Supplementary Method S6: Drug class composition, bloominess, and ARG co-occurrence — visualisation and statistical framework

---

1. Overview and rationale

To investigate whether different classes of antimicrobial resistance genes (ARGs) are associated with different patterns of host-organism bloominess and ARG co-occurrence, we developed a two-part analysis: (1) a drug class proportion heatmap visualising the distribution of ARG loci across a two-dimensional grid of bloominess tertile × total ARG burden, and (2) a chi-squared goodness-of-fit test quantifying whether each drug class is distributed non-randomly across that grid.

This approach was applied to variable-class MAGs only (prevalence $> 10\%$ and $< 85\%$). Rare MAGs were excluded because their very low prevalence means bloominess estimates are based on few observations and are less reliable. Core MAGs were excluded because their near-universal presence means they do not vary meaningfully across hosts in a way that allows meaningful co-occurrence interpretation.

---

2. ARG locus definition

ARG annotations were obtained from AMRFinder (see main Methods). Because AMRFinder can produce multiple hits within the same resistance operon or mobile cassette — each annotated as a separate gene — raw hit counts overestimate independent ARG acquisition events.

To correct for this, individual ARG hits were collapsed into **ARG loci** using a 10 kb proximity window. Specifically:

1. For each MAG, all ARG hits were sorted by their midpoint position (`mid = (start + stop) / 2`) within each contig.
2. Hits were assigned to the same locus if consecutive hits on the same contig had midpoints within 10,000 bp of each other.
3. The total number of distinct loci per MAG was used as the measure of ARG burden throughout.

The 10 kb window was chosen to span typical operons and resistance cassettes while separating independently acquired elements on the same contig.

---

3. The drug class proportion heatmap

3.1 Construction

For each variable-class MAG with at least one ARG locus, the following two properties were computed:

- **Bloominess tertile (bloom_bin):** Bloominess values (see Supplementary Method S5) were divided into three equally-sized groups (tertiles) using `ntile()` across all variable MAGs with non-missing bloominess. Groups were labelled Low, Medium, and High bloominess. Tertiles were computed within the variable-class population only, so the labels are relative to this group.
- **Total ARG locus bin (arg_bin):** The total number of ARG loci per MAG (all drug classes combined) was binned as: 0, 1, 2, 3, 4+. This defines the x-axis of the heatmap.

These two variables define a grid of 3 × 5 cells (15 cells total, though the arg_bin = 0 column contains no ARG loci by definition and is retained only for MAG context).

ARG loci were then assigned a drug class from the `class` field of the AMRFinder output. For each drug class separately, the **proportion of all ARG loci in each grid cell belonging to that class** was computed:

```
prop[bloom_bin i, arg_bin j, drug_class k]
    = n_loci[i, j, k] / sum_k(n_loci[i, j, k])
```

where `n_loci[i, j, k]` is the count of loci of class k in cell (i, j), and the denominator is the total loci of all classes in that cell.

This proportion is the fill value in the heatmap. The top 10 most abundant drug classes were retained; all others were collapsed into "Other".

3.2 What each cell represents

A single cell, for example (bloom_bin = High, arg_bin = 1, drug_class = BETA-LACTAM), answers the question:

> "Of all ARG loci found in high-bloominess MAGs that carry exactly one total ARG locus, what fraction are beta-lactam loci?"

If this proportion is high, beta-lactam tends to be the sole ARG in high-bloominess MAGs. If this proportion is low or absent, beta-lactam is rare in that bloom × burden context.

3.3 Why total ARG loci (not class-specific loci) is used on the x-axis

The x-axis intentionally represents **total ARG burden across all classes**, not the number of loci for the specific drug class being shown in each facet.

This is the key design choice, and it gives the visualisation two interpretable dimensions simultaneously:

- **Left-to-right (co-occurrence):** If a drug class's fill is concentrated in the left columns (arg_bin = 1), MAGs carrying that class tend to have only one ARG locus total — this class occurs *alone*. If the fill is concentrated in the right columns (arg_bin = 3, 4+), MAGs carrying that class tend to have many total ARGs — this class *co-occurs* with other resistance determinants.
- **Top-to-bottom (bloominess association):** If a drug class's fill is concentrated in the upper rows (High bloominess), that class tends to be found in more bloomy MAGs. If concentrated in the lower rows (Low bloominess), the class is disproportionately found in less-bloomy organisms.

These two dimensions are independent enough to reveal distinct biological patterns simultaneously, without requiring separate analyses.

3.4 Unit of observation and pseudoreplication

The fill values in the heatmap are computed at the **locus level**: each ARG locus contributes one count to one cell. A MAG with five beta-lactam loci contributes five locus-counts to a single cell (since all loci from one MAG share the same bloom_bin and arg_bin). This locus-level representation was used for the visualisation because it reflects the composition of the resistome pool, and because it naturally captures the proportion of resistance attributable to each class.

However, locus-level counts are not independent observations within a MAG and cannot be used directly for statistical inference without pseudoreplication bias. For statistical testing, the unit of observation was therefore switched to **carrier MAGs** (see Section 4).

---

4. Chi-squared goodness-of-fit test

4.1 Rationale

The chi-squared goodness-of-fit test was applied to quantify whether each drug class is distributed non-randomly across the 12 non-zero cells of the bloom_bin × arg_bin grid (three bloominess tertiles × four ARG burden bins: 1, 2, 3, 4+).

4.2 Unit of observation

For the statistical test, the unit was a **carrier MAG**: a variable-class MAG carrying at least one locus of the drug class being tested. Each carrier MAG contributes one count to one cell, based on its bloom_bin and arg_bin (total ARG burden). This eliminates the pseudoreplication issue inherent in locus-level counts.

4.3 Null hypothesis

The null hypothesis is that carrier MAGs of drug class k are distributed across the bloom_bin × arg_bin grid **in exact proportion to the overall density of variable-class MAGs per cell**:

```
Expected[i, j, k] = n_carriers[k] × (n_variable_MAGs[i, j] / total_variable_MAGs_with_ARGs)
```

where `n_variable_MAGs[i, j]` is the total number of variable MAGs in cell (i, j) (from the 5C heatmap, restricted to arg_bin > 0), and `n_carriers[k]` is the total number of carrier MAGs of drug class k among variable MAGs.

Under this null, a drug class with no particular preference for any bloom or ARG burden context would spread proportionally across the grid. Deviations from this expected distribution indicate that the drug class is disproportionately found in certain quadrants — either certain bloominess levels, certain ARG burden levels, or both.

4.4 Observed counts and test statistic

For each drug class, a vector of 12 observed counts was assembled (one per non-zero cell, ordered by bloom_bin then arg_bin). The corresponding expected count vector was computed from the null distribution above. The chi-squared statistic was computed as:

```
X² = Σ (observed - expected)² / expected
```

with degrees of freedom = (number of non-zero cells) $- 1 = 11$.

4.5 Monte Carlo simulation for sparse cells

Several drug classes have small numbers of carrier MAGs (as few as 9–16), which produces expected cell counts well below the conventional threshold of 5 required for the theoretical chi-squared distribution to be reliable. To account for this, empirical p-values were generated by **Monte Carlo simulation** (B = 9,999 replicates). In each replicate, the observed carrier MAGs were redistributed across cells by random multinomial sampling under the null probability vector, and the chi-squared statistic was recomputed. The empirical p-value is the proportion of simulated statistics exceeding the observed statistic.

This approach makes no assumption about the sampling distribution of the test statistic and is valid regardless of cell sparsity.

4.6 Multiple testing correction

Ten drug classes were tested independently (the top 10 most abundant in variable-class MAGs). To control the false discovery rate, p-values were adjusted using the **Benjamini-Hochberg procedure**. Results are reported as both raw simulated p-values and BH-adjusted p-values. Significance thresholds used: *** p_adj < 0.001, ** p_adj < 0.01, * p_adj < 0.05.

4.7 Standardised residuals

To identify which cells drive a significant result, **standardised residuals** were computed for each cell:

```
std_resid[i, j] = (observed[i, j] - expected[i, j]) / sqrt(expected[i, j])
```

Positive residuals indicate cells where a drug class is over-represented relative to null (more carrier MAGs than expected); negative residuals indicate under-representation. Residuals with |std_resid| > 2 are conventionally considered noteworthy.

4.8 Implementation

The full analysis is implemented in `09_analysis/Figure5/run_figure5_chisq.R`. Results are saved to `09_analysis/Figure5/FigS_drug_class_chisq_results.csv`.

---

5. Interpretation framework

5.1 Co-occurrence tendency

A drug class enriched in high arg_bin cells (right columns of the heatmap) is found predominantly in MAGs with high total ARG burden — meaning those MAGs also carry resistance determinants from other classes. This indicates that the drug class tends to co-occur with other resistance mechanisms, whether by co-selection on the same mobile element or by accumulation in MAGs from high-resistance contexts.

A drug class enriched in low arg_bin cells (left columns, particularly arg_bin = 1) tends to occur as the sole resistance mechanism in those MAGs, suggesting independent acquisition or maintenance without co-selection.

5.2 Bloominess association

A drug class enriched in the High bloominess row tends to be found in organisms that undergo episodic abundance spikes. A drug class enriched in Low bloominess rows tends to be found in organisms with more stable, lower-amplitude abundance profiles.

These associations are correlational. Two biological explanations are consistent with the data:

- **Fitness cost hypothesis:** Carriage of certain ARGs imposes a metabolic or structural cost that reduces competitive fitness, leading to lower bloom capacity. Under this model, the ARG itself modulates the MAG's bloom behaviour.
- **Taxonomic association:** Certain bacterial lineages may be both naturally bloomy and prone to acquiring specific ARG classes, independently. Under this model, the host taxon determines both bloominess and ARG class, and the observed association is mediated by taxonomy.

5.3 Risk interpretation and the role of taxonomy

Critically, the risk interpretation holds under **both** explanations above. From a resistome risk perspective, the key question is not *why* a drug class tends to occur in bloomy organisms, but *whether* it does. An ARG class found predominantly in bloomy taxa has greater potential for episodic amplification: even if it is rare at baseline, the host organism's bloom behaviour means resistance abundance can spike dramatically in a subset of hosts, flocks, or environmental contexts.

Bloominess therefore captures a dimension of ARG risk that prevalence and mean abundance cannot: the potential for dynamic, transient bursts of high resistance abundance that would be missed by cross-sectional or time-averaged surveillance. Whether the association between a drug class and bloominess reflects a fitness cost, a taxonomic co-association, or both, the practical implication is the same.

References

- Feldgarden, M. et al. (2021). AMRFinderPlus and the Reference Gene Catalog facilitate examination of the genomic links among antimicrobial resistance, stress response, and virulence. *Scientific Reports*, 11, 12728. https://doi.org/10.1038/s41598-021-91456-0
- Benjamini, Y. & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, 57(1), 289–300.

- Hope, A. C. A. (1968). A simplified Monte Carlo significance test procedure. *Journal of the Royal Statistical Society, Series B*, 30(3), 582–598.