Microbial Species Abundance Distributions Guide Human Population Size Estimation from Sewage Metagenomes

Fangqiong Ling[1,2,3,4,*], Likai Chen[5], Lin Zhang[1], Xiaoqian Yu[6], Claire Duvallet[7,8], Siavash Isazadeh[7,8], Chengzhen Dai[9], Shinkyu Park[9], Katya Frois-Moniz[7,8], Fabio Duarte[9], Carlo Ratti[9], and Eric J. Alm[7,8,10,*]

[1]Washington University in St. Louis, Department of Energy, Environmental and Chemical Engineering, St. Louis, MO, USA
[2]Washington University in St. Louis, Department of Computer Science and Engineering, St. Louis, MO, USA
[3]Washington University in St. Louis, Division of Biological and Biomedical Sciences, St. Louis, MO, USA
[4]Washington University in St. Louis, Division of Computational and Data Science, St. Louis, MO, USA
[5]Washington University in St. Louis, Department of Mathematics, St. Louis, MO, USA
[6]Massachusetts Institute of Technology, Department of Biology, Boston, MA, USA
[7]Massachusetts Institute of Technology, Department of Biological Engineering, Boston, MA, USA
[8]Massachusetts Institute of Technology, Center for Microbiome Informatics and Therapeutics, Boston, MA, USA
[9]Massachusetts Institute of Technology, SENSEable City Lab, Boston, MA, USA
[10]Eli and Edythe L. Broad Institute of MIT and Harvard, Boston, MA, USA
[*]Authors to whom correspondence should be addressed.

Abstract

The metagenome embedded in urban sewage is an attractive new data source to understand urban ecology and assess human health status at scales beyond a single host. However, using census-based population size instead of real-time population estimates can mislead the interpretation of data acquired from sewage, hindering assessment of representativeness, inference of prevalence, or comparisons of taxa across sites. Here, we develop a new method to estimate human population size in light of recent developments in species-abundance distributions of microbial ecosystems. Using a population-scale human gut microbiome sample of over 1,100 people, we found that taxon-abundance distributions of gut-associated multi-person microbiomes exhibited generalizable relationships in response to human population size. We present a new non-parametric model, MicrobiomeCensus, for estimating human population size from sewage samples. MicrobiomeCensus harnesses the inter-individual variability in human gut microbiomes and performs maximum likelihood estimation based on simultaneous deviation of multiple taxa's relative abundances from their population means. MicrobiomeCensus outperformed generic algorithms in data-driven simulation benchmarks and detected population size differences in field data. This research provides a mathematical framework for inferring population sizes in real time from sewage samples, paving the way for more accurate ecological and public health studies utilizing the sewage metagenome.

Introduction

The metagenome embedded in urban sewage is an attractive new data source to understand urban ecology and assess human health status at scales beyond a single host[1–3]. Sewage microbiomes are found to share a variety of taxa with human gut microbiomes, where the baseline communities are characterized by a dominance of human-associated commensal organisms from the *Bacteroidetes* and *Firmicutes* phyla[1,3,4]. Human viruses like SARS-CoV-2 and polioviruses were detected in sewage samples during the pandemic and silent spreads, respectively, and found to correlate to reported cases, suggesting that sewage samples could be useful for understanding the dynamics in the human-associated symbionts at a population level[5,6]. Sewage has several advantages as samples of the population's collective symbionts. For instance, sewage samples are naturally aggregated, wastewater infrastructures are highly accessible, and data on human symbionts can be collected without visits to clinics, thus utilizing sewage samples can reduce costs and avoid biases associated with stigma and accessibility[2,7]. Consequently, SARS-CoV-2 surveillance utilizing sewage samples are underway globally and incorporated into the U.S. Centers for Disease Control and Prevention surveillance framework[8].

A pressing challenge in utilizing sewage for ecological and public health studies is the lack of methods to directly estimate human population size from sewage. Specifically, virus monitoring at finer spatial granularity, e.g., single university dorms and nursing homes, are informative for guiding contact tracing and protecting populations at higher risk, but real-time population size estimations at such fine granularity arenot yet available. For a given area, the census population (*de jure* population) can be larger than the number of people who contributed feces to sewage at a given time (*de facto* population)[9]. Conversely, the de jure population can also be smaller than the de facto population due to the presence of undocumented individuals[10]. Population proxies that are currently used for monitoring at wastewater-treatment plants, such as the loading of pepper mild mottle viruses, likely have high error at the neighborhood level because of their large variability in human fecal viromes (10^6-10^9 virions per gram of dry weight fecal matter)[11]. Consequently, it is difficult to assess the representativeness of a sewage sample, infer the taxon abundance differences across time and space, or interpret errors. Lack of population size information could lead to false negatives in assessing virus eradication, because an absence of biomarkers might be caused by a sewage sample that under-represents the population size. Despite its importance, few studies have explicitly explored ways to estimate real-time human population size from sewage samples independent from census estimates[12].

Macroecological theories of biodiversity may offer clues to decipher and even enumerate the sources of a sewage microbiome. While we are only beginning to view sewage as samples of human symbionts beyond one person, generating multi-host microbiomes resembles a fundamental random multiplicative process that can give rise to

many universal patterns seen in ecology. It has been suggested that the approximately lognormal shape of the Species-Abundance Distribution (SAD) might result from multiplicative processes[13]. Although ecological processes such as growth and stochastic interactions have a multiplicative nature and could lead to a central limiting pattern, Sizling et al. showed that lognormal SADs can be generated solely from summing the abundances from multiple non-overlapping sub-assemblages to form new assemblages[14]. Likewise, adding multiple sub-assemblages can also give rise to common Species-Area Relationships[14]. For microbial ecosystems, Shoemaker et al. examined the abilities of widely known and successful models of SADs in predicting microbial SADs and found that Poisson Lognormal distributions outperform other distributions across environmental, engineered, and host-associated microbial communities, highlighting the underpinning role of lognormal processes in shaping microbial diversity[15].

In this study, we conceptualize a sewage microbiome as a multi-person microbiome, where the number of human contributors can vary. We hypothesize that the species abundance distribution in the multi-person microbiome will vary as a function of the human population size, which would arise from summing taxon abundances from multiple hosts analogous to the Central Limit Theorem. We use human gut microbiome data comprising over a thousand human subjects and machine learning algorithms to explore these relationships. Upon discovering a generalizable relationship, we develop MicrobiomeCensus, a nonparametric model that utilizes relative taxon abundances in the microbiome to predict the number of people contributing to a sewage sample. MicrobiomeCensus utilizes a multivariate T statistic to capture the simultaneous deviation of multiple taxa's abundances from their means in a human population and performs maximum likelihood estimation. We provide proof on the validity of our approach. Next, we examine model performance through a simulation benchmark using human microbiome data. Last, we apply our model to data derived from real-world sewage. Our nonparametric method does not assume any underlying distributions of microbial abundances and can make inferences with just the computational power of a laptop computer.

Results

Species abundance distributions of multi-person microbiomes vary by population size

Here, we present MicrobiomeCensus, a nonparametric model that utilizes relative taxon abundances in the microbiome to predict the number of people contributing to a sewage sample. We establish MicrobiomeCensus in four steps. First, we demonstrate the usefulness of human microbiome features in estimating population size through a simulation mimicking an ideal mixing scenario in sewage. Then, we propose a T statistic to capture the simultaneous deviation of multiple taxa's abundances from their means in a human population, build a maximum likelihood model, and provide proof on the validity of our approach. Next, we examine model performance

through a simulation benchmark using human microbiome data. Last, we apply our model to real-world sewage. Our nonparametric method does not assume any underlying distributions of microbial abundances and can make inferences with just the computational power of a laptop computer.

We consider the fraction of microorganisms observed in sewage that are human-associated anaerobes as an "average gut microbiome" sampled from residents of a catchment area. Hence, our task becomes to find the underlying relationship between the number of contributors and the observed microbiome profiles in sewage samples. We define an "ideal sewage mixture" scenario to illustrate our case, where the sewage sample consists only of gut-associated microorganisms and is an even mix of n different individuals' feces (Figure 1). We denote the gut microbiome profile of an individual as $X_i = (X_{i_1}, X_{i_2}, \ldots, X_{i_p})^\intercal$, where each X_{i_j} represents the relative abundance of individual i and operational taxonomic unit (OTU) j; hence, our ideal sewage mixture can be represented as a mean from individuals $1, \ldots, n$:

$$\bar{X}_n = \sum_{i=1}^{n} X_i / n \tag{1}$$

where $X_1, X_2, \ldots, X_n$ are microbiome profiles from individuals $1, \ldots, n$. Under the ideal sewage mixture scenario, if we can quantitatively capture the departure of the sewage microbiome profile from the population mean of the human gut microbiomes of people constituting the catchment area, we will be able to estimate the population size.

Using a dataset comprised of 1,100 individuals' gut microbiome taxonomic profiles[16], we created synthetic mixture samples of different numbers of contributors through bootstrapping (Figure 1A). First, examined from an ecological perspective, the shape of the ranked abundance curves of the gut microbiomes differed when the means of multiple individuals were examined: when the number of contributors increased, dominance (Figure 1B). For the single-person microbiomes, log-series and lognormal distributions explained 94% and 93% of the variations in the SADs, respectively, compared with 89% for Poisson lognormal, 87% for Zipf multinomial and 80% for the broken-stick model. Multi-person microbiomes were best predicted by log-series or lognormal models, but as the population increased to over a hundred, the multi-person SADs were best described by only lognormal SADs (Table S1). The predictive performance of the lognormal is expected to be good for the gut microbial communities across different sizes because they can reflect processes of a multiplicative nature[15].

We explored the distributions of the relative abundances of gut bacteria as a function of population size. As expected, the distribution of a taxon's relative abundance changes with population size (Figure 1C). For instance, for OTU-2397, a *Bifidobacterium* taxon, the relative abundance distribution was approximately log-normal when the relative abundance in single-host samples was considered, yet converged to a Normal distribution when mixtures of

multiple hosts were considered. Although the means of the distributions of the same taxon under different population sizes were close, the variation in the data changed. A smaller variance was observed when the number of contributors increased (Figure 1D). Notably, different taxa varied in the rates at which their variances decreased with population size (Figure 1E), suggesting that a model that considers multiple features would be useful in predicting the number of contributors.

Classifiers utilizing microbial taxon abundance features alone detects single-person and multi-person microbiomes

Next, we set up a classification task using the taxon relative abundances to separate synthetic communities constituting one, ten, and a hundred people. With algorithms of varying complexity, namely Logistic Regression (LR), Support Vector Machine (SVM), and Random Forest (RF) classifier, classification accuracies of 29.6%, 97.2%, and 100% were achieved (Figure 2). Between RF and SVM, RF showed higher sensitivity and specificity in classifying all population groups (Table S2). This experiment suggests the usefulness of microbiome features in predicting human population counts from mixture samples.

MicrobiomeCensus is a statistical model that estimates population size from microbial taxon abundances

While the classification tasks described above demonstrated the usefulness of taxa's relative abundances in predicting the population size, a complex model like RF provides little explanatory power. We then ask, since the variance in the relative abundance of a given taxon decreases with population size, can we devise a statistic that captures the simultaneous deviation of several taxa's abundances from their means, and estimate population size utilizing the statistic? Further, will this new method perform well despite inter-personal variation in gut microbiomes?

Our new method, MicrobiomeCensus, involves a T statistic to capture the simultaneous deviation of multiple taxa's abundances from their means in relation to the variance of those taxa in the population (Figure 3A):

$$T = ||\Lambda_0^{-1}(\bar{X}_n - u)||_2^2, \tag{2}$$

where $\bar{X}_n = \sum_{i=1}^{n} X_i/n$ denotes the observed microbiome profile in ideal sewage, u denotes the population mean for the catchment area, and Λ_0 denotes the diagonal of the covariance matrix, $\Sigma_0 = (\sigma_{ij})_{1\leq i,j\leq p}$, i.e., where $\sigma_i = \sigma_{ii}^{1/2}, 1 \leq i \leq p$.

In developing this new method, we utilize the variance change by population, but without an *a priori* assumption about the gut bacterium species taxon abundance distributions and the covariance between species. Our analysis showed that the T statistic changed monotonically with increasing population size, indicating the promise of a

population estimation model (Figure 3B). While our T statistic is based on Hotelling's T-squared statistic[17], which is often used in multivariate T-tests, we extend its application beyond the problem of the significance of the multivariate means.

Leveraging our T statistic, we build a maximum likelihood model to estimate population size from an unseen sample. Here, the parameter of interest is the population size, the test statistic is our T statistic, and a point estimate is made by maximizing the likelihood of the observed T statistic in that sample. We performed training and validation using 50% of the human microbiome data and held out the rest of the data for testing. Our model achieved a training error as low as 0.13 (mean absolute percentage error, MAPE) when up to 250 features are included. The model's training performance increased when more features were included, yet the validation error did not profoundly change with an increasing number of features (Figure 3C). Upon training and validation, we chose the top 120 OTUs and tested the performance of the tuned model on a test set held out during training/validation. The model's MAPE was 0.21 (Figure 3D and E, testing errors at each population size evaluated are provided in Table S3), indicating that our model generalized well across different hosts. We then used all data and tuned hyperparameter to acquire a final model. When applying the final model on the same testing data, our model achieved a testing error of 0.162 (Figure 3D).

It is worth noting that in this algorithm, for each population size, we need to calculate the sampling distribution of the T statistic only once, hence it is not time-consuming, regardless of the true population size. We also note that an RF regression model could not be trained in a reasonable time on the same dataset, even with high-performance computing (Methods). Our model performed remarkably better than a ten-fold cross-validated RF regression model utilizing a reduced dataset, which gave an MAPE of 0.320, while the training time for our model was only a fraction of that of the RF regression model (Figure S1).

MicrobiomeCensus detects human population size differences in sewage samples

With the newly developed population model, we set out to apply our model to sewage samples. Ideally, we would like to apply the model to samples generated from a fully controlled experiment with known human hosts contributing at a given time, yet such an experiment presents logistic challenges beyond the field's current abilities. Instead, we applied our model to sewage samples taken using one of two methods, either a snapshot (grab sample) sample taken from the sewage stream over 5 minutes, or an accumulative (composite sample) taken at a constant rate over 3 hours during morning peak human defecation[18] (Figure S2). We hypothesized that the composite samples would represent more people than snapshot samples. Taking grab samples, we sampled at 1-hr intervals at one manhole (n=25); using the accumulative method, we sampled at three campus buildings (classroom, dormitory, and

family housing) multiple times over three months (n=76). To remove sequences possibly contributed by the water, we applied a taxonomic filter to retain families associated with the gut microbiome and normalized the species abundance by the retained sequencing reads (Methods, Table S4). We applied our final model to the sewage data set. Our model estimated 1-9 people's waste was captured by the snapshot samples (mean=3, s.d.=3), and 3-27 people were represented by the composite samples (mean=9, s.d.=7), where the composite samples represented significantly more people ($p < 0.0001$) (Figure 3F). The hypothesis that composite samples represent more people is well supported by our model results.

Sub-species diversity in sewage samples reflects adding microbiomes from multiple people

Independent from our MicrobiomeCensus model, we found that certain human gut-associated species were frequently detected in sewage samples by using shotgun metagenomics, e.g., *Bacteroides vulgatus*, *Provotella copri*, and *Eubacterium rectale*. Further, their sub-species diversity, as indicated by nucleotide diversity and the number of polymorphic sites in housekeeping genes, was dramatically higher in sewage samples than in the gut microbiomes of individual human subjects (Figure 4A-F and Supplementary Results).

To examine the effect of increasing population size on sub-species genetic variation in representative gut-associated microbial species, we simulated aggregate human gut samples using a sample without replacement procedure and computed the nucleotide diversity and numbers of polymorphic sites for the aggregate samples at different population sizes. This resulted in SNV profiles from 64 species. Our simulation showed increases in both nucleotide diversity and the number of polymorphic sites as more human gut samples were aggregated (Figure 3 G and H). For instance, the nucleotide diversity and number of polymorphic sites in *Eubacterium rectale* increased from 0.029 (s.d. 0.026) to 0.149 (s.d. 0.002) and 64 (s.d. 54.33) to 1274 (s.d. 18.41), respectively, when the population size increased from 1 to 300. Further, the number of polymorphic sites strongly correlated with the population size (Pearson correlation coefficient > 0.8) in 49 of the 64 species (Table S7), suggesting the potential that the SNV profiles of a wide range of gut species could be developed into feature space for population size estimation. Our simulation further shows that the number of polymorphic sites increased with population size more slowly than nucleotide diversity, indicating its potential to reflect more subtle changes in population size (Figure 4G and H). Despite the need for further model developments, the analysis here shows the potential of the sub-species diversity of gut anaerobes as a feature space to be developed into a population size estimation model, independent from the taxon abundance-based model described here.

Discussion

MicrobiomeCensus showed excellent performance in our simulation benchmark. In particular, the study subjects that we utilized in the training and testing sets are random samples out of 1,110 men and women across a wide range of age without any stratification, hence the model's testing performance indicates its generalizability. Our study is founded on the observations that healthy gut microbiomes are resilient, with inter-individual variability outweighing variability within individuals over time[19–21]. There are caveats to our approach; potentially, diets and regional effects on human microbiome composition could introduce noises to the prediction[?,22]. In applications to sewage, future studies on water matrix effects should be performed to understand and further account for noises from the sewage collection network. In further validating and applying the model, we recognize that both the responsible engagement of citizen scientists and privacy protection are crucial for advancing sewage-based ecological and public health studies.

Utilizing sewage to understand population-level dynamics of human symbionts presents an interesting scenario of sampling meta-communities. The gut microbiomes of humans can be viewed as local communities, and gut microbiomes of people living in a neighborhood could be viewed as a kind of regional meta-communities, because these communities are linked by dispersal that can take place among people connected by social networks and through a shared built environment. The meta-community framework is considered to provide useful new conceptual tools to understand the largely unexplained inter-personal variability in gut microbiomes, with expansions of the theory to consider biotic interactions suggested by Miller, Svanbäck, and Bohannan[23] In considering a sample of meta-communities, Leibold and Chase asked provocatively "what is a community?" and observed that the definition of a community is usually "user-defined and could be context-dependent" –"one community ecologist might explore the patterns of coexistence and species interactions among species within a delimited area, the other might ask the same question but define a community that encompasses more area and thus types of species, as well as different degrees of movements and heterogeneity patterns"[24]. The ambiguity between samples of meta-communities and local communities is particularly challenging for samples of microbial communities, because dispersal boundaries are difficult to delineate. Despite the conceptual importance, empirical methods that explicitly test whether a microbiome sample is a sample of a meta-community or a local community has not been available. MicrobiomeCensus directly distinguishes samples of meta-communities and local communities by enumerating the number of hosts contributing to a microbiome. While MicrobiomeCensus is trained on gut microbiome data, the procedure may have wide applications in other microbial ecosystems.

In response to the COVID-19 pandemic now affecting the human population globally, sewage-based virus

monitoring is underway(Bivins et al. 2020). Our analysis calls for attention to the denominator used in normalizing the biomarker measurements. While in practice, loading-based population proxies such as the copy numbers of pepper mild mottle viruses are used to normalize data generated from sewage, such proxies would likely have high error at the neighborhood level because of their variability in human fecal viromes (10^6-10 virions per gram of dry weight in fecal matters)[11], while they likely have reasonable performance when the population size is sufficiently large and the means of biomarker loadings converge under the Central Limit Theorem. Thus, the relationships between sewage measurements and true viral prevalence in small populations are hard to establish despite the need for sentinel population studies. Our model has immediate application in detecting false negatives, because it alerts us to the possibility that an absence of biomarkers might be caused by a sewage sample that under-represents a population. With further developments incorporating local training data, the model can potentially generate a denominator that can help turn biomarker measurements into estimates of prevalence and enable the application of epidemiology models at finer spatio-temporal resolutions.

Methods

Proof. Recall $X_1, X_2, \ldots, X_n$ are independent and identically distributed (i.i.d) random vectors in $\mathbb{R}^p$ with mean $u \in \mathbb{R}^p$ and variance $\mathrm{Var}(X_i) = \Sigma_0 \in \mathbb{R}^{p\times p}$. (Note $X_i = (X_{i1}, \ldots, X_{ip})^\mathsf{T}$ and in our application, each X_{ij} represents the value for person i and bacteria j.) Denote $\Sigma_0 = (\sigma_{ij})_{1\le i,j\le p}$ and $\sigma_i = \sigma_{ii}^{1/2}, 1 \le i \le p$. Let $\Lambda = \mathrm{diag}(\sigma_i, 1 \le i \le p)$. If both Σ_0 and u are given, then we can construct our statistic:

$$T_n = \|\Lambda_0^{-1}(\bar{X}_n - u)\|_2^2,$$

where $\bar{X}_n = \sum_{i=1}^n X_i/n$. For notation's simplicity, consider $Y_i = \Lambda_0^{-1}(X_i - u)$, the normalized version of X_i. Then

$$T_n = \|\bar{Y}_n\|_2^2,$$

where $\bar{Y}_n = \sum_{i=1}^n Y_i/n$. Then the covariance matrix Σ for Y_i is the correlation matrix of X_i, with expression $\Sigma = \Lambda^{-1}\Sigma_0\Lambda^{-1}$.

We need the following condition on Y_i for the main theorem.

ASSUMPTION 1 *Let* $\delta > 0$. *Assume*

$$K_\delta^{2+\delta} = \mathbb{E}\left|\frac{\|Y_1\|_2^2 - p}{\|\Sigma\|_F}\right|^{2+\delta} < \infty, \quad \text{and} \quad D_\delta^{2+\delta} = \mathbb{E}\left|\frac{Y_1^\top Y_2}{\|\Sigma\|_F}\right|^{2+\delta} < \infty. \tag{3}$$

REMARK 1 *Above conditions naturally hold if* $Y_{1i}, 1 \le i \le p$, *are independent and* $\max_{1\le i\le p} \|Y_{1i}\|_{2+\delta} \le M < \infty$. *Actually under this setting,* $\Sigma = I_p$ *and thus* $\|\Sigma\|_F = p^{1/2}$. *By Lemma 1,*

$$\mathbb{E}\Big(\big|\|Y_1\|_2^2 - p\big|^{2+\delta}\Big) \le (1+\delta)^{2+\delta}\Big(\sum_{i=1}^{p}\|Y_{1i}^2 - 1\|_{2+\delta}^2\Big)^{(2+\delta)/2} \lesssim p^{(2+\delta)/2},$$

where the constant in $\lesssim$ *only depends on* δ. *This justifies* K_δ *part in condition (3). Similarly by Lemma 1,*

$$\mathbb{E}\Big(|Y_1^\top Y_2|^{2+\delta}\Big) \le (1+\delta)^{2+\delta}\Big(\sum_{i=1}^{p}\|Y_{1i}Y_{2i}\|_{2+\delta}^2\Big)^{(2+\delta)/2} \lesssim p^{(2+\delta)/2}.$$

And thus D_δ *part in condition (3) holds.*

Theorem 1 *Assume Assumption 1 holds with some* $\delta > 0$, *also assume*

$$K_0^2/n + K_\delta^q/n^{q-1} + D_\delta^q/n^{\delta/2} \to 0, \tag{4}$$

where $q = 2+\delta$. *Then for* $Z \sim N(0,\Sigma)$, *we have*

$$\sup_{t\in\mathbb{R}}\Big|\mathbb{P}(nT_n \le t) - \mathbb{P}(\|Z\|_2^2 \le t)\Big| \to 0.$$

It is worth noticing that under settings in Remark 1, condition (4) holds. The proof follows from Theorem 2.2 in Xu et al.[25].

Based on above theorem, we would have the following result for justification of our sub-sampling approach. Let $A_1, A_2, \ldots, A_J$ be i.i.d uniformly sampled from the class $\mathscr{A}_m = \{A : A \subset \{1,2,\ldots,n\}, |A| = m\}$. Assume the sampling process are independent from our data $(X_i)_i$.

Let

$$\hat{F}_m(t) = J^{-1}\sum_{j=1}^{J}\mathbf{1}_{m\|\Lambda_0^{-1}(\bar{X}_{A_j} - \bar{X}_n)\|_2^2 \le t(1-m/n)},$$

where $\bar{X}_{A_j} = \sum_{i\in A_j} X_i/m$. Following result comes from Theorem 3.5 in Xu et al.[25].

Theorem 2 *Assume Assumption 1 holds with some* $\delta > 0$, *also assume* $m \to \infty$, $m = o(n)$ *and (4) is satisfied with n replaced by m. Then for* $J \to \infty$, *we have*

$$\sup_{t\in\mathbb{R}}|\hat{F}_m(t) - \mathbb{P}(\|Z\|_2^2 \le t)| \to 0.$$

Therefore under conditions in Theorems 1 and 2, we have

$$\sup_{t\in\mathbb{R}} |\hat{F}_m(t) - \mathbb{P}(mT_m \le t)| \to 0. \tag{5}$$

Note T_m is an infeasible estimator since u and Λ_0 are typically unknown. Therefore we need to estimate u and Λ_0, using our data $X_1,\ldots,X_{n_0}$ where n_0 is the number of observations we have. Consider the estimate

$$\hat{u} = \bar{X}_{n_0}, \quad \text{and} \quad \hat{\Lambda}^2_{0,j} = \sum_{i=1}^{n_0} (X_{i,j} - \bar{X}_{n_0,j})^2/n_0, \quad 1 \le j \le p,$$

where $\Lambda_{0,j}$ is the jth entity of Λ_0. If X_i has heavy tail, we can also consider robust m-estimator for $\hat{u}$ and $\hat{\Sigma}_0$, see for example, Catoni[26].

LEMMA 1 (BURKHOLDER[27], RIO[28]) *Let $q > 1$, $q' = \min\{q,2\}$. Let $D_T = \sum_{t=1}^{T} \xi_t$, where $\xi_t \in \mathscr{L}^q$ are martingale differences. Then*

$$\|D_T\|_q^{q'} \le K_q^{q'} \sum_{t=1}^{T} \|\xi_t\|_q^{q'}, \text{ where } K_q = \max\left\{(q-1)^{-1}, \sqrt{q-1}\right\}.$$

Bootstrap procedure. Below we describe the bootstrap procedure we use to approximate the distribution of T for different census counts. Recall that $X_1,...,X_m$ represents arrays of taxon relative abundances in the gut microbiome of human subject $1,\ldots,m$, and T is defined in (Eq. 2).

Step 1. Estimate the population mean, $\hat{u}$, and diagonal matrix, $\hat{\Lambda}_0$, from a sample human gut microbiome.

Step 2. For each census count N, generate $X_1^*,\ldots,X_m^*$ which is equivalent to drawing a simple random sample with replacement from $\{X_1^*,\ldots,X_m^*\}$. Compute $\hat{T_1}^*$ on the resulting bootstrap sample.

Step 3. Repeat Step 2 many times, B, (herein 10,000 times) to get $\hat{T_1}^*, \ldots, \hat{T_B}^*$.

Step 4. Estimate the distribution of $\hat{T}^*$ at census count N, using a Gaussian kernel.

Step 5. Repeat Steps 2-4 for all the census counts $1,\ldots,N$ considered, herein integers from 1 to 300. It should be noted that per Theorem 2 we require bootstrap sample size much smaller than total sample size, thus up to 300-person samples were simulated here because the gut microbiome dataset we utilized consisted of a total of 1010 people. The range can be expanded if a larger dataset is available.

Maximum Likelihood Estimation. We use a maximum likelihood estimation (MLE) procedure to achieve point estimates of the population size from a new mixture sample, X_0. The MLE procedure first computes T_0 from X_0, and then computes the likelihoods that T_0 was drawn from population sizes from 1 to B, respectively, using the

sampling distributions generated from the bootstrap procedures described above. Next, the population size that yields the highest likelihood is chosen. For a point estimate N, the confidence interval for the population size, N, is $[1, N]$.

Model training, validation, and testing. We synthesized a mixed data set from a gut microbiome dataset of 1,135 healthy human hosts from the Lifelines Deep study[16], which was the largest single-center study of population-level human microbiome variations from a single sequencing center at the time of this study. The data set consisted of 661 women and 474 men. We considered OTUs defined by 99% similarity of partial ribosomal RNA gene sequences (Methods of OTU clustering are described in detail in Supplementary Methods). After quality filtering, we retained 1,110 samples that had more than 4,000 sequencing reads/sample. We split the entire dataset approximately in half, using 550 subjects to generate the training/validation set and the other 550 subjects to generate the test set. We then used the aforementioned ideal sewage mixture approach to generate synthetic populations of up to 300 individuals, which is the relevant range for population estimation in upstream sewage. The training error was computed using the entire training data set. Five repeated holdout validations using a 50-50 split in the training set were performed to tune the hyperparameter for feature selection. The training and cross-validation errors were evaluated at integers from 1 to 100, using the error definition:

$$\delta = |\frac{N_{predicted} - N_{actual}}{N_{actual}}| \times 100\%, \tag{6}$$

and the model's performance across all the population sizes was characterized by the mean absolute percentage error (MAPE):

$$MAPE = \frac{1}{n}\sum_{n=1}^{100} |\frac{N_{predicted} - N_{actual}}{N_{actual}}| \times 100\%. \tag{7}$$

After training and validation, the hyperparameter (in this case, the top k abundant OTUs) that yielded the best performance in the validation step was used in the model. The tuned model was then tested on the test set. Our synthetic sewage microbiome approach captured the actual microbiome variation among individual hosts and demonstrated the model's generalizability.

Human gut microbiome 16S rRNA amplicon data source. The single-person and multi-person microbiome data were drawn from a gut microbiome dataset of 1,135 healthy human hosts from the Lifelines Deep study[16], which was the largest study of population-level human microbiome variations from a single sequencing center at the time of this study. The data set consisted of 661 women and 474 men. We considered operational taxonomic

units defined by 99% similarity of partial ribosomal RNA gene sequences. After quality filtering, we retained 1,100 samples that had more than 4,000 sequencing reads/sample. The rarefaction depth was chosen to balance sample size and sequencing depth.

16S rRNA gene amplicon sequencing data analysis. Operational taxonomic units defined at 99% sequencing similarity were generated from the combined dataset by first denoising the samples with DADA2[29], and then clustering the outputted exact sequence variants with the q2-vsearch plugin of QIIME2[30]. Taxonomic assignments were performed using a multinomial naïve Bayes classifier against SILVA 132[31,32]. All 16S rRNA gene amplicon analyses were performed in the QIIME2 platform (QIIME 2019.10)[33].

Species Abundance Distribution. We examined the relationships between the performances of several widely used SAD models and the number of contributors (population size) to a multi-person microbiome. Multi-person microbiomes were generated by sampling N individuals from the quality-filtered gut microbiome 16S rRNA dataset and summing the abundances of the same taxa. At each population size, 10,000 repeats were performed. The repeats were chosen according to the constraints of computational efficiency. The SADs evaluated included the Lognormal, Poisson Lognormal, Broken-stick, Log series and the Zipf model, which were shown to have varied successes in predicting microbial SADs[15]. We examined the fit using a rank-by-rank approach as previously described by Shoemaker et al.[15]. First, maximum-likelihood coefficients for each of the SADs described above were estimated using the R package sads[34]. Next, SADs were predicted using each model, and tabulated as RADs. Then, we used a least-squares regression to assess the relationship between the performance of the predicted SADs against the observations and recorded the coefficient of determination (R-squared). Last, R-squared values from model fits of each SAD model were summarized as the means, and the models that resulted in the highest R-squared values for each simulated community were recorded.

Field data. We conducted a field sampling campaign, collecting sewage samples daily at manholes near three buildings (two dormitory buildings and one office building) on the campus of Massachusetts Institute of Technology. Seventy-six sewage samples were collected through a continuous peristaltic pump sampler operated at the morning peak (7-10 a.m. near the dormitory buildings and 8-11 a.m. near the office building) at 4 mL/min for 3 hours. Wastewater was filtered through sterile 0.22-μm mixed cellulose filters to collect microbial biomass. Environmental DNA was extracted with a Qiagen PowerSoil DNA extraction kit according to the manufacturer's protocol. The DNA was amplified for the V4 region of the 16S rRNA gene and sequenced in a Miseq paired-end format at the MIT BioMicro Center, according to a previously published protocol[35]. Included as a comparison are a set of snapshot sewage samples taken using a peristaltic pump sampler at 100 mL/min for 5 minutes over a day (10 a.m.

on Wednesday April 8, 2015, to 9 a.m. on Thursday April 9, 2015). The sampling methods for snapshot samples are described in detail by Matus et al.[4]

Application to sewage data. 1he 16S rRNA gene amplicon sequencing data from the field sewage samples were trimmed to the same region, 16S V4 (534-786) with the LifeLines Deep data using Cutadapt 1.12[36]. Forward reads were trimmed to 175bps, and reverse reads were first trimmed to 175bps and then further trimmed to 155bps during quality screening. We created a taxonomic filter based on the composition of the gut microbiome data set, which consisted of the abundant family-level taxa that accounted for 99% of the sequencing reads in the human gut microbiome data set, and excluded those that might have an ecological niche in tap water (*Enterobacteriaceae* and *Burkholderiaceae*). This exclusion resulted in 25 bacterial families and one archaeal family in our taxonomic filter, including *Lachnospiraceae*, *Ruminococcaceae*, *Bifidobacteriaceae*, *Erysipelotrichaceae*, *Bacteroidaceae*, and others (Table S4). We applied our taxonomic filter to the sewage sequencing data, which retained 73.9% of the sequencing reads. This retention rate is consistent with our previous report of the human microbiome fraction in residential sewage samples[4]. We then normalized the relative abundance of taxa against the remaining sequencing reads in each sample. Welch's two-sample t-tests were performed to retain the OTUs whose means did not differ significantly from the human microbiome data set ($p > 0.05$).

Deployment of generic machine learning models. Logistic regression, support vector machine, and random forest classifiers were employed to perform the classification task for population sizes of 1, 10, and 100. Model training, cross-validation, and testing were performed using the R Caret platform with the default setting[37]. For the support vector machine, the radial basis function kernel was employed. Ten-fold cross-validation and five repeats were performed for all the models considered. Model performance was evaluated using accuracy, sensitivity, and specificity. Based on the classifier performance, the RF regression model was used for comparison with our new model's performance. Initially, we trained the model using the same training data set used in training our maximum likelihood model, however, the computation was infeasible, even with a 36-thread, 3TB-memory computing cluster. We then introduced gaps in the population size range, using populations from the vector $(1, 5, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, 120, 150, 180, 240, 300)^T$ while maintaining the same sample size at each population size (10,000 samples). The training was performed in R Caret, using 10-fold cross-validation. Ten variables were randomly sampled as candidates at each split, mtry=10. The performance was evaluated using the same testing set that was used to evaluate the maximum likelihood model.

References

1. Maritz, J. M., Ten Eyck, T. A., Elizabeth Alter, S. & Carlton, J. M. Patterns of protist diversity associated with raw sewage in New York City. *The ISME J.* **13**, 2750–2763, 10.1038/s41396-019-0467-z (2019). Number: 11 Publisher: Nature Publishing Group.

2. Berchenko, Y. *et al.* Estimation of polio infection prevalence from environmental surveillance data. *Sci. Transl. Medicine* **9**, 10.1126/scitranslmed.aaf6786 (2017). Publisher: American Association for the Advancement of Science Section: Reports.

3. Newton, R. J. *et al.* Sewage Reflects the Microbiomes of Human Populations. *mBio* **6**, 10.1128/mBio.02574-14 (2015).

4. Matus, M. *et al.* 24-hour multi-omics analysis of residential sewage reflects human activity and informs public health. *bioRxiv* 728022, 10.1101/728022 (2019). Publisher: Cold Spring Harbor Laboratory Section: New Results.

5. Medema, G., Heijnen, L., Elsinga, G., Italiaander, R. & Brouwer, A. Presence of SARS-Coronavirus-2 RNA in Sewage and Correlation with Reported COVID-19 Prevalence in the Early Stage of the Epidemic in The Netherlands. *Environ. Sci. & Technol. Lett.* 10.1021/acs.estlett.0c00357. Publisher: American Chemical Society.

6. Manor, Y. *et al.* Intensified environmental surveillance supporting the response to wild poliovirus type 1 silent circulation in Israel, 2013. *Euro Surveillance: Bull. Eur. Sur Les Maladies Transm. = Eur. Commun. Dis. Bull.* **19**, 20708, 10.2807/1560-7917.es2014.19.7.20708 (2014).

7. Murakami, M., Hata, A., Honda, R. & Watanabe, T. Letter to the Editor: Wastewater-Based Epidemiology Can Overcome Representativeness and Stigma Issues Related to COVID-19. *Environ. Sci. & Technol.* **54**, 5311, 10.1021/acs.est.0c02172 (2020).

8. CDC. National Wastewater Surveillance System (2020).

9. Daughton, C. G. Real-time estimation of small-area populations with human biomarkers in sewage. *The Sci. Total. Environ.* **414**, 6–21, 10.1016/j.scitotenv.2011.11.015 (2012).

10. Fazel-Zarandi, M. M., Feinstein, J. S. & Kaplan, E. H. The number of undocumented immigrants in the United States: Estimates based on demographic modeling with data from 1990 to 2016. *PloS One* **13**, e0201193, 10.1371/journal.pone.0201193 (2018).

418 **11.** Zhang, T. *et al.* RNA Viral Community in Human Feces: Prevalence of Plant Pathogenic Viruses. *PLOS Biol.* **4**,
419 e3, 10.1371/journal.pbio.0040003 (2005). Publisher: Public Library of Science.

420 **12.** Yang, Z., Xu, G., Reboud, J., Kasprzyk-Hordern, B. & Cooper, J. M. Monitoring Genetic Population Biomarkers
421 for Wastewater-Based Epidemiology. *Anal. Chem.* **89**, 9941–9945, 10.1021/acs.analchem.7b02257 (2017).

422 **13.** Putman, R. J. & Putman, R. *Community Ecology* (Springer Science & Business Media, 1994). Google-Books-ID:
423 xwjjMzGUVgwC.

424 **14.** Šizling, A. L., Storch, D., Šizlingová, E., Reif, J. & Gaston, K. J. Species abundance distribution results
425 from a spatial analogy of central limit theorem. *Proc. Natl. Acad. Sci. United States Am.* **106**, 6691–6695,
426 10.1073/pnas.0810096106 (2009).

427 **15.** Shoemaker, W. R., Locey, K. J. & Lennon, J. T. A macroecological theory of microbial biodiversity. *Nat. Ecol.*
428 *& Evol.* **1**, 1–6, 10.1038/s41559-017-0107 (2017). Number: 5 Publisher: Nature Publishing Group.

429 **16.** Zhernakova, A. *et al.* Population-based metagenomics analysis reveals markers for gut microbiome composition
430 and diversity. *Science* **352**, 565, 10.1126/science.aad3369 (2016). Publisher: NIH Public Access.

431 **17.** Hotelling, H. & Frankel, L. R. The Transformation of Statistics to Simplify their Distribution. *The Annals Math.*
432 *Stat.* **9**, 87–96 (1938). Publisher: Institute of Mathematical Statistics.

433 **18.** Heaton, K. W. *et al.* Defecation frequency and timing, and stool form in the general population: a prospective
434 study. *Gut* **33**, 818–824, 10.1136/gut.33.6.818 (1992).

435 **19.** David, L. A. *et al.* Host lifestyle affects human microbiota on daily timescales. *Genome Biol.* **15**, R89,
436 10.1186/gb-2014-15-7-r89 (2014).

437 **20.** Lozupone, C. A., Stombaugh, J. I., Gordon, J. I., Jansson, J. K. & Knight, R. Diversity, stability and resilience
438 of the human gut microbiota. *Nature* **489**, 220–230, 10.1038/nature11550 (2012).

439 **21.** Mehta, R. S. *et al.* Stability of the human faecal microbiome in a cohort of adult men. *Nat. Microbiol.* **3**,
440 347–355, 10.1038/s41564-017-0096-0 (2018).

441 **22.** Johnson, A. J. *et al.* Daily Sampling Reveals Personalized Diet-Microbiome Associations in Humans. *Cell Host*
442 *& Microbe* **25**, 789–802.e5, 10.1016/j.chom.2019.05.005 (2019).

443 **23.** Miller, E. T., Svanbäck, R. & Bohannan, B. J. M. Microbiomes as Metacommunities: Understanding Host-
444 Associated Microbes through Metacommunity Ecology. *Trends Ecol. & Evol.* **33**, 926–935, 10.1016/j.tree.2018.
445 09.002 (2018).

24. Leibold, M. A. & Chase, J. M. *Metacommunity Ecology, Volume 59* (Princeton University Press, 2017). Google-Books-ID: IEkqDwAAQBAJ.

25. Xu, M., Zhang, D. & Wu, W. B. L2 asymptotics for high-dimensional data. *arXiv preprint arXiv:1405.7244* (2014).

26. Catoni, O. Challenging the empirical mean and empirical variance: a deviation study. In *Annales de l'IHP Probabilités et statistiques*, vol. 48, 1148–1185 (2012).

27. Burkholder, D. L. Sharp inequalities for martingales and stochastic integrals. *Astérisque* **157**, 75–94 (1988).

28. Rio, E. Moment inequalities for sums of dependent random variables under projective conditions. *J. Theor. Probab.* **22**, 146–163 (2009).

29. Callahan, B. J. *et al.* Dada2: high-resolution sample inference from illumina amplicon data. *Nat. methods* **13**, 581–583 (2016).

30. Rognes, T., Flouri, T., Nichols, B., Quince, C. & Mahé, F. VSEARCH: A versatile open source tool for metagenomics. *PeerJ* **4**, e2584, 10.7717/peerj.2584 (2016).

31. Quast, C. *et al.* The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Res.* **41**, D590–596, 10.1093/nar/gks1219 (2013).

32. Bokulich, N. A. *et al.* Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. *Microbiome* **6**, 90, 10.1186/s40168-018-0470-z (2018).

33. Bolyen, E. *et al.* Reproducible, interactive, scalable and extensible microbiome data science using qiime 2. *Nat. biotechnology* **37**, 852–857 (2019).

34. Prado, P. I., Miranda, M. D. & Chalom, A. *sads: Maximum Likelihood Models for Species Abundance Distributions* (2018). R package version 0.4.2.

35. Preheim, S. P. *et al.* Computational methods for high-throughput comparative analyses of natural microbial communities. *Methods Enzymol.* **531**, 353–370, 10.1016/B978-0-12-407863-5.00018-6 (2013).

36. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10–12, 10.14806/ej.17.1.200 (2011).

37. Kuhn, M. Building Predictive Models in R Using the caret Package. *J. Stat. Softw.* **028** (2008). Publisher: Foundation for Open Access Statistics.

Acknowledgements

We thank Cho C. Yiu, David E Hingston, and Joseph S. Monteiro from the MIT Facilities department for assistance with sewage sampling. We thank Noriko Endo, Sean Gibbons, Tami Lieberman, Xiaofang Jiang and Shijie Zhao for valuable discussions. We thank Mariana Matus and Newsha Ghaeli for acquisition and access of 24-hr sewage time series data. The sampling and sequencing experiments were performed with funding from the Kuwait Foundation for Advancement of Sciences. The computational analyses were performed with resources provided by the WUSTL McKelvey School of Engineering. Fangqiong Ling was partially supported by an Alfred P. Sloan Foundation Microbiology of the Built Environment Postdoctoral Fellowship. Lin Zhang was supported by a Washington University Faculty Startup Fund to Fangqiong Ling.

Author contributions statement

F.L. and E.J.A. designed the study; F.L. performed simulation; L.C. performed mathematical proofs; F.L., X.Y. and L.Z. performed sequence analysis; F.L., S.I., C.Dai., and S.P. performed sewage experiments; C. Duvallet, K. F-M, F.D. C.R. and E.J.A. coordinated the acquisition of sequencing data; F.L., L.C.,and E.J.A wrote the manuscript.

Competing interests

E.J.A has an equity stake in Biobot Analytics. C. Duvallet is employed by Biobot Analytics.

Code and data availability

Source code will be made avaiable through https://github.com/linglab-washu/population-model upon publication. Sewage metagenomic data will be made available at National Center for Biotechnology Information Short Read Archive at BioProject PRJNA683921 upon the time of publication.

Figure Legends

Figure 1. An ideal sewage mixture simulation shows the potential of microbiome taxon abundance profiles as population census information sources. (A) We generated an "ideal sewage mixture" consisting of gut microbiomes from different numbers of people. (B) Ranked abundance curves for gut microbiomes of one person and mixtures of multiple people exhibit different levels of dominance and diversity. Blue lines show the rank abundance curves in stool samples (one person), red lines show 10-person mixtures, and saffron lines show 100-person mixtures. In each scenario, ten examples are shown. All samples were rarefied to the same sequencing depths (4,000 seqs/sample). (C)

The probability density function of the relative abundance of one taxon for different population sizes. OTU-2379, a *Bifidobacterium* taxon, was used as an example. Maroon dashed lines indicate the sample means. (D) Multiple taxa's abundance variances in one-person samples and 100-person samples. The dominant taxa are shown (top100) and are sorted by their ranks in variance. (E) The ratios of the variances of one-person samples and 100-person samples across dominant gut microbial taxa.

Figure 2. Classifier performance of models utilizing gut microbiome taxon abundances.

Figure 3. MicrobiomeCensus statistic definition, model training, validation, and application. (A) Example of computing the T statistic. (B) Simulation results for T with different population sizes. Grey points are simulation results. Red bars are means of 10,000 repeats performed for each population size. (C) Model training and tuning. We built the MicrobiomeCensus model using our T statistic and a maximum likelihood procedure. The training set consisted of 10,000 samples for population sizes ranging from 1-300, and 50% of the data were used to train and validate the model. Training and validation errors from different feature subsets are shown. Training errors are shown as red lines, and validation errors are shown as blue lines. (D) Model performance on simulation benchmark. After training and validation, the model utilized the top 120 abundance features. Model performance was tested on synthetic data generated from 550 different subjects not previously seen by the model. The training set consisted of 10,000 samples with population sizes from 1-300, and the testing set consisted of 10,000 repeats at the evaluated population sizes. The training error, testing error, and the error of the final model are shown. (E) Model performance evaluated using a testing set. Black solid dots indicate the means of the predicted values, and error bars indicate the standard deviations of the predicted values. (F) Application of the microbiome population model in sewage. Seventy-six composite samples (blue) were taken from three manholes on the MIT campus, and each sample was taken over 3 hours during the morning peak water usage hours. Twenty-five snapshot samples (grey) were taken using a peristaltic pump for 5 minutes at 1-hour intervals throughout a day.

Figure 4. Sub-species diversity in gut-associated bacterial species as a potential marker for human population size. (A-F) Comparison of sub-species diversity of gut-associated bacteria in human gut microbiome samples (LifelinesDeep) and MIT sewage samples. Nucleotide diversity and numbers of polymorphic sites were computed from ten phylogenetic marker genes. (G) and (H) Simulation results showing intra-species diversity in response to increasing population size, as represented by the number of polymorphic sites (G) and nucleotide diversity (H).

Figure 1.

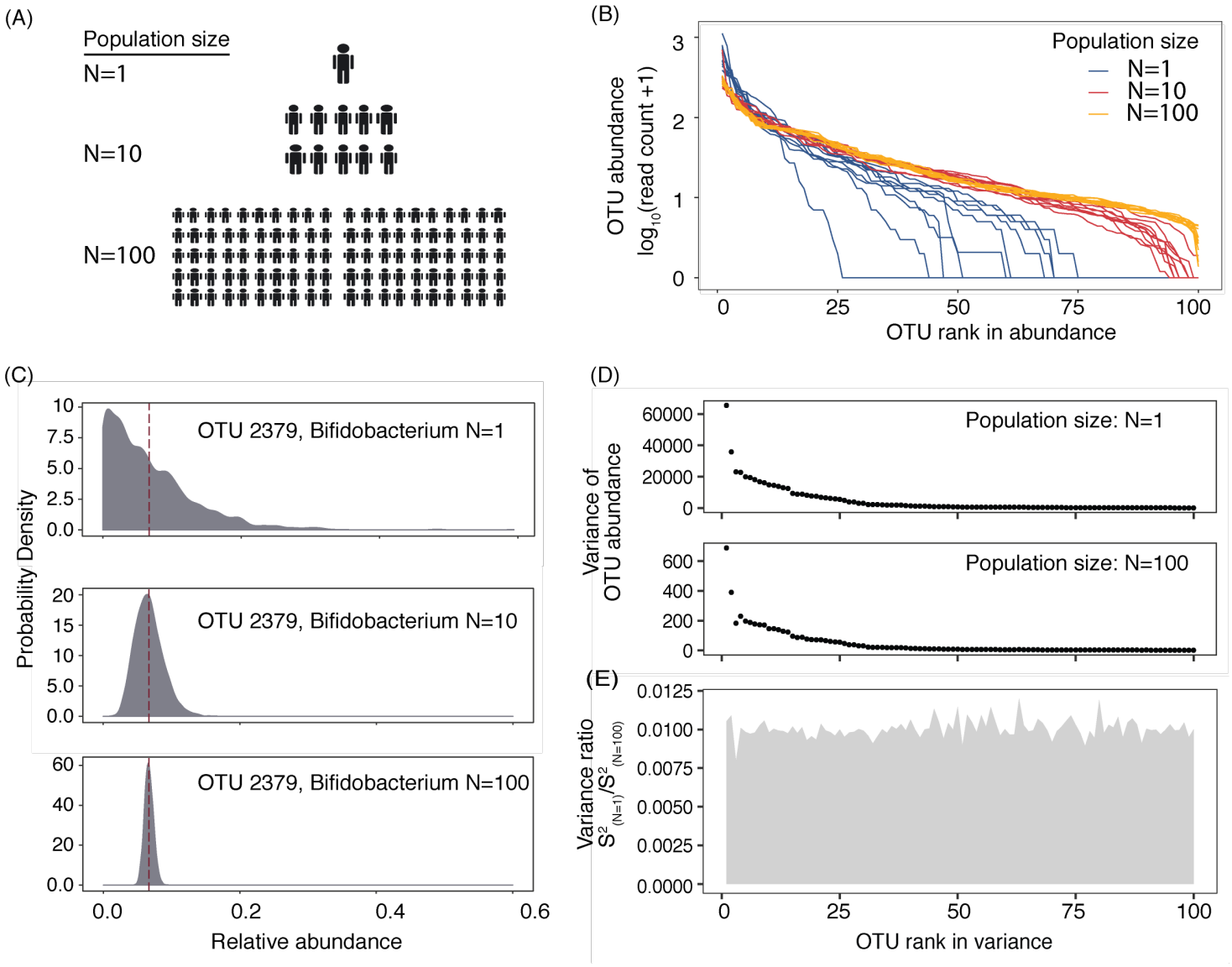

Figure 2.

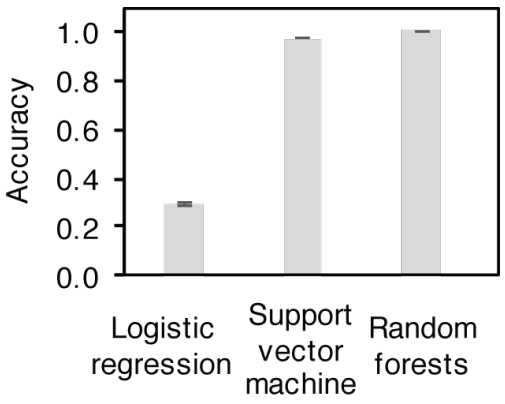

Figure 3.

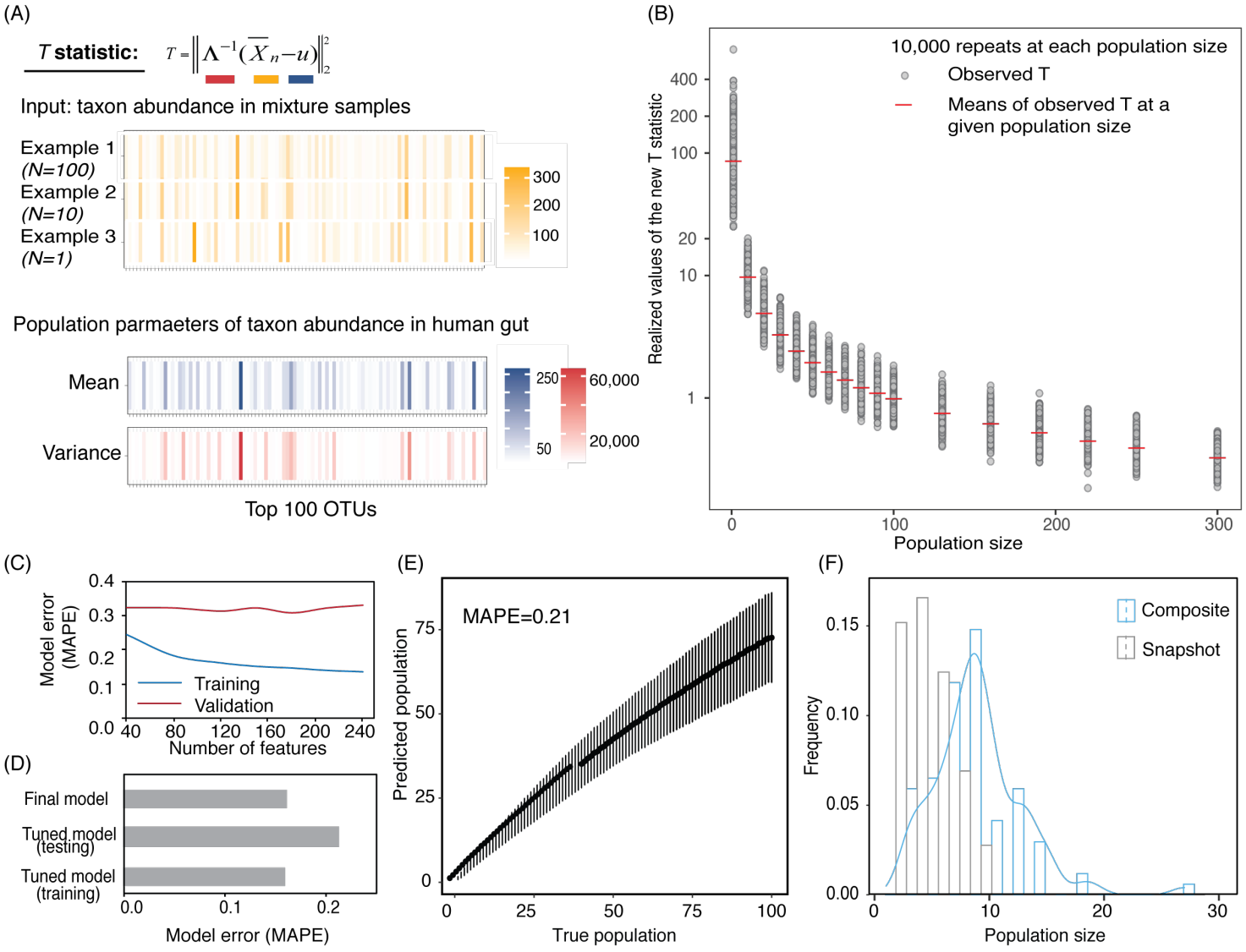

Figure 4.