Supplementary Materials for “Characterization of the Clinical Value of Alpha-diversity metrics in Microbiome Studies”

Nataša Mortvanski[1,2,*], José Luis Villanueva-Cañas[2,3] and Climent Casals-Pascual[1,4,5]

[1]Barcelona Institute for Global Health (ISGlobal), Barcelona, Spain.
[2]Pompeu Fabra University (UPF), Barcelona, Spain.
[3]Molecular Biology CORE (CDB), Hospital Clínic of Barcelona, Barcelona, Spain.
[4]Department of Clinical Microbiology, Hospital Clínic of Barcelona, Barcelona, Spain.
[5]University of Barcelona (UB), Barcelona, Spain.

*Corresponding author: natasa.mortvanski01@estudiant.upf.edu

1 Supplementary Notes

1.1 Data processing in QIIME2

All QIIME2 artefacts analysed on Qiita platform went through the same standardised preprocessing procedure. Trimmed to a length of 100bp, quality control was done by Deblur 2021.09, Greengenes database was used as reference phylogeny. In order to avoid doing data preparation for all these studies from scratch, we decided to keep working with these artefacts.

However, Deblur is not the most compatible with analysing paired-end sequencing data, such as data obtained from Hospital Clínic (paired reads need to be joined before denoising). In this case it was easier to use the DADA2 pipeline for quality control. CDI data from BioProject database (Khanna et al. 2016) on the other hand is produced by single-end sequencing technology. Since it was being compared with American Gut Project data (McDonald et al. 2018) and other CDI datasets (Weingarden et al. 2015, Khanna et al. 2017) obtained from Qiita, we wanted to analyse these datasets in the same way. That is why we processed it using Deblur.

Pre-fitted sklearn-based taxonomy classifier scikit-learn_0.24.1 was used for taxonomy assignment on Qiita platform (the most recent update of QIIME2 - qiime2 2022.2.1), while scikit-learn_0.23.1 was used for Khanna et al. (2016) CDI dataset and Hospital Clínic's data (analysed on local machine with QIIME2 version 2020.8.0).

1.2 Overview of different alpha diversity metrics

We conducted a literature search to obtain the definitions of different alpha metrics that were used in the analysis. There are two main aspects on which those indices are based, namely richness and evenness. Some of the indices are based on a combination of both. Below are the results sorted accordingly.

1.2.1 Richness metrics

Richness indices estimate the number of different species in a sample. The simplest measure for richness is the number of species or Operational Taxonomic Units (OTU). However, simply counting the number of present species is strongly affected by the bias introduced by undersampling and sequencing. This bias is even worse when the species evenness is low. There are numerous different metrics that are trying to capture or estimate richness. Here are some of the metrics that can be calculated using QIIME2 that we selected to use in our analysis:

Chao1 index is a nonparametric estimator of species richness that is correcting the observed richness for the number of lost species, estimated considering the distribution of the rarest species (Bent 2008, Finotello 2018). Chao's index for estimation of species richness is given by the equation (Thukral 2017, Website: CD Genomics):

$$S_{(max)Chao} = S_{obs} + (a^2 + b^2)$$

, where Smax = maximum no. of species, Sobs = number of species observed in different samples, a = singletons (number of species represented by one individual each), b = doubletons (number of species represented by two individuals each).

Margalef's index measures the species richness in a given area or community. It is defined as:

$$R_{MAR} = \frac{S-1}{lnN}$$

, where, S is the total number of species and N is the total number of individuals in the sample (Thukral 2017).

Menhinick's index is defined as the ratio of the total number of species (S) to square root of number of individuals in the sample (N) (Thukral 2017):

$$R_{MEN} = \frac{S}{\sqrt{N}}$$

Fisher alpha is measuring the relationship between the number of species and the relative abundance of each species (Finotello 2018). This index is based upon the logarithmic distribution of number of individuals of different species:

$$S = \alpha ln(1 + \frac{N}{\alpha})$$

where, S is the total number of species and N is the total number of individuals in the sample. The value of Fisher's alpha is computed by iteration (Thukral 2017).

Faith's phylogenetic diversity is the sum of OTU branch lengths. It takes into account phylogenetic distance between OTUs. The greater the number of unique, phylogenetically more distant OTUs, the higher this index will be (Finotello 2018).

1.2.2 Evenness metrics

Evenness indices measure how evenly the relative abundances are distributed across the different species. Besides being a valuable indicator of biodiversity, evenness also determines the stability and resilience of an ecosystem. Some indices estimate unevenness or dominance, which is complementary to evenness.

Gini index (Bendel 1989, Zheng 2008) is also defined in reference to the Lorenz curve which results from a plot of the cumulative proportion of the population to the cumulative proportion of the variable. The Gini coefficient can be, as in the figure, defined geometrically as the ratio of two geometrical areas in the unit box: (a) the area between the line of perfect equality (45 degree line in the unit box) and the Lorenz curve, which is called area A and (b) the area under the 45 degree line, or areas A + B. Because areas A + B represents the half of the unit box, that is, A+B = 1/2 , the Gini Coefficient, G, can be written as:

$$G = \frac{A}{A+B} = 2A = 1 - 2B$$

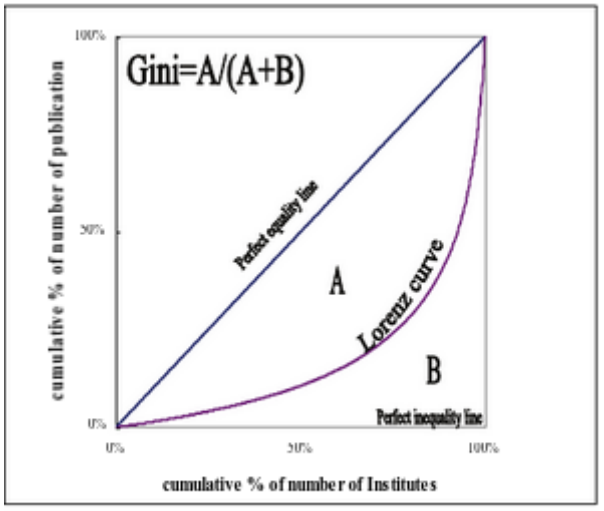


Figure 1: The Lorenz curve and Gini coefficient

Therefore a Gini coefficient is a number between 0 and 1 that measures the degree of inequality with 0 being maximum equality and 1 being maximum inequality.

Strong's index (Strong 2002) assesses species relative dominance concentration. It was, in part, based on the Gini index and Lorenz curve or partial order approach, but without the need to calculate area. It is defined as:

$$D_W = max_i [(b_i/Q) - i/R]$$

, where R = the number of species in the sample, i = the i-th species in the data set (i = 1 through R), b = the sequential cumulative totaling of i-th species abundance values (a_k) ranked largest to smallest (i.e., b = largest a_k, b = b + second largest a_k, b = b + third largest a_k . . .), Q = sum of species abundance values (Σa_k), where k = 1 through R;

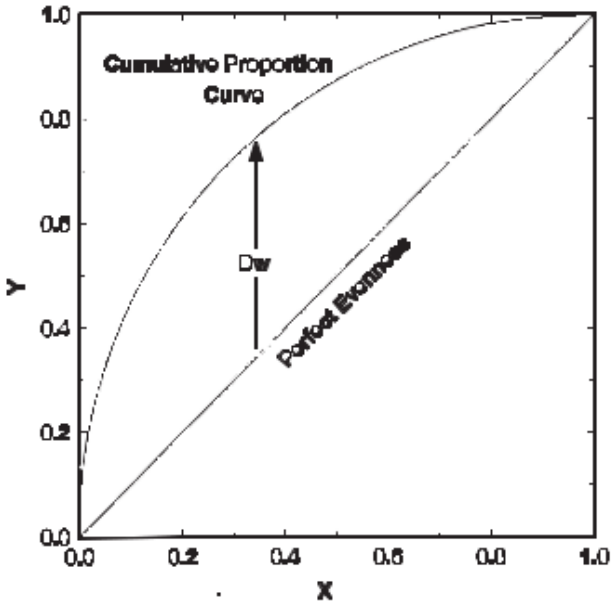


Figure 1. A cumulative proportion graph with an example of a dominance (D_W) measure. The x-axis (i/R) and y-axis (b_i/Q) of the diagram represent the right and left halves of the D_W equation. See Methods for definitions of individual parameters.

Pielou evenness is calculated as the ratio of the observed diversity to the maximum possible diversity having the same number of species (Pielou 1966). The formula is:

$$J' = H'/H'max$$

It has Shannon's formula in both numerator and denominator. $H'max = logS$, where S is the number of species.

1.2.3 Both richness and evenness

Shannon's index (Shannon–Weaver index, Shannon entropy) (Thukral 2017) was originally developed for communication systems and is based on information theory, however, it can also be used to define the biological diversity of the communities. The information content, H, therefore can be written as a function of probability:

$$H' = -\sum \frac{n_i}{N} ln \frac{n_i}{N} = -\sum p_i lnp_i$$

, where a message consists of N number of alphabets (number of species), pi is the probability of each letter in a message consisting of m alphabets (species), ni is the number of individuals of the i th letter (specie).

The minimum value of H' is 0 when all the individuals in the sample belong to the same species. This community has minimal redundancy and therefore maximum entropy (Bent 2008). This index is reaching the maximum if all the species in the sample are represented by equal number of individuals:

$$H'max = logS$$

Simpson's index reflects the probability that any two organisms sampled will be the same phylotype by capturing both richness and relative abundance (Finotello 2018). If there are k species consisting of n individuals, distributed among different species as n1, n2, n3, …nk, then the probability (p1) of the first individual belonging to a species will be:

$$p_i = \frac{n_i}{n}$$

Probability (p1,2) that the second individual drawn from the sample without replacement also belongs to the same species will be:

$$p_{1,2} = (\frac{n_1}{n} * \frac{n_1 - 1}{n-1})$$

The sum of the probabilities for all the species is a measure of the concentration (or abundance) (C) of the species:

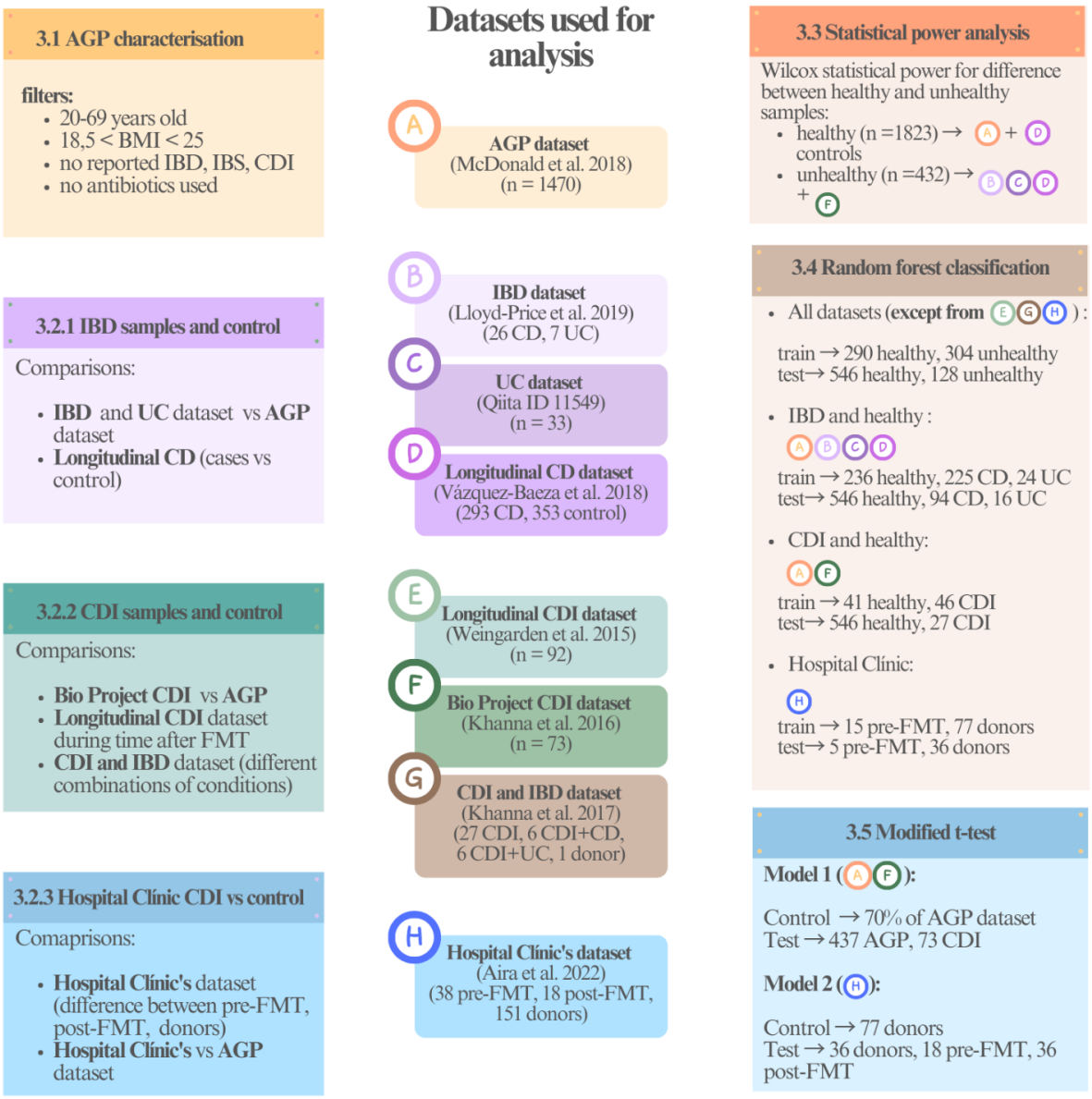
$$C_{Simpson} = (\frac{\sum n_1(n_i - 1)}{n(n-1)})$$

If the sample size is large, then the probability (p1,2) that the second individual drawn from the sample with replacement also belongs to the same species (C') will be:

$$C_{Simpson} = \sum p_i^2 = \frac{\sum n_i^2}{n^2}$$

The maximum value of Simpson's concentration is 1 when all the individuals in the sample belong to the same species. It follows logarithmic distribution which means that similar increments on the assessment scale do not represent equal changes in dominance concentration.

1.3 Study design



2 Supplementary Tables

Reference	Qiita ID	Study title	16S reg.	Raref. depth	N (used in analysis)	Meta data	Country	Techn.
McDonald *et al.* 2018	10317	American Gut: an Open Platform for Citizen Science Microbiome Research	V4	5000	**1470** (healthy)	Yes	America/ Europe	Illumina MiSeq
Lloyd-Price *et al.* 2019	11484	Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases	V4	1000	**33** (26 CD, 7 UC)	Yes	America	Illumina MiSeq HiSeq2000 or 2500 2x101
No publication (PI: Robert Knight) (Website: Qiita)	11549	Metaomics Reveals Microbiome Based Proteolysis as a Driver of Ulcerative Colitis Severity	V4	4500	**33** (UC)	Yes	America (UCSD)	Illumina MiSeq
Vázquez-Baeza *et al*. 2018	2538	Guiding longitudinal sampling in IBD cohorts	V4	7000	**646** (293 CD, 353 control)	Yes	America (UNC)	**Illumina HiSeq 2000**
Weingarden *et al*. 2015	1924	Dynamic changes in short- and long-term bacterial composition following fecal microbiota transplantation for recurrent Clostridium difficile infection	V4	15000	**92** (CDI)	Yes	America (UMN)	Illumina MiSeq 2 × 150 bp
Khanna *et al*. 2017	10057	Changes in microbial ecology after fecal microbiota transplantation for recurrent C. difficile infection affected by underlying inflammatory bowel disease	V4	30000	**40** (27 CDI, 6 CDI+CD, 6 CDI+UC, 1 donor)	Yes	America	Illumina MiSeq

Supplementary Table S1. Properties of selected studies from Qiita repository. All artefacts from Qiita had the same preprocessing steps (*Deblur 2021.09 (Reference phylogeny for SEPP: Greengenes_13.8, BIOM: all.biom)* | *Trimming (length: 100)*)

Study	Study Accession	16S reg.	N (used in analysis)	Rarefact. depth	Metadata	Country	Techn.
Khanna *et al*. 2016	PRJNA342347	V4	**73** (CDI)	6260	Yes	America	Illumina MiSeq (Single-Read)

Supplementary Table S2. Properties of selected study from BioProject repository.

Dataset	16S region	n (Pre-FMT)	n (Post-FMT)	n (donors)	Rarefaction depth	Technology
CDI samples	**V3-V4 (?)**	38	18	38	5500	Illumina MiSeq
Catalan biobank	**V3-V4 (?)**	/	/	113	15000	Illumina MiSeq

Supplementary Table S3. Properties of data from Hospital Clínic.

metric	**statistic**	**p.value**	**skewness**	**kurtosis**
Faith PD	0.9937	**4.46e-12**	0.2673	2.7841
Margalef	0.9915	**1.83e-07**	0.3456	2.9911
Menhinick	0.9915	**1.83e-07**	0.3456	2.9911
Chao1	0.9831	**1.51e-15**	0.5375	3.3656
Fisher alpha	0.9742	**7.79e-06**	0.6494	3.4489
Gini index	0.9604	**1.38e-19**	-0.7684	3.6214
Strong	0.9559	**1.21e-20**	0.8152	3.4930
Shannon entropy	0.9454	**1.14e-30**	-0.8837	3.4804
Pielou evenness	0.8938	**6.89e-23**	-1.2674	4.4065
Simpson	0.7173	**2.53e-44**	-2.2548	8.2933

Supplementary Table S4. American Gut Project data alpha metrics' distribution statistics. Normality test (Shapiro-Wilk), skewness and kurtosis of different metrics

parameter	**group1**	**group2**	**p.value**	**p.adjusted**
Chao1	CD	healthy	**2.10e-18**	**3.16e-17**
Margalef	CD	healthy	**3.15e-09**	**1.57e-08**
Faith PD	CD	healthy	**4.41e-08**	**1.70e-07**
Gini index	CD	healthy	**9.39e-08**	**2.82e-07**
Strong	CD	healthy	**4.32e-06**	**1.08e-05**
Fisher alpha	CD	healthy	**0.00007**	**0.00015**
Pielou evenness	CD	healthy	**0.00831**	**0.01558**
Menhinick	CD	healthy	0.07199	0.10798
Shannon entropy	CD	healthy	0.14985	0.20434
Simpson	CD	healthy	0.75357	0.76939

Supplementary Table S5. Results of the Mann-Whitney-Wilcoxon test for difference in means of healthy population and Crohn's disease samples from Lloyd-Price et al. 2019

parameter	group1	group2	p.value	p.adjusted
Chao1	UC	healthy	**3.27e-27**	**9.82e-26**
Margalef	UC	healthy	**1.40e-14**	**1.40e-13**
Menhinick	UC	healthy	**1.41e-12**	**1.06e-11**
Fisher alpha	UC	healthy	**9.57e-12**	**5.74e-11**
Faith PD	UC	healthy	**4.52e-08**	**1.70e-07**
Gini index	UC	healthy	**8.05e-08**	**2.68e-07**
Strong	UC	healthy	**0.00072**	**0.00144**
Shannon entropy	UC	healthy	**0.01894**	**0.03157**
Pielou evenness	UC	healthy	0.20476	0.26708
Simpson	UC	healthy	0.48789	0.56295

Supplementary Table S6. Results of the Mann-Whitney-Wilcoxon test for difference in means of healthy population and Ulcerative colitis samples (Lloyd-Price et al. 2019 and Qiita ID: 11549)

parameter	group1	group2	p.value	p.adjusted
Fisher alpha	control(AGP)	control_2	**1.93e-187**	**1.93e-185**
Gini index	control(AGP)	control_2	**2.40e-30**	**2.66e-29**
Margalef	control(AGP)	control_2	**4.85e-13**	**1.87e-12**
Menhinick	control(AGP)	control_2	**5.61e-10**	**1.44e-09**
Chao1	control(AGP)	control_2	**7.50e-06**	**1.32e-05**
Faith PD	control(AGP)	control_2	**0.00045**	**0.00066**
Simpson	control(AGP)	control_2	0.20806	0.22615
Pielou evenness	control(AGP)	control_2	0.4135	0.43526
Shannon entropy	control(AGP)	control_2	0.52903	0.54539
Strong	control(AGP)	control_2	0.77432	0.77432

Supplementary Table S7. Results of the Mann-Whitney-Wilcoxon test for difference in means of controls from longitudinal CD study (Vázquez-Baeza et al. 2018) and AGP controls

parameter	group1	group2	p.value	p.adjusted
Faith PD	CD_2	CD_1	**2.01e-14**	**3.01e-13**
Gini index	CD_2	CD_1	**2.14e-09**	**5.35e-09**
Menhinick	CD_2	CD_1	**6.30e-08**	**1.35e-07**
Strong	CD_2	CD_1	**9.74e-07**	**1.83e-06**
Pielou evenness	CD_2	CD_1	**0.00015**	**0.00024**
Chao1	CD_2	CD_1	**0.00586**	**0.00732**
Margalef	CD_2	CD_1	**0.01065**	**0.01278**
Simpson	CD_2	CD_1	0.22596	0.2421
Fisher alpha	CD_2	CD_1	0.42371	0.43832
Shannon entropy	CD_2	CD_1	0.68844	0.68844

Supplementary Table S8. Results of the Mann-Whitney-Wilcoxon test for difference in means of CD samples from longitudinal study (Vázquez-Baeza et al. 2018) and CD samples from first data set (Lloyd-Price et al. 2019)

parameter	group1	group2	p.value	p.adjusted
Simpson	crohns	control	**0.00007**	**0.00010**
Pielou evenness	crohns	control	**0.00010**	**0.00014**
Shannon entropy	crohns	control	**0.00067**	**0.00091**
Chao1	crohns	control	**0.00074**	**0.00097**
Strong	crohns	control	**0.00087**	**0.00109**
Gini index	crohns	control	**0.00926**	**0.01068**
Faith PD	crohns	control	**0.04426**	**0.04918**
Margalef	crohns	control	0.06702	0.06702
Menhinick	crohns	control	0.06702	0.06702
Fisher alpha	crohns	control	0.06702	0.06702

Supplementary Table S9. Results of the Mann-Whitney-Wilcoxon test for difference in means of controls and Crohn's samples in Vázquez-Baeza et al. 2018

parameter	group1	group2	p.value	p.adjusted
Chao1	crohns (surgery)	control	**1.39e-24**	**4.17e-23**
Margalef	crohns (surgery)	control	**2.14e-23**	**3.22e-22**
Menhinick	crohns (surgery)	control	**4.20e-23**	**4.20e-22**
Fisher alpha	crohns (surgery)	control	**1.23e-21**	**6.15e-21**
Faith PD	crohns (surgery)	control	**1.23e-21**	**6.15e-21**
Gini index	crohns (surgery)	control	**1.23e-21**	**6.15e-21**
Strong	crohns (surgery)	control	**1.04e-19**	**4.47e-19**
Pielou evenness	crohns (surgery)	control	**1.27e-19**	**4.77e-19**
Shannon entropy	crohns (surgery)	control	**1.65e-17**	**5.50e-17**
Simpson	crohns (surgery)	control	**9.04e-11**	**2.26e-10**

Supplementary Table S10. Results of the Mann-Whitney-Wilcoxon test for difference in means of controls and Crohn's samples that undergone surgery in Vázquez-Baeza et al. 2018

parameter	group1	group2	p.value	p.adjusted
Chao1	healthy	CDI	**2.92e-47**	**2.92e-46**
Faith PD	healthy	CDI	**9.10e-47**	**4.55e-46**
Fisher alpha	healthy	CDI	**2.69e-45**	**8.97e-45**
Gini index	healthy	CDI	**1.41e-44**	**3.52e-44**
Margalef	healthy	CDI	**8.07e-44**	**1.61e-43**
Menhinick	healthy	CDI	**1.06e-43**	**1.77e-43**
Shannon entropy	healthy	CDI	**9.76e-21**	**1.39e-20**
Simpson	healthy	CDI	**7.75e-12**	**9.69e-12**
Pielou evenness	healthy	CDI	**0.00004**	**0.00004**
Strong	healthy	CDI	0.99646	0.99646

Supplementary Table S11. Results of the Mann-Whitney-Wilcoxon test for difference between controls and CDI samples in Khanna et al. 2016

parameter	group1	group2	p.value	p.adjusted
Chao1	donor	CDIpost	**2.84e-20**	**8.52e-19**
Chao1	donor	CDIpre	**1.46e-15**	**2.18e-14**
Faith PD	donor	CDIpost	**6.45e-15**	**6.45e-14**
Faith PD	donor	CDIpre	**1.34e-14**	**1.00e-13**
Fisher alpha	donor	CDIpost	**7.26e-13**	**4.36e-12**
Fisher alpha	donor	CDIpre	**1.18e-12**	**5.88e-12**
Gini index	donor	CDIpost	**5.29e-12**	**2.27e-11**
Gini index	donor	CDIpre	**1.04e-11**	**3.90e-11**
Margalef	donor	CDIpost	**3.19e-11**	**1.06e-10**
Margalef	donor	CDIpre	**6.77e-11**	**1.85e-10**
Pielou evenness	donor	CDIpost	**6.77e-11**	**1.85e-10**
Pielou evenness	donor	CDIpre	**9.17e-11**	**2.29e-10**
Shannon entropy	donor	CDIpost	**2.67e-10**	**6.16e-10**
Shannon entropy	donor	CDIpre	**9.59e-10**	**2.05e-09**
Simpson	donor	CDIpost	**2.05e-09**	**4.11e-09**
Simpson	donor	CDIpre	**7.63e-09**	**1.43e-08**
Strong	donor	CDIpost	**8.33e-08**	**1.47e-07**
Strong	donor	CDIpre	**2.16e-07**	**3.59e-07**
Menhinick	donor	CDIpre	**0.00001**	**0.00002**
Gini index	CDIpre	CDIpost	**0.00876**	**0.01314**
Shannon entropy	CDIpre	CDIpost	**0.00925**	**0.01321**
Chao1	CDIpre	CDIpost	**0.01016**	**0.01385**
Fisher alpha	CDIpre	CDIpost	**0.01271**	**0.01469**
Margalef	CDIpre	CDIpost	**0.01271**	**0.01469**
Menhinick	CDIpre	CDIpost	**0.01271**	**0.01469**
Simpson	CDIpre	CDIpost	**0.01273**	**0.01469**
Menhinick	donor	CDIpost	0.09159	0.10177
Pielou evenness	CDIpre	CDIpost	0.09762	0.10459
Faith PD	CDIpre	CDIpost	0.14901	0.15415
Strong	CDIpre	CDIpost	0.31623	0.31623

Supplementary Table S12. Results of the Mann-Whitney-Wilcoxon test for difference in means of different conditions (healthy donors, CDi pre-FMT, CDI post-FMT) in Hospital Clínic's dataset (Aira et al. 2022)

model	accuracy_condition	accuracy_healthy_or_not
all alpha metrics	0.88	0.89
Faith + Gini	0.85	0.86
Menhinick + Gini	0.84	0.85
Chao1 + Gini	0.84	0.87
Fisher + Gini	0.83	0.85
Margalef + Gini	0.82	0.85
Chao1 + Pielou	0.72	0.74
Chao1 + Simpson	0.72	0.74
Menhinick + Strong	0.72	0.74
Menhinick + Pielou	0.71	0.73
Menhinick + Shannon	0.71	0.72
Fisher + Pielou	0.71	0.74
Fisher + Shannon	0.71	0.73
Margalef + Shannon	0.70	0.74
Margalef + Pielou	0.70	0.73
Margalef + Simpson	0.70	0.73
Faith + Pielou	0.69	0.70
Chao1 + Shannon	0.69	0.73
Fisher + Simpson	0.69	0.71
Menhinick + Simpson	0.68	0.71
Chao1 + Strong	0.68	0.72
Margalef + Strong	0.68	0.71
Faith + Shannon	0.68	0.68
Faith + Simpson	0.68	0.70
Fisher + Strong	0.67	0.71
Faith + Strong	0.64	0.66

Supplementary Table S13. Accuracy of prediction of different models of random forest classifier trained on a dataset consisting of IBD (CD and UC), CDI and healthy samples

model	accuracy_condition	accuracy_healthy_or_not
all alpha metrics	0.87	0.88
Faith + Gini	0.86	0.86
Menhinick + Gini	0.85	0.86
Fisher + Gini	0.84	0.86
Margalef + Gini	0.83	0.85
Chao1 + Gini	0.82	0.84
Fisher + Simpson	0.70	0.70
Menhinick + Pielou	0.70	0.69
Chao1 + Pielou	0.69	0.70
Margalef + Pielou	0.69	0.70
Fisher + Pielou	0.69	0.69
chao1 + shannon_entropy	0.69	0.70
margalef + simpson	0.68	0.70
chao1 + simpson	0.68	0.69
Menhinick + Simpson	0.68	0.68
Fisher + Shannon	0.67	0.67
Menhinick + Strong	0.67	0.66
Menhinick + Shannon	0.67	0.66
Margalef + Shannon	0.66	0.66
Chao1 + Strong	0.65	0.67
Margalef + Strong	0.64	0.65
Fisher + Strong	0.64	0.65
Faith + Simpson	0.63	0.63
Faith + Pielou	0.62	0.61
Faith + Shannon	0.61	0.62
Faith + Strong	0.59	0.62

Supplementary Table S14. Accuracy of prediction of different models of random forest classifier trained on dataset consisting of IBD (CD and UC) and healthy samples

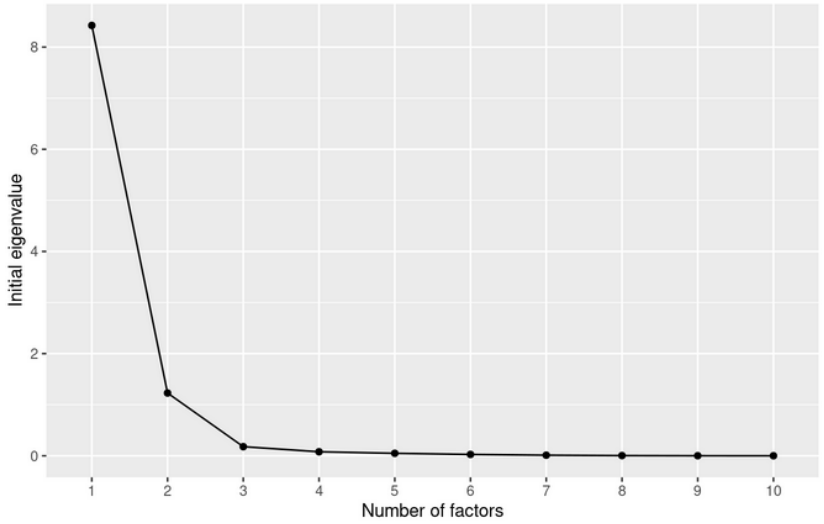
model	accuracy
Chao1 + Gini	1.00
Margalef + Gini	1.00
Menhinick + Gini	1.00
Fisher + Gini	1.00
Faith + Gini	0.99
Faith + Strong	0.98
Faith + Pielou	0.98
Faith + Shannon	0.98
Faith + Simpson	0.98
All alpha metrics	0.98
Chao1 + Simpson	0.97
Menhinick + Pielou	0.97
Menhinick + Simpson	0.97
Chao1 + Strong	0.97
Chao1 + Pielou	0.97
Chao1 + Shannon	0.97
Menhinick + Shannon	0.97
Menhinick + Strong	0.96
Margalef + Strong	0.96
Margalef + Pielou	0.95
Fisher + Strong	0.95
Fisher + Pielou	0.95
Fisher + Shannon	0.95
Fisher + Simpson	0.95
Margalef + Shannon	0.95
Margalef + Simpson	0.95

Supplementary Table S15. Accuracy of prediction of different models of random forest classifier trained on dataset consisting of CDI and healthy samples

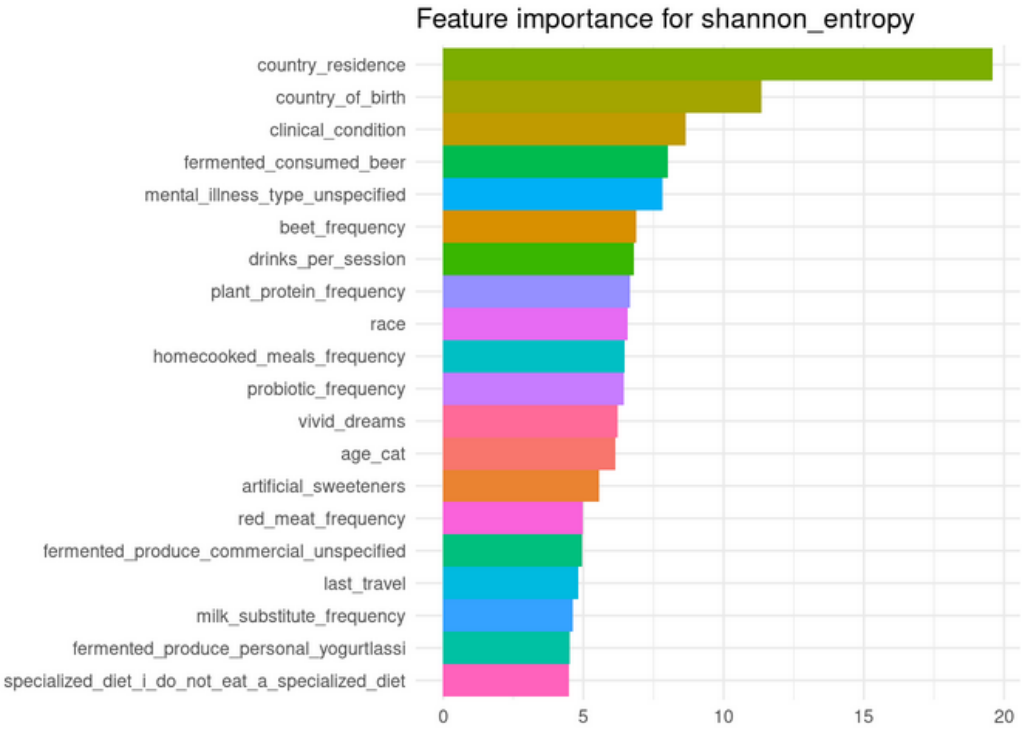
model	accuracy
Chao1 + Gini	1.00
Margalef + Gini	1.00
Menhinick + Gini	1.00
Menhinick + Strong	1.00
Menhinick + Shannon	1.00
Fisher + Gini	1.00
Faith + Gini	1.00
All alpha metrics	1.00
Chao1 + Strong	0.97
Menhinick + Pielou	0.97
Menhinick + Simpson	0.97
Fisher + Shannon	0.97
Fisher + Simpson	0.97
Faith + Strong	0.97
Faith + Pielou	0.97
Faith + Shannon	0.97
Faith + Simpson	0.97
Chao1 + Pielou	0.95
Chao1 + Simpson	0.95
Margalef + Strong	0.95
Margalef + Pielou	0.95
Fisher + Pielou	0.95
Chao1 + Shannon	0.92
Margalef + Shannon	0.92
Margalef + Simpson	0.92
Fisher + Strong	0.92

Supplementary Table S16. Accuracy of prediction of different models of random forest classifier trained on dataset consisting of Hospital Clínic's CDI samples before FMT and healthy donor samples

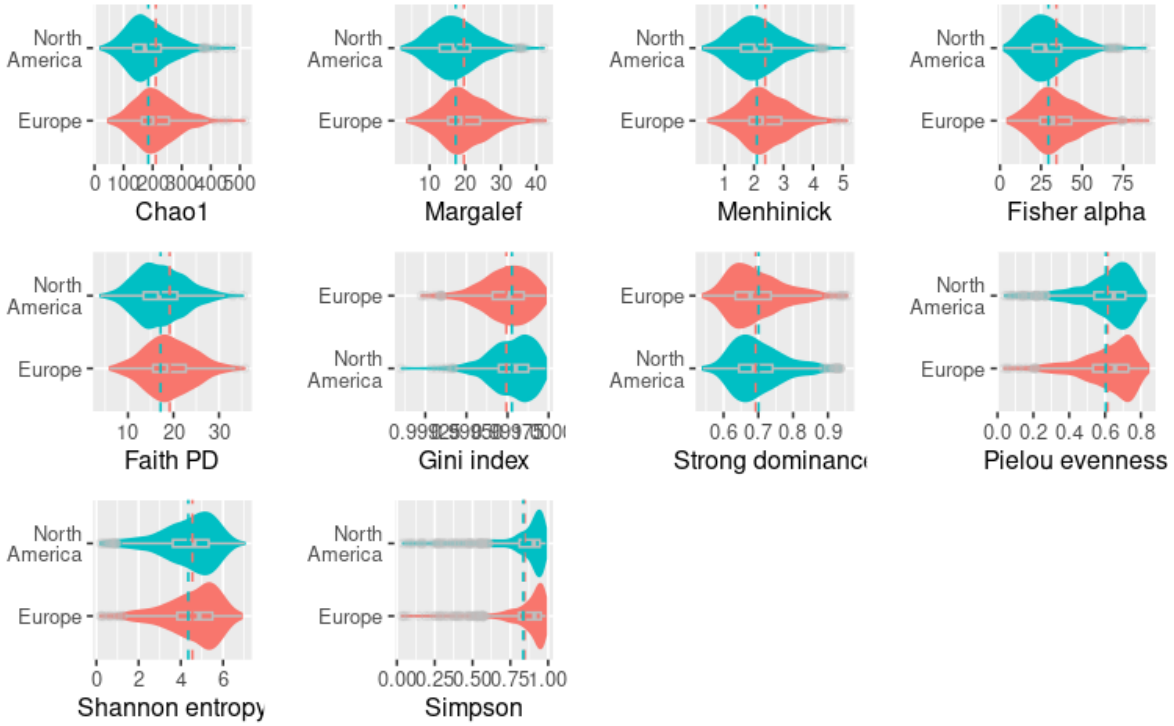
3 Supplementary Figures



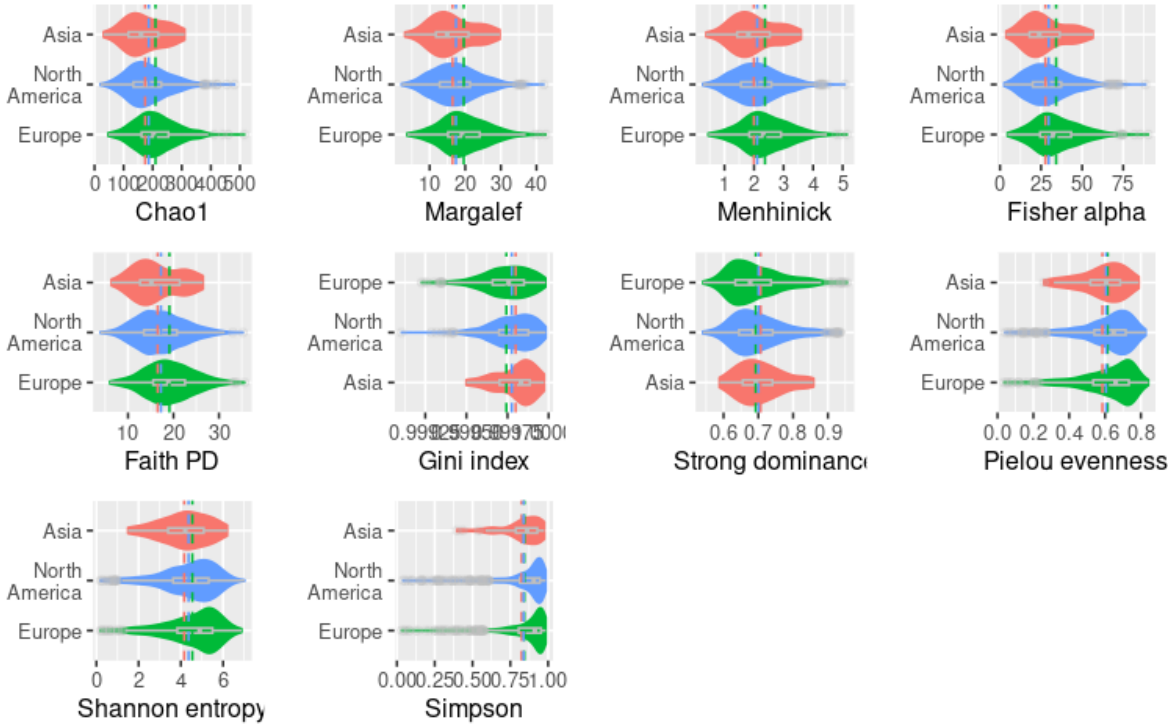
Supplementary Figure S1. Scree plot based on exploratory factor analysis of 10 alpha diversity indices computed on AGP dataset



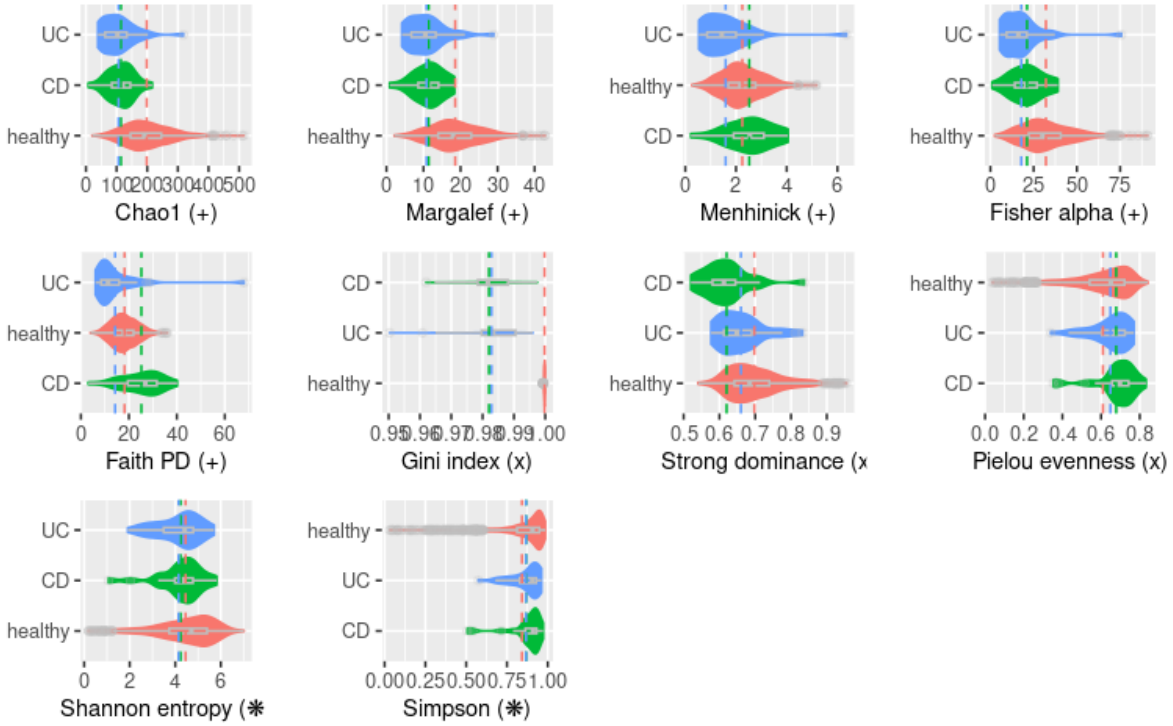
Supplementary Figure S2. Feature importance of AGP metadata categories for estimating alpha diversity metrics (in this case Shannon entropy) obtained by Random Forest classifier



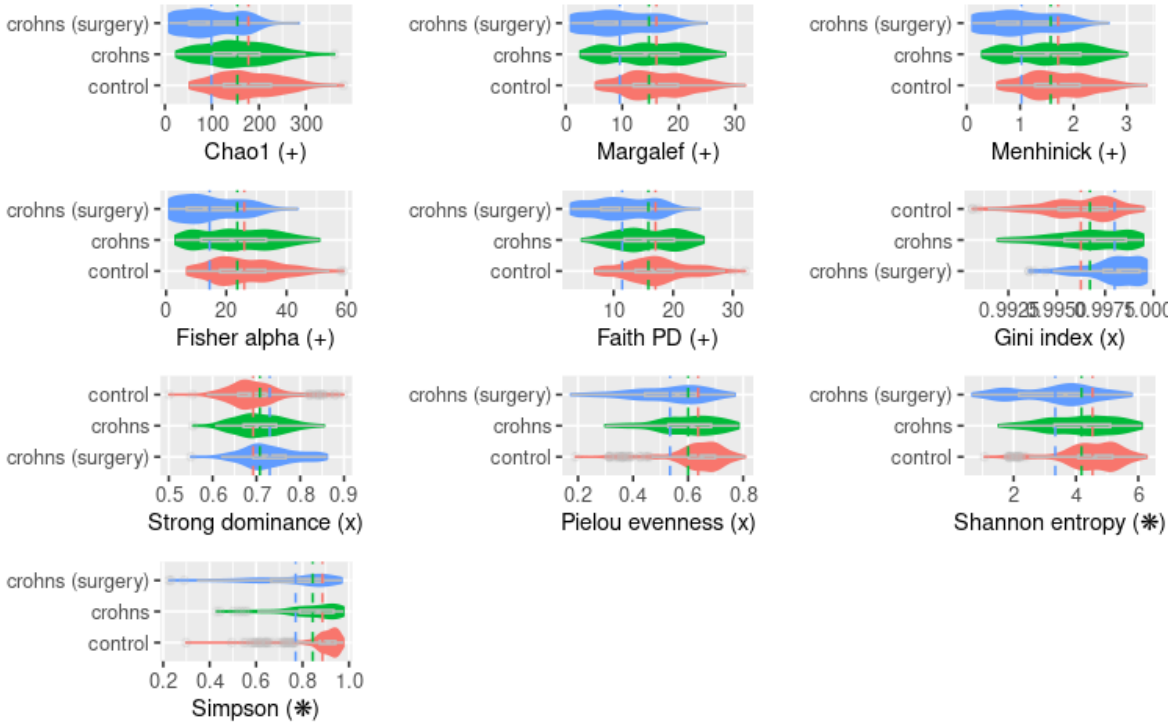
Supplementary Figure S3. Difference in distributions and means of alpha metrics in groups of AGP samples with different countries of residence (grouped by continents)



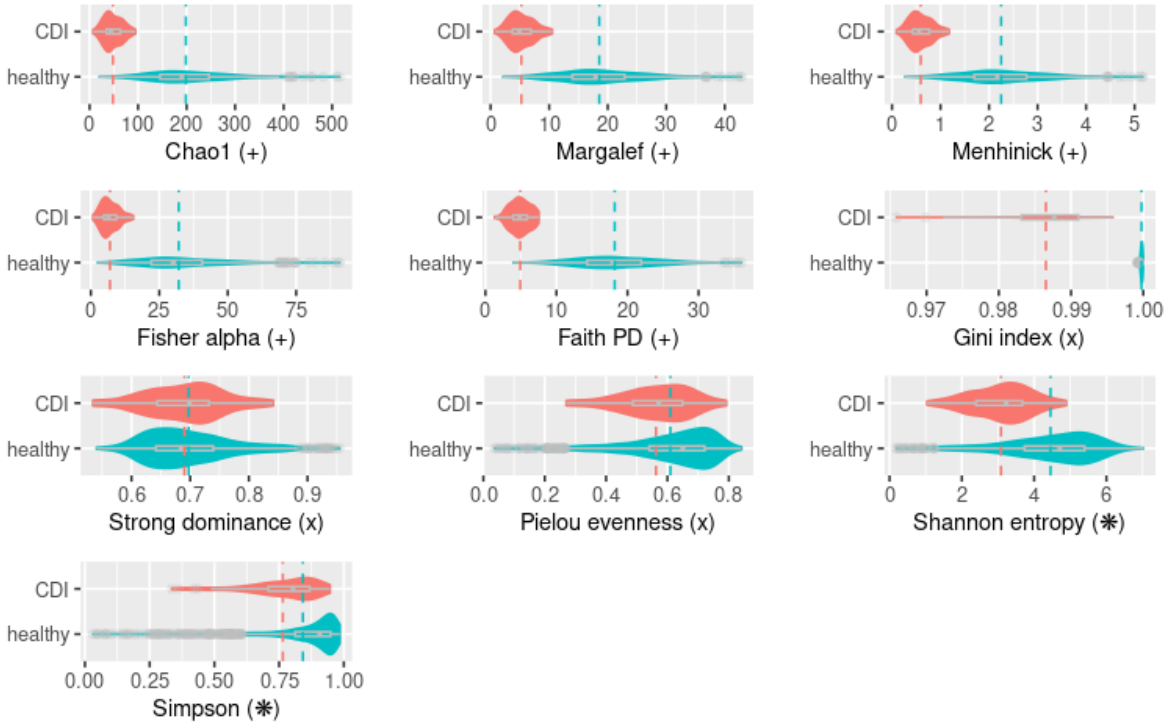
Supplementary Figure S4. Difference in distributions and means of alpha metrics in groups of AGP samples with different countries of birth (grouped by continents)



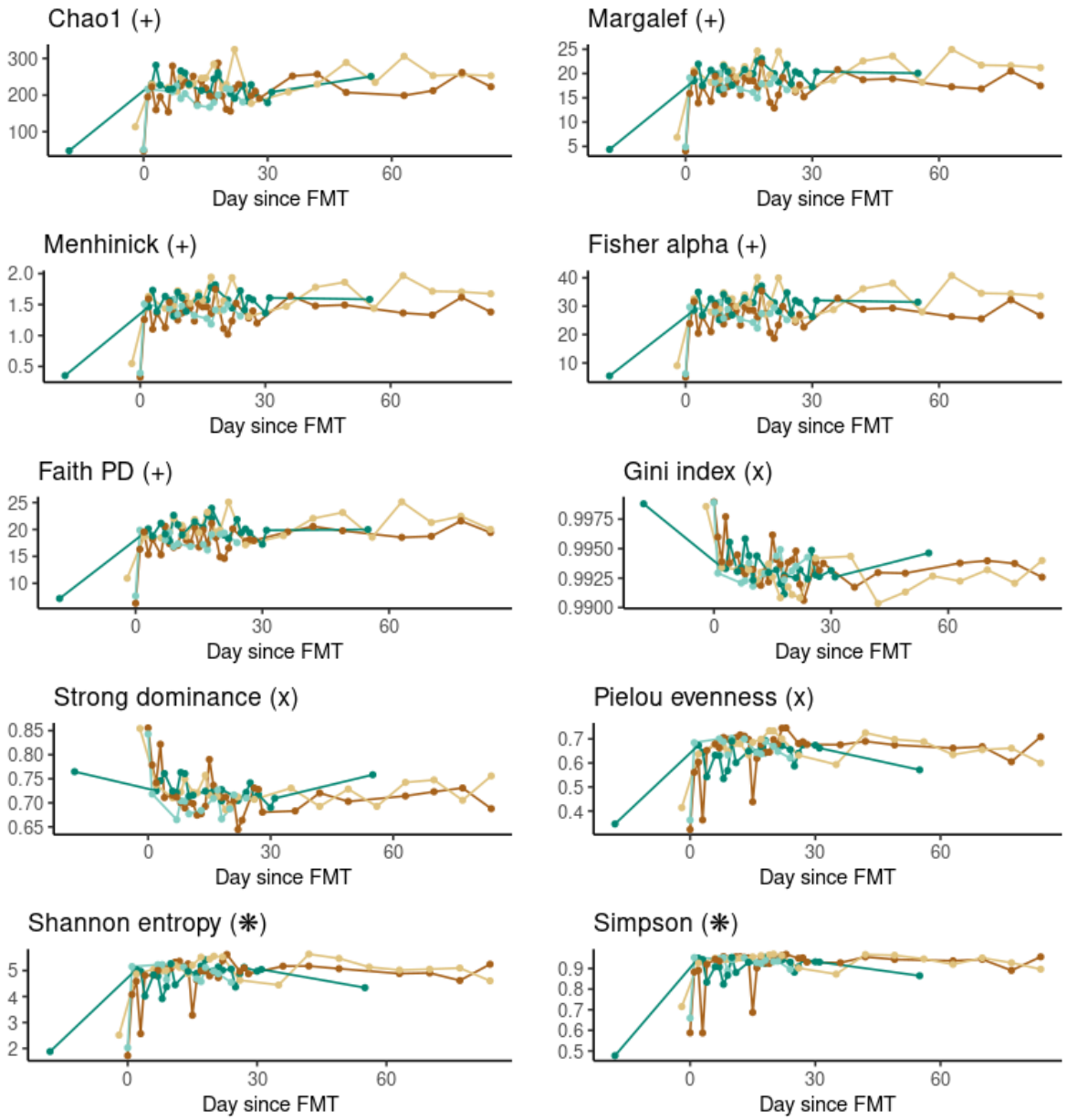
Supplementary Figure S5. Difference in distributions and means of different alpha metrics between healthy and IBD dataset (Lloyd-Price et al. 2019 and Qiita ID: 11549)



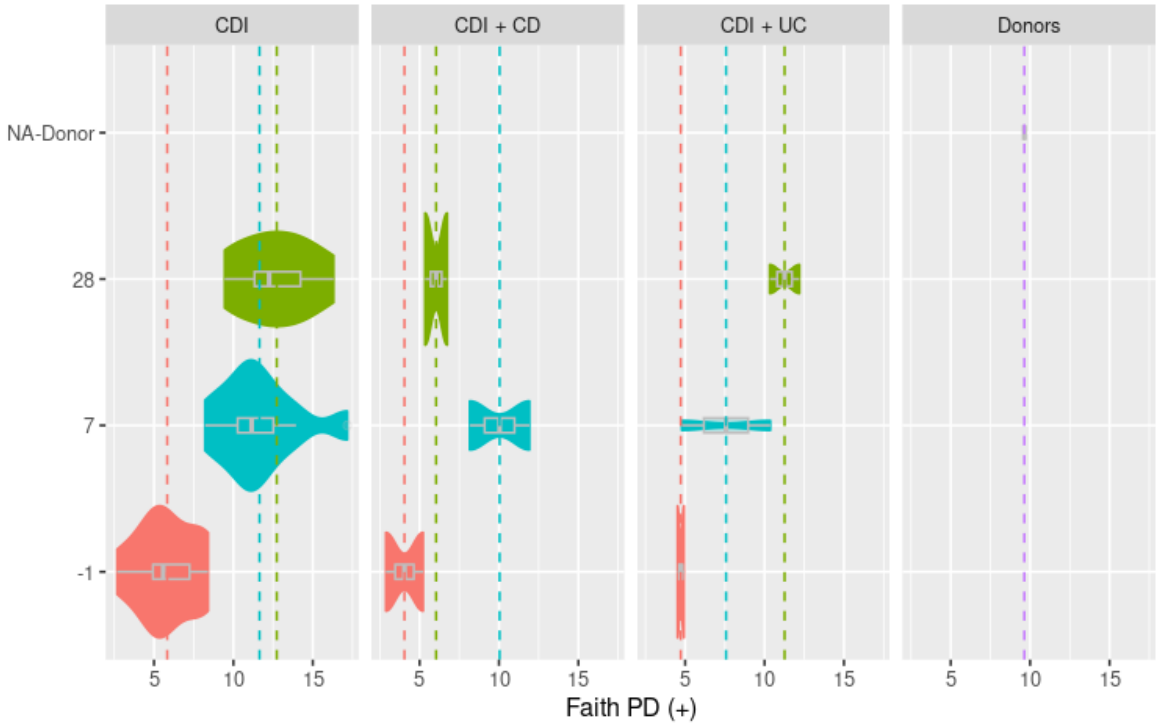
Supplementary Figure S6. Difference in distributions and means of different alpha metrics between controls and Crohn's samples (Vázquez-Baeza et al. 2018)

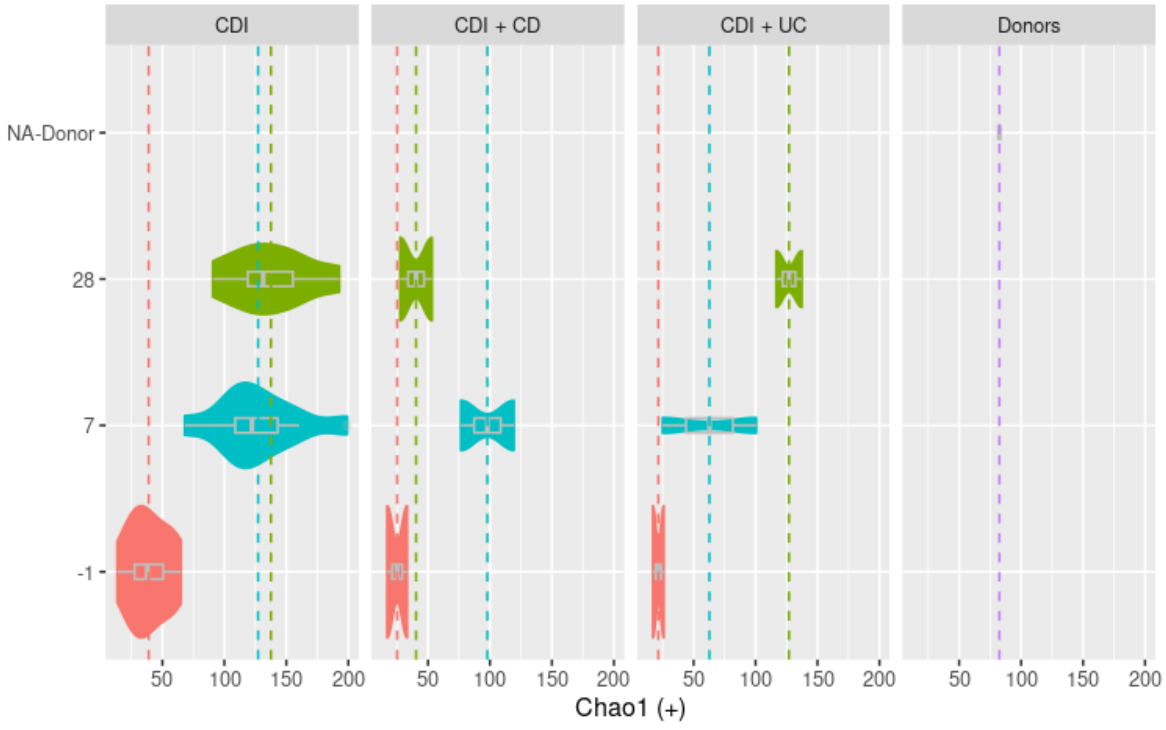


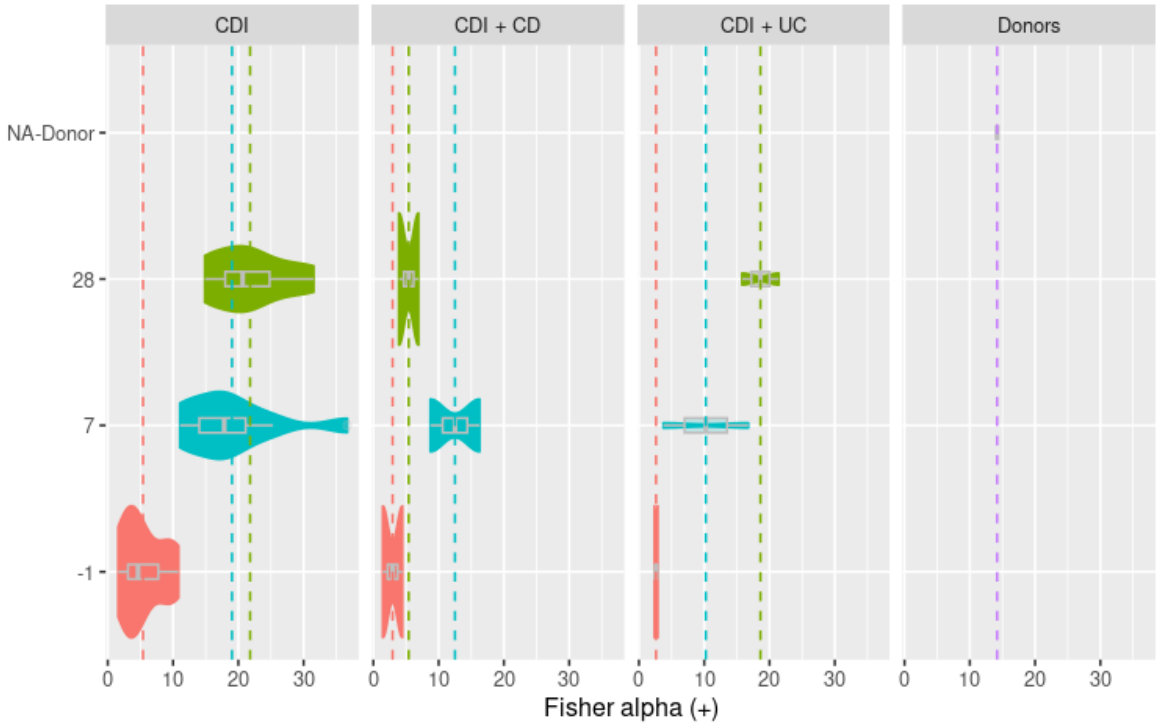
Supplementary Figure S7. Difference in distributions and means of different alpha metrics between controls and CDI samples (Khanna et al. 2016)

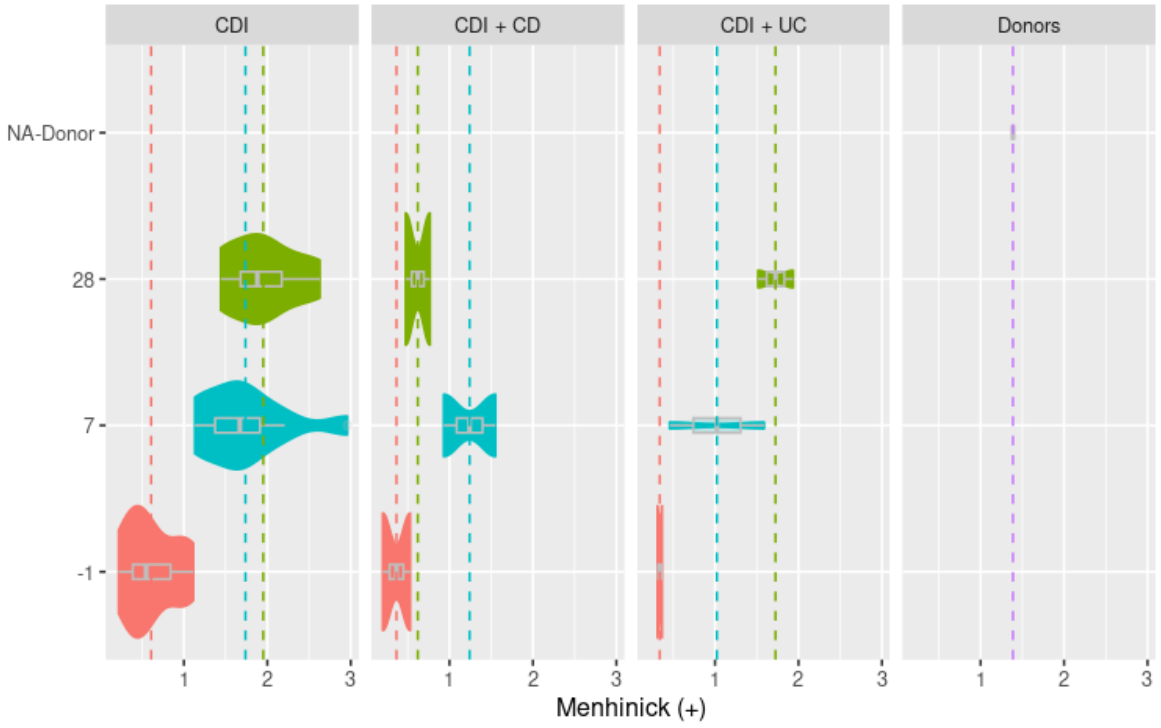


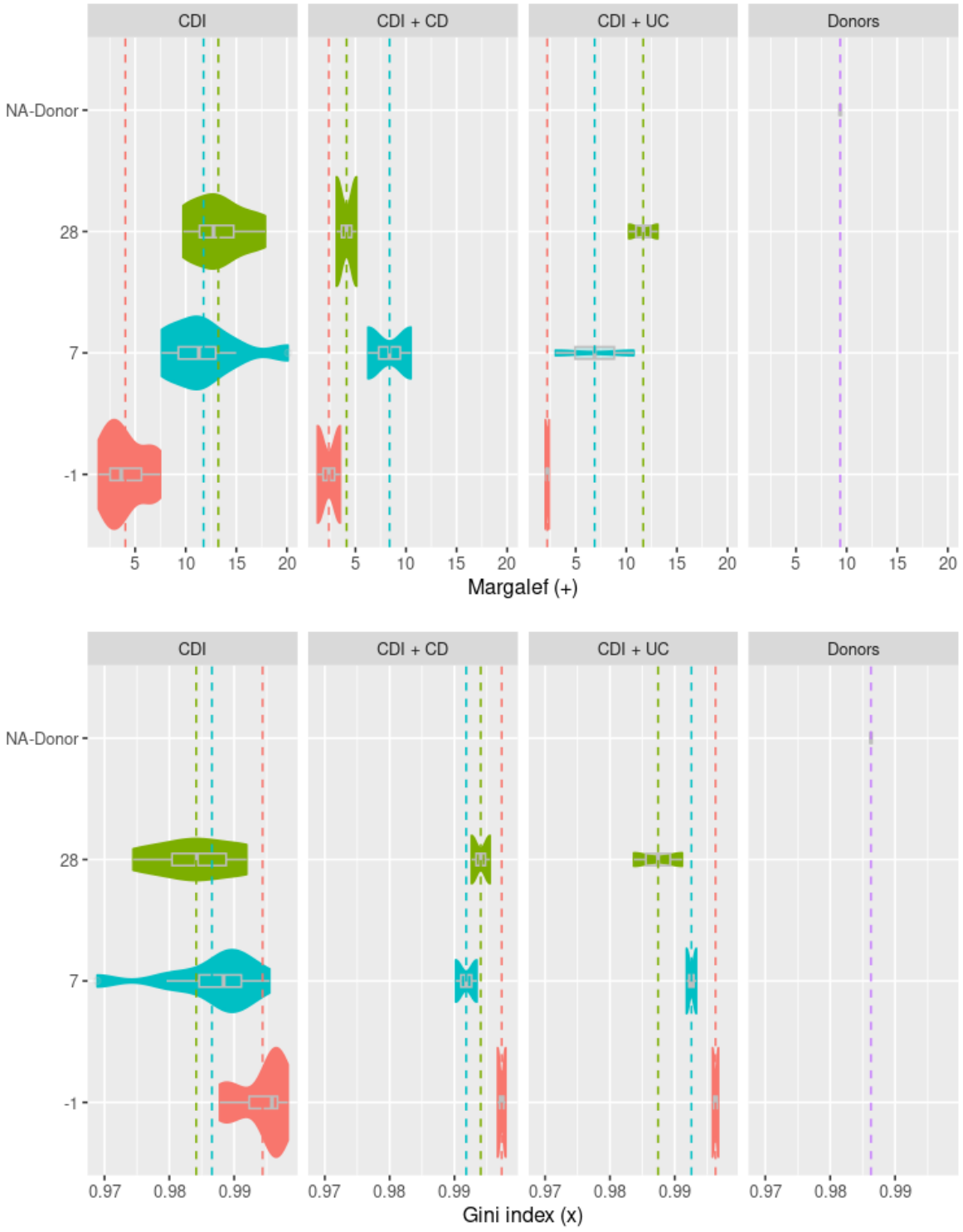
Supplementary Figure S8. Progression of alpha diversity metrics' value in time of CDI samples after FMT (Weingarden et al. 2015)

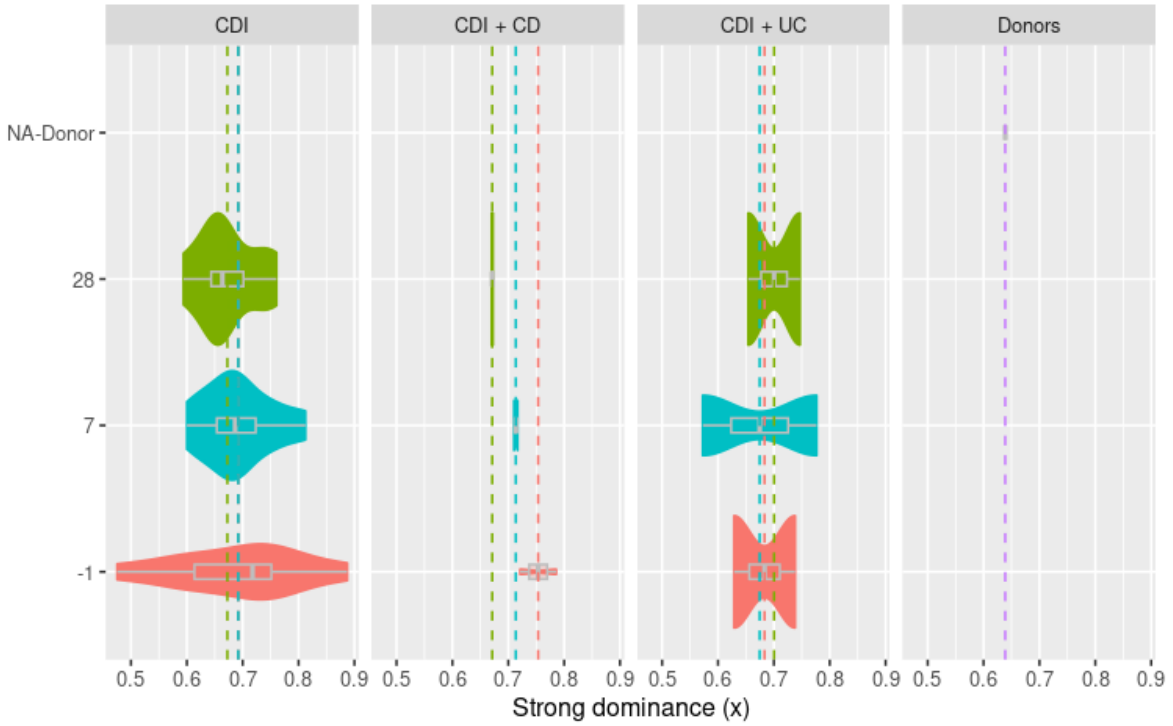

CDI
CDI + CD
CDI + UC
Donors
NA-Donor
28
7
-1
5
10
15
Faith PD (+)

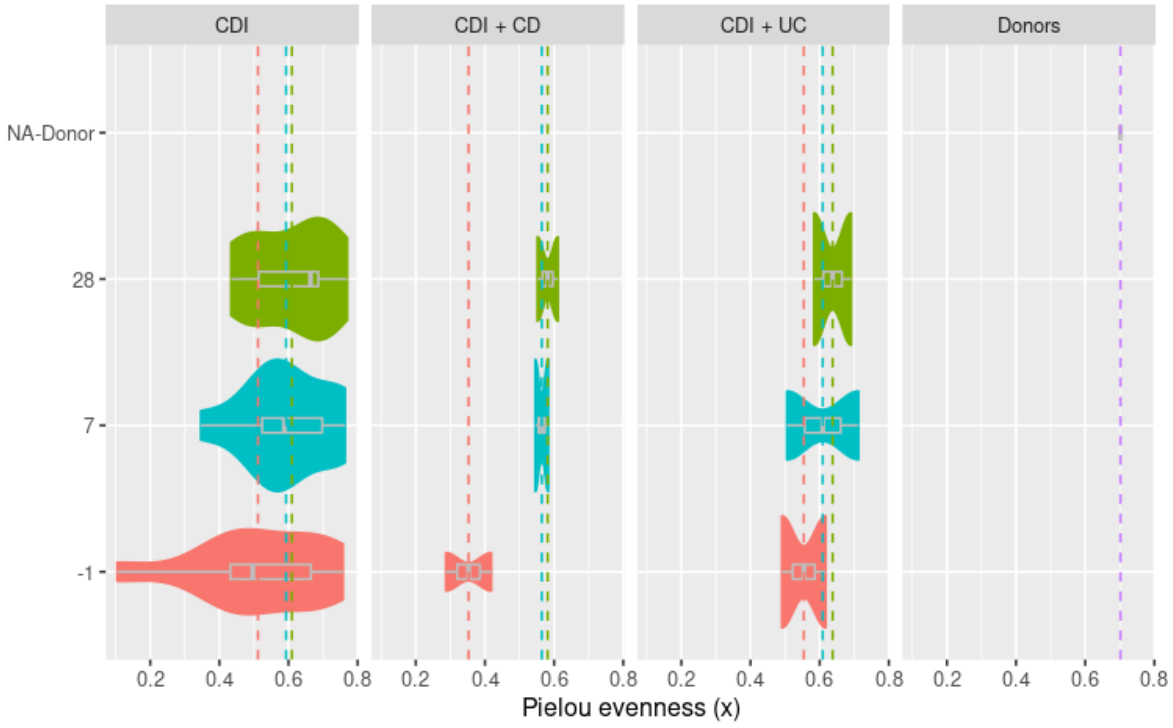

CDI
CDI + CD
CDI + UC
Donors
NA-Donor
28
7
-1
50
100
150
200
Chao1 (+)

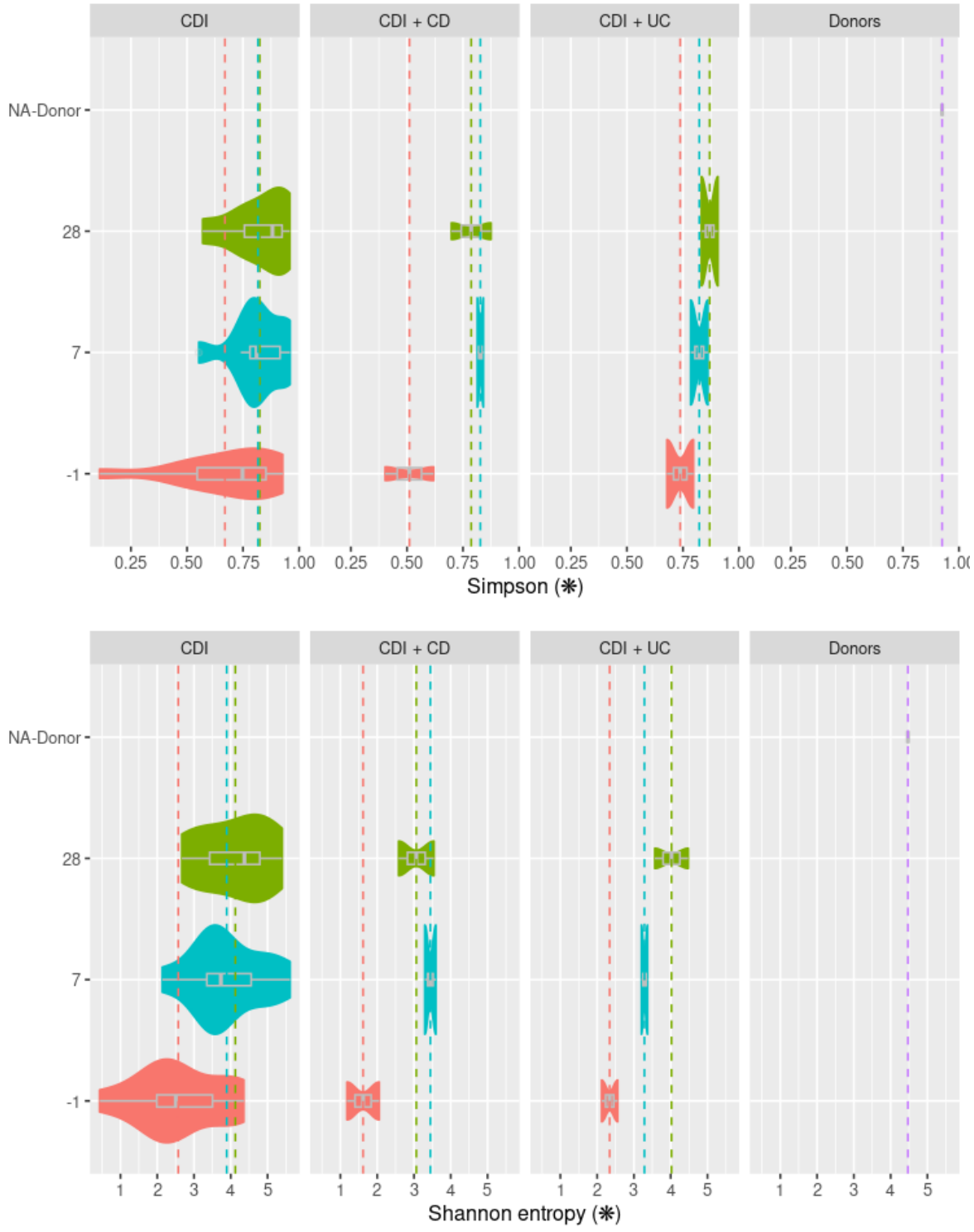

CDI
CDI + CD
CDI + UC
Donors
NA-Donor
28
7
-1
0
10
20
30
Fisher alpha (+)


CDI
CDI + CD
CDI + UC
Donors
NA-Donor
28
7
-1
1
2
3
Menhinick (+)


CDI
CDI + CD
CDI + UC
Donors
NA-Donor
28
7
-1
5
10
15
20
Margalef (+)
0.97
0.98
0.99
Gini index (x)


CDI
CDI + CD
CDI + UC
Donors
NA-Donor
28
7
-1
0.5
0.6
0.7
0.8
0.9
Strong dominance (x)


CDI
CDI + CD
CDI + UC
Donors
NA-Donor
28
7
-1
0.2
0.4
0.6
0.8
Pielou evenness (x)



Supplementary Figure S9. Difference of different alpha diversity value depending on day since FMT in CDI patients with different underlying conditions (none, CD or UC) from Khanna et al. 2017

References

Aira A et al. New Procedure to Maintain Fecal Microbiota in a Dry Matrix Ready to Encapsulate. Front Cell Infect Microbiol. 2022. 12:899257. Published 2022 Jun 10. doi:10.3389/fcimb.2022.899257

Bendel RB et al. Comparison of skewness coefficient. coefficient of variation, and Gini coefficient as inequality measures within populations. Oecologia. 1989;78(3):394-400. doi:10.1007/BF00379115

Bent SJ, Forney LJ. The tragedy of the uncommon: understanding limitations in the analysis of microbial diversity. ISME J. 2008 Jul;2(7):689-95. doi: 10.1038/ismej.2008.44. Epub 2008 May 8. PMID: 18463690.

Finotello F, Mastrorilli E, Di Camillo B. Measuring the diversity of the human microbiota with targeted next-generation sequencing. Brief Bioinform. 2018;19(4):679-692. doi:10.1093/bib/bbw119

Khanna S et al. Changes in microbial ecology after fecal microbiota transplantation for recurrent C. difficile infection affected by underlying inflammatory bowel disease. Microbiome. 2017;5(1):55. Published 2017 May 15. doi:10.1186/s40168-017-0269-3

Khanna S et al. Gut microbiome predictors of treatment response and recurrence in primary Clostridium difficile infection. Aliment Pharmacol Ther. 2016;44(7):715-727. doi:10.1111/apt.13750

Lloyd-Price J et al. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. Nature. 2019;569(7758):655-662. doi:10.1038/s41586-019-1237-9

McDonald D et al. American Gut: an Open Platform for Citizen Science Microbiome Research. mSystems. 2018;3(3):e00031-18. Published 2018 May 15. doi:10.1128/mSystems.00031-18

Pielou EC. The measurement of diversity in different types of biological collections. Journal of Theoretical Biology. 1966;13:131–144.doi:10.1016/0022-5193(66)90013-0

Strong WL. Assessing species abundance unevenness within and between plant communities. Community Ecology.2002;3(2):237–246.doi:10.1556/comec.3.2002.2.9

Thukral A. A review on measurement of Alpha diversity in biology. Agricultural Research Journal. 2017; 54:1. doi:10.5958/2395-146X.2017.00001.1

Vázquez-Baeza Y et al. Guiding longitudinal sampling in IBD cohorts. Gut. 2018;67(9):1743-1745. doi:10.1136/gutjnl-2017-315352

Website: CD Genomics. The Use and Types of Alpha-Diversity Metrics in Microbial NGS.https://www.cd-genomics.com/microbioseq/the-use-and-types-of-alpha-diversity-metrics-in-microbial-ngs.html (17 April 2023, date last accessed)

Website: Qiita. Metaomics Reveals Microbiome Based Proteolysis as a Driver of Ulcerative Colitis Severity. https://qiita.ucsd.edu/study/description/11549 (16 January 2023, date last accessed)

Weingarden A et al. Dynamic changes in short- and long-term bacterial composition following fecal microbiota transplantation for recurrent Clostridium difficile infection. Microbiome. 2015;3:10. Published 2015 Mar 30. doi:10.1186/s40168-015-0070-0

Zheng M et al. Using Lorenz Curve and Gini Coefficient to Reflect the Inequality Degree of S&T Publications: An Examination of the Institutional Distribution of Publications in China and other Countries. Economics. 2008.