Batching inflates error rates in microbial colonization of *Caenorhabditis elegans*: an empirical investigation of the statistical properties of count-based data

Megan Taylor[1], Nic Vega[1,2]

1 Biology Department, Emory University, Atlanta GA 3033

2 Physics Department, Emory University, Atlanta GA 30322

Correspondence: Nic Vega, nvega@emory.edu

Keywords: Central limit theorem, biological averaging, finite sample statistics, *Caenorhabditis elegans*

Abstract

Heterogeneity is ubiquitous across individuals in biological data, and sample batching, a form of biological averaging, inevitably loses information about this heterogeneity. The consequences for inference from biologically averaged data are frequently opaque. Here we investigate a case where biological averaging is common - count-based measurement of bacterial load in individual *Caenorhabditis elegans* - to empirically determine the consequences of batching. We find that both central measures and measures of variation on individual-based data contain biologically relevant information that is useful for distinguishing between groups, and that batch-based inference readily produces both false positive and false negative results in these comparisons. These results support the use of individual rather than batched samples when possible, illustrate the importance of understanding the distributions of individual-based data in experimental systems, and indicate the need to consider effect size when drawing conclusions from these data.

Introduction

A fundamental concern in data analysis is that of statistical robustness – sensitivity of testing to violation of the underlying assumptions. Robustness is inherently a property of the test applied and the data on which it is used; in particular, robustness of parametric testing against violations of normality in a common concern, as the normal distribution is rarely a good fit to real data (Micceri 1989; Sorrentino 2010; Blanca *et al.* 2013; Mar 2019). Fortunately, many common parametric tests are at least somewhat robust to departures from normality (Rasmussen 1986; Sawilowsky and Blair 1992; Knief and Forstmeier 2021) when other requirements of these tests are met (Erceg-Hurn and Mirosevich 2008), and sturdy non-parametric alternatives are widely implemented (Rasmussen 1986; Potvin and Roff 1993). Further, the central limit theorem (CLT) is commonly invoked to permit the normality assumption for average values. However, the CLT is an asymptotic approximation of the behavior of sample means as the number of measurements becomes large. Formally, the classical CLT does not hold exactly when sampling from a finite population (the finite population CLT instead applies) (Plane and Gordon 1982; Bellhouse 2001), and practically, the definition of "large" depends on the distribution of values within the population being sampled (Smith and Wells 2006; CHANG *et al.* 2008).

Sample pooling or batching of individuals, a form of population averaging, is common practice in some areas of biological research. Sample pooling is used to minimize cost and effort in first-pass analysis of population-level frequencies, for example in population genetics (Dubreuil *et al.* 1999; Lynch *et al.* 2014), vector-based transmission (Ebert *et al.* 2010), and surveillance for contaminants and transmissible pathogens in environmental and agricultural systems (Arnold *et al.* 2011; Bignert *et al.* 2014). In epidemiological surveying, analysis of pathogen load in pooled samples is often used to increase efficiency of screening (Caudill 2010); this approach entered the public eye during the COVID-19 pandemic (Deckert *et al.* 2020). In laboratory experiments, it is sometimes desirable to pool multiple samples when the amount of biological material generated per sample is small (Kendziorski *et al.* 2005; Mary-Huard *et al.* 2007). For this reason, batch-based measurements are common in studies of host-microbe associations in small organisms such as the nematode *Caenorhabditis elegans* (Walker *et al.* 2022). In these experiments, "batches" are created by collecting groups of individual hosts (usually 5-50 hosts per batch) into individual volumes for mechanical disruption and quantification, often by plating on solid agar to determine colony forming unit (CFU) counts. The resulting counts are then divided by the number of hosts in the batch to arrive at an estimate for individual bacterial load. This approach can be used to "average over" heterogeneity in individual microbial burden.

Treating batch-based or pooled data in this way amounts to an assumption that heterogeneity among samples is "error" which can and should be smoothed via averaging (Churchill and Oliver 2001; Han *et al.* 2004; Lamichhane *et al.* 2017). The biological averaging assumption indicates that the value obtained from a pooled sample will be comparable to the arithmetic mean of values in the samples within that pool (Mary-Huard *et al.* 2007). Loss of information on inter-individual variation within the batch inevitably follows (Bignert *et al.* 1993). When the individual data are normally or log-normally distributed, it is possible to estimate the inter-individual variance from the variance of pooled data (Caudill 2010); however, this is rarely the case in biological data in general (Micceri 1989), and is very far from being true in data for host-microbe associations.

For these reasons, there has been increasing emphasis on the utility of individual measurements. Batching can alter the dimensionality of microbiome communities (Tsilimigras and Fodor 2016), change

measurements of diversity (Rodríguez-Ruano *et al.* 2020) and bias measures of total microbial load (Taylor *et al.* 2022). Further, there is increasing evidence that, even within isogenic and synchronized populations of model organisms, there are biologically significant and replicable patterns of inter-individual heterogeneity (Baeriswyl *et al.* 2009; Gruber *et al.* 2009; Chen *et al.* 2013; Diaz and Viney 2014; Kinser *et al.* 2021), indicating that much of this variation is likely to be information rather than noise or "error".

Here we quantify the effects of batching on measurements of host-microbe association in the *C. elegans* model, when data are given as counts (colony forming units, or CFUs) per individual. We demonstrate that batching consistently inflates false-positive rates in between-group comparisons, and that inflation of false-negative rates is possible under specific conditions. We investigate the structure of microbial load data for this model host to determine the data properties that allow for inflation of error rates, and clarify what information is relevant for determining between-group differences in measurements of individual hosts. Based on these results, we provide simple recommendations for experimental design, considering the statistical properties of microbial load data and typical within- and between-group heterogeneity for this host specifically.

Batching skews inferred microbial loads and inflates false positive rates

Two major issues for batch-based data are: (1) batching is an average (arithmetic mean) over individuals, which inherently involves some loss of information about the distribution of individual observations and (2) it is difficult to interpret the arithmetic mean as a population measure if data are not symmetric and centrally distributed. This can be demonstrated experimentally by comparing single-worm colonization data to batch-inferred CFU/worm data from the same experiment. For example, in N2 worms colonized with *Staphylococcus aureus* Newman, batching pulls estimated CFU/worm toward the upper quantiles of the distribution, and this effect is more marked in larger batches (20 individuals/batch vs 5 individuals/batch, **Figure 1A**). This reflects the right-hand skew in the single-worm data (median<mean; here, median=5,500 and mean=20,884).

We can assess the practical importance of this effect in simulations, by drawing from single-worm CFU data to create simulated “batch digests”. Here we show two runs of the same experiment, performed on different days (n=48 worms per run, **Figure 1B**). In these experiments, N2 adults were colonized on NGM agar plates with *S. enterica* LT2 for 48 hours before assessing colonization. The single-worm data are not significantly different day to day (t-test p=0.93, Mann-Whitney U p=0.37), but the distribution is slightly different across samples. On both days, the data have a right-hand skew (median < mean), but the mean values and the degree of skew are not perfectly identical across replicates (day 1 mean 27,497 and median 3,200; day 2 mean 29,207 and median 1,600).

We can perform a series of bootstraps (resampling with replacement) on these data to determine whether batch digests would have an impact on assessment of day-to-day variation. When we resample these raw data to create simulated batches of different sizes (5, 10, 20, or 50 worms per batch; n=25 simulated batch digests for each batch size), we find that our simulated batch data **(Figure 1C-E)** behave as expected from **Figure 1A**. Batching averages across individual hosts, and in accordance with expectations from central limit theorem, the variance between measurements decreases and the inferred CFU/worm is pulled toward the arithmetic mean of the individual-worm data, which is located

toward the upper quantiles of the right-skewed distribution. Further, day to day differences in single-worm data clearly affect the behavior of the distributions as batch sizes increase. Comparison of CFU/worm across replicates (days) in bootstrap-based simulated batch data (**Figure 1C-E**) indicated that differences between replicates are amplified as batch size increases (t-test batch 5 p=0.61, batch 10 p=0.95, batch 20 p=0.12, batch 50 p=0.0997; Mann-Whitney U batch 5 p= 0.07, batch 10 p= 0.29, batch 20 p= 0.8, batch 50 p= 0.005). The situation was not improved by instead comparing $\log_{10}$-transformed counts (t-test batch 5 p=0.23, batch 10 p=0.2, batch 20 p=0.013, batch 50 p=0.004; Mann-Whitney U tests unchanged). For these data, increasing batch size altered the conclusions of analysis, indicating with increasing frequency that there is a significant difference between replicates.

As in previous work from this lab (Taylor *et al.* 2022), batch digests introduced spurious differences between replicates in the *Salmonella* CFU/worm data (false positive). However, a larger batch size was required to produce spurious differences here (using the *Salmonella* data set) than in the previous example, which used single-worm colonization data for two different bacteria (MYb53 and MYb120 from the *C. elegans* native microbiome (Dirksen *et al.* 2016)). This raised the question of why, and what features of these data are responsible for this phenomenon. Understanding the origins of batch-based errors in these data will be important for determining when batching is and is not admissible.

Quantifying run-to-run heterogeneity in CFU/worm count data

At some magnitude, we will want to consider a shift in the distribution of values as a true difference between groups, whether or not the arithmetic means differ. This raises two questions: First, how much of a change in average and in distribution is expected across replicates of the same experiment (as any between-group differences smaller than this will not be interesting), and what is the best way to quantify this expectation? Second, how does the distribution of the data alter the error structure of between-group comparisons, and what does this mean for the ability to accurately detect differences?

To address the first question, we re-analyzed existing data to determine the distributional properties of CFU/worm data. A large set of single-worm measurements collected as part of a study on heterogeneity in a minimal native worm microbiome (Taylor and Vega 2021) was filtered to remove all replicates containing less than 18 individuals to ensure sufficient sampling of inter-individual heterogeneity, leaving data representing eight host genotypes with at least two days of data for comparisons. Mono-colonization data for *Salmonella enterica, Staphylococcus aureus, and Microbacterium oxydans* in N2 adults were added to this data set, for a total of 11 experimental conditions with replicates performed on separate days **(Figure 2A**).

Measures of central tendency (mean, median, Cohen's D) and coefficients of variation for CFU/worm data were more similar day-to-day across replicates of the same experiment ("same") than across experimental conditions ("different", **Figure 2B**), suggesting that these statistics contain information useful for distinguishing between groups. This comparison is not significant in the higher moments of the distributions (skewness and kurtosis, although these measures are known to be sensitive to outliers, and Hogg's Q1, a more robust measure of skew, shows a marginally significant difference) (Hogg 1974; Hill and Dixon 1982). These higher moments may therefore be more usefully considered as general properties of count-based CFU/worm data. Overall, these data tend strongly to be heterogeneous (0.5<CV<1.5), asymmetric (nearly always skewed to the right) and heavy-tailed (positive excess kurtosis);

multiple modes (two, occasionally three) are often, but not always, apparent in a given sample (**Figures S1-S2 and Table S1, File S1**).

Data heterogeneity and false positives

It is easy to demonstrate (**Figure 1; Figure S3 and Tables S2-S4, File S1**) that batching, a form of biological averaging, produces data that behave mathematically as an average with all the properties that entails. When two populations have the same mean, even when the distributions of individuals within these populations are very different, batch-based sampling can result in no difference between samples being detected (false negative; **Table S4, File S1**). However, in the real data examples so far, batching tends to produce inflated false positive rates. While sampling and small number effects produce an expected rate of false positives in simulated data with a common mean (**Table S4, File S1**), sampling of variable and skewed data is insufficient to produce *elevated* false-positive rates. It is illustrative to look at the properties of real data to isolate features that will promote false positives.

First, in the simple simulations (**SI**), all data sets have exactly the same mean; the null hypothesis of the t-test (no difference in means) is true even when the null hypothesis of the Mann-Whitney test (no difference in the sum of ranks) is not. For real data, even under ideal conditions, this is unlikely to be precisely true. It is reasonable that replicates of the same process, for example on different days, will differ slightly due to inescapable run-to-run differences (small shifts in timing, a few percent pipetting error, etc). In the *Salmonella* colonization data, for example, average colonization on day 1 was 94.1% of the average colonization on day 2 two weeks later (81.2% when data are zero-censored) and 53% of the average colonization from a third run almost a year later with a different operator (**Figure 2A**, all Mann-Whitney U tests $p > 0.05$). This is typical of these data; normalized distance between average values across replicates showed a high-density region around 0.25 with few comparisons above this value, corresponding to a ~50% difference when these averages are similar (**Figure 2B**).

Based on these results, we returned to simulations to gain insight into inflation of false positives in batch data. As before (**Figure S3, File S1**), these simulations used the Beta(1,5) distribution as a baseline; additionally, these simulations assumed that there is day-to-day variation in the parameters from which the data are generated. Specifically, new parameters were generated for each data set *i*=(1,2) in each run as $(\alpha_i, \beta_i) = \left(\alpha\left(1 + U(-0.1,\ 0.1)\right),\ \beta\left(1 + U(-0.1,\ 0.1)\right)\right)$. The between-set distances for moments of the single-"worm" data in these simulations were consistent with those observed between replicates in real data (**Figure 2B, Figure 3**). Also consistent with results from real data, in simulations with day-to-day variation, batching promoted false-positive results in pairwise comparisons (**Table 1**). The increased rate of false positives with batching was not due to dispersal of the means in batch data (**Figure S4, File S1**). Nor did batching strongly affect the between-run distance in skew or any properties of the excess kurtosis (**Figure S4, File S1**). The nature of the skewness changed with batching, as the positive skew seen in the individual data decreases as batch size increases. However, the largest change was in the variation (shown as CV, **Figure S4, File S1**). Consistent with expectations from previous sections and from the central limit theorem, variation decreased as batch size increased. As data points clustered closer to the daily mean, run-to-run variation in that mean (here, due to stochastic variation in the parameters of the underlying distribution) was interpreted as a real difference between groups. Notably, the t-test and the Mann-Whitney test behaved similarly; the parametric test is reacting to the

collapse of uncertainty about the value of the mean, and the nonparametric test is reacting to the coalescence of the distribution of data points toward the mean.

The higher moments of the data affected the results of between-"run" comparisons. The impact of kurtosis (tail weight) was minimal for these samples, as can be seen by comparing the effects of batching when parameters are centered around Beta(1,5) vs. Beta(0.5, 2.5) (**Table 1**). Likewise, the direction of skew had minimal effect, as when the values of α and β were reversed (**Table 1**). However, the magnitudes of both variance and skewness were important, and to some extent it was possible to trade one for another; symmetric distributions were more likely to produce significant differences at a given batch size than skewed distributions with similar variance (compare for example Beta(1,5) with var~0.02 and skewness ~1.18 vs. Beta(5,5) with var~0.02 and skewness 0, **Table 1**). The reason for this is surprisingly straightforward: for symmetric distributions, the data points are clustered closer to the true within-day mean (which is not identical in these simulations due to day-to-day variation in underlying parameters) than for asymmetric distributions. This is true even in individual data but is exacerbated by batching (**Figure S5, File S1**). This result recalls the Berry-Esseen theorem in statistics, which indicates that the rate at which the arithmetic mean converges to a normal distribution is a function of the sample size (as $n^{-1/2}$) as well as the ratio of the third and second absolute normalized moments of the distribution of the random sample (Berry 1941; Esseen 1942; Korolev and Shevtsova 2010).

Introducing experimental realism: False-positive rates and day-to-day variation in data with replicates

Real experimental data have additional structure. Data are generally combined across multiple replicates, and most experiments are not large enough to create many batches of 20-50 individuals per condition. Simulations were therefore performed to assess the effects of replication and sampling limitation on false-positive rates (**Table S5 and Figures S6-S7, File S1**). Briefly, these simulations indicated that combining replicates across sampling days did not decrease false-positive inflation due to batching.

Data heterogeneity and false negatives

If false positives can be obtained when heterogeneous data from the same base distribution are regressed toward their mean (**Figure S4, File S1**), false negative rate inflation can occur when batching is performed on samples from different distributions but with similar means. We concluded earlier that averages were overall more similar across days within an experiment than across experiments. However, there is substantial overlap (**Figure 2**), and our data set includes several instances where data with very different distributions have similar means.

Based on previous results, we expected to get false negative inflation in batching if data sets were differently variable and/or skewed but had sufficiently similar means. An example was provided by two replicates from the multispecies community data representing host genotypes Bristol N2 (6/21/19; n=21; average 8,277 CFU/worm) and *dbl-1* (6/26/19; n=33; average 7,799 CFU/worm). Here the N2 replicate was moderately variable and skewed (CV 0.88, skewness 0.89) as compared with the more heterogeneous *dbl-1* data (CV 2.06, skewness 3.37) (**Figure 3A**), and the two raw data sets were distinguishable from one another (Mann-Whitney U p=0.03). We resampled these data to simulate batching as before (n=21 batches per batch size, to match the smaller of the original data sets, with

229 batch size up to 20 worms/batch due to the small number of individuals, **Figure 3 B-D**). Over 10,000 runs
230 of this resampling, we observed $p<0.05$ in (38.24%, 19.70%, 16.85%) of runs at batch sizes of 5, 10, and
231 20 respectively. As a second example, we compared *S. enterica* colonization (day 2, n=25, mean 33,880,
232 CV 3.03, skewness 4.22) vs. minimal microbiome colonization in N2 adults (day 2/13/19, n=24, mean
233 34,069, CV 0.49, skewness –0.07). Comparison of the original data sets (**Figure 3E**) indicated a difference
234 between conditions (Mann-Whitney U $p= 0.0009$). When we resampled these data as before (10,000
235 runs, each run with n=24 data points at each batch size to match the smaller data set, **Figure 3F-H**), we
236 again saw false-negative rate inflation; $p<0.05$ is achieved in 93.41%, 51.11%, and 8.21% of runs at batch
237 sizes of 5, 10, and 20 respectively. False negative rates were therefore inflated by batching, but unlike
238 false positive rates, there was not a simple monotonic relationship with batch size.

239 It is reasonable to ask whether these results are likely to recur in different samples from the same
240 process. When we introduced variation by allowing noisy resampling, such that each data point was
241 drawn with +/-5% error in counts, fraction of runs with $p<0.05$ was (1.000, 0.928, 0.503, 0.085) at batch
242 sizes (1,5,10,20). Increasing noise to +/-10% did not change these rates appreciably (1.0000, 0.930,
243 0.504, 0.089). Adding sampling noise therefore did not prevent inflation of false negative rates. Day-to-
244 day variation in samples could therefore reasonably affect false negatives, but the direction of effect is
245 likely to be a matter of happenstance. If the true distributions of conditions A and B have vaguely similar
246 means but different variance and skew, across any number of days of sampling, false negatives will be a
247 potential issue in batch data when data set averages happen to be sufficiently similar. Importantly, if
248 only batch data are taken and the underlying distribution is not known, there is no obvious way to
249 determine whether this is likely to have occurred for a given experiment.

250

251 **Discussion**

252 Experimentalists are regularly warned by statisticians that if an inappropriate test is used, the
253 conclusions of data analysis may be incorrect (Bridge and Sawilowsky 1999; Othman *et al.* 2004; Vickers
254 2005; Erceg-Hurn and Mirosevich 2008; Zuur *et al.* 2010; Forstmeier *et al.* 2017). The results shown here
255 are an example of the empiricist's addendum – if the data are problematic, the test is the lesser
256 problem. Commonly, the data themselves are low-quality due to sampling issues, bias, mislabeling,
257 contamination, etc. (Cortes *et al.* 1995; Clausen and Willis 2021; Schloss) In the case illustrated here,
258 data quality issues are not necessary to produce spurious results. Instead, this case belongs to the family
259 of problems where sampling the wrong thing – or measuring the right thing in the wrong way –
260 produces measures unsuitable for answering the question being asked (Riniolo and Porges 2000).

261 For count data with the structure observed in CFU/worm data sets, error rates are inflated by batching.
262 For both false positives and false negatives, errors arise because batching produces data points
263 representing the arithmetic mean of the individuals included in the batch. As batch size increases, the
264 central limit theorem indicates that the distribution of these data will lose information about the original
265 distribution and become centered around the sample mean. When distributions are very similar, as in
266 experimental replicates, run-to-run variation can create variation in sample means that may be mistaken
267 for a real difference. When distributions are different, if the means of two groups happen to be similar,
268 batching-associated loss of information about the distributions of the data can lead to failure to
269 distinguish between groups. Critically, if information about the distribution of individuals is missing and

270 particularly when effect size is small, there is no obvious way to determine whether a given comparison
271 is likely to experience inflated rates of error.

272 The data shown here illustrate the fact that batching, by producing a biological average, generates
273 values that converge toward the arithmetic mean according to the central limit theorem. This should not
274 be interpreted to mean that these batch-inferred values are distributed as Gaussian – even without
275 testing for normality, it is visually obvious that they are not (**Figures 1, 3**). It is known that convergence
276 to normality as sample size increases should be slow for variable and skewed data (Korolev and
277 Shevtsova 2010). Further, other empirical work has demonstrated that discreteness of data and
278 skewness and kurtosis of a distribution dramatically increase the sample size required for the population
279 of means to approach normality (Smith and Wells 2006; CHANG *et al.* 2008). For log-scale, skewed, and
280 heavy-tailed data typical of bacterial load measurements, even at large batch sizes the distribution of
281 the population of means is clearly contaminated (**Figure 3**), with marked asymmetry and even
282 multimodality. We do not attempt to describe this progression formally. For the purposes of the present
283 analysis, it is sufficient to note that the population of means loses biologically relevant information
284 about the underlying heterogeneity among individuals, while failing to gain the advantages conferred
285 when normality can be assumed.

286 As part of this analysis, we describe common features of CFU/worm data and confirm measures that are
287 useful for distinguishing between groups. The lower moments of the data (mean and variance) and the
288 central quantiles are most strongly associated with specific combinations of host strain and microbial
289 colonist(s), suggesting that both central measures and measures of variation contain information that is
290 useful for distinguishing biologically relevant groups. Further, these data are almost always right-hand
291 skewed and frequently heavy-tailed – these appear to be general features of this data type. However,
292 due to the relatively small *n* within replicates in these data, it is not at present possible to rule out the
293 possibility that specific measures of asymmetry and/or tail weight might also contain distinguishing
294 information in larger samples. Hogg's Q1 and Q2, which were designed as stable measures of skewness
295 and tail weight, are promising for sufficiently large samples where the 5% and 95% quantiles can be
296 repeatably described. Interestingly, single-species colonization by the two pathogens in this data set is
297 associated with higher CV and broader range than colonization by our minimal microbiome, as well as
298 with a more pronounced multi-modality; additional data, including information on single-species
299 colonization by commensals, will be required to determine generality of this observation.

300 For the experimentalist, these results provide guidelines to minimize the risk of false-negative and false-
301 positive comparisons. For count-based data, heterogeneity within even populations of isogenic,
302 synchronized worms means that individual-worm data should be preferred over batch-based data if
303 possible. Effect size should be considered along with statistical significance when making comparisons
304 between groups (Sullivan and Feinn 2012), as small but statistically significant differences are likely to be
305 enriched in false-positive comparisons, particularly if batching cannot be avoided. The underlying
306 distribution of the data (variation and centrality/skewness in particular) and the run-to-run variation will
307 be important – the results shown here (**Figure 2**) indicated typical run-to-run variation of around 50% in
308 average CFU/worm (normalized distance of means <0.25, or Cohen's *d* <1.3), suggesting that effect sizes
309 smaller than this should be disregarded when comparing batch CFU data.

310 More broadly, these results suggest a need for the experimentalist in this area to consider carefully the
311 question that is being asked. First, is the question about what happens on average? Batching by its

nature averages across individuals. If the question is not about the population average (arithmetic mean), batch-based data are perhaps not best suited to provide an answer. There are many cases where the arithmetic mean is not the most informative quantity to describe a process. For example, "bet-hedging" in a fluctuating environment is canonically represented by geometric mean fitness (Starrfelt and Kokko 2012), and transmission of infectious disease is often driven by "superspreading" with a long-tailed distribution of secondary infections (Woolhouse *et al.* 1997; Galvani and May 2005; Lloyd-Smith *et al.* 2005; Paull *et al.* 2012; VanderWaal and Ezenwa 2016). Further, instability of the arithmetic mean in asymmetric and "lumpy" data has been historically described (Ansell 1973; Hill and Dixon 1982). Colonization data are heterogeneous, often ranging over two or more orders of magnitude across individuals, and can show multiple modes (**Figure S1, File S1**); the arithmetic mean is heavily influenced by the largest values, and batch-inferred values for bacterial load mostly contain information about the upper quantiles of the distribution (**Figure 1**).

Second, is it acceptable not to observe individual heterogeneity? Even if we assume the average is the correct summary statistic for a given question, variation may still be relevant. We observe (**Figure 2**) that heterogeneity in bacterial load has characteristic features that are repeatable across replicates but which change across experiments, inconsistent with the idea that this heterogeneity is simply attributable to sampling noise. For example, colonization by this eight-species minimal community tends to (but does not always) produce a heavy-bodied "stingray"-shaped distribution of bacterial loads, as compared with the more extended distributions sometimes observed for mono-colonization in this data set (**Figure 2**). As stated above, measures of variation among individuals appear to be characteristic of a given colonization condition, indicating that sacrificing individual heterogeneity will lead to loss of information that is useful for distinguishing between groups.

It may be useful to extend these analyses to other common data types with different structures. In *C. elegans*, count data are commonly taken in studies of fecundity (offspring per hermaphrodite), and mortality data are taken as counts and presented as fractional survival over time. Both data types are well known to have issues related to hidden heterogeneity within populations (Wu *et al.* 2006; Suda *et al.* 2009; Baeriswyl *et al.* 2009; Gruber *et al.* 2009; Pincus *et al.* 2011; Chen *et al.* 2013; Diaz and Viney 2014; Kinser *et al.* 2021; Travers *et al.*). These data also have characteristic complications, notably censoring, with well-known relevance for analysis (Escobar and Meeker 1992; Gijbels 2010). If (as seems probable) variation in these data contains information about the underlying biology, understanding inter-individual and inter-sample heterogeneity will be important not only statistically, to allow appropriate choice and interpretation of analyses, but also conceptually, to allow experimental design that captures this information and conclusions that take heterogeneity into account. Understanding the nature of heterogeneity in data therefore presents an opportunity to maximize the potential of the experimental model by improving accuracy and repeatability of results.

Acknowledgements

This research was funded by Emory University and NSF [grant number PHY2014173]. The authors would like to thank C. LaRock for generously providing us with bacterial strains, and L. Waller and L. Morran for discussions and feedback on this manuscript.

353 **Methods**

354 *Strains and materials*

355 *C. elegans* (Bristol N2; *dbl-1(nk3) V*, CGC NU3) were obtained from the *Caenorhabditis* Genetic Center,
356 which is funded by NIH Office of Research Infrastructure Programs (P40 OD010440). *Salmonella enterica*
357 LT2 was obtained from ATCC (ATCC 700720). *Staphylococcus aureus* MSSA Newman was a gift from Chris
358 LaRock (Emory University).

359 *Worm colonization and CFU counting*

360 Nematodes were grown, maintained, and manipulated using standard techniques (Stiernagle 2006).
361 Briefly, breeding stocks were maintained on NGM plates + OP50 at 25°C and synchronized using a
362 standard bleach/NaOH protocol where eggs were allowed to hatch out in M9 worm buffer overnight
363 (~16h) with shaking (200 RPM) at 25°C. Starved L1 larvae were transferred to 10cm NGM plates
364 containing lawns of *E. coli pos-1* RNAi and incubated at 25°C for 3 days to produce reproductively sterile
365 adults. Worms were then transferred to liquid S-medium + 200 µg/ml gentamicin + 50 µg/ml
366 chloramphenicol + 2X heat-killed OP50 (to trigger feeding) for 24 hours, resulting in largely germ-free
367 adults. Adult worms were washed via sucrose floatation before colonization.

368 For colonization, *S. aureus* and *S. enterica* were grown from glycerol stocks to stationary phase
369 overnight in 1 mL LB media at 37°C with shaking at 200 RPM. 10 cm NGM agar plates were seeded with
370 100 µL of culture, which was distributed with plating beads to cover the surface and incubated 24 hours
371 at 25°C to form lawns. Adult worms prepared as described above were aliquoted onto plates in 25-50 µL
372 drops of buffer (M9 worm buffer + 0.1% Triton X-100 to prevent sticking; ~500-1000 worms/plate) and
373 incubated at 25°C in the dark for 24 hours (*S. aureus*) or 48 hours (*S. enterica*) to allow colonization.
374 Colonized worms were prepared for mechanical disruption as previously described, then sorted
375 individually or in batches into wells of a deep 2 mL 96-well plate (Axygen) using a BioSorter large object
376 sorter fitted with a 250 µM FOCA (Union Biometrica). Sorted worms were disrupted in 96-well plate
377 format as previously described (Taylor and Vega 2021; Taylor *et al.* 2022).

378 Indicated sets of CFU/worm data were previously published, and data and methods are available from
379 the indicated manuscript (Taylor and Vega 2021).

380 *Data analysis and simulations*

381 All data analysis and simulations were performed in R 4.1.3 using structures from base R (R Core Team
382 2022) and tidyverse (Wickham *et al.* 2019). Differences between groups were assessed using functions
383 *t.test()* and *wilcox.test()* from *stats* to perform two-sided testing against a null hypothesis of no
384 difference. Calculation of higher moments of data used functions *skewness()* and *kurtosis()* from
385 package *e1071* (Meyer *et al.* 2021). Gaussian mixture models were fitted to $\log_{10}$-transformed data using
386 package *mclust* (Scrucca *et al.* 2016). Plotting was carried out using functionality from packages *ggplot2*
387 (Wickham 2016 p. 2), *cowplot*, (Wilke 2020) and *ggpubr* (Kassambara 2020).

388 *Data availability*

389 Strains are available upon request. All code and data used in this manuscript are available on GitHub
390 (https://github.com/veganm/WormCFUHeterogeneity2022).

References

Ansell M. J. G., 1973 Robustness of Location Estimators to Asymmetry. Journal of the Royal Statistical Society. Series C (Applied Statistics) 22: 249–254. https://doi.org/10.2307/2346926

Arnold M. E., J. J. Carrique-Mas, I. McLaren, and R. H. Davies, 2011 A comparison of pooled and individual bird sampling for detection of Salmonella in commercial egg laying flocks. Preventive Veterinary Medicine 99: 176–184. https://doi.org/10.1016/j.prevetmed.2010.12.007

Baeriswyl S., M. Diard, T. Mosser, M. Leroy, X. Manière, *et al.*, 2009 Modulation of aging profiles in isogenic populations of Caenorhabditis elegans by bacteria causing different extrinsic mortality rates. Biogerontology 11: 53. https://doi.org/10.1007/s10522-009-9228-0

Bellhouse D. R., 2001 The Central Limit Theorem Under Simple Random Sampling. The American Statistician 55: 352–357. https://doi.org/10.1198/000313001753272330

Berry A. C., 1941 The accuracy of the Gaussian approximation to the sum of independent variates. Trans. Amer. Math. Soc. 49: 122–136. https://doi.org/10.1090/S0002-9947-1941-0003498-3

Bignert A., A. Göthberg, S. Jensen, K. Litzén, T. Odsjö, *et al.*, 1993 The need for adequate biological sampling in ecotoxicological investigations: a retrospective study of twenty years pollution monitoring. Science of The Total Environment 128: 121–139. https://doi.org/10.1016/0048-9697(93)90215-R

Bignert A., U. Eriksson, E. Nyberg, A. Miller, and S. Danielsson, 2014 Consequences of using pooled versus individual samples for designing environmental monitoring sampling strategies. Chemosphere 94: 177–182. https://doi.org/10.1016/j.chemosphere.2013.09.096

Blanca M. J., J. Arnau, D. López-Montiel, R. Bono, and R. Bendayan, 2013 Skewness and Kurtosis in Real Data Samples. Methodology 9: 78–84. https://doi.org/10.1027/1614-2241/a000057

Bridge P. D., and S. S. Sawilowsky, 1999 Increasing Physicians' Awareness of the Impact of Statistics on Research Outcomes: Comparative Power of the t-test and Wilcoxon Rank-Sum Test in Small Samples Applied Research. Journal of Clinical Epidemiology 52: 229–235. https://doi.org/10.1016/S0895-4356(98)00168-1

Caudill S. P., 2010 Characterizing populations of individuals using pooled samples. J Expo Sci Environ Epidemiol 20: 29–37. https://doi.org/10.1038/jes.2008.72

CHANG H.-J., C.-H. WU, J.-F. HO, and P.-Y. CHEN, 2008 On Sample Size in Using Central Limit Theorem for Gamma Distribution. International journal of information and management sciences 19: 153–174.

Chen H., F. Zajitschek, and A. A. Maklakov, 2013 Why ageing stops: heterogeneity explains late-life mortality deceleration in nematodes. Biology Letters 9: 20130217. https://doi.org/10.1098/rsbl.2013.0217

Churchill G. A., and B. Oliver, 2001 Sex, flies and microarrays. Nat Genet 29: 355–356. https://doi.org/10.1038/ng1201-355

Clausen D. S., and A. D. Willis, 2021 Evaluating replicability in microbiome data. Biostatistics kxab048. https://doi.org/10.1093/biostatistics/kxab048

Cortes C., L. D. Jackel, and W.-P. Chiang, 1995 Limits on learning machine accuracy imposed by data quality, pp. 57–62 in *Proceedings of the First International Conference on Knowledge Discovery and Data Mining*, KDD'95. AAAI Press, Montréal, Québec, Canada.

433 Deckert A., T. Bärnighausen, and N. N. Kyei, 2020 Simulation of pooled-sample analysis strategies for
434 COVID-19 mass testing. Bull World Health Organ 98: 590–598.
435 https://doi.org/10.2471/BLT.20.257188

436 Diaz S. A., and M. Viney, 2014 Genotypic-specific variance in Caenorhabditis elegans lifetime fecundity.
437 Ecology and Evolution 4: 2058–2069. https://doi.org/10.1002/ece3.1057

438 Dirksen P., S. A. Marsh, I. Braker, N. Heitland, S. Wagner, *et al.*, 2016 The native microbiome of the
439 nematode Caenorhabditis elegans: gateway to a new host-microbiome model. BMC Biology 14:
440 38. https://doi.org/10.1186/s12915-016-0258-1

441 Dubreuil P., C. Rebourg, M. Merlino, and A. Charcosset, 1999 Evaluation of a DNA pooled-sampling
442 strategy for estimating the RFLP diversity of maize populations. Plant Molecular Biology
443 Reporter 17: 123–138. https://doi.org/10.1023/A:1007571101815

444 Ebert T. A., R. Brlansky, and M. Rogers, 2010 Reexamining the Pooled Sampling Approach for Estimating
445 Prevalence of Infected Insect Vectors. Annals of the Entomological Society of America 103: 827–
446 837. https://doi.org/10.1603/AN09158

447 Erceg-Hurn D. M., and V. M. Mirosevich, 2008 Modern robust statistical methods: An easy way to
448 maximize the accuracy and power of your research. American Psychologist 63: 591–601.
449 https://doi.org/10.1037/0003-066X.63.7.591

450 Escobar L. A., and W. Q. Meeker, 1992 Assessing Influence in Regression Analysis with Censored Data.
451 Biometrics 48: 507–528. https://doi.org/10.2307/2532306

452 Esseen C.-G., 1942 On the Liapunov limit error in the theory of probability. Ark. Mat. Astr. Fys. 28: 1–19.

453 Forstmeier W., E.-J. Wagenmakers, and T. H. Parker, 2017 Detecting and avoiding likely false-positive
454 findings – a practical guide. Biological Reviews 92: 1941–1968.
455 https://doi.org/10.1111/brv.12315

456 Galvani A. P., and R. M. May, 2005 Dimensions of superspreading. Nature 438: 293–295.
457 https://doi.org/10.1038/438293a

458 Gijbels I., 2010 Censored data. WIREs Computational Statistics 2: 178–188.
459 https://doi.org/10.1002/wics.80

460 Gruber J., L. F. Ng, S. K. Poovathingal, and B. Halliwell, 2009 Deceptively simple but simply deceptive –
461 Caenorhabditis elegans lifespan studies: Considerations for aging and antioxidant effects. FEBS
462 Letters 583: 3377–3387. https://doi.org/10.1016/j.febslet.2009.09.051

463 Han E.-S., Y. Wu, R. McCarter, J. F. Nelson, A. Richardson, *et al.*, 2004 Reproducibility, Sources of
464 Variability, Pooling, and Sample Size: Important Considerations for the Design of High-Density
465 Oligonucleotide Array Experiments. The Journals of Gerontology: Series A 59: B306–B315.
466 https://doi.org/10.1093/gerona/59.4.B306

467 Hill M., and W. J. Dixon, 1982 Robustness in Real Life: A Study of Clinical Laboratory Data. Biometrics 38:
468 377–396. https://doi.org/10.2307/2530452

469 Hogg R. V., 1974 Adaptive Robust Procedures: A Partial Review and Some Suggestions for Future
470 Applications and Theory. Journal of the American Statistical Association 69: 909–923.
471 https://doi.org/10.2307/2286160

472 Kassambara A., 2020 *ggpubr: "ggplot2" Based Publication Ready Plots*.

473 Kendziorski C., R. A. Irizarry, K.-S. Chen, J. D. Haag, and M. N. Gould, 2005 On the utility of pooling
474 biological samples in microarray experiments. Proceedings of the National Academy of Sciences
475 102: 4252–4257. https://doi.org/10.1073/pnas.0500607102

476 Kinser H. E., M. C. Mosley, I. B. Plutzer, and Z. Pincus, 2021 Global, cell non-autonomous gene regulation
477 drives individual lifespan among isogenic C. elegans, (J. Gruber, and J. K. Tyler, Eds.). eLife 10:
478 e65026. https://doi.org/10.7554/eLife.65026

479 Knief U., and W. Forstmeier, 2021 Violating the normality assumption may be the lesser of two evils.
480 Behav Res 53: 2576–2590. https://doi.org/10.3758/s13428-021-01587-5

481 Korolev V. Yu., and I. G. Shevtsova, 2010 On the Upper Bound for the Absolute Constant in the Berry–
482 Esseen Inequality. Theory Probab. Appl. 54: 638–658.
483 https://doi.org/10.1137/S0040585X97984449

484 Lamichhane S., U. K. Sundekilde, T. Blædel, T. K. Dalsgaard, L. H. Larsen, *et al.*, 2017 Optimizing sampling
485 strategies for NMR-based metabolomics of human feces: pooled vs. unpooled analyses. Anal.
486 Methods 9: 4476–4480. https://doi.org/10.1039/C7AY01465A

487 Lloyd-Smith J. O., S. J. Schreiber, P. E. Kopp, and W. M. Getz, 2005 Superspreading and the effect of
488 individual variation on disease emergence. Nature 438: 355–359.
489 https://doi.org/10.1038/nature04153

490 Lynch M., D. Bost, S. Wilson, T. Maruki, and S. Harrison, 2014 Population-Genetic Inference from Pooled-
491 Sequencing Data. Genome Biol Evol 6: 1210–1218. https://doi.org/10.1093/gbe/evu085

492 Mar J. C., 2019 The rise of the distributions: why non-normality is important for understanding the
493 transcriptome and beyond. Biophys Rev 11: 89–94. https://doi.org/10.1007/s12551-018-0494-4

Mary-Huard T., J.-J. Daudin, M. Baccini, A. Biggeri, and A. Bar-Hen, 2007 Biases induced by pooling samples in microarray experiments. Bioinformatics 23: i313–i318. https://doi.org/10.1093/bioinformatics/btm182

Meyer D., E. Dimitriadou, K. Hornik, A. Weingessel, and F. Leisch, 2021 *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*.

Micceri T., 1989 The unicorn, the normal curve, and other improbable creatures. Psychological Bulletin 105: 156–166. https://doi.org/10.1037/0033-2909.105.1.156

Othman A. R., H. j. Keselman, A. R. Padmanabhan, R. R. Wilcox, and K. Fradette, 2004 Comparing measures of the 'typical' score across treatment groups. British Journal of Mathematical and Statistical Psychology 57: 215–234. https://doi.org/10.1348/0007110042307159

Paull S. H., S. Song, K. M. McClure, L. C. Sackett, A. M. Kilpatrick, *et al.*, 2012 From superspreaders to disease hotspots: linking transmission across hosts and space. Frontiers in Ecology and the Environment 10: 75–82. https://doi.org/10.1890/110111

Pincus Z., T. Smith-Vikos, and F. J. Slack, 2011 MicroRNA Predictors of Longevity in Caenorhabditis elegans. PLOS Genetics 7: e1002306. https://doi.org/10.1371/journal.pgen.1002306

Plane D. R., and K. R. Gordon, 1982 A Simple Proof of the Nonapplicability of the Central Limit Theorem to Finite Populations. The American Statistician 36: 175–176. https://doi.org/10.1080/00031305.1982.10482823

Potvin C., and D. A. Roff, 1993 Distribution-Free and Robust Statistical Methods: Viable Alternatives to Parametric Statistics. Ecology 74: 1617–1628. https://doi.org/10.2307/1939920

R Core Team, 2022 *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rasmussen J. L., 1986 An evaluation of parametric and non-parametric tests on modified and non-modified data. British Journal of Mathematical and Statistical Psychology 39: 213–220. https://doi.org/10.1111/j.2044-8317.1986.tb00858.x

Riniolo T. C., and S. W. Porges, 2000 Evaluating group distributional characteristics: Why psychophysiologists should be interested in qualitative departures from the normal distribution. Psychophysiology 37: 21–28. https://doi.org/10.1111/1469-8986.3710021

Rodríguez-Ruano S. M., E. Juhaňáková, J. Vávra, and E. Nováková, 2020 Methodological Insight Into Mosquito Microbiome Studies. Frontiers in Cellular and Infection Microbiology 10.

Sawilowsky S. S., and R. C. Blair, 1992 A more realistic look at the robustness and Type II error properties of the t test to departures from population normality. Psychological Bulletin 111: 352–360. https://doi.org/10.1037/0033-2909.111.2.352

Schloss P. D., Identifying and Overcoming Threats to Reproducibility, Replicability, Robustness, and Generalizability in Microbiome Research. mBio 9: e00525-18. https://doi.org/10.1128/mBio.00525-18

Scrucca L., M. Fop, T. B. Murphy, and A. E. Raftery, 2016 mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models. R J 8: 289–317.

Smith Z., and C. Wells, 2006 Central Limit Theorem and Sample Size, in Kerhonkson, New York.

Sorrentino R. P., 2010 Large standard deviations and logarithmic-normality. Fly 4: 327–332. https://doi.org/10.4161/fly.4.4.13260

535 Starrfelt J., and H. Kokko, 2012 Bet-hedging—a triple trade-off between means, variances and
536 correlations. Biological Reviews 87: 742–755. https://doi.org/10.1111/j.1469-185X.2012.00225.x

537 Stiernagle T., 2006 Maintenance of C. elegans, (T. C. elegans R. Community, Ed.). WormBook.
538 https://doi.org/10.1895/wormbook.1.101.1

539 Suda H., T. Shoyama, and Y. Shimizu, 2009 Analyzing observed or hidden heterogeneity on survival and
540 mortality in an isogenic C. elegans cohort. Biophysics 5: 59–66.
541 https://doi.org/10.2142/biophysics.5.59

542 Sullivan G. M., and R. Feinn, 2012 Using Effect Size—or Why the *P* Value Is Not Enough. Journal of
543 Graduate Medical Education 4: 279–282. https://doi.org/10.4300/JGME-D-12-00156.1

544 Taylor M., and N. M. Vega, 2021 Host Immunity Alters Community Ecology and Stability of the
545 Microbiome in a Caenorhabditis elegans Model. mSystems 6.
546 https://doi.org/10.1128/mSystems.00608-20

547 Taylor M., S. Spandana Boddu, and N. Vega, 2022 Using Single-Worm Data to Quantify Heterogeneity in
548 Caenorhabditis elegans-Bacterial Interactions | Protocol. JoVE. https://doi.org/10.3791/64027

549 Travers L. M., H. Carlsson, M. I. Lind, and A. A. Maklakov, Beneficial cumulative effects of old parental
550 age on offspring fitness. Proc Biol Sci 288: 20211843. https://doi.org/10.1098/rspb.2021.1843

551 Tsilimigras M. C. B., and A. A. Fodor, 2016 Compositional data analysis of the microbiome:
552 fundamentals, tools, and challenges. Annals of Epidemiology 26: 330–335.
553 https://doi.org/10.1016/j.annepidem.2016.03.002

554 VanderWaal K. L., and V. O. Ezenwa, 2016 Heterogeneity in pathogen transmission: mechanisms and
555 methodology. Functional Ecology 30: 1606–1622. https://doi.org/10.1111/1365-2435.12645

556 Vickers A. J., 2005 Parametric versus non-parametric statistics in the analysis of randomized trials with
557 non-normally distributed data. BMC Medical Research Methodology 5: 35.
558 https://doi.org/10.1186/1471-2288-5-35

559 Walker A. C., R. Bhargava, A. S. Vaziriyan-Sani, A. S. Brust, and D. M. Czyz, 2022 Quantification of
560 Bacterial Loads in Caenorhabditis elegans. Bio-protocol 12: e4291–e4291.

561 Wickham H., 2016 *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

562 Wickham H., M. Averick, J. Bryan, W. Chang, L. D. McGowan, *et al.*, 2019 Welcome to the tidyverse.
563 Journal of Open Source Software 4: 1686. https://doi.org/10.21105/joss.01686

564 Wilke C. O., 2020 *cowplot: Streamlined Plot Theme and Plot Annotations for "ggplot2."*

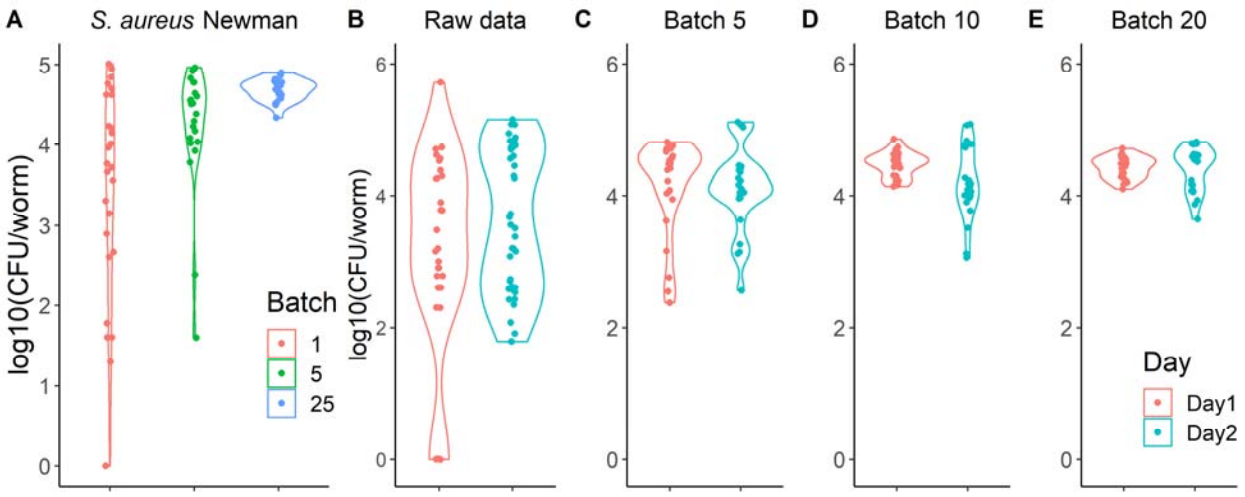
565 Woolhouse M. E. J., C. Dye, J.-F. Etard, T. Smith, J. D. Charlwood, *et al.*, 1997 Heterogeneities in the
566 transmission of infectious agents: Implications for the design of control programs. PNAS 94:
567 338–342. https://doi.org/10.1073/pnas.94.1.338

568 Wu D., S. L. Rea, A. I. Yashin, and T. E. Johnson, 2006 Visualizing hidden heterogeneity in isogenic
569 populations of C. elegans. Experimental Gerontology 41: 261–270.
570 https://doi.org/10.1016/j.exger.2006.01.003

571 Zuur A. F., E. N. Ieno, and C. S. Elphick, 2010 A protocol for data exploration to avoid common statistical
572 problems. Methods in Ecology and Evolution 1: 3–14. https://doi.org/10.1111/j.2041-
573 210X.2009.00001.x

574
575
576

577 **Figure Legends**



578

579 *Figure 1. Batching individuals for CFU/worm quantification approximates the sample mean and amplifies*
580 *differences in day-to-day variation in the distribution of data. (A) Experimental batching skews estimated*
581 *CFU/worm toward the upper quantiles of the data. Here, synchronized N2 adult hermaphrodites were*
582 *colonized by feeding on NGM plates with lawns of Staphylococcus aureus Newman for 20 hours. Worms*
583 *were cleaned and permeabilized according to standard protocols, then mechanically disrupted as*
584 *individual worms, batches of 5 worms, or batches of 25 worms for dilution plating and CFU/worm*
585 *quantification (n=24 data points each). Data are CFU/worm, as measured for individuals or as inferred by*
586 *dividing the CFU count for each batch digest by the number of individuals in that batch. (B-E) Simulated*
587 *batch data, created by resampling a raw CFU/worm dataset. Two independent runs of single-worm data*
588 *for N2 adults colonized with Salmonella enterica LT2 (days 1 and 2, B) were resampled with replacement*
589 *within day to create 25 simulated data points at each batch size (C-E).*

590

591

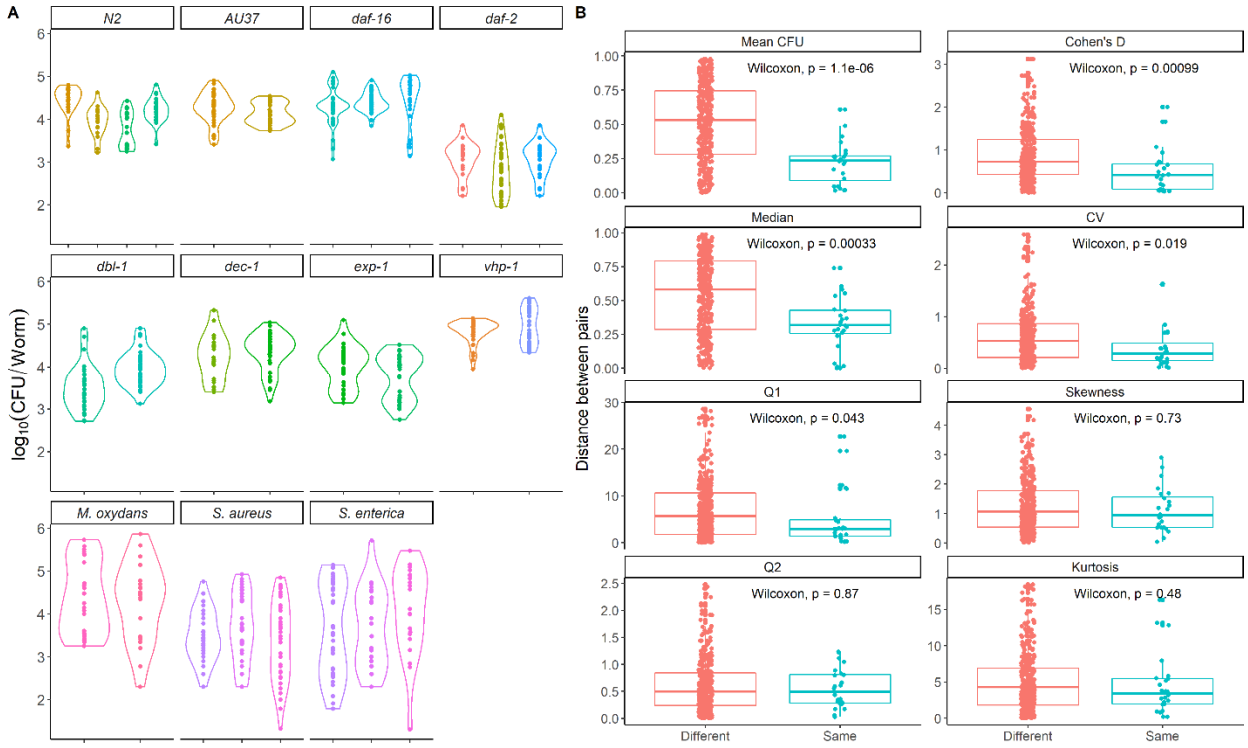


Figure 2. Comparisons of CFU/worm data within and between conditions. (A) CFU/worm data for each replicate (experimental day) within each condition (host genotype for minimal community data, colonizing bacterial species for mono-colonization data). Each point represents an individual worm. (B) Pairwise differences in summary statistics within the multi-day data set, grouped according to whether the runs compared are replicates of the same condition from two different days (same host genotype for minimal microbiome data or same colonizing microbe for single-species colonization data, "same") or from different conditions. Cohen's d is calculated as (M1-M2)/s, where Mi is the mean of sample 1 or 2 (M1>M2), and s is the smaller of the two standard deviations. Hogg's statistics are calculated as $Q1=(U_{0.05}-M_{0.5})/(M_{0.5}-L_{0.05})$ *and* $Q2=(U_{0.05}-L_{0.05})/(U_{0.5}-L_{0.5})$*, where* $U_{0.05}$ *and* $L_{0.05}$ *are the means of the upper and lower 5%,* $M_{0.5}$ *is the mean of the middle 50%, and* $U_{0.5}$ *and* $L_{0.5}$ *are the means of the upper and lower 50% of the data. Q1 is a measure of skewness, and Q2 is a measure of tail weight. Distances between means and medians are the rescaled quantity obtained by dividing the absolute difference between the statistics of two replicates by the sum of the statistics of those two replicates. Distances in the scale-free quantities CV, Q1, Q2, skewness, and kurtosis are the absolute value of the difference across pairs of replicates.*

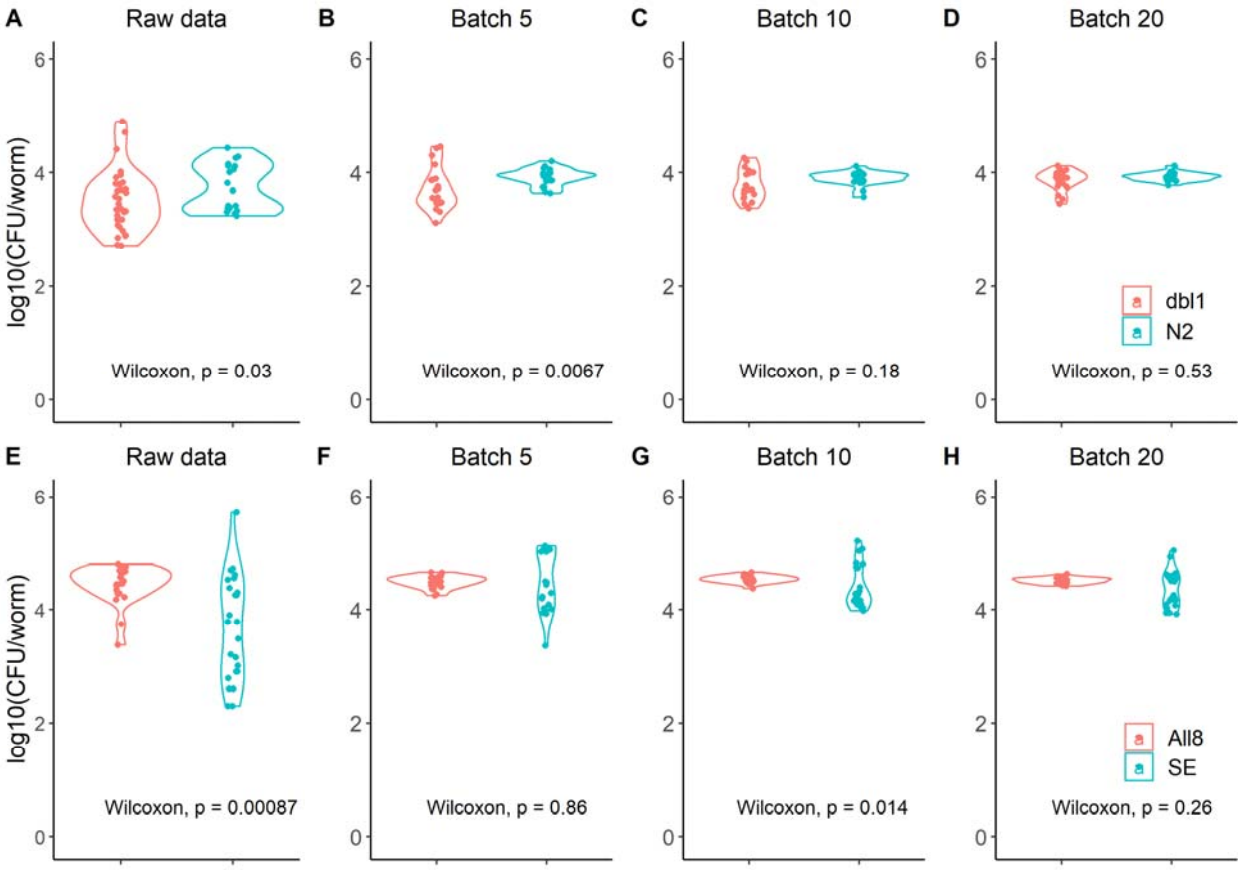


Figure 3. Raw data (A, E) and one run of simulated batch digests generated by bootstrapping (B-D) multispecies total CFU/worm data for N2 vs. dbl-1 adults and (F-H) multispecies (all-8) total CFU/worm vs S. enterica (SE) total CFU/worm in N2 adults. All raw data sets represent a single day of sampling (n=21-24 worms).

615 **Tables**

616

		Batch Size				
		1	5	10	20	50
Beta(1, 5)	**t-test**	0.07	0.15	0.23	0.33	0.515
	Mann-Whitney	0.06	0.14	0.21	0.32	0.506
Beta(0.5, 2.5)	**t-test**	0.08	0.09	0.14	0.22	0.401
	Mann-Whitney	0.07	0.09	0.12	0.23	0.391
Beta(5,1)	**t-test**	0.08	0.13	0.20	0.32	0.502
	Mann-Whitney	0.07	0.13	0.20	0.31	0.488
Beta(2.5, 0.5)	**t-test**	0.05	0.10	0.14	0.25	0.402
	Mann-Whitney	0.06	0.09	0.14	0.24	0.394
Beta(5, 5)	**t-test**	0.10	0.25	0.40	0.53	0.692
	Mann-Whitney	0.10	0.25	0.39	0.51	0.679
Beta(2.5, 2.5)	**t-test**	0.07	0.19	0.29	0.44	0.603
	Mann-Whitney	0.07	0.17	0.28	0.42	0.591
Beta(1, 1)	**t-test**	0.07	0.11	0.18	0.30	0.471
	Mann-Whitney	0.06	0.10	0.18	0.29	0.458

617

618 *Table 1. Results of comparisons of two sets of data drawn from beta distributions with the indicated*
619 *baseline parameters, where parameter values (αi,βi) are drawn independently within each run of*
620 *simulations (1000 runs, max 10^5 CFU/worm, 24 data points at each batch size, parameter errors drawn*
621 *from U[-0.1, 0.1]). Data shown are fraction of runs where $p<0.05$ (false positive).*

622