Supplementary Method S5: Coverage profiling, the bloominess metric, and robustness analysis

---

1. Coverage profiling with anvi'o

Following dereplication and quality filtering, all Illumina samples were mapped to the final MAG catalogue to generate the per-sample abundance profiles used in all downstream analyses.

1a. Read preparation

Paired-end Illumina reads were concatenated across sequencing runs for each sample. For samples spread across multiple run directories, R1 and R2 FASTQ files were identified by sample ID and merged using `cat` into per-sample combined FASTQ files (`01_merge_illumina_runs.sh`). A per-sample manifest was written recording which source files contributed to each merged output.

1b. Mapping to the MAG catalogue

A minimap2 index was pre-built from the fixed MAG catalogue FASTA (`NICE_MAG_COLLECTION-fixed.fa`). Illumina paired-end reads were mapped to this index using **minimap2** with the short-read alignment preset (`-ax sr`; script: `02_map_to_mags.sh` for single-sample runs; `04_map_and_profile_batched.sh` for batched processing of three samples in parallel). Alignments were sorted with `samtools sort` and indexed with `samtools index`. A per-sample mapping summary was written recording the number of mapped reads and any detected pair-count mismatches.

1c. anvi'o profiling and merging

Each BAM file was profiled with `anvi-profile` (`03_profile_mags.sh`; `04_map_and_profile_batched.sh`), providing the MAG contigs database (`NICE_MAGs-CONTIGS.db`), 8 threads, and a minimum contig length of 1,000 bp for analysis. Profiles were run three in parallel to balance throughput with memory requirements.

Individual profiles were then merged with `anvi-merge` (`05_merge_profiles.sh`), producing the merged profile database used for all abundance extraction. BAM mapping statistics (total reads, mapped reads, percentage mapped, properly paired reads) were summarised across samples using `samtools flagstat` (`06_bam_flagstat_summary.sh`) to verify alignment quality.

The key output used for all downstream analyses was the `mean_coverage_Q2_Q3` table exported from the merged anvi'o profile.

---

2. Abundance metric

The primary abundance measure for all MAG-level analyses was the **mean interquartile coverage** (`mean_coverage_Q2_Q3`), which anvi'o defines as the mean per-base read depth calculated over the middle two quartiles (Q2–Q3) of covered positions for each MAG in each sample. This metric reduces the influence of uneven coverage across high-copy regions, positional biases at contig ends, and mapping artefacts, and is more comparable across MAGs of different sizes than raw mean coverage or percentage recruitment.

---

3. Presence and prevalence definitions

Presence

A MAG was considered present in a sample if it met a dual threshold:

`mean_coverage_Q2_Q3  0.5×  AND  detection  0.2`

Coverage threshold (0.5×): excludes low-level mapping noise from non-specific alignment while retaining biologically meaningful low-abundance populations.

Detection threshold (0.2): detection in anvi'o is the fraction of contig positions covered by at least one read. A detection threshold of 0.2 requires that at least 20% of the MAG's positions are covered, filtering out cases where a few reads produce deep but spatially concentrated signal — a pattern associated with strain-level cross-mapping between related MAGs rather than true presence.

Prevalence

Prevalence was calculated per MAG as the proportion of samples in which the MAG was present:

```
prevalence = n_samples_present / n_samples_total
```

All prevalence calculations use the dual-threshold presence definition above.

---

4. Core microbiome definitions

MAGs were classified into ecological groups based on prevalence and within-host abundance:

Class	Prevalence criterion
Core	85% of samples
Variable	$>$ 10% and $<$ 85% of samples
Rare	10% of samples
Singleton	present in exactly 1 sample

Singleton MAGs were excluded from core analyses but retained for separate accessory analyses.

---

5. The bloominess metric

Bloominess quantifies the degree of skew in a MAG's within-host abundance distribution across the samples in which it is present. Specifically, it measures whether a MAG is disproportionately abundant in a small subset of samples (bloomy: the mean exceeds the median) or whether its abundance is evenly distributed across all present samples (non-bloomy: mean and median are similar).

The metric is defined as:

```
bloominess = log10(mean_present +  ) - log10(median_present +  )
```

where:

- **mean_present** = arithmetic mean of `mean_coverage_Q2_Q3` across samples in which the MAG is present (meeting the dual presence threshold)
- **median_present** = median of the same values
- **= 10** is a small constant added for numerical stability at the log transformation

Bloominess is only computed for MAGs present in more than one sample (`n_present > 1`), since a single observation provides no distributional information.

A bloominess value of 0 indicates a perfectly symmetric abundance distribution; positive values indicate right-skew (bloom-like episodes of high abundance in some samples); values approaching 0 from above indicate stable, flat abundance profiles across all samples where the MAG is detected. The log-ratio formulation makes the metric scale-invariant: a MAG detected at 2× in most samples and 200× in a few (100-fold

bloom) has the same bloominess as one detected at 20× and 2000× (same fold-difference). This allows comparison across MAGs spanning a wide range of absolute abundances.

---

6. Robustness of bloominess to threshold choice

Because the presence threshold (Section 3) determines which samples enter the mean and median calculation, we tested whether bloominess rankings were sensitive to the choice of coverage threshold. The analysis is implemented in `09_analysis/prevalence_thresholds_bloominess.Rmd` and the Figure 6 visualisations are in `09_analysis/Figure6_bloominess_robustness.Rmd`.

Threshold grid

Bloominess was recomputed across a grid of four coverage thresholds while holding the detection threshold fixed at 0.2:

Coverage threshold	Label
0.1×	cov0.1_det0.2
0.25×	cov0.25_det0.2
0.5×	cov0.5_det0.2 (baseline)
1.0×	cov1.0_det0.2

For each threshold, the full bloominess pipeline was re-run: the presence flag was recomputed, and mean, median, and bloominess values were recalculated using only samples passing the updated threshold. MAGs classified by simplified prevalence groups for this analysis (core 85% prevalence; rare 10%; variable: all others) without the additional coverage floor, enabling classification to remain consistent across all four threshold settings regardless of absolute coverage level.

Figure 6 panels

Panel A — Rank correlation: Bloominess ranks at the baseline threshold (cov0.5_det0.2) were compared against ranks at the strictest threshold (cov1.0_det0.2). Each MAG is shown as a point coloured by prevalence class and sized by number of samples present. Strong concordance between ranks across the two thresholds indicates that the relative ordering of MAGs by bloominess is robust to threshold choice.

Panel B — Rank instability vs support: For each MAG, the standard deviation of bloominess across all four threshold settings was computed and plotted against the mean number of samples in which the MAG was present across thresholds. A LOESS smoother was overlaid. This panel tests whether low-support MAGs (present in few samples) show greater sensitivity to threshold choice than well-supported MAGs.

Panel C — Ecological space preservation: The prevalence × bloominess landscape was plotted for all four threshold settings on the same axes, with a LOESS smoother per threshold. Overlap between the smoothed curves indicates that the overall ecological structure (the relationship between prevalence and bloominess) is preserved regardless of threshold.

References

- Eren, A. M. et al. (2021). Community-led, integrated, reproducible multi-omics with anvi'o. *Nature Microbiology*, 6, 3–6. https://doi.org/10.1038/s41564-020-00834-3
- Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18), 3094–3100. https://doi.org/10.1093/bioinformatics/bty191