AMERICAN SOCIETY FOR MICROBIOLOGY mSphere®

RESEARCH ARTICLE
Applied and Environmental Science

PhenoGMM: Gaussian Mixture Modeling of Cytometry Data Quantifies Changes in Microbial Community Structure

Peter Rubbens,[a,c] Ruben Props,[b] Frederiek-Maarten Kerckhof,[b] Nico Boon,[b] Willem Waegeman[a]

[a]KERMIT, Department of Data Analysis and Mathematical Modelling, Ghent University, Ghent, Belgium
[b]Center for Microbial Ecology and Technology (CMET), Ghent University, Ghent, Belgium
[c]Flanders Marine Institute (VLIZ), Ostend, Belgium

ABSTRACT Microbial flow cytometry can rapidly characterize the status of microbial communities. Upon measurement, large amounts of quantitative single-cell data are generated, which need to be analyzed appropriately. Cytometric fingerprinting approaches are often used for this purpose. Traditional approaches either require a manual annotation of regions of interest, do not fully consider the multivariate characteristics of the data, or result in many community-describing variables. To address these shortcomings, we propose an automated model-based fingerprinting approach based on Gaussian mixture models, which we call PhenoGMM. The method successfully quantifies changes in microbial community structure based on flow cytometry data, which can be expressed in terms of cytometric diversity. We evaluate the performance of PhenoGMM using data sets from both synthetic and natural ecosystems and compare the method with a generic binning fingerprinting approach. PhenoGMM supports the rapid and quantitative screening of microbial community structure and dynamics.

IMPORTANCE Microorganisms are vital components in various ecosystems on Earth. In order to investigate the microbial diversity, researchers have largely relied on the analysis of 16S rRNA gene sequences from DNA. Flow cytometry has been proposed as an alternative technology to characterize microbial community diversity and dynamics. The technology enables a fast measurement of optical properties of individual cells. So-called fingerprinting techniques are needed in order to describe microbial community diversity and dynamics based on flow cytometry data. In this work, we propose a more advanced fingerprinting strategy based on Gaussian mixture models. We evaluated our workflow on data sets from both synthetic and natural ecosystems, illustrating its general applicability for the analysis of microbial flow cytometry data. PhenoGMM supports a rapid and quantitative analysis of microbial community structure using flow cytometry.

KEYWORDS diversity, fingerprint, flow cytometry, machine learning, microbial communities, mixture model

Various tools have been developed to study and monitor microbial communities. With the emergence of 16S rRNA gene sequencing, researchers have uncovered the genotypic diversity of microbial communities to a large extent (1). However, microorganisms with the same genotype can still present different phenotypes, displaying so-called phenotypic heterogeneity (2). Therefore, instead of solely focusing on genotypic information, there is a need to combine omics data with phenotypic information (3). One such tool to study the phenotypic identity of microbial communities is flow cytometry (FCM). FCM is a high-throughput technique, measuring hundreds to thousands of individual cells in mere seconds. These measurements result in a multivariate

Citation Rubbens P, Props R, Kerckhof F-M, Boon N, Waegeman W. 2021. PhenoGMM: Gaussian mixture modeling of cytometry data quantifies changes in microbial community structure. mSphere 6:e00530-20. https://doi.org/10.1128/mSphere.00530-20.

Editor Katherine McMahon, University of Wisconsin-Madison

Address correspondence to Peter Rubbens, peter.rubbens@vliz.be.

PhenoGMM, a new cytometric fingerprinting approach proposed by @prubbens and colleagues, allows researchers to describe the structure of microbial communities based on flow cytometry.

Received 5 June 2020
Accepted 18 December 2020
Published 3 February 2021

description of each cell, derived from both scatter and fluorescence signals. The first is related to cell size and morphology, while the latter depends on either autofluorescence properties or the interaction between the cell and a specific stain.

Many algorithms exist in the field of immunophenotyping cytometry to identify separated cell populations, i.e., cells that share similar phenotypic characteristics as measured by FCM and that therefore can be grouped together. These algorithms are extensively benchmarked for different human FCM and mass cytometry data sets (4, 5). However, microbial cytometry data have a number of different characteristics. This originates from the fact that bacterial cells are typically much smaller in both cell size and volume than eukaryotic cells (6), which complicates their detection. In addition, no general antibody-based panels have been established for microbial cells due to the high complexity of microbial communities (7). One has to rely on general DNA stains, for which it is difficult to develop multicolor approaches (8). Therefore, the number of variables describing an individual bacterial cell is typically much lower than that for, for example, a human cell. As the number of bacterial taxa is much larger than the number of differentiating signals, cytometric distributions of these taxa can highly overlap (9–11). This is why automated cell population identification algorithms cannot be directly applied for the analysis of bacterial cytometry data. Consequently, data analysis pipelines should be designed to consider these characteristics.

To do so, microbiologists commonly rely on so-called cytometric fingerprinting techniques (12, 13). Such a fingerprint allows researchers to derive community-level variables in terms of the number of bins or clusters (i.e., gates), cell counts per cluster, and the position of those clusters (14), despite the fact that there are no or only a few clearly separated cell populations. The approaches that are currently used for the analysis of bacterial communities can be broadly divided into two categories: (i) manual annotation of clusters (12, 15) and (ii) automated approaches that employ binning strategies (13, 16–18). Both categories of methods have a number of drawbacks: (i) manual gating of regions of interest is laborious in time and operator dependent, (ii) traditional binning approaches result in a large number of variables (e.g., a fixed grid of dimensions 100 × 100 will result in 10,000 sample-describing variables), and because of that, (iii) only bivariate interactions of cytometry channels are considered when employing such a binning approach.

After a fingerprint has been constructed, communities are described by a contingency table that contains the abundances of groups of cells that are similar. Based on this, changes in microbial community structure can be quantified. This approach has been successfully applied to characterize dynamics of the microbiome in a multitude of environments, including pure (19, 20), synthetic (21), drinking water (22, 23), wastewater (12, 15), freshwater (24, 25), marine (16, 26), salivary (27), soil (28), and gut (29, 30) microbial communities. Cytometric fingerprint data can be summarized in what has been proposed as the cytometric or phenotypic diversity of a microbial community (13, 18). These are estimations of the diversity of a microbial community based on the cell counts per gate, bin, or cluster. If many clusters or bins contain cells, a community can be considered "rich." If the cells are equally distributed over those clusters, a community can be considered "even." Recent reports have shown a moderate to strong correlation between the cytometric diversity and genotypic diversity derived from 16S rRNA gene amplicon sequencing data (13, 16, 25, 30).

Our methodology makes use of Gaussian mixture models (GMMs). GMMs have been successfully applied to cytometry data before to identify separated cell populations in an automated way (31, 32). Considering microbial communities, Hyrkas et al. have shown that their GMM approach outperformed state-of-the-art immunophenotyping cytometry algorithms for the automated identification of phytoplankton populations (33). A similar approach has been recently proposed by Ludwig et al. to identify separated bacterial populations using two-dimensional cytometry data (34). By overclustering the data, GMMs can be adjusted to describe the distribution of multivariate data without the need to identify separated populations. As such, GMMs can be used as an effective fingerprinting strategy. Two additional advantages are the fact that multivariate data can be modeled at once and that the number of mixtures needed to describe the data is much lower than the number of variables resulting from traditional binning approaches.

In this work, we propose an extension of current fingerprinting approaches that we have called PhenoGMM. The methodology is able to describe the potentially many overlapping cell populations in microbial FCM data. We demonstrate that changes in community structure can be quantified based on the cytometric fingerprints derived from PhenoGMM. We evaluate its performance for synthetic and natural freshwater microbial communities and compare its performance with that of a generic binning approach. The methodology has been integrated in the R package PhenoFlow (13).

RESULTS

PhenoGMM correctly quantifies the community structure of *in silico* synthetic microbial communities. In the first experiment, we evaluated the capacity of PhenoGMM to estimate the intracommunity diversity (i.e., α-diversity) of synthetic microbial communities. To this end, we simulated 400 different synthetic microbial community compositions and artificially aggregated the data of bacterial strains that were measured individually by FCM according to these compositions. Three hundred communities made up a training set; the other 100 communities made up the test set. The number of strains varied randomly between two and 20 (the total number of available strains). Community compositions were simulated using a Dirichlet distribution for three different values of the concentration parameter a (i.e., $a = 0.1$, 1, and 10). This parameter determines how evenly the weight is spread among the different strains. If a is small, only a few species are dominantly present. If a is large, the weight will be more evenly spread among the different strains. Its effect on the sampled proportions for $a = 0.1$, 1, and 10 is illustrated using Lorenz curves. These depict the cumulative proportion of abundance versus the cumulative proportion of bacterial species (see Fig. S1 in the supplemental material).

We compared PhenoGMM with an approach that we have called PhenoGrid for this work. The latter represents common cytometric fingerprinting approaches in microbial ecology that employ a binning approach to one or more bivariate combinations of the data. A GMM of $K = 128$ mixtures or a fixed binning grid of dimensions $3 \times 128 \times 128$ (i.e., number of bivariate combinations × number of intervals first channel × number of intervals second channel) was fitted to a combined representation of the 300 communities in the training set. The resulting fingerprint templates were then used to retrieve cell counts per mixture or bin to describe each community in the test set. α-Diversity metrics were determined based on the resulting cell count contingency tables, as defined by the Hill numbers D_q. The sensitivity parameter q determines the importance that is given to rare species or populations, with a low q giving more weight to rare species. α-Diversity was determined for $q = 0$ (richness), $q = 1$ (exponent of the Shannon entropy), and $q = 2$ (inverse Simpson index). Estimations of α-diversity were correlated with the "true" α-diversity values, which were based on the predefined compositions with which the communities in the training and test sets were simulated. Correlations were quantified using Kendall's rank correlation coefficient τ_B and summarized in Fig. 1. PhenoGMM resulted in moderate to highly correlated α-diversity estimations and showed a better correspondence to the predefined community compositions compared to PhenoGrid. Estimations were just above the significance level ($P = 0.05$) for the latter. The performance mainly depended on the sensitivity parameter q. Estimations resulted in higher correlations for PhenoGMM when $q > 0$, i.e., when more weight was given to more abundant strains. This means that PhenoGMM captured the structure rather than the identity of a microbial community. This effect was less clear for PhenoGrid.

We further evaluated to what extent a mixture corresponded to one or more bacterial strains. To do so, we constructed a fingerprint using 20 mixtures for the setting in which the concentration parameter of the Dirichlet distribution was set to $a = 1$.

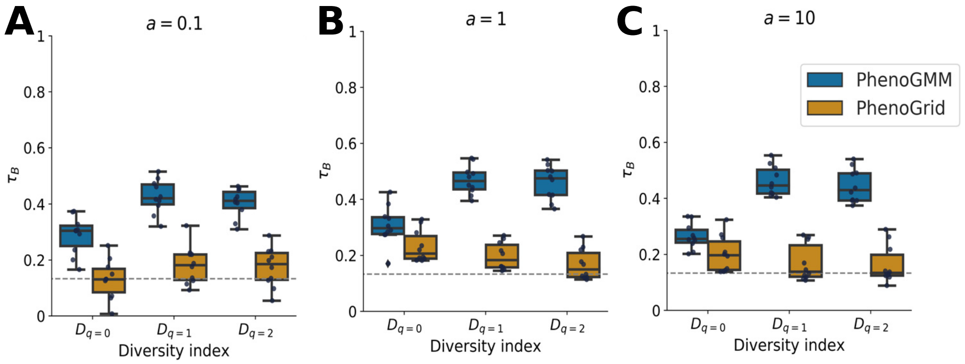


FIG 1 Summary of α-diversity estimations for *in silico* synthetic microbial communities, quantified by Kendall's τ_B, for PhenoGMM and PhenoGrid. Both workflows were run 10 times. Kendall's τ_B was calculated between true and estimated values. Each boxplot displays the 25% and 75% quartiles of the τ_B, and the whiskers show the full range of τ_B. Each dot represents the resulting value from an individual run. (A) $a = 0.1$; (B) $a = 1$; (C) $a = 10$. The dashed line indicates the strength of τ_B at $P = 0.05$.

Relative cell counts per mixture were correlated with variations in individual abundances of bacterial strains (Fig. 2A). In most cases multiple mixtures were correlated with multiple strains (Fig. 2B), which could be explained by the fact that the cytometric characterizations of the considered bacterial strains overlapped in various degrees. At the same time, no mixture was correlated with all bacterial strains, demonstrating that despite the overlapping structure in a cytometric fingerprint, variations in the mixtures could be successfully related to variations in individual strains.

To scrutinize the results and potentially facilitate future synthetic microcosm experiments, we performed a predictive modeling analysis. Cytometric fingerprints from PhenoGMM and PhenoGrid were fed to a Random Forest model in order to predict the community structure according to which microbial communities were assembled in the test set. Cytometric fingerprints from both approaches either resulted in comparable predictions according to Kendall's τ_B (Fig. 3) or were slightly in favor of PhenoGMM according to R^2 (Fig. S2). Random Forest predictions resulted in stronger correlations

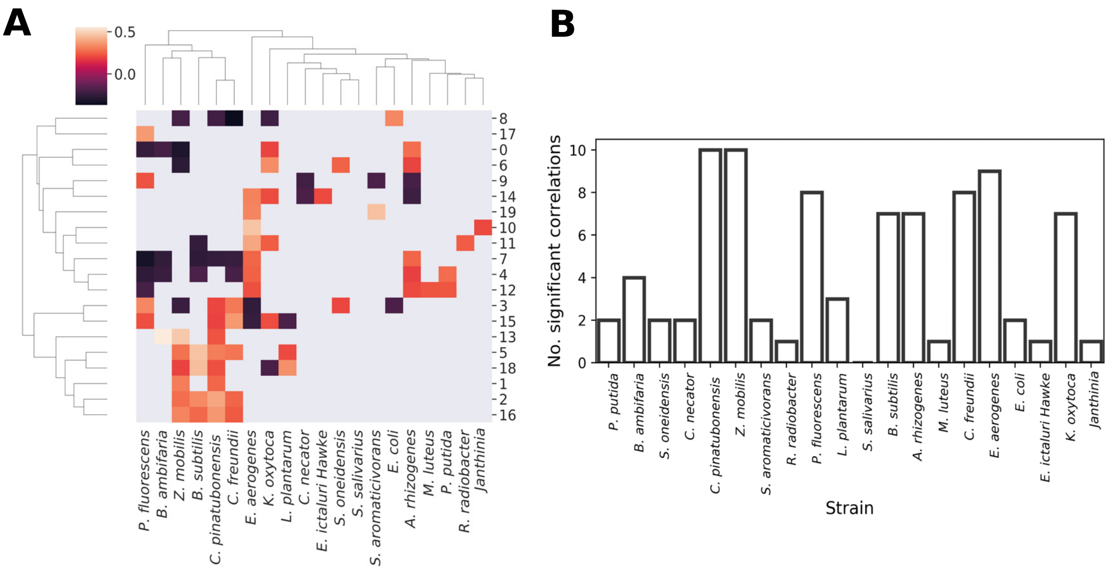


FIG 2 Summary of the correspondence between individual bacterial strains and the cell counts for each Gaussian mixture. (A) Kendall's τ_B between relative cell counts per mixture (rows) and relative abundances of bacterial strains (columns). Values are given if P is ≤ 0.05, after performing a Benjamini-Hochberg correction for multiple hypothesis testing. (B) Number of significant correlations per bacterial strain.

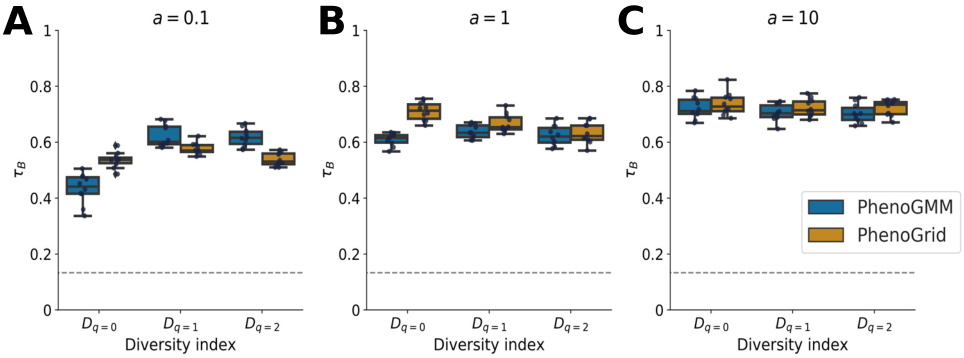


FIG 3 Summary of Random Forest predictions of α-diversity for *in silico* synthetic microbial communities, quantified by Kendall's τ_B, for PhenoGMM and PhenoGrid. Both workflows were run 10 times. Kendall's τ_B was calculated between true and estimated values. Each boxplot displays the 25% and 75% quartiles of the τ_B, and the whiskers show the full range of τ_B. Each dot represents the resulting value from an individual run. (A) $a = 0.1$; (B) $a = 1$; (C) $a = 10$. The dashed line indicates the strength of τ_B at $P = 0.05$.

with the simulated community compositions compared to directly applying the Hill numbers to the relative cell contingency table.

We estimated the time to run PhenoGMM for $a = 1$ and $D_{q=1}$ in function of the number of mixtures K. As there were 300 samples in our training set, this amounted to fitting a GMM to 1.5 million cells. The time in seconds was determined in function of K (Fig. S3A). Most importantly, the entire analysis remained under 1 h. Training a Random Forest model on the fitted GMM resulted in an average increase of 24.4% of the runtime for $K = 256$ (Fig. S3B).

In order to provide guidance concerning use of the model, the most important parameters were varied one by one (i.e., the number of included detectors D, the number of mixtures K, the number of cells sampled per file to fit a GMM denoted as N_CELLS_MIN, the number of cells sampled per individual sample to determine the cell counts per mixture denoted as N_CELLS_REP, a learning curve in function of N_SAMPLES, and the TYPE of covariance matrix used to fit a GMM). The performance was quantified using R^2 ($D_{q=1}$), based on the communities for a concentration parameter of $a = 1$, for the same Random Forest analysis as described above (Fig. S4). The results indicated that:

- Including additional detectors improved the performance.
- Generally, the higher the number of mixtures K, the better the performance, which saturated after a specific threshold.
- PhenoGMM was quite robust for the number of included cells to fit a GMM.
- PhenoGMM was quite robust for the number of included cells per sample.
- The predictive performance did not saturate yet at after 300 samples.
- PhenoGMM was quite robust for the type of used covariance matrix, although the "full" type (i.e., each mixture has its own covariance matrix) resulted in the best predictions.

PhenoGMM retrieves the community structure of natural freshwater microbial communities. In the second experiment, we evaluated whether and to what extent it was possible to quantify the diversity of natural freshwater microbial communities using FCM in combination with PhenoGMM. We used two data sets, of which the first describes the dynamics of a cooling water microbiome during two surveys of a research nuclear reactor (surveys I and II) and the second describes the microbiomes of three different freshwater lake systems (i.e., Michigan inland lakes ["Inland"], Lake Michigan, and Muskegon Lake, respectively). The same approach as before was applied, and PhenoGMM and PhenoGrid were compared. Samples were first aggregated to determine a fingerprint template based on either a GMM or a gridded binning approach. Next, cell counts per mixture or bin were retrieved per sample, based on which the α-diversity

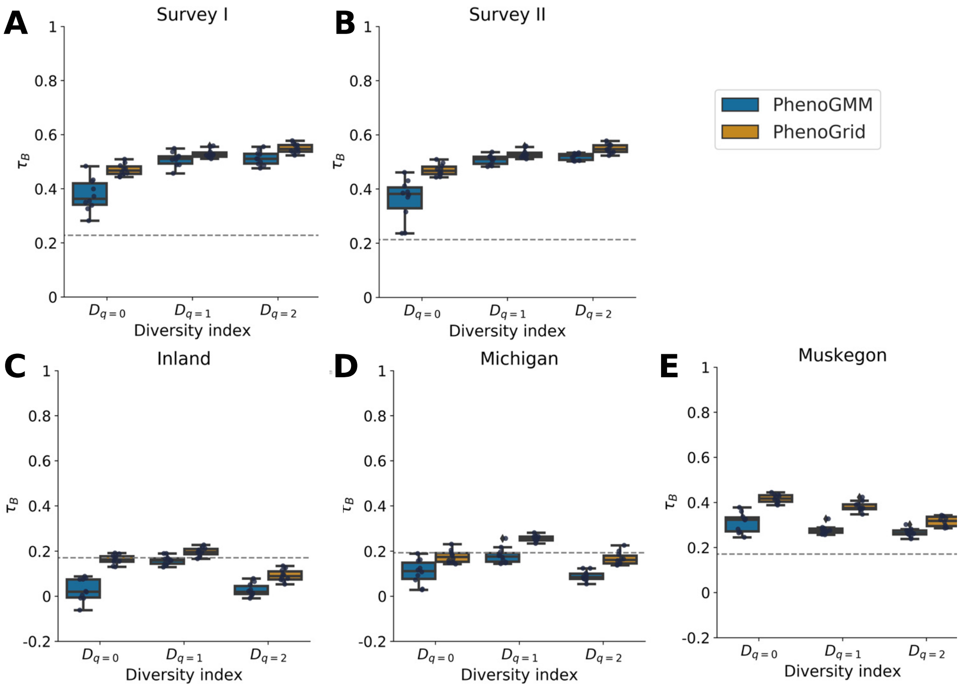


FIG 4 Summary of α-diversity estimations for the cooling water and freshwater lake microbiomes, evaluated by Kendall's τ_B, using PhenoGMM and PhenoGrid. Both methods were run 10 times. Kendall's τ_B was calculated between true and estimated diversity values. Each boxplot displays the 25% and 75% quartiles of the τ_B, and the whiskers show the full range of τ_B. Each dot represents the resulting value from an individual run. (A and B) Results for the cooling water microbiome. (A) Survey I. (B) Survey II. (C to E) Results for the freshwater lake microbiome. (C) Inland lakes. (D) Lake Michigan. (E) Muskegon Lake. The dashed line indicates the strength of τ_B at $P = 0.05$.

values were calculated. To estimate how well both methods were able to retrieve the taxonomic structure of the microbial community, these values were compared with α-diversity estimations based on 16S rRNA gene amplicon sequencing. The correspondence was again evaluated using Kendall's τ_B and summarized in Fig. 4.

Diversity estimations were highly significant for the cooling water microbiome for both approaches. The α-diversity of the microbial communities in Muskegon Lake could be successfully retrieved as well. For $q = 1$ (the exponent of the Shannon entropy), estimations were significant based on PhenoGrid, but not for PhenoGMM. In most cases, PhenoGrid outperformed PhenoGMM, indicating that more mixtures or additional detectors might be needed to make it competitive with PhenoGrid in this setting. To summarize, PhenoGMM successfully quantified the community structure of most considered natural communities, but its ability depended on the ecosystem of study and its specific implementation. In the current implementation, PhenoGrid seems to be favored, with small to moderate differences between the two approaches.

PhenoGMM quantifies intercommunity differences. We also evaluated the possibility to quantify intercommunity diversity (i.e., β-diversity) for both approaches. We used the Bray-Curtis dissimilarity to quantify these differences based on the resulting cell contingency tables for each data set. A Mantel test was used to calculate the correlation between the dissimilarity matrix based on the cytometric fingerprints and the one derived from the *in silico* synthetic microbial community composition or the composition based on 16S rRNA gene sequencing for the cooling water and freshwater lake data sets. This was done for PhenoGMM and PhenoGrid (Fig. 5). Both approaches resulted in strong correlations ($P < 0.001$ for all considered communities, except Lake Michigan, for which $P < 0.05$). PhenoGMM-based fingerprints resulted in higher correlations for the simulated synthetic microbial communities, Lake Michigan and

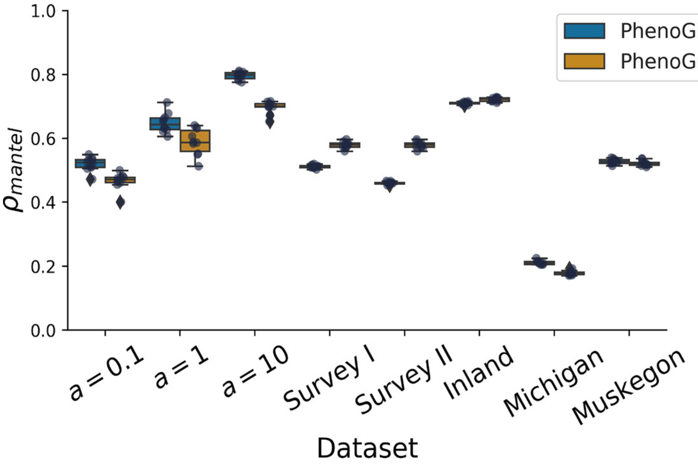


FIG 5 Summary of β-diversity estimations for all data sets, evaluated by ρ_{mantel}, for both PhenoGMM and PhenoGrid. Both methods were run 10 times. ρ_{mantel} was calculated between the Bray-Curtis dissimilarity matrices based on cytometric fingerprints and the simulated synthetic community composition or 16S rRNA gene amplicon sequencing (cooling water and freshwater lake communities). Each boxplot displays the 25% and 75% quartiles of ρ_{mantel}, and the whiskers show the full range of ρ_{mantel}.

Muskegon Lake, while PhenoGrid-based fingerprints resulted in higher correlations for the cooling water and inland lake system microbiome.

DISCUSSION

In this paper we propose a data-driven cytometric fingerprinting strategy based on Gaussian mixture models (GMMs), which we have called PhenoGMM. Our approach allows the derivation of information-rich variables from microbial cytometry data in order to describe the community structure. One of its advantages is that the method reduces the number of community-describing variables considerably compared to traditional binning approaches. We evaluated the performance of PhenoGMM in terms of the α-diversity, as quantified by the Hill numbers D_q for $q = 0$, 1, and 2. These are the equivalents of the richness, exponent of the Shannon entropy, and inverse Simpson index. We also evaluated to what extent intercommunity differences can be quantified to perform β-diversity estimations. Both synthetic and natural microbial communities were considered. We compared PhenoGMM with the performance of a generic traditional binning approach that is representative for common approaches for cytometry fingerprinting in microbial ecology, which we have called PhenoGrid for this work.

In the first part of the paper, we constructed synthetic microbial communities *in silico* by aggregating cytometric characterizations of individual bacterial strains according to predefined compositions. This allowed us to simulate microbial community compositions in a highly precise and controlled way. These predefined compositions were used to calculate α- and β-diversity values. Cytometric diversity, based on the resulting cell counts for PhenoGMM and PhenoGrid, was benchmarked with the predefined values, by calculating Kendall's rank correlation coefficient τ_B between the two sets of values. Both approaches resulted in moderate to strong correlations. PhenoGMM resulted in stronger or equally accurate estimations compared to PhenoGrid. The exponent of the Shannon entropy ($D_{q=1}$) and the inverse Simpson index ($D_{q=2}$) were better estimated compared to the richness of the community, indicating that community structure rather than identity is captured by the fingerprints. The total analysis time of PhenoGMM remained under 1 h for the analysis of 1.5 million cells.

In the second part, we evaluated to what extent PhenoGMM was able to retrieve the structure of natural communities. Two types of ecosystems were considered, the cooling water microbiome during two surveys of a research nuclear reactor and the microbiome of three freshwater lake systems. Correlations with the taxonomic diversity based on 16S rRNA gene amplicon sequencing data have been demonstrated in previous work (13, 25) and were therefore used as a benchmark. Depending on the ecosystem of study, correlations of different strengths were observed. Considering the cooling water microbiome, moderate to strong correlations were reported for both surveys. Differences between PhenoGMM and PhenoGrid were small, but results were in favor of PhenoGrid. When considering freshwater lakes, only the data from Muskegon Lake resulted in significant correlations for PhenoGrid. The exponent of the Shannon entropy resulted in significant correlations as well for the other two lake systems, indicating again that cytometric fingerprinting approaches capture community structure rather than identity.

Note that we do not expect to find a "perfect" correlation between the cytometric and taxonomic diversity. Besides the fact that 16S rRNA gene amplicon sequencing is subject to a number of biases (35, 36), microbial FCM is sensitive to both taxonomic and physiological changes. Therefore, the strength of the correspondence between cytometric and taxonomic diversity will vary from experiment to experiment and from system to system, with multiple factors affecting the strength of the correspondence. First, the freshwater lake microbiome displays larger values in richness and evenness compared to the cooling water microbiome (25). Second, the levels of trophicity differ between the considered data sets, which could be affecting the estimations. Third, the sampling coverage is different between the data sets. The cooling water microbiome contains many measurements over a few days in a highly dynamic system, compared to the freshwater lake microbiome that contains samples spanning a much larger range in time (years) and space (multiple locations).

Estimations of β-diversity (i.e., intercommunity diversity) could be successfully quantified as well, by calculating Bray-Curtis dissimilarities between the cytometric fingerprints of different communities. A Mantel test demonstrated that correlations were significant for all data sets and strong in most cases, indicating that in some cases it could be more worthwhile to investigate inter- rather than intracommunity differences.

Few reports exist that quantitatively evaluate fingerprinting approaches for the analysis of microbial cytometry data. Most fingerprinting strategies make use of manual annotation of clusters or of fixed binning approaches (see, e.g., the report by Koch et al. [14] which qualitatively discusses different existing methods). In almost all cases, only bivariate interactions are inspected. PhenoGMM allows modeling the full parameter space at once. This is interesting, because although it is difficult to develop multicolor approaches for bacterial analyses, these are possible (see, e.g., the work by Barbesti et al. [37]). In addition, our research group has demonstrated that additional detectors that capture signals due to spillover can assist in the discrimination between bacterial species (38). Therefore, the parameter space in which bacterial cells can be described is increasing, and PhenoGMM is able to model this straightforwardly. Because it is an adaptive strategy as well, by defining small clusters in regions of high density and vice versa, it reduces the number of sample-describing variables considerably compared to fixed binning approaches. In that sense, it shares some properties with FlowFP. This is an adaptive binning approach, in which bins are smaller when the density of the data is higher and vice versa. However, the bins are still hyperrectangular in shape, while PhenoGMM allows clusters to be of any shape. Other adaptive binning strategies have been proposed for microbial FCM data as well (24); however, these are still limited to bivariate interactions.

Our approach comes with a number of caveats. First, PhenoGMM fits a fingerprint template based on the concatenation of measured samples. New samples are characterized based on this template. In the case that multiple samples diverge considerably from those that were used to determine the template (for example, in the case that an experiment was conducted under different conditions), we recommend refitting the model. Second, we overcluster the data to model the multiple and potentially overlapping cell distributions due to the differences in physiology and the many species that can be present in a microbial community. This makes it difficult to determine the exact number of mixtures. As the number of mixtures *K* increases, the performance saturates gradually, and more mixtures will not improve estimations. Therefore, *K* should be chosen high enough but might differ from experiment to experiment. PhenoGMM can also be tailored toward the identification of separated cell populations, for example, to identify phytoplankton populations (33), for the identification of so-called high- and low-nucleic-acid groups (39), or to identify distinct bacterial populations when the resolution of the data is high enough (34). In this case, if the number of populations is known beforehand, *K* can be chosen accordingly; if this is not known, one can use decision rules such as the Bayesian information criterion (BIC) to determine the optimal number of mixtures (34). Third, due to overclustering of the data, mixtures can be highly correlated with each other.

Our *in silico* benchmark study made use of cytometric characterizations of individual bacterial strains. Individual cultures are known to exhibit considerable heterogeneity due to cell size diversity and cell cycle variations (40). Our research group has recently shown that the cytometric diversity of an individual culture reduces when it is part of a coculture (21). Therefore, data used for the *in silico* community creation setup cannot be used to study environmental samples, as we hypothesize that members of natural communities will have a different cytometric fingerprint from strains that were grown and measured individually. Yet, we believe that our *in silico* community assembly approach is useful, as it allows a precise simulation of variations in cytometric community structure.

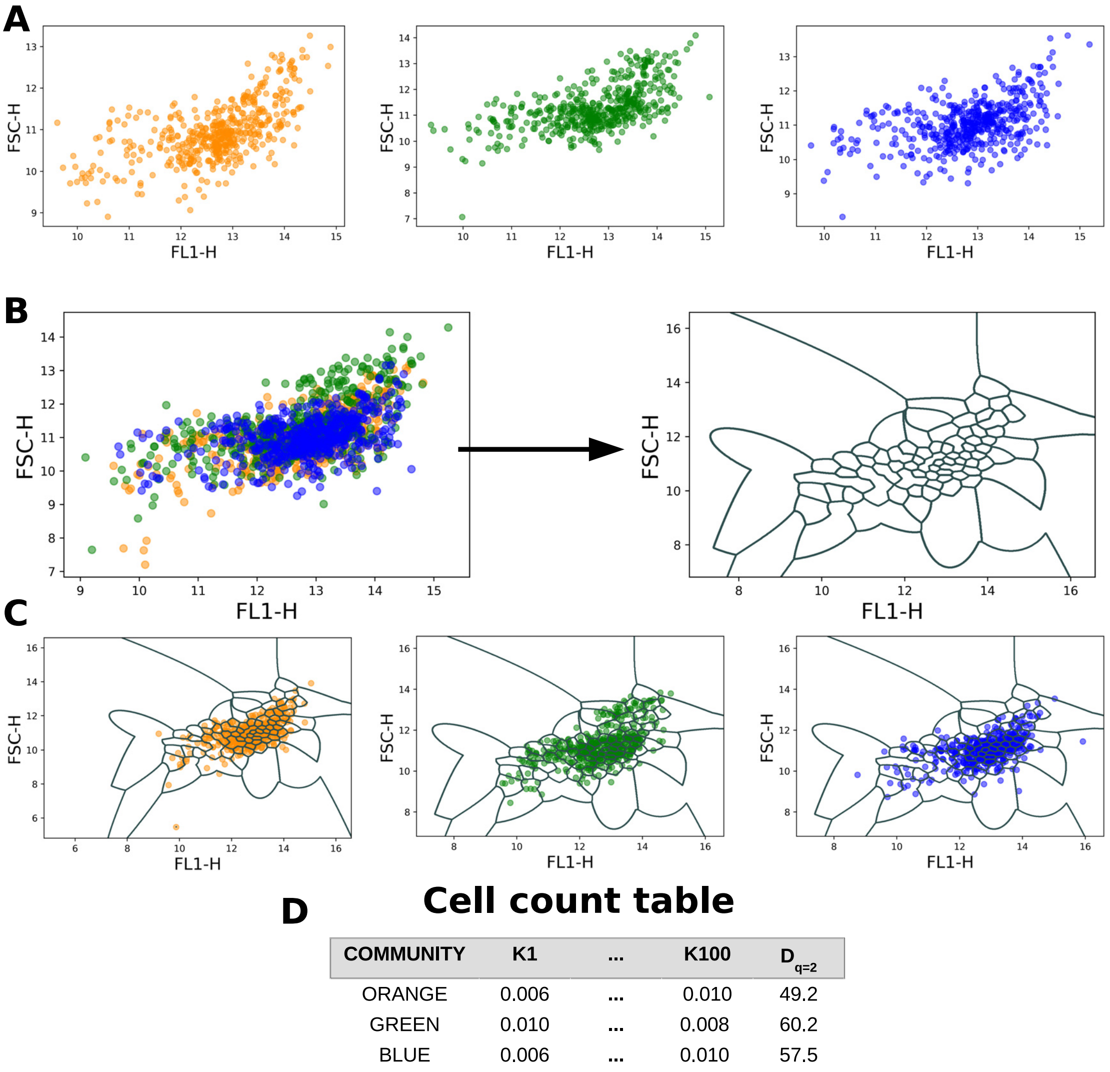
To conclude, PhenoGMM can be used to derive information-rich variables from microbial FCM measurements. Microbial community structure can be quantified by computing cytometric diversity metrics based on the PhenoGMM-based fingerprints. The method has a number of advantages compared to traditional cytometric fingerprinting approaches. To facilitate its use by the scientific community, it has been integrated in the R package PhenoFlow (13). Technological advancements have enabled an automated data acquisition, resulting in a detailed characterization of the microbial community online (i.e., samples are measured at routine intervals between 5 and 15 min) or in real time (i.e., near-continuous measurements) (41, 42). Therefore, we see great potential to use FCM as a monitoring technique to rapidly and frequently investigate microbial community dynamics, which can be supported by PhenoGMM. It has to be noted that quantification of diversity should serve as a starting point to test ecological hypotheses rather than as a final outcome of an experiment (43). Microbial FCM, in combination with PhenoGMM, has the potential to be an effective strategy to serve this research line in microbial ecology.

MATERIALS AND METHODS

Methodology. In this work, multiple data sets were analyzed. Each data set contained multiple FCM samples, either individual bacterial strains or natural communities. Bacterial cells were described by scatter and fluorescence signals, for which the latter resulted from the use of a nucleic acid stain (SYBR green I). Experimental details per data set are laid out in detail below. We first describe the methodology of PhenoGMM.

Preprocessing. Two preprocessing steps are applied to all cytometry samples before further analysis of the data. First, all individual FCM channels are transformed by $f(x) = \mathrm{asinh}(x)$. Next, background due to debris and noise is removed using a fixed digital gating strategy (13, 44). In other words, a single gate is applied to separate bacterial cells from background and is used for all samples. This gate is fixed within a specific experiment but can differ from data set to data set.

Cytometric fingerprinting using Gaussian mixture models. When the preprocessing is completed, a fingerprint template or model needs to be determined that is able to describe all the samples within an experiment. Therefore, samples are first subsampled to the same number of cells per sample (N_CELLS_MIN), in order to not bias the Gaussian mixture model (GMM) toward a specific sample, and concatenated in a training set. This number can be either the lowest number of cells present in one sample or a number of choice. A rough guideline can be to not let the training set be larger than 1×10^6 cells, but this depends on computational resources. If *n* denotes the total number of samples, then the total number of cells (N_CELLS) in the training set will be determined as N_CELLS = $n \times$ N_REP $\times$ N_CELLS_MIN, in which N_REP denotes the number of technical replicates of a specific sample. Typically, forward (FSC) and side (SSC) scatter channels are included, along with one or more targeted fluorescence channels (denoted as FL*X*, in which *X* indicates the number of a specific fluorescence



COMMUNITY	K1	...	K100	$D_{q=2}$
ORANGE	0.006	...	0.010	49.2
GREEN	0.010	...	0.008	60.2
BLUE	0.006	...	0.010	57.5

FIG 6 Illustration of PhenoGMM for two channels (FL1-H and FSC-H) using $K = 100$ mixtures. (A) The analysis starts from cytometric measurements of three bacterial communities of interest, noted as "ORANGE" ($S = 6$), "GREEN" ($S = 8$), and "BLUE" ($S = 15$). (B) Data for the three communities are concatenated into one data frame, to which a GMM with (in this case) $K = 100$ mixtures is fitted. This results in a fingerprint template, which is depicted on the right. (C) The fingerprint template is used to derive relative cell counts per cluster and per bacterial community. (D) This results in a "count" table, which can be used to rapidly quantify the cytometric diversity based on equations 2 to 4 (in this case D_2).

detector). Unless noted otherwise, channels FSC-H, SSC-H, and FL1-H (530/30 nm) are included for data analysis.

Once this training set is created, a GMM of K mixtures is fitted to the data. If $\mathbf{X}$ denotes the entire data matrix or training set containing N cells, then $\mathbf{X}$ consists of cells written as $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$, of which each cell is described by D variables (i.e., the number of signals collected from the flow cytometer). Cell i is described as $\boldsymbol{x}_i = \{x_{i1}, \ldots, x_{iD}\}$. A GMM consists of a superposition of normal distributions N, of which each distribution has its own mean μ and covariance matrix Σ. Each mixture has a mixing coefficient or weight π, which represents the fraction of data each mixture is describing. The distribution p can be written as follows:

$$p(\mathbf{X}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{X}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k). \tag{1}$$

The set of parameters $\Theta = [\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k]_{k=1}^{K}$ is estimated by the expectation-maximization (EM) algorithm (45). Once a GMM has been trained on the concatenated data, the fingerprint template is determined and one can assign all cells per sample to the mixture for which it has the highest posterior probability. This time, replicate samples are subsampled to a specific number of cells or the lowest number of cells of the replicates that are part of that specific sample and pooled. This number is denoted as

N_CELLS_REP. After clustering, the number of cells per mixture is counted, after which the relative number of cells per mixture and sample is retrieved, further defined as the cytometric fingerprints. An illustration of PhenoGMM can be seen in Fig. 6.

We used the GaussianMixture() function of the scikit-learn machine learning library to implement our method (46). This function contains four different ways to estimate the covariance matrix of each mixture:

- diag: each mixture has its own diagonal covariance matrix.
- full: each mixture has its own general covariance matrix.
- spherical: each mixture has its own single variance. (Note this is not the same as *k*-means clustering. In this case, all mixtures would share the same single variance.)
- tied: all mixtures share the same general covariance matrix.

Unless otherwise noted, we let each mixture have its own general covariance matrix (full). mClust was used to integrate PhenoGMM in the R package PhenoFlow (47).

Defining cytometric diversity. Cytometric fingerprints allow definition of the cytometric diversity of a microbial community (13, 18). If one considers each predefined gate or mixture as a phenotypic unit, one can calculate both intra- and intercommunity diversity metrics, also known as α- and β-diversity. The first quantifies the diversity within a sample, and the latter quantifies the diversity between samples. Various diversity metrics exist in ecology to calculate α-diversity. In this work, we apply the Hill numbers $D_q(\mathbf{p})$ to quantify community diversity (48), as proposed by Leinster and Cobbold (49) and Daly et al. (50). If we let $\mathbf{p} = p_1, \ldots, p_S$ represent the vector of relative abundances, describing the abundance of S bacterial species or populations, then we can define the richness ($D_{q=0}$) and evenness ($D_{q=1}$, $D_{q=2}$) of a microbial community as follows:

$$D_{q=0}(\mathbf{p}) = S, \tag{2}$$

$$D_{q=1}(\mathbf{p}) = \exp(-\sum_{i=1}^{S} p_i \ln p_i), \tag{3}$$

$$D_{q=2}(\mathbf{p}) = \frac{1}{\sum_{i=1}^{S} p_i^2}. \tag{4}$$

q denotes the order of the Hill-number, which is part of a general family that can be denoted as $D_q(\mathbf{p})$, and expresses the weight that is given to more abundant species. $D_{q=1}$ is the equivalent of the exponential of the Shannon entropy, and $D_{q=2}$ is the equivalent of the inverse Simpson index (50).

β-Diversity metrics quantify the difference in compositions between different communities. We quantify the dissimilarity between samples using the Bray-Curtis dissimilarity (51). If we let BC_{AB} denote the dissimilarity between communities A and B, BC_{AB} is calculated using the following equation:

$$BC_{AB} = \frac{\sum_{i=1}^{S} |p_{A,i} - p_{B,i}|}{\sum_{i=1}^{S} |p_{A,i} + p_{B,i}|} \tag{5}$$

Predictive modeling. FCM fingerprints can be used as input variables to train a machine learning model. We use Random Forest regression (52), an ensemble of decision trees, to predict α-diversity values, based on the *in silico* assembling strategy to estimate the structure of synthetic microbial communities (see below). A randomized grid search is performed to search for an optimal hyperparameter combination (53). This means that a number of random combinations of hyperparameter values were evaluated. The maximum number of variables that are considered at an individual split for a decision tree is randomly drawn from {1, . . ., *K*}, and the minimum number of samples for a specific leaf is randomly drawn between {1, . . ., 5}. One hundred different combinations were evaluated using 5-fold cross-validation, and predictions were reported for a separate test set.

Data sets. (i) *In silico* synthetic bacterial communities. Data from 20 individual bacterial strains, which were grown in the laboratory and measured by FCM, were collected from reference 10. In brief, individual bacterial cultures were sampled after 24 h of incubation and stained with SYBR green I, and two technical replicates per strain were measured on an Accuri C6 (BD Biosciences). Fluorescence was measured by the targeted detector (FL1, 530/30 nm) and three additional detectors, next to forward (FSC) and side (SSC) scatter. After background removal, additional automated denoising was performed using the FlowAI package (v1.4.4., default settings; target channel, FL1; changepoint detection, 150 [54]). A full experimental overview can be found in reference 10. The lowest number of cells collected after background removal amounted to 13,166 cells. The data are available via FlowRepository (accession ID: FR-FCM-ZZSH).

(ii) Cooling water microbiome. Data were used as presented in reference 13. Samples were collected from the cooling water of a discontinuously operated research nuclear reactor. This reactor underwent four phases: control, startup, operational, and shutdown. Samples were taken from two surveys in time (surveys I and II) and analyzed via FCM and 16S rRNA gene amplicon sequencing ($n_{\text{survey I}} = 36$ and

$n_{\text{survey II}}$ = 31). The procedure and data preprocessing are described in reference 13. In brief, samples were stained with SYBR green I, and three technical replicates were analyzed using an Accuri C6 (BD Biosciences). Fluorescence was measured by the targeted detector (FL1, 530/30 nm) and three additional detectors, next to forward (FSC) and side (SSC) scatter. The data are available via FlowRepository (accession ID: FR-FCM-ZZNA). The lowest number of cells collected after background removal amounted to 10,565 cells. Taxonomic identification of microbial communities was done at the operational taxonomic unit (OTU) level at 97% similarity after preprocessing. All the samples were subsampled down to the minimum sequencing depth and normalized afterward. Sequences are available from the NCBI Sequence Read Archive (SRA) under accession ID SRP066190.

(iii) Freshwater lake microbiome. Data were collected as presented in reference 55. A total of 173 samples, from three types of freshwater lake systems, were analyzed through 16S rRNA gene amplicon sequencing and FCM. Samples originated from three different freshwater lake systems: (i) 49 samples from Lake Michigan (2013 and 2015), (ii) 62 samples from Muskegon Lake (2013 to 2015; one of Lake Michigan's estuaries); and (iii) 62 samples from 12 inland lakes in southeastern Michigan (2014 to 2015). Field sampling, DNA extraction, DNA sequencing, and processing are described in reference 56. Fastq files were submitted to NCBI SRA under BioProject accession numbers PRJNA412984 and PRJNA414423. Taxonomic identification of microbial communities was done for each of the three lake data sets separately and treated with a minimum OTU abundance threshold cutoff of one sequence in 3% of the samples. Sequences were clustered into OTUs at 97% similarity. Each of the three data sets was rarefied to an even sequencing depth, which was 4,491 sequences for Muskegon Lake samples, 5,724 sequences for the Lake Michigan samples, and 9,037 sequences for the inland lake samples. The relative abundance at the OTU level was calculated by taking the count value and dividing it by the sequencing depth of the sample. Flow cytometry procedures are described in reference 25. In brief, samples were stained with SYBR green I, and three technical replicates were measured on an Accuri C6 (BD Biosciences). Fluorescence was measured by the targeted detector (FL1, 530/30 nm) and three additional detectors, next to forward (FSC) and side (SSC) scatter. FCM data are available via FlowRepository (accession IDs: FR-FCM-ZY9J and FR-FCM-ZYZN). The lowest number of cells collected after denoising amounted to 2,342 cells.

Method evaluation. Our proposed fingerprinting approach based on GMMs was compared to a generic fixed binning approach, which we have called PhenoGrid. In brief, we implemented a binning grid of $L = 128 \times 128$ for each bivariate FCM channel combination, after which cell fractions per bin were determined. The resulting cell fractions were next vectorized, concatenated, and normalized. Both PhenoGMM and PhenoGrid result in multiple variables that describe relative cell counts, either per mixture or per bin. These methods were evaluated to estimate the structure of both synthetic and natural communities.

(i) α-Diversity estimations of *in silico* synthetic microbial communities. In the first setup, we assessed how well PhenoGMM was able to capture variations in the structure of synthetic microbial communities. To do so, we first performed an *in silico* community assembly strategy. In other words, cytometric characterizations of individual bacterial strains were artificially aggregated according to simulated compositions (10). These compositions were determined according to the following strategy:

1. Sample at random a number S'_i that represents the number of different members that will constitute the microbial community *i*. S'_i lies between two and 20 (the total number of strains that are available).
2. The Dirichlet distribution can be used to model the joint distribution of individual fractions of multiple species (57). We applied the Dirichlet distribution to randomly simulate the composition of microbial community *i*. The evenness of the composition depends on the concentration parameter *a* of the Dirichlet distribution, which determines how evenly the weight will be spread over multiple community members. If *a* is low, only a few members will make up a large part of the community (low evenness). If *a* is high, the fraction of each member contributing to the community composition will be close to equal (high evenness).

Four hundred community compositions (300 training and 100 test communities) were simulated for three different values of $a = 0.1$, 1, and 10. The simulated compositions were visualized using Lorenz curves (see Fig. S1 in the supplemental material). Next, *in silico* synthetic bacterial communities were assembled by aggregating the cytometric characterizations of individual bacterial strains according to these simulated compositions. Diversity values could be calculated with high accuracy based on the simulated compositions by calculating the Hill numbers for $q = 0$, 1, and 2 and Bray-Curtis dissimilarities for these compositions. These were then correlated with the relative cell abundances that resulted from PhenoGMM and PhenoGrid. The strength of the correlation was assessed using Kendall's τ_B and a Mantel test. The Random Forest prediction experiment was additionally evaluated using the R_2 (see below).

(ii) α-Diversity estimations of natural microbial communities. Cytometric diversity estimations for natural communities (i.e., the cooling water and freshwater lake microbiome) were evaluated in a different way. To benchmark PhenoGMM, these values were correlated with α- and β-diversity values based on 16S rRNA gene amplicon sequencing, motivated by previous reported correlations between the cytometric and taxonomic diversity (13, 25). The strength of the correlation was assessed using Kendall's τ_B and a Mantel test (see below).

(iii) Performance evaluation.

- α-Diversity estimations were quantified by calculating Kendall's rank correlation coefficient τ_B between the true and estimated values. The τ_B implementation, which is able to deal with ties, was calculated as follows:

$$\tau_B = \frac{N_c - N_d}{\sqrt{(N_c + N_d + N_t) \times (N_c + N_d + N_u)}}. \qquad (6)$$

  N_c denotes the number of concordant pairs between true and predicted values, N_d the number of discordant pairs, N_t the number of ties in the true values, and N_u the number of ties in the predicted values. Values range from -1 (perfect negative association) to $+1$ (perfect positive association), and a value of 0 indicates the absence of an association. This was done using the KENDALLTAU() function in Scipy (v1.0.0) (59).
- Random Forest predictions were evaluated by calculating the R^2 between true ($\mathbf{y} = \{y_1, \ldots, y_n\}$) and predicted ($\hat{\mathbf{y}} = \{\hat{y}_1, \ldots, \hat{y}_n\}$) values:

$$R^2(\mathbf{y}, \hat{\mathbf{y}}) = 1 - \frac{\sum_{i=0}^{n-1} (y_i - \hat{y}_i)^2}{\sum_{i=0}^{n-1} (y_i - \bar{y})^2}, \qquad (7)$$

  in which y denotes the average value of $\mathbf{y}$. If $R^2 = 1$, predictions are correctly estimated. If $R_2 < 0$, predictions are worse than random guessing. The R2_SCORE()-function from the scikit-learn machine learning library was used (46).
- β-Diversity estimations were evaluated by calculating the correlation between Bray-Curtis dissimilarity matrices (BC) based on FCM and 16S rRNA gene sequencing data using a Mantel test (58). This test assesses the alternative hypothesis that the distances between samples based on cytometry data are linearly correlated with those based on 16S rRNA gene sequencing data. It makes use of the cross-product term Z_M across the two matrices for each element ij:

$$Z_M = \sum_{i=1}^{n} \sum_{j=1}^{n} BC_{ij}^{\mathrm{FCM}} \times BC_{ij}^{\mathrm{16S}}. \qquad (8)$$

- The test statistic Z_M is normalized and then compared to a null distribution, based on 1,000 permutations.

Code and data availability. All code and data supporting this article are freely available on GitHub at https://github.com/prubbens/PhenoGMM. The functionality of PhenoGMM has been incorporated in the R package PhenoFlow: https://github.com/CMET-UGent/Phenoflow_package. Raw flow cytometry data are freely available on FlowRepository (accession numbers FR-FCM-ZZSH, FR-FCM-ZZNA, FR-FCM-ZY9J, and FR-FCM-ZYZN). Raw sequences are available via the NCBI Sequence Read Archive (accession numbers SRP066190, PRJNA412984, and PRJNA414423).

SUPPLEMENTAL MATERIAL

Supplemental material is available online only.

FIG S1, PDF file, 2 MB.

FIG S2, PDF file, 0.4 MB.

FIG S3, PDF file, 0.5 MB.

FIG S4, PDF file, 2.1 MB.

ACKNOWLEDGMENTS

The computational resources (Stevin Supercomputer Infrastructure) and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by Ghent University, the Hercules Foundation, and the Flemish Government department EWI.

P.R. is supported by Ghent University (BOFSTA2015000501). R.P. is supported by Ghent University (BOFDOC2015000601). W.W. also received funding from the Flemish Government under the Onderzoeksprogramma Artificiele Intelligentie (AI) Vlaanderen program.

REFERENCES

1. Van Dijk E, Auger H, Jaszczyszyn Y, Thermes C. 2014. Ten years of next-generation sequencing technology. Trends Genet 30:418–426. https://doi.org/10.1016/j.tig.2014.07.001.

2. Ackermann M. 2015. A functional perspective on phenotypic heterogeneity in microorganisms. Nat Rev Microbiol 13:497–508. https://doi.org/10.1038/nrmicro3491.

3. De Vrieze J, Boon N, Verstraete W. 2018. Taking the technical microbiome into the next decade. Environ Microbiol 20:1991–2000. https://doi.org/10.1111/1462-2920.14269.
4. Aghaeepour N, Finak G, Hoos H, Mosmann TR, Brinkman R, Gottardo R, Scheuermann RH, DREAM Consortium. 2013. Critical assessment of automated flow cytometry data analysis techniques. Nat Methods 10:228–238. https://doi.org/10.1038/nmeth.2365.
5. Weber LM, Robinson MD. 2016. Comparison of clustering methods for high-dimensional single-cell flow and mass cytometry data. Cytometry A 89A:1084–1096. https://doi.org/10.1002/cyto.a.23030.
6. Robinson JP. 2018. Overview of flow cytometry and microbiology. Curr Protoc Cytom 84:e37. https://doi.org/10.1002/cpcy.37.
7. Koch C, Müller S. 2018. Personalized microbiome dynamics - cytometric fingerprints for routine diagnostics. Mol Aspects Med 59:123–134. https://doi.org/10.1016/j.mam.2017.06.005.
8. Buysschaert B, Byloos B, Leys N, Van Houdt R, Boon N. 2016. Reevaluating multicolor flow cytometry to assess microbial viability. Appl Microbiol Biotechnol 100:9037–9051. https://doi.org/10.1007/s00253-016-7837-5.
9. Cichocki N, Hübschmann T, Schattenberg F, Kerckhof F-M, Overmann J, Müller S. 2020. Bacterial mock communities as standards for reproducible cytometric microbiome analysis. Nat Protoc 15:2788–2812. https://doi.org/10.1038/s41596-020-0362-0.
10. Rubbens P, Props R, Boon N, Waegeman W. 2017. Flow cytometric single-cell identification of populations in synthetic bacterial communities. PLoS One 12:e0169754. https://doi.org/10.1371/journal.pone.0169754.
11. Wilkins MF, Hardy SA, Boddy L, Morris CW. 2001. Comparison of five clustering algorithms to classify phytoplankton from flow cytometry data. Cytometry 44:210–217. https://doi.org/10.1002/1097-0320(20010701)44:3<210::AID-CYTO1113>3.0.CO;2-Y.
12. Koch C, Günther S, Desta AF, Hübschmann T, Müller S. 2013. Cytometric fingerprinting for analyzing microbial intracommunity structure variation and identifying subcommunity function. Nat Protoc 8:190–202. https://doi.org/10.1038/nprot.2012.149.
13. Props R, Monsieurs P, Mysara M, Clement L, Boon N. 2016. Measuring the biodiversity of microbial communities by flow cytometry. Methods Ecol Evol 7:1376–1385. https://doi.org/10.1111/2041-210X.12607.
14. Koch C, Harnisch F, Schröder U, Müller S. 2014. Cytometric fingerprints: evaluation of new tools for analyzing microbial community dynamics. Front Microbiol 5:273. https://doi.org/10.3389/fmicb.2014.00273.
15. Günther S, Koch C, Hübschmann T, Röske I, Müller RA, Bley T, Harms H, Müller S. 2012. Correlation of community dynamics and process parameters as a tool for the prediction of the stability of wastewater treatment. Environ Sci Technol 46:84–92. https://doi.org/10.1021/es2010682.
16. García FC, Alonso-Sáez L, Morán XAG, López-Urrutia Á. 2015. Seasonality in molecular and cytometric diversity of marine bacterioplankton: the re-shuffling of bacterial taxa by vertical mixing. Environ Microbiol 17:4133–4142. https://doi.org/10.1111/1462-2920.12984.
17. Koch C, Fetzer I, Harms H, Müller S. 2013. CHIC—an automated approach for the detection of dynamic variations in complex microbial communities. Cytometry A 83A:561–567. https://doi.org/10.1002/cyto.a.22286.
18. Li WKW. 1997. Cytometric diversity in marine ultraphytoplankton. Limnol Oceanogr 42:874–880. https://doi.org/10.4319/lo.1997.42.5.0874.
19. Buysschaert B, Kerckhof F-M, Vandamme P, De Baets B, Boon N. 2018. Flow cytometric fingerprinting for microbial strain discrimination and physiological characterization. Cytometry A 93A:201–212. https://doi.org/10.1002/cyto.a.23302.
20. Melzer S, Winter G, Jäger K, Hübschmann T, Hause G, Syrowatka F, Harms H, Tárnok A, Müller S. 2015. Cytometric patterns reveal growth states of Shewanella putrefaciens. Microb Biotechnol 8:379–391. https://doi.org/10.1111/1751-7915.12154.
21. Heyse J, Buysschaert B, Props R, Rubbens P, Skirtach AG, Waegeman W, Boon N. 2019. Coculturing bacteria leads to reduced phenotypic heterogeneities. Appl Environ Microbiol 85:e02814-18. https://doi.org/10.1128/AEM.02814-18.
22. De Roy K, Clement L, Thas O, Wang Y, Boon N. 2012. Flow cytometry for fast microbial community fingerprinting. Water Res 46:907–919. https://doi.org/10.1016/j.watres.2011.11.076.
23. Props R, Rubbens P, Besmer M, Buysschaert B, Sigrist J, Weilenmann H, Waegeman W, Boon N, Hammes F. 2018. Detection of microbial disturbances in a drinking water microbial community through continuous acquisition and advanced analysis of flow cytometry data. Water Res 145:73–82. https://doi.org/10.1016/j.watres.2018.08.013.
24. Amalfitano S, Fazi S, Ejarque E, Freixa A, Romaní AM, Butturini A. 2018. Deconvolution model to resolve cytometric microbial community patterns in flowing waters. Cytometry A 93A:194–200. https://doi.org/10.1002/cyto.a.23304.
25. Props R, Schmidt ML, Heyse J, Vanderploeg HA, Boon N, Denef VJ. 2018. Flow cytometric monitoring of bacterioplankton phenotypic diversity predicts high population-specific feeding rates by invasive dreissenid mussels. Environ Microbiol 20:521–534. https://doi.org/10.1111/1462-2920.13953.
26. Li WKW. 2002. Macroecological patterns of phytoplankton in the northwestern North Atlantic Ocean. Nature 419:154–157. https://doi.org/10.1038/nature00994.
27. van Gelder S, Röhrig N, Schattenberg F, Cichocki N, Schumann J, Schmalz G, Haak R, Ziebolz D, Müller S. 2018. A cytometric approach to follow variation and dynamics of the salivary microbiota. Methods 134-135:67–79. https://doi.org/10.1016/j.ymeth.2017.08.009.
28. Menyhárt L, Nagy S, Lepossa A. 2018. Rapid analysis of photoautotroph microbial communities in soils by flow cytometric barcoding and fingerprinting. Appl Soil Ecol 130:237–240. https://doi.org/10.1016/j.apsoil.2018.06.013.
29. Zimmermann J, Hübschmann T, Schattenberg F, Schumann J, Durek P, Riedel R, Friedrich M, Glauben R, Siegmund B, Radbruch A, Müller S, Chang HD. 2016. High-resolution microbiota flow cytometry reveals dynamic colitis-associated changes in fecal bacterial composition. Eur J Immunol 46:1300–1303. https://doi.org/10.1002/eji.201646297.
30. Rubbens P, Props R, Kerckhof F-M, Boon N, Waegeman W. 2021. Cytometric fingerprints of gut microbiota predict Crohn's disease state. ISME J 15:354–358. https://doi.org/10.1038/s41396-020-00762-4.
31. Boedigheimer MJ, Ferbas J. 2008. Mixture modeling approach to flow cytometry data. Cytometry A 73A:421–429. https://doi.org/10.1002/cyto.a.20553.
32. Reiter M, Rota P, Kleber F, Diem M, Groeneveld-Krentz S, Dworzak M. 2016. Clustering of cell populations in flow cytometry data using a combination of Gaussian mixtures. Pattern Recognit 60:1029–1040. https://doi.org/10.1016/j.patcog.2016.04.004.
33. Hyrkas J, Clayton S, Ribalet F, Halperin D, Armbrust EV, Howe B. 2016. Scalable clustering algorithms for continuous environmental flow cytometry. Bioinformatics 32:417–423. https://doi.org/10.1093/bioinformatics/btv594.
34. Ludwig J, zu Siederdissen CH, Liu Z, Stadler PF, Müller S. 2019. flowEMMi: an automated model-based clustering tool for microbial cytometric data. BMC Bioinformatics 20:643. https://doi.org/10.1186/s12859-019-3152-3.
35. Louca S, Doebeli M, Parfrey LW. 2018. Correcting for 16S rRNA gene copy numbers in microbiome surveys remains an unsolved problem. Microbiome 6:41. https://doi.org/10.1186/s40168-018-0420-9.
36. McCarthy A, Chiang E, Schmidt ML, Denef VJ. 2015. RNA preservation agents and nucleic acid extraction method bias perceived bacterial community composition. PLoS One 10:e0121659. https://doi.org/10.1371/journal.pone.0121659.
37. Barbesti S, Citterio S, Labra M, Baroni MD, Neri MG, Sgorbati S. 2000. Two and three-color fluorescence flow cytometric analysis of immunoidentified viable bacteria. Cytometry 40:214–218. https://doi.org/10.1002/1097-0320(20000701)40:3<214::AID-CYTO6>3.0.CO;2-M.
38. Rubbens P, Props R, Garcia-Timermans C, Boon N, Waegeman W. 2017. Stripping flow cytometry: how many detectors do we need for bacterial identification? Cytometry A 91A:1184–1191. https://doi.org/10.1002/cyto.a.23284.
39. García FC, López-Urrutia Á, Morán XAG. 2014. Automated clustering of heterotrophic bacterioplankton in flow cytometry data. Aquat Microb Ecol 72:175–185. https://doi.org/10.3354/ame01691.
40. Vives-Rego J, Resina O, Comas J, Loren G, Julià O. 2003. Statistical analysis and biological interpretation of the flow cytometric heterogeneity observed in bacterial axenic cultures. J Microbiol Methods 53:43–50. https://doi.org/10.1016/S0167-7012(02)00219-1.
41. Besmer MD, Hammes F. 2016. Short-term microbial dynamics in a drinking water plant treating groundwater with occasional high microbial loads. Water Res 107:11–18. https://doi.org/10.1016/j.watres.2016.10.041.
42. Hammes F, Broger T, Weilenmann H-U, Vital M, Helbing J, Bosshart U, Huber P, Peter Odermatt R, Sonnleitner B. 2012. Development and laboratory-scale testing of a fully automated online flow cytometer for drinking water analysis. Cytometry A 81A:508–516. https://doi.org/10.1002/cyto.a.22048.
43. Shade A. 2017. Diversity is the question, not the answer. ISME J 11:1–6. https://doi.org/10.1038/ismej.2016.118.
44. Prest EI, Hammes F, Kötzsch S, van Loosdrecht MCM, Vrouwenvelder JS. 2013. Monitoring microbiological changes in drinking water systems using a fast and reproducible flow cytometric method. Water Res 47:7131–7142. https://doi.org/10.1016/j.watres.2013.07.051.

45. Bishop CM. 2006. Pattern recognition and machine learning. Springer-Verlag, Berlin, Germany.
46. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E. 2011. Scikit-learn: machine learning in Python. J Mach Learn Res 12:2825–2830.
47. Scrucca L, Fop M, Murphy TB, Raftery AE. 2016. mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. R J 8:289–317. https://doi.org/10.32614/RJ-2016-021.
48. Hill MO. 1973. Diversity and evenness: a unifying notation and its consequences. Ecology 54:427–432. https://doi.org/10.2307/1934352.
49. Leinster T, Cobbold CA. 2012. Measuring diversity: the importance of species similarity. Ecology 93:477–489. https://doi.org/10.1890/10-2402.1.
50. Daly A, Baetens J, De Baets B. 2018. Ecological diversity: measuring the unmeasurable. Mathematics 6:119. https://doi.org/10.3390/math6070119.
51. Bray JR, Curtis JT. 1957. An ordination of the upland forest communities of southern Wisconsin. Ecol Monogr 27:325–349. https://doi.org/10.2307/1942268.
52. Breiman L. 2001. Random forests. Mach Learn 45:5–32. https://doi.org/10.1023/A:1010933404324.
53. Bergstra J, Bengio J. 2012. Random search for hyper-parameter optimization. J Mach Learn Res 13:281–305.
54. Monaco G, Chen H, Poidinger M, Chen J, De Magalhães JP, Larbi A. 2016. FlowAI: automatic and interactive anomaly discerning tools for flow cytometry data. Bioinformatics 32:2473–2480. https://doi.org/10.1093/bioinformatics/btw191.
55. Rubbens P, Schmidt ML, Props R, Biddanda BA, Boon N, Waegeman W, Denef VJ. 2019. Randomized lasso links microbial taxa with aquatic functional groups inferred from flow cytometry. mSystems 4:e00093-19. https://doi.org/10.1128/mSystems.00093-19.
56. Chiang E, Schmidt ML, Berry MA, Biddanda BA, Burtner A, Johengen TH, Palladino D, Denef VJ. 2018. Verrucomicrobia are prevalent in north-temperate freshwater lakes and display class-level preferences between lake habitats. PLoS One 13:e0195112. https://doi.org/10.1371/journal.pone.0195112.
57. Friedman J, Alm EJ. 2012. Inferring correlation networks from genomic survey data. PLoS Comput Biol 8:e1002687. https://doi.org/10.1371/journal.pcbi.1002687.
58. Mantel N. 1967. The detection of disease clustering and a generalized regression approach. Cancer Res 27:209–220.
59. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J, van der Walt SJ, Brett M, Wilson J, Millman KJ, Mayorov N, Nelson ARJ, Jones E, Kern R, Larson E, Carey CJ, Polat I, Feng Y, Moore EW, VanderPlas J, Laxalde D, Perktold J, Cimrman R, Henriksen I, Quintero EA, Harris CR, Archibald AM, Ribeiro AH, Pedregosa F, van Mulbregt P, SciPy 1.0 Contributors. 2020. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods 17:261–272. https://doi.org/10.1038/s41592-019-0686-2.