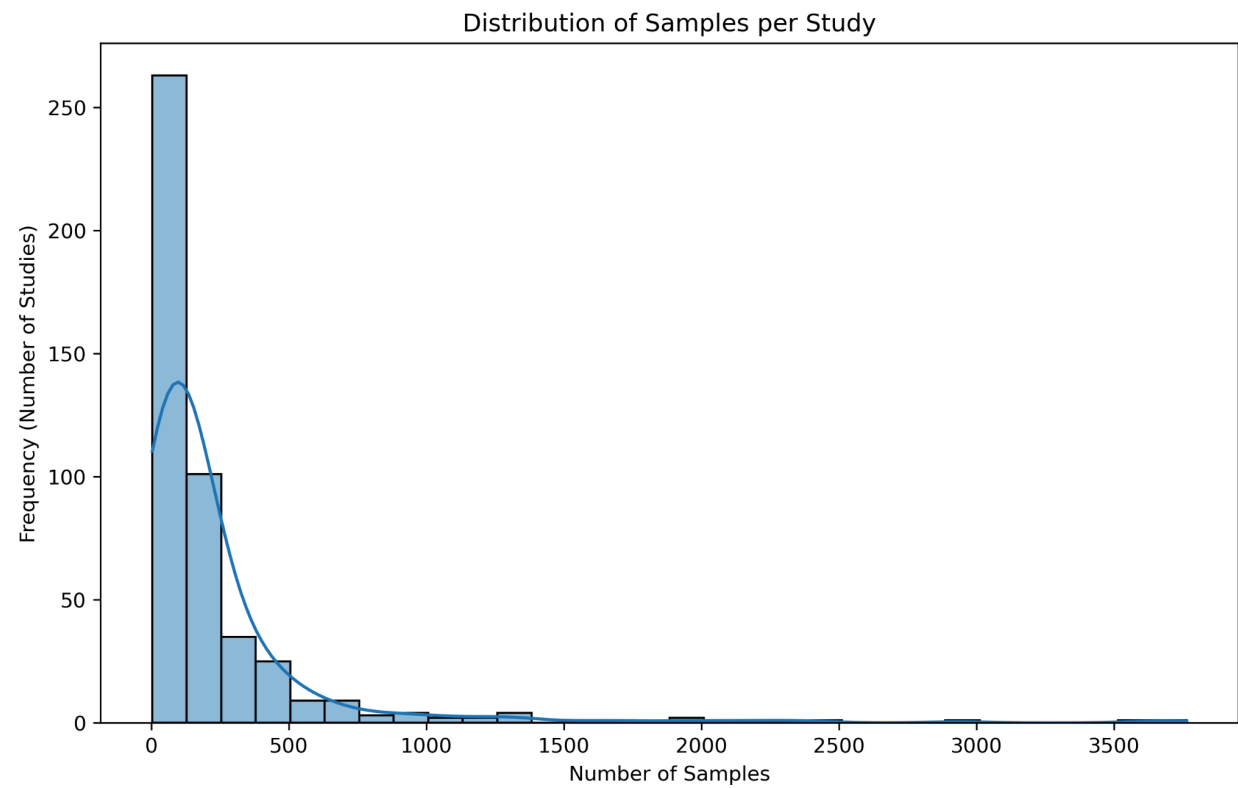


Used environment: Human

Exploratory analysis

Dataset composition

Number of samples: 108276, number of studies: 467
Samples per study: Min: 4, Max: 3765, Mean: 232, Median: 110

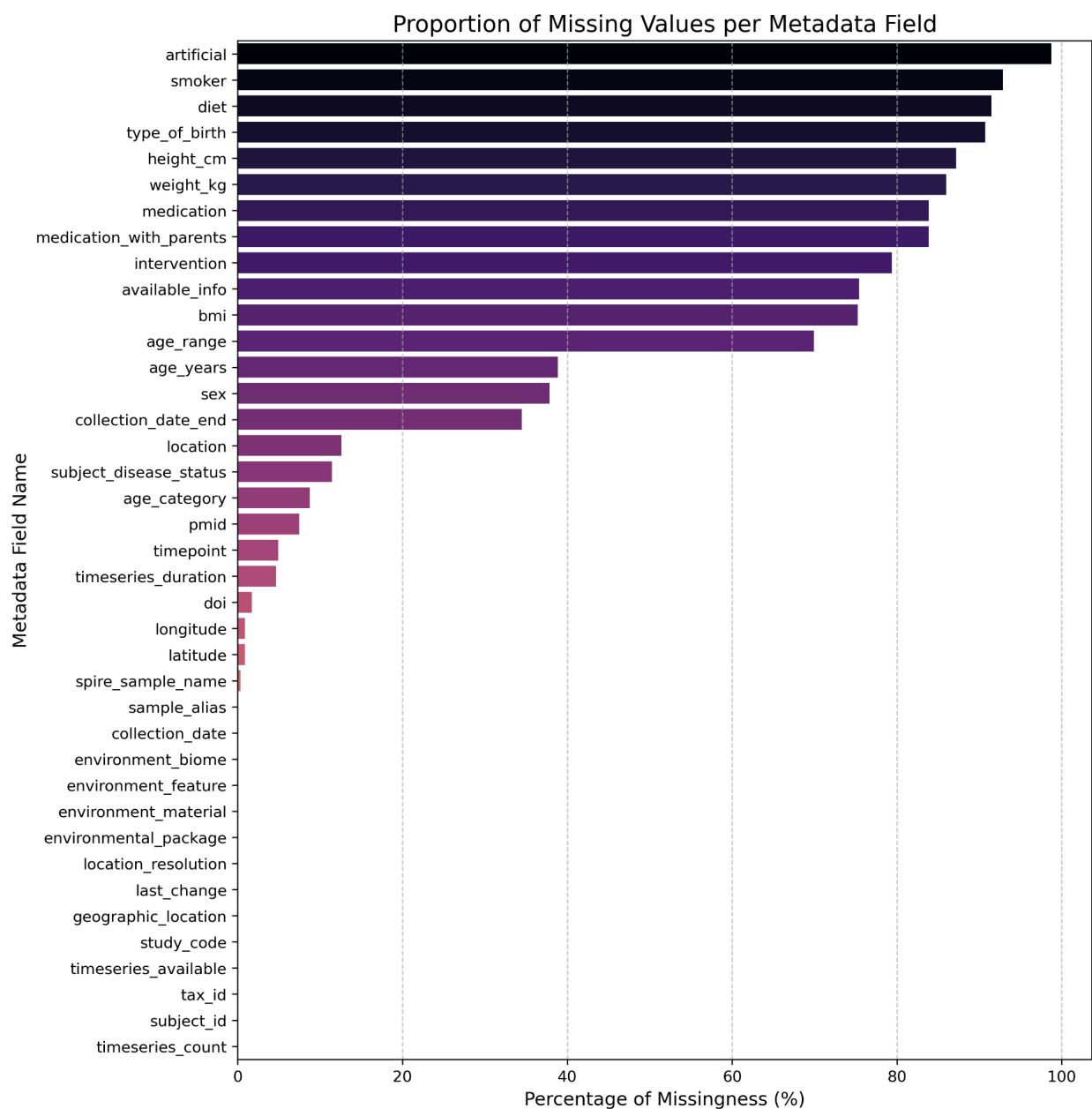


The dataset is imbalanced, with few studies having a large amount of samples; the frequency plot is highly skewed to the left. The mean number of studies is 232, but the median is twice smaller – 110, which highlights that most studies have a moderate number of samples, and one huge study in the dataset, which could influence the overall analysis.

Missing values

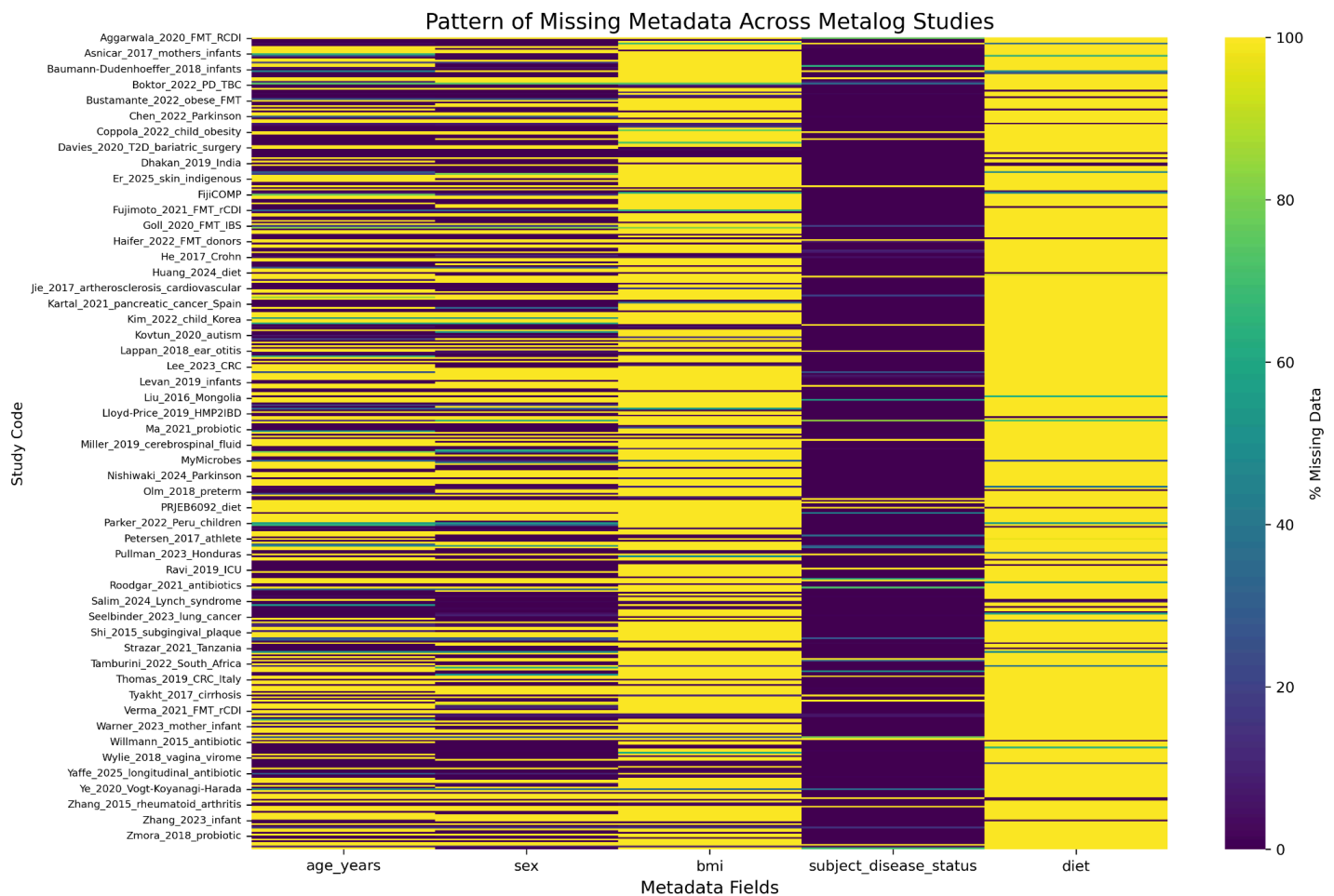
Metadata Field	Missingness (%)
artificial	98.71
smoker	92.85
diet	91.47
type_of_birth	90.71
height_cm	87.17
weight_kg	85.96
medication	83.82
medication_with_parents	83.82

intervention	79.36
available_info	75.42
bmi	75.24
age_range	69.93
age_years	38.86
sex	37.84
collection_date_end	34.51
location	12.59
subject_disease_status	11.46
age_category	8.78
pmid	7.45
timepoint	4.94
timeseries_duration	4.66
doi	1.7
longitude	0.9
latitude	0.9
spire_sample_name	0.34
sample_alias	0
collection_date	0
environment_biome	0
environment_feature	0
environment_material	0
environmental_package	0
location_resolution	0
last_change	0
geographic_location	0
study_code	0
timeseries_available	0
tax_id	0
subject_id	0
timeseries_count	0



A lot of the metadata fields are missing in more than 50% of the studies. These fields are probably specific to certain studies and were not collected in other studies.

Core identifiers like *geographic_location* are complete across the entire dataset (0% of missingness), which makes them perfect ground for further analysis. *age_years* and *sex* have roughly 38% missingness, making them good candidates for explanatory variables.



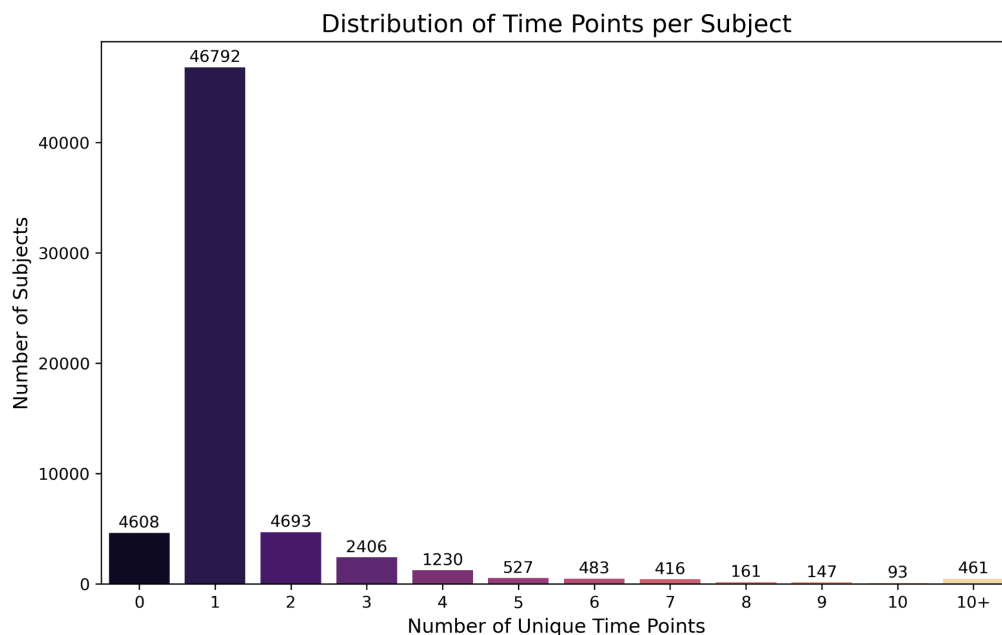
As we can see on the heatmap, a lot of studies have solid colors for certain metadata fields (violet or yellow), and only a few of them have some intermediate colors. This fact indicates that the missing values definitely have patterns, which are dictated by the study design.

Longitudinal and repeated data

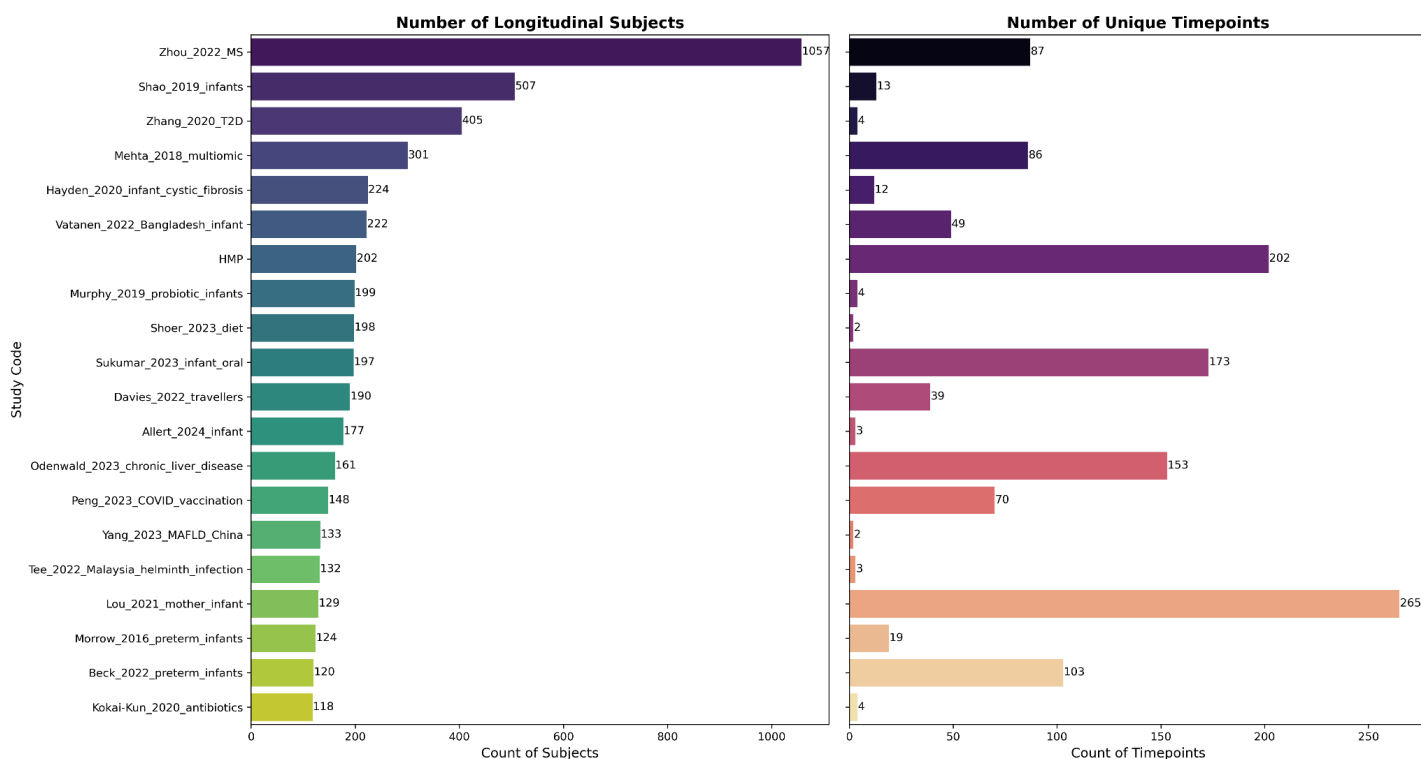
Number of studies with longitudinal data	204
Number of longitudinal subjects	10617
Total samples involved in longitudinal studies	48233
Maximum time points for a single subject	193
Maximum unique time points in a single study	311
Study with the most time points	Shaffer_2023_infants_vaccination

There are 204 studies with longitudinal subjects. Overall, the dataset contains 48233 samples collected longitudinally across 10617 subjects. For a single subject, the maximum is 193 timepoints, and the Shaffer_2023_infants_vaccination study has the biggest number of unique timepoints for sample collections: 311.

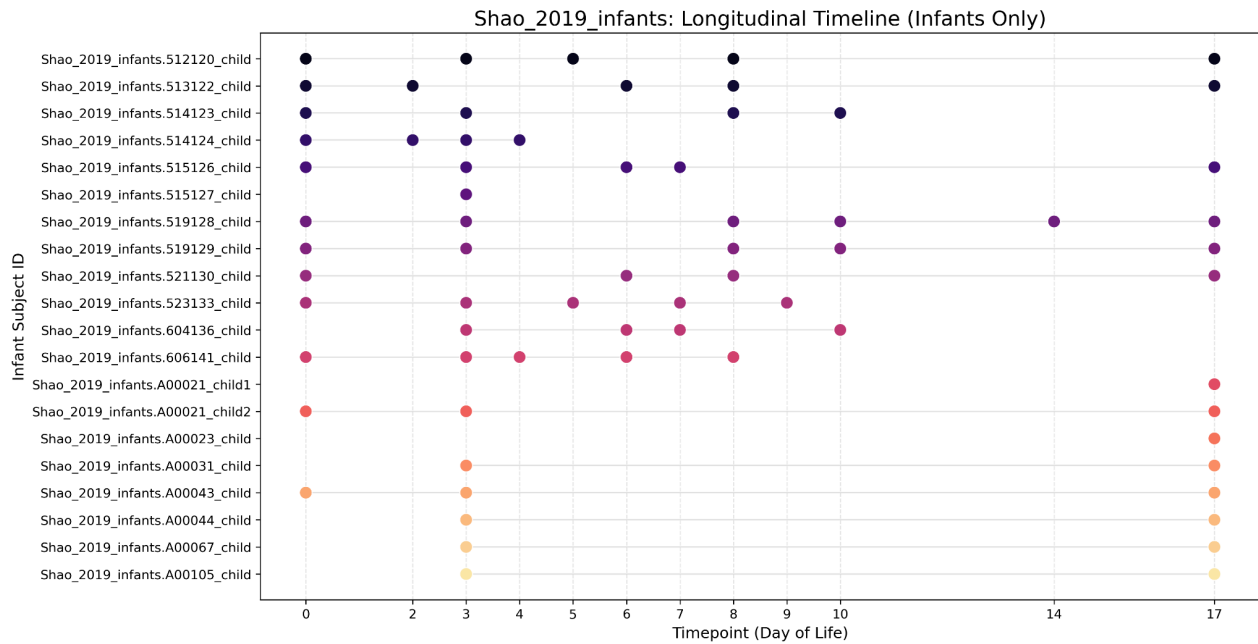
In most cases, longitudinal subjects have no more than 7 unique timepoints. Although there is a portion of subjects with more than 10 unique timepoints.



Longitudinal Study Heterogeneity: Subject Volume vs. Sampling Depth (first 20 studies)



For the top 20 longitudinal studies by the number of subjects, unique timepoints are distributed randomly with 87 timepoints for the biggest study, and 202 timepoints for Top7 study.



A closer look at the Shao_2019_infants study reveals high temporal heterogeneity. The sample collection is characterized by irregular intervals and missing timepoints for specific subjects, which could introduce bias in downstream longitudinal modeling. To construct a balanced dataset, a subsetting strategy will be required to focus on the most complete timepoints.

Number of duplicate sample_alias entries: 0

Number of samples with identical metadata profiles: 20844

While every sample has unique identifier, there are over 20 thousands samples with identical metadata profile (same study, subject, timepoint, and disease status). This indicates that whether the same sample was sequenced several times, or the same subject was sampled at the same timepoint, but from different locations, or the same sample was processed for both 16S rRNA and Metagenomics.

For the downstream analysis, these must be deduplicated by selecting a single representative sample per subject/timepoint to avoid inflating the statistical power of specific studies.

Studies with the most profile duplicates:	
Li_2025_Singapore_skin	3572
HMP	2005
Zhou_2022_MS	1850
Larson_2022_elderly	1348
NIH_skin	1116
Zmora_2018_probiotic	894
Kumbhari_2024_IBD	650
Shoer_2023_diet	620
Hannigan_2015_skin_virome	577
Shen_2023_skin	562

Research question and subsetting

Does the fecal microbial signature of Crohn's Disease patients exhibit sexual dimorphism, and how does this signature evolve between adolescence and adulthood?

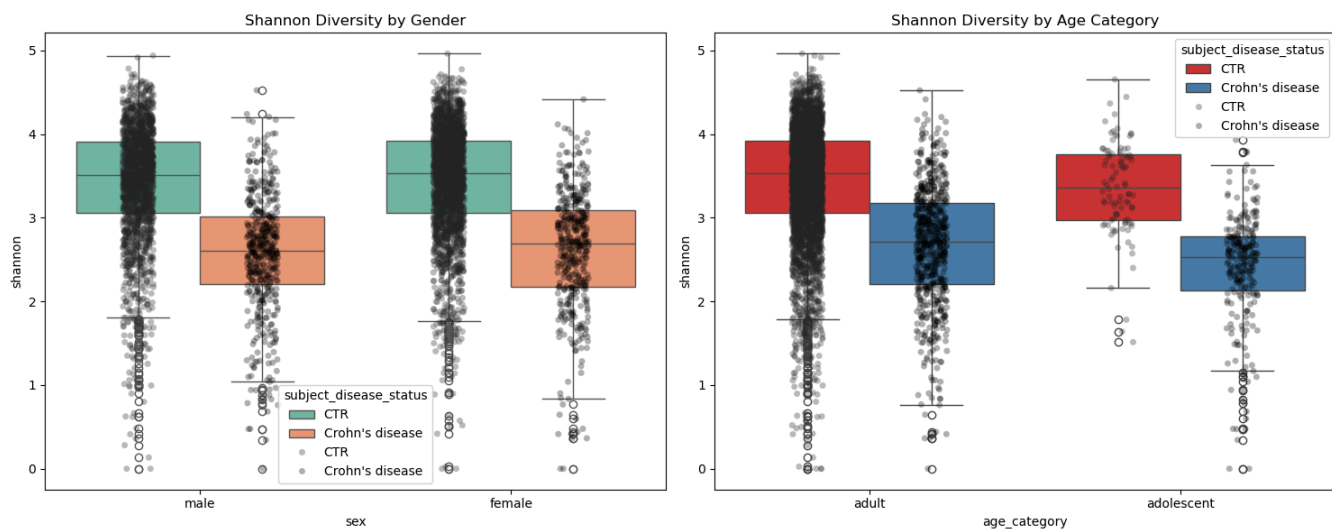
The data has been subset to delete metadata profile duplicates, and include only those control and Crohn's disease samples, which have information about sex, belong to the adult or adolescent groups, and were collected from fecal material.

Overall, missingness in covariates were 5.9% for age category, 25.4% for sex, and 0% for sample material. Final stats for the dataset is presented in the table below:

age_category	adolescent		adult		All
sex	female	male	female	male	
subject_disease_status					
CTR	53	52	2987	2828	5920
Crohn's disease	85	166	387	357	995
All	138	218	3374	3185	6915

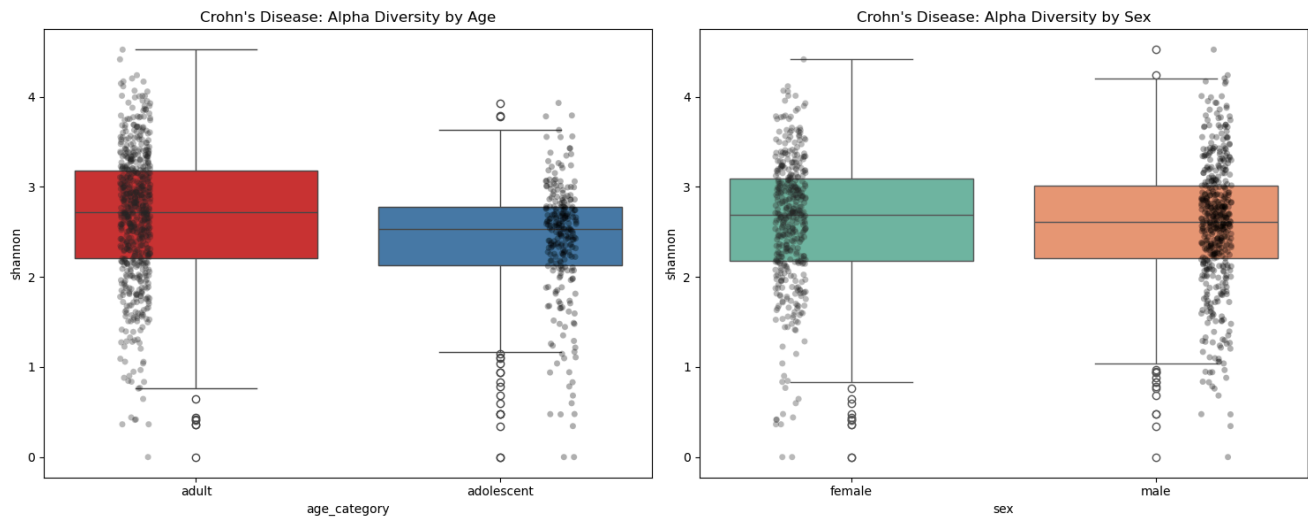
Comparative analysis

At the end of the abundance data addition, we have 5150 samples to analyze.



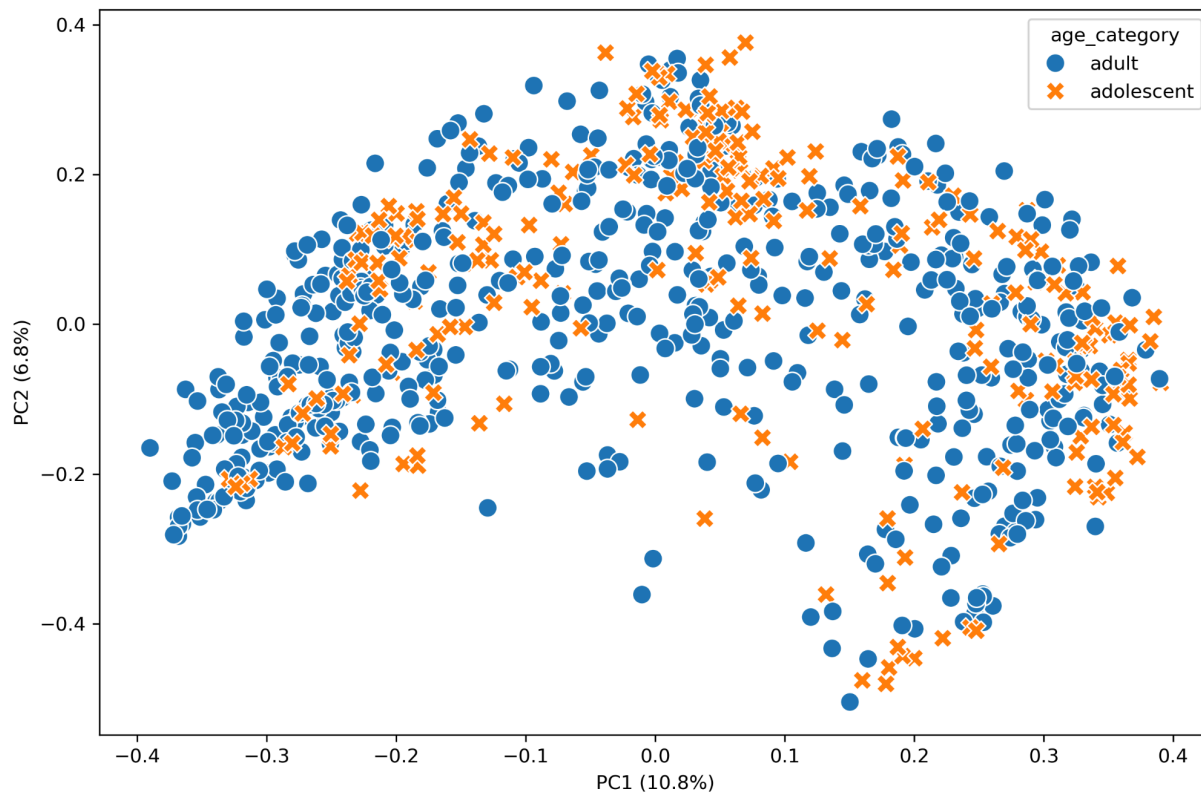
Comparison	Raw P	Adj P	Significant
Overall	2.95E-191	1.47E-190	TRUE
Gender: female	2.94E-94	3.67E-94	TRUE
Gender: male	4.74E-97	7.91E-97	TRUE
Age: adult	4.20E-124	1.05E-123	TRUE
Age: adolescent	5.55E-29	5.55E-29	TRUE

Alpha-diversity analysis (Shannon Index) using the Mann-Whitney U test with Benjamini-Hochberg (FDR) correction confirmed that CD patients possess a significantly different microbial signature compared to healthy controls. This divergence from the "healthy" state remains highly significant ($p < 0.05$) when stratified by sex, indicating that while the disease signature is robust, its presence is consistently felt across both males and females.



Comparison (within Crohn's)	Raw P	Adj P	Significant
Crohn's: Adult vs Adolescent	1.92E-07	3.84E-07	TRUE
Crohn's: Male vs Female	2.46E-01	2.46E-01	FALSE

Further analysis specifically within the Crohn's Disease cohort revealed a significant difference in alpha diversity between adolescents and adults.



PERMANOVA P-value: 0.0010

Test Statistic (Pseudo-F): 13.1620

While visual inspection of the Bray-Curtis PCoA plot showed overlapping clusters, PERMANOVA analysis revealed a highly significant difference in microbial community composition between adults and adolescents with Crohn's disease ($F = 13.16$, $p = 0.001$). This indicates that while both groups share a similar 'core' microbiome, there are distinct shifts in community structure associated with age category that are statistically robust after 999 permutations.

Conclusion

These findings indicate that the CD microbial signature is not static; it evolves significantly as patients transition from adolescence to adulthood, characterized by shifts in community structure that are statistically robust after 999 permutations.