G3, 2026, 16(11), jkag249
https://doi.org/10.1093/g3journal/jkag249
Advance Access Publication Date: 2 September 2026
Investigation

Signal, noise, and sampling: how pool size and replication shape metabolomic inference

David L. Hubert ,* Dylan L. Porter, Ryan D. Robinson , Maya E. Mijares , Elmira Ahmadian , Kenneth R. Arnold , Mark A. Phillips

Department of Integrative Biology, Oregon State University, Corvallis, OR 97331, United States

*Corresponding author: Department of Integrative Biology, Oregon State University, Corvallis, OR 97331, United States. Email: hubertd@oregonstate.edu

Metabolomics provides direct insight into physiological state, but for small organisms such as *Drosophila melanogaster,* it typically requires pooling individuals to obtain sufficient material. Pool sizes vary widely across studies with little justification, and the impact pooling and biological replication have on metabolomic characterization and signal detection remains poorly understood. We evaluated the effects of pool size and biological replication on metabolomic profiles and signal detection using 2 complementary designs in *D. melanogaster*. First, we tested how pooling (5, 50, or 100 individuals) affects metabolomic structure and reproducibility in inbred and outbred populations. Second, we tested how pool size interacts with replicate number to affect detection of diet-associated metabolite changes under a high-sugar perturbation. Pool size-shaped metabolomic profiles: pools of 5 individuals consistently differed from larger pools, which improved reproducibility in a dataset-dependent manner. In the dietary experiment, smaller pools showed reduced sensitivity, detecting fewer true diet-associated metabolites without increasing false discoveries, and replicate downsampling showed that pool size and replication independently shape signal retention. Detection depended on effect size and variability: metabolites with larger, more stable effects were consistently retained, while smaller more variable effects were rapidly lost under reduced sampling. Beta-binomial modeling confirmed that detection probability reflects a balance between signal strength and measurement variability, with pool size and replicate number independently shaping this relationship. Together, these results show that metabolomic inference depends on the interplay of signal, noise, and sampling design, with pool size and replication jointly shaping the detectability, stability, and interpretation of biological signals.

Keywords: metabolomics; sampling design; pool size; biological replication; signal detection

Introduction

Metabolomics, the comprehensive analysis of small-molecule metabolites in biological systems, offers a functional readout of cellular physiology (Johnson et al. 2016). The metabolome serves as a bridge between genotype and phenotype, reflecting the combined influence of genetic variation, environmental conditions, and organismal state (Nicholson et al. 1999). Because metabolites act as building blocks, enzyme cofactors, and mediators of energy production and signaling, metabolomic profiling provides a direct window into the molecular processes underlying biological function. Advances in high-resolution mass spectrometry and computational analysis have made it possible to characterize metabolite profiles at increasingly large scales, expanding the use of metabolomics across physiology, disease biology, and evolutionary studies (Tautenhahn et al. 2012; Clish 2015).

Drosophila melanogaster has become a central model for this work due to its short generation time, tractable genetics, and long history in experimental biology. Integrating metabolomics with genetic and systems-level approaches has revealed substantial variation in metabolic architecture across lines and populations, demonstrating how genetic background shapes physiological traits (Ayroles et al. 2009; Mackay et al. 2012; Zhou et al. 2020). Metabolomic profiling has also been incorporated into experimental evolution studies, where replicated populations evolving under defined conditions allow direct links to be drawn between selection, environment, and physiological change (Phillips et al. 2022; Erkosar et al. 2023; Hubert et al. 2025). However, despite the increasingly common use of metabolomics in Drosophila studies, a basic aspect of experimental design remains highly variable across studies: the number of individuals pooled per sample.

As individual flies yield limited material, pooled sampling is standard practice. However, pool sizes differ widely in the literature, ranging from only a few individuals to several dozen or more (eg Zhao et al. 2022; Hubert et al. 2025; Vecchié et al. 2026), often without clear justification. This variation introduces an underappreciated source of heterogeneity. Smaller pools may be more sensitive to stochastic differences among individuals or may fail to capture low-abundance metabolites, whereas larger pools may provide more stable estimates but at increased cost and reduced throughput. As metabolomics becomes more central to comparative and evolutionary work, understanding how pooling decisions influence both metabolome characterization and the detection of biological signal is increasingly important.

Here, we evaluate how pooled sample size affects metabolomic characterization and signal detection across 2 complementary experimental contexts in *D. melanogaster*. In the first, we tested the effect of pool size (5, 50, or 100 individuals) in both inbred and outbred populations sampled at 2 adult ages. Here, outbred populations

Received on 27 July 2026; accepted on 26 August 2026
© The Author(s) 2026. Published by Oxford University Press on behalf of The Genetics Society of America.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

capture substantial standing genetic variation, whereas inbred populations provide a reduced-variance baseline, allowing us to evaluate how genetic heterogeneity interacts with sampling scale. This design allows us to assess how genetic background influences the degree to which larger pools improve the stability and consistency of metabolomic measurements.

In a second experiment, we focused on inbred populations to test how pool size interacts with replication and treatment effects under a defined dietary perturbation. Replicate populations were maintained on either control or high-sugar diets, and metabolomic profiles were again generated across a gradient of pool sizes within each replicate population. We then systematically downsampled the number of replicates and evaluated metabolomic signal retention across pool sizes at each iteration. This design allows us to evaluate how both pooling and biological replication influence the ability to detect consistent metabolomic differences associated with diet treatment, and to quantify the trade-offs between increasing pool size vs increasing the number of replicates. Together, these experiments address a practical but largely unresolved question in metabolomics: how sampling decisions shape the stability of metabolomic measurements and the power to detect biologically meaningful signals.

Methods

Study system

For this experiment, 2 distinct *Drosophila melanogaster* populations representing different genetic backgrounds were used: inbred and outbred flies. In Experiment 1, the outbred populations were maintained on standard food and used as a laboratory outbred population for approximately 35 generations at the time of this work (CRB). These populations were created by combining 10 females and 5 males from 100 lines from the Drosophila Genetic Reference panel (Mackay et al. 2012; Supplementary Table 1). These lines were derived from a single natural population in Raleigh, North Carolina. Combining lines effectively creates outbred populations that, at least in part, capture the variation of this source natural population. The inbred population was a OR wild-type (ORWT) strain obtained from Carolina Biological Supply Company (Burlington, NC, USA). In Experiment 1, both populations were maintained on a 28 d life cycle, where eggs were collected, transferred to vials, and incubated for 14 d, then freshly eclosed adult flies were placed into cages for 14 d before the next egg collection was conducted. All populations were provided with identical standard food replaced every 2 d, and were kept under standard laboratory conditions. To minimize environmental variation, cage positions on the shelf were rotated whenever food was replaced to account for any differences in light or temperature.

In Experiment 2, to test how pool size interacts with biological replication and diet treatment to influence metabolomic profiles in an inbred background, we performed a dietary perturbation experiment using the ORWT inbred *Drosophila melanogaster* strain. Replicate ORWT populations were founded from the laboratory stock and maintained under 2 dietary conditions: standard food containing approximately 7% molasses (STD) and high-sugar food containing 20% molasses (HSD). Eight independent replicate populations were maintained per diet treatment. Each replicate population comprised 40 vials containing 50 to 80 adult flies per vial. Flies were collected on their assigned treatment and kept in the same vials in the incubator until sampling to minimize microenvironmental variation. Incubator conditions were 23 °C, 12 h light: 12 h dark.

For clarity, we distinguish 3 levels of sampling and replication used throughout this study. A biological replicate population refers to an independently founded and maintained population (eg a separate cage or vial set derived from a common source population or genetic background), representing an independent instance of population-level biological variation; Experiment 1 used 3 replicate populations per genetic background, and Experiment 2 used 8 replicate populations per diet treatment. Pool size (or pooled sample) refers to the number of individual flies (5, 50, or 100) physically combined into a single sample prior to homogenization and metabolomic extraction; pooling aggregates individual-level variation within a single collection event and is not itself a form of replication. A technical replicate refers to repeated processing or measurement of the same biological material to assess analytical or instrument-level variability; in this study, technical variability was monitored using a pooled quality-control (QC) sample analyzed throughout each batch (see Metabolomic Characterization and Data Processing), rather than through repeated measurement of individual experimental samples. Unless otherwise noted, the experimental unit for all statistical analyses is the individual pooled sample (one sample per pool size, within each replicate population, within each genetic background or diet treatment, and, in Experiment 1, age), and biological replicate population identity was included as a random effect or stratification variable where appropriate to account for non-independence among samples drawn from the same replicate population.

Sample collection

For Experiment 1, a single population from each genetic background (CRB and ORWT) was expanded into 3 independent biological replicate populations. These replicate populations were allowed to develop and reproduce, and eggs collected from each were used to establish experimental cohorts. Samples for metabolomic analysis were collected at 2 time points, 14 and 40 d post-egg deposition, representing distinct young and old adult stages associated with known age-related metabolomic differences (Hoffman et al. 2014; Zhao et al. 2022). For day 14 sampling, adult females were collected directly from vials, after which remaining flies were transferred to cages (~1,000 individuals per population) and maintained until day 40, when the second sampling was performed. Food was available up until sample collection in all cases. For each replicate population, samples were collected at the same time of day at 3 pool sizes (5, 50, and 100 individuals), selected to span the range commonly used in metabolomic studies and to maximize potential differences in detection (3 replicates × 2 genetic backgrounds × 3 pool sizes × 2 ages = 36 samples).

For Experiment 2, to assess how pool size interacts with replication and diet treatment, replicate ORWT populations were maintained under standard (STD) or high-sugar (HSD) diets (8 replicates per diet). Adult females were sampled 1 to 2 d post-eclosion for each replicate, standardizing developmental stage despite differences in development time between treatments. Samples were collected at the same time of day at pool sizes of 5, 50, and 100 individuals (2 diets × 8 replicates × 3 pool sizes = 48 samples), and food was available up until collection.

For both experiments, flies were anesthetized on ice, transferred to a chilled ceramic plate, and sexed. Only females were used to eliminate sex-specific metabolic differences (Millington and Rideout 2018). Individuals were pooled into the designated sample sizes, flash-frozen in liquid nitrogen, and stored at −80 °C until metabolomic analysis.

Metabolomic characterization and data processing

For untargeted metabolomics analyses, high-resolution mass spectrometry (AB Sciex 5600 tripleTOF) was used. The whole fly

samples were homogenized using a bead beater and extracted with a mixture of ice-cold methanol in water (1 mL, 4:1, v/v), incubated at −20 °C for 1 h, then the homogenates centrifuged at 15,000 × *g* at 4 °C for 10 min. The 800 µL of supernatant was dried by speed-vac and then reconstituted with 200 µL of acetonitrile: water (1:1, v/v). The mixture was centrifuged at 15,000 × *g* at 4 °C for 10 min then supernatant was transferred into LC vials. QC sample was pooled from all samples to monitor analytical variability and to correct for any signal drift across batches of samples in untargeted metabolomics. Samples were analyzed by ultra-performance liquid chromatography coupled with electrospray quadrupole-time of flight (UPLC-QToF) mass spectrometry in the positive and negative ion modes with C18 and HILIC method using an Applied Biosystems Sciex 5600 instrument (AB Sciex, Framingham, MA, USA). Metabolites were identified based on accurate mass, MS/MS spectra, isotope pattern, and retention time using PeakView and MultiQuant software (AB Sciex) compared with metabolites in the facility's metabolite IROA library (650 standards). When standards for comparison were not available, online databases, METLIN, HMDB, for metabolite identification were used (Choi et al. 2019).

Normalization was conducted independently for each experimental dataset. Relative concentration values for metabolites were log10-transformed to approximate a Gaussian distribution. Prior to normalization, metabolites with high technical variability were removed based on pooled QC samples (coefficient of variation > 0.30). Data were then mean-centered within each analytical run using one of 2 approaches depending on the downstream analysis: (i) mean-centering by metabolite across samples to normalize metabolite-specific signal intensity differences while preserving variation among samples or (ii) mean-centering within samples to normalize overall signal intensity differences while preserving relative metabolite composition within each sample. Because extraction volumes were fixed across pool sizes (see above), differences in pool size translate into differences in extract concentration rather than extraction volume; mean-centering removes this intensity-scale effect prior to all downstream similarity and signal-detection analyses, so that reported pool-size effects reflect differences in metabolomic profile structure rather than raw signal intensity.

Metabolomic similarity and profile structure

To evaluate the effect of pool size on overall metabolomic composition, a common similarity analysis framework was applied to both experiments, with model terms adjusted to reflect the experimental design of each dataset. All analyses were performed separately for each LC–MS panel.

Three complementary approaches were used to assess similarity. First, principal component analysis (PCA) was performed using the prcomp function in R to visualize clustering of samples by pool size in multivariate metabolomic space. Prior to PCA, metabolite values were mean-centered across samples so that principal components capture variation among samples rather than differences in baseline metabolite abundance. Additional scaling was not applied, as metabolites were already on a comparable scale following log-transformation and normalization (Hubert et al. 2025).

Second, pairwise Euclidean distances were calculated in the full metabolomic space between all pool-size pairs (5 to 50, 5 to 100, and 50 to 100). Distances were computed within each relevant biological grouping (see below), yielding one distance value per pool-size pair per replicate. These distances quantify the magnitude of metabolomic divergence between pool sizes, where a large 5 to 50 distance relative to a small 50 to 100 distance indicates that most profile change occurs at the first increase in pool size. Given the limited number of replicates, these comparisons are treated as descriptive and complement the formal statistical analyses.

Third, permutational multivariate analysis of variance (PERMANOVA) was performed using pseudo-F statistic implemented in the adonis2 function from the *vegan* package (Oksanen 2022) to test whether samples with different pool sizes have systematically different metabolomic profiles, rather than just differing by chance. Euclidean distance matrices were calculated from the normalized metabolite data for each panel. Models were fitted using marginal (Type III) tests so that the effect of pool size was evaluated after accounting for other experimental factors. Permutations ($n = 999$) were stratified within replicate populations using the how function from the *permute* package (Simpson 2022) to account for non-independence of samples within replicates, and a fixed random seed was used for reproducibility. The primary output was R^2, representing the proportion of total variance explained by each factor.

Finally, multivariate dispersion was assessed using the betadisper function from the *vegan* package to evaluate the reproducibility of metabolomic profiles across pool sizes. This test asks whether metabolomic measurements become more consistent, and less spread out, as pool size increases. Within each grouping, samples were assigned to pool-size categories, and the distance of each sample to its group centroid in multivariate space was calculated. A decrease in mean distance to centroid with increasing pool size indicates improved reproducibility. Differences in dispersion were tested using permutation tests (permutest, 999 permutations), with replicate population included as a blocking factor.

Experiment-specific model structures

For **Experiment 1: Effects of Pool Size, Genetic Background, and Age on Metabolomic Structure**, analyses were conducted within each strain × age × replicate combination for distance calculations and dispersion analyses. PERMANOVA models included pool size, strain, and age as fixed effects, allowing the contribution of pool size to be evaluated after accounting for genetic background and age-related differences. PCA was performed separately for each strain to avoid confounding pool-size effects with between-strain variation. To characterize age-related changes at the level of individual metabolites, a linear mixed-effects model was fit separately for each metabolite in each panel:

$$y_{ijk} = \beta_0 + \beta_1 Age_j + \beta_2 \log PoolSize_k + \beta_3 Strain + b_i + \varepsilon_{ijk}$$
$$b_i \sim N(0, \sigma_b^2)$$

where Age is a 2-level categorical predictor (14 vs 40 d, with day 14 as the reference level, so that a positive coefficient indicates an increase and a negative coefficient a decrease with age), Strain is included as a fixed effect to account for the genetic background difference between the CRB and ORWT populations, and log-transformed pool size is included as a covariate for consistency with the pool-size model used elsewhere in this study. A random intercept for each Strain by Population replicate accounts for repeated sampling of the same replicate population across ages and pool sizes. Models were fit in R using the "lme4" and "lmerTest" packages, with *t*-statistics and *P*-values based on Satterthwaite's approximation. *P*-values for the Age coefficient were corrected for multiple testing within each panel using the Benjamini–Hochberg procedure (Benjamini and Hochberg 1995), and metabolites with false discovery rate (FDR) < 0.05 were

considered to show a significant age effect.

For **Experiment 2: Effects of Pool Size and Replication on Dietary Metabolomic Signal Detection**, analyses were conducted within each diet × replicate combination for distance calculations and dispersion analyses. PERMANOVA models included pool size and diet as fixed effects, reflecting the absence of strain and age variation in this experiment. PCA was performed across all samples within each panel, as genetic background was held constant.

Pool size analysis

To assess whether pool size affects the ability to detect genuine diet-related metabolomic changes, a per-metabolite signal detection analysis was performed using our 4 LC-MS panels. Input data were log-transformed metabolite values mean-centered across samples within each metabolite (as described above). For each metabolite in each panel, a linear model was fitted within each pool size:

$$y_{ij} = \beta_0 + \beta_1 \, \mathrm{Diet}_j + \varepsilon_{ij}$$

where y_{ij} is the mean-centered, log-transformed abundance of metabolite i in sample j, Diet_j is a 2-level categorical predictor (HSD vs STD), and ε_{ij} is the residual error term. Models were fitted separately within each pool size ($n = 5$, $n = 50$, $n = 100$), using all 16 observations at that pool size (8 replicate populations × 2 diet groups). Running separate models per pool size ensures that results at each pool size reflect only the data collected at that pool size, allowing direct comparison of what a researcher would conclude from each pool size independently. To correct for multiple comparisons, *P*-values for the Diet coefficient were adjusted using the Benjamini–Hochberg FDR procedure applied within each pool size across all metabolites within a given panel. A metabolite was considered to have a significant diet effect at a given pool size if its FDR-adjusted *P*-value was below 0.05.

Results at $n = 100$ were treated as a reference condition, as this pool size provides the most stable and reproducible metabolomic profiles based on the similarity analysis. We note that this reference condition is not an independently validated ground truth; it simply represents the condition with the greatest input material and, based on the similarity analysis, the highest expected statistical power within this experiment. Accordingly, metabolites classified as false positives were not necessarily spurious, and metabolites classified as false negatives at $n = 100$ do not necessarily represent genuine diet effects, only effects that were not detected under the reference condition. For each metabolite, the diet effect at $n = 5$ and $n = 50$ was classified relative to $n = 100$ using 4 categories: a true positive (TP) if the metabolite was significant at both pool sizes in the same direction; a false negative (FN) if the metabolite was significant at $n = 100$ but not at the smaller pool size; a false positive (FP) if the metabolite was significant at the smaller pool size but not at $n = 100$; and a true negative (TN) if the metabolite was not significant at either pool size. From these counts, 2 summary statistics were derived: sensitivity (TP/[TP + FN]), representing the proportion of reference-condition diet-significant metabolites recovered at a smaller pool size; and false discovery rate (FP/[FP + TP]), representing the proportion of detected effects that are not supported by the $n = 100$ reference.

Next, to assess whether pool size affects the reliability of individual effect size estimates, the diet coefficient estimated at $n = 5$ and $n = 50$ was compared against the corresponding estimate at $n = 100$ for all metabolites. Pearson correlation between effect sizes at each pool size and at $n = 100$ was computed separately for all metabolites and for the subset significant at $n = 100$, to assess whether the reliability of effect size estimates is consistent across the full metabolome.

Lastly, we quantified the sensitivity of individual metabolites to pool size by fitting a linear mixed-effects model across all pool sizes:

$$y_{ijk} = \beta_0 + \beta_1 \log(\mathrm{PoolSize}_k) + \beta_2 \mathrm{Diet}_j + b_i + \varepsilon_{ijk}$$

where y_{ijk} is the mean-centered, log-transformed metabolite abundance, $\log(\mathrm{PoolSize}_k)$ is the natural log of pool size, and Diet_j is a fixed effect (HSD vs STD). The term $b_i \sim \mathcal{N}(0, \sigma_b^2)$ represents a random intercept for replicate population. Pool size was modeled on the log scale because sampling effort increases nonlinearly across the discrete levels used ($n = 5$, 50, 100), and we were specifically interested in capturing diminishing returns in precision with increasing pool size. Models were fit in R using the "lme4" package, with t-statistics and *P*-values obtained using "lmerTest" (Bates et al. 2015). We quantified pool-size sensitivity as the absolute value of the *t*-statistic associated with log(PoolSize). Multiple testing correction was performed using the Benjamini–Hochberg procedure within each panel. Sensitivity here reflects how many of the truly detectable diet effects a smaller pool size would find, while false discovery rate reflects how often a detected effect at a smaller pool size fails to hold up at the larger, more reliable pool size.

Downsampling and pool size

Replicate downsampling was used to assess the robustness of diet-associated metabolite signals to reduced biological replication. Analyses were performed separately for each normalized metabolomics dataset (C18 positive, C18 negative, HILIC positive, and HILIC negative) using data mean-centered by metabolite. Replicate populations were removed in a balanced manner across diets, with 0 to 5 replicate populations removed per diet. Because the number of possible replicate-removal combinations varies enormously across removal levels (from 1 to several thousand), a capped random-resampling design was used: at removal levels where the full combinatorial space exceeded 64 combinations, 64 combinations were sampled at random without replacement; at levels with 64 or fewer possible combinations, all were evaluated. This produced a uniform number of iterations (64) across downsampling levels, avoiding the highly unequal sampling precision that exhaustive enumeration would otherwise introduce across removal levels (Supplementary Table 2). For each downsampled dataset and each pool size (5, 50, and 100), metabolite abundance was modeled as a function of diet using simple linear models (metabolite ~ Diet). Diet effect sizes, test statistics, and *P*-values were extracted for each metabolite, and *P*-values were adjusted for multiple testing using the Benjamini–Hochberg FDR procedure within each dataset × pool size × iteration combination. Metabolites with FDR < 0.05 were considered significant. Results were summarized across iterations to quantify changes in significant metabolite counts, true- and false-positive recovery relative to the reference condition (pool size = 100, full replication), concordance of diet effect sizes across pool sizes at full replication, and retention of significant metabolites across downsampling levels for metabolites grouped into high-, medium-, and low-effect bins based on reference absolute effect size.

Functional signal retention

To evaluate the retention of interpretable biological signal under replicate downsampling and varying pool size, metabolites were

assigned to functional modules based on biochemical identity and established metabolic pathway context using curated classification rules informed by central metabolic processes (Xia and Wishart 2010; Kanehisa et al. 2012; Supplementary Table 3). Modules represented major metabolic processes, and metabolites not assigned to a defined module were excluded from module-level analyses. A reference set of diet-associated metabolites was defined from the full dataset (pool size = 100, maximum replication) using an FDR threshold of 0.05. Module-level signal preservation was quantified by calculating, for each downsampling condition, the proportion of reference metabolites that remained significant across iterations. Recovery frequency was first calculated at the metabolite level and then averaged within modules to obtain the percentage of metabolites recovered per module. Results were summarized across replicate downsampling levels and pool sizes and visualized as heatmaps to assess the robustness of functional metabolic signals.

Determinants of metabolite signal retention

To identify the factors underlying variation in metabolite signal retention across downsampling conditions, we modeled the probability that individual metabolites remained significant as a function of effect size, variability, pool size, and replication. For each metabolite within each dataset, pool size, and replicate downsampling condition, signal retention was quantified as the proportion of downsampling iterations in which the metabolite was detected as significant (FDR < 0.05).

Reference effect size for each metabolite was defined as the absolute value of the diet coefficient estimated under the reference condition (pool size = 100, maximum replication). To quantify variability in metabolite estimates across downsampling, we calculated the standard deviation of the diet coefficient across all iterations within each dataset × pool size × replicate condition. This measure captures the stability of effect size estimates under reduced sampling.

Metabolite-level signal retention was modeled using a beta-binomial generalized linear mixed model (GLMM). The response was specified as the number of iterations in which a metabolite was detected as significant vs the number in which it was not, within each dataset × pool size × replicate-number combination. Fixed effects included reference effect size, variability (standard deviation of the diet coefficient), pool size (categorical: 5, 50, 100, with 100 as the reference level), replicate number (3 to 7 replicates remaining; the full-replication reference condition was excluded from the regression), and the interaction between pool size and replicate number. In practical terms, this model asks which factors, effect size, estimate variability, pool size, or replicate number, make a metabolite more or less likely to be reliably detected as significant. Dataset and metabolite identity were included as random intercepts to account for non-independence among measurements from the same dataset and inherent differences among metabolites, and the beta-binomial dispersion parameter was allowed to vary by dataset. Models were fit using the glmmTMB package in R with a logit link. Model fit was assessed using simulation-based residual diagnostics (DHARMa package), including tests for dispersion and zero-inflation. The significance of the pool size × replicate number interaction was evaluated using a likelihood ratio test comparing the full model to an additive model excluding the interaction term.

Quantitative guidance for pool size vs replication

To provide quantitative guidance for balancing pool size against biological replication under realistic resource constraints, we modeled the total number of diet-significant metabolites detected per downsampling iteration (summed across all metabolites within a dataset) as a function of replicate number, pool size, and their interaction, with a random intercept for dataset. Because this count is bounded below by zero and a subset of iterations detected no significant metabolites at all, more than expected under a standard Poisson model, we used a zero-inflated Poisson GLMM (glmmTMB) with a zero-inflation submodel on pool size. In practical terms, this tells us how much benefit an additional biological replicate provides at each pool size, relative to the number of diet-associated metabolites detected. Because this model evaluates the total yield of significant metabolites per iteration rather than the retention of any single metabolite, its interaction term addresses a distinct question from the one tested in the per-metabolite retention model above and need not reach the same conclusion. Model fit was compared against a standard Poisson GLMM lacking the zero-inflation term using a likelihood ratio test, and adequacy of the final model was assessed using simulation-based residual diagnostics (DHARMa package), including tests for dispersion and zero-inflation. The significance of the replicate number × pool size interaction was similarly evaluated using a likelihood ratio test comparing the full model to an additive model excluding the interaction term.

Use of AI tools

Generative AI tools, ChatGPT (OpenAI, GPT-5.3) and Claude (Anthropic, Claude Sonnet 5), were used to assist with text editing, code debugging, and data analysis workflows during manuscript preparation. In all cases, all AI-generated output was reviewed, verified, and edited by the authors, who take full responsibility for the accuracy and content of the final manuscript.

Results

Experiment 1: Effects of pool size, genetic background, and age on metabolomic structure

LC-MS and QC

A total of 219 metabolites were detected across the 4 LC-MS panels, of which 213 passed quality-control filtering (Table 1). The C18 negative panel was excluded from Experiment 1 analyses due to an insufficient number of retained metabolites ($n = 2$), leaving 211 metabolites used for downstream analyses. Panel size varied considerably among the remaining 3 panels, from 19 metabolites in C18 positive to 142 in HILIC negative; this difference is relevant to interpreting panel-specific results, as analyses based on the smaller C18-positive panel are more sensitive to individual metabolite-level variability and carry a lower multiple-testing burden than the larger panels.

Metabolomic similarity and profile structure

PCA revealed consistent separation of samples by pool size across all 3 LC-MS panels and both strains (Fig. 1). In all panels, $n = 5$ samples were consistently separated from $n = 50$ and $n = 100$ samples along PC1, which explained 46% to 62% of total variance, while $n = 50$ and $n = 100$ samples clustered more closely together. Separation by age was also visible along PC2, particularly in the HILIC panels, consistent with known age-related metabolomic changes. To further characterize this age-related structure, we fit a per-metabolite linear mixed model (age, strain, and log-transformed pool size as fixed effects, with a random intercept for each strain by population replicate) across all 4 panels, applying FDR correction within panel. This identified 135 metabolites with a

Table 1. Metabolite counts by LC-MS panel and experiment before and after quality-control filtering. For each panel, defined by chromatography column (C18 or HILIC) and ionization mode (positive or negative), the number of metabolites detected (raw) and retained after quality-control filtering (metabolites with a coefficient of variation > 0.30 in pooled QC samples were removed; post QC) is shown separately for Experiment 1 (inbred vs outbred populations) and Experiment 2 (dietary perturbation). The C18 negative panel in Experiment 1 was excluded from downstream analyses due to an insufficient number of retained metabolites ($n = 2$).

Experiment 1: Effects of pool size, genetic background, and age on metabolomic structure			
Column	**Ion mode**	**Metabolites (raw)**	**Metabolites (post-QC)**
C18	Negative	2	2*
C18	Positive	19	19
HILIC	Negative	147	142
HILIC	Positive	51	50
Total		219	213

*2 is not enough for downstream analyses, so this panel was dropped leaving 211 for further analyses.

Experiment 2: Effects of pool size and replication on dietary metabolomic signal detection			
C18	Negative	62	62
C18	Positive	69	69
HILIC	Negative	61	52
HILIC	Positive	18	18
Total		210	201

significant age effect (2 in C18 negative, 14 in C18 positive, 87 in HILIC negative, and 32 in HILIC positive; Supplementary Table 8).

Pairwise Euclidean distances were consistent with the PCA, with $n = 5$ profiles being more dissimilar to larger pools than the larger pools were to each other (Fig. 2, Supplementary Table 4). The effect was strongest in the HILIC panels, where 5 to 100 distances were 2.3 to 3.9 times larger than 50 to 100 distances across all strains and ages, with HILIC negative showing the greatest absolute distances reflecting its larger metabolite count. In C18 positive, the same directional pattern held in most conditions, though the difference between pairs was less pronounced (ratio range 1.6 to 2.8×), with the notable exception of older CRB flies where 5 to 50 and 50 to 100 distances were nearly equivalent (0.66 and 0.64, respectively). Across all panels, 5 to 50 distances were consistently intermediate, indicating that most profile divergence occurs at the first pooling step, with comparatively little additional change from $n = 50$ to $n = 100$.

PERMANOVA confirmed that pool size explained a statistically significant proportion of metabolomic variance in all 3 panels, though the magnitude varied considerably (Fig. 3a). The effect was largest in HILIC negative ($R^2 = 0.333$, $P = 0.001$) and HILIC positive ($R^2 = 0.151$, $P = 0.001$), and notably smaller in C18 positive ($R^2 = 0.057$, $P = 0.001$), where strain was the overwhelmingly dominant source of variation ($R^2 = 0.695$, $P = 0.001$). Strain was a significant predictor across all panels ($R^2 = 0.262$ to 0.695, all $P = 0.001$), as was age ($R^2 = 0.029$ to 0.189, all $P = 0.001$), with the largest age effect in HILIC negative. The weaker pool size signal in C18 positive likely reflects the fact that strain differences account for so much of the total variance in that panel that the pool size effect, while present, represents a smaller relative fraction. Importantly, pool size remained significant after marginal adjustment for both strain and age in all panels, confirming that its effect on metabolomic composition is independent of biological group differences.

Betadisper analysis showed that the decline in multivariate dispersion with increasing pool size was clearest in HILIC negative, consistent with the largest pool size R^2 in that panel, and more variable elsewhere (Fig. 3b). In HILIC negative, mean dispersion at $n = 5$ ranged from 1.39 to 2.01 across strain and age subgroups, falling to 1.03 to 1.36 at $n = 50$ and 0.79 to 1.16 at $n = 100$, indicating a consistent improvement in reproducibility with larger pool sizes. In HILIC positive, a directional decline was present in the ORWT strain ($n = 5$: 1.06 to 1.30; $n = 50$: 0.84 to 1.00) but was less consistent in CRB, where dispersion was similar across all pool sizes (0.82 to 0.89). In C18 positive, dispersion patterns were similarly flat across pool sizes for both strains (CRB: 0.71 to 0.79; ORWT: 0.55 to 0.80 across pool sizes), consistent with the weak overall pool-size effect in this panel. In HILIC negative, both strains showed declining dispersion from $n = 5$ to $n = 50$ (CRB: 1.39–1.59 to 1.03–1.10; ORWT: 1.49–2.01 to 1.17–1.36), but diverged at $n = 100$: ORWT continued to decline (0.79 to 1.11), whereas CRB dispersion increased slightly relative to $n = 50$ (1.14 to 1.16), suggesting a less consistent improvement in reproducibility for the outbred population at the largest pool size in this panel.

Experiment 2: Effects of pool size and replication on dietary metabolomic signal detection

LC-MS and QC

A total of 210 metabolites were detected across the 4 LC-MS panels, of which 201 passed quality-control filtering and were used for downstream analyses (Table 1). Panel size again varied considerably, from 18 metabolites in HILIC positive to 69 in C18 positive; results from HILIC positive, the smallest panel, should be interpreted with this in mind, as they are more sensitive to individual metabolite-level variability and carry a lower multiple-testing burden than the other panels (see also the caveats noted for HILIC positive sensitivity and FDR estimates below).

Metabolomic similarity and profile structure

PCA showed consistent separation by pool size across all metabolite panels, with $n = 5$ samples clearly separated from $n = 50$ and $n = 100$, which clustered more closely (Fig. 4), mirroring the pattern observed in Experiment 1. PC1 explained 51% to 64% of total variance across panels. Along PC2, samples showed separation by diet, most clearly in the C18 panels and to a lesser extent in HILIC negative, while HILIC positive exhibited greater scatter and less distinct clustering. This likely reflects the smaller number of metabolites in this panel ($n = 18$) relative to the others.

Pairwise Euclidean distances confirmed this pattern, with $n = 5$ profiles substantially more dissimilar to larger pool sizes than $n = 50$ and $n = 100$ were to each other (Fig. 5, Supplementary Table 5). Across all panels, mean 5 to 100 distances were approximately 1.8 to 2.8 times larger than 50 to 100 distances, with 5 to 50 distances consistently intermediate. These patterns were consistent across diets, indicating that most profile divergence occurs at the first increase in pool size.

PERMANOVA further showed that pool size explained a large proportion of metabolomic variance in all panels (Fig. 6a), accounting for 39% to 54% of total variance depending on the panel. Diet was also a significant predictor, explaining 11% to 27% of variance, with stronger effects in the C18 panels relative to the HILIC panels. The persistence of a strong pool size effect after accounting for diet indicates that pooling influences metabolomic composition independently of treatment effects.

Betadisper analysis revealed that improvements in reproducibility with increasing pool size were most pronounced in the C18

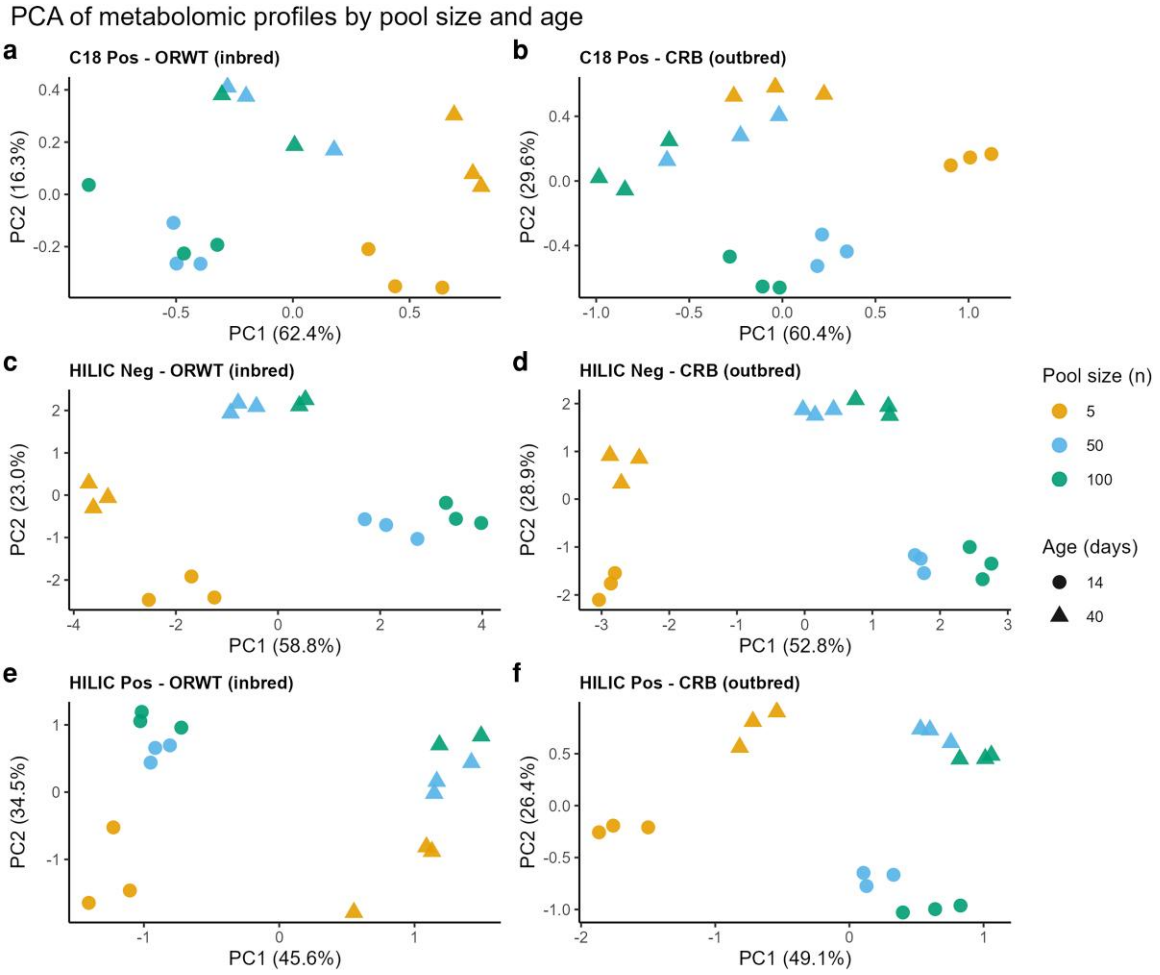

Fig. 1. PCA of metabolomic profiles by pool size and age across 3 metabolite panels. Points represent individual samples colored by pool size and shaped by age. PCA was performed on mean-centered metabolite values without additional scaling. Variance explained by each PC is shown in parentheses on axis labels.

panels (Fig. 6b). In both C18 panels, dispersion declined substantially from $n = 5$ to $n = 50$, with little additional improvement at $n = 100$, indicating diminishing returns beyond intermediate pool sizes. In contrast, the HILIC panels showed weaker and less consistent patterns, with HILIC negative exhibiting relatively stable dispersion across pool sizes and HILIC positive showing variable behavior, including an increase in dispersion at $n = 100$ under the HSD treatment. This latter pattern should be interpreted cautiously given the small number of metabolites in this panel.

Together, these results confirm that pool size has a strong effect on metabolomic profile structure, with the largest changes occurring between $n = 5$ and $n = 50$ and comparatively little further change from $n = 50$ to $n = 100$, a pattern consistent with diminishing returns at larger pool sizes. Notably, this substantial pool-size effect emerged despite the use of a single, genetically uniform inbred background (ORWT), indicating that pool size shapes metabolomic measurement even among individuals that are nominally genetically identical, likely reflecting non-genetic sources of individual variation such as developmental stochasticity and microenvironmental differences. Compared to Experiment 1, the magnitude of the pool size effect was substantially larger in this dataset, and the presence of a strong diet signal further demonstrates that pooling effects persist even in the presence of pronounced biological differences.

Detection of diet-associated metabolites across pool sizes

To assess the practical consequences of pool size choice on biological inference, we examined whether pool size affects the ability to detect diet-related changes in individual metabolite levels. For each metabolite and pool size, a linear model was fitted to test the effect of diet, and results at $n = 100$ were treated as a reference condition (see Methods). Sensitivity was substantially lower at $n = 5$ than at $n = 50$ across all panels (Fig. 7, a and b, Supplementary Table 6b). At $n = 5$, sensitivity ranged from 50% in both HILIC negative and HILIC positive to 73% in C18 positive, meaning that between 27% and 50% of reference-condition diet effects would be missed depending on the panel. At $n = 50$, sensitivity improved substantially across all panels, reaching 89% in C18 negative, 90% in C18 positive, 75% in HILIC negative, and 100% in HILIC positive, though the HILIC positive result should be interpreted cautiously given the small number of metabolites in that panel (18 metabolites) and the limited number of significant effects at $n = 100$ ($n = 10$).

FDR was low and broadly consistent across panels at both pool sizes (Fig. 7b, Supplementary Table 6b). In C18 negative, FDR was 19% at $n = 5$ and 14% at $n = 50$; in C18 positive, FDR was 10% at $n = 5$ and remained similar at $n = 50$ (10%). HILIC negative showed FDR of 17% at $n = 5$ rising modestly to 21% at $n = 50$. HILIC positive showed FDR of 29% at $n = 5$ dropping to 17% at $n = 50$, though as noted above the absolute counts are small.

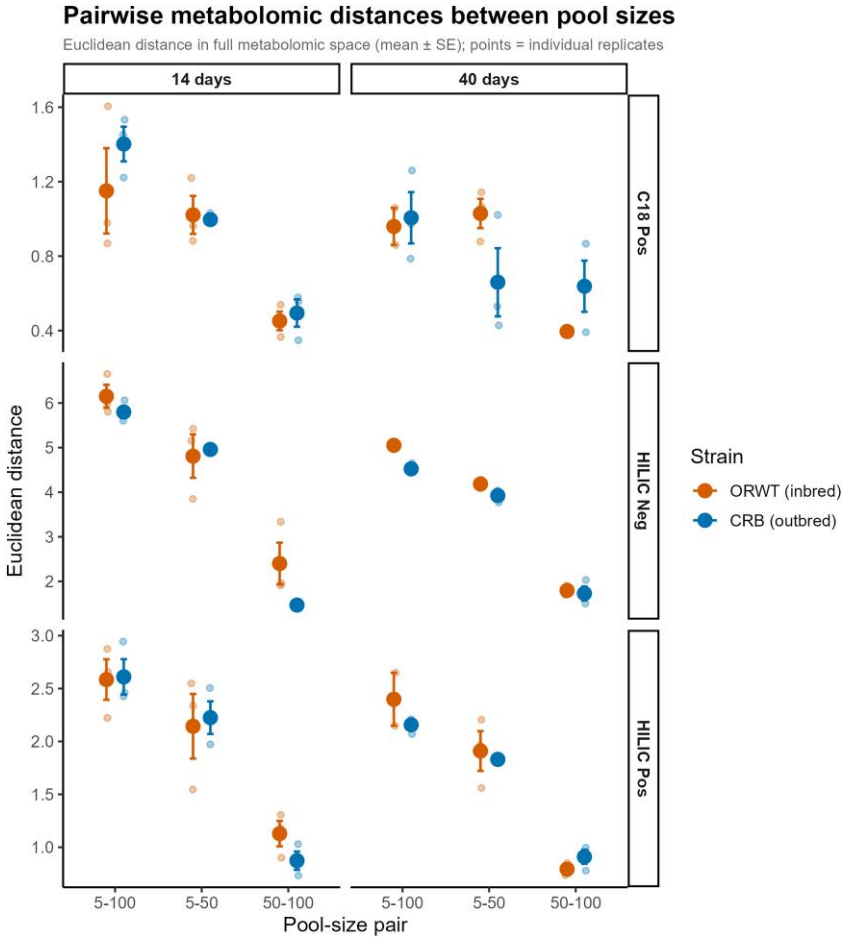

Fig. 2. Pairwise Euclidean distances between pool sizes in the full metabolomic space. Mean distance (± SE) between all pool-size pairs (5 to 100, 5 to 50, 50 to 100) is shown for each metabolite panel, colored by strain. Individual replicate values are shown as transparent points.

Effect size attenuation analysis confirmed that pool size does not systematically bias the direction or magnitude of diet effect estimates, but does introduce substantial noise at $n = 5$ (Fig. 7c, Supplementary Table 6a). Across all metabolites, Pearson correlations between diet effect sizes at smaller pool sizes and at $n = 100$ were consistently high at $n = 50$ ($r = 0.92$ to 0.98 across panels), with points clustering tightly around the 1:1 line. At $n = 5$, correlations were lower and more variable across panels ($r = 0.48$ to 0.88), with HILIC positive showing the greatest scatter ($r = 0.48$) and C18 positive the highest reliability ($r = 0.87$). Notably, metabolites with the largest diet effect sizes at $n = 100$ showed the greatest consistency across pool sizes, clustering close to the 1:1 line even at $n = 5$. The increased scatter at $n = 5$ was concentrated among metabolites with modest effect sizes, suggesting that pools of $n = 5$ are adequate for detecting the strongest diet signals but produce unreliable estimates for moderate and small effects. Correlations were similar whether computed across all metabolites or restricted to those significant at $n = 100$, indicating that this reliability pattern is representative of the metabolome as a whole rather than an artifact of pre-selecting metabolites with large effects.

The supplementary sensitivity analysis confirmed that pool size has a strong and nearly universal effect on individual metabolite values (Supplementary Fig. 1, Supplementary Table 6d). Across all panels, the majority of metabolites showed |t|-values well above the approximate significance threshold of 2 for the log(pool size) term, and almost all metabolites showed a positive relationship with pool size. This directional consistency indicates that pool size introduces a systematic upward shift in measured metabolite levels rather than random noise, likely reflecting the averaging of individual stochasticity in small pools. The C18 panels showed notably higher pool-size sensitivity than the HILIC panels, consistent with the stronger pool size effects in these panels observed in the similarity analysis.

Impact of replicate downsampling on signal detection

To evaluate the robustness of diet-associated metabolite signals, we systematically reduced replicate number across pool sizes and quantified the retention of significant metabolites relative to a reference condition (pool size = 100, full replication) (Supplementary Table 7). Across all datasets, the number of diet-significant metabolites declined with decreasing replicate number, with more pronounced losses observed at smaller pool sizes (Fig. 8). At the largest pool size (100 individuals), a high proportion of the reference signal was retained at moderate replication, with gradual declines as replicate number decreased. In contrast, smaller pool sizes accelerated signal loss, particularly at pool size = 5, where the number of detected metabolites declined sharply even when several replicates remained.

Classification of significant metabolites relative to the reference condition revealed that reductions in replicate number primarily resulted in the loss of true positives rather than an increase in false positives. Across all datasets and pool sizes, the proportion of true positives decreased steadily with replicate

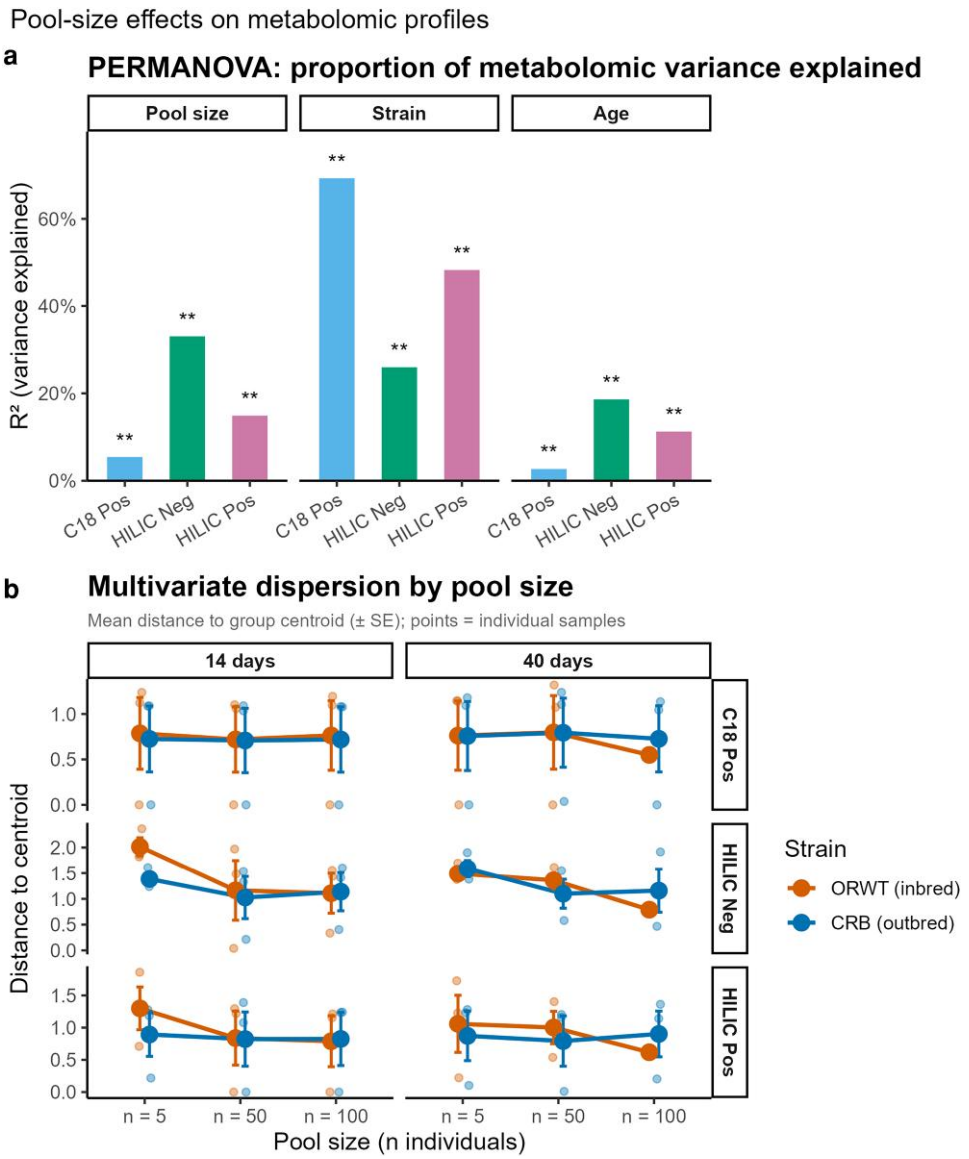

Fig. 3. Pool-size effects on metabolomic profiles. a) Proportion of total metabolomic variance explained by pool size, strain, and age for each panel based on PERMANOVA results. b) Multivariate dispersion by pool size, shown as mean distance to group centroid (± SE) within each strain × age combination. Individual sample values are shown as transparent points.

removal, while false positives remained comparatively low and showed only modest increases at low replication levels. These patterns indicate that, similar to our findings of reduced pool size, reduced replication primarily limits statistical power rather than inflating spurious detections.

Effect size retention across pool sizes under replicate downsampling

To determine whether signal stability depended on the magnitude of the underlying biological effect, metabolites were grouped into high-, medium-, and low-effect bins based on absolute diet effect size under the reference condition. Signal retention differed markedly across these categories (Fig. 9). High-effect metabolites remained detectable across a wide range of downsampling conditions, whereas medium-effect metabolites showed more gradual declines and low-effect metabolites were rapidly lost even with modest reductions in replication.

While all datasets exhibited a consistent directional trend of decreasing detection with replicate loss, the rate of signal decline varied substantially across datasets. In some datasets, high- and medium-effect metabolites were retained across a broader range of downsampling levels, whereas in others, detection declined more rapidly across all effect-size categories. These differences indicate that, in addition to effect size, dataset-specific properties influence the robustness of metabolite detection under reduced replication.

In addition to effects on statistical significance, replicate downsampling also influenced the magnitude and consistency of estimated diet effect sizes. At full replication, effect size estimates were highly concordant across pool sizes, with strong agreement between estimates from pool size of 50 relative to the reference pool size of 100. However, as replicate number decreased, effect size estimates became increasingly variable, with this effect strongly modulated by pool size. Smaller pool sizes, particularly pool size = 5, exhibited substantially greater variability and more

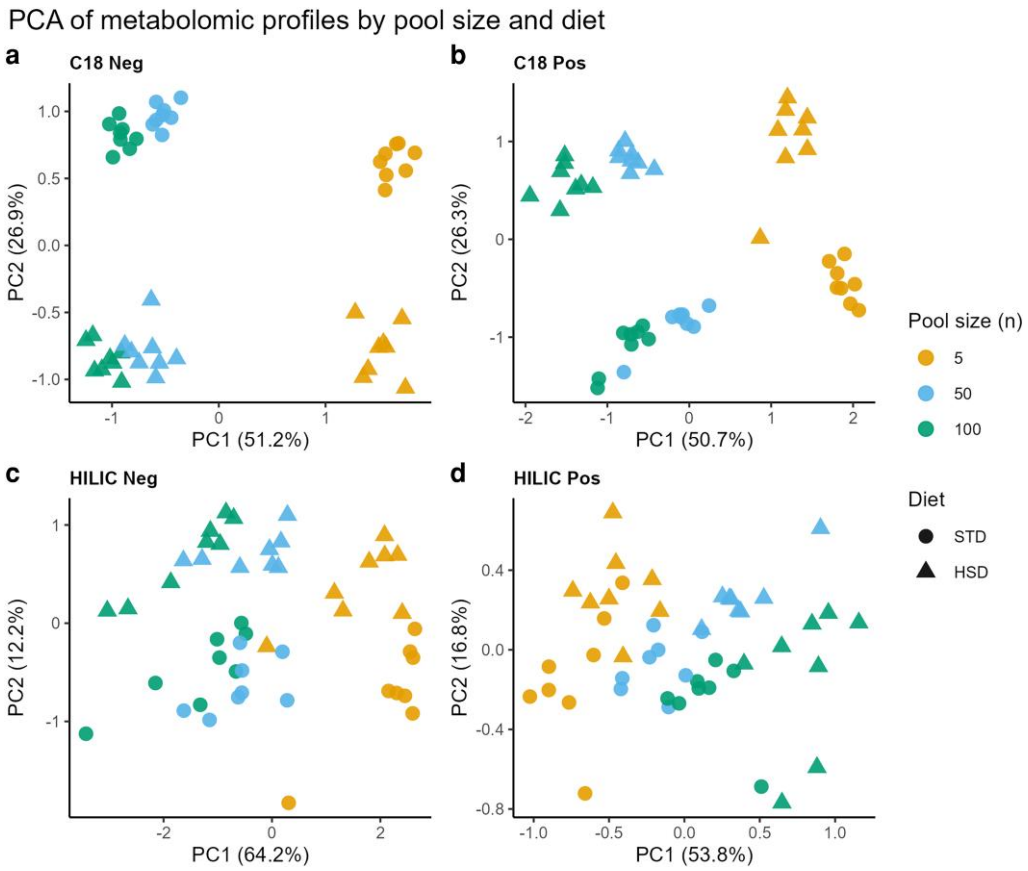

Fig. 4. PCA of metabolomic profiles coded by pool size and diet across 4 metabolite panels. Points represent individual samples colored by pool size and shaped by diet. PCA was performed on mean-centered metabolite values without additional scaling. Variance explained by each PC is shown in parentheses on axis labels.

rapid degradation of effect size estimates compared to larger pool sizes.

Consistent with this pattern, the rate of signal loss under downsampling differed markedly across pool sizes. Larger pool sizes retained diet-associated signals across a broader range of replicate reductions, whereas smaller pool sizes showed accelerated loss of both effect size stability and statistical detection. This effect was most pronounced for metabolites with smaller effect sizes, which were rapidly lost under conditions of both low replication and small pool size.

Overall, these results indicate that both biological replication and pool size jointly determine the robustness of diet-associated metabolite signals. These results demonstrate that pool size and replication are not interchangeable, but instead each play a distinct, statistically independent role in determining the reliability of metabolomic inference (see Determinants of metabolite signal retention, below). Larger pool sizes reduce variability and stabilize effect size estimates, thereby slowing the rate at which signals are lost under reduced replication, while increased replication improves detection across all conditions. Together, these factors jointly determine the effective signal-to-noise ratio and, consequently, the ability to detect consistent biological effects.

Functional signal retention across sampling conditions

Diet significantly altered metabolite profiles across multiple functional modules in the reference dataset (pool size = 100, maximum replication). The strongest responses were observed in lipid-associated pathways, including fatty acid oxidation and fatty acid overflow, as well as central carbon metabolism and one-carbon metabolism (Fig. 10). These coordinated changes are consistent with increased carbon flux and lipid remodeling under high-sugar conditions, indicating system-wide metabolic reprogramming.

To evaluate the robustness of these signals to reduced sampling effort, we quantified the retention of reference metabolites across replicate downsampling and pool size reduction (Fig. 11). Across all modules, the proportion of recovered metabolites declined with decreasing replicate number, with additional losses observed at smaller pool sizes. At the largest pool size (100 individuals), most modules retained a high proportion of the reference signal at moderate replication, with gradual declines as replicate number decreased. In contrast, smaller pool sizes accelerated signal loss, particularly at pool size = 5, where recovery declined sharply even at higher replicate numbers.

The extent of signal loss varied across metabolic modules. Lipid-associated pathways, including fatty acid oxidation and fatty acid overflow, as well as one-carbon metabolism, showed relatively high retention across conditions, remaining detectable under reduced replication and smaller pool sizes. In contrast, redox metabolism and nucleotide turnover exhibited rapid declines in recovery, indicating greater sensitivity to reduced sampling effort. Central carbon metabolism and amino acid metabolism showed intermediate behavior, with moderate decreases in retention as replication and pool size were reduced.

Together, these results again demonstrate that both replicate number and pool size influence the detection of metabolite-level

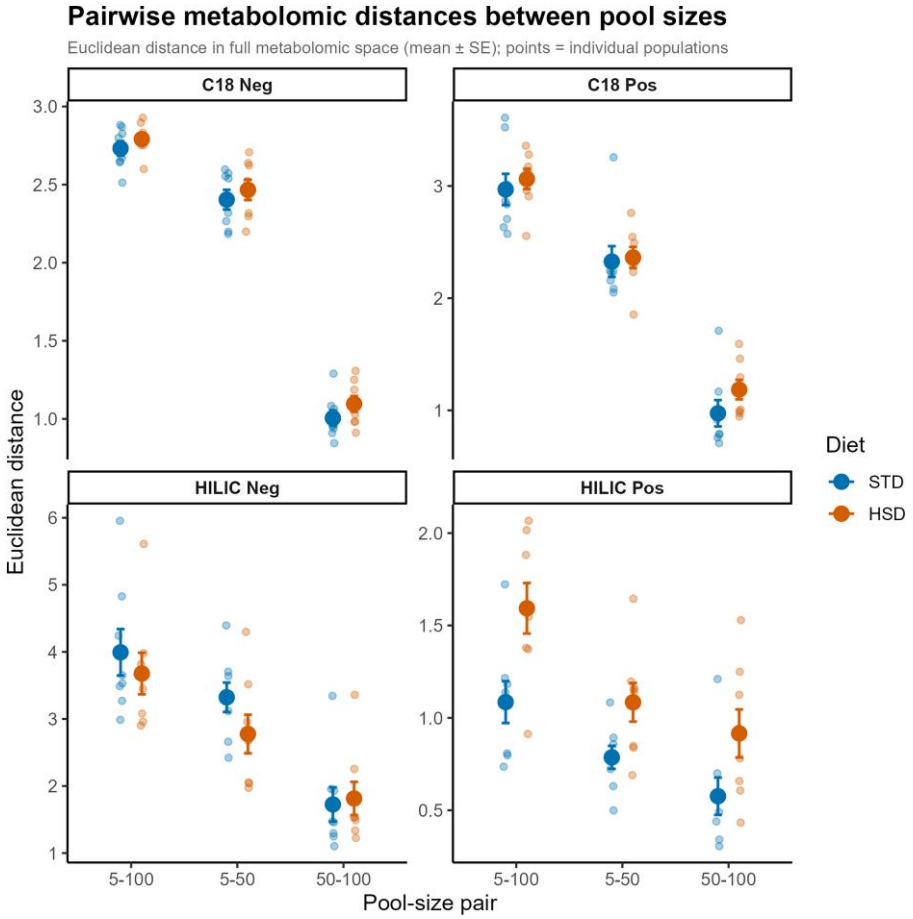

Fig. 5. Pairwise Euclidean distances between pool sizes in the full metabolomic space. Mean distance (± SE) between all pool-size pairs (5 to 100, 5 to 50, 50 to 100) is shown for each metabolite panel, colored by diet.Individual replicate values are shown as transparent points.

signals. Importantly, the robustness of signal retention differed across metabolic pathways, indicating that some biological processes are more resilient to experimental downsampling than others.

Determinants of metabolite signal retention

To identify the factors underlying variation in metabolite signal retention across downsampling conditions, we modeled detection frequency as a function of effect size, variability, pool size, and replication using a beta-binomial generalized linear mixed model (Table 2). Across all datasets, retention was strongly positively associated with the magnitude of the underlying diet effect ($z = 15.24$, $P < 2 \times 10^{-16}$); each 0.1-unit increase in the absolute reference diet effect size nearly tripled the odds that a metabolite's signal was retained (odds ratio = 2.95), indicating that metabolites with larger effect sizes are substantially more robust to replicate downsampling.

In contrast, variability in effect size estimates across iterations was negatively associated with retention ($z = -8.14$, $P = 4.00 \times 10^{-16}$); each 0.01-unit increase in the standard deviation of the diet coefficient across iterations reduced the odds of retention by about 15% (odds ratio = 0.85). This relationship provides a mechanistic explanation for differences in signal robustness, linking reduced detection to increased estimate variability rather than systematic bias.

Pool size and replicate number also had strong effects on detection. Pool size = 5 showed a substantial reduction in retention relative to pool size = 100 ($z = -5.91$, $P = 3.44 \times 10^{-9}$), whereas pool size = 50 was not statistically distinguishable from pool size = 100 ($z = -1.29$, $P = 0.20$). Because pool size was modeled as a categorical predictor rather than a continuous or smoothed term, this pattern, a large gap between pool size 5 and the 2 larger pool sizes with no detectable difference between 50 and 100, indicates diminishing returns above an intermediate pool size without implying a formally tested nonlinear function of pool size. Increasing replicate number strongly improved detection ($z = 7.35$, $P = 1.96 \times 10^{-13}$). The interaction between pool size and replicate number was not statistically significant (likelihood ratio test: $\chi^2 = 1.33$, df = 2, $P = 0.51$), indicating that pool size and replicate number act additively rather than interactively on individual metabolite retention. Each additional replicate increased the odds of signal retention by approximately 35% ($\exp(0.3029) \approx 1.35$), after accounting for pool size, effect size, and variability.

Together, these results demonstrate that metabolite detection is governed by a balance between biological signal strength and measurement variability, with pool size and replication each independently modulating the ability to detect consistent effects. Differences in signal robustness across datasets are therefore explained not only by experimental design but also by underlying variation in effect size and stability among metabolites.

Quantitative guidance for pool size vs replication

To provide quantitative guidance on the trade-off between pool size and biological replication, we modeled the total number of

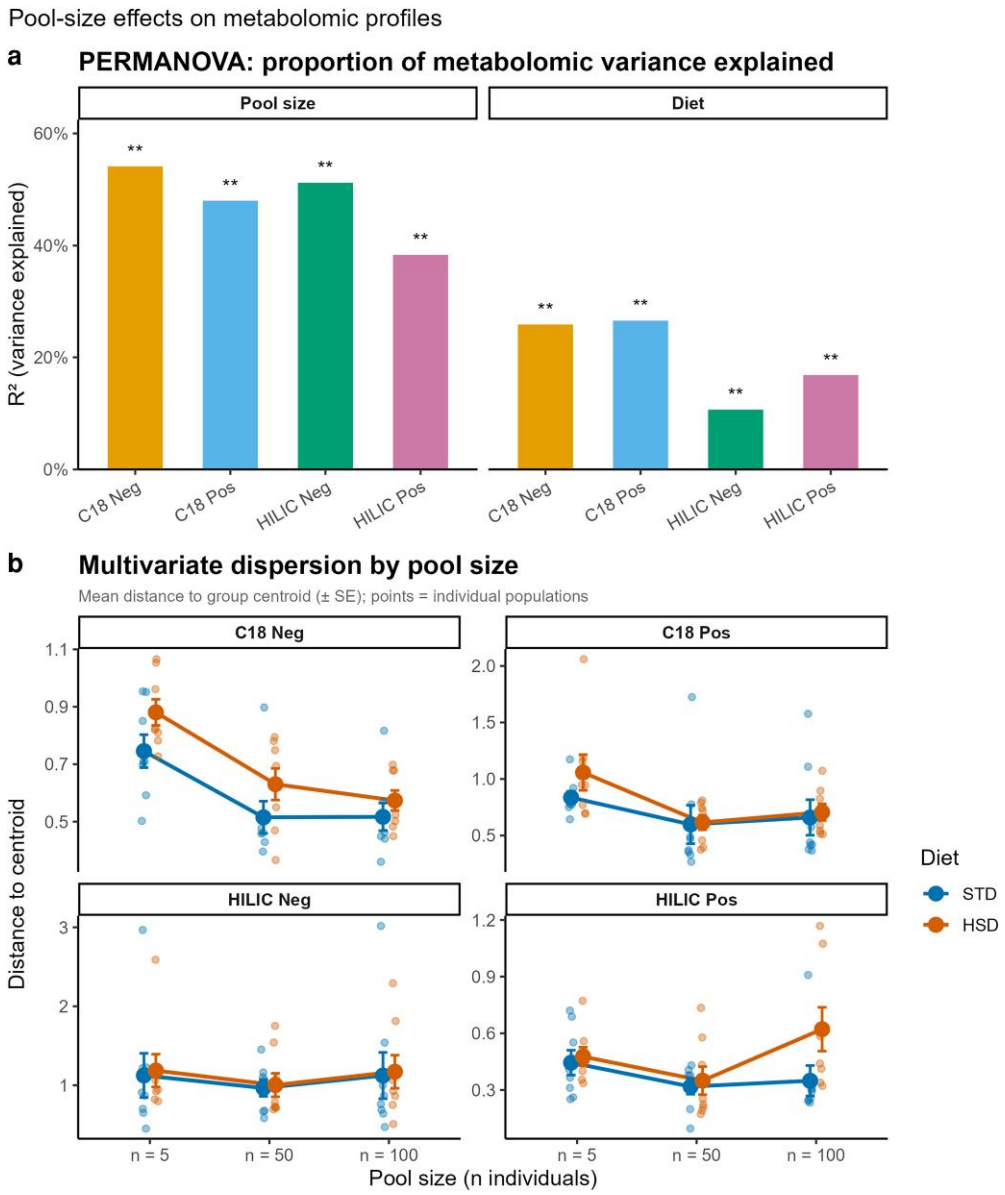

Fig. 6. Pool-size effects on metabolomic profiles. a) Proportion of total metabolomic variance explained by pool size and diet for each panel based on PERMANOVA results.(B) Multivariate dispersion by pool size, shown as mean distance to group centroid (± SE) within each diet group.

diet-significant metabolites detected per downsampling iteration using a zero-inflated Poisson GLMM (see Methods). The zero-inflation component was strongly supported relative to a standard Poisson model (likelihood ratio test: $\chi^2 = 725$, df = 3, $P < 2.2 \times 10^{-16}$), and model diagnostics indicated an adequate fit (dispersion = 0.95; ratio of observed to simulated zero-count iterations = 1.17). The interaction between replicate number and pool size was strongly supported (likelihood ratio test: $\chi^2 = 415.7$, df = 2, $P < 2.2 \times 10^{-16}$). This contrasts with the per-metabolite retention model (Table 2), where the interaction was not significant. The 2 models test different quantities: the probability that an individual metabolite is retained vs the total yield of significant metabolites within an iteration. Small pool-dependent differences distributed across many metabolites may therefore affect total metabolite yield even when the interaction is not detectable at the individual-metabolite level. Each additional biological replicate increased the expected number of detected significant metabolites, expressed as an incidence rate ratio (IRR = exp(β)), by approximately 30% at pool size 5 (IRR = exp(0.263) = 1.30), 16% at pool size 50 (IRR = exp(0.263 – 0.112) = 1.16), and 13% at pool size 100 (IRR = exp(0.263 – 0.137) = 1.13), indicating that the marginal benefit of additional replication was largest at small pool sizes and diminished as pool size increased. The zero-inflation submodel further indicated that the probability of an iteration detecting no significant metabolites at all was approximately 6.0% at pool size 5, dropping to approximately 3.0% at pool size 50 (z-test on the zero-inflation coefficient, $P = 0.001$); pool size 100 was not significantly different from pool size 5 in this respect ($P = 0.24$), a result we interpret cautiously given the small number of zero-count iterations available to estimate this comparison.

Discussion

Pooling is a necessary but highly variable component of omics-based experimental design (eg genomics, transcriptomics, proteomics, lipidomics, and metabolomics), particularly in small

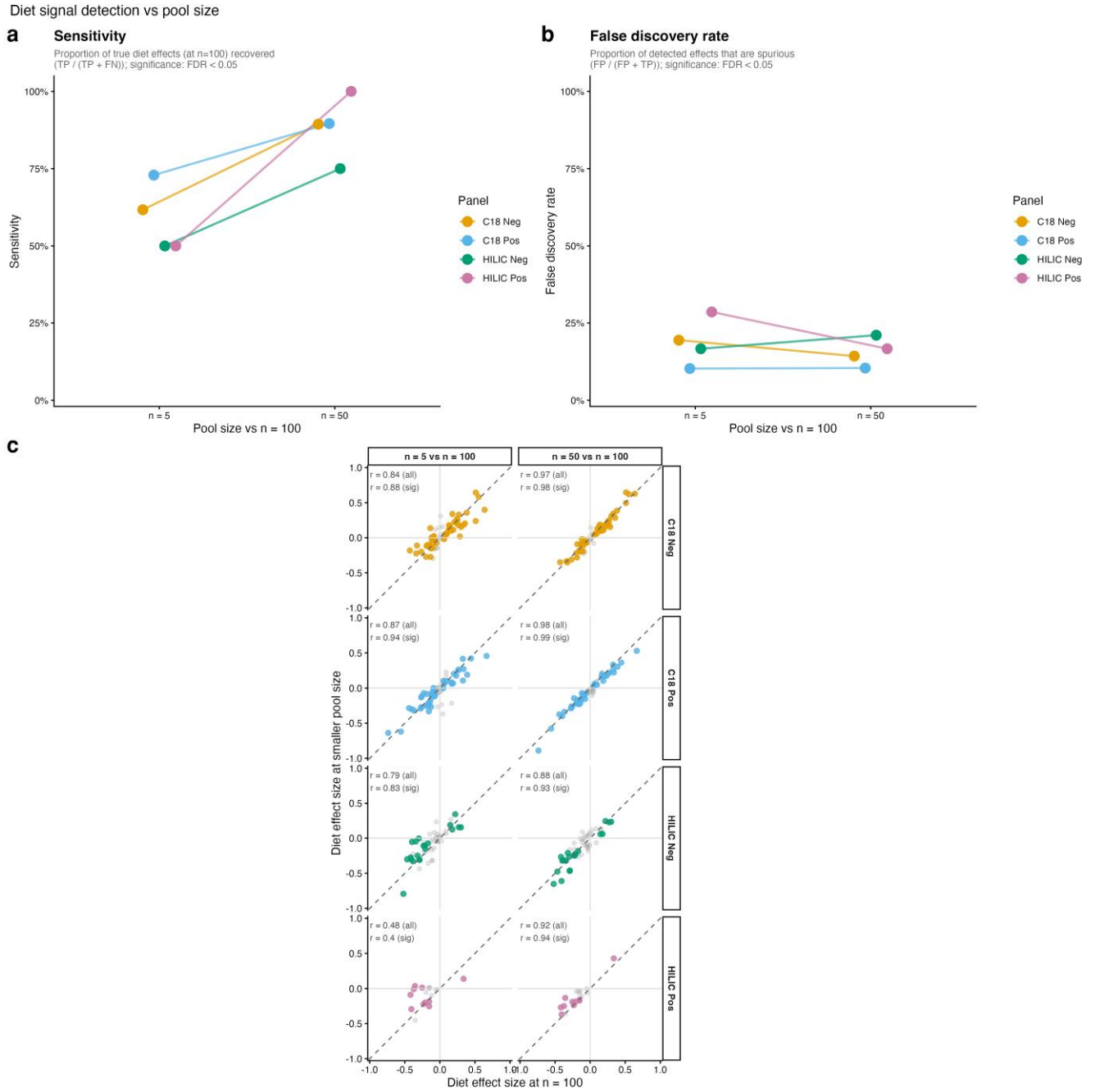

Fig. 7. Diet signal detection vs pool size. a) Sensitivity—proportion of true diet effects (significant at $n = 100$, FDR < 0.05) recovered at $n = 5$ or $n = 50$. b) False discovery rate—proportion of detected effects not supported at $n = 100$. Each point is one LC-MS panel; lines connect the 2 pool-size comparisons. c) Diet effect sizes at smaller pool sizes vs $n = 100$ for all metabolites. Colored points are significant at $n = 100$ (FDR < 0.05); gray points are not. Pearson r shown for all metabolites and the significant subset.

organisms such as *Drosophila melanogaster* (Zhao et al. 2022; Hubert et al. 2025; Vecchié et al. 2026). Despite its widespread use across these fields, the consequences of pool size for biological inference remain especially poorly defined for metabolomics, where measurements integrate the combined physiological state of pooled individuals in ways that are less well characterized than for other omics layers. Here, using 2 complementary experimental designs, we show that pool size and biological replication jointly determine both the structure of metabolomic profiles and the ability to detect biologically meaningful signals. Importantly, the structural effects of pool size were dataset-dependent, with the most substantial differences occurring between very small pools ($n = 5$) and larger pools ($n \geq 50$). At the level of individual metabolites, pool size and biological replication contributed independently, rather than interactively, to detection odds, although the total yield of significant metabolites per downsampling iteration showed a statistically supported, pool-size-dependent benefit of additional replication (see Quantitative guidance for pool size vs replication, above).

Pool size fundamentally alters metabolomic profile structure

In Experiment 1, we found that pool size fundamentally alters the structure of metabolomic profiles. Across all panels, samples

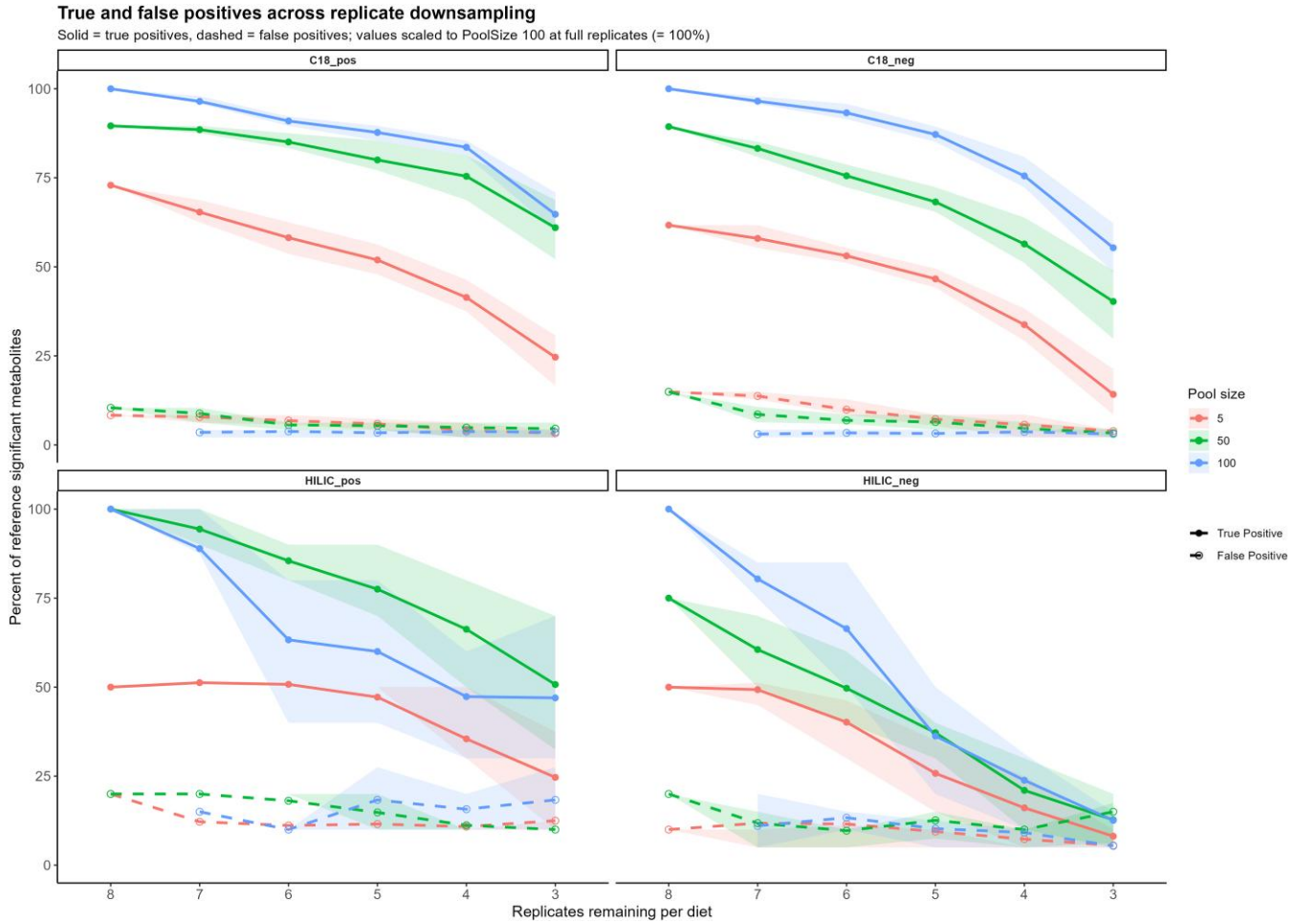

Fig. 8. Retention of true and false positive metabolite detection across replicate and pool-size downsampling. The proportion of diet-associated metabolites identified under downsampling was expressed as a percentage of the reference set defined at PoolSize = 100 with full replicates (8 per diet; 100%). True positives (solid lines) represent metabolites that were significant (FDR < 0.05) in both the downsampled and reference datasets, while false positives (dashed lines) were significant only in the downsampled condition. Lines show the mean percentage across resampled combinations of replicate removal (up to 64 iterations per removal level), and shaded ribbons indicate the interquartile range (25th–75th percentile), reflecting variability across downsampling iterations. Across all metabolite panels, reductions in replicate number and pool size led to a progressive loss of true positives, while false positives remained comparatively low, indicating reduced statistical power rather than systematic inflation of spurious detections under downsampling.

pooled at $n = 5$ were consistently separated from those pooled at $n = 50$ and $n = 100$, with much smaller differences observed between the 2 larger pool sizes. Pairwise distance analyses further showed that most of the divergence in metabolomic profiles occurs at the first increase in pool size, from 5 to 50 individuals, with relatively little additional change from 50 to 100. These results indicate that small pools capture a substantially different representation of the metabolome than larger pools, likely reflecting increased influence of individual-level stochasticity and heterogeneity (Kendziorski et al. 2005). In contrast, larger pools appear to more reliably capture the average physiological state of the population.

The age-related metabolite changes we detected are broadly consistent with prior Drosophila aging metabolomics studies. Among long-chain acylcarnitines, linolenyl-, linoleyl-, oleoyl-, and stearoylcarnitine each declined significantly with age, paralleling the age-related decline in acylcarnitine shuttle activity reported by Hoffman et al. (2014); this pattern was not uniform across the carnitine pool, as acetylcarnitine and succinylcarnitine instead increased significantly with age. Glycolytic feeder molecules, including glucose 6-phosphate, glucose 1-phosphate, and galactose 1-phosphate, also declined significantly with age, consistent with Hoffman et al. (2014). We additionally found significant age effects among kynurenine/tryptophan pathway metabolites (kynurenine, tryptophan) and arginine–ornithine pathway metabolites (ornithine, N-acetylornithine), the 2 pathways Zhao et al. (2022) identified as most strongly associated with age-at-death parameters in DGRP flies. While these studies differ from ours in tissue scope and experimental design, the compound- and pathway-level overlap with what we observe here lends support to the biological interpretability of the age-related signals detected in this study.

Importantly, the inclusion of both inbred and outbred populations provides insight into how genetic background interacts with metabolomic variation. As expected, strain explained a substantial proportion of total variance across all panels, indicating strong underlying genetic effects on metabolomic profiles. The impact of pool size on reproducibility was broadly similar between inbred and outbred populations in C18 positive and HILIC positive, but diverged somewhat in HILIC negative, where the inbred population showed continued improvement at the largest pool size while the outbred population did not, indicating that the relationship between pool size, reproducibility, and genetic background is panel-dependent rather than uniform across the metabolome.

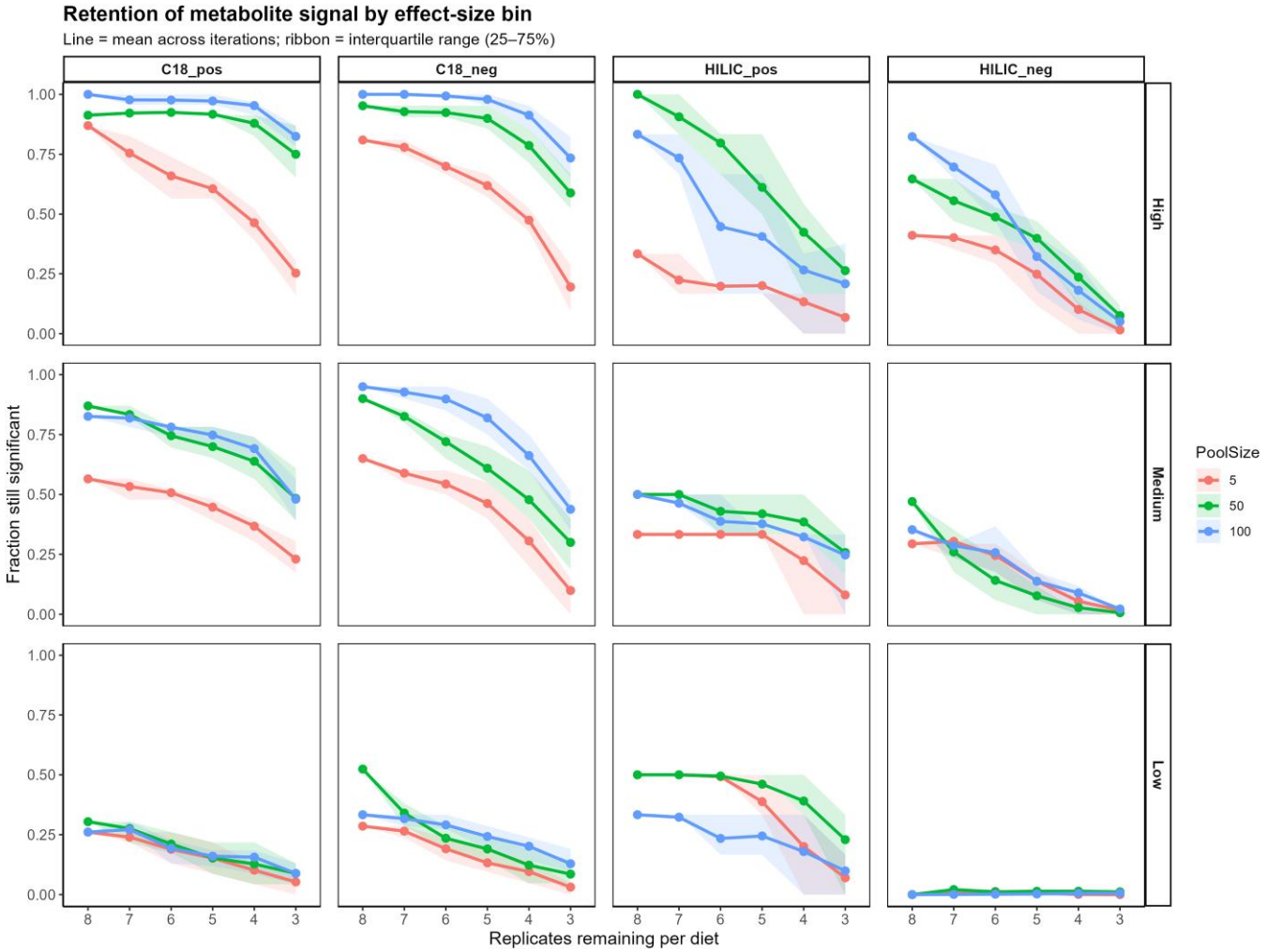

Fig. 9. Retention of metabolite signal across replicate and pool-size downsampling, stratified by effect-size bin. Metabolites were grouped into high, medium, and low effect-size bins based on tertiles of absolute diet effect sizes estimated from the full dataset (PoolSize = 100, full replicates), calculated separately within each metabolite panel. For each bin, the proportion of metabolites remaining significant (FDR < 0.05) was evaluated across up to 64 resampled combinations of replicate downsampling and pool sizes (5, 50, and 100). Lines represent the mean fraction of metabolites remaining significant across all downsampling iterations, and shaded ribbons indicate the interquartile range (25th–75th percentile), reflecting sensitivity to which replicate populations were removed. Across all panels, metabolites with larger effect sizes exhibited greater robustness to reductions in replicate number and pool size, whereas low-effect metabolites rapidly lost significance under downsampling.

This pattern runs counter to the expectation that outbred populations, which harbor greater individual-level genetic variation, would show the more consistent benefit from increased pool size; instead, our results suggest that the benefits of increased pool size are not determined solely by genetic variation (Mackay et al. 2012). Pooling effects appear to depend on additional dataset-specific factors, including metabolite composition and measurement characteristics.

The magnitude of the pool size effect, however, varied substantially across datasets. Pool size explained a large proportion of variance in the HILIC panels, particularly HILIC negative, but a much smaller proportion in the C18-positive panel, where strain differences dominated the overall signal. Similarly, improvements in reproducibility with increasing pool size were most evident in HILIC negative, more variable in HILIC positive, and minimal in C18 positive. These results demonstrate that while pooling can improve the stability of metabolomic measurements, its impact is not uniform and depends on the underlying structure of the dataset, including the types of metabolites measured and the relative contribution of biological vs technical variation.

In Experiment 2, we directly evaluated how pool size affects biological inference by quantifying the detection of diet-associated metabolite changes. Smaller pool sizes, particularly $n = 5$, showed substantially reduced sensitivity, missing a large fraction of diet-responsive metabolites identified at $n = 100$. In contrast, false discovery rates remained relatively low and stable across pool sizes, indicating that reduced pooling primarily results in the loss of true signals rather than the introduction of spurious ones. These findings demonstrate that small pool sizes do not necessarily bias results, but instead reduce statistical power, leading to underestimation of the extent of biological change. This principle extends beyond metabolomics to other high-dimensional omics approaches. In genomics, for example, pooled sequencing (Pool-seq) designs face an analogous trade-off between pool size and the ability to detect allele-frequency differences associated with a trait (Carvalho et al. 2023; Long et al. 2026), and similar power limitations apply in proteomics (Diz et al. 2009).

Consistent with this interpretation, effect size comparisons showed that while direction and magnitude of diet effects were

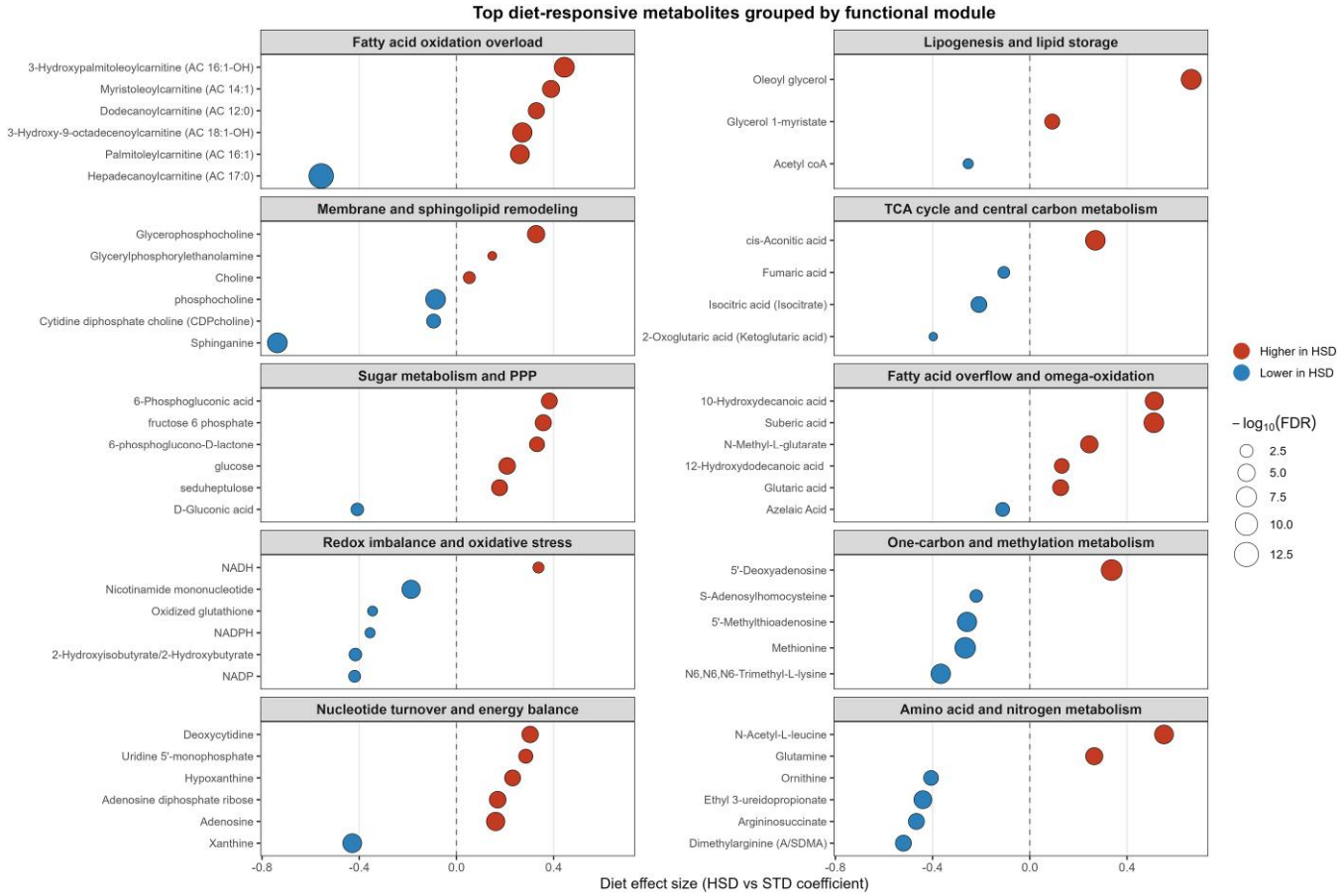

Fig. 10. Metabolic responses to diet and preservation of signals under downsampling. Top diet-responsive metabolites from the reference dataset (pool size = 100, maximum replication) grouped by functional module. Points represent individual metabolites, with position indicating effect size (HSD vs STD), size reflecting statistical significance (–log10 FDR), and color indicating direction of change.

largely preserved across pool sizes, estimates from $n = 5$ were substantially noisier and less reliable than those from larger pools. This increased variability was concentrated among metabolites with moderate and small effect sizes, whereas metabolites with large diet effects remained relatively stable even at small pool sizes. Together, these results indicate that small pools may be sufficient to detect the strongest signals, but are inadequate for capturing the full spectrum of biologically relevant variation.

Pool size and replication jointly determine signal detection

Replicate downsampling further revealed that both pool size and biological replication independently determine the robustness of metabolite detection. Across all datasets, reducing replicate number led to a progressive loss of significant metabolites, with smaller pool sizes accelerating this decline. Notably, the loss of signal under downsampling was driven primarily by a reduction in true positives rather than an increase in false positives, reinforcing the conclusion that sampling limitations primarily affect power. Importantly, pool size strongly modulated the rate of signal loss: larger pools retained diet-associated signals across a broader range of replicate reductions, whereas pool size = 5 exhibited rapid degradation of both statistical significance and effect size stability. These results demonstrate that pool size and replication are not interchangeable, but instead play complementary roles in determining the reliability of metabolomic inference.

The robustness of metabolite detection was also strongly dependent on effect size. Metabolites with large diet effects were consistently retained across downsampling conditions, whereas those with smaller effects were rapidly lost as replication and pool size decreased. While this pattern was consistent in direction across datasets, the rate at which signal was lost varied substantially, indicating that dataset-specific properties influence the sensitivity of metabolite detection.

Signal detection is governed by biological effect size and variability

To identify the factors underlying these differences, we modeled metabolite detection frequency as a function of effect size, variability, pool size, and replication. Retention was strongly positively associated with effect size and negatively associated with variability in effect estimates, indicating that metabolites with stronger and more stable signals are more robust to reduced sampling. These results demonstrate that signal detection is governed by a balance between biological signal strength and measurement variability, with experimental design modulating this relationship. Differences in robustness across datasets are therefore explained not only by sampling design but also by variation in the underlying signal-to-noise structure of the metabolome.

At the level of biological interpretation, functional module analysis revealed that not all metabolic pathways are equally robust to reduced sampling effort. Lipid-associated pathways and one-carbon metabolism showed relatively high retention across

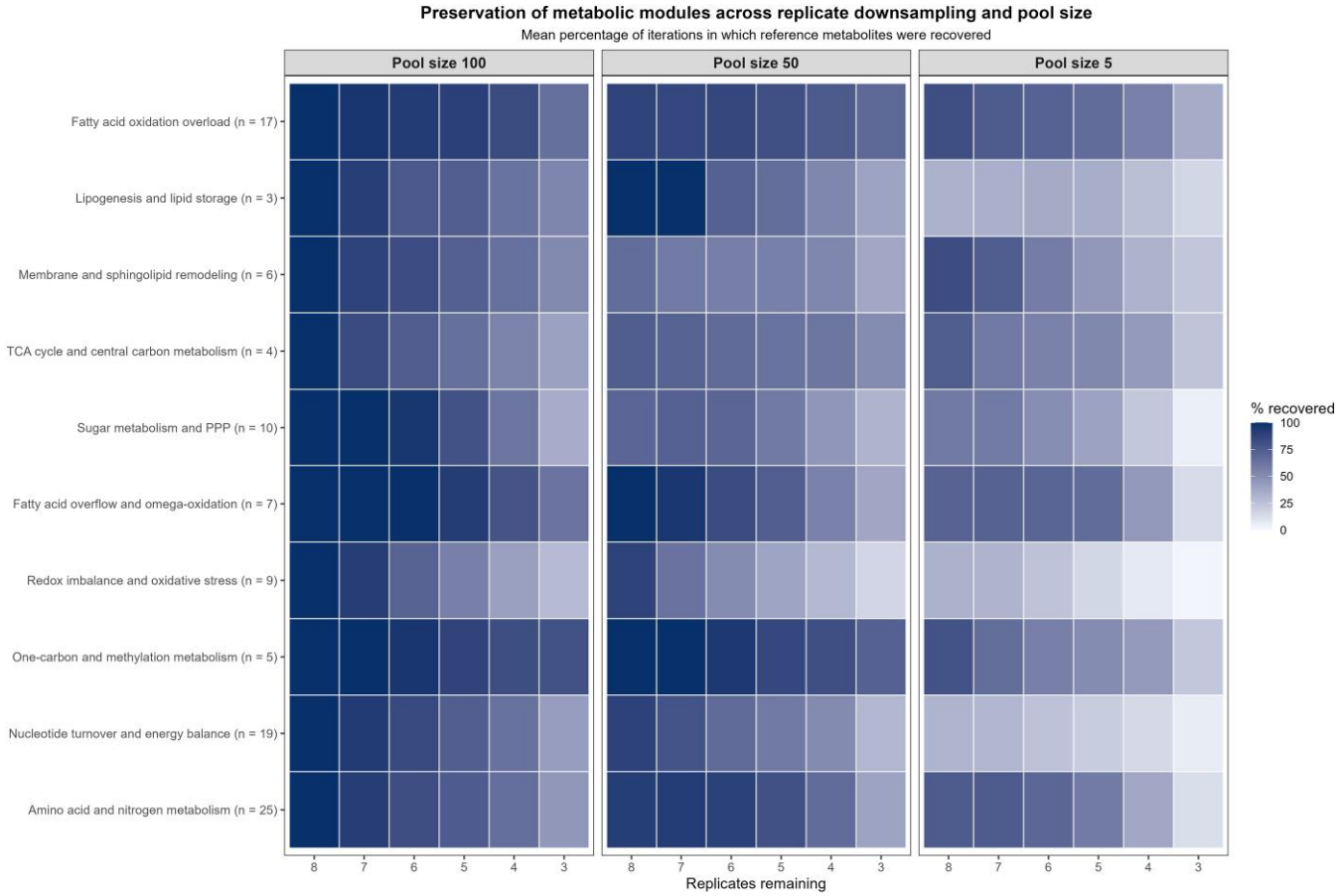

Fig. 11. Module preservation heatmaps by pool size. Heatmaps showing the percentage of reference metabolites recovered (FDR < 0.05) across decreasing replicate number and pool sizes (100, 50, and 5 individuals; panels). Values represent the mean recovery across iterations. Metabolites are grouped into functional modules (module sizes indicated on the y-axis). Recovery declines with reduced replication and smaller pool sizes, with lipid and one-carbon metabolism remaining relatively robust, whereas redox and nucleotide metabolism show rapid loss of signal.

Table 2. Determinants of metabolite signal retention under replicate downsampling. Beta-binomial generalized linear mixed model (GLMM) evaluating the effects of reference effect size, variability (standard deviation of the diet coefficient), pool size, replicate number, and their interaction on metabolite detection frequency (the proportion of downsampling iterations in which a metabolite was significant). Dataset and metabolite identity were included as random intercepts, and the dispersion parameter was allowed to vary by dataset.

Term	Estimate	SE	z	P-value
(Intercept)	−2.8821	0.614	−4.69	2.67×10^{-06}
Reference effect size (diet coefficient)	10.8262	0.7104	15.24	$<2 \times 10^{-16}$
Effect size variability (SD of diet coefficient)	−15.8372	1.946	−8.14	4.00×10^{-16}
Pool size = 5	−1.6251	0.275	−5.91	3.44×10^{-09}
Pool size = 50	−0.3321	0.2573	−1.29	1.97×10^{-01}
Replicates remaining	0.3029	0.0412	7.35	1.96×10^{-13}
Pool size = 5 × Replicates remaining	0.0588	0.0517	1.14	2.56×10^{-01}
Pool size = 50 × Replicates remaining	0.0189	0.0492	0.39	7.00×10^{-01}

downsampling conditions, whereas redox metabolism and nucleotide turnover were more sensitive to reductions in pool size and replication. These differences suggest that some biological processes generate stronger or more coherent metabolomic signals, making them more resilient to experimental limitations. Importantly, this demonstrates that sampling decisions influence not only the number of detected metabolites but also which biological processes are ultimately inferred to be affected.

The pathways we identified as most strongly responsive to the high-sugar diet, lipid-associated metabolism (fatty acid oxidation and overflow) and central carbon metabolism, are broadly consistent with previously reported effects of high-sugar diets in *Drosophila*. Lipid remodeling under chronic high-sugar feeding, including changes in triglyceride and diacylglycerol content, has been directly characterized by lipidomic profiling of fat body, heart, and hemolymph (Tuthill et al. 2020), and increased TCA cycle activity together with depleted one-carbon metabolism has been reported under high-sugar conditions in a separate metabolomic study of fasting and refeeding (Wilinski et al. 2019). While these studies differ from ours in tissue scope and experimental

design, the direction of the pathway-level changes they report is consistent with what we observe here, lending support to the biological interpretability of the diet-responsive signals detected in this study.

Implications for metabolomics experimental design

Together, these findings have important implications for experimental design in metabolomics. First, very small pool sizes (eg $n = 5$) should be avoided when the goal is to detect moderate or subtle biological effects, as they substantially reduce sensitivity and increase variability in effect size estimates. Second, increasing pool size from 5 to 50 individuals provides a large improvement in both profile stability and signal detection, whereas further increases from 50 to 100 yield diminishing returns. Third, biological replication remains critical across all conditions, and reductions in replicate number lead to consistent losses in detectable signal regardless of pool size. Quantitatively, each additional biological replicate increased the number of detected significant metabolites by approximately 30% at pool size 5, compared with only 13% to 16% at pool sizes 50 and 100, indicating that the marginal statistical benefit of added replication is greatest when pool sizes are small. For researchers working within a fixed total sample budget, this suggests that prioritizing biological replication over pool size will generally yield larger gains in statistical power, particularly when replicate numbers are already low. These results indicate that optimal experimental design requires balancing pool size and replication, rather than prioritizing one at the expense of the other.

Several limitations of this study should be considered. Foremost, throughout this manuscript, the $n = 100$ pool size condition is used as a reference point for classifying detection outcomes (eg true/false positives and negatives) and for defining reference effect sizes; this reference condition reflects the pool size with the greatest input material and highest expected statistical power in this study, not an independently validated ground truth, and classifications made relative to it should be interpreted accordingly. Additionally, because each downsampling iteration is a resample drawn from a finite, shared pool of biological replicates rather than an independent sample, iterations within a given pool size and removal level are not statistically independent; this likely results in underestimated standard errors and anti-conservative P-values for the models reported here, though point estimates remain valid. This affects our 2 significance claims asymmetrically in a way that supports rather than undermines our conclusions: the zero-inflated Poisson interaction ($\chi^2 = 415.7$, $P < 2.2 \times 10^{-16}$) is significant by a margin far larger than plausible pseudoreplication-driven inflation could produce, and the non-significant beta-binomial interaction ($\chi^2 = 1.33$, $P = 0.51$) would, if anything, become more clearly non-significant under a more conservative correction. We therefore interpret the direction and relative magnitude of these effects with confidence, while treating the exact P-values as approximate rather than precise. Our analyses were conducted in *Drosophila melanogaster* using whole-body metabolomics, and the extent to which these findings generalize to other organisms, tissues, or targeted metabolomics approaches remains to be determined. More broadly, our findings are derived from a single extraction protocol, LC-MS platform (UPLC-QToF, C18/HILIC chromatography, positive and negative ionization), and data-processing pipeline. Given the substantial variability in sample preparation, chromatography, ionization, instrumentation, and processing pipelines across laboratories, the specific quantitative relationships we report between pool size, replication, and signal detection should not be assumed to generalize directly to other platforms. We expect the qualitative pattern, that pool size and replication both shape signal detection and trade off in their statistical benefits, to generalize broadly, since it reflects basic statistical properties of sampling rather than features specific to our workflow. The precise magnitude of these effects, however, will likely depend on the platform and pipeline used. In addition, pool sizes were evaluated at discrete levels (5, 50, and 100), and intermediate values may provide further insight into the shape of these relationships. While we observed clear differences in signal robustness across datasets, our results indicate that these differences arise from variation in both effect size distributions and measurement variability among metabolites. This suggests that the detectability of metabolite signals is not solely determined by experimental design but also by intrinsic properties of the measured metabolome, including the magnitude and stability of underlying biological effects.

Conclusion

Our results demonstrate that pooling is not a neutral technical decision, but a key determinant of both metabolomic measurement and biological inference. Pool size and replication jointly shape the detectability, stability, and interpretability of metabolomic signals, with the magnitude of these effects varying across datasets. As metabolomics continues to expand in evolutionary and systems biology, careful consideration of sampling design will be essential for ensuring robust and reproducible insights into the molecular basis of biological variation.

Data availability

All code and supplemental materials used in this study are publicly available on GitHub (https://github.com/hubertdl/Metabolomics_sampling_project_D.melanogaster) and archived in Zenodo (https://doi.org/10.5281/zenodo.22085589). Processed data and intermediate files required to reproduce all analyses and figures are included in the repository. Raw metabolite abundance data and sample metadata for both experiments are available at Zenodo (https://doi.org/10.5281/zenodo.21461823) and MassIVE (https://doi.org/doi:10.25345/C5F47H81 V).

Supplemental material available at G3 online.

Acknowledgments

We thank Jaewoo Choi and the Oregon State University Mass Spectrometry Center (USDA # 2024-70410-43767).

Funding

This work was funded through the National Institute of General Medical Sciences, Maximizing Investigator's Research Award number: NIH R35GM155286.

Conflicts of interest

The authors declare no competing interests.

Literature cited

Ayroles JF *et al.* 2009. Systems genetics of complex traits in *Drosophila melanogaster*. Nat Genet. 41:299–307. https://doi.org/10.1038/ng.332.

Bates D, Mächler M, Bolker BM, Walker SC. 2015. Fitting linear mixed-effects models using lme4. J Stat Softw. 67:1–48. https://doi.org/10.18637/jss.v067.i01.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Series B. 57:289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.

Carvalho J, Faria R, Butlin RK, Sousa VC. 2023. poolHelper: an R package to help in designing Pool-Seq studies. Methods Ecol Evol. 14:2300–2307. https://doi.org/10.1111/2041-210X.14185.

Choi J *et al.* 2019. Tissue-specific metabolic profiles after prolonged cardiac arrest reveal brain metabolome dysfunction predominantly after resuscitation. J Am Heart Assoc. 8:e012809. https://doi.org/10.1161/JAHA.119.012809.

Clish CB. 2015. Metabolomics: an emerging but powerful tool for precision medicine. Cold Spring Harb Mol Case Stud. 1:a000588. https://doi.org/10.1101/mcs.a000588.

Diz AP, Truebano M, Skibinski DO. 2009. The consequences of sample pooling in proteomics: an empirical study. Electrophoresis. 30:2967–2975. https://doi.org/10.1002/elps.200900210.

Erkosar B *et al.* 2023. Evolutionary adaptation to juvenile malnutrition impacts adult metabolism and impairs adult fitness in *Drosophila*. Elife. 12:e92465. https://doi.org/10.7554/eLife.92465.

Hoffman JM *et al.* 2014. Effects of age, sex, and genotype on high-sensitivity metabolomic profiles in the fruit fly, *Drosophila melanogaster*. Aging Cell. 13:596–604. https://doi.org/10.1111/acel.12215.

Hubert D *et al.* 2025. Selection for early reproduction leads to accelerated aging and extensive metabolic remodeling in *Drosophila melanogaster*. Genome Biol Evol. 17:evaf082. https://doi.org/10.1093/gbe/evaf082.

Johnson CH, Ivanisevic J, Siuzdak G. 2016. Metabolomics: beyond biomarkers and towards mechanisms. Nat Rev Mol Cell Biol. 17:451–459. https://doi.org/10.1038/nrm.2016.25.

Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M. 2012. KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res. 40:D109–D114. https://doi.org/10.1093/nar/gkr988.

Kendziorski C, Irizarry RA, Chen KS, Haag JD, Gould MN. 2005. On the utility of pooling biological samples in microarray experiments. Proc Natl Acad Sci U S A. 102:4252–4257. https://doi.org/10.1073/pnas.0500607102.

Long AD, Hanson KM, Macdonald SJ. 2026. The illusion of polygenicity in pool-seq genetic mapping studies: insufficient power can mask simple genetic architectures. Genetics. 233:iyag068. https://doi.org/10.1093/genetics/iyag068.

Mackay TFC *et al.* 2012. The *Drosophila melanogaster* genetic reference panel. Nature. 482:173–178. https://doi.org/10.1038/nature10811.

Millington JW, Rideout EJ. 2018. Sex differences in *Drosophila* development and physiology. Curr Opin Physiol. 6:46–56. https://doi.org/10.1016/j.cophys.2018.04.002.

Nicholson JK, Lindon JC, Holmes E. 1999. 'Metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli. Xenobiotica. 29:1181–1189. https://doi.org/10.1080/004982599238047.

Oksanen J. 2022. vegan: community ecology package. R package version 2.6-4. https://CRAN.R-project.org/package=vegan. Accessed March 13, 2026.

Phillips MA *et al.* 2022. Combining metabolomics and experimental evolution reveals mechanisms underlying longevity differences in *Drosophila melanogaster*. Int J Mol Sci. 23:1067. https://doi.org/10.3390/ijms23031067.

Simpson GL. 2022. Permute: functions for generating restricted permutations. R package version 0.9-7. https://cran.r-project.org/package=permute. Accessed March 13, 2026.

Tautenhahn R, Patti GJ, Rinehart D, Siuzdak G. 2012. XCMS online: a web-based platform to process untargeted metabolomic data. Anal Chem. 84:5035–5039. https://doi.org/10.1021/ac300698c.

Tuthill BF II, Searcy LA, Yost RA, Musselman LP. 2020. Tissue-specific analysis of lipid species in *Drosophila* during overnutrition by UHPLC-MS/MS and MALDI-MSI. J Lipid Res. 61:275–290. https://doi.org/10.1194/jlr.RA119000198.

Vecchié D, Anholt RRH, Mackay TFC, De Luca M. 2026. Selection for postponed senescence in *Drosophila melanogaster* reveals distinct metabolic aging trajectories modifiable by lisinopril. Aging Cell. 25:e70375. https://doi.org/10.1111/acel.70375.

Wilinski D *et al.* 2019. Rapid metabolic shifts occur during the transition between hunger and satiety in *Drosophila melanogaster*. Nat Commun. 10:4052. https://doi.org/10.1038/s41467-019-11933-z.

Xia J, Wishart DS. 2010. MetPA: a web-based metabolomics tool for pathway analysis and visualization. Bioinformatics. 26:2342–2344. https://doi.org/10.1093/bioinformatics/btq418.

Zhao X *et al.* 2022. The metabolome as a biomarker of aging in *Drosophila melanogaster*. Aging Cell. 21:e13548. https://doi.org/10.1111/acel.13548.

Zhou S *et al.* 2020. Systems genetics of the *Drosophila* metabolome. Genome Res. 30:392–405. https://doi.org/10.1101/gr.243030.118.

Editor: G. Morota