Coursework Declaration and Feedback Form

The Student should complete and sign this part

Student Number: 3047154z	Student Name: Zhen Zeng
Programme of Study : MSc in Computer Systems Engineering	
Course Code: ENG5059P	Course Name: MSc Project
Name of **First** Supervisor: Umer Ijaz	Name of **Second** Supervisor: Chinyere Rosemary Ngwu
Title of Project: **Effect of Pre-filtering Strategies on Microbiome Diversity**	

Declaration of Originality and Submission Information

I affirm that this submission is all my own work in accordance with the University of Glasgow Regulations and the School of Engineering requirements Signed (Student) : ZHEN ZENG	E N G 5 0 5 9 P
Date of Submission : October 20th 2025	

Feedback from Lecturer to Student – to be completed by Lecturer or Demonstrator	
Grade Awarded: Feedback (as appropriate to the coursework which was assessed):	
Lecturer/Demonstrator:	Date returned to the Teaching Office:

Effect of Pre-filtering Strategies on Microbiome Diversity

Student Name: Zhen Zeng
Student ID: 3047154Z
Supervisor: Dr Umer Zeeshan Ijaz

October 20th 2025

ACKNOWLEDGEMENT

This dissertation would not have been possible without the guidance and support of many individuals.

I would like to express my sincere gratitude to Dr Umer Ijaz, my supervisor, for his continuous guidance, constructive feedback, and encouragement throughout this research.

I also wish to thank Gladys Maria Pangga for her valuable assistance during data analysis and discussions, and all members of the Ijaz Lab for their technical and academic support.

Special thanks go to the School of Engineering at the University of Glasgow for providing the research environment and resources that enabled this work.

I am also grateful to my classmates and friends for their kind help and insightful discussions during this MSc journey.

Finally, I would like to express heartfelt thanks to my family for their patience, understanding, and constant encouragement, which gave me the strength to complete this work.

ABSTRACT

Feature filtering is a critical but under-reported step in microbiome data preprocessing.
This study systematically evaluated how single-feature filters influence α-diversity, β-diversity, and PCoA geometry in 16S rRNA amplicon data.
A unified diagnostic framework was applied to two complementary designs: a cross-sectional granular-biofilm community (Project 1) and a longitudinal broiler-cecum microbiome (Project 2).
Seven families of filters were compared, including abundance-based, variability-based, distribution-based, relationship-based, transformation-specific, model-based, and network-based categories.
Geometric preservation was assessed using Mantel correlation, symmetric Procrustes correlation, and PERMANOVA R^2, across Bray–Curtis and UniFrac distances.
Results showed that abundance- and prevalence-oriented filters most consistently preserved β-diversity geometry, with high Mantel and Procrustes concordance and minimal variation in PERMANOVA R^2.
Variability- and mild distribution-based filters also behaved stably when thresholds aligned with data sparsity.
In contrast, distribution-shape, transformation-specific, and model-linked rules occasionally distorted ordinations or inflated apparent R^2.
Design-specific effects were observed: stratum-specific thresholds were preferable for cross-sectional datasets differing in sparsity (DNA vs cDNA in P1), while time-aware filters improved phase separation in longitudinal designs (P2) without altering geometry.
These findings emphasise that filter selection, rather than intensity, governs the stability of ecological structure.
Abundance- or prevalence-based single filters are recommended as default preprocessing options, with diagnostic reporting of Mantel, Procrustes, and R^2 values to ensure transparency and reproducibility.
This work provides evidence-based guidance for choosing single-filter strategies that balance feature reduction with ecological fidelity, establishing a reproducible baseline for future microbiome analyses.

Keywords: 16S rRNA amplicon; feature filtering; β-diversity; PERMANOVA R^2; Procrustes analysis; microbiome methods.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

1.1 Background and Workflow Overview

Microbial ecologists increasingly rely on DNA sequencing to characterize community composition across environments and hosts [20,22]. Two primary strategies are used: amplicon/marker-gene sequencing (e.g., 16S rRNA) and shotgun metagenomics [4,5]. For bacteria and archaea, the 16S rRNA gene is the standard phylogenetic marker because it is ubiquitous and contains nine hypervariable regions (V1–V9) that support taxonomic and phylogenetic resolution [5,16,27]. A typical amplicon workflow starts with sample collection and DNA extraction, continues with PCR of a selected region (often V3–V4), and proceeds to high-throughput sequencing and bioinformatics processing [9, 16]. Processing includes pre-processing and quality control, clustering or denoising to define OTUs (≈97% similarity) or ASVs (single-nucleotide resolution), taxonomic assignment against SILVA/Greengenes/RDP, and construction of a feature table (samples × OTUs/ASVs) for downstream analysis[4,9,27]. ASV pipelines (e.g., DADA2 [10], Deblur[2]) often show higher sensitivity and specificity than OTU clustering, although both remain in use in standard operating procedures (e.g., QIIME 2) [5,16,27].

1.2 Common Tools and Metrics

Downstream analysis summarizes within-sample α-diversity and between-sample β-diversity, then visualizes structure with ordination and tests effects with distance-based statistics [9]. α-diversity commonly uses Observed OTUs, Chao1, Shannon, and Inverse Simpson, with convenient implementations in phyloseq [9,16,18,20]. β-diversity typically relies on Bray–Curtis [8] (compositional overlap) and UniFrac [17] (unweighted/weighted; phylogenetically aware), computed via vegan and related R packages [9,20,23]. These dissimilarities are most often embedded with principal coordinates analysis (PCoA [12]) or non-metric MDS; PCoA is the default ordination in many pipelines, including QIIME 2 [4,7,9]. For hypothesis testing, PERMANOVA [3] estimates the proportion of variance in the distance matrix explained by factors of interest and is widely used in community ecology [9,20]. Within this standard pipeline, the present thesis focuses on preprocessing feature filtering—the removal of low-information or unstable features before normalization, distance calculation, and ordination—because this step can materially influence α/β-diversity, ordination geometry, and subsequent statistical inference [4,9,16].

Microbiome feature tables are high-dimensional, sparse, and compositional. Most taxa are infrequent, so matrices contain many zeros; many 16S datasets even observe far <10% of possible entries, which destabilizes distance estimates and ordinations (zero inflation, over-dispersion) [4,9]. Compositionality arises because each sample's total counts are constrained by sequencing depth; closure induces dependencies among taxa, so shifts in one relative abundance force changes in others even if absolute biology is unchanged—ignoring this yields spurious correlations and misleading geometry in downstream analyses [9,27]. Biology and technology add further noise: intragenomic 16S copy number varies across taxa, biasing quantitative comparisons; PCR and sequencing errors create artefactual variants if not removed [16,27]. Collectively, sparsity, zero inflation, and closure can bias β-diversity (e.g., Bray–Curtis/UniFrac), distort PCoA, and undermine distance-based inference. Hence compositionally aware methods and careful preprocessing are recommended before computing distances and running tests [4,9].

A major practical symptom is the proliferation of low-abundance, error-prone features. Amplicon PCR can yield spurious OTUs that inflate richness and skew diversity metrics; empirically, low-abundance OTUs show higher coefficients of variation, indicating poorer detection accuracy [20]. Including such features reduces statistical power, inflates the

multiple-testing burden, and can break assumptions of normalization when zeros/near-zeros dominate [13]. Therefore, pipelines commonly pre-filter rare or low-prevalence features prior to normalization, distance calculation, ordination, and PERMANOVA, in order to stabilize geometry and improve interpretability [13,20]. In this thesis, feature filtering is treated as a principled upstream step—motivated by sparsity, zero inflation, compositional closure, and technical artefacts—to make α/β-diversity estimates, PCoA structure, and distance-based tests more reliable [4,9].

1.3 A concise mini-review of filter families

Feature filtering methods can be grouped by the property they target. Abundance-based filters operate on detectability or overall level (total count, mean, or prevalence) and remove ultra-rare or sporadic features that contribute mostly noise. A common rule is the mean-abundance screen, which drops features with very low average counts across samples [13]. Variability-based filters keep features that change enough to be informative and down-weight flat or noise-dominated profiles; the IQR(interquartile range) filter uses interquartile range as a robust dispersion measure [13]. Distribution-based filters look at the shape of the profile (e.g., entropy, zero proportion, skewness) to exclude degenerate or highly zero-inflated patterns [13]. Relationship-based filters align selection with the study question; for example, time_correlation retains features whose trajectories track a continuous time variable in longitudinal designs [13]. These four families act as light-touch denoisers. They reduce stochastic zeros and stabilize distances prior to Bray–Curtis/UniFrac and ordination, which is recommended for sparse and compositional 16S tables[4,9,13].

Other families tailor filtering to transformations, generative assumptions, or ecological structure. Transformation-specific filters are applied after a chosen transform so that keep/remove decisions reflect the analysis scale actually used downstream. The mad_vst rule identifies features with sufficient post-transform dispersion under a variance-stabilizing transform (VST) [13]. Model-based filters use simple statistical models to separate signal from noise; the zinb_aic rule keeps features better explained by a zero-inflated negative binomial model fit (lower AIC), and limit-of-quantification variants screen features near detection boundaries [13]. Network-based filters use co-occurrence or association structure; network_connectivity retains features that are embedded in well-connected modules, improving interpretability and reducing idiosyncratic singletons before multivariate tests [13]. In practice, no single family is universally optimal. Analyses often combine a general noise-control step (abundance/variability/distribution) with a design-aware step (relationship-based for longitudinal data, or transformation-/model-/network-aware when those assumptions are central). This sequencing aligns with compositionally aware workflows and helps stabilize ordination geometry and distance-based inference in microbiome studies[4,9,13]. Full definitions, inputs, and threshold choices are provided in Methods (§2).

1.4 Research gaps

Despite mature 16S pipelines, pre-filtering practices remain heterogeneous and under-documented. Many studies remove “low-abundance/low-prevalence” features but do not justify cut-offs or report threshold sweeps, while upstream choices (primer region, OTU vs ASV, reference database) add further variability and can shift α/β-diversity and even reverse conclusions across pipelines (standardization/bias)[5, 20]. At the same time, most commonly used distances (Bray–Curtis; weighted/unweighted UniFrac) are not subcompositionally coherent; combined with relative scaling or rarefaction, this can induce spurious associations and depth-dependent distance shifts if compositionality is ignored [1,9,19,27,34]. In practice, reports often discuss filtering qualitatively but do not quantify how choices affect the geometry of β-diversity embeddings or the stability of distance-based inference. Clear guidance exists for compositionally aware analysis, but the field still lacks consensus on how

strict filters should be and how to demonstrate robustness to those choices [9].
A second gap is limited cross-design benchmarking and scarce, standardized diagnostics. In longitudinal data, between-subject differences can dominate ordinations and repeated-measures structure is often ignored; dimensionality-reduction methods that jointly handle phylogeny and compositionality remain under-developed, and there is no single method suitable for all designs [4]. As a result, filter choices are rarely tested for their impact on key diagnostics such as Mantel correlation and Procrustes m^2 between pre- and post-filtering ordinations, or on the stability of PERMANOVA R^2 across thresholds and alternative distances (Bray–Curtis, UniFrac, Aitchison) [4,9]. Finally, time-aware and strata-aware filtering are underused: longitudinal studies seldom screen by time-correlation, and strata with different sparsity or detection limits (e.g., DNA vs cDNA) often share a single global threshold, risking biased contrasts; residual technical/biological artefacts (multi-copy 16S, read-length limits, database gaps, contamination) further complicate inference [13,16,27]. These gaps motivate a diagnostics-driven, design-robust assessment that (i) compares cross-sectional and longitudinal cohorts with a common filter catalogue, (ii) reports threshold sweeps and alternative distances, and (iii) incorporates time- and strata-aware rules with transparent reporting [4,9,13].

1.5 Data context and motivation

This thesis conducts a secondary analysis of two published datasets with contrasting designs. Project 1 (P1) is a cross-sectional study of anaerobic granular biofilms with DNA and cDNA layers; the source reports 24 samples and 2,175 ASVs, and the rarefied working table used here contains 24 × 1,829 ASVs [32]. Project 2 (P2) is a longitudinal broiler cecum cohort spanning days 3–35; the source reports n = 379 with 18,588 OTUs (97%), while the present analysis uses a retained subset (337 samples; 2,309 OTUs) after the upstream QC and study-specific inclusion rules described in Section 2 (Methods). A common filter catalogue is applied to both datasets to enable side-by-side comparisons of α/β-diversity, ordination similarity, and PERMANOVA R^2 across alternative distance measures, while holding implementation details constant. Unless otherwise noted, dataset characteristics are taken as reported in the original publications.

1.6 Objectives

This thesis has four methodological objectives and two comparative objectives.

i. Filter catalogue (methodological). Compile a reproducible catalogue of microbiome pre-filters spanning seven families (abundance, variability, distribution, transformation-specific, model-based, network-based, and relationship-based), with clear operational definitions and documented parameters.
ii. Quantified impacts (methodological). Quantify the downstream effects of single-filter application on within-sample diversity (α-diversity), between-sample structure (β-diversity), and ordination geometry, using Mantel correlation (distance concordance), symmetric Procrustes correlation (geometric concordance), and PERMANOVA R^2.
iii. Design contrast across studies (comparative). Evaluate whether conclusions generalise across contrasting study designs by applying the same catalogue to a cross-sectional dataset (Project 1: anaerobic granular biofilm, DNA/cDNA strata) and a longitudinal dataset (Project 2: broiler chicken cecum).
iv. Stratum/time contrast within studies (comparative). Assess whether findings hold across biological strata in P1 (DNA vs cDNA) and across temporal structure in P2 (days 3–35), using the same filters and diagnostics.
v. Diagnostics and reporting (methodological). Provide standardised diagnostics per filter (Mantel, Procrustes, PERMANOVA R^2) together with a parameter

summary and software/version details sufficient for independent reproduction.

vi. Practice-oriented guidance (methodological). Synthesize evidence into data-driven guidance on which single filters most effectively compress features while preserving β-diversity geometry in cross-sectional and longitudinal settings.

1.7 Contributions and Significance

This thesis provides a systematic evaluation of feature filtering—a key yet under-reported step in 16S workflows. A standardized catalogue covering seven filter families (abundance, variability, distribution, transformation-specific, model-based, network-based, and relationship-based) is compiled and applied to two public datasets: a cross-sectional anaerobic granular biofilm study (Project 1; DNA and cDNA strata) and a longitudinal broiler cecum study (Project 2). The analysis quantifies how filter choices affect β-diversity structure, ordination geometry (via Mantel and symmetric Procrustes correlations), and the stability of distance-based inference (PERMANOVA R^2). Evidence is translated into clear, single-filter guidance for choosing thresholds that reduce noise while preserving geometry and reliable inference across both designs. Code, parameter definitions, and thresholds are documented to support transparent reuse [4,9,13].

Section 2 (Methodology) describes the two datasets, defines the seven filter families and their parameterization, and details the distance measures and ordination methods used for evaluation. All analyses were conducted in R with standard toolchains (e.g., phyloseq for data integration; vegan for distance computation, ordination, and PERMANOVA).

Section 3 (Results) reports filter effects on α- and β-diversity, PCoA geometry, and PERMANOVA R^2, using single-threshold settings per filter and, where applicable, limited sensitivity checks (e.g., selected alternative thresholds or distance metrics).

Section 4 (Discussion) interprets single-filter results for P1 and P2, indicating when DNA and cDNA strata warrant distinct thresholds and when time-aware criteria are appropriate in longitudinal settings, and it summarises methodological implications and limitations.

Section 5 (Conclusions) summarises the most stable strategies and provides practical recommendations for routine 16S preprocessing.

CHAPTER 2

METHODOLOGY

2.1 Data description

This study analysed two microbiome datasets derived from distinct environments and experimental designs. The first dataset (Project 1; P1) comprises microbial samples collected from granular biofilm reactors and represents a cross-sectional design. The second dataset (Project 2; P2) involves time-series faecal sampling of broiler chickens under different dietary treatments and represents a longitudinal design[14,[32]].

2.1.1 Dataset 1 – Granular Biofilm Community (Cross-sectional)

In P1, samples were collected from anaerobic granular biofilm reactors operated under controlled laboratory conditions. The reactors were designed to study microbial granulation in response to nutrient loading and hydraulic retention time. Samples included different biomass fractions (e.g., top and bottom granules, planktonic phase) with varying settling velocities and particle sizes[32].

The microbial composition was profiled using 16S rRNA gene sequencing (V4 region). Raw reads were processed into an amplicon sequence variant (ASV) table with taxonomic assignment against a reference database. Metadata includes reactor ID, biomass type, particle size, and settling velocity. In total, the working table contains 24 samples and 1,829 ASVs before any filtering. This data set provides a relatively stable endpoint community in which between-sample variation is driven mainly by operational and environmental differences[32].

2.1.2 Dataset 2 – Chicken Gut Microbiome (Temporal)

P2 investigated gut microbial dynamics of broiler chickens subjected to different dietary treatments. Faecal samples were collected from days 3–35, covering transitions between diet phases, with multiple diet groups and repeated time points per bird[14].

Profiling used 16S rRNA gene sequencing (V4 region). The present analysis uses a 97% OTU table (not ASVs) with taxonomic assignments. Metadata include treatment (diet) group, day of sampling, and bird ID, enabling analysis of temporal patterns and individual-level variation. In total, the working table contains 337 samples and 2,309 OTUs before any filtering. Because temporal autocorrelation can influence diversity patterns, this dataset permits evaluation of relationship-based filters that incorporate time as a covariate[14].

2.1.3 Rationale for Dataset Selection

By including both a cross-sectional dataset and a longitudinal dataset, this study evaluates whether pre-filtering strategies preserve diversity patterns consistently across designs. Sampling design shapes the structure of the feature table and the statistical relationships between samples, so contrasting contexts are informative. The differing metadata structures also permit evaluation of relationship-based filters (e.g., ANOVA-based criteria for cross-sectional contrasts; correlation-with-time criteria for longitudinal trajectories).

2.2 Laboratory data and primary data provision

The present analysis did not reprocess raw FASTQ files. All work started from the feature tables and matched metadata released by the original studies.

Project 1 (P1). Inputs: feature_w_tax.biom (feature-by-sample counts with

taxonomy), Paul_metadata2.xlsx (metadata), and tree.nwk (phylogeny). Upstream settings (instrument, trimming, chimera handling, classifier/training set) remain as reported in the source [7,10,32].

Project 2 (P2). Inputs: feature_w_tax.biom (feature-by-sample counts with taxonomy), meta_table.xlsx (metadata), and tree.nwk (phylogeny). Upstream settings remain as reported in the source[14].

Table-level curation applied in this study. Only taxonomy-based artifact removal was performed: features annotated as Mitochondria or Chloroplast and features unassigned at all taxonomic ranks were excluded, following community guidance[7].

2.3 Filtering Strategies Evaluated

This study tested different feature filtering strategies. They were grouped according to the types described in literature [13]. The goal of these strategies is to remove features that are very unlikely to give useful biological information. This can make the later analysis more stable and easier to understand. Both datasets used the same filtering categories, but the filters used in each dataset were not always exactly the same. This is because the datasets are different and also due to some computing limits.

2.3.1 Abundance-Based Filters

Abundance-based filters remove features that have very small counts or appear in only a few samples. These features are often just sequencing noise or not biologically useful [13]. For both Project 1 and Project 2, all filters in this group were applied without problem. These are:

total_count

Let $X \in \mathbb{R}^{n \times p}$ be a matrix of raw count data, where n is the number of samples and p is the number of features. For each sample $i \in 1, \ldots, n$, the total count (library size or sequencing depth) is calculated as:

$$s_i = \sum_{j=1}^{p} x_{ij},$$

where x_{ij} is the raw count of feature j in sample i.

A user-defined threshold τ is used to remove low-depth samples that may be dominated by noise or may not provide sufficient information for meaningful analysis. Samples are retained if their total counts exceed this threshold:

$$\text{keep}_i = \begin{cases} 1, & \text{if } s_i \geq \tau, \\ 0, & \text{otherwise.} \end{cases}$$

The filtering result is a binary vector indicating which samples are retained, and the total count s_i is recorded as a sample-level quality metric.

prevalence

Keep only features present in a minimum part of all samples.

Define an indicator matrix $Z \in \{0,1\}^{n \times p}$ such that:

$$z_{ij} = \begin{cases} 1, & \text{if } x_{ij} > 0, \\ 0, & \text{otherwise.} \end{cases}$$

For each feature $j \in 1, \ldots, p$, the prevalence is computed as the fraction of samples in which the feature is present:

$$s_j = \frac{1}{n}\sum_{i=1}^{n} z_{ij}$$

A user-defined threshold $\pi \in [0,1]$ is used to filter out rare features that occur in only a small proportion of samples:

$$\text{keep}_j = \begin{cases} 1, & \text{if } s_j \geq \pi, \\ 0, & \text{otherwise.} \end{cases}$$

mean_abundance

For feature j, the mean abundance is computed as:

$$\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$$

Given a threshold parameter $T = p_{\text{mean\_abundance\_min}}$, the filter selects features by:

$$\text{keep}_j = \begin{cases} \text{TRUE}, & \bar{x}_j \geq T \\ \text{FALSE}, & \text{otherwise} \end{cases}$$

mean_detection

For feature j, define the set of detected counts as:

$$D_j = \{x_{ij} \mid x_{ij} > 0,\ i = 1, \ldots, n\}$$

The mean detection value for feature j is computed as:

$$\bar{d}_j = \begin{cases} \frac{1}{|D_j|}\sum_{x \in D_j} x, & |D_j| > 0 \\ 0, & \text{if } |D_j| = 0 \end{cases}$$

Given a threshold parameter $T = p_{\text{mean detection min}}$, the filter selects features by:

$$\text{keep}_j = \begin{cases} \text{TRUE}, & \bar{d}_j \geq T \\ \text{FALSE}, & \text{otherwise} \end{cases}$$

low_count

For feature j, define a threshold $t = p_{\text{low\_count\_threshold}}$. The filter computes the fraction of samples where the count x_{ij} is at least t:

$$f_j = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\,(x_{ij} \geq t)$$

Given a minimum fraction parameter $f_{\min} = p_{\min\ \text{,fraction\_samples}}$, the filter retains features that are expressed above the threshold t in at least $f_{\min}$ proportion of samples:

$$\text{keep}_j = \begin{cases} \text{TRUE}, & f_j \geq f_{\min} \\ \text{FALSE}, & \text{otherwise} \end{cases}$$

min_count_in_ x samples

For each feature $j \in \{1, \ldots, m\}$, count the number of samples with counts greater than or equal to $p_{\text{min\_reads\_count}}$:

$$s_j = \sum_{i=1}^{n} \mathbf{1}\left(x_{ji} \geq p_{\text{min\_reads\_count}}\right)$$

The filter decision is based on:

$$\text{keep}_j = \begin{cases} \text{TRUE} & \text{if } s_j \geq p_{\text{min\_samples\_count}} \\ \text{FALSE} & \text{otherwise} \end{cases}$$

min_count_in_fraction

For feature j, define the fraction of samples with counts greater than or equal to a threshold $t = p_{\text{min\_count,threshold}}$ as:

$$f_j = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(x_{ij} \geq t)$$

Using a minimum fraction parameter $f_{\text{min}} = p_{\text{min\_fraction\_samples}}$, features are retained if:

$$\text{keep}_j = \begin{cases} \text{TRUE}, & f_j \geq f_{\text{min}} \\ \text{FALSE}, & \text{otherwise} \end{cases}$$

percentage_below_lod

For each feature $j = 1, \ldots, m$, calculate the proportion of samples with counts below the LOD threshold:

$$b_j = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{X_{ij} < p_{\text{lod\_threshold\_val}}\}$$

Determine which features to keep based on whether their proportion is less than or equal to the maximum allowed:

$$\text{keep}_j = \begin{cases} \text{TRUE} & \text{if } b_j \leq p_{\text{lod\_proportion\_max\_val}} \\ \text{FALSE} & \text{otherwise} \end{cases}$$

2.3.2 Variability-Based Filters

Variability-based filters keep features that change a lot between samples, because these are more likely to show differences between biological groups [5]. Both datasets used all filters in this group:

mean_variance

For each feature j, the mean-to-variance ratio is computed as:

$$\text{MVR}_j = \frac{\sigma_j^2}{\mu_j + \epsilon}$$

where μ_j is the sample mean, σ_j^2 is the sample variance, and ϵ is a small positive constant for numerical stability.

Features are retained if their mean-to-variance ratio exceeds a threshold τ:

$$\text{keep}_j = \begin{cases} 1, & \text{if MVR}_j \geq \tau \\ 0, & \text{otherwise} \end{cases}$$

This filter removes features with low variability relative to their mean abundance.

dispersion

For each feature j, the fold change between two experimental groups is computed as:

$$\text{FC}_j = \frac{\mu_j^{(A)}}{\mu_j^{(B)} + \epsilon}$$

where $\mu_j^{(A)}$ and $\mu_j^{(B)}$ are the mean counts in groups A and B respectively, and ϵ is a small constant to prevent division by zero.
Features are retained if their absolute $\log_2$ fold change exceeds a user-defined threshold τ:

$$\text{keep}_j = \begin{cases} 1, & \text{if } |\log_2(\text{FC}_j)| \geq \log_2(\tau) \\ 0, & \text{otherwise} \end{cases}$$

IQR (interquartile range)
For each feature j, the interquartile range is computed as:

$$\text{IQR}_j = Q_3^{(j)} - Q_1^{(j)}$$

where $Q_1^{(j)}$ and $Q_3^{(j)}$ are the first and third quartiles of the counts for feature j across all samples.
Features are retained if their IQR exceeds a user-defined threshold τ:

$$\text{keep}_j = \begin{cases} 1, & \text{if IQR}_j \geq \tau \\ 0, & \text{otherwise} \end{cases}$$

This filter identifies features with sufficient variability across samples using IQR as a robust measure of dispersion.
CV (coefficient of variation)
For each feature j, the coefficient of variation is computed as:

$$\text{CV}_j = \begin{cases} \dfrac{\sigma_j}{\mu_j} & \text{if } \mu_j > 0 \\ 0 & \text{if } \mu_j = 0 \end{cases}$$

where μ_j is the sample mean and σ_j is the sample standard deviation.
Features are retained if their CV exceeds a user-defined threshold τ:

$$\text{keep}_j = \begin{cases} 1, & \text{if CV}_j \geq \tau \\ 0, & \text{otherwise} \end{cases}$$

This filter identifies features with sufficient relative variability across samples using the coefficient of variation as a normalized measure of dispersion.
intra_group_variance
For each feature j, the mean intra-group variance is computed as:

$$\bar{\sigma_j^2} = \frac{1}{K}\sum_{k=1}^{K} \sigma_{j,k}^2$$

where $\sigma_{j,k}^2$ is the variance of feature j within group k, and K is the number of experimental groups.
Features are retained if their mean intra-group variance exceeds threshold τ:

$$\text{keep}_j = \begin{cases} 1, & \text{if } \bar{\sigma_j^2} \geq \tau \\ 0, & \text{otherwise} \end{cases}$$

combined_mean_variance:
use both abundance and variance rules together to choose features.
For each feature j, the mean abundance μ_j and sample variance σ_j^2 are computed as:

$$\mu_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}\,, \sigma_j^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_{ij} - \mu_j)^2$$

Features are retained only if they meet both user-defined thresholds:

$$\text{keep}_j = \begin{cases} 1, & \text{if } \mu_j \geq \tau_\mu \text{ and } \sigma_j^2 \geq \tau_{\sigma^2} \\ 0, & \text{otherwise} \end{cases}$$

variance

For each feature j, the variance is computed as:

$$\sigma_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2$$

where $\bar{x}_j$ is the mean of feature j.

Features are retained if their variance exceeds threshold T:

$$\text{keep}_j = \begin{cases} 1, & \text{if } \sigma_j^2 \geq T \\ 0, & \text{otherwise} \end{cases}$$

This fundamental variability filter removes features with minimal expression changes across samples, focusing analysis on features exhibiting biologically relevant variability.

2.3.3 Distribution-Based Filters

These filters look at the shape of feature value distribution [13]. Most of them were used successfully in both datasets.

entropy

For each feature j, the Shannon entropy is computed from normalized count probabilities:

$$H_j = -\sum_{i=1}^{n} p_i \log_2 p_i \text{ where } p_i = \frac{x_i}{\sum_{k=1}^{n} x_k} > 0$$

Features are retained if their entropy exceeds threshold τ:

$$\text{keep}_j = \begin{cases} 1, & \text{if } H_j \geq \tau \\ 0, & \text{otherwise} \end{cases}$$

zero_proportion

For each feature j, the proportion of zero counts across samples is computed as:

$$z_j = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(x_i = 0)$$

Features are retained if their zero proportion does not exceed threshold θ:

$$\text{keep}_j = \begin{cases} 1, & \text{if } z_j \leq \theta \\ 0, & \text{otherwise} \end{cases}$$

gini

For each feature j, the Gini coefficient is computed from sorted count values:

$$G_j = \frac{2\sum_{i=1}^{n} i x_i^{(j)} - (n+1)\sum_{i=1}^{n} x_i^{(j)}}{n\sum_{i=1}^{n} x_i^{(j)}}$$

where $x_i^{(j)}$ represents the sorted count values in ascending order.

Features are retained if their Gini coefficient does not exceed threshold τ:

$$\text{keep}_j = \begin{cases} 1, & \text{if } G_j \leq \tau \\ 0, & \text{otherwise} \end{cases}$$

bimodality

For each feature j, the bimodality score is computed as the proportion of zero counts:

$$\text{bimod}_j = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\,(x_{ij} = 0)$$

Features are retained if their proportion of zeros exceeds threshold τ:

$$\text{keep}_j = \begin{cases} 1, & \text{if bimod}_j \geq \tau \\ 0, & \text{otherwise} \end{cases}$$

This filter identifies features with potential bimodal distributions characterized by distinct "on" and "off" states across samples.

skewness

For each feature j, the skewness of its distribution is computed as:

$$\text{skewness}_j = \frac{E[(X_j - \mu_j)^3]}{\sigma_j^3}$$

where μ_j is the mean and σ_j is the standard deviation of feature j's counts.

Features are retained if their absolute skewness does not exceed threshold β:

$$\text{keep}_j = \begin{cases} 1, & \text{if } |\,\text{skewness}_j\,| \leq \beta \\ 0, & \text{otherwise} \end{cases}$$

combined_skewness_kurtosis

For each feature j, the combined statistic is computed from skewness and kurtosis:

$$v_j = |\,s_j\,| + k_j$$

where s_j is the skewness and k_j is the kurtosis of feature j's count distribution.

Features are retained if their combined statistic does not exceed threshold T:

$$\text{keep}_j = \begin{cases} 1, & \text{if } v_j \leq T \\ 0, & \text{otherwise} \end{cases}$$

zero_inflation

For each feature j, the zero inflation is computed as:

$$z_j = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(x_{ij} = 0)$$

Features are retained if their zero inflation does not exceed threshold $z_{\max}$:

$$\text{keep}_j = \begin{cases} 1, & \text{if } z_j \leq z_{\max} \\ 0, & \text{otherwise} \end{cases}$$

enhanced_low_count_percentage

This filter employs a dynamic threshold approach to identify features with meaningful expression patterns. The low count threshold is determined from the distribution of all positive counts:

$$\text{LOD}_{\text{dyn}} = \text{Quantile}(\text{AllPos}, p_{\text{low\_count\_threshold\_percentile}})$$

For each feature, two proportions are computed:

$$\text{PropBelowLOD}_j = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(x_{ij} < \text{LOD}_{\text{dyn}}), \text{PropPositive}_j = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(x_{ij} > 0)$$

Features are retained if they meet both criteria:

$$\text{keep}_j = (\text{PropPositive}_j \geq p_{\min\ ,\text{positive\_samples\_prop}}) \wedge (\text{PropBelowLOD}_j \leq 1 - p_{\min\ ,\text{positive\_samples\_prop}})$$

2.3.4 Relationship-Based Filters

These filters use the link between features and experimental variables [13].

In Project 1, only anova_p, fold_change, and group_fold_diff were used. The last two used a two-step grouping: first into DNA and cDNA, then into "bottom" and "top" groups.

In Project 2, the same four filters were used, plus time_correlation, autocorrelation_time, correlation_with_covariate, and low_correlation_with_covariate, which use time as one of the variables.

fold_change

For each feature j, the fold change between two experimental groups is computed as:

$$\text{FC}_j = \frac{\mu_j^{(A)}}{\mu_j^{(B)} + \epsilon}$$

where $\mu_j^{(A)}$ and $\mu_j^{(B)}$ are the mean counts in groups A and B respectively.

Features are retained if their absolute $\log_2$ fold change exceeds threshold τ:

$$\text{keep}_j = \begin{cases} 1, & \text{if } |\log_2(\text{FC}_j)| \geq \log_2(\tau) \\ 0, & \text{otherwise} \end{cases}$$

time_correlation

For each feature j, the Pearson correlation coefficient with a continuous time variable is computed:

$$r_j = \text{cor}(X_{:j}, \mathbf{t}) = \frac{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(t_i - \bar{t})}{\sqrt{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}\sqrt{\sum_{i=1}^{n}(t_i - \bar{t})^2}}$$

Features are retained if their absolute correlation with time exceeds threshold τ:

$$\text{keep}_j = \begin{cases} 1, & \text{if } |r_j| \geq \tau \\ 0, & \text{otherwise} \end{cases}$$

autocorrelation_time

For each feature j, the lag-ℓ autocorrelation is computed as:

$$r_j^{(t)} = \frac{\sum_{i=1}^{n-\ell}(x_i - \bar{x})(x_{i+\ell} - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \text{ for } n > \ell$$

Features are retained if their absolute autocorrelation exceeds threshold τ:

$$\text{keep}_j = \begin{cases} 1, & \text{if } |r_j^{(t)}| \geq \tau \\ 0, & \text{otherwise} \end{cases}$$

anova_p

For each feature j, a one-way ANOVA test is performed to assess differences in mean counts across predefined groups:

$$H_0: \mu_{1j} = \mu_{2j} = \cdots = \mu_{kj} \text{ vs } H_a: \text{at least one } \mu_{ij} \text{ differs}$$

Features are retained if their ANOVA p-value does not exceed threshold α:

$$\text{keep}_j = \begin{cases} 1, & \text{if } p_j \leq \alpha \\ 0, & \text{otherwise} \end{cases}$$

group_fold_diff

For each feature j, the fold difference between two experimental groups is computed as:

$$\text{fold\_diff}_j = \bar{x}_{j,1} - \bar{x}_{j,2}$$

where $\bar{x}_{j,1}$ and $\bar{x}_{j,2}$ are the mean counts in groups 1 and 2 respectively.

Features are retained if their absolute fold difference exceeds threshold T:

$$\text{keep}_j = \begin{cases} 1, & \text{if } |\text{ fold\_diff}_j | \geq T \\ 0, & \text{otherwise} \end{cases}$$

cohens_d_group

For each feature j, Cohen's d effect size is computed as:

$$d_j = \frac{| m_1 - m_2 |}{s_p}$$

where m_1 and m_2 are group means, and s_p is the pooled standard deviation:

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

Features are retained if their Cohen's d exceeds threshold T:

$$\text{keep}_j = \begin{cases} 1, & \text{if } d_j \geq T \\ 0, & \text{otherwise} \end{cases}$$

correlation_with_covariate

For each feature j, the Pearson correlation with a numeric covariate is computed as:

$$r_j = \frac{\sum_{i=1}^{n} (x_{ij} - \bar{x}_j)(c_i - \bar{c})}{\sqrt{\sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2} \sqrt{\sum_{i=1}^{n} (c_i - \bar{c})^2}}$$

Features are retained if their absolute correlation exceeds threshold t:

$$\text{keep}_j = \begin{cases} 1, & \text{if } |\, r_j \,| \geq t \\ 0, & \text{otherwise} \end{cases}$$

low_correlation_with_covariate

This filter removes features that show insufficient association with a specified numeric covariate. For each feature, the absolute correlation is computed:

$$\text{abs\_cor}_j = | \text{ cor}(\mathbf{x}_j, \text{covariate}) |$$

Features are retained if their absolute correlation meets the minimum threshold:

$$\text{keep}_j = (\text{abs\_cor}_j \geq p_{\text{min ,abs\_correlation}}) \vee (\text{abs\_cor}_j = \text{NA})$$

2.3.5 Transformation-Specific Filters

These filters work on data after variance-stabilising or other transformations [13].

mad_vst

Count data first undergo variance stabilizing transformation (VST) to address mean-dependent variance:

$$X^{\text{VST}} = \text{VST}(\text{round}(X^{\top}))^{\top}$$

For each feature j, the median absolute deviation is computed on VST-transformed data:

$$\text{MAD}_j = \text{median}(|\, x_{ij}^{\text{VST}} - \text{median}(x_j^{\text{VST}})\,|)$$

Features are retained if their MAD exceeds threshold τ:

$$\text{keep}_j = \begin{cases} 1, & \text{if MAD}_j \geq \tau \\ 0, & \text{otherwise} \end{cases}$$

mad_rlog

Raw counts are transformed using DESeq2's regularized logarithm transformation:

$$X^{\text{rlog}} = \text{rlog}(\text{round}(X^{\top}))^{\top}$$

For each feature j, the median absolute deviation is computed on rlog-transformed data:

$$\text{MAD}_j = \text{median}(|\, x_{ij}^{\text{rlog}} - \text{median}(x_j^{\text{rlog}})\,|)$$

Features are retained based on a MAD threshold τ:

$$\text{keep}_j = \begin{cases} 1, & \text{if MAD}_j \geq \tau \\ 0, & \text{otherwise} \end{cases}$$

dropout_after_vst

After applying variance stabilizing transformation (VST) to the count matrix, the dropout proportion for each feature is computed as:

$$d_i = \frac{1}{n}\sum_{j=1}^{n} \mathbb{I}(\tilde{x}_{ij} = 0)$$

where $\tilde{x}_{ij}$ represents the VST-transformed value.

Features are retained if their dropout proportion does not exceed threshold δ:

$$\text{keep}_i = \begin{cases} 1, & \text{if } d_i \leq \delta \\ 0, & \text{otherwise} \end{cases}$$

variance_ filter

Features are selected based on variance after applying variance stabilizing transformation (VST):

$$\mathbf{Y} = \text{VST}(\mathbf{X}^T), v_j = \text{Var}(\mathbf{Y}_{j,:})$$

The top $k = \lceil m \times p_{\text{top\_n\_percent, variance}} \rceil$ features with highest variances are retained:

$$\text{keep}_j = \begin{cases} 1, & \text{if feature } j \text{ in top } k \text{ by variance} \\ 0, & \text{otherwise} \end{cases}$$

extreme_values_robust_z_score

For VST-transformed data, robust Z-scores are computed for each feature:

$$Z_{ij} = \begin{cases} \dfrac{x_{ij} - \text{median}(\mathbf{x}_j)}{\text{MAD}(\mathbf{x}_j)}, & \text{if MAD}(\mathbf{x}_j) > \varepsilon \\ 0, & \text{otherwise} \end{cases}$$

Features are retained if their maximum absolute Z-score across samples does not exceed

threshold τ:

$$\text{keep}_j = \begin{cases} 1, & \text{if } \max_i \mid Z_{ij} \mid \leq \tau \\ 0, & \text{otherwise} \end{cases}$$

mad_filter

Features are selected based on median absolute deviation after variance stabilizing transformation:

$$\text{MAD}_j = \text{mad}(\mathbf{X}_{j,\cdot}^T)$$

The top $N_{\text{keep}} = \max(1, \lceil m \times p_{\text{top\_n\_percent\_mad}} \rceil)$ features by MAD are retained:

$$\text{keep}_j = \begin{cases} 1, & \text{if feature } j \text{ in top } N_{\text{keep}} \text{ by MAD} \\ 0, & \text{otherwise} \end{cases}$$

cv_filter

For VST-transformed data, the coefficient of variation is computed as:

$$\text{CV}_j = \frac{\sigma_j}{\mu_j + \epsilon}$$

Features are retained if their CV exceeds threshold $p_{\text{min\_cv\_val}}$:

$$\text{keep}_j = \begin{cases} 1, & \text{if CV}_j \geq p_{\text{min\_cv\_val}} \\ 0, & \text{otherwise} \end{cases}$$

iqr_filter

For VST-transformed data, the interquartile range is computed as:

$$\text{IQR}_j = Q_{3j} - Q_{1j}$$

The top $N_{\text{keep}} = \max(1, \lceil m \times p_{\text{top\_n\_percent\_IQR}} \rceil)$ features by IQR are retained:

$$\text{keep}_j = \begin{cases} 1, & \text{if feature } j \text{ in top } N_{\text{keep}} \text{ by IQR} \\ 0, & \text{otherwise} \end{cases}$$

2.3.6 Model-Based and Other Filters

These filters use statistical models to find important features [13]. In both projects, only the loq (limit of quantification) filter was used. Other filters here were not used because of convergence problems or heavy computation.

LOQ(Limit of Quantification)

For each feature j, the limit of quantification is computed as:

$$\text{LOQ}_j = \mu_j + q \cdot \sigma_j$$

where μ_j is the mean abundance, σ_j is the standard deviation, and q is a quantile multiplier parameter controlling confidence level.

Features are retained based on their LOQ values:

$$\text{keep}_j = \begin{cases} 1, & \text{if LOQ}_j \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

2.3.7 Network-Based Filters

This type keeps features that have enough links to other taxa in a co-occurrence network [13]. The network_connectivity filter worked in both datasets and kept features with enough network connections.

network_connectivity

This filter evaluates features based on their connectivity within a weighted co-expression network. After applying variance stabilizing transformation, an adjacency matrix is constructed using soft-thresholded correlation coefficients:

$$A_{ij} = | \operatorname{cor}(\mathbf{D}_{.i}, \mathbf{D}_{.j}) |^{\beta}$$

The network degree for each feature is computed as:

$$\text{degree}_j = \sum_{i=1}^{m'} B_{ij}$$

where B_{ij} is the binary adjacency matrix thresholded at $p_{\text{network\_corr\_threshold}}$.
Features are retained if their network degree meets the minimum threshold:

$$\text{keep}_j = \begin{cases} 1, & \text{if degree}_j \geq p_{\text{min\_network\_degree}} \\ 0, & \text{otherwise} \end{cases}$$

This approach identifies features with strong co-expression relationships, reflecting their potential involvement in biological modules or pathways, while filtering out isolated features that may represent noise.

2.4 Diversity and Ordination Metrics

In this study, statistical methods were used to check how different pre-filtering strategies change microbiome diversity and community structure. The evaluation included diversity measures and the extent to which community patterns were preserved after filtering.

Alpha and beta diversity

Alpha diversity measures richness and evenness within a microbial community in a single sample. The Shannon diversity index was used because it looks at both richness and evenness. The number of observed features (ASVs/OTUs) was also used, and it shows richness only. These measures are important because they can show if the community is balanced or if a few taxa dominate, which can be linked to changes in how the community is formed [21].

Beta diversity measures differences in microbial community between samples. Distance measures such as Bray–Curtis, Jaccard, unweighted UniFrac and weighted UniFrac were calculated. These measures show both taxonomic and phylogenetic differences. They help to see how communities change under different environmental or experimental conditions [33].

PERMANOVA (adonis2)

PERMANOVA (Permutational Multivariate Analysis of Variance) was used to test if beta diversity is different between defined groups. This test is non-parametric, uses permutations of the distance matrix, and checks if the group centroids in multivariate space are different [29]. In this study, the adonis2 function from the R package vegan was used, with 999 permutations. PERMANOVA is common in microbial ecology to see if environmental or experimental factors explain a significant part of the variation in community composition.

Mantel Test

The Mantel test was used to check the correlation between two distance matrices: one from the unfiltered data and one from each filtered dataset. A high Mantel correlation means the filtering kept the original community dissimilarity pattern [36]. The significance of this correlation was tested with permutations. This method is often used in ecology to check if spatial or environmental patterns in the data stay consistent.

Procrustes Test

Procrustes analysis was used to see how similar the ordinations are between unfiltered and

filtered data. The method optimally aligns the two ordinations and reports a correlation value that shows the similarity in community structure [33]. A permutation test was then done to check if the correlation is higher than by chance. This method is useful with the Mantel test because it focuses on the visual and geometric match between community structures.

PERMANOVA R^2

PERMANOVA R^2 is the proportion (percentage) of variability in the distance matrix that is explained between groups defined by the study design. For each design factor, R^2 and its permutation p-value were computed before and after filtering. If the post-filter R^2 remains statistically significant ($p < 0.05$) and close to the baseline value, this indicates that the between-group β-diversity structure is preserved despite the removal of some features [29]. Conversely, large shifts in R^2 or loss of significance may imply distortion or attenuation of group separation introduced by filtering.

Summary

Using alpha and beta diversity measures together with PERMANOVA, Mantel, Procrustes, and PERMANOVA R^2, this study made a full evaluation of how filtering strategies change microbiome data. These methods are common in microbial ecology and work well to find both the size and type of changes in community structure after preprocessing.

2.5 Data Preprocessing and Normalization

No raw FASTQ reprocessing was performed; downstream analyses used the public feature tables and matched metadata. In the source workflows, raw 16S reads underwent quality filtering (e.g., q2-quality-filter in QIIME 2 [7]), trimming/paired-end merging, chimera removal (uchime in VSEARCH [28]), and either denoising to infer ASVs (e.g., DADA2 [10]) or 97% clustering to generate OTUs, as reported. Taxonomy was assigned with a naive Bayes classifier trained on SILVA 138 for the V4 region [25]. Where applied, tables were rarefied to a common depth and/or converted to relative abundance. Upstream parameters differed between datasets per the original studies.

2.6 Software and Computational Environment

All statistical analyses in this study were done in R (version 4.5.0) [26].

Alpha diversity metrics such as the Shannon index, and beta diversity metrics such as Bray–Curtis dissimilarity, were calculated with the phyloseq package (version 1.52.0)[18]. Ordination methods like Principal Coordinates Analysis (PCoA) were used to show the differences in microbial community composition. PERMANOVA, Mantel, and Procrustes tests were done with the vegan package (version 2.7-1) [23].

All analysis were run on a Windows 11 workstation with 32 GB RAM and an Intel Core i7-12700 processor. Figures were **produced** with ggplot2 package (version 3.5.2) [35]. All R scripts were saved in a version-controlled repository so the analysis can be repeated.

CHAPTER 3

RESULTS

3.1 Dataset overview

This chapter assesses how single-filter strategies influence community diversity and ordination geometry across two 16S rRNA datasets with contrasting designs. A harmonised preprocessing workflow (Section 2;[7,18,23]) yields like-for-like unfiltered baselines used throughout. Full dataset summaries are provided in Appendix A (Table A.1), which lists sample and feature counts, design information, and key contextual notes from the source studies.

Project 1 — Granular biofilm community (cross-sectional)

Design and baseline. Project 1 comprises 24 samples from laboratory UASB reactors contrasting biomass fractions (floating vs settled) and nucleic-acid type (DNA vs cDNA). The retained unfiltered baseline is 24 × 1,829 (samples × features). Covariates distinguish nucleic-acid type and biomass position. Bray–Curtis dissimilarities and PCoA provide the primary β-diversity frame for figures, with UniFrac used as a robustness check consistent with the source practice [32].

Design contrast. Relative to Project 2, this cross-sectional snapshot tests whether filters preserve between-group geometry at a single time point.

Project 2 — Broiler chicken caecal microbiome (longitudinal)

Design and baseline. Project 2 is a day-to-day longitudinal series spanning day 3–35 with ≈12 birds per day (original 379 samples × 18,588 OTUs at 97% clustering). Following the harmonised preprocessing (Section 2), the retained unfiltered baseline is 337 samples × 2,309 OTUs, used for all downstream filter evaluations. Available covariates include Day (time), diet/phase (Starter → Grower → Finisher), body weight, feed intake, and FCR. Analyses are conducted at the OTU level using Bray–Curtis as the primary distance and PCoA as the default ordination; for displays and tests requiring discrete groups, Diet is used as a factor in the order *Starter, Grower, Finisher*. For relationship-based filters, Day is treated as a numeric covariate[14].

Published characteristics. The source study reported a rapid rise in α-diversity up to approximately day 12, stabilisation after ~day 20, temporal shifts aligned with diet-phase changes, and time/week explaining ~16–17% of Bray–Curtis variance with smaller R^2 for Unweighted UniFrac; these anchors motivate treating time and diet phase as primary structuring variables in subsequent α-/β-diversity evaluations [18错误!未找到引用源。].

Design contrast. In contrast to Project 1, the temporal structure here supports direct assessment of whether filtering preserves time-ordered gradients and diet-phase separation in both α- and β-diversity.

Cross-project note. Unless stated otherwise, downstream sections reference the unfiltered baselines P1: 24 × 1,829 and P2: 337 × 2,309. Bray–Curtis + PCoA are used throughout for comparability; UniFrac results are referenced as robustness checks aligned with prior practice.

3.2 Filtering outcomes

The two projects start from markedly different data scales—P1: 24 × 1,829 and P2: 337 × 2,309—but both were subjected to the same filter catalogue to quantify feature-space compression prior to ecological evaluation (§3.3–§3.4). Family-level summaries are provided below; per-filter counts and stability statistics (Mantel, Procrustes, PERMANOVA R^2, ΔR^2) are reported in Appendix A, Table A.2 (P1/P2).

Project 1 — Granular biofilm community

Across the compact cross-sectional matrix, abundance/sparsity-oriented filters (e.g., *prevalence*, *low_count*, *percentage_below_lod*) delivered the largest reductions, often removing the bulk of ultra-rare or sub-LOD features while keeping a workable table (see Table A.2–P1). Variability and distribution families showed moderate reductions on average, with *IQR* a representative robust screen. Transformation-specific rules spanned a wide range—from very strict (e.g., *mad_rlog*) to minimal impact (e.g., utilities such as *mean_detection*). Relationship-based procedures that rely on explicit contrasts (e.g., *anova_p*; within-stratum *group_fold_diff*) are summarised separately because their denominators differ from whole-table filters; see Table A.2–P1.

Project 2 — Broiler caecal microbiome

In this larger longitudinal dataset, feature-space reductions ranged from negligible to very strong, with abundance-based (low-count/sparsity) rules driving the largest compression (e.g., percentage_below_lod retained ~13% of features), while several mean/dispersion utilities had little or no effect; variability and distribution families were intermediate on average, though shape-extreme rules (e.g., skewness, gini) could be highly stringent. Transformation-specific filters spanned a wide impact range depending on the chosen transform (from aggressive pruning such as mad_vst to modest changes), and relationship-based filters in P2 were evaluated against time (Day), retaining features aligned with longitudinal signal (e.g., anova_p, time_correlation). By contrast, pairwise group tools (e.g., fold_change, group_fold_diff, cohens_d_group) were not applicable because Diet has three levels (Starter, Grower, Finisher) and no single two-group contrast captures the full design. Full per-filter retention counts and stability statistics (Mantel, Procrustes, PERMANOVA R^2, ΔR^2) are reported in Appendix A, Table A.2–P2.

Across both datasets, abundance-based (sparsity) filters produce the largest feature-space compression, whereas most mean/dispersion utilities have minimal impact. Sections 3.3–3.4 assess whether these reductions preserve or distort α- and β-diversity using Mantel, Procrustes, and PERMANOVA relative to the unfiltered baselines.

3.3 Effects of filtering on α-diversity

Project 1 — Granular biofilm community

Metrics and contrasts

Within-sample diversity was assessed using Shannon entropy (evenness-weighted) and Chao1 richness (estimated ASV richness) on the rarefied table. Groups comprise four biomass fractions by nucleic-acid type (Floating–cDNA, Settled–DNA, Settled–cDNA, Floating–DNA). Pairwise differences above the boxplots are from one-way ANOVA with post-hoc tests; asterisks indicate significance (* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$).

Unfiltered baseline

In the unfiltered data (Fig. 3.3A–P1, "Original"), Shannon shows a clear separation of groups: Floating–DNA has the highest diversity, Settled–cDNA the lowest, with Floating–cDNA and Settled–DNA in between. Multiple pairwise contrasts are significant, consistent with a marked evenness difference between floating vs settled fractions. Chao1 differences are modest: Settled–DNA tends to be the lowest, while Floating–DNA and Settled–cDNA are higher and broadly overlapping; only a subset of pairwise contrasts reaches significance.

Filters that preserve the baseline pattern

These filters remove low-signal (rare or sub-LOD) features yet keep the between-group ordering seen at baseline.

• prevalence (retained 509/1,829 ASVs, −72.2%; Fig. 3.3–Prevalence): Shannon ordering and significance pattern are maintained. Chao1 medians shift slightly but group ranking is unchanged; spreads are comparable to the baseline.

• low_count (364/1,829, −80.1%; Fig. 3.3–Low-count): Again preserves the Shannon separation (floating > settled; Settled–cDNA lowest). Chao1 shows small downward shifts

with overlapping boxes; significance remains similar to the baseline.
• percentage_below_lod (253/1,829, −86.2%; Fig. 3.3–Percentage_below_lod): Despite the strongest compression of the feature space, Shannon separation is retained and several pairwise contrasts remain significant. Chao1 medians are modestly reduced; variability increases slightly in some groups but not enough to alter the ordering.
• IQR (333/1,829, −81.8%; Fig. 3.3–IQR): Preserves the baseline ranking for both metrics. Shannon differences remain clear; Chao1 differences remain modest.

Borderline adjustments

These filters nudge the α-diversity distributions but do not overturn the qualitative conclusions.
• entropy (618/1,829, −66.2%): Shannon ordering is conserved with small shifts. For Chao1, group medians are very similar to the baseline—if anything, slightly higher; the interquartile ranges are not consistently narrower and are a bit wider for some groups. Overall pattern is intact.
• mad_vst (183/1,829, −90.0%): Both Shannon and Chao1 distributions contract. Group ranking is still recognisable, but effect sizes are altered; treated here as a boundary case rather than a preferred option for preserving α-diversity scale.

Filters that distort α-diversity distributions

These either collapse the scale of the metrics or re-order groups in a way that departs from the baseline.
• gini (184/1,829, −89.9%): Richness and Shannon drop across all groups, the scale compresses, and most pairwise significance disappears. The ordering becomes unstable—indicative of over-filtering toward highly uneven features.
• mad_rlog (48/1,829, −97.4%): Richness contracts sharply and Shannon shifts downward with partial re-ordering. Although some pairwise contrasts remain significant, the overall distribution is markedly altered relative to the baseline; classified here as distorting.

Interpretation

Filters that target sparsity without aggressively reshaping the abundance structure—prevalence, low_count, percentage_below_lod, IQR—reduce the feature space by ~70–86% while leaving the baseline α-diversity narrative intact: Floating–DNA is most even, Settled–cDNA least; richness differences are present but modest. In contrast, gini and mad_rlog materially change the α-diversity scale and are not suitable when preservation of within-sample diversity is a priority.

Fig. 3.3A–P1 α-diversity under representative filters (Project 1)

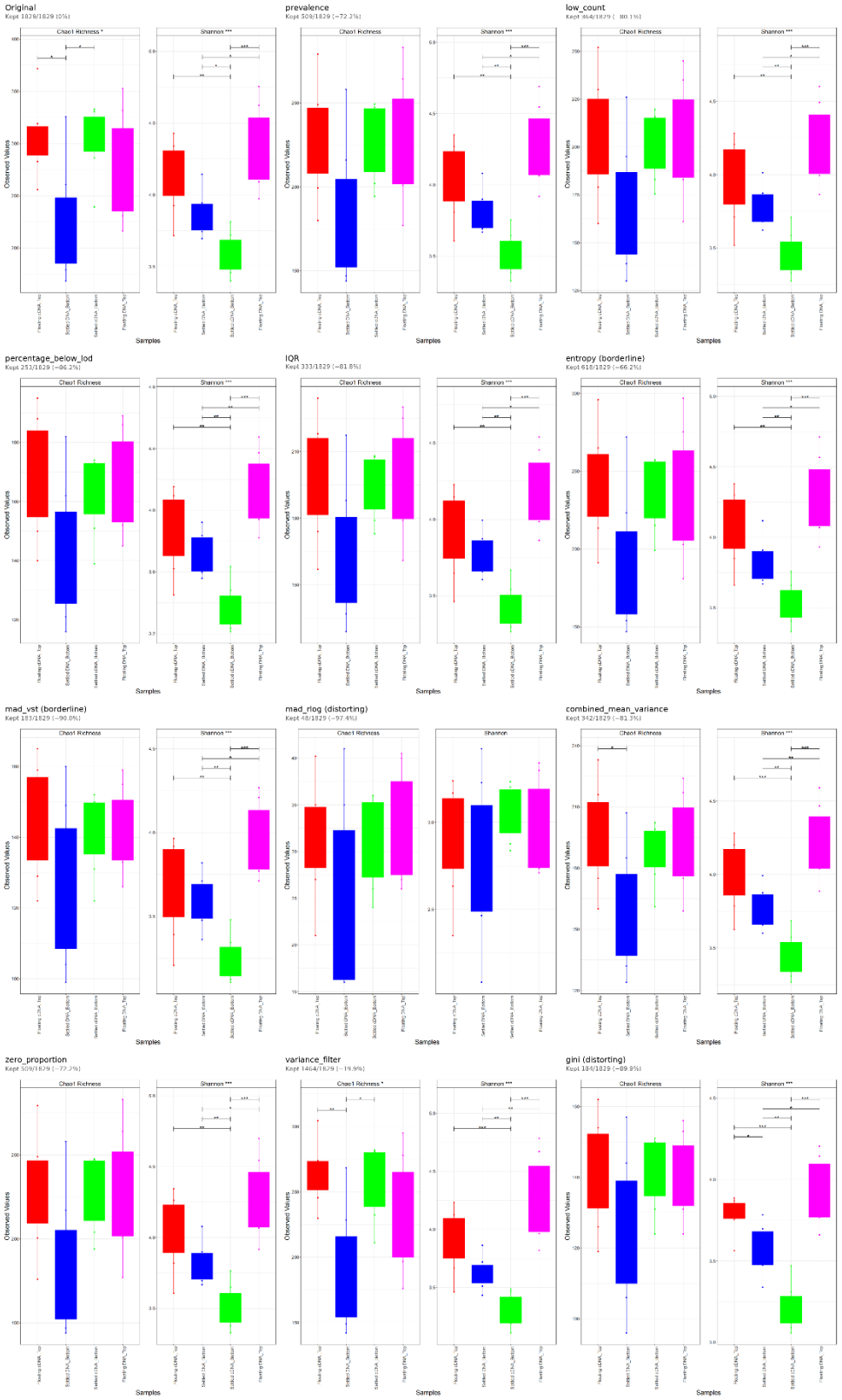

Fig. 3.3A–P1. α-diversity under representative filters (Project 1).
Each panel shows Chao1 (left) and Shannon (right) with one-way ANOVA post-hoc significance bars (p<0.05, ** p<0.01, *** p<0.001). Panel titles indicate the applied filter; retained features are reported below the title. Exact Mantel/Procrustes/PERMANOVA R^2 values are in Appendix A, Table A.2–P1.

Project 2 — Broiler caecal microbiome

Metrics and contrasts

Within-sample diversity was assessed using Shannon entropy and Chao1/Observed richness at the OTU level. Boxplots are grouped by diet phase (Starter, Grower, Finisher) with one-way ANOVA post-hoc significance bars (* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$).

Unfiltered baseline

In the unfiltered data (Fig. 3.3B–P2), both Shannon and richness rise Starter < Grower < Finisher, with most pairwise contrasts significant. This matches the expected early increase (≈ day 3–12) followed by higher levels thereafter.

Filters that retain the baseline pattern.

Each retained the baseline ordering while removing rare/sub-LOD features.

• **percentage_below_lod** — retains 302/2,309 (−86.9%). The **Starter < Grower < Finisher** ordering is unchanged for both metrics; spreads are close to baseline. Chao1 shows slightly slimmer boxes for Grower and Finisher, but inter-quartile overlap remains limited.

• **min_count_in_fraction** — 365/2,309 (−84.2%). Phase separation is visually strong: Chao1 medians are well spaced with only light whisker overlap; Shannon boxes shift upward monotonically with clear gaps between adjacent medians.

• **combined_mean_variance** — 483/2,309 (−79.1%). Both metrics retain the baseline ranking with **minimal scale change**; Grower and Finisher boxes remain distinct, and Starter sits clearly lower. Point clouds show similar dispersion to the baseline.

• **cv_filter** — 451/2,309 (−80.5%). Ordering is preserved; boxes are **slightly contracted** relative to baseline, especially for Shannon, but adjacent phases remain separable by median and central IQR.

• **IQR** — 705/2,309 (−69.5%). The phase gradient is again maintained. Chao1 shows a small tightening of Grower's IQR; Shannon medians rise stepwise with limited IQR overlap, matching the expected pattern.

Borderline adjustments

These nudge the α-diversity distributions but do not overturn the qualitative conclusions.

• **low_count** — 475/2,309 (−79.4%).Chao1 and Shannon both retain the Starter < Grower < Finisher ordering with visibly good separation of medians; the Starter box remains broad for Shannon, and adjacent phases show moderate IQR overlap. Overall, α-patterns are close to baseline, aligning with a borderline classification once β-geometry is considered.

• **zero_inflation** — 1,005/2,309 (−56.5%). Phase medians are well spaced for both metrics; boxes are slightly wider than baseline for Grower/Finisher, but **ordering and central tendency are preserved**. Visual separation remains, consistent with a borderline rather than distorting effect.

Filters that distort α-diversity distributions

These either collapse the scale or change relative group separation.

• **skewness** — 22/2,309 (−99.0%). Both panels show **extreme compression**: Grower and Finisher boxes are **very thin and nearly coincident** for Shannon; Chao1 values cluster in a narrow band with minimal spread, indicating collapse of α-scale.

• **gini** — 168/2,309 (−92.7%). Distributions are **depressed and partly convergent**: Chao1 medians drop far below the baseline range and Shannon boxes are compact with **reduced vertical separation** between phases, consistent with a distorting effect.

• **bimodality** — 2,093/2,309 (−9.4%). **Chao1 boxes for Starter, Grower, and Finisher are largely overlapping** with similar medians and whiskers; Shannon shows stepwise medians but broad IQRs that **substantially overlap**. The panels illustrate **loss of phase separation** despite small feature removal.

• **extreme_values_robust_z_score** — 2,129/2,309 (−7.8%). Medians increase

monotonically, yet **IQRs tighten** and adjacent phases **overlap in spread** for both metrics; separation is **visually modest** rather than clear, matching the weak stability statistics.

Interpretation
Filters that reduce sparsity without re-weighting abundances—percentage_below_lod, min_count_in_fraction, combined_mean_variance, cv_filter, IQR—remove ~69–87% of features while preserving the Starter < Grower < Finisher α-diversity gradient and usable dynamic range. low_count and zero_inflation largely preserve α-patterns but warrant caution given β-geometry (see §§3.4–3.5). skewness, gini, bimodality, and extreme_values_robust_z_score compress α-scale and erode phase separation, and are therefore unsuitable when within-sample diversity preservation is required.

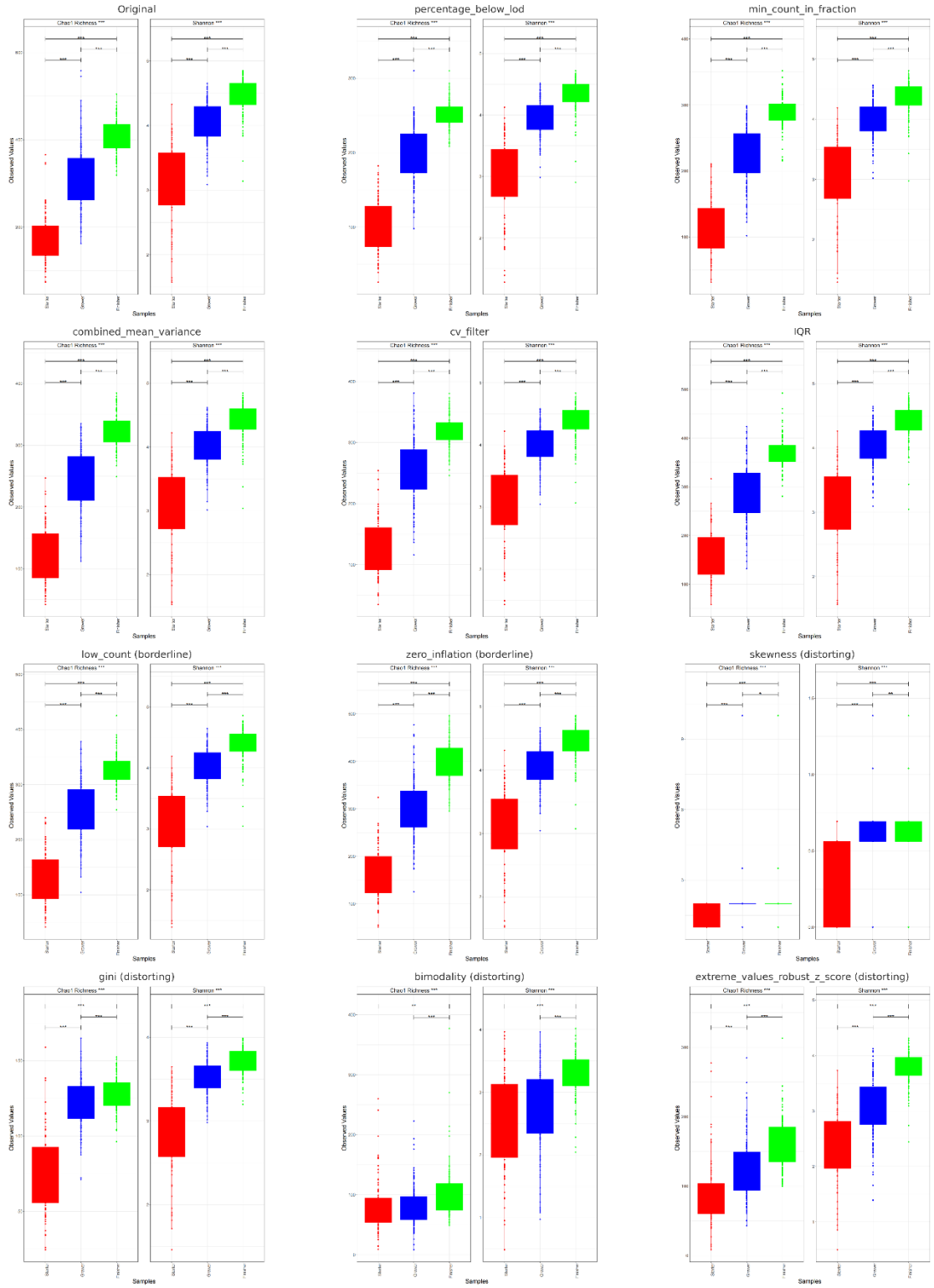

Fig. 3.3B–P2. α-diversity under representative filters (Project 2).

Boxplots of Chao1 richness and Shannon entropy (OTU level, rarefied) by diet phase (Starter, Grower, Finisher). Exact Mantel, Procrustes, and PERMANOVA R^2 values are reported in Appendix A, Table A.2–P2.

Figure note: Asterisks denote significance for one-way ANOVA post-hoc tests: * = $p < 0.05$; ** = $p < 0.01$; *** = $p < 0.001$. Y-axes are harmonised within metric across panels; diet order and colours are fixed as Starter (red) → Grower (blue) → Finisher (green).

3.4 Effects of filtering on β-diversity

Between-sample structure was evaluated with Bray–Curtis distances and PCoA. Preservation relative to the unfiltered baseline is quantified by: (i) Mantel correlation (999 permutations) between distance matrices, (ii) symmetric Procrustes correlation between ordinations, and (iii) PERMANOVA R^2 (variance explained by group). We describe visible changes in PCoA geometry (centroid shifts, rotations, dispersion) alongside ΔR^2. Filters are classified as geometry-preserving, borderline, or distorting based on these metrics and on visual agreement with each baseline ordination.

Project 1 — Granular biofilm community

Baseline ordination

The unfiltered PCoA resolves four compact, well-separated clusters corresponding to the experimental fractions (Floating–cDNA, Settled–DNA, Settled–cDNA, Floating–DNA). Group centroids are widely spaced on both axes and ellipses are tight; this configuration is used as the geometric reference.

Filters that preserve geometry

Filters that primarily remove low-information or zero-inflated features leave the baseline layout essentially unchanged—panels are visually indistinguishable from the baseline and Mantel/Procrustes are ≈1, with ΔR^2 small and often positive. Representative examples include:

- **Prevalence** and **Percentage_below_lod**: four clusters remain in the same relative positions with near-identical shapes; ΔR^2 increases slightly, consistent with a modest sharpening of contrasts.
- **IQR** and **Variance**: minor thinning of point clouds but stable centroids and orientation; R^2 stays at baseline levels.

Borderline changes

A second group shows small, visible deviations while broadly maintaining the four-cluster arrangement:

- **Entropy**: mild reshaping and slight centroid drift (a gentle shear/stretch) without re-ordering of groups; Procrustes < 1 but high, and ΔR^2 near zero.
- Similar, subtle effects are seen with other light variability-based filters (e.g., variance-type selections).

Distorting filters

Filters that re-weight distributional shape or over-emphasize extremes consistently reconfigure the ordination:

- **Gini**: clusters are displaced and rescaled, indicating over-weighting of highly uneven features.
- **Bimodality**: axes bend and the relative placement of groups changes despite limited feature removal.
- **mad_rlog**: the most aggressive change in geometry; centroids move and ellipses inflate, yielding a layout that no longer matches the baseline.

Interpretation

Geometry-preserving compression is achieved by sparsity-oriented filters (prevalence, Percentage_below_lod, IQR, variance-type). In contrast, shape-reweighting filters (gini, bimodality) and transform-specific **mad_rlog** alter the Bray–Curtis geometry and should be avoided when preserving the baseline ordination is the priority.

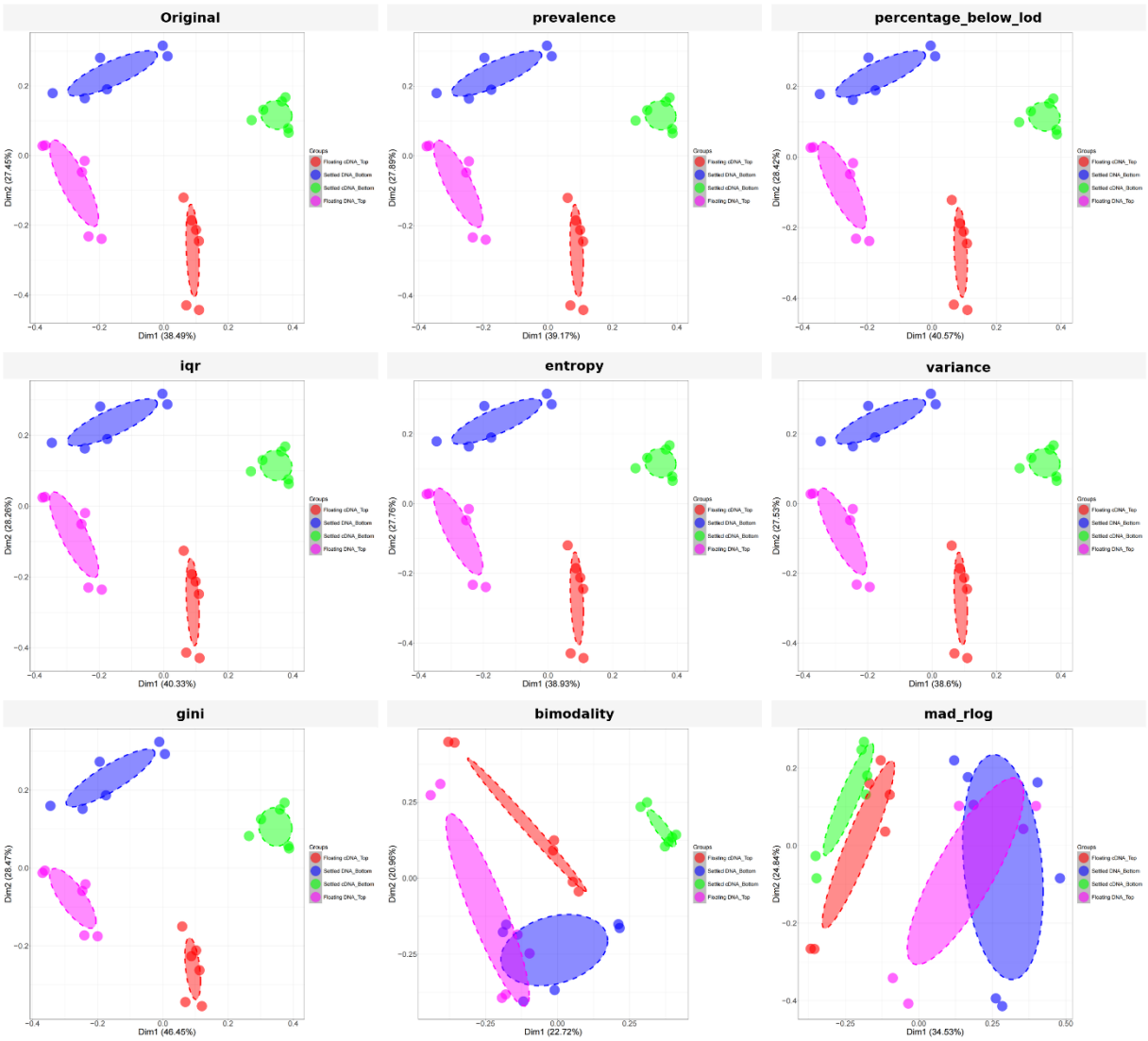

Fig. 3.4A–P1. Bray–Curtis PCoA under representative filters (Project 1).

Project 2 (Broiler caecal microbiome)

Baseline

The unfiltered panel shows a clear Starter → Grower → Finisher trajectory: Starter (red) occupies the left–lower region, Grower (blue) spans the centre, and Finisher (green) sits to the right–upper. Dashed ovals mark stable local neighbourhoods used for visual anchoring.

Filters that preserve geometry

These panels keep the same global trajectory and local neighbourhoods; point clouds tighten slightly but relative positions of phases are unchanged.

Percentage_below_lod: the arc is intact; Grower tightens a little around the central band, and Finisher remains upper-right.

Min count in fraction: phase separation is even crisper; the red–blue transition is clean while Finisher stays well apart.

Combined mean–variance: visually almost indistinguishable from the baseline; only subtle contraction of spread.

IQR and CV filter: mild shrinkage of the three clouds without any re-ordering; the dashed ovals still sit in the same relative locations.

Time correlation and ANOVA p: geometry matches the baseline arc, but phase centroids are sharper and between-phase gaps are a bit larger.

Borderline changes

The ordering is kept, but local neighbourhoods shift enough to be noticeable.

Low count: the Starter–Grower transition bends slightly; the dashed ovals indicate of the Grower core compared with the baseline.

Zero inflation: the three phases remain in order, though centroids move modestly and spreads widen for Grower/Finisher.

Distorting filters

These re-weight the geometry and no longer reflect the baseline Bray–Curtis space.

Skewness: point clouds collapse into a narrow band and the time arc is largely lost; Starter, Grower, and Finisher intermix around the centre.

Gini: the space is warped toward dominant taxa; the mid-phase (blue) shifts and the red–green progression bends away from the baseline path.

Interpretation

For longitudinal PCoA where preserving the Starter→Grower→Finisher gradient is essential, prefer percentage_below_lod, min count in fraction, combined mean–variance, IQR, and CV filter. If a touch more visual separation is desired without changing geometry, time correlation or ANOVA p are acceptable (report Mantel/Procrustes/ΔR^2). Low count and zero inflation are borderline but usable with a note. Avoid skewness and gini when fidelity to the baseline Bray–Curtis geometry is required.

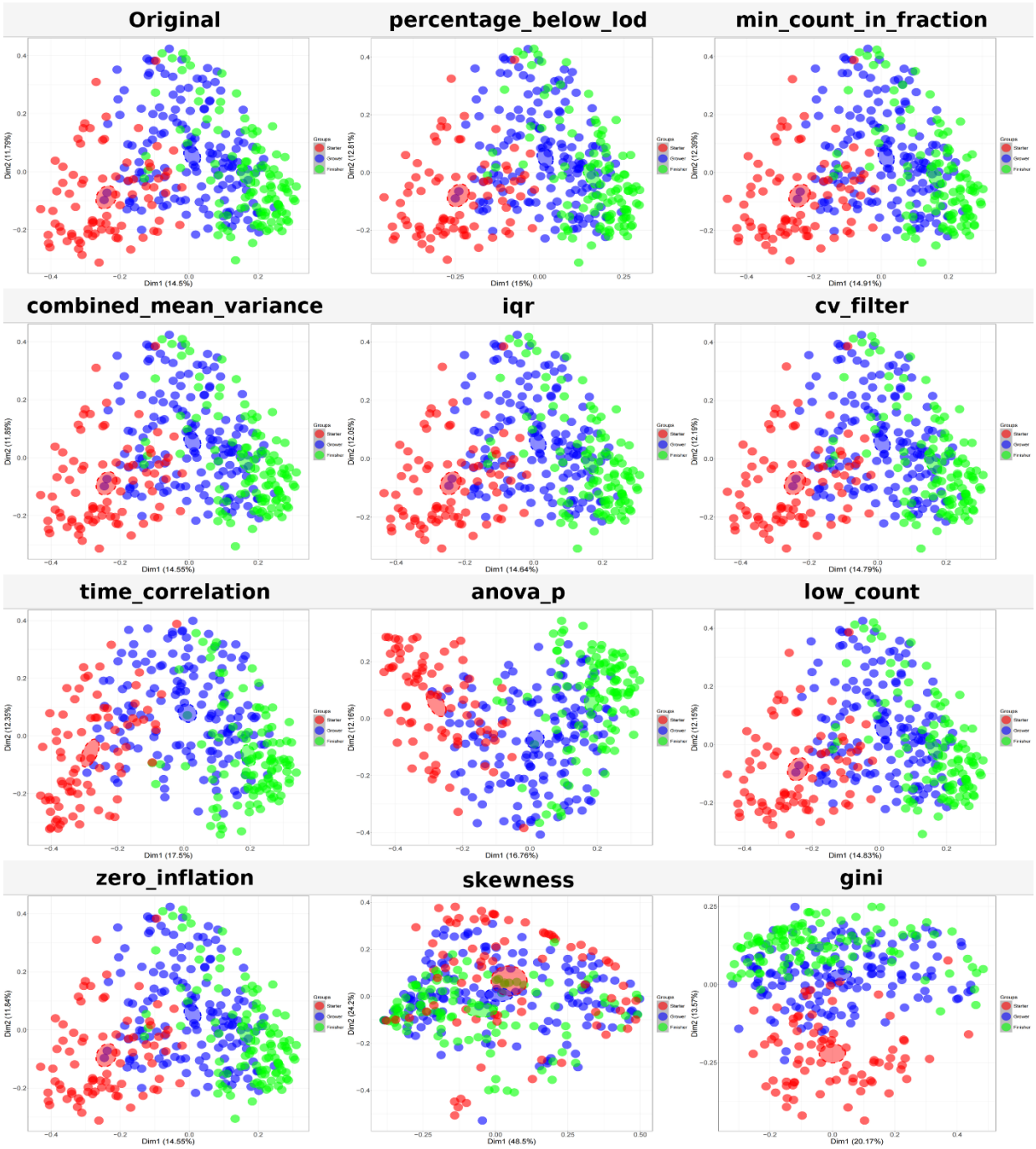

Fig. 3.4B–P2. Bray–Curtis PCoA under representative filters (Project 2).

3.5 Statistical test results

This section synthesises Mantel correlations (distance-matrix concordance), symmetric Procrustes correlations (ordination geometry), and PERMANOVA R^2 (variance explained by group) relative to each project's unfiltered Bray–Curtis baseline (999 permutations throughout). Full statistics for all filters are reported in Appendix A, Table A.2 (P1/P2), with visual ordinations in Figures 3.3–3.4.

3.5.1 Mantel correlation and Procrustes agreement

Project 1 — Granular biofilm (cross-sectional)

Distance preservation. Distance matrices were exceptionally stable for most single-filter selections: abundance/variability/sparsity-oriented methods yielded Mantel $r \geqslant 0.999$. The largest departures occurred with gini, bimodality, and mad_rlog.

Geometry. Ordination geometry was more sensitive than distances. Prevalence, low_count, percentage_below_lod, IQR, combined_mean_variance and the sparsity filters had Procrustes ≈ 0.999–1.000 (layouts indistinguishable from baseline). Moderate reshaping appeared for entropy and variance_filter, and was lower for dispersion. Clear re-projections were seen for bimodality, gini, extreme_values_robust_z_score, cv_filter, and mad_rlog.
Relationship-based (within-stratum) selectors.
group_fold_diff preserved geometry in cDNA (Procrustes ≈ 0.997; Mantel ≈ 0.99998) and retained the two-cluster layout in DNA albeit with rotation/elongation (≈ 0.68; Mantel ≈ 0.99997). fold_change showed evident re-weighting: cDNA ≈ 0.94 (Mantel ≈ 0.99); DNA ≈ 0.72 (Mantel ≈ 0.95).

Project 2 — Broiler caecal (longitudinal)

Geometry preserved (near-isometric). Filters that trim sparsity/low variance—percentage_below_lod, min_count_in_fraction, mean_abundance, prevalence, IQR, combined_mean_variance, cv_filter—reproduce the baseline Starter→Grower→Finisher trajectory with only subtle tightening of point clouds (see Table A.2–P2).
Borderline changes. low_count and zero_inflation maintain the time arc but shift local neighbourhoods enough to be visible (Table A.2–P2).
Geometry distorted. Shape-reweighting/extreme-value/network filters—skewness, gini, bimodality, zero_proportion, extreme_values_robust_z_score, autocorrelation_time, network_connectivity—warp the ordination (centroid drift/axis bending/dispersion inflation; Table A.2–P2).
Interpretation rule (both projects). For cross-sectional P1, Procrustes near 1.0 indicates faithful cluster geometry; for longitudinal P2, Procrustes $\geqslant 0.95$ reliably signals preservation of the Starter→Grower→Finisher trajectory. Values below these levels—especially $\leqslant 0.90$—denote genuine shape change rather than mere rotation or translation.

3.5.2 PERMANOVA R²

Project 1 — Granular biofilm (cross-sectional)

Baseline. PERMANOVA $R^2 = 0.624 (p = 0.001)$.
Small, beneficial gains with preserved geometry. Filters like *percentage_below_lod*, *low_count*, *prevalence*, and *IQR* yield slight positive ΔR^2consistent with noise reduction, while maintaining the baseline ordination.
Inflation with distortion. *gini* and *combined_skewness_kurtosis* inflate R^2yet warp the geometry, indicating selective up-weighting of high-contrast subsets rather than strengthening the original pattern.
Loss of fit with distortion. *bimodality*, *mad_rlog*, and *extreme_values_robust_z_score* reduce model fit and visibly distort the ordination.
(Exact values are reported in Table A.2–P1.)

Project 2 — Broiler caecal (longitudinal)

Baseline. Diet explains β-diversity with PERMANOVA $R^2 = 0.13327$.
Geometry preserved, negligible change. Filters that primarily reduce sparsity—e.g., *percentage_below_lod*, *min_count_in_fraction*, *IQR*, *cv_filter*, and *combined_mean_variance*—retain the baseline Bray–Curtis geometry with only trivial changes in R^2(well within sampling-level fluctuation).
Geometry preserved, contrast enhanced. Time-aware selectors—*time_correlation*, *correlation_with_covariate*, *low_correlation_with_covariate*, and *anova_p*—sharpen phase separation while preserving ordination geometry; R^2rises modestly (about $+0.02$–0.03), reflecting contrast enhancement rather than re-projection.
Apparent gains that are not meaningful. *autocorrelation_time* and *network_connectivity* increase R^2but warp the ordination (poor Mantel/Procrustes agreement), whereas *skewness*

and *gini* both reduce explained variation and distort geometry.

Interpretation rule (both projects). Changes in R^2 are only interpretable when geometry is preserved. Increases under poor Mantel/Procrustes reflect re-weighting artefacts, not improved separation of the original community structure.

Synthesis across projects

Best overall (preserve distances and geometry; small $+\Delta R^2$). prevalence, low_count, percentage_below_lod, IQR, combined_mean_variance and sparsity-oriented distributional filters; in P2 the time-aware time_correlation / (low_)correlation_with_covariate / anova_p also qualify, acting as contrast enhancers without re-projection.

Use with caution (moderate reshaping). entropy, variance_filter, dispersion, mad_vst, cv_filter—layouts remain recognisable but show measurable re-weighting.

Avoid when preservation is required (clear distortion; unstable R^2. bimodality, gini, combined_skewness_kurtosis, mad_rlog, extreme_values_robust_z_score; in P2 also zero_proportion, autocorrelation_time, network_connectivity.

These conclusions align with the visual PCoA assessments: sparsity/low-signal filters can markedly reduce feature space while maintaining the original β-diversity pattern, whereas shape-reweighting and extreme-focused methods often alter geometry and yield misleading changes in PERMANOVA R^2.

3.6 Summary of findings

Project 1 — Granular biofilm (cross-sectional)

Data and baseline. The cross-sectional granular-biofilm dataset comprised 24 samples × 1,829 ASVs. The unfiltered Bray–Curtis PCoA resolved four compact, well-separated fractions (Floating–cDNA, Settled–DNA, Settled–cDNA, Floating–DNA), which served as the reference configuration for all comparisons.

Compression without distortion. Filters targeting sparsity/low signal achieved large reductions while preserving β-geometry and, in several cases, slightly sharpening group separation.

- **Preserving set.** *prevalence, low_count, percentage_below_lod, IQR, combined_mean_variance* and distributional sparsity filters (*zero_proportion, zero_inflation, enhanced_low_count_percentage*) removed ~70–86% of features with Mantel ≈ 1 and Procrustes $\gtrsim 0.999$.
- **PERMANOVA.** R^2 was unchanged or modestly higher (Δ up to +0.022; e.g., *percentage_below_lod* ~0.646; *low_count* ~0.640; *prevalence* ~0.632; *IQR* ~0.625–0.636), consistent with noise reduction without re-projection.

Filters that altered geometry. Methods emphasising distributional shape or extreme values frequently re-projected the ordination.

- **Distorting set.** *bimodality, gini, combined_skewness_kurtosis, extreme_values_robust_z_score, cv_filter, mad_rlog* produced visible rotations/scale changes (several with Procrustes ≤ 0.85) and either inflated R^2 spuriously or reduced it substantially (e.g., *bimodality* ~0.416; *mad_rlog* ~0.492; *extreme_values_robust_z_score* ~0.609).
- **Moderate-impact utilities.** *entropy, mean_variance, dispersion,* and post-VST *variance_filter* showed small to noticeable reshaping; these are usable with caution when strict geometry preservation is required.
- **No-effect utilities.** *loq, mean_detection, min_count_in_x_samples* had no practical effect (feature counts unchanged).

α-diversity patterns. Under geometry-preserving filters, Shannon ordering remained clear (Floating–DNA highest; Settled–cDNA lowest), and Chao1 richness differences were modest and broadly ordered. Aggressive, distortion-prone filters (e.g., *mad_rlog, gini*) compressed richness and/or re-ranked Shannon, mirroring their β-diversity effects.

Relationship-based filters (within-stratum). Evaluated against stratum-matched baselines.

- *group_fold_diff* closely matched the baseline ordination (especially cDNA).
- *fold_change* introduced centroid shifts and ellipse elongations, most evident in DNA, indicating re-weighting towards large mean differences.

Practical guidance for Project 1.

1. **Recommended defaults (preservation paramount):** *prevalence, low_count, percentage_below_lod, IQR, combined_mean_variance, zero_proportion, zero_inflation, enhanced_low_count_percentage*.
2. **Use sparingly/with checks:** *entropy, mean_variance, dispersion, variance_filter, network_connectivity*.
3. **Avoid for preservation goals:** *bimodality, gini, combined_skewness_kurtosis, mad_rlog, extreme_values_robust_z_score, cv_filter*.
4. **For pairwise analyses:** prefer *group_fold_diff* over *fold_change*.

Overall for P1. Removing rare/low-information features compresses the feature space substantially without altering community structure; shape- or extreme-based filters tend to re-project distances and can mislead downstream interpretation. These results form the benchmark for evaluating the same families in the longitudinal setting of P2.

Project 2 — Broiler caecal (longitudinal)

Data and baseline. The longitudinal dataset comprised 337 samples × 2,309 OTUs. α-diversity increased rapidly early in life and plateaued by ~day 12; Bray–Curtis PCoA exhibited a Starter → Grower → Finisher trajectory. The unfiltered diet $R^2 = 0.13327$ served as the reference.

Safe compression with preserved geometry. Filters that reduce sparsity or capture stable variability preserved the temporal arc and left ΔR^2 negligible.

- **Preserving set.** *percentage_below_lod (302/2,309; −86.9%)*, *min_count_in_fraction (365/2,309; −84.2%)*, *IQR (705/2,309; −69.5%)*, *combined_mean_variance, cv_filter, variance/intra_group_variance,* and *mean_abundance/prevalence* maintained Mantel $\gtrsim 0.99$ and Procrustes $\gtrsim 0.99$ with trivial ΔR^2.
- **Time-aware selectors (contrast enhancement, not re-projection).** *time_correlation, correlation_with_covariate, low_correlation_with_covariate, anova_p* preserved geometry (Procrustes ≈ 0.956–0.959) while legitimately increasing R^2 to ≈ 0.156–0.160, in line with literature anchors that time/week explains ~16–17% of Bray–Curtis variance [vegan], [phyloseq].

Borderline changes. *low_count* (Procrustes ≈ 0.921) and *zero_inflation* (≈ 0.930) retain the global trajectory but shift local neighbourhoods; these are usable with caveats, especially for fine-grained longitudinal inferences.

Distorting filters. Methods that re-weight distributional shape or rely on structural heuristics altered the Bray–Curtis geometry.

- **Shape re-weighting:** *skewness, gini, bimodality, zero_proportion* caused bending/re-ordering of the temporal path (often Procrustes ≤ 0.93); any R^2 changes (including increases) are not interpretable.
- **Structural/other:** *extreme_values_robust_z_score* and *network_connectivity* also distorted geometry; *autocorrelation_time* provides a characteristic longitudinal failure mode—R^2 increases (0.16394) while Mantel/Procrustes degrade (0.891/0.931), indicating emphasis on persistence rather than the monotonic day gradient.

Dispersion and robustness. Observed beta-dispersion shifts around diet transitions reinforce that dispersion inferences should be statistically tested (e.g., *betadisper*), and that **R^2 changes are meaningful only when geometry is preserved**. Bray–Curtis remains the primary metric; Unweighted UniFrac typically yields smaller R^2 and should be used as a sensitivity check.

Practical guidance for Project 2 (single-filter, not a pipeline).

- **Default choices for safe compression:** *percentage_below_lod* or *min_count_in_fraction*; when more conservative, *IQR*.
- **Recommended set (geometry preserved):**
  • **Abundance-based:** *total_count, prevalence, mean_abundance, mean_detection, min_count_in_fraction, min_count_in_x_samples, percentage_below_lod*.
  • **Variability-based:** *IQR, combined_mean_variance, variance, intra_group_variance, mean_variance, dispersion, CV*.
  • **Distribution/other:** *entropy, enhanced_low_count_percentage*.
  • **Transformation-specific:** *cv_filter, variance_filter, mad_filter, iqr_filter, dropout_after_vst*.
  • **Time-aware (contrast enhancement, geometry preserved):** *time_correlation, correlation_with_covariate, low_correlation_with_covariate, anova_p*.
- **Use with caution:** *low_count* (Procrustes 0.921; ΔR^2 +0.00136), *zero_inflation* (0.930; +0.00030), *mad_vst* (0.967; −0.00369).
- **Avoid for longitudinal preservation:** *skewness, gini, bimodality, zero_proportion, autocorrelation_time, network_connectivity, extreme_values_robust_z_score* (Procrustes ≤ 0.93 and/or misleading R^2 changes).

Operational rule. Across P2, **R^2 is interpretable only when Mantel/Procrustes confirm geometric preservation**. Within this rule, sparsity/variability filters and time-aware selectors provide safe compression for time-ordered analyses; shape-/network-/autocorrelation-driven filters do not.

Cross-project perspective. Despite contrasting designs (P1 discrete groups vs P2 temporal arc), the best-performing filters coincide: sparsity/low-signal and stable-variability criteria compress aggressively yet preserve geometry; shape- or extreme-focused criteria distort. For P2, time-aware selectors provide legitimate contrast enhancement at Procrustes ≈ 0.956–0.959, aligning diet R^2 with literature anchors. A consistent interpretative principle holds in both settings: treat **R^2 changes as meaningful only when geometry is preserved**.

CHAPTER 4

DISCUSSION

4.1 Summary of principal findings

This study assessed how single-feature filters altered α-diversity, β-diversity, PCoA geometry, and PERMANOVA R^2 in two complementary designs. In P1 (cross-sectional anaerobic granular biofilm; DNA vs cDNA), light abundance and prevalence screens generally stabilised β-diversity relative to the unfiltered baseline. Mantel correlations remained high, symmetric Procrustes indicated near-isometric alignment, and PERMANOVA R^2 for floating/settled and DNA/cDNA contrasts changed only modestly. α-diversity behaved as expected: observed richness decreased, while evenness-weighted indices were preserved or slightly clarified by removal of ultra-rare features. These patterns are consistent with the published characterisation of DNA–cDNA strata and the heterogeneity between floating and settled fractions in the underlying study [32].

In P2 (longitudinal broiler cecum, days 3–35), conservative sparsity-oriented filters (abundance/prevalence and mild variability/distribution criteria) preserved the baseline Bray–Curtis geometry and the Starter→Grower→Finisher trajectory. Mantel and Procrustes diagnostics indicated minimal geometric change, and PERMANOVA R^2 for diet-phase contrasts was stable. α-diversity trends were retained, mirroring the known early rise and later stabilisation of diversity in this system [14]. Time-aware relationship filters (e.g., day-correlation) tightened phase separation and preserved trajectories when tuned conservatively; however, aggressive thresholds occasionally compressed neighbourhood structure. These outcomes align with the longitudinal dynamics and diet-linked dispersion reported for the source dataset[14].

Across both projects, single-filter analyses indicated that conservative, sparsity-oriented choices were the most geometry-stable and information-efficient. In particular, abundance- and prevalence-based single filters tended to maintain high Mantel and symmetric Procrustes agreement and to preserve PERMANOVA R^2 while limiting feature loss. Under severe zero inflation, mild distribution-oriented single filters showed similar stability; by contrast, aggressive distributional single filters and certain transformation- or model-linked selections produced mixed effects and sometimes distorted ordinations. These single-filter patterns suggest design-aware but conservative defaults—e.g., stratum-specific thresholds for P1's DNA/cDNA layers and cautiously tuned time-aware criteria for P2's repeated-measures setting—without implying that combined pipelines were evaluated. These observations motivate a mechanism-based explanation developed in §4.2.

4.2 Interpretation in light of data properties

Observed patterns arise from sparsity, zero inflation, and compositionality that characterise amplicon tables. Many taxa are rare or sample-unique, creating stochastic zeros that destabilise dissimilarities and inflate dispersion. Light abundance–prevalence screening reduces these stochastic contributions, so pairwise distances change little and ordinations rotate rather than deform. This explains why geometry and PERMANOVA R^2 remained stable under conservative filters in both projects. These mechanisms are consistent with the diagnostics-driven aims outlined for this analysis.

Compositional closure further couples all features through a unit-sum constraint. Adding or removing components redistributes mass and can shift distances even when absolute abundances are unchanged. Filtering therefore interacts with transformations. Variance-stabilising steps dampen mean–variance coupling, but over-filtering can alter the variance structure and reweight distances. Log-ratio transforms mitigate closure by operating in Aitchison geometry; however, they require careful zero handling given the excess-zero

nature of amplicon counts [11,19,24,34].Removing near-structural zeros before log-ratio transformation yields more coherent coordinates and stable layouts, while aggressive distributional filters can induce non-isometric re-projection. These principles align with the motivating literature on compositional microbiome analysis [4,9].

Metric choice also conditions stability. Bray–Curtis emphasises abundant taxa and was robust under sparsity-oriented filters, consistent with preserved Mantel and symmetric Procrustes. Unweighted UniFrac amplifies rare features; consequently, filters that reshape distribution tails or target extremes can warp neighbourhoods and alter PERMANOVA R^2 despite small distance changes. Aitchison distances were most stable once ultra-rare features were excluded, and zeros were addressed appropriately. The empirical distortions seen for tail-reweighting, autocorrelation-based, or network-degree selections illustrate these mechanisms by bending trajectories or inflating apparent separation without true structural change.

Together, these considerations link data properties to the observed stability range, clarifying why conservative sparsity filters preserve geometry while aggressive shape-based criteria risk misleading inferences.

4.3 Design-specific insights (P1 vs P2)

Design shaped how filters behaved and which diagnostics remained stable. P1 was cross-sectional with parallel DNA and cDNA strata. P2 was longitudinal with repeated measures across days 3–35. These contexts differ in sparsity profiles, dependence structures, and the risk of over-removal. Consequently, thresholds that were benign in one setting could be disruptive in the other. The following contrasts explain when shared versus design-aware strategies were preferable.

In P1, conservative abundance–prevalence screens preserved geometry across DNA and cDNA with high Mantel and symmetric Procrustes agreement. PERMANOVA R^2 for floating versus settled fractions remained stable, indicating retained group structure despite feature attrition. Shared thresholds were adequate when library sizes and zero inflation were comparable between strata. However, cDNA often displayed higher zero inflation and lower counts, reflecting transcriptional specificity. Stratum-specific thresholds were therefore safer, preventing removal of active but low-abundance taxa central to cDNA contrasts. Over-filtering cDNA compressed ordinations and reduced R^2, consistent with signal loss in biologically meaningful rare features [32]. Active but low-abundance taxa in cDNA can carry functional signal; their removal reduces contrast even when DNA-level structure remains.

In P2, time-aware relationship filters helped preserve within-subject trajectories and phase transitions when tuned conservatively. Mantel and Procrustes metrics indicated small geometric change relative to the baseline, and diet-phase PERMANOVA R^2 remained stable. Between-subject heterogeneity was substantial, especially under rare-taxa-sensitive metrics. Light abundance–prevalence screens reduced spurious day-to-day noise yet retained individual trajectories. Aggressive smoothing or tail-truncation distorted neighbourhoods, attenuating phase separation and underestimating cross-sectional variability important for inference [14].

Convergence across designs appeared for conservative sparsity screens, which consistently stabilised geometry with minimal loss. Divergence emerged for distributional and relationship filters, which required stratum-specific or time-aware tuning. These design-specific patterns motivated the sensitivity analyses and distance substitutions presented next.

4.4 Sensitivity and robustness

Threshold sweeps identified a broad stability band for light filtering and a narrow band for aggressive criteria. Within the stable band, Bray–Curtis ordinations changed little, Mantel correlations remained high, and symmetric Procrustes indicated rotations without

deformation. PERMANOVA R^2 for primary contrasts was preserved, suggesting that group structure was robust to moderate feature loss. Outside this band, compressed ordinations and reordered neighbourhoods indicate that informative, moderately rare taxa were removed.
Distance substitutions showed metric-dependent behaviour. Bray–Curtis aligned closely with weighted UniFrac under conservative thresholds, reflecting emphasis on abundant taxa and shared structure. Unweighted UniFrac remained sensitive to rare features; light filtering improved agreement with Bray–Curtis, but aggressive tail trimming sometimes exaggerated separation or erased weak signals. Aitchison distances were notably stable once near-structural zeros were addressed, consistent with ratio geometry. When log-ratio analysis followed zero handling, trajectories and between-group relations were maintained despite feature attrition.
Design influenced robustness but did not overturn conclusions. In P1, stratum-specific thresholds reduced DNA–cDNA discrepancies across metrics. In P2, conservative time-aware filters preserved within-subject trajectories under Bray–Curtis and weighted UniFrac, while unweighted UniFrac showed mixed responses. Across both projects, diagnostics converged on the same stability band: light abundance–prevalence single filters, with mild distribution-oriented single filters showing comparable stability under severe zero inflation. Combined filters were not evaluated in this study.
Implementation choices were consistent with these results. Distance calculations and ordinations were implemented in R using phyloseq [18] and vegan [23], with upstream preprocessing following QIIME 2 [7]. These sensitivity findings support conservative defaults and transparent reporting of tested ranges and diagnostic outcomes.

4.5 Methodological implications and practical guidance

Evidence from single-filter trials supports the following recommendation for future workflows: use a light abundance–prevalence single filter; when zero inflation is severe, consider a mild distribution-oriented single filter. This recommendation is extrapolated from single-filter evidence and does not imply that combined or sequential filters were empirically evaluated here. Avoid aggressive tail trimming that compresses ordinations or reduces R^2; these choices prioritise geometry preservation and minimise feature loss.
Design-aware refinements remain relevant. In cross-sectional, multi-layer contexts like P1, stratum-specific thresholds are advisable when library sizes or sparsity differ between DNA and cDNA. In longitudinal contexts like P2, time-aware single filters can stabilise trajectories when tuned conservatively; avoid oversmoothing that erases between-subject variation. Align filters with transformation choices. Under variance-stabilising transforms, filter lightly to respect the mean–variance structure. Under log-ratio analysis, remove near-structural zeros and document zero handling to ensure coherent Aitchison geometry. For phylogenetic distances, note that unweighted UniFrac is sensitive to rare taxa and may require gentler thresholds than Bray–Curtis or weighted UniFrac.
Reporting should enable exact re-execution. Provide a compact record of thresholds tested and selected, features removed, and resulting table dimensions. Track the change in zero proportion. Report Mantel and symmetric Procrustes relative to the unfiltered baseline. Report PERMANOVA R^2 before and after filtering, and ΔR^2 for primary contrasts. Document distance metrics, transformation choices, random seeds, and full software versions. Archive scripts and parameter dictionaries, consistent with open, reproducible practice [13]. A concise checklist is: thresholds and sweep range; retained dimensions; zero-proportion shift; Mantel/Procrustes vs baseline; PERMANOVA R^2 and ΔR^2; metrics and transforms; versions, seeds, and code repository.

4.6 Limitations and future directions

This study was a secondary analysis and therefore inherited upstream choices from the original sources for P1 and P2, including sample processing, denoising, chimera removal,

and taxonomic assignment[14, 32]. These steps influence sparsity, zero inflation, and library-size variation, which in turn condition filtering effects and metric stability. Some filters may remove rare but biologically meaningful taxa, particularly in cDNA strata or early successional stages, leading to attenuated PERMANOVA R^2 or compressed ordinations. Copy-number variation in 16S rRNA genes can bias apparent abundances and distance calculations. Potential contamination, batch effects, and uneven sequencing depth remain additional risks despite conservative filtering and diagnostic checks.

Several aspects were not exhaustively evaluated. Filters were applied individually rather than as optimised combinations; interactions among abundance, distributional, and relationship-based criteria may yield different stability profiles. Transformation choices were considered, but formal model selection for transformation–filter pairs was outside scope. Time-aware methods were tuned conservatively; richer repeated-measures ordinations that incorporate subject-level random effects could better preserve trajectories in longitudinal designs.

Future work should incorporate negative and mock controls to calibrate zero handling and benchmark thresholds. Multi-omics integration could test whether 16S filtering preserves concordant signals in metagenomes or metabolomes. Data-driven threshold selection—e.g., stability selection, cross-validation, or Bayesian shrinkage—should be compared against heuristic rules. Finally, a transparent parameter dictionary, versioned code, and containerised workflows should be standardised to strengthen reproducibility across laboratories [13].

CHAPTER 5

CONCLUSION

This study systematically evaluated how filters affect the stability of α-diversity, β-diversity, and PCoA geometry in 16S rRNA amplicon data. A unified diagnostic framework was applied to two complementary designs: a cross-sectional granular-biofilm community (Project 1) and a longitudinal broiler-cecum microbiome (Project 2). The comparison spanned seven filter families: abundance-based, variability-based, distribution-based, relationship-based, transformation-specific, model-based and other, and network-based. Geometry and structure were assessed using Mantel correlation, symmetric Procrustes correlation, and PERMANOVA R^2; distances included Bray–Curtis and UniFrac variants. Standard tools were used for data handling and statistics (QIIME 2; phyloseq; vegan) [7,18,23]. Primary datasets followed published designs and upstream choices[14,32].

Across both projects, abundance- and prevalence-oriented single filters preserved β-diversity geometry most consistently. Mantel and symmetric Procrustes correlations remained high, indicating near-baseline concordance of distances and ordinations. When geometry was preserved, PERMANOVA R^2 values were stable or changed only slightly in interpretable ways, supporting the conclusion that group structure remained intact. Variability-based and mild distribution-based filters behaved similarly when tuned in alignment with data characteristics, primarily by attenuating stochastic zeros and low-signal noise.

Design-specific insights emerged from each dataset. In P1 (DNA vs cDNA), differences in sparsity profiles supported stratum-specific thresholds rather than uniform criteria, in order to avoid removal of low-abundance but biologically active cDNA features [32]. In P2 (days 3–35), time-aware single filters modestly enhanced phase separation (Starter→Grower→Finisher) without re-projecting samples into new configurations. Mantel and Procrustes diagnostics confirmed geometric stability, while changes in R^2 were small and ecologically interpretable[14] (see Table A.2–P2).

These patterns align with known properties of amplicon data: sparsity, zero inflation, and compositional closure [4,9,11,19,24,34]. Reducing random zeros plausibly stabilises inter-sample distances and ordinations by suppressing noise from rare features. Bray–Curtis and weighted UniFrac, which down-weight rare taxa, tended to preserve geometry under appropriate filters. Aitchison geometry was not computed here; however, it provides a conceptual rationale for how principled zero handling and compositional scaling can improve interpretability when such transformations are applied in future analyses. Together, these considerations explain why filters targeting sparsity and low signal preserved geometry, whereas some distribution-shape, transformation-specific, or model-linked rules occasionally altered R^2 while distorting ordinations.

Practical guidance follows directly from the evidence. Abundance- or prevalence-based filters are recommended as default single-filter options for exploratory preprocessing. Variability- or mild distribution-based filters are suitable when zero inflation is pronounced, provided geometry diagnostics remain favourable. Interpretation of PERMANOVA R^2 is valid only when high Mantel and strong symmetric Procrustes correlations confirm preserved geometry. Parameter choices should reflect design context: stratum-specific in P1 to protect cDNA signal, and time-aware in P2 to maintain longitudinal trajectories. Transparent reporting of filter names, parameter values, features removed, zero-proportion changes, and ΔR^2 relative to the unfiltered baseline supports reproducibility and reuse (see Fig. 3.4B; Table A.2–P2).

This work had a defined scope. Filters were assessed individually, and conclusions therefore apply to single-filter use; combined or sequential pipelines were not empirically evaluated in this analysis. Secondary results inherit upstream choices and constraints as reported in the

source studies [14,32]. Future work could examine interactions among filter families, evaluate transformation-specific criteria more broadly, and compare behaviour across additional distance metrics.

In summary, this thesis provides evidence-based guidance for selecting single-filter strategies that balance feature reduction with ecological fidelity. The diagnostics and recommendations presented here establish a reproducible preprocessing baseline across cross-sectional and longitudinal designs, and they remain adaptable to evolving methods and pipelines[7,14,18,23].

REFERENCES

[1]. Aitchison, J. (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological)*, *44*(2), 139-160. https://doi.org/10.1111/j.2517-6161.1982.tb01195.x

[2]. Amir, A., McDonald, D., Navas-Molina, J. A., Kopylova, E., Morton, J. T., Zech Xu, Z., ... & Knight, R. (2017). Deblur rapidly resolves single-nucleotide community sequence patterns. *MSystems*, *2*(2), 10-1128. https://doi.org/10.1128/mSystems.00191-16

[3]. Anderson, M. J. (2001). A new method for non-parametric multivariate analysis of variance. *Austral ecology*, *26*(1), 32-46. https://doi.org/10.1111/j.1442-9993.2001.01070.pp.x

[4]. Armstrong, G., Rahman, G., Martino, C., McDonald, D., Gonzalez, A., Mishne, G., & Knight, R. (2022). Applications and comparison of dimensionality reduction methods for microbiome data. *Frontiers in bioinformatics*, *2*, 821861. https://doi.org/10.3389/fbinf.2022.821861

[5]. Bharti, R., & Grimm, D. G. (2021). Current challenges and best-practice protocols for microbiome analysis. *Briefings in bioinformatics*, *22*(1), 178-193. https://doi.org/10.1093/bib/bbz155

[6]. Rai, S. N., Qian, C., Pan, J., Rai, J. P., Song, M., Bagaitkar, J., ... & McClain, C. J. (2021). Microbiome data analysis with applications to pre-clinical studies using QIIME2: Statistical considerations. Genes & diseases, 8(2), 215-223. https://doi.org/10.1016/j.gendis.2019.12.005

[7]. Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C. C., Al-Ghalith, G. A., ... & Caporaso, J. G. (2019). Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nature biotechnology*, *37*(8), 852-857. https://doi.org/10.1038/s41587-019-0209-9

[8]. Bray, J. R., & Curtis, J. T. (1957). An ordination of the upland forest communities of southern Wisconsin. Ecological monographs, 27(4), 326-349. https://doi.org/10.2307/1942268

[9]. Calle, M. L. (2019). Statistical analysis of metagenomics data. Genomics & informatics, 17(1), e6. https://doi.org/10.5808/GI.2019.17.1.e6

[10]. Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. Nature methods, 13(7), 581-583. https://doi.org/10.1038/nmeth.3869

[11]. Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V., & Egozcue, J. J. (2017). Microbiome datasets are compositional: and this is not optional. Frontiers in microbiology, 8, 2224. https://doi.org/10.3389/fmicb.2017.02224

[12]. Gower, J. C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika, 53(3-4), 325-338. https://doi.org/10.1093/biomet/53.3-4.325

[13]. Ijaz, U. Z. (2025). *Streamlining Omics Data: A Deep Dive into Effective Feature Filtering Methodologies*. In preparation.

[14]. Ijaz, U. Z., Sivaloganathan, L., McKenna, A., Richmond, A., Kelly, C., Linton, M., ... & Gundogdu, O. (2018). Comprehensive longitudinal microbiome analysis of the chicken cecum reveals a shift from competitive to environmental drivers and a window of opportunity for Campylobacter. Frontiers in microbiology, 9, 2452. https://doi.org/10.3389/fmicb.2018.02452

[15]. Johnson, J. S., Spakowicz, D. J., Hong, B. Y., Petersen, L. M., Demkowicz, P., Chen, L., ... & Weinstock, G. M. (2019). Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. Nature communications, 10(1), 5029. https://doi.org/10.1038/s41467-019-13036-1

[16]. Kamble, A., Sawant, S., & Singh, H. (2020). 16S ribosomal RNA gene-based metagenomics: A review. *Biomedical Research Journal*, *7*(1), 5-11. DOI: 10.4103/BMRJ.BMRJ_4_20

[17]. Lozupone, C., & Knight, R. (2005). UniFrac: a new phylogenetic method for comparing microbial communities. Applied and environmental microbiology, 71(12), 8228-8235. https://doi.org/10.1128/AEM.71.12.8228-8235.2005

[18]. McMurdie, P. J., & Holmes, S. (2013). phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. PloS one, 8(4), e61217. https://doi.org/10.1371/journal.pone.0061217

[19]. McMurdie, P. J., & Holmes, S. (2014). Waste not, want not: why rarefying microbiome data is inadmissible. PLoS computational biology, 10(4), e1003531. https://doi.org/10.1371/journal.pcbi.1003531

[20]. Nikodemova, M., Holzhausen, E. A., Deblois, C. L., Barnet, J. H., Peppard, P. E., Suen, G., & Malecki, K. M. (2023). The effect of low-abundance OTU filtering methods on the reliability and variability of microbial composition assessed by 16S rRNA amplicon sequencing. Frontiers in cellular and infection microbiology, 13, 1165295. https://doi.org/10.3389/fcimb.2023.1165295

[21]. Ning, D., Deng, Y., Tiedje, J. M., & Zhou, J. (2019). A general framework for quantitatively assessing ecological stochasticity. *Proceedings of the National Academy of Sciences*, *116*(34), 16892-16898. https://doi.org/10.1073/pnas.1904623116

[22]. Okuda, S., Tsuchiya, Y., Kiriyama, C., Itoh, M., & Morisaki, H. (2012). Virtual metagenome reconstruction from 16S rRNA gene sequences. *Nature communications*, *3*(1), 1203.https://doi.org/10.1038/ncomms2203

[23]. Oksanen, J., Blanchet, F. G., Friendly, M., et al. (2022). *vegan: Community Ecology Package* (Version 2.7–1). https://CRAN.R-project.org/package=vegan

[24]. Paulson, J. N., Stine, O. C., Bravo, H. C., & Pop, M. (2013). Differential abundance analysis for microbial marker-gene surveys. Nature methods, 10(12), 1200-1202. https://doi.org/10.1038/nmeth.2658

[25]. Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., ... & Glöckner, F. O. (2012). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic acids research, 41(D1), D590-D596. https://doi.org/10.1093/nar/gks1219

[26]. R Core Team. (2023). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

[27]. Regueira-Iglesias, A., Balsa-Castro, C., Blanco-Pintos, T., & Tomás, I. (2023). Critical review of 16S rRNA gene sequencing workflow in microbiome studies: From primer selection to advanced data analysis. Molecular Oral Microbiology, 38(5), 347-399. https://doi.org/10.1111/omi.12434

[28]. Rognes, T., Flouri, T., Nichols, B., Quince, C., & Mahé, F. (2016). VSEARCH: a versatile open source tool for metagenomics. PeerJ, 4, e2584. https://doi.org/10.7717/peerj.2584

[29]. Stegen, J. C., Lin, X., Konopka, A. E., & Fredrickson, J. K. (2012). Stochastic and deterministic assembly processes in subsurface microbial communities. The ISME journal, 6(9), 1653-1664. https://doi.org/10.1038/ismej.2012.22

[30]. Stegen, J. C., Lin, X., Fredrickson, J. K., & Konopka, A. E. (2015). Estimating and mapping ecological processes influencing microbial community assembly. Frontiers in microbiology, 6, 370. https://doi.org/10.3389/fmicb.2015.00370

[31]. Thompson, L. R., Sanders, J. G., McDonald, D., Amir, A., Ladau, J., Locey, K. J., ... & Knight, R. (2017). A communal catalogue reveals Earth's multiscale microbial diversity. Nature, 551(7681), 457-463. https://doi.org/10.1038/nature24621

[32]. Trego, A. C., McAteer, P. G., Nzeteu, C., Mahony, T., Abram, F., Ijaz, U. Z., & O'Flaherty, V. (2021). Combined stochastic and deterministic processes drive community assembly of anaerobic microbiomes during granule flotation. Frontiers in microbiology, 12, 666584. https://doi.org/10.3389/fmicb.2021.666584

[33]. Webb, C. O., Ackerly, D. D., McPeek, M. A., & Donoghue, M. J. (2002). Phylogenies and community ecology. *Annual review of ecology and systematics*, *33*(1), 475-505. https://doi.org/10.1146/annurev.ecolsys.33.010802.150448

[34]. Weiss, S., Xu, Z. Z., Peddada, S., Amir, A., Bittinger, K., Gonzalez, A., ... & Knight, R. (2017). Normalization and microbial differential abundance strategies depend upon data characteristics. Microbiome, 5(1), 27. https://doi.org/10.1186/s40168-017-0237-y

[35]. Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis.* Springer. https://ggplot2.tidyverse.org

[36]. Vass, M., Székely, A. J., Lindström, E. S., & Langenheder, S. (2020). Using null models to compare bacterial and microeukaryotic metacommunity assembly under shifting environmental conditions. Scientific Reports, 10(1), 2455. https://doi.org/10.1038/s41598-020-59182-1

APPENDIX

Table A.1 — Summary of datasets and retained baselines

Dataset	Design	Samples (original)	Samples (retained)	Features (original)	Features (retained)	Notes
Project 1	Cross-sectional (floating vs. settled; DNA/cDNA)	24	24	–	1,829	Source study reports significant DNA / position effects on β-diversity (PERMANOVA p = 0.001).
Project 2	Longitudinal (day 3–35; ≈ 12 replicates per day)	379	337	18,588 OTUs	2,309	Early α-diversity rise → plateau; shift in community drivers (days 12–20); diet phase aligns with β-dispersion.

Table A.1. Summary of the datasets analysed in this study and the harmonised baselines retained after preprocessing. Sample and feature counts correspond to the unfiltered matrices used for downstream diversity and filtering analyses.

Table A.2–P1. Feature-retention and β-diversity stability by single filters in Project 1 (ASV level; Bray–Curtis + PCoA).

Category	Filter	Features (after/before)	Reduction %	Mantel r	Procrustes	PERMANOVA R^2	ΔR^2 vs. base
Abundance-based	total_count	813/1829	55.50%	0.999997	0.642	0.627	0.003
	prevalence	509/1829	72.20%	0.999945	0.9999	0.632	0.008
	mean_abundance	588/1829	67.90%	0.999991	≈1.000	0.629	0.005
	mean_detection	1829/1829	0.00%	1	1	0.624	0
	low_count	364/1829	80.10%	0.9995	0.9997	0.64	0.016
	min_count_in_fraction	321/1829	82.40%	0.9995	0.867	0.642	0.018
	min_count_in_x_samples	1829/1829	0.00%	1	1	0.624	0
	percentage_below_lod	253/1829	86.20%	0.9992	0.9995	0.646	0.022
Variability-based	mean_variance	1829/1829	0.00%	1	0.864	0.624	0
	dispersion	1827/1829	0.10%	0.99999	0.689	0.625	0
	IQR	333/1829	81.80%	0.999	0.9993	0.625	0
	CV	1825/1829	0.20%	0.99998	0.9999	0.625	0.001
	intra_group_variance	1259/1829	31.20%	1	1	0.625	0.001
	combined_mean_variance	342/1829	81.30%	0.9999	0.9998	0.636	0.012
	variance	1118/1829	38.90%	0.99999	0.9997	0.626	0.001
Distribution-based	entropy	618/1829	66.20%	0.99998	0.864	0.63	0.006
	zero_proportion	509/1829	72.20%	0.9999	0.9999	0.632	0.008
	gini	184/1829	89.90%	0.97	0.754	0.703	0.079
	bimodality	1727/1829	5.60%	0.888	0.514	0.416	−0.208

	skewness	294/1829	83.90%	0.981	0.9945	0.688	0.064
	combined_skewness_kurtosis	140/1829	92.30%	0.95	0.983	0.736	0.112
	zero_inflation	509/1829	72.20%	0.9999	0.9999	0.632	0.008
	enhanced_low_count_percentage	498/1829	72.80%	0.9999	0.9999	0.633	0.009
Transformation-specific	mad_vst	183/1829	90.00%	0.995	0.874	0.67	0.046
	mad_rlog	48/1829	97.40%	0.819	0.853	0.492	−0.132
	dropout_after_vst	1829/1829	0.00%	1	1	0.624	0
	variance_filter	1464/1829	19.90%	0.985	0.9951	0.618	−0.006
	extreme_values_robust_z_score	1715/1829	6.20%	0.92	0.576	0.609	−0.015
	mad_filter	1464/1829	19.90%	1	1	0.625	0.001
	cv_filter	173/1829	90.50%	0.992	0.572	0.674	0.05
	iqr_filter	1464/1829	19.90%	1	1	0.625	0.001
Model-based	loq	1829/1829	0.00%	1	1	0.624	0
Network-based	network_connectivity	1618/1829	11.50%	0.998	0.9991	0.642	0.018

Design & baseline. Cross-sectional granular biofilm community; unfiltered baseline **24 × 1,829** (samples × features).
Columns. **Features (after/before)** reports retained ASVs over the **1,829-feature** baseline; **Reduction %** is the percentage reduction relative to **1,829**. **Mantel r** compares pre- vs post-filter Bray–Curtis distance matrices; **Procrustes** is the symmetric Procrustes correlation between baseline and filtered PCoA ordinations. **PERMANOVA** R^2uses a factorial design with **Nucleic acid (DNA vs cDNA)** and **Biomass position (floating vs settled)**; **ΔR^2 vs. base** is the difference from the unfiltered baseline ($R^2 = 0.624$).
Classes. Filters are grouped as **Abundance-based**, **Variability-based**, **Distribution-based**, **Transformation-specific**, **Model-based/Other**, and **Network-based**.
Scope. This table enumerates **non-relationship filters**. Relationship-based

procedures (e.g., *anova_p*, *group_fold_diff*, *fold_change*) are reported separately where applicable.
Analysis settings. Distance = **Bray–Curtis**; Ordination = **PCoA**; permutations for Mantel and PERMANOVA = **999**.
*Rounding. Mantel r is reported to **five** decimal places; Procrustes to **three** (four when ≥ 0.995); PERMANOVA R^2 and ΔR^2 to **five**; Reduction to **one** decimal place.Sample retention.* No filters removed samples.

Table A.2–P2. Feature-retention and β-diversity stability by single filters in Project 2 (OTU level; Bray–Curtis + PCoA).

Category	Filter	Features (after/before)	Reduction %	Mantel r	Procrustes	PERMANOVA R^2	ΔR^2 vs. base
Abundance-based filters	total_count	1636/2309	29.20%	0.9999999	0.9979	0.13328	0.00001
	prevalence	1005/2309	56.50%	0.9996455	0.9979	0.13357	0.0003
	mean_abundance	737/2309	68.10%	0.9999907	1	0.13349	0.00022
	mean_detection	2309/2309	0.00%	1	0.9981	0.13327	0
	low_count	475/2309	79.40%	0.9966352	0.921	0.13463	0.00136
	min_count_in_fraction	365/2309	84.20%	0.9927164	0.9964	0.13458	0.00131
	min_count_in_x_samples	2309/2309	0.00%	1	0.9998	0.13327	0
	percentage_below_lod	302/2309	86.90%	0.9884977	0.9866	0.1341	0.00083
Variability-based filters	mean_variance	2285/2309	1.00%	1	0.995	0.13327	0
	dispersion	2285/2309	1.00%	1	0.9897	0.13327	0
	IQR	705/2309	69.50%	0.9976191	0.9968	0.13376	0.00049
	CV	2309/2309	0.00%	1	0.9981	0.13327	0
	intra_group_variance	1177/2309	49.00%	0.9999991	1	0.13334	0.00007
	combined_mean_variance	483/2309	79.10%	0.9999439	0.9981	0.13342	0.00015
	variance	1079/2309	53.30%	0.9999986	0.9998	0.13336	0.00009
Distribution-	entropy	1963/2309	15.00%	0.9999968	0.9979	0.13331	0.00004

based filters							
	zero_proportion	1005/2309	56.50%	0.9996455	0.8464	0.13357	0.0003
	gini	168/2309	92.70%	0.864286	0.7734	0.10896	−0.02431
	bimodality	2093/2309	9.40%	0.6168042	0.702	0.13874	0.00547
	skewness	22/2309	99.00%	0.5228126	0.3055	0.05109	−0.08218
	zero_inflation	1005/2309	56.50%	0.9996455	0.9304	0.13357	0.0003
	enhanced_low_count_percentage	1005/2309	56.50%	0.9996455	0.9949	0.13357	0.0003
Relationship-based filters	time_correlation	999/2309	56.70%	0.9565882	0.9555	0.16041	0.02714
	autocorrelation_time	579/2309	74.90%	0.8914561	0.9311	0.16394	0.03067
	anova_p	983/2309	57.40%	0.9674892	0.9558	0.15625	0.02298
	correlation_with_covariate	999/2309	56.70%	0.9565882	0.9594	0.16041	0.02714
	low_correlation_with_covariate	999/2309	56.70%	0.9565882	0.9562	0.16041	0.02714
Transformation-specific filters	mad_vst	387/2309	83.20%	0.9865278	0.9673	0.12958	−0.00369
	dropout_after_vst	2309/2309	0.00%	1	1	0.13327	0

	variance_filter	1848/2309	20.00%	1	1	0.13328	0.00001
	extreme_values_robust_z_score	2129/2309	7.80%	0.6988687	0.7748	0.13321	−0.00006
	mad_filter	1848/2309	20.00%	0.9999145	0.9999	0.13278	−0.00049
	cv_filter	451/2309	80.50%	0.9965903	0.9957	0.13425	0.00098
	iqr_filter	1848/2309	20.00%	0.9996421	0.9998	0.13322	−0.00005
Model-based and Other Filters	loq	2309/2309	0.00%	1	0.9998	0.13327	0
Network-based filters	network_connectivity	530/2309	77.10%	0.7513228	0.818	0.20179	0.06852

Design & baseline. Longitudinal broiler caecum; unfiltered baseline **337 × 2,309** (samples × features).
Columns. **Features (after/before)** reports retained features over the 2,309-feature baseline; **Reduction %** is the percentage reduction relative to **2,309**. **Mantel r** compares pre- vs post-filter Bray–Curtis distance matrices; **Procrustes** is the symmetric Procrustes correlation between baseline and filtered PCoA ordinations. **PERMANOVA** R^2uses **Diet** (Starter → Grower → Finisher) as the grouping factor; ΔR^2 **vs. base** is the difference from the unfiltered baseline ($R^2 = 0.13327$).
Analysis settings. Distance = **Bray–Curtis**; Ordination = **PCoA**; permutations for Mantel and PERMANOVA = **999**.

Table A.3–P1. Relationship-based filters

Category	Filter	Features (after/before)	Reduction %	Mantel r	Procrustes corr.	PERMANOVA R^2
Relationship-based	anova_p	246 / 1829	86.60%	0.97	0.89	0.7169
Relationship-based	group_fold_diff (cDNA)	619 / 1829	66.20%	1	0.9967	0.5366
Relationship-based	group_fold_diff (DNA)	537 / 1829	70.70%	1	0.6822	0.3415
Relationship-based	fold_change (cDNA)	1750 / 1829	4.30%	0.9864	0.9354	0.5821

Relationship-based	fold_change (DNA)	1765 / 1829	3.50%	0.952	0.7185	0.4254