-stage CD, with different microbiome compositions, inflammation severities, CD behaviors and subtypes (Fig 2D).

Characterizing overall dysbiotic shifts during CD progression

In the progression modeling analysis, we performed supervised learning to detect disease-related microorganisms. However, due to the use of the ℓ_1 regularization (see Eq (6)), if multiple microorganisms had similar microbial profiles across samples, only one microorganism was retained (that is, we intended to construct a parsimonious model to minimize the chance of overfitting). To comprehensively search for disease-related microorganisms, we performed a Spearman's rank test to detect OTUs that showed significant changes in relative abundance along at least one identified progression path. We used the DS-FDR method [39] to control the false discovery rate and filtered out OTUs with average relative abundance < 0.001 and Spearman's rank correlation coefficient $|\rho| < 0.3$ (i.e., those with a weak or no correlation). At an FDR of 0.01, a total of 90 species-level OTUs were retained (Fig 4 and S5 Fig).

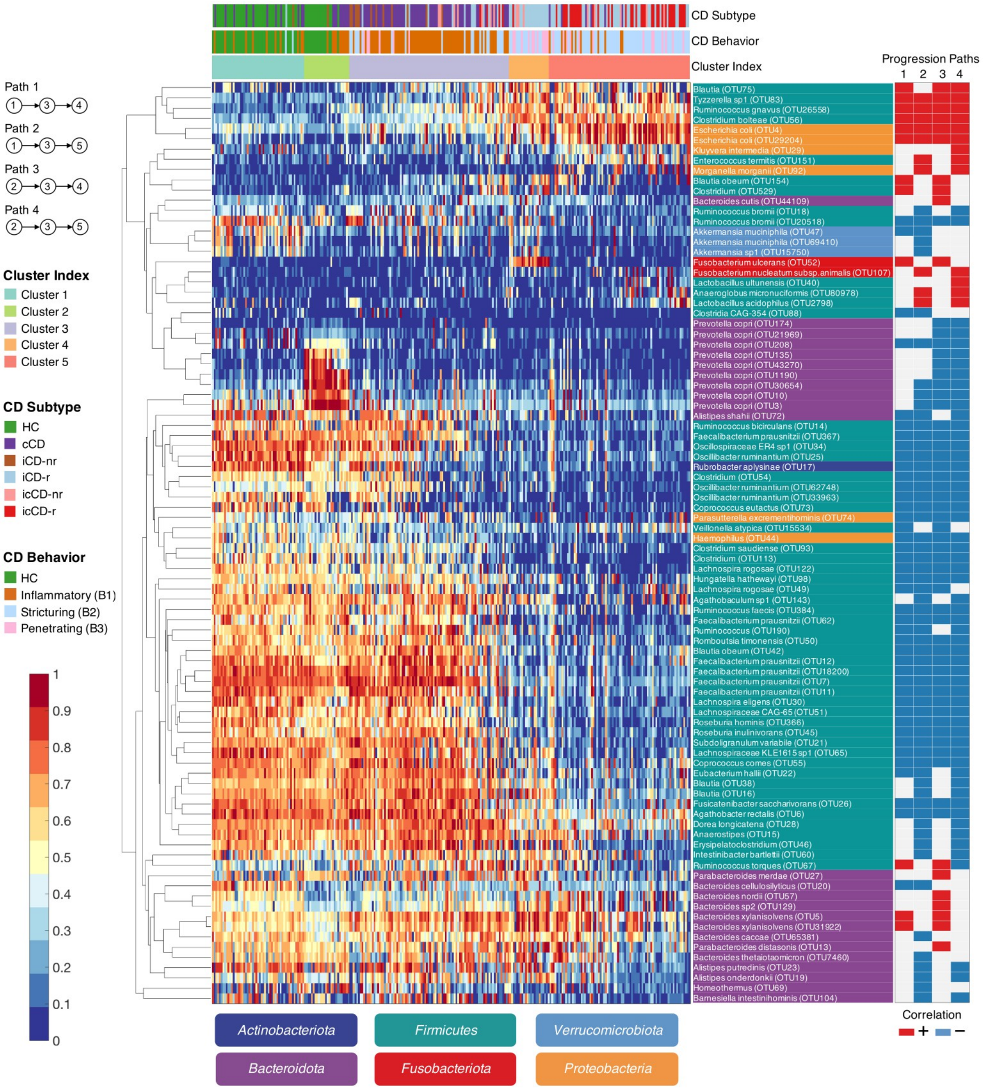

Fig 4. Heatmap of microorganisms for which the relative abundances were detected to be highly correlated with at least one of the four identified CD progression paths. Each row represents an OTU, and each column represents a sample. The samples were first ordered by cluster labels and then by progression distances. For the purpose of visualization, the relative abundance of each OTU was log-transformed and scaled into the range of [0, 1]. See S5 Fig for additional details.

https://doi.org/10.1371/journal.pcbi.1010373.g004

Consistent with previous findings [7, 40, 41], we observed an overall decrease of beneficial bacteria including *Faecalibacterium prausnitzii*, *Roseburia*, *Subdoligranulum* and *Lachnospiraceae*, as well as members of *Ruminococcus* and *Oscillospiraceae* as disease severity progressed (S6 Fig). By producing butyric acid, the beneficial bacteria, such as *F. prausnitzii*, may protect the host by up-regulating anti-inflammatory cytokines [42]. Thus, the reduction of these clades may impair the ability of the host to repair the epithelium and regulate inflammation. In contrast, the relative abundances of pro-inflammatory bacteria, including *Escherichia coli* and *Ruminococcus gnavus*, were significantly increased along the disease progression paths (S7 Fig). Our data confirms previous findings that suggested that *E. coli* and *R. gnavus* may play a role in CD development [43]. In addition, OTUs classified as

Tyzzerella sp. and *Clostridium bolteae* were found by our pipeline to be associated with CD progression through all paths (S7 Fig).

Importantly, our analysis also detected path-specific microbial variations. A decrease in *Prevotella copri* was associated with the path starting from ET-P, while *Clostridia CAG-354*, *Bacteroides cellulosilyticus* and *Akkermansia muciniphila* decreased with disease progression from ET-B. Notably, *Ruminococcus torques*, which is known to degrade gastrointestinal mucin [44, 45] and is more frequently found in relatives of CD patients compared with healthy individuals [46], was positively correlated with the disease progression paths leading to Cluster 4. *Fusobacterium ulcerans*, which has been previously isolated from skin ulcers, was also increased in a specific manner along progression paths 1 and 3 leading to Cluster 4. Microbial changes leading to Cluster 5—a severe disease status—included a decrease in the beneficial gut commensal *Anaerostipes*, which may protect against colon cancer by producing butyric acid [47], and an increase in the oral commensals *Fusobacterium nucleatum subsp. animalis*, *Lactobacillus acidophilus* and *Anaeroglobus micronuciformes* along both paths 2 and 4, leading to Cluster 5.

Inferring microbial interaction networks associated with CD progression

Once a microbial progression model was constructed, we projected each sample back onto the identified progression paths. Here, the projection of a sample was defined as a point on a progression path that is closest to the sample. By using the healthy controls as the baseline, the static samples were ordered along a path according to the extent to which the disease progressed from an inflammatory phenotype toward intestinal stricture and penetration (S8 Fig). The ordered samples can be viewed as *pseudo-time series* data, which provides a unique opportunity to perform a microbial interaction network analysis to identify key bacteria potentially responsible for the alterations of microbiota associated with disease development. In this study, we used the generalized Lotka-Volterra (gLV) method [48, 49] to infer pairwise interactions between microorganisms (see Methods). Following the work of [50], we evaluated the influence of each microorganism (or node) affecting others on the network by its out-degree—the number of edges directed out of the node. We found that a decrease in *Prevotella copri* was an important event associated with disease development along all paths (Fig 5 and S9 Fig). Our analysis also revealed that a decrease in *Parasutterella excrementihominis* and *Veillonella atypica* and an increase in *E. coli* are key events in the progression toward Cluster 4, while the progression leading to Cluster 5 appears to be primarily driven by an increase in *F. nucleatum subsp. animalis*.

Characterizing functional shifts associated with CD progression

The constructed progression model also enabled us to investigate how the shifts in functional potential of the microbiome were associated with disease development. To this end, we applied PICRUSt2 [51] to predict the functional content of microbial communities and performed a Spearman's rank correlation test to identify KEGG pathways that showed significant changes in pathway activities along at least one progression path (see Methods). As with the dysbiosis analysis, we employed the DS-FDR method [39] to control the false discovery rate and filtered out the functional pathways with average relative abundance < 0.001 and Spearman's rank correlation coefficient $|\rho| < 0.3$. At an FDR of 0.01, a total of 101 KEGG pathways were identified (S10 and S11 Figs and S4 Table). Consistent with previous studies [52–55], we found that the activities of pathways such as galactose metabolism, pentose and glucuronate interconversions, sulfur metabolism, glyoxylate and dicarboxylate metabolism, nitrogen metabolism, and phenylalanine metabolism were significantly increased with CD severity. Conversely, pathways

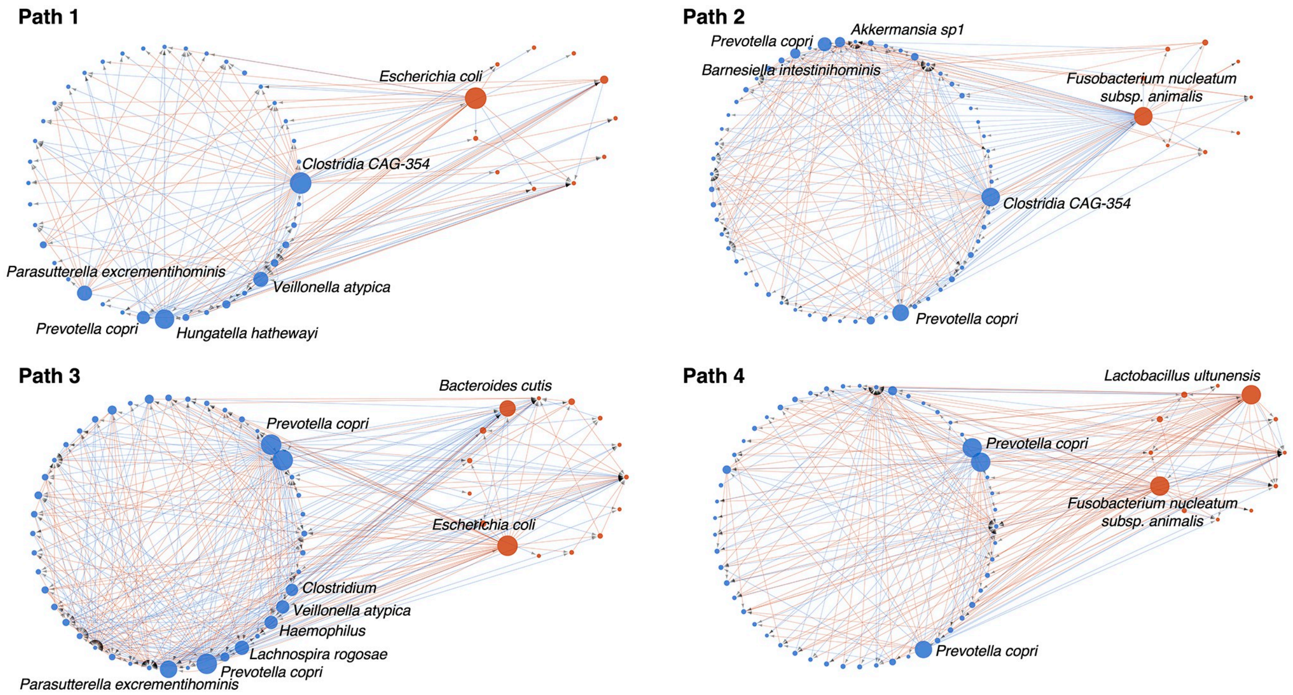

Fig 5. Microbial interaction networks inferred by the gLV method applied to pseudo-time series data recovered from four identified CD progression paths. Each node represents an OTU, its size is proportional to the number of edges directed out of the node (i.e., out-degree), and its face color represents the sign of the correlation of the relative abundance of the OTU with a progression path (red: positive, blue: negative). Only the nodes with out-degrees larger than 10 were annotated. Since compositionality was not considered in the analysis, artificial links might arise. See S9 Fig for detailed annotations.

https://doi.org/10.1371/journal.pcbi.1010373.g005

including fatty acid biosynthesis, phenylalanine, tyrosine and tryptophan biosynthesis, and D-glutamine and D-glutamate metabolism were negatively correlated with the disease progression paths. Through comparative analysis of pathway activities along different progression trajectories, we found that progression paths leading to Cluster 4 were associated with decreased amino acid and nucleotide metabolism along with increased metabolism of several carbohydrates, glycan degradation and primary and secondary bile acid biosynthesis, while the progression paths that end at Cluster 5 were linked to a decline of antimicrobial biosynthesis and an enrichment in glutathione metabolism, and xylene and dioxin degradation. Path 4 in particular was associated with an increase in two-component systems, ABC transporters, the phosphotransferase system, and the butyrate and propionate metabolic pathways. These results highlight the ability of the proposed bioinformatics pipeline to identify microbiome functional shifts along disease progression paths toward distinct disease phenotypes.

Longitudinal microbiome shifts in individual subjects during disease progression

The study cohort was collected from participants every three months for up to two years, which provided us with an opportunity to examine variations in the microbiome of individual patients across the sampling period. Since it is not reliable to estimate variations using a small number of samples, we excluded from the analysis the individuals with < 5 serial samples. In total, data from 34 participants were examined. Since both ET-B and ET-P can be used as the disease origin, the progression distance of a sample could vary depending on the origin used. To address this issue, we scaled the curve distance between Clusters 2 and 3 (i.e., the shorter

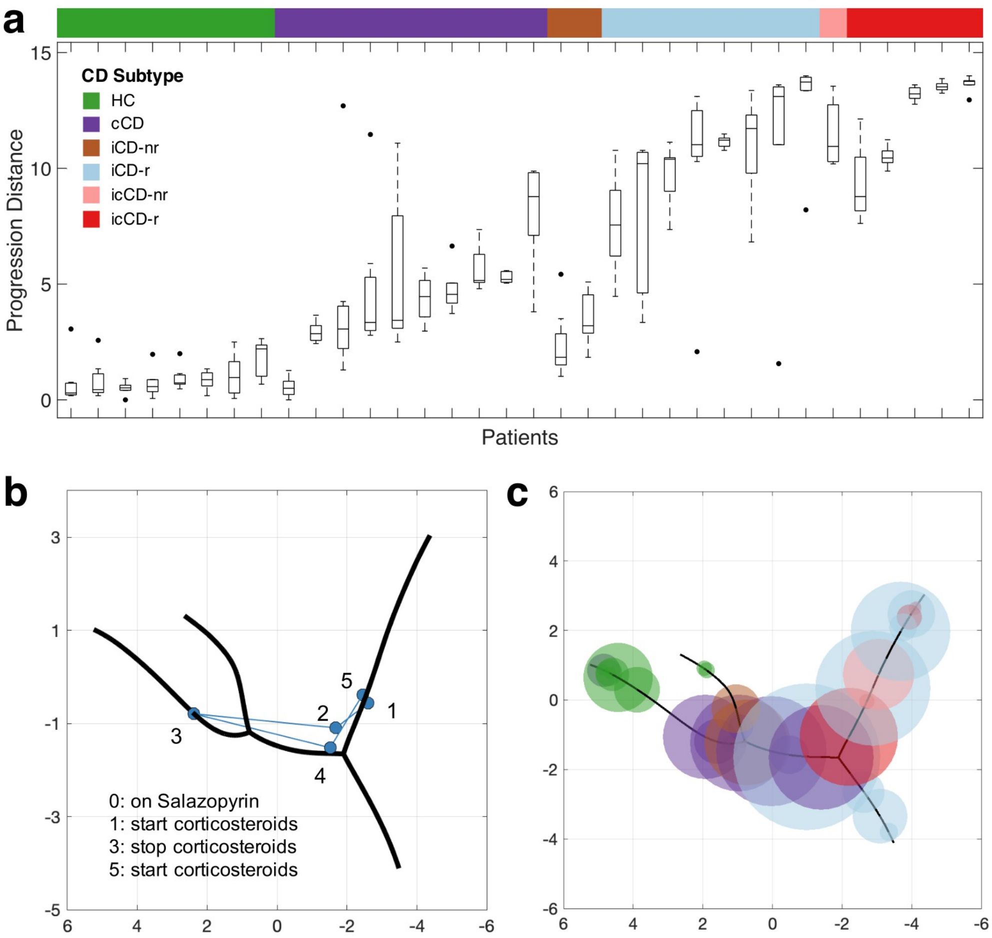

Fig 6. Microbial community dynamics analysis of individual patients. (**a**) The progression distances of the samples collected from individual participants over a two-year period. The participants were first ordered by CD subtypes and then by median progression distances. (**b**) The microbiome composition of a patient was significantly altered by medication. Sample 0 contained only a few hundreds of reads and thus was omitted. (**c**) Samples collected from a two-year study provided only a partial picture of microbial community dynamics associated with disease development. Each circle represents a patient, the face color represents the CD subtype, and the radius equals 1.5 MAD of the progression distances of the samples collected from the patient. MAD: median absolute deviation.

https://doi.org/10.1371/journal.pcbi.1010373.g006

branch) by a constant of 1.66 in downstream analysis so that the calculation of the progression distance of a given sample is independent of the disease origin used. Our analysis showed that CD samples significantly deviated from HCs (p-value < 0.001, Student's t-test, Fig 6A). Specifically, the icCD-r cases attained the largest progression distances, in line with reported clinical severity, followed by icCD-nr and iCD-r. We can also see that the microbial communities of healthy individuals are much more stable than those of CD patients. In addition to physiological conditions, other factors (e.g., dietary changes and medications) can also significantly alter the human gut microbiome. For example, in a patient who was diagnosed with cCD and placed on corticosteroids, we found that in the period of receiving the medication the gut microbiome moved backward along the progression path toward the healthy microbiome, but after corticosteroids were stopped, the microbiome moved forward along the progression path toward the initial microbial status, at which point corticosteroids were again administered (Fig 6B). This result indicates that the constructed progression model correlates well with the clinical trajectory of CD and supports the potential of such tools to monitor treatment responses and disease remission.

To further assess the sample variation of each patient, we computed the median absolute deviation (MAD) of the progression distances of the samples from each patient (1.5 MAD is a robust measure of one standard deviation) and mapped the samples back onto the progression model (Fig 6C). We found that the radii of the circles representing CD patients ranged from 0.13 to 2.13 (average: 0.83), which equals to 5.9%–9.1% of the total length of the progression path, respectively. This analysis shows that while the gut microbiome of CD patients is highly unstable and influenced by various factors, the samples collected from a two-year longitudinal study provided *only* a partial picture of the progressive clinical course of CD. This underscores the importance of the development of novel approaches, such as the bioinformatics pipeline proposed in this study, to overcome the sampling limitations that impede longitudinal studies of microbiome-related chronic diseases.

Discussion

As with any biological system, a microbial community is a dynamic system undergoing constant change in response to internal and external stimuli. The composition of the human gut microbiota, for example, can be modulated by the introduction or extinction of particular microbial groups, or by a change in population structure caused by various factors [56]. In turn, such changes can have significant implications for human health [57]. The delineation of system dynamics of a microbial community can provide a wealth of insights not accessible through a static experiment, and lay a critical foundation for the development of probiotic, prebiotic, antibiotic, and other strategies to manipulate the microbiome. However, due to the difficulty in obtaining longitudinal samples, most existing microbiome studies have been cross-sectional and largely descriptive. Here, we present a novel computational strategy that leverages massive static sample data to study microbial dynamics associated with chronic human disease development. We applied the developed pipeline to a Crohn's disease microbiome dataset and constructed one of the first microbial progression models of the disease. Our analysis revealed that CD may have two disease origins depending on the enterotypes of individual patients, and two disease endpoints with distinct clinical characteristics and microbial compositions. Since there is currently no established progression model for comparison, model validation poses a challenge. Our strategy was to align the model with established clinical and molecular traits. Our analysis suggested that the constructed model recapitulated the longitudinal progression of microbial dysbiosis during the known clinical trajectory of CD.

This study has several limitations worth discussing. It has been reported that the compositional nature of microbiome data could induce biases in data analysis [58–60]. While there are several strategies (e.g., modeling with data after centered log-ratio (CLR) transformation) that can be used to alleviate the issue [59], the analysis of compositional microbiome data remains a challenge [60]. Due to the lack of microbial biomass data of samples, our approach assumed that there was no significant variation in absolute abundances between samples. Thus, the compositionality was not factored into the modeling. Out of interest, we performed an experiment using CLR transformed data for modeling, and we observed a similar double bifurcating structure. However, we should point out that in the microbial interaction network analysis, artificial links might arise since compositionality was not considered. In the software package, we provided users with an option to use CLR transformed data for the proposed analysis. As more data becomes available, we will perform in-depth analysis to assess potential biases that compositional data might introduce into a model. In this study, we used the *k*-means method to detect patient groups with homogenous microbial compositions. While *k*-means is one of the most widely methods for clustering analysis, there are several other methods that might be more suitable for microbiome data sets (e.g., Dirichlet multinomial mixtures [61] and partition around medoids [62]). Another limitation of the study is that the model was derived from a dataset with a relatively small sample size. When larger datasets become available, the development of robust models encompassing the microbiome variability across individuals and populations will become possible. In this study, we used the sequence data obtained from the V4 hyper-variable region of the 16S rRNA gene for OTU table construction, which may not provide sufficient taxonomic resolution at the species level and can affect model resolution. A possible way to address the issue is to use whole metagenome or full-length 16S rRNA gene sequence data to estimate microbial compositions, which could significantly refine constructed models. We should emphasize that the constructed model ultimately needs to be verified through wet-lab experiments. However, interrogation of a computational model will allow researchers to generate and test novel hypotheses *in silico* and help to prioritize resources and inform focused and detailed investigations experimentally.

We expect that our approach will find wide applications. Although here we focused on Crohn's disease, the approach can be used to study other microbe-related chronic diseases, where the lack of longitudinal data is a *ubiquitous* problem. Compared to resource-intensive or impractical time-course studies, it is much easier to conduct a cross-sectional study; researchers only need to be concerned with recruiting patients presenting various stages of a disease, and the recent development of sequencing technology has already made large-scale sequencing projects feasible. The application of our approach to large cross-sectional populations will significantly advance our understanding of microbiome dynamics during chronic disease development and help to identify novel diagnostic and therapeutic strategies.

Materials and methods

Data pre-processing and OTU table construction

We used the QIIME pipeline (v1.9.0) [17] for data pre-processing and OTU table construction. Specifically, we first removed low-quality reads by filtering out sequences that contained ambiguous bases, had a Phred quality score less than 3, or had more than three consecutive low-quality base calls. A total of 90,456,980 sequences were retained for further analysis. Then, we performed a taxonomy-independent analysis using USEARCH [63] to group sequences into OTUs at the 3% distance level, and removed chimeric sequences using UCHIME [64]. We calculated the relative abundance of each sample by dividing the number of reads in each OTU by the total read counts in the sample. We added a small constant 10^{-6} to the relative abundances and performed a 10-base logarithmic transformation [65]. We performed the taxonomy annotation of each OTU by using BLAST [66] against the Genome Taxonomy Database (v86) [67], and conducted a functional analysis by using PICRUSt2 [51]. The functional analysis yielded 8,602 KEGG orthologies, which were grouped into 204 KEGG pathways [68] by MinPath [69].

Bioinformatics pipeline for microbial community dynamics analysis

Feature selection to identify disease related microorganisms. We used the LOGO algorithm [18] to identify disease-related microorganisms. It represents one of the most competitive feature-selection algorithms derived to date, with excellent accuracy and computational efficiency. The basic idea is to decompose a complex nonlinear problem into a set of locally linear ones through local learning, and then learn feature relevance globally within the large margin framework. Below, we present a detailed description of the method.

Let $\{(\mathbf{x}_n, y_n)\}_{n=1}^{N}$ be a dataset, where $\mathbf{x}_n$ is the n-th data sample and y_n is the corresponding label. We aim to select a subset of features so that the class labels of unseen samples can be correctly predicted. We start by defining the margin of $\mathbf{x}_n$. Given a distance function, we find two nearest neighbors of $\mathbf{x}_n$, one from the same class (called nearest hit or NH), and the other from a different class (called nearest miss or NM). The margin of $\mathbf{x}_n$ is defined as $\rho_n = d(\mathbf{x}_n, \mathrm{NM}(\mathbf{x}_n)) - d(\mathbf{x}_n, \mathrm{NH}(\mathbf{x}_n))$, where $d(\cdot)$ is a distance function. In this study, we used the Manhattan distance to define the margin, while other distance function can also be used. By the large margin theory [70], a classifier that minimizes a margin-based error function usually generalizes well on unseen test data. Let $\mathbf{w} \geq 0$ be a feature weight vector, where the magnitude of each element represents the relevance of the corresponding feature. Our goal is to find a weighted subspace specified by $\mathbf{w}$ so that a margin-based error function in the induced space is minimized.

The margin of $\mathbf{x}_n$, computed with respect to $\mathbf{w}$, is given by

$$\rho_n(\mathbf{w}) = d(\mathbf{x}_n, \mathrm{NM}(\mathbf{x}_n)|\mathbf{w}) - d(\mathbf{x}_n, \mathrm{NH}(\mathbf{x}_n)|\mathbf{w}) \triangleq \mathbf{w}^T\mathbf{z}_n \ , \tag{1}$$

where $\mathbf{z}_n = |\mathbf{x}_n - \mathrm{NM}(\mathbf{x}_n)| - |\mathbf{x}_n - \mathrm{NH}(\mathbf{x}_n)|$, and $|\cdot|$ is an element-wise absolute operator. A major issue with the above margin definition is that the nearest neighbors of a given sample are unknown before learning. To account for the uncertainty in defining local information, we develop a probabilistic model, where the nearest neighbors of a given sample are treated as *hidden* variables. Following the principles of the expectation-maximization algorithm [71], we estimate the margin by computing the expectation of $\rho_n(\mathbf{w})$ via averaging out the hidden variables:

$$\rho_n(\mathbf{w}) = \mathbf{w}^T(\textstyle\sum_{i\in\mathcal{M}_n} P(\mathbf{x}_i = \mathrm{NM}(\mathbf{x}_n)|\mathbf{w})|\mathbf{x}_n - \mathbf{x}_i| - \sum_{i\in\mathcal{H}_n} P(\mathbf{x}_i = \mathrm{NH}(\mathbf{x}_n)|\mathbf{w})|\mathbf{x}_n - \mathbf{x}_i|) \triangleq \mathbf{w}^T\bar{\mathbf{z}}_n \ , \tag{2}$$

where $\mathcal{M}_n = \{i : 1 \leq i \leq N, y_i \neq y_n\}$, $\mathcal{H}_n = \{i : 1 \leq i \leq N, y_i = y_n\}$, and $P(\mathbf{x}_i = \mathrm{NM}(\mathbf{x}_n)|\mathbf{w})$ and $P(\mathbf{x}_i = \mathrm{NH}(\mathbf{x}_n)|\mathbf{w})$ are the probabilities of sample $\mathbf{x}_i$ being the nearest miss or hit of $\mathbf{x}_n$, respectively. The probabilities are estimated via the standard kernel density estimation:

$$P(\mathbf{x}_i = \mathrm{NM}(\mathbf{x}_n)|\mathbf{w}) = \frac{K(d(\mathbf{x}_n, \mathbf{x}_i|\mathbf{w}))}{\sum_{j\in\mathcal{M}_n} K(d(\mathbf{x}_n, \mathbf{x}_j|\mathbf{w}))}, \forall i \in \mathcal{M}_n \ , \tag{3}$$

and

$$P(\mathbf{x}_i = \mathrm{NH}(\mathbf{x}_n)|\mathbf{w}) = \frac{K(d(\mathbf{x}_n, \mathbf{x}_i|\mathbf{w}))}{\sum_{j\in\mathcal{H}_n} K(d(\mathbf{x}_n, \mathbf{x}_j|\mathbf{w}))}, \forall i \in \mathcal{H}_n \ , \tag{4}$$

where $K(\cdot)$ is a kernel function. In this study, we employed the Epanechnikov kernel [72], given by

$$K(d(\mathbf{x}_n, \mathbf{x}_i|\mathbf{w})) = \begin{cases} \frac{3}{4}\left(1 - \left(\frac{d(\mathbf{x}_n, \mathbf{x}_i|\mathbf{w})}{d(\mathbf{x}_n, \hat{\mathbf{x}}_{k+1}|\mathbf{w})}\right)^2\right) & \text{if } \dfrac{d(\mathbf{x}_n, \mathbf{x}_i|\mathbf{w})}{d(\mathbf{x}_n, \hat{\mathbf{x}}_{k+1}|\mathbf{w})} \leq 1 \\ 0 & \text{otherwise,} \end{cases} \tag{5}$$

where $\hat{\mathbf{x}}_{k+1}$ is the $(k + 1)$-th nearest neighbor of $\mathbf{x}_n$ in a feasible set. To reduce the number of parameters to be tuned, we simply set $k = 10$.

Once we define the margins, we solve the problem of learning feature weights within the large-margin framework. Specifically, we perform the estimation using the logistic-regression formulation, and obtain the following optimization problem:

$$\min_{\mathbf{w}} \sum_{n=1}^{N} \log\ (1 + \exp(-\mathbf{w}^T \bar{\mathbf{z}}_n)) + \lambda \|\mathbf{w}\|_1,\ \text{subject to}\ \mathbf{w} \geq 0\ . \tag{6}$$

Here, we impose an ℓ_1 constraint on **w** to achieve a sparse solution [73], and λ is a regularization parameter that can be estimated by using ten-fold cross-validation. Problem (6) can be solved iteratively. Briefly, we first make a guess on **w**. Then, we find the nearest neighbors of each sample and compute $\bar{\mathbf{z}}_n$. Finally, we update **w** by solving Problem (6). The iterations are carried out until convergence.

Embedded structure learning to delineate microbial community dynamics. After we identified groups of samples with similar microbiome compositions, we built a model to mathematically describe the microbial dynamics associated with disease development. To this end, principal curve fitting methods were used. Formally, a principal curve is a nonlinear generalization of the first principal-component line passing through data cloud. In the last decade, a dozen methods have been developed for principal curve fitting. However, they are generally limited to learn a curve that is embedded in a low-dimensional space and does not intersect itself, which is quite restrictive for real applications. We have recently developed a new graphic model-based method, referred to as DDRTree [24], that addresses some limitations of prior work.

Let $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_N]$ be a dataset in the input space $\mathcal{X} \subset \mathbb{R}^D$ and $\mathbf{x}_n$ be the n-th sample. We assume that the structure to be learned lies in a latent space $\mathcal{Y} \subset \mathbb{R}^d$ with $d \ll D$, and use an undirected graph $G = (V, E)$ to represent the structure, where $V = \{v_1, \cdots, v_N\}$ is a set of vertices and E is a set of edges. We introduce a set of latent variables $\mathbf{Z} = [\mathbf{z}_1, \cdots, \mathbf{z}_N]$ to explicitly represent the graph, and associate $\mathbf{z}_n$ with vertex v_n. Following the work of Gaussian mixture models [21], we assume that the observed data **X** are generated through a random process. Specifically, we first randomly select a data point residing on the graph, then corrupt the data with some random noise, and finally map the corrupted data back onto the input space. Let $\mathbf{Y} = [\mathbf{y}_1, \cdots, \mathbf{y}_N]$ be the corrupted data. Our goal is to find the latent variables **Z** and a mapping function $f : \mathbb{R}^d \rightarrow \mathbb{R}^D$ that projects data in the latent space back onto the input space so that the reconstruction error is minimized. Without explicitly specifying a form for f, it is generally difficult to learn the structure of a graph. For the purpose of the study, we use a linear mapping function $f(\mathbf{y}_n) = \mathbf{W}\mathbf{y}_n$, where $\mathbf{W} \in \mathbb{R}^{D \times d}$ is a projection matrix and $\mathbf{W}^T\mathbf{W} = \mathbf{I}$. To avoid the difficulty of learning a general graph, we consider G to be a minimum spanning tree (MST) [74], where the costs of the edges are defined to be the squared Euclidean distances of the latent variables. By combining all the above considerations, we obtain the following formulation:

$$\begin{aligned} \min_{\mathbf{W},\mathbf{Z},\mathbf{Y},\{b_{ij}\},\{p_{ij}\}} \quad & \sum_{n=1}^{N} \|\mathbf{x}_n - \mathbf{W}\mathbf{y}_n\|^2 + \gamma \sum_{i,j=1}^{N} p_{ij}(\|\mathbf{y}_i - \mathbf{z}_j\|^2 + \sigma\ \log\ p_{ij}) \\ \text{subject to} \quad & \sum_{i,j=1}^{N} b_{ij} \|\mathbf{z}_i - \mathbf{z}_j\|^2 \leq \ell, \mathbf{W}^T\mathbf{W} = \mathbf{I}, \sum_{j=1}^{N} p_{ij} = 1, p_{ij} \geq 0,\ \forall i, j, \end{aligned} \tag{7}$$

where p_{ij} is the probability of assigning $\mathbf{y}_i$ to $\mathbf{z}_i$, σ is a parameter for soft assignment using negative entropy regularization [75], γ is a parameter that controls the tradeoff between the data reconstruction error and the quantization error, and b_{ij} is constrained to be a feasible solution of an MST that takes a value of 1 if $(v_i, v_j) \in E$ and 0 otherwise. The above formulation can be interpreted as fitting a dataset by using an MST with a length bounded by ℓ (see

Fig 1C). For ease of optimization, we moved the length constraint to the objective function:

$$\begin{aligned} \min_{\mathbf{W},\mathbf{Z},\mathbf{Y},\{b_{ij}\},\{p_{ij}\}} \quad & \sum_{n=1}^{N} \|\mathbf{x}_n - \mathbf{W}\mathbf{y}_n\|^2 + \frac{\lambda}{2}\sum_{i,j=1}^{N} b_{ij}\|\mathbf{z}_i - \mathbf{z}_j\|^2 + \gamma \sum_{i,j=1}^{N} p_{ij}(\|\mathbf{y}_i - \mathbf{z}_j\|^2 + \sigma \ \log \ p_{ij}) \\ \text{subject to} \quad & \mathbf{W}^T\mathbf{W} = \mathbf{I}, \sum_{j=1}^{N} p_{ij} = 1, p_{ij} \geq 0, \ \forall i, j, \end{aligned} \tag{8}$$

where λ is a regularization parameter. For the purpose of data visualization, we projected the samples onto a three-dimensional space (i.e., $d = 3$). In order not to tune too many parameters, following the work of [24], we set $\gamma = 2$ and estimated the kernel width σ and the regularization parameter λ by using the elbow method [25]. Problem (8) can be efficiently solved by using alternating structure optimization [24, 76]. Briefly, we first fix $\{b_{ij}\}$ and $\{p_{ij}\}$ and find a solution for **W**, **Z**, and **Y** via convex optimization. Then, we fix **W**, **Z**, and **Y** and find a solution for $\{b_{ij}\}$ by solving an MST problem using Kruskal's method [77] and solve $\{p_{ij}\}$ analytically. The two steps iterate until convergence.

Constructing a microbial progression model. We combined the clustering and principal-tree results to build a progression model and extract progression paths. We represented a progression model as an undirected graph, where the vertices were the centroids of the clusters identified in the cluster analysis and they were connected based on the progression trend inferred from the principal curve. Specifically, we first projected each sample back onto the principal tree, and then extracted the progression paths by finding the shortest path from a designated root vertex to all the leaf vertices of the principal tree. In this study, we used the leaf node of the healthy control samples as the root vertex to represent the origin of the disease. By using the same procedure, we mapped the centroids of the clusters onto the principal tree and constructed an undirected graph. Two projected centroids were connected if there were no other centroids between them along a progression path, and the length of the edge was proportional to the curve distance of the two centroids measured along the progression path.

Microbial interaction network analysis

By using the pseudo-time series data recovered from the identified progression paths, we built generalized Lotka-Volterra (gLV) models to study microbial interactions associated with disease development. The gLV model has been successfully applied to several longitudinal studies to uncover pairwise interactions between microorganisms and to identify key bacteria possibly responsible for the alterations of microbiota associated with the development of a disease [48, 49]. Let $x_i(t)$ be the relative abundance of the i-th OTU, measured at time t, $1 \leq t \leq T$. A gLV model can be represented as a set of first-order ordinary differential equations, given by

$$\frac{dx_i(t)}{dt} = x_i(t)\left(\alpha_i + \sum_{j=1}^{J} \beta_{ij} x_j(t)\right), 1 \leq i \leq J, \tag{9}$$

where J is the number of OTUs, α_i is the growth rate of the i-th OTU, and β_{ij} is the strength of the pairwise interaction between the i-th and j-th OTUs. To simulate a biologically realistic ecological system where interacting species may have a wide range of relationships including competition, cooperation, or neutralism, we assume that the growth rates are positive (i.e., $\alpha_i > 0$) and the self-intersection rates are negative (i.e., $\beta_{ii} < 0$) [78]. Dividing both sides of

Eq (9) by $x_i(t)$ yields

$$\frac{d\ \ln\ x_i(t)}{dt} = \alpha_i + \sum_{j=1}^{J} \beta_{ij} x_j(t), 1 \leq i \leq J\ , \tag{10}$$

which can be further approximated as a linear system, given by

$$\frac{d\ \ln\ x_i(t)}{dt} \approx (\ln\ x_i(t))' \approx \alpha_i + \sum_{j=1}^{J} \beta_{ij} x_j(t), 1 \leq i \leq J\ , \tag{11}$$

where $(\ln x_i(t))'$ is the gradient of $\ln x_i(t)$ at time t. We used a two-step estimation procedure [79] to solve the above linear system. Specifically, we first estimated the log-transformed relative abundances and the corresponding gradients of each OTU along a progression path using cubic smoothing spline, and then estimated the variables of a gLV model using Bayesian Adaptive Lasso [79] implemented by the MDSINE package [78] with default settings. Once the linear system was solved, we built a gLV interaction network of the OTUs for each identified disease progression path.

Alpha diversity estimation

We used alpha diversity, specifically Chao1 index [27] and Shannon index [28], to assess the species richness of the gut microbial communities of individual patients. Since the estimation of alpha diversity can be biased for communities with different sequencing depth [80], we performed a rarefaction analysis by sampling 10,000 reads from each community and then calculated the corresponding alpha diversity. The process was repeated 1,000 times and the average value was reported.

Statistical analysis

We performed the Spearman's rank correlation analysis to test the association between a ranked variable and a measurement variable (e.g., the change in the relative abundance of an OTU along an identified progression path), and the Wilcoxon rank-sum test to evaluate the difference of the microbial compositions between two groups of samples. We performed the χ^2 test to explore the dependence between two sets of categorical variables. If necessary, p-values were adjusted by the DS-FDR method [39] for multiple testing correction. We performed the ANOVA analysis to compare the alpha diversities of the identified clusters.

Supporting information

S1 Fig. Removing samples containing less than 10^4 reads. A total of 37 samples were excluded from downstream analysis.
(PDF)

S2 Fig. Identifying disease-related microorganisms using the LOGO algorithm. (**a**)The regularization parameter λ was estimated through ten-fold cross-validation. (**b**) By using a cutoff of 0.001, a total of 172 OTUs were identified to be related to disease development.
(PDF)

S3 Fig. Estimating regularization parameter λ and kernel width *σ* of the DDRTree algorithm using the elbow method. The optimal *σ* and λ were estimated to be 0.5 and 150, respectively.
(PDF)

S4 Fig. Comparison of inflammation activities of patients with the same CD behaviors in Cluster 4 and Cluster 5. Active inflammation was measured by fecal calprotectin >150 μg/g.
(PDF)

S5 Fig. OTUs with significant changes in relative abundance along at least one progression path.
(PDF)

S6 Fig. Spearman's rank correlation analysis of selected OTUs for which the relative abundances were significantly decreased along the four modeled progression paths.
(PDF)

S7 Fig. Spearman's rank correlation analysis of selected OTUs for which the relative abundances were significantly increased along the four modeled progression paths.
(PDF)

S8 Fig. Toy example illustrating how to use static samples to form pseudo-time series data. Each point presents a sample, and the solid line represents the identified progression paths. The static samples were projected onto the identified progression paths. Here, the projection of a sample was defined as a point on a progression path that is closest to the sample. By using the healthy controls as the baseline, the static samples were ordered along a path according to the extent to which the disease progressed from an inflammatory phenotype toward intestinal stricture and penetration. The ordered samples can be viewed as pseudo-time series data.
(PDF)

S9 Fig. Microbial interaction networks inferred by the gLV method applied to pseudo-time series data recovered from modeled disease progression paths. Each node represents an OTU, its size is proportional to the number of edges directed out of the node (i.e., out-degree), and its face color represents the sign of the correlation of the relative abundance of the OTU with a progression path (red: positive, blue: negative).
(PDF)

S10 Fig. Heatmap of KEGG pathways that were significantly disrupted along at least one of the modeled disease progression paths. Each row represents a pathway, and each column represents a patient sample. The samples were first ordered by cluster labels and then by progression distances. For the purpose of visualization, the pathway activity was log-transformed and scaled into the range of [0, 1].
(PDF)

S11 Fig. KEGG pathways that were significantly disrupted along at least one disease progression path.
(PDF)

S1 Table. Summary of the study cohort used in the analysis.
(PDF)

S2 Table. Detailed clinical information of the study cohort used in the analysis.
(XLSX)

S3 Table. Pairwise comparisons of alpha diversities of identified clusters. The level of significance was assessed by ANOVA. ns: not significant.
(PDF)

S4 Table. KEGG pathways that were significantly disrupted along at least one of the modeled progression paths.
(XLSX)

Author Contributions

Conceptualization: Lu Li, Robert J. Genco, Yijun Sun.

Data curation: Lu Li, Yijun Sun.

Formal analysis: Lu Li, Jiho Sohn, Steve Goodison, Patricia I. Diaz, Yijun Sun.

Funding acquisition: Robert J. Genco, Jean Wactawski-Wende, Yijun Sun.

Investigation: Jiho Sohn, Yijun Sun.

Methodology: Lu Li, Yijun Sun.

Project administration: Jean Wactawski-Wende, Patricia I. Diaz, Yijun Sun.

Supervision: Robert J. Genco, Jean Wactawski-Wende, Patricia I. Diaz, Yijun Sun.

Writing – original draft: Lu Li, Jiho Sohn, Steve Goodison, Patricia I. Diaz, Yijun Sun.

Writing – review & editing: Lu Li, Jiho Sohn, Jean Wactawski-Wende, Steve Goodison, Patricia I. Diaz, Yijun Sun.

References

1. Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, Gordon JI. The Human Microbiome Project. Nature. 2007; 449:804–810. https://doi.org/10.1038/nature06244 PMID: 17943116
2. Gilbert JA, Blaser MJ, Caporaso JG, Jansson JK, Lynch SV, Knight R. Current understanding of the human microbiome. Nature Medicine. 2018; 24(4):392–400. https://doi.org/10.1038/nm.4517 PMID: 29634682
3. Cryan JF, O'Riordan KJ, Sandhu K, Peterson V, Dinan TG. The gut microbiome in neurological disorders. The Lancet Neurology. 2020; 19(2):179–194. https://doi.org/10.1016/S1474-4422(19)30356-4 PMID: 31753762
4. Miraglia F, Colla E. Microbiome, Parkinson's disease and molecular mimicry. Cells. 2019; 8(3):222. https://doi.org/10.3390/cells8030222 PMID: 30866550
5. Thomas AM, Manghi P, Asnicar F, Pasolli E, Armanini F, Zolfo M, et al. Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation. Nature Medicine. 2019; 25(4):667–678. https://doi.org/10.1038/s41591-019-0405-7 PMID: 30936548
6. The Integrative HMP (iHMP) Research Network Consortium. The Integrative Human Microbiome Project: dynamic analysis of microbiome-host omics profiles during periods of human health and disease. Cell Host & Microbe. 2015; 16(3):276–289.
7. Halfvarson J, Brislawn CJ, Lamendella R, Vázquez-Baeza Y, Walters WA, Bramer LM, et al. Dynamics of the human gut microbiome in inflammatory bowel disease. Nature Microbiology. 2017; 2(5):17004. https://doi.org/10.1038/nmicrobiol.2017.4 PMID: 28191884
8. Lloyd-Price J, Arze C, Ananthakrishnan AN, Schirmer M, Avila-Pacheco J, Poon TW, et al. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. Nature. 2019; 569(7758):655–662. https://doi.org/10.1038/s41586-019-1237-9 PMID: 31142855
9. Yost S, Duran-Pinedo AE, Teles R, Krishnan K, Frias-Lopez J. Functional signatures of oral dysbiosis during periodontitis progression revealed by microbial metatranscriptome analysis. Genome Medicine. 2015; 7(1):1–19. https://doi.org/10.1186/s13073-015-0231-6
10. Gevers D, Kugathasan S, Denson LA, Vázquez-Baeza Y, Van Treuren W, Ren B, et al. The treatment-naive microbiome in new-onset Crohn's disease. Cell Host & Microbe. 2014; 15(3):382–392. https://doi.org/10.1016/j.chom.2014.02.005 PMID: 24629344
11. Baumgart DC, Sandborn WJ. Crohn's disease. The Lancet. 2012; 380(9853):1590–1605. https://doi.org/10.1016/S0140-6736(12)60026-9

12. Lo B, Vester-Andersen M, Vind I, Prosberg M, Dubinsky M, Siegel C, et al. Changes in disease behaviour and location in patients with Crohn's disease after seven years of follow-up: a Danish population-based inception cohort. Journal of Crohn's and Colitis. 2017; 12(3):265–272. https://doi.org/10.1093/ecco-jcc/jjx138

13. Freeman HJ. Natural history and clinical behavior of Crohn's disease extending beyond two decades. Journal of Clinical Gastroenterology. 2003; 37(3):216–219. https://doi.org/10.1097/00004836-200309000-00005 PMID: 12960719

14. Khanna S, Raffals LE. The microbiome in Crohn's disease: role in pathogenesis and role of microbiome replacement therapies. Gastroenterology Clinics. 2017; 46(3):481–492. https://doi.org/10.1016/j.gtc.2017.05.004 PMID: 28838410

15. Lavoie S, Conway KL, Lassen KG, Jijon HB, Pan H, Chun E, et al. The Crohn's disease polymorphism, ATG16L1 T300A, alters the gut microbiota and enhances the local Th1/Th17 response. eLife. 2019; 8: e39982. https://doi.org/10.7554/eLife.39982 PMID: 30666959

16. Satsangi J, Silverberg M, Vermeire S, Colombel J. The Montreal classification of inflammatory bowel disease: controversies, consensus, and implications. Gut. 2006; 55(6):749–753. https://doi.org/10.1136/gut.2005.082909 PMID: 16698746

17. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, et al. QIIME allows analysis of high-throughput community sequencing data. Nature Methods. 2010; 7(5):335–336. https://doi.org/10.1038/nmeth.f.303 PMID: 20383131

18. Sun Y, Todorovic S, Goodison S. Local-learning-based feature selection for high-dimensional data analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2010; 32(9):1610–1626. https://doi.org/10.1109/TPAMI.2009.190 PMID: 20634556

19. Jain AK. Data clustering: 50 years beyond *K*-means. Pattern Recognition Letters. 2010; 31(8):651–666. https://doi.org/10.1016/j.patrec.2009.09.011

20. Tibshirani R, Walther G, Hastie T. Estimating the number of clusters in a data set via the gap statistic. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2001; 63(2):411–423. https://doi.org/10.1111/1467-9868.00293

21. Bishop CM. *Pattern Recognition and Machine Learning*. Berlin: Springer; 2006.

22. Monti S, Tamayo P, Mesirov J, Golub T. Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. Machine Learning. 2003; 52(1-2):91–118. https://doi.org/10.1023/A:1023949509487

23. Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics. 1987; 20:53–65. https://doi.org/10.1016/0377-0427(87)90125-7

24. Mao Q, Wang L, Goodison S, Sun Y. Dimensionality reduction via graph structure learning. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,; 2015. p. 765–774.

25. Thorndike RL. Who belongs in the family? Psychometrika. 1953; 18(4):267–276. https://doi.org/10.1007/BF02289263

26. Cosnes J, Cattan S, Blain A, Beaugerie L, Carbonnel F, Parc R, et al. Long-term evolution of disease behavior of Crohn's disease. Inflammatory Bowel Diseases. 2002; 8(4):244–250. https://doi.org/10.1097/00054725-200207000-00002 PMID: 12131607

27. Chao A, Chiu CH. Species richness: estimation and comparison. Wiley StatsRef: Statistics Reference Online. 2016; p. 1–26.

28. Hughes JB, Hellmann JJ, Ricketts TH, Bohannan BJ. Counting the uncountable: statistical approaches to estimating microbial diversity. Applied and Environmental Microbiology. 2001; 67(10):4399–4406. https://doi.org/10.1128/AEM.67.10.4399-4406.2001 PMID: 11571135

29. Kostic AD, Xavier RJ, Gevers D. The microbiome in inflammatory bowel disease: current status and the future ahead. Gastroenterology. 2014; 146(6):1489–1499. https://doi.org/10.1053/j.gastro.2014.02.009 PMID: 24560869

30. Norman JM, Handley SA, Baldridge MT, Droit L, Liu CY, Keller BC, et al. Disease-specific alterations in the enteric virome in inflammatory bowel disease. Cell. 2015; 160(3):447–460. https://doi.org/10.1016/j.cell.2015.01.002 PMID: 25619688

31. Imhann F, Vila AV, Bonder MJ, Fu J, Gevers D, Visschedijk MC, et al. Interplay of host genetics and gut microbiota underlying the onset and clinical presentation of inflammatory bowel disease. Gut. 2018; 67(1):108–119. https://doi.org/10.1136/gutjnl-2016-312135 PMID: 27802154

32. Roager HM, Licht TR, Poulsen SK, Larsen TM, Bahl MI. Microbial enterotypes, inferred by the prevotella-to-bacteroides ratio, remained stable during a 6-month randomized controlled diet intervention with

the new nordic diet. Applied and Environmental Microbiology. 2014; 80(3):1142–1149. https://doi.org/10.1128/AEM.03549-13 PMID: 24296500

33. Claesson MJ, Jeffery IB, Conde S, Power SE, O'Connor EM, Cusack S, et al. Gut microbiota composition correlates with diet and health in the elderly. Nature. 2012; 488:178–184. https://doi.org/10.1038/nature11319 PMID: 22797518

34. Gorvitovskaia A, Holmes SP, Huse SM. Interpreting Prevotella and Bacteroides as biomarkers of diet and lifestyle. Microbiome. 2016; 4(1):1–12. https://doi.org/10.1186/s40168-016-0160-7 PMID: 27068581

35. Tyakht AV, Kostryukova ES, Popenko AS, Belenikin MS, Pavlenko AV, Larin AK, et al. Human gut microbiota community structures in urban and rural populations in Russia. Nature Communications. 2013; 4(1):1–9. https://doi.org/10.1038/ncomms3469

36. Mobeen F, Sharma V, Tulika P. Enterotype variations of the healthy human gut microbiome in different geographical regions. Bioinformation. 2018; 14(9):560. https://doi.org/10.6026/97320630014560 PMID: 31223215

37. Costa F, Mumolo M, Ceccarelli L, Bellini M, Romano M, Sterpi C, et al. Calprotectin is a stronger predictive marker of relapse in ulcerative colitis than in Crohn's disease. Gut. 2005; 54(3):364–368. https://doi.org/10.1136/gut.2004.043406 PMID: 15710984

38. Lewis JD. The utility of biomarkers in the diagnosis and therapy of inflammatory bowel disease. Gastroenterology. 2011; 140(6):1817–1826. https://doi.org/10.1053/j.gastro.2010.11.058 PMID: 21530748

39. Jiang L, Amir A, Morton JT, Heller R, Arias-Castro E, Knight R. Discrete false-discovery rate improves identification of differentially abundant microbes. mSystems. 2017; 2(6):e00092–17. https://doi.org/10.1128/mSystems.00092-17 PMID: 29181446

40. Willing BP, Dicksved J, Halfvarson J, Andersson AF, Lucio M, Zheng Z, et al. A pyrosequencing study in twins shows that gastrointestinal microbial profiles vary with inflammatory bowel disease phenotypes. Gastroenterology. 2010; 139(6):1844–1854. https://doi.org/10.1053/j.gastro.2010.08.049 PMID: 20816835

41. Pascal V, Pozuelo M, Borruel N, Casellas F, Campos D, Santiago A, et al. A microbial signature for Crohn's disease. Gut. 2017; 66(5):813–822. https://doi.org/10.1136/gutjnl-2016-313235 PMID: 28179361

42. Qiu X, Zhang M, Yang X, Hong N, Yu C. Faecalibacterium prausnitzii upregulates regulatory T cells and anti-inflammatory cytokines in treating TNBS-induced colitis. Journal of Crohn's and Colitis. 2013; 7 (11):e558–e568. https://doi.org/10.1016/j.crohns.2013.04.002 PMID: 23643066

43. Henke MT, Kenny DJ, Cassilly CD, Vlamakis H, Xavier RJ, Clardy J. Ruminococcus gnavus, a member of the human gut microbiome associated with Crohn's disease, produces an inflammatory polysaccharide. Proceedings of the National Academy of Sciences. 2019; 116(26):12672–12677. https://doi.org/10.1073/pnas.1904099116 PMID: 31182571

44. Colina AR, Aumont F, Deslauriers N, Belhumeur P, de Repentigny L. Evidence for degradation of gastrointestinal mucin by Candida albicans secretory aspartyl proteinase. Infection and Immunity. 1996; 64 (11):4514–4519. https://doi.org/10.1128/iai.64.11.4514-4519.1996 PMID: 8890200

45. Dethlefsen L, Eckburg PB, Bik EM, Relman DA. Assembly of the human intestinal microbiota. Trends in Ecology & Evolution. 2006; 21(9):517–523. https://doi.org/10.1016/j.tree.2006.06.013 PMID: 16820245

46. Joossens M, Huys G, Cnockaert M, De Preter V, Verbeke K, Rutgeerts P, et al. Dysbiosis of the faecal microbiota in patients with Crohn's disease and their unaffected relatives. Gut. 2011; 60:631–637. https://doi.org/10.1136/gut.2010.223263 PMID: 21209126

47. Ai D, Pan H, Li X, Gao Y, Liu G, Xia LC. Identifying gut microbiota associated with colorectal cancer using a zero-inflated lognormal model. Frontiers in Microbiology. 2019; 10:826. https://doi.org/10.3389/fmicb.2019.00826 PMID: 31068913

48. Stein RR, Bucci V, Toussaint NC, Buffie CG, Rätsch G, Pamer EG, et al. Ecological modeling from time-series inference: insight into dynamics and stability of intestinal microbiota. PLoS Computational Biology. 2013; 9(12):e1003388. https://doi.org/10.1371/journal.pcbi.1003388 PMID: 24348232

49. Buffie CG, Bucci V, Stein RR, McKenney PT, Ling L, Gobourne A, et al. Precision microbiome reconstitution restores bile acid mediated resistance to Clostridium difficile. Nature. 2015; 517(7533):205–208. https://doi.org/10.1038/nature13828 PMID: 25337874

50. Fisher CK, Mehta P. Identifying keystone species in the human gut microbiome from metagenomic timeseries using sparse linear regression. PLoS ONE. 2014; 9(7):e102451. https://doi.org/10.1371/journal.pone.0102451 PMID: 25054627

51. Douglas GM, Maffei VJ, Zaneveld JR, Yurgel SN, Brown JR, Taylor CM, et al. PICRUSt2 for prediction of metagenome functions. Nature Biotechnology. 2020; 38(6):685–688. https://doi.org/10.1038/s41587-020-0548-6 PMID: 32483366

52. Jansson J, Willing B, Lucio M, Fekete A, Dicksved J, Halfvarson J, et al. Metabolomics reveals metabolic biomarkers of Crohn's disease. PLoS ONE. 2009; 4(7):e6386. https://doi.org/10.1371/journal.pone.0006386 PMID: 19636438

53. Morgan XC, Tickle TL, Sokol H, Gevers D, Devaney KL, Ward DV, et al. Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. Genome Biology. 2012; 13:R79. https://doi.org/10.1186/gb-2012-13-9-r79 PMID: 23013615

54. Lewis JD, Chen EZ, Baldassano RN, Otley AR, Griffiths AM, Lee D, et al. Inflammation, antibiotics, and diet as environmental stressors of the gut microbiome in pediatric Crohn's disease. Cell Host & Microbe. 2015; 18(4):489–500. https://doi.org/10.1016/j.chom.2015.09.008 PMID: 26468751

55. Kim J, Kim MS, Koh AY, Xie Y, Zhan X. FMAP: functional mapping and analysis pipeline for metagenomics and metatranscriptomics studies. BMC Bioinformatics. 2016; 17:420. https://doi.org/10.1186/s12859-016-1278-0 PMID: 27724866

56. Blaser MJ, Kirschner D. The equilibria that allow bacterial persistence in human hosts. Nature. 2007; 449(7164):843–849. https://doi.org/10.1038/nature06198 PMID: 17943121

57. Cho I, Blaser MJ. The human microbiome: at the interface of health and disease. Nature Reviews Genetics. 2012; 13(4):260–270. https://doi.org/10.1038/nrg3182 PMID: 22411464

58. Gloor GB, Macklaim JM, Pawlowsky-Glahn V, Egozcue JJ. Microbiome datasets are compositional: and this is not optional. Frontiers in Microbiology. 2017; 8:2224. https://doi.org/10.3389/fmicb.2017.02224 PMID: 29187837

59. Greenacre M. *Compositional Data Analysis in Practice*. New York: Chapman and Hall/CRC; 2018.

60. Tsilimigras MC, Fodor AA. Compositional data analysis of the microbiome: fundamentals, tools, and challenges. Annals of Epidemiology. 2016; 26(5):330–335. https://doi.org/10.1016/j.annepidem.2016.03.002 PMID: 27255738

61. Holmes I, Harris K, Quince C. Dirichlet multinomial mixtures: generative models for microbial metagenomics. PLOS One. 2012; 7(2):e30126. https://doi.org/10.1371/journal.pone.0030126 PMID: 22319561

62. DiGiulio DB, Callahan BJ, McMurdie PJ, Costello EK, Lyell DJ, Robaczewska A, et al. Temporal and spatial variation of the human microbiota during pregnancy. Proceedings of the National Academy of Sciences. 2015; 112(35):11060–11065. https://doi.org/10.1073/pnas.1502875112 PMID: 26283357

63. Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinformatics. 2010; 26(19):2460–2461. https://doi.org/10.1093/bioinformatics/btq461 PMID: 20709691

64. Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R. UCHIME improves sensitivity and speed of chimera detection. Bioinformatics. 2011; 27(16):2194–2200. https://doi.org/10.1093/bioinformatics/btr381 PMID: 21700674

65. Zeller G, Tap J, Voigt AY, Sunagawa S, Kultima JR, Costea PI, et al. Potential of fecal microbiota for early-stage detection of colorectal cancer. Molecular Systems Biology. 2014; 10(11):766. https://doi.org/10.15252/msb.20145645 PMID: 25432777

66. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. Journal of Molecular Biology. 1990; 215(3):403–410. https://doi.org/10.1016/S0022-2836(05)80360-2 PMID: 2231712

67. Woodcroft BJ, Singleton CM, Boyd JA, Evans PN, Emerson JB, Zayed AA, et al. Genome-centric view of carbon processing in thawing permafrost. Nature. 2018; 560(7716):49–54. https://doi.org/10.1038/s41586-018-0338-1 PMID: 30013118

68. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Research. 2000; 28(1):27–30. https://doi.org/10.1093/nar/28.1.27 PMID: 10592173

69. Ye Y, Doak TG. A parsimony approach to biological pathway reconstruction/inference for genomes and metagenomes. PLoS Computational Biology. 2009; 5(8):e1000465. https://doi.org/10.1371/journal.pcbi.1000465 PMID: 19680427

70. Vapnik V. *The Nature of Statistical Learning Theory*. Berlin: Springer; 2013.

71. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society: Series B (Methodological). 1977; 39(1):1–22.

72. Epanechnikov VA. Non-parametric estimation of a multivariate probability density. Theory of Probability & Its Applications. 1969; 14(1):153–158. https://doi.org/10.1137/1114019

73. Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological). 1996; 58(1):267–288.

74. Wu BY, Chao KM. *Spanning Trees and Optimization Problems*. New York: CRC Press; 2004.

75. Friedman JH, Meulman JJ. Clustering objects on subsets of attributes (with discussion). Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2004; 66(4):815–849. https://doi.org/10.1111/j.1467-9868.2004.02059.x

76. Ando RK, Zhang T, Bartlett P. A framework for learning predictive structures from multiple tasks and unlabeled data. Journal of Machine Learning Research. 2005; 6(11):1817–1853.

77. Kruskal JB. On the shortest spanning subtree of a graph and the traveling salesman problem. Proceedings of the American Mathematical Society. 1956; 7(1):48–50. https://doi.org/10.1090/S0002-9939-1956-0078686-7

78. Bucci V, Tzen B, Li N, Simmons M, Tanoue T, Bogart E, et al. MDSINE: Microbial Dynamical Systems INference Engine for microbiome time-series analyses. Genome Biology. 2016; 17(1):121. https://doi.org/10.1186/s13059-016-0980-6 PMID: 27259475

79. Leng C, Tran MN, Nott D. Bayesian adaptive lasso. Annals of the Institute of Statistical Mathematics. 2014; 66(2):221–244. https://doi.org/10.1007/s10463-013-0429-6

80. Gotelli NJ, Colwell RK. Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. Ecology Letters. 2001; 4(4):379–391. https://doi.org/10.1046/j.1461-0248.2001.00230.x