**PhenoGMM**: Gaussian mixture modelling of microbial cytometry data enables efficient predictions of biodiversity

Peter Rubbens¹*, Ruben Props², Frederiek-Maarten Kerckhof², Nico Boon², Willem Waegeman¹

¹KEMIT, Department of Data Analysis and Mathematical Modelling, Ghent University, Coupure Links 653, B-9000, Gent, Belgium
²Center for Microbial Ecology and Technology (CMET), Ghent University, Coupure Links 653, B-9000, Gent, Belgium.

*To whom correspondence should be addressed.

**Abstract**

**Motivation**: Microbial flow cytometry allows to rapidly characterize microbial community diversity and dynamics. Recent research has demonstrated a strong connection between the cytometric diversity and taxonomic diversity based on 16S rRNA gene amplicon sequencing data. This creates the opportunity to integrate both types of data to study and predict the microbial community diversity in an automated and efficient way. However, microbial flow cytometry data results in a number of unique challenges that need to be addressed.

**Results**: The results of our work are threefold: i) We expand current microbial cytometry fingerprinting approaches by using a model-based fingerprinting approach based upon Gaussian Mixture Models, which we called PhenoGMM. ii) We show that microbial diversity can be rapidly estimated by PhenoGMM. In combination with a supervised machine learning model, diversity estimations based on 16S rRNA gene amplicon sequencing data can be predicted. iii) We evaluate our method extensively by using multiple datasets from different ecosystems and compare its predictive power with a generic binning fingerprinting approach that is commonly used in microbial flow cytometry. These results confirm the strong connection between the genetic make-up of a microbial community and its phenotypic properties as measured by flow cytometry.

**Availability**: All code and data supporting this manuscript is freely available on GitHub at: [https://github.com/prubbens/PhenoGMM](https://github.com/prubbens/PhenoGMM). Raw flow cytometry data is freely available on FlowRepository and raw sequences via the NCBI Sequence Read Archive. The functionality of PhenoGMM has been incorporated in the R package PhenoFlow: [https://github.com/CMET-UGent/Phenoflow_package](https://github.com/CMET-UGent/Phenoflow_package).

**Contact**: Peter.Rubbens@UGent.be

**Supplementary information**: Supplementary data are available in attachment to this submission.

---

**1 Introduction**

Life as we know it would not be possible without microorganisms (Gilbert and Neufeld, 2014). Microbial communities are driving forces of biogeochemical processes such as the carbon and nitrogen cycle (Falkowski et al., 2008), they maintain human health (Young, 2017) and they are used for the creation of a vast array of products, such as chemicals, antibiotics and food (Blaser et al., 2016). The field of microbial ecology is interested in the diversity of a community and its relation to ecosystem functionality (Konopka et al., 2015). 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 (Van Dijk et al., 2014). Although advances have been made to perform sequencing in real-time (Ardui et al., 2018), most 16S rRNA gene amplicon sequencing surveys are still expensive (Sims et al., 2014) and laborious in time (van Dorst et al., 2014).

Instead of solely focusing on genotypic information, there is a need to combine omics data with phenotypic information (De Vrieze et al., 2018). One of such tools to study the phenotypic identity of microbial communities is flow cytometry (FCM). FCM is a high-throughput technique, able to measure hundreds to thousands of individual cells in mere seconds. These measurements result in a multivariate 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. Common for microbial FCM is to use a stain that interacts with the nucleic acid content of a cell (Koch et al., 2013b; Van Nevel et al., 2013).

Many algorithms exist in the field of immunophenotyping cytometry to identify separated cell populations (see e.g. the extensive benchmark
studies by Aghaeepour et al. (2013) and Weber and Robinson (2016). However, the number of variables that describe a bacterial cell is typically much lower than is common for cytometry setups studying mammalian cells. As a result, cytometric distributions of individual bacterial populations tend to overlap (Rubbens et al., 2017a), as the number of bacterial populations in a community is typically much larger than the number of differentiating signals. Therefore, bacterial cytometry data is characterized by overlapping cell populations and these algorithms cannot be applied for the analysis bacterial cytometry data. Consequently, data analysis pipelines should be designed to address these characteristics.

In previous research, microbiologists have relied on cytometric fingerprinting techniques (Koch et al., 2013b; Props et al., 2016). The approaches that are currently used for the analysis of bacterial communities can be broadly divided in two categories: i) manual annotation of clusters (Günther et al., 2012; Koch et al., 2013b) and ii) automated approaches that employ binning strategies (Li, 1997; Koch et al., 2013a; García et al., 2015; Props et al., 2016). Such a fingerprint allows 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 (Koch et al., 2014), despite the fact that there are no or only a few clearly separated cell populations. These methods have a number of drawbacks: i) Manual gating of regions of interest is laborious in time and operator dependent, ii) only bivariate interactions of cytometry channels are considered and iii) 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).

After a fingerprint has been constructed, one can calculate what is called the cytometric or phenotypic diversity of a community (Li, 1997). These are estimations of the diversity of a microbial community based on the cell counts per cluster. If many clusters contain cells, a community can be considered as ‘rich’. If the cells are equally distributed over those clusters, a community can be considered as ‘even’. Recent reports have shown a significant correlation between the cytometric diversity and genotypic diversity derived from 16S gene sequencing data (García et al., 2015; Props et al., 2016, 2018). In other words, there is a strong connection between the genetic make-up of a microbial community and its phenotypic properties, which can be quantified. This result has been backed up by molecular identification using DGGE of sorted subpopulations (Park et al., 2005; Koch et al., 2013b), the sequencing of sorted individual cells or subpopulations (Zimmermann et al., 2016; Stepanauskas et al., 2017; Günther et al., 2018) and by using a bottom-up approach in which individual bacterial populations resulted in unique cytometric characterizations, which can be automatically identified using machine learning models (Rubbens et al., 2017a).

We propose an extension of current fingerprinting approaches that is able to deal with overlapping cell populations. Our workflow, which is called ‘PhenoGMM’, makes use of Gaussian mixture models (GMM). GMMs have been successfully applied to cytometry data before to identify separated cell populations in an automated way (Boedighheimer and Ferbas, 2008; Reiter et al., 2016). Interestingly, Hýrkaš et al. (2015) have shown that their GMM approach outperformed state-of-the-art immunophenotyping cytometry algorithms for the automated identification of phytoplankton populations. By overclustering the data, GMMs can also be used to describe the distribution of the data, and therefore PhenoGMM is able to deal with overlapping cell populations. In addition, the number of mixtures that are needed to describe the data is much lower compared to the number of variables that result from traditional binning approaches. This facilitates the use of supervised machine learning models. We demonstrate that bacterial diversity can be predicted based on cytometric fingerprints derived from PhenoGMM. We illustrate our method using multiple datasets, in which we use diversity values derived from 16S rRNA gene amplicon sequencing data as target values that need to be predicted. We compare its performance with the predictive power of a generic traditional binning approach, which we have called PhenoGrid. Finally, we highlight a number of possible extensions concerning the integration of FCM with 16S rRNA gene sequencing, such as the calculation of β-diversity values, the imputation of missing diversity values and the prediction of individual OTU abundances based on FCM data.

Materials and Methods
Methodology

Preprocessing
Two steps are carried out for all measurements before further analysis of the data. First, all individual channels are transformed using \( f(x) = \sinh(x) \). Next, background due to debris and noise has been removed using a fixed digital gating strategy (Prest et al., 2013; Props et al., 2016). In other words, a single gate is applied to separate bacterial cells from background and is used for all samples.

Cytometric fingerprinting using Gaussian Mixture Models
In order to create a fingerprint template that can be used to extract variables describing a specific sample, all samples in the dataset in the training set need to be concatenated. Files are first subsampled to the same number of cells per file (\( N_{\text{CELLS \_MIN}} \)), in order not to bias the Gaussian Mixture Model (GMM) towards a specific sample. This number can either be 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 \( 1 \times 10^6 \) cells, depending on computational resources. If \( n \) denotes the total number of samples, then the total number of cells (\( N_{\text{CELLS}} \)) in the training set will be determined as \( N_{\text{CELLS}} = n \times N_{\text{REP}} \times N_{\text{CELLS \_MIN}} \), in which \( N_{\text{REP}} \) denotes the number of technical repicates of a specific sample. Typically, forward (FSC) and side scatter (SSC) channels are included, along with one or two targeted fluorescence channels (denoted as FL\( X \), in which \( X \) indicates the number of a specific fluorescence detector). Unless noted otherwise, channels FSC-H, SSC-H and FL1-H (488 nm) were included for data analysis.

Once this training set is created, a GMM of \( K \) mixtures can be fitted to the data. If \( X \) denotes the entire datamatrix or training set containing \( N \) cells, then \( X \) consists of cells written as \( x_1, \ldots, x_N \), 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 \( x_i = \{ x_{i1}, \ldots, x_{iD} \} \). A GMM consists out of a superposition of normal distributions \( N \), of which each distribution has its own mean \( \mu \) and covariance matrix \( \Sigma \). Each mixture has a mixing coefficient or weight \( \pi \), which represents the fraction of data each mixture is describing. The distribution \( p \), which describes the GMM, can be written as follows:

\[
p(X) = \sum_{k=1}^{K} \pi_k N(X | \mu_k, \Sigma_k).
\] (1)

The set of parameters \( \Theta = \{ \pi_k, \mu_k, \Sigma_k \}_{k=1}^{K} \) is determined by the expectation-maximization (EM) algorithm (Bishop, 2006). Once a GMM has been trained on the concatenated data, one can cluster the cells in each sample separately using the trained GMM. For this step, either a specific number of cells of choice are sampled per replicate, or the lowest number of cells of the replicates that are part of that specific sample, denoted as \( N_{\text{CELLS \_REP}} \). After clustering, we count the number of cells per cluster, after which the relative number of cells per cluster and sample can be retrieved. The resulting variables can be used for different purposes: i) to calculate diversity metrics in an unsupervised way or ii) as input variables...
to be included in predictive models. An illustration of PhenoGMM can be seen in Fig. 1. We used the GaussianMixture() function of the scikit-learn machine learning library to implement our method (Pedregosa et al., 2011). This function contains four different ways in which the covariance matrix of each mixture is determined:

- 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.\(^1\)
- 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 (Scrucca et al., 2016).

Defining \(\alpha\) and \(\beta\)-diversity

Both 16S gene amplicon sequencing and flow cytometry fingerprints give rise to a compositional representation of a microbial community. The first is determined by counting the number of similar sequences at a certain taxonomic level, e.g., a taxonomic unit, the latter by counting the number of cells present in a predefined gate or cluster in the cytometric fingerprint, the phenotypic unit. Based on abundance data, one can calculate both \(\alpha\)- and \(\beta\)-diversity metrics. The first quantifies the diversity within a sample, the latter the diversity between samples. As various diversity metrics exist in ecology to calculate \(\alpha\)-diversity; we use the Hill numbers to quantify community diversity (Hill, 1973), as proposed by recent reviews of Leinster and Cobbold (2012) and Daly et al. (2018). If we let \(p_1, \ldots, p_s\) represent the vector of relative abundances, describing the abundance of \(S\) bacterial populations, then we can define the richness (\(D_0\)) and evenness (\(D_1, D_2\)) of a microbial community as follows:

\[
D_{q=0}(\mathbf{p}) = S, \quad (2)
\]

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

\[
D_{q=2}(\mathbf{p}) = \frac{1}{\sum_{i=1}^{S} p_i^2}, \quad (4)
\]

\(q\) denotes the order of the Hill-number, which is part of a general family, which can be denoted as \(D_q^p\). It expresses the weight that is given to more abundant populations.

\(\beta\)-diversity quantifies the difference in compositions between different samples. Typically, this is calculated by performing ordination on a dissimilarity matrix that contains the dissimilarities or distances between samples. We quantify the dissimilarity between samples using the Bray-Curtis dissimilarity (Bray and Curtis, 1957). If we let \(BC_{AB}\) denote the dissimilarity between samples \(A\) and \(B\), \(BC_{AB}\) is calculated using the following equation:

\[
BC_{AB} = \frac{\sum_{i=1}^{S} |p_{A,i} - p_{B,i}|}{\sum_{i=1}^{S} (|p_{A,i}| + |p_{B,i}|)} \quad (5)
\]

Predictive modeling

FCM fingerprints can be used as input variables to train a machine learning model. We used Random Forest regression (Breiman, 2001), an ensemble of decision trees, to predict \(\alpha\)-diversity metrics. A randomized grid search was performed to search for an optimal hyperparameter combination (Bergstra and Bengio, 2012). This means that a 100 random combinations of hyperparameter values were evaluated. The maximum number of variables that are considered at an individual split for a decision tree was randomly drawn from \([1, ..., K]\), the minimum number of samples for a specific leaf was randomly drawn between \([1, ..., 5]\). The cross-validation strategy differed per experiment, and is described accordingly.

Datasets

Dataset 1: In Silico Bacterial Communities

Data from 20 individual bacterial populations that were measured through FCM were collected from Rubbens et al. (2017a). The data is available at FlowRepository (accession ID: FR-FCM-ZZSH). In brief, bacterial populations were measured after 24h of incubation, stained with SYBR Green I and two technical replicates per population 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 scatter (SSH) information that was collected as well. Additional automated denosing was performed using the FlowAI package (v1.4.4., default settings, target channel: FL1, changepoint detection: 150, Monaco et al. (2016)). A full experimental overview can be found in Rubbens et al. (2017a). The lowest number of cells collected after background removal amounted to 13166 cells.

Dataset 2: Cooling water microbiome

Data was used as presented in Props et al. (2016). 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 (Survey I and II) and analyzed through 16S rRNA gene amplicon sequencing (\(n = 77\)) and FCM (\(n = 153\)). The sequencing and flow cytometric procedures are extensively described in Props et al. (2016). Taxonomic identification of the microbial communities was done at the operational taxonomic unit (OTU) level at 97% similarity. Sequences are available from the NCBI Sequence Read Archive (SRA) under (accession ID: SRP066190), flow cytometry data is available from FlowRepository (accession ID: FR-FCM-ZZNA). The lowest number of cells collected after background removal amounted to 10565 cells.

Dataset 3: Freshwater lake system microbiome

A total of 173 samples collected from three types of freshwater lake systems were analyzed. Data were used as presented in (Rubbens et al., 2019). All samples were analyzed through 16S rRNA gene amplicon sequencing and FCM. Samples originate from three different freshwater lake systems: (1) 49 samples from Lake Michigan (2013 & 2015), (2) 62 samples from Muskegon Lake (2013-2015; one of Lake Michigan’s estuaries), and (3) 62 samples from twelve Inland lakes in Southeastern Michigan (2014-2015). Field sampling, DNA extraction, DNA sequencing and processing are described in Chiang et al. (2018). Fastq files were submitted to NCBI SRA under BioProject accession number PRJNA412984 and PRJNA414423. Taxonomic identification of microbial communities was done for each of the three lake datasets separately and treated with an OTU abundance threshold cutoff of either 1 sequence in 3% of the samples. For comparison of taxonomic abundances across samples, each the three datasets were then 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. Next, 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 extensively described in Props et al. (2016). In brief, samples were stained with SYBR Green I and three technical replicates were measured on an...
Fig. 1. 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 of the three communities are concatenated into one dataframe, to which is a GMM with \( K = 100 \) mixtures is fitted. This results in a cluster structure, 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 2-4 (in this case \( D_2 \)). E: Based on the count table derived using PhenoGMM, one can try and predict diversity metrics based another type of data such as 16S rRNA gene sequencing, using a machine learning model.

Experimental setup

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 parameter combination, after which relative cell fractions per bin were determined. The resulting cell fractions were next concatenated into one vector.

Both PhenoGMM and PhenoGrid result in multiple variables that describe cell counts, either per cluster or bin. This can be used to perform:

1. Unsupervised \( \alpha \)-diversity estimation, by directly calculating \( D_0 \), \( D_1 \) and \( D_2 \) according to equations 2, 3 and 4 based on the count cell vectors.
2. Unsupervised \( \beta \)-diversity estimations, by calculating Bray-Curtis dissimilarities (equation 5) between the cytometric fingerprints.
3. Supervised \( \alpha \)-diversity predictions, with cytometric fingerprints as input variables to predict true target variables \( D_0 \), \( D_1 \) and \( D_2 \) based on 16S rRNA gene sequencing data, by means of Random Forest regression.
4. Supervised taxon abundance predictions, with cytometric fingerprints as input variables to predict true taxon abundances, based on 16S rRNA gene sequencing data, by means of Random Forest regression.

Research question 1: Does PhenoGMM allow \( \alpha \)-diversity estimations of in silico synthetic microbial communities?

The main goal is to estimate (i.e., unsupervised) or predict (i.e., supervised) \( \alpha \)-diversity metrics based on cytometric fingerprinting of the data. Dataset 1 contains the cytometric characterization of individual bacterial populations. By using a data-aggregation step, it is possible to create bacterial communities of different compositions. As it is known which cell belongs to which species, diversity indices can be calculated with high accuracy by simply counting the number of bacterial populations that are present in a community (\( D_0 \)) or by counting the fraction of cells that comes from every population (\( D_1 \), \( D_2 \)). A training set representing 300
different \textit{in silico} compositions, and a test set containing 100 different compositions, were created in the following way:

1. Sample uniformly at random a number $S'_i$ between two and 20; this is the number of populations that will make up community $i$.
2. Select randomly which $S'_i$ populations will make up the total community (from $S = 20$ populations).
3. Use the Dirichlet distribution to randomly sample a specific composition that sums to 1, containing the selected populations. The Dirichlet distribution can be used to model the joint distribution of individual fractions of multiple species (Friedman and Alm, 2012).

The evenness of the composition depends on the concentration parameter $a$, which determines how evenly the weight will be spread over multiple species. If $a$ is low, only a few species will make up a large part of the community. If $a$ is high, the fraction of each population will be almost equally divided.

Using these compositions, \textit{in silico} communities can be sampled accordingly. This results in a training and test set containing 300 and 100 cytometric representations of bacterial communities respectively, ranging from two to 20 populations, with varying compositions. This experiment was repeated for $a = 0.1, 1, 10$. Random forests were trained using 5-fold cross-validation. Both unsupervised and supervised $\alpha$-diversity estimations were reported for the test set.

Research question 2: Does \textit{PhenoGMM} allow $\alpha$-diversity predictions based on 16S rRNA gene sequencing data for \textit{freshwater microbial communities}?

Analogous to experiment 1, the main goal is here to both estimate and predict $\alpha$-diversity metrics based on cytometric fingerprinting of the data. However, different from dataset 1, we will now estimate $\alpha$-diversity values based on 16S rRNA gene amplicon sequencing. Dataset 2 and 3 contain natural communities, which were measured both by FCM and 16S rRNA gene amplicon sequencing. These values were used as target variables to predict. 10-fold cross-validation was used to select hyperparameters for the Random Forest model, for which predictive performance of the validation sets is reported. Unsupervised estimations were reported based on the full dataset.

Extensions

1. We quantified the correlation between $\beta$-diversity estimations based on FCM and 16S rRNA gene amplicon sequencing for all datasets.
2. Missing diversity values based on 16S rRNA gene amplicon sequencing were imputed based on \textit{PhenoGMM} for dataset 2.
3. Individual abundances of the first twenty bacterial populations in the composition (either sampled \textit{in silico} or based on 16S rRNA gene sequencing) were predicted based on cytometric fingerprints for all datasets.

Performance evaluation

- Unsupervised and supervised $\alpha$-diversity estimations were quantified by calculating the Kendall’s rank correlation coefficient $\tau$ between the true and estimated values. The $\tau_B$ implementation, which is able to deal with ties, is calculated as follows:

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

$N_c$ 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).

- Supervised predictions are evaluated by calculating the $R^2$ between true ($y = \{y_1, \ldots, y_n\}$) and predicted ($\hat{y} = \{\hat{y}_1, \ldots, \hat{y}_n\}$) values:

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

in which $\bar{y}$ denotes the average value of $y$. If $R^2 = 1$, predictions were correctly estimated. If $R^2 < 0$, predictions are worse than random guessing. The $r^2$-score()-function from the scikit-learn machine learning library was used.

- Unsupervised $\beta$-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 (Mantel, 1967). 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 $ij$:

$$
Z_M = \sum_{i=1}^{n} \sum_{j=1}^{n} BC_{ij}^{FCM} \times BC_{ij}^{16S}.
$$

The test statistic $Z_M$ is normalized and then compared to a null distribution, based on 1000 permutations.

\textbf{Results}

\textit{PhenoGMM} allows to predict $\alpha$-diversity of \textit{in silico} synthetic microbial communities

300 different bacterial communities were assembled by aggregating cytometric characterizations of individual populations in varying compositions (creating communities \textit{in silico}), constituting the training set. This allowed to simulate community compositions in an accurate way, as cell labels according to taxonomy are known for every individual cell. Based on these compositions, diversity metrics could be accurately determined, and were used as target variables to evaluate diversity estimations and predictions. We repeated the experiment for three different values of $a$, in which $a$ determines how evenly the weight is spread amongst the different populations. If $a$ is small, only a few species will be dominantly present, if $a$ is large, chances are high that the weight is evenly spread amongst the different populations. This is illustrated using Lorenz curves, which depict the cumulative proportion of abundance versus the cumulative proportion of bacterial species (SI Fig. 1).

100 additional bacterial communities were assembled using the same aggregation strategy, making up the test set. Cytometric fingerprints were determined on the concatenated representation of the samples in the training set, to which a GMM of $BC$ was compared with a generic fixed binning approach called \textit{PhenoGrid} (Table 1). To compare supervised with unsupervised diversity estimations and predictions, we used Kendall’s rank correlation coefficient $\tau_B$ and the $R^2$-score() function from scikit-learn. The copyright holder for this preprint (which was this version posted May 18, 2019. ; https://doi.org/10.1101/641464 doi: bioRxiv preprint available under a CC-BY-NC 4.0 International license.
expected, unsupervised estimations resulted in lower correlations with true diversity metrics compared to supervised predictions, although still significant in most cases (Kendall’s \(\tau_B\), level of significance \(\alpha = 0.05\)). The only exceptions were \(D_1\) and \(D_2\) for \(\alpha = 10\) when using PhenoGrid. PhenoGMM resulted in better unsupervised \(\alpha\)-diversity estimations than PhenoGrid, but both approaches resulted in a comparable supervised performance. \(R^2\) values were considerably higher than zero, and slightly in favor of PhenoGMM (SI Table 1). We note that the predictive performance mainly depended on \(\alpha\) and the diversity metric of choice. For example, the hardest setting was the one in which \(\alpha = 0.1\) and \(D_0\) the target variable to predict. In this case only a few populations made up a large part of the community (low \(\alpha\)), but an equal weight is attributed to all species when defining diversity. Generally, if the abundance of populations is taken into account \((q > 0)\), \(\alpha\)-diversity predictions were better. In other words, FCM is able to capture community structure rather than the identity of the community.

### Computational efficiency

We timed different steps in the workflow of PhenoGMM for \(\alpha = 1\) and \(D_1\). The time in seconds was determined in function of the number of mixtures \(K\) (Fig. 2A). Each analysis was run on a separate node of a computer infrastructure, with 2.6 Ghz CPU and 20GB of RAM for each node. The timing consists out of the following steps: fitting a GMM, using this model to extract variables per sample and calculating \(D_1\) directly according to equation 3 (Fig. 2A), or with the addition of fitting a Random Forest model to predict \(D_1\) (Fig. 2B). We sampled 5000 cells per sample. As we have 300 samples in our training set, this amounts to fitting a GMM to 1.5 million cells. Most importantly, the entire analysis remains under one hour. Most of the time is spent on fitting the GMM. Training a Random Forest model on the fitted GMM comes with an average increase of 24.4% of the runtime for \(K = 256\). The predictive performance of both PhenoGMM and PhenoGrid was evaluated in function of the total runtime, indicating that PhenoGMM needs much less time than PhenoGrid to reach its optimal performance (SI Fig. 2).

### Influence of hyperparameters on \(\alpha\)-diversity estimations

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 cluster 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 \(R^2(D_1)\) for \(\alpha = 1\) for a supervised analysis (SI Fig. 3). The results indicate that considering the predictive performance:

- **K**: generally, the higher \(K\), the better the performance, which saturates after a specific threshold.
- **N CELLS MIN**: predictions are quite robust for this parameter.
- **N CELLS REP**: predictions are quite robust for this parameter.
- **N SAMPLES**: predictive performance did not saturate yet at N SAMPLES = 300.
- **TYPE**: predictions are quite robust for the type of covariance matrix, but the ‘full’ type resulted in the best predictions.

### PhenoGMM allows to predict \(\alpha\)-diversity for freshwater microbial communities

\(\alpha\)-diversity predictions were made based on cytometric fingerprinting of natural microbial communities, which were either part of a cooling water system (i.e. Survey I, II or combined), or a freshwater lake system (i.e. Inland, Michigan, Muskegon or all of them combined). \(\alpha\)-diversity values, based on 16S rRNA gene amplicon sequencing, were used as target variables to predict. Supervised predictions were the result of Random Forest regression, which was tuned using ten-fold cross-validation. Values are reported for the model that returned the best combined predictions on the validation folds using Kendall’s \(\tau_B\) (Table 2) and \(R^2\) (SI Table 2). Diversity predictions were feasible (i.e., significant according to \(\tau_B\) for \(\alpha = 0.05\) and considerably higher than zero for \(R^2\)), but depending on the dataset and diversity index. For example, predictions of \(D_0\) were easier to make compared to \(D_1\) or \(D_2\) for the Inland lake-system and were better for the cooling water system than for the lake systems. The predictive performance of PhenoGMM \((K = 256)\) was better or similar compared to PhenoGrid \((K = 128 \times 128)\).

Unsupervised diversity estimations were evaluated as well (SI Table 3). Diversity estimations were highly significant for the cooling water microbiome, but were insignificant in a number of cases for the freshwater lake systems according to both approaches \((D_0\) and \(D_2\) for the Inland lakes and Muskegon lake; Kendall’s \(\tau_B\), \(\alpha = 0.05\)). PhenoGMM outperformed PhenoGrid in most cases, indicating that even more mixtures might be needed to make it competitive with PhenoGrid in this setting. We conclude that FCM shows a strong connection with 16S rRNA gene sequencing data. FCM is sensitive for the community structure and can be used to adequately perform microbial diversity estimations and predictions of natural communities.

### PhenoGMM allows estimations of \(\beta\)-diversity

\(\beta\)-diversity, which quantifies the difference in community composition between different samples, can be determined as well using both 16S rRNA gene amplicon sequencing and FCM. This was done by calculating Bray-Curtis dissimilarities between all communities based on relative fractions per OTU or mixture. A mantel test was used to calculate the correlation between Bray-Curtis dissimilarity matrices, derived from the two types of data using both PhenoGMM and PhenoGrid (SI Table 4).
Both approaches resulted in statistically significant correlations (Mantel test, \( \alpha = 0.05 \)). PhenoGMM resulted in better (synthetic microbial communities) or similar (freshwater communities) \( \beta \)-diversity estimations compared to PhenoGrid.

**PhenoGMM can be used to impute missing \( \alpha \)-diversity values**

It is common practice in the field of microbial ecology to analyze only a subset of samples by 16S rRNA gene amplicon sequencing. This was also the case for the cooling water dataset, for which all samples (\( n = 153 \)) were analyzed through FCM, but additionally roughly half (\( n = 77 \)) by both FCM and 16S rRNA gene amplicon sequencing. PhenoGMM allowed to make inference concerning the \( \alpha \)-diversity of these missing samples. An example is given for \( D_2 \), for both survey I and II (Fig. 3). Predictions are the average of ten runs of PhenoGMM. This illustrates how FCM can be integrated with 16S rRNA gene sequencing in order to frequently monitor a microbial community of interest, and reduce the number of samples that have to be analyzed by 16S rRNA gene sequencing at the same time.

**PhenoGMM allows to predict individual bacterial abundances**

The fact that biodiversity can be estimated from cytometric data implies that the taxonomic structure of a microbial community is captured by the cytometric fingerprint. This opens up the opportunity to predict individual bacterial abundances based on estimations for the validation folds, using 10-fold cross-validation. Values denote the average of ten runs, along with corresponding standard deviation (SE). Values are bolded if the mean value of one approach is significantly higher than the mean value of the other approach according to a student’s \( t \)-test (\( \alpha = 0.05 \)).

| Dataset          | \( \tau_B(D_0) \)  | \( \tau_B(D_1) \)  | \( \tau_B(D_2) \)  | \( \tau_B(D_0) \)  | \( \tau_B(D_1) \)  | \( \tau_B(D_2) \)  |
|------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|
| Survey I         | 0.40 ± 0.03       | 0.49 ± 0.06       | 0.53 ± 0.05       | 0.27 ± 0.07       | 0.48 ± 0.04       | 0.49 ± 0.06       |
| Survey II        | 0.66 ± 0.04       | 0.61 ± 0.04       | 0.62 ± 0.03       | 0.62 ± 0.03       | 0.59 ± 0.04       | 0.59 ± 0.04       |
| Survey I+II      | 0.55 ± 0.03       | 0.63 ± 0.014      | 0.64 ± 0.04       | 0.52 ± 0.04       | 0.59 ± 0.03       | 0.61 ± 0.04       |
| Inland           | 0.52 ± 0.07       | 0.33 ± 0.05       | 0.22 ± 0.02       | NS                | 0.27 ± 0.04       | NS                |
| Michigan         | 0.26 ± 0.07       | 0.42 ± 0.05       | 0.39 ± 0.05       | NS                | 0.35 ± 0.06       | 0.36 ± 0.03       |
| Muskegon         | 0.35 ± 0.04       | 0.29 ± 0.04       | 0.19 ± 0.08       | NS                | NS                | NS                |
| All lake systems | 0.510 ± 0.018     | 0.48 ± 0.02       | 0.44 ± 0.02       | 0.38 ± 0.03       | 0.38 ± 0.02       | 0.34 ± 0.03       |

NS: Not significant (average \( P > 0.05 \)).

Fig. 2. Benchmarking of PhenoGMM vs PhenoGrid. Each model was run five times. A: time \( t \), in seconds, versus \( K \), the number of mixtures or dimension of the grid. B: Performance expressed in \( R^2(D_1) \) in function of \( K \).

Table 2. Summary of supervised \( \alpha \)-diversity estimations for different \( \alpha \), quantified by \( \tau_B \), using both PhenoGMM versus PhenoGrid. Performance was quantified based on estimations for the validation folds, using 10-fold cross-validation. Values denote the average \( \tau_B \) of 10 different runs, along with corresponding standard deviation (SE). Values are bolded if the mean value of one approach is significantly higher than the mean value of the other approach according to a student’s \( t \)-test (\( \alpha = 0.05 \)).
For natural communities, we note that it was possible to predict taxon abundances for 67-95% of the evaluated taxa.

Discussion

In this paper we have extensively shown that flow cytometry (FCM) can be used to estimate and predict microbial biodiversity. To do so, we proposed a more advanced fingerprinting strategy based on Gaussian Mixture Models (GMMs), called PhenoGMM. Our approach allows to create meaningful variables and reduces the number sample-describing variables considerably compared to traditional binning approaches. This makes the use of predictive models, in this study by means of Random Forest regression, much more feasible. We evaluated the performance of PhenoGMM both for unsupervised estimations and supervised predictions of biodiversity using multiple datasets. We compared it with the performance of a generic traditional binning approach, which we in this work called PhenoGrid.

In the first part of the paper, we constructed communities in silico by aggregating cytometric characterizations of individual bacterial populations in different compositions. This allowed us to simulate microbial community compositions in a highly precise and controlled way. In the second part we showed that flow cytometry can be used to predict biodiversity values based on 16S rRNA gene sequencing data. Upon making predictions, PhenoGMM resulted in either more or equally accurate predictions compared to PhenoGrid for all datasets. Unsupervised estimations of α-diversity resulted in higher correlations with the target diversity values for PhenoGMM for the synthetic communities, while estimations were better for PhenoGrid for natural communities, for which the diversity was determined based on 16S rRNA gene amplicon sequencing. Total analysis time of PhenoGMM remains under one hour.

Many algorithms exist for the analysis of cytometry data. However, most of these methods are developed for an automated analysis of immunophenotyping data, in which many separated cell populations can be identified. Microbial cytometry data has a number of different characteristics, which is why most of these approaches are not applicable. The reason is that bacterial cells are typically much smaller in both cell size and volume compared to eukaryotic cells (Robinson, 2018). In addition, no general antibody-based panels have been established for microbial cells due to the high complexity of microbial communities (Koch and Müller, 2018). One has to rely on general DNA stains, for which it is difficult to develop multicolor approaches (Buysschaert et al., 2016). Therefore, the number of variables describing an individual bacterial cell is typically much lower than e.g. a human cell. As a result, cytometric distributions of bacterial populations tend to overlap, as the number of bacterial populations is larger than the number of differentiating signals. GMMs allow to model overlapping cell distributions. However, as distributions overlap, it is hard to determine the exact number of populations. That is why we overcluster the data by choosing a sufficient number of K mixtures. As K increases, the performance saturates gradually, and more mixtures will not improve predictions.

Few reports exist that quantitatively evaluate fingerprinting approaches for the analysis of microbial data. A brief comparison study with n = 21 samples has been recently conducted (Menyhárt et al., 2018), illustrating a better performance for FlowFP (Rogers and Holyst, 2009), compared to the use of FlowCyBar (Koch et al., 2013b). FlowFP is quite similar compared to PhenoGMM, as it makes use of an adaptive binning approach, in which bins are smaller when the density of the data is higher, while FlowCyBar makes use of manually annotated clusters. However, the bins are still rectangular in shape, while PhenoGMM allows clusters to be of any shape. Most fingerprinting strategies make use of manual annotation of clusters or of fixed binning approaches (see e.g. the report by Koch et al. (2014) which qualitatively discusses different existing methods). In almost all cases, only bivariate interactions are inspected. PhenoGMM allows to model the full parameter space at once. This is interesting, because although it is hard to develop multicolor approaches for bacterial analyses, they are possible (see e.g. the work by Barbesti et al. (2000)). In addition our research group has established that additional detectors that capture signals due to spillover can assist in the discrimination between bacterial species (Rubbens et al., 2017b). Therefore, the parameter space in which bacterial cells can be described is increasing, and PhenoGMM is able to model this straightforwardly. Because it is in 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. Other adaptive binning strategies have been proposed for microbial FCM data as well, however these still only investigate bivariate interactions (Amalfitano et al., 2018; Huang et al., 2018).

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 case multiple samples diverge considerably from those which were used to determine the template (for example in case an experiment was conducted in different conditions), we recommend to refit the model. Second, PhenoGMM overclusters the data, which might result in a number of correlated

Fig. 3. Predictions of ten different runs for D2 for which no 16S rRNA gene amplicon sequencing data was available for the cooling water dataset (Survey I and II), using PhenoGMM.
variables. We recommend therefore researchers to use a classification or regression method that is able to deal with multicollinearity, which is why we used Random Forest regression in this work. Other methods that might be suitable are regularized regression methods, such as the Lasso or ElasticNet (Tibshirani, 1996; Zou and Hastie, 2005). Third, although the performance tends to saturate once $K$ is high enough, this threshold seems to application dependent, and one needs to validate the settings of the approach.

Our in silico benchmark study made use of cytometric characterizations of individual bacterial populations. These populations are known to exhibit considerable heterogeneity due to cell size diversity and cell cycle variations (Vives-Rego et al., 2003). Our research group has recently shown that the cytometric diversity of an individual population reduces when that population is part of a co-culture (Heyse et al., 2019). 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 as opposed to populations that were grown and measured individually. Yet we believe that our in silico approach is useful, as it allows to simulate variations in cytometric community structure with high precision.

In this study we focused mainly on estimations of $\alpha$-diversity (i.e., within-sample diversity), but quantification of $\beta$-diversity (i.e., between-sample diversity) can be successfully performed as well. In addition, it is possible to predict variations in the abundance of a specific bacterial populations. This might be interesting for certain biotechnological applications, in which researchers or engineers are not interested in the total diversity of the community but in the behavior of a specific bacterial population.

PhenoGMM allows to infer diversity metrics efficiently, both in an unsupervised and supervised setting. Technological advancements have enabled an automation of the data acquisition, resulting in a detailed characterization of the microbial community on-line (i.e., samples are measured at routine intervals between 5-15 min) or even in real-time (i.e., near-continuous measurements) (Hammes et al., 2012; Besner and Hammes, 2016). Therefore we see great potential to use FCM as a monitor technique to rapidly investigate microbial community dynamics. In this work we have confirmed the strong correspondence between FCM and the genetic make-up of a community, quantified by 16S rRNA gene sequencing. Therefore, FCM can be integrated with other types of data and machine learning models can be used to exploit the relationship between the two. One fruitful approach would be to routinely monitor the microbial community using FCM, and additionally analyze states 'of high interest' by for example 16S rRNA gene amplicon sequencing. Cytometric fingerprinting in combination with a supervised machine learning model can then be used to predict the diversity of missing samples (conform Fig. 3). The use of predictive models can also be used to perform classification at the community level, to for example categorize communities according to the system they are part of (De Roy et al., 2012; Dhoble et al., 2018), or to identify a case versus control status.

Acknowledgements

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.

Funding

PR is supported by Ghent University (BOFSTA2015000501). RP is supported by Ghent University (BOFDOC2015000601).
Tibshirani, R. (1996). Regression Selection and Shrinkage via the Lasso. *Journal of the Royal Statistical Society B, 58*(1), 267–288.

Van Dijk, E., Auger, H., Jaszczyszyn, Y., and Thermes, C. (2014). Ten years of next-generation sequencing technology. *Trends in Genetics, 30*(9), 418–426.

van Dorst, J., Bissett, A., Palmer, A. S., Brown, M., Snape, I., Stark, I. S., Raymond, B., McKinlay, J., Ji, M., Winsley, T., and Ferrari, B. C. (2014). Community fingerprinting in a sequencing world. *FEMS Microbiology Ecology, 89*(2), 316–330.

Van Nevel, S., Koetzsch, S., Weilenmann, H. U., Boon, N., and Hammes, F. (2013). Routine bacterial analysis with automated flow cytometry. *Journal of Microbiological Methods, 94*(2), 73–76.

Vives-Rego, J., Resina, O., Comas, J., Loren, G., and Julià, O. (2003). Statistical analysis and biological interpretation of the flow cytometric heterogeneity observed in bacterial axenic cultures. *Journal of Microbiological Methods, 53*(1), 43–50.

Weber, L. M. and Robinson, M. D. (2016). Comparison of clustering methods for high-dimensional single-cell flow and mass cytometry data. *Cytometry Part A, 89*(12), 1084–1096.

Young, V. B. (2017). The role of the microbiome in human health and disease: An introduction for clinicians. *BMJ, 356*.

Zimmermann, J., Hubschmann, T., Schattenberg, F., Schumann, J., Durek, P., Riedel, R., Friedrich, M., Glauben, R., Siegmund, B., Radbruch, A., Müller, S., and Chang, H. D. (2016). High-resolution microbiota flow cytometry reveals dynamic colitis-associated changes in fecal bacterial composition. *European Journal of Immunology, 46*(5), 1300–1303.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67*(2), 301–320.