Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Biotechnology

Recent advances in microbial community analysis from machine learning of multiparametric flow cytometry data☆

Birge D Özel Duygan[1,2] and Jan R van der Meer[1]

Dynamic analysis of microbial composition is crucial for understanding community functioning and detecting dysbiosis. Compositional information is mostly obtained through sequencing of taxonomic markers or whole meta-genomes, which may be productively complemented by real-time quantitative community multiparametric flow cytometry data (FCM). Patterns and clusters in FCM community data can be distinguished and compared by unsupervised machine learning. Alternatively, FCM data from preselected individual strain phenotypes can be used for supervised machine-training in order to differentiate similar cell types within communities. Both types of machine learning can quantitatively deconvolute community FCM data sets and rapidly analyse global changes in response to treatment. Procedures may further be optimized for recurrent microbiome samples to simultaneously quantify physiological and compositional states.

Addresses
[1] Department of Fundamental Microbiology, University of Lausanne, Lausanne, 1015, Switzerland

[2] Institute of Microbiology, CHUV, Lausanne, 1011, Switzerland

Corresponding authors: Özel Duygan, Birge D (birgeozel@gmail.com), van der Meer, Jan R (janroelof.vandermeer@unil.ch)

Current Opinion in Biotechnology 2022, **75**:102688

This review comes from a themed issue on **Systems biology**

Edited by **Mark P Styczynski** and **Neda Bagheri**

For complete overview of the section, please refer to the article collection, "Systems Biology"

Available online 2nd February 2022

https://doi.org/10.1016/j.copbio.2022.102688

Introduction

The microbial communities that live within and around us are typically species-rich and unevenly diverse; they are unique for hosts and environments, yet dynamic, evolving, and adapting [1–3]. Monitoring the compositional structure of microbial communities is important for understanding of their interactions with their host or the environment, and in response to changes thereof [4,5]. Microbial composition analysis mostly entails quantifying abundances of key populations or taxa, ideally accompanied with their physiological states and metabolic activities [6].

Current microbiome analysis is largely dominated by 'omics'-methods, such as metagenomics (e.g. 16S rRNA gene amplicon sequencing from community-isolated DNA or whole community shotgun sequencing), metatranscriptomics and metaproteomics, and metabolomics. These permit a direct inference of microbial taxa, gene expression, proteins and metabolic reactions within communities [7]. Particularly, both 16S rRNA gene amplicon and deep shotgun sequencing have become extremely popular for microbiome analysis [8,9]. They have tremendously contributed to the understanding of the extent of microbial diversity, the differences and commonalities in species distributions among habitats, hosts and even individuals [10]. Although crucial, the methods have some important drawbacks. Firstly, several studies have pointed out the potential biases in interpreting microbiome structure and function from metagenomics alone [11,12], and have suggested they should be complemented by methods providing cell mass and absolute microbial abundances. Secondly, it is complicated to deduce cell physiologies and growth stages from metagenomics and—transcriptomics methods, yet both are important for microbiome functional interpretations [13]. Finally, most omics methods are not easily optimized to near-real-time results. They require relatively long processing and analysis times (i.e. weeks to months), expensive instruments or outsourcing, and expert bioinformatics knowledge. Therefore, there is clearly substantial room for alternative and complementary methods in microbiome research.

As we elaborate on below, flow cytometry (FCM), in particular combined with machine learning analysis of multiparametric single cell data sets, may provide rapid, quantitative, phenotypically and possibly even taxonomically relevant information of microbiome samples. For recurring samples, one can envision standardizing and optimizing procedures to such an extent that

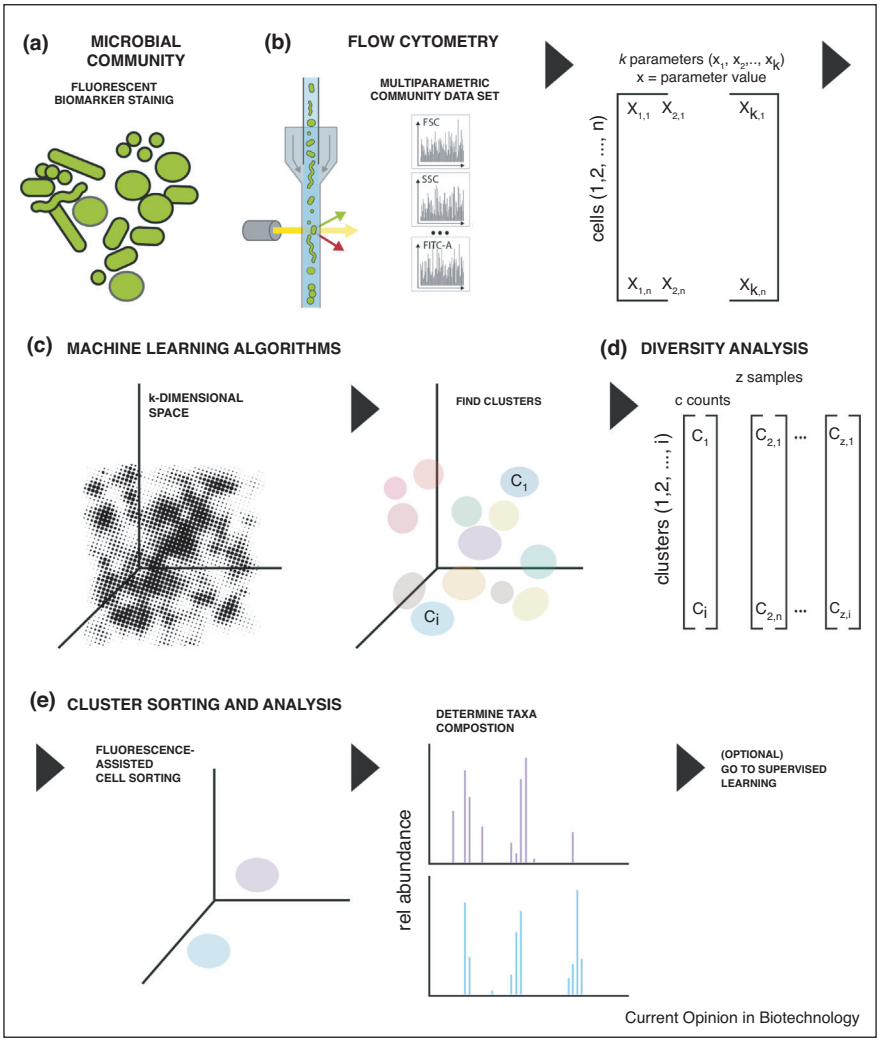
☆ Given his role as Editor in Chief, Jan Roelof van der Meer, had no involvement in the peer review of the article and has no access to information regarding its peer-review. Full responsibility for the editorial process of this article was delegated to Mark P. Styczynski.

comprehensive near-instant microbiome compositional analysis would be possible, which would be extremely useful for diagnostics and treatment of microbiome-related disorders or biotechnology process optimization, as has been suggested previously [14].

Flow cytometry and microbiome analysis

FCM is the current gold standard for single cell quantification and allows rapid, sensitive and high throughput analysis of microbial cells in suspension [15]. FCM has a long history in the analysis of microbial cell cultures and communities, with its advantages of absolute cell quantification, wide dynamic range (10^2–10^7 cells per ml), fast turnaround time, and simplicity with respect to methodology [16,17]. Most modern instruments use a variety of light beams to hit each passing cell, and measure optical diffraction (light scatter) as well as a range of emitted spectral wavelengths or bandwidths (Figure 1a,b). There is a wide range of well-studied fluorophores that can target cellular biomarkers (e.g. SYBR Green I increasing fluorescence emission upon complexing double-stranded nucleic acids) [15,16,18]. Different cellular markers can be detected simultaneously by different fluorophores, within the limits of spectral overlap, compensation techniques or by full spectral analysis. Five to seven fluorophores can be combined unproblematically [19]. FCM

Figure 1



Outline of an unsupervised machine learning approach to analyse flow cytometry community data. Fluorescent-stained cellular biomarkers in the microbial community sample **(a)** are analysed by FCM **(b)** to produce multiparametric datasets of each analyzed cell in the community. Machine learning is used to find clusters or patterns in the datasets **(c)**, which reduces the data to cell abundance per cluster lists **(d)**. These can be deployed for classical diversity analysis. Optionally, identified (bivariate) clusters are physically sorted **(e)** and their taxa composition is determined by 16S rRNA gene amplicon sequencing. The clusters and compositional taxa information can be used as input for further supervised learning. Figure panels (a) and (b) inspired from Ref. [46••] Figure 1.

instruments and methods are flexible [17], permitting single or automized high-density sample throughput and even semi-continuous measurements of automated, auto-stained samples [20]. Some FCMs are equipped with mass-coupled detectors (mass cytometers) to measure masses of passing particles, further widening the potential to detect different biomarkers by, for example, metal-carried antibodies [21]. Although this has so far been applied mostly on eukaryotic cells, imaging cytometers can add further measurable parameters from the cell images they take [22]. Important advantages of FCM for the perspective of microbiome analysis are thus the multiple detected morphological, cellular and/or physiological features measured on a relatively large number of cells in a community. Such large parametric data sets are excellently suitable for machine-learned training and interpretation (Figure 1b,c).

Machine-learned interpretation of multiparametric FCM datasets

Machine learning is a vast domain of advanced computational and statistical methods with the purpose to facilitate interpretation of big datasets in all fields, including microbiology [23•]. Machine learning models can help to represent complex data relationships, albeit without necessarily showing underlying causality (see also Section 'Cytometric and taxonomic relatedness') [24]. *A priori* one would expect FCM community data sets to be complex, consisting typically of 10^5 cell events and multiple species, with each cell being characterized by between 5–15 cytometric features (Figure 1b). The data set is further expected to display some level of (but *a priori* unknown) redundancy, with cells from the same species measured multiple times (Figure 1c). Broadly speaking, two types of machine learning methods have been used to unravel microbial FCM data, categorized as unsupervised and supervised learning. Briefly, unsupervised learning algorithms help to cluster similar data points in a k-dimensional dataset. In supervised learning, a model is trained on a labeled dataset (i.e. a list with class labels and their corresponding input variables, like *species* and *cytometric parameters*) and then produces a classifier. The classifier is a complex non-linear mathematical formula predicting the probability of given input variables to belong to any of the defined output classes. Now how do these machine learning approaches translate into multiparametric FCM community data analysis?

As an example, consider a microbial community as a set of S strains with y physiological states in n relative abundances ($S_{y,n}$). From pure culture studies with fluorescently stained biomarkers (e.g. DNA, RNA, lipids), it is known that coherent physiological states will produce FCM signals that are nearly normal (Gaussian) distributed [25]. One would thus expect an FCM community data set to comprise a composite of all these ($S_{y,n}$) individual phenotypic characters, centered on their respective Gaussian means. The question is then whether we can deconvolute any community multiparametric data set (n cells, each with k parameter values, Figure 1b) back into the individual strains, physiologies and abundances? Given that species are not evenly distributed in communities, many of the individual Gaussians may be difficult to detect, when the population of a given species in a given physiological state contained within the sample is too small (e.g. Figure 1c). Some others may actually be overlapping between species and physiological states, limiting their proper discrimination [26••]. Cells within a community may also display more phenotypic heterogeneity than what is expected from pure cultures, leading to diffuse 'Gaussian' signals [27–30]. The goal of machine learning methods is thus to interpret this community composition or 'cytometric fingerprint' as it has been called [31]; the global makeup of the taxonomic and physiological states of the cells within the community.

Unsupervised learning of microbial FCM data

Unsupervised learning methods aim to reduce the entirety of the cytometric fingerprint (i.e. the list of all cells with their 5–15 FCM parameter values) into defined and quantitated patterns (Figure 1c,d). This is less intuitive than it seems because – as explained above, the FCM data sets comprise multiple underlying strain-dependent, phenotype-dependent, parameter-dependent and density-dependent Gaussians, which would require hyperdimensional comparisons. Fingerprinting therefore often first entails a definition of detectable hyperdimensional clusters (Figure 1c), and then, a comparison of those clusters, for instance, by the numbers of cells they encompass [18,25,26••] (Figure 1d). Since the exact number of clusters (from the constituing strains and their physiologies) in most samples is not *a priori* known, its determination remains an approximation, and a wide variety of clustering approaches exist [26••]. For more theoretical considerations on clustering approaches, the interested reader is referred to Refs. [32,33]. In order to simplify the clustering analysis, the data can be split into a single or multiple bivariate comparisons (each consisting of, for example, two measured FCM parameters) [31]. Bivariate data are subsampled to the same community size, then manually (i.e. gated) [34], arbitrarily (i.e. discretizing the 2D-area in bins, or representing bivariate cell densities as grey-scaled images with pixel resolution as bins) [35] or automatically clustered (i.e. detecting real Gaussian density distributions) [25]. Once the clusters have been identified, the numbers or density of cells within them are quantified (Figure 1d). Subsequently, the resulting list of categories (i.e. bins or clusters) with their respective cell abundances can be compared across samples, for example, by calculating sample distances

using classical diversity matrices, based on the defined cluster categories [26••].

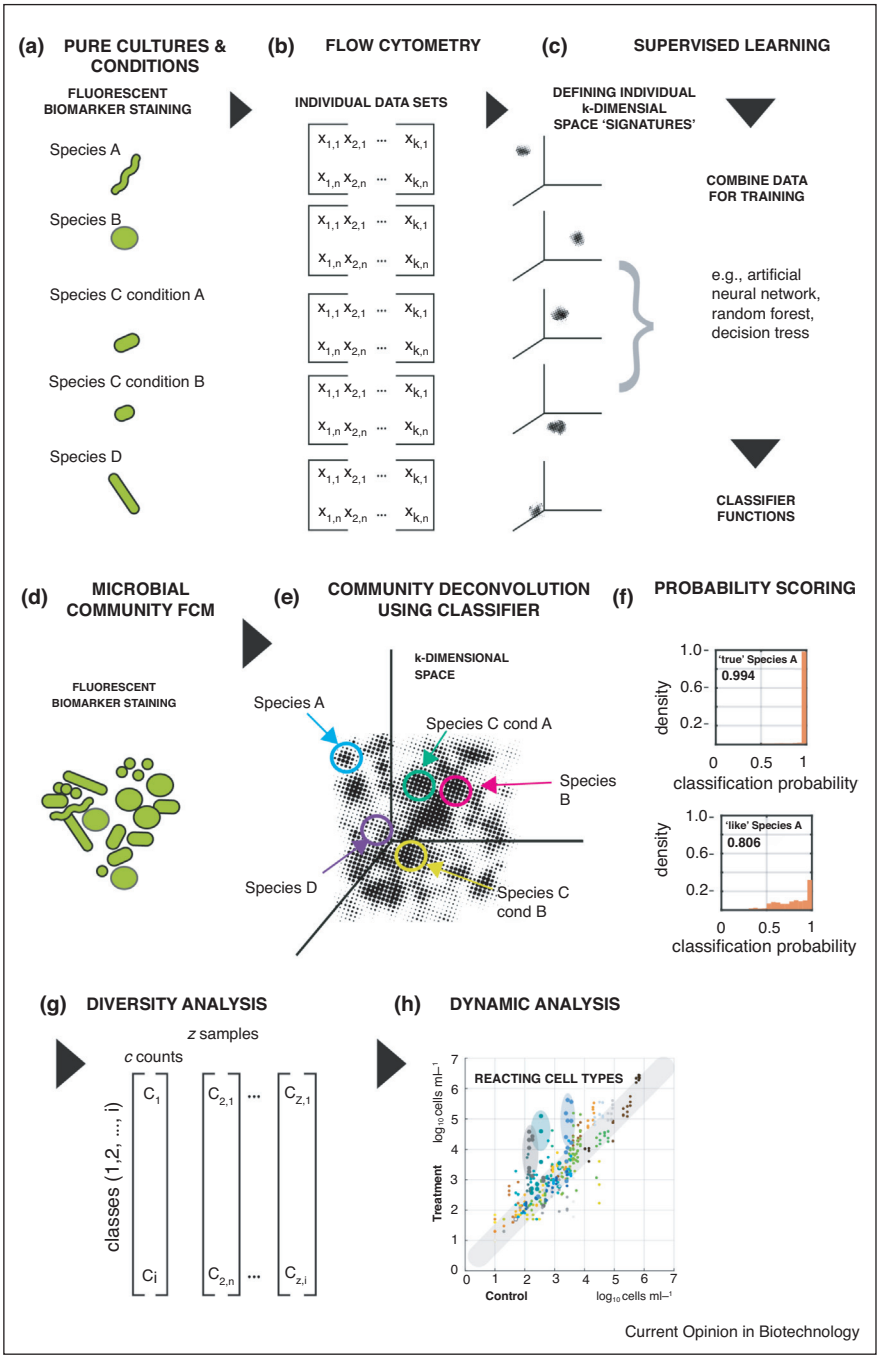
To gain better multivariate comparative power, recent approaches cluster the FCM data directly in hyperdimensional space, for instance, by Gaussian mixture models (PhenoGMM [36•]) or hyperdimensional bins [37], which considerably improves their accuracy. Other approaches use neural network competitive training and mapping to self-organize data clusters [19,38]. The cluster definitions obtained from cytometric fingerprinting can themselves be used for supervised training to better predict recurrent community 'types' or changes (Figure 1e) [26••,39,40•]. The limitation of the unsupervised clustering methods on FCM data is that it is not *a priori* possible to understand what identified clusters consist of, nor to know in advance the amount of clusters to expect. As all clustering methods rely on some underlying estimation of data clusters (e.g. equal cluster sizes or hyperdimensional space separation) [32,33], this adds a black-box aspect to unsupervised clustering and a sense of non-causality (see Section 'Cytometric and taxonomic relatedness' below).

Supervised learning of microbial FCM data

Although the diversity of the microbial world as a whole may be endless, the diversity in any microbiome sample clearly is not, and 95% of sampled cells in microbiomes typically constitute anywhere between 50–100 genotypes, depending on sample size [41,42]. Conceivably, therefore, microbiome samples may be characterized from the phenotypes of their individual constituting strains. This is the concept behind supervised learning of microbial FCM data that aims to learn from training on known categories (Figure 2). First, individual microbiome species are grown as axenic pure cultures, for instance, under conditions that represent the microbiome's natural state (Figure 2a). The cultures are stained and individually analyzed by FCM in order to detect the means and variations of each of the measured FCM parameters (Figure 2b). In case the culture displays multiple physiological states, these may be identified from separate subpopulation Gaussians in the FCM data (see also Figure 2b,c). The FCM parameter values of each distinct strain and physiological state or apparent subpopulations, each with the same amount of cells, now build the data sets that the supervised learning algorithm will train on. The algorithm trains, validates and tests how it can best differentiate all the provided strains, subpopulations and states (the output *classes*) (Figure 2c), and builds a classifier. Next, the classifier function can be deployed to analyze unseen FCM data from the same microbiome (Figure 2d) that contains the strains used for training (Figure 2e). The output is a prediction of the probability for every cell in the sample to fall into each of the trained classes (Figure 2f). The class attribution list is then used as the basis for diversity measures (Figure 2g) or dynamic analysis of cell type abundances (Figure 2h). The classifier function can also be used to categorize cells from unknown microbiomes (for which the same set of FCM parameters has been measured as was used for training). But in that case it is more difficult to interpret the output class attribution, because the probabilities of cell to class assignment may be lower as a result of dissimilarities to the used strain standards (Figure 2f).

So far, few studies have attempted supervised learning to classify individual cells in FCM data; however, the results are promising, because they attempt to link a causality back to the community sample data [43–45,46••,47]. A pioneering study with marine algae profited from their autofluorescence properties and larger cell sizes than bacteria to train an artificial neural network and discriminate 72 different taxa at around 80% accuracy [43]. Another earlier study with four bacterial strains found 68–99% successful recognition with a support vector machine classifier based on five discrete light scattering properties [48]. Three recent studies focused on differentiating (mostly) bacterial communities, by using classifiers built on pure culture FCM data [45,46••,47]. Rubbens *et al.* [47], deployed basic FCM scatter parameters plus a single general nucleic-acid stain, to test how well *in silico* mixtures of FCM cell data were differentiated as a function of the number of training sets. Their linear discriminatory analysis (LDA) and random-forest decision trees showed that any two strains randomly picked from all the data sets could be relatively well discriminated (80% correct). However, the accuracy decreased to around an average of 40–50% when all strain data sets (n = 20) were trained simultaneously, with random-forest giving overall better accuracy than LDA. Later re-analysis of the same data with a new distance calculation improved the discrimination of cell data by a few percent [45]. In our own work [46••,49], we trained an artificial neural network model on a set of 32 standards (consisting of 14 bacterial isolates, some of which with two or more subpopulations and physiological conditions, one yeast, and eight fluorescent bead standards) with 7 FCM parameters including a single nucleic-acid staining dye. Standards were on average well differentiated (recall of 80%), albeit with considerable variation (recalls of between 27.3–99.8%) [46••]. Some of the poorer differentiation seemed the result of taxonomic similarity between the isolates chosen as standards, but not in all cases, and this will require further investigation. Bacterial cells from pure culture spiked into freshwater communities were recognized by the classifier with accuracies of between 56–80% and high probabilities (average >85% on individually assigned cells) [46••]. Importantly, that study also showed that different cell physiologies can be correctly recognized (>80%), even within background of a diverse microbial community. In addition, even unknown

Figure 2



Outline of a supervised machine learning approach for interpretation of flow cytometry community data. Individual pure cultures of taxa occurring in the sample **(a)** are analysed separately by multibiomarker flow cytometry **(b)**, cleaned to define coherent main populations **(c)**, then combined and used for training, testing and validation of supervised input-output class assignments. When satisfactory classifiers are obtained, these are then deployed to deconvolute similarly stained **(d)** and FCM-passed community samples **(e)** into corresponding output classes with the highest probabilities **(f)**. The corresponding lists of cells-per-class attribution are normalized and compared among samples to analyse community diversity **(g)** or attribute dynamic changes **(h)** to output-class subpopulations ('cell types') under influence of treatment or condition. Example in (h) modified from Ref. [50] Figure 5C. Figure panels inspired from Ref. [46••] Figure 1.

communities could be meaningfully analyzed with a single classifier based on these 32 standards [50,51•]. Collectively, these studies showed promising results for supervised learning-based cell type recognition, even within diverse communities. As this is a rather uncharted territory, the molecular mechanisms of the observed successes and the reasons for underlying pitfalls will need to be investigated further.

Cytometric and taxonomic relatedness

Overall, several studies have now demonstrated that machine-learning applications can be extremely valuable for FCM microbial community analysis, in order to help recognize biologically and ecologically meaningful patterns and clusters, or even distinguish and quantify occurrences of specific cell types by comparison to predefined standards. In this manner, cytometric fingerprints have been used to detect ecologically relevant community shifts in process-engineered communities [39,52], to find strain to strain variation [20], to detect different physiological states [53,54], to measure changes in murine fecal microbiota depending on disease-state [42], to diagnose Crohn's disease from human stool samples [55•], or to detect community changes occurring in contaminated sites [56]. Flow cytometry community analysis has further been deployed to study stability and resilience [57], neutral mechanisms and niche differentiation [58••], and nestedness of subcommunity diversity in continuous reactors seeded with wastewater microbiomes [59]. Combined clustering and supervised learning enabled to detect community changes in shrimp-aquaculture pointing to disproportionate taxa [40•], and link abundance changes in identified cell-types in freshwater communities to fragrance biodegradation [51•] or to antibiotic contamination [50]. To some extent, both unsupervised and supervised methods are converging. Cell sorting was used to isolate identified (unsupervised) clusters in microbial community FCM data and identify their composition by 16S rRNA gene amplicon sequencing [42]. This showed that they are not monospecies, but enriched for one or a few strains. The same strains also appeared in multiple clusters, probably due to their different physiological and phenotypic states. A very recent study showed how identified cluster taxonomies may be used for supervised learning of the FCM data, to more reliably predict occurrences and constellations of particular taxa within communities as a consequence of process technology [40•]. This could be useful, for instance, when specific taxa point to microbiome dysbiosis. The opposite has also been tried: applying classifiers trained in supervised learning on pure culture standards to interpret changes in microbial communities with unknown species composition. This procedure was also quite effective in detecting community shifts and responses to xenobiotic compound input, which were highly correlated to 16S rRNA gene amplicon community sequencing analysis [46••,50].

Both supervised and unsupervised machine learning methods thus detect relevant community changes, ecological principles and cell-type occurrences in FCM community data. Their application, however, leads to the more general question of how we as microbiologists can deal with accuracies and probabilities? If the phenotype of a cell (as characterized by k FCM parameters) is 95% similar to that of a predefined standard, we may have no difficulty to accept that this truly is the same phenotype, but how do we interpret a classification of 50% or 80%? Would this still be the same genotype but seen under different physiological conditions, or would this cell belong to a genotype from a different genus or family? Theoretically, only a small dozen of different measured cellular parameters in FCM is sufficient to cover even community samples with high phenotypic richness (e.g. 1000 species and states). For example, measuring 7 independent cell parameters, each with at least 5 different mean values can already create a variation of 5^7 = 78 125 combinations, arguably sufficient to capture the richness in most microbiomes. However, how well can strain differences be captured? In extension of this, how do we interpret and compare global clustering patterns across different unknown microbiome samples [39]? How do we interpret technical FCM 'gates' [59] in terms of taxonomy? These questions arise but we currently don't have enough information to answer them properly. *De facto*, sequencing methods face the same issue; however, there is adequate experience and technical know-how to understand what a 99.99% probability in base-calling means, and to interpret 95% identity of the 16S rRNA gene. Therefore, it may be only a matter of time and more basic comprehension before we gain such confidence with FCM computational methods.

Conclusions

Microbial community analysis is dominated by metagenomics, proteomics and metabolomics, which are robust and accurate, yet complicated, time-consuming, and lacking absolute microbial cell quantification. FCM has been instrumental in rapid cell quantification and simultaneous recording of suitable cell biomarkers, which can be fruitfully interpreted by unsupervised or supervised methods, as many studies have now attested. The advantage by FCM is that the global sample profile is correlated with the taxonomic composition of the community sample but also to the cells' physiological states. This quantitative, taxonomic and physiological profiling offered by FCM and machine learning could present an important advancement in microbial community diagnostics [14]. However, we still need to gain more confidence to link both 'omics' and FCM approaches and find causalities for the observed strain physiologies in the community samples. This may be accomplished by analyzing more and different microbiome samples with both FCM and sequence-based approaches, and further by choosing other or more biomarkers that can link omics-type data

to cell physiological states. We can improve sensitivity and specificity of the methods by better understanding the variation in FCM signals originating from different growth conditions (e.g. in axenic cultures) as well as of sorting cells from community samples. This would then also help us to understand if any microbial community can be reasonably analyzed with a set of 'universal' pre-defined cell standards or pre-established clusters, or if specific and unique classifiers and clusters need to be build for each and every new microbiome or dataset. A comprehensive understanding of the probability assignment of cells to similar pre-defined cell type classes can particularly expand the field of supervised machine learning applications on FCM data. Many packages for machine learning tools on FCM are now available, even for non-expert users and with detailed explanations (e.g. Ref. [26••]). This will help to acquire more user experience, and improve data analysis to make appropriate causal inferences on the occurrences of microbial populations of interest and cell physiologies as well as community composition and functioning.

Conflict of interest statement

Nothing declared.

Data availability

Data will be made available on request.

Acknowledgements

Both authors acknowledge support from the Swiss National Centre in Competence Research *NCCR Microbiomes* and further from an Innotrek grant to B.D.Ö.D.

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Schloissnig S, Arumugam M, Sunagawa S, Mitreva M, Tap J, Zhu A, Waller A, Mende DR, Kultima JR, Martin J *et al.*: **Genomic variation landscape of the human gut microbiome**. *Nature* 2013, **493**:45-50.
2. Ellegaard KM, Engel P: **Genomic diversity landscape of the honey bee gut microbiota**. *Nat Commun* 2019, **10**:446.
3. Ley RE, Peterson DA, Gordon JI: **Ecological and evolutionary forces shaping microbial diversity in the human intestine**. *Cell* 2006, **124**:837-848.
4. Neville BA, Forster SC, Lawley TD: **Commensal Koch's postulates: establishing causation in human microbiota research**. *Curr Opin Microbiol* 2018, **42**:47-52.
5. Raymann K, Shaffer Z, Moran NA: **Antibiotic exposure perturbs the gut microbiota and elevates mortality in honeybees**. *PLoS Biol* 2017, **15**:e2001861.
6. Kesnerova L, Mars RAT, Ellegaard KM, Troilo M, Sauer U, Engel P: **Disentangling metabolic functions of bacteria in the honey bee gut**. *PLoS Biol* 2017, **15**:e2003467.
7. Zuniga C, Zaramela L, Zengler K: **Elucidation of complexity and prediction of interactions in microbial communities**. *Microb Biotechnol* 2017, **10**:1500-1522.
8. Sunagawa S, Mende DR, Zeller G, Izquierdo-Carrasco F, Berger SA, Kultima JR, Coelho LP, Arumugam M, Tap J, Nielsen HB *et al.*: **Metagenomic species profiling using universal phylogenetic marker genes**. *Nat Methods* 2013, **10**:1196-1199.
9. Ruscheweyh HJ, Milanese A, Paoli L, Sintsova A, Mende DR, Zeller G, Sunagawa S: **mOTUs: profiling taxonomic composition, transcriptional activity and strain populations of microbial communities**. *Curr Protoc* 2021, **1**:e218.
10. Thompson LR, Sanders JG, McDonald D, Amir A, Ladau J, Locey KJ, Prill RJ, Tripathi A, Gibbons SM, Ackermann G *et al.*: **A communal catalogue reveals earth's multiscale microbial diversity**. *Nature* 2017, **551**:457-463.
11. Rivett DW, Bell T: **Abundance determines the functional role of bacterial phylotypes in complex communities**. *Nat Microbiol* 2018, **3**:767-772.
12. Vandeputte D, Kathagen G, D'Hoe K, Vieira-Silva S, Valles-Colomer M, Sabino J, Wang J, Tito RY, De Commer L, Darzi Y *et al.*: **Quantitative microbiome profiling links gut community variation to microbial load**. *Nature* 2017, **551**:507-511.
13. Salazar G, Paoli L, Alberti A, Huerta-Cepas J, Ruscheweyh HJ, Cuenca M, Field CM, Coelho LP, Cruaud C, Engelen S *et al.*: **Gene expression changes and community turnover differentially shape the global ocean metatranscriptome**. *Cell* 2019, **179**:1068-1083 e1021.
14. Koch C, Müller S: **Personalized microbiome dynamics - cytometric fingerprints for routine diagnostics**. *Mol Aspects Med* 2018, **59**:123-134.
15. Czechowska K, Johnson D, Van Der Meer JR: **Use of flow cytometric methods for single-cell analysis in environmental microbiology**. *Curr Opin Microbiol* 2008, **11**:205-212.
16. Müller S, Nebe-von-Caron G: **Functional single-cell analyses: flow cytometry and cell sorting of microbial populations and communities**. *FEMS Microbiol Rev* 2010, **34**:554-587.
17. Wang Y, Hammes F, De Roy K, Verstraete W, Boon N: **Past, present and future applications of flow cytometry in aquatic microbiology**. *Trends Biotechnol* 2010, **28**:416-424.
18. Montante S, Brinkman RR: **Flow cytometry data analysis: recent tools and algorithms**. *Int J Lab Hematol* 2019, **41(Suppl. 1)**:56-62.
19. Saeys Y, Van Gassen S, Lambrecht BN: **Computational flow cytometry: helping to make sense of high-dimensional immunology data**. *Nat Rev Immunol* 2016, **16**:449-462.
20. Buysschaert B, Kerckhof FM, Vandamme P, De Baets B, Boon N: **Flow cytometric fingerprinting for microbial strain discrimination and physiological characterization**. *Cytometry A* 2018, **93**:201-212.
21. Bandura DR, Baranov VI, Ornatsky OI, Antonov A, Kinach R, Lou X, Pavlov S, Vorobiev S, Dick JE, Tanner SD: **Mass cytometry: technique for real time single cell multitarget immunoassay based on inductively coupled plasma time-of-flight mass spectrometry**. *Anal Chem* 2009, **81**:6813-6822.
22. Dunker S, Boho D, Waldchen J, Mader P: **Combining high-throughput imaging flow cytometry and deep learning for efficient species and life-cycle stage identification of phytoplankton**. *BMC Ecol* 2018, **18**:51.
23. • Goodswen SJ, Barratt JLN, Kennedy PJ, Kaufer A, Calarco L, Ellis JT: **Machine learning and applications in microbiology**. *FEMS Microbiol Rev* 2021, **45**:fuab015 http://dx.doi.org/10.1093/femsre/fuab015
This article comprehensively reviews key points in machine learning and its applications in microbiology (from clinical microbiology to microbial ecology).
24. Babic B, Gerke S, Evgeniou T, Cohen IG: **Beware explanations from AI in health care**. *Science* 2021, **373**:284-286.
25. Ludwig J, Zu Siederdissen CH, Liu Z, Stadler PF, Müller S: **flowEMMi: an automated model-based clustering tool for microbial cytometric data**. *BMC Bioinformatics* 2019, **20**:643.

26. •• Rubbens P, Props R: **Computational analysis of microbial flow cytometry data**. *mSystems* 2021, **6** e00895-00820
An excellent overview on the full computational analysis pipeline (data pretreatment, machine learning) of flow cytometry data in microbial ecology with an online R resource workflow demonstration.

27. Heyse J, Buysschaert B, Props R, Rubbens P, Skirtach AG, Waegeman W, Boon N: **Coculturing bacteria leads to reduced phenotypic heterogeneities**. *Appl Environ Microbiol* 2019, **85**: e02814-e02818.

28. Ackermann M: **A functional perspective on phenotypic heterogeneity in microorganisms**. *Nat Rev Microbiol* 2015, **13**:497-508.

29. Zimmermann M, Escrig S, Hübschmann T, Kirf MK, Brand A, Inglis RF, Musat N, Müller S, Meibom A, Ackermann M *et al.*: **Phenotypic heterogeneity in metabolic traits among single cells of a rare bacterial species in its natural environment quantified with a combination of flow cell sorting and NanoSIMS**. *Front Microbiol* 2015, **6**:243.

30. Garcia-Timermans C, Rubbens P, Heyse J, Kerckhof FM, Props R, Skirtach AG, Waegeman W, Boon N: **Discriminating bacterial phenotypes at the population and single-cell level: a comparison of flow cytometry and Raman spectroscopy fingerprinting**. *Cytometry A* 2020, **97**:713-726.

31. Koch C, Günther S, Desta AF, Hübschmann T, Müller S: **Cytometric fingerprinting for analyzing microbial intracommunity structure variation and identifying subcommunity function**. *Nat Protoc* 2013, **8**:190-202.

32. Perez-Suarez A, Martinez-Trinidad JF, Carrasco-Ochoa JA: **A review of conceptual clustering algorithms**. *Artif Intell Rev* 2019, **52**:1267-1296.

33. Xu D, Tian Y: **A comprehensive survey of clustering algorithms**. *Ann Data Sci* 2015, **2**:165-193.

34. Koch C, Fetzer I, Schmidt T, Harms H, Müller S: **Monitoring functions in managed microbial systems by cytometric bar coding**. *Environ Sci Technol* 2013, **47**:1753-1760.

35. Koch C, Fetzer I, Harms H, Müller S: **CHIC - an automated approach for the detection of dynamic variations in complex microbial communities**. *Cytometry A* 2013, **83a**:561-567.

36. • Rubbens P, Props R, Kerckhof FM, Boon N, Waegeman W: **PhenoGMM: Gaussian mixture modeling of cytometry data quantifies changes in microbial community structure**. *mSphere* 2021, **6** e00530-00520
An automated computational approach based on Gaussian mixture models to cluster FCM data directly in hyperdimensional space. The authors illustrate the applicability for quantitative community diversity on synthetic and natural microbial ecosystems.

37. Roederer M, Treister A, Moore W, Herzenberg LA: **Probability binning comparison: a metric for quantitating univariate distribution differences**. *Cytometry* 2001, **45**:37-46.

38. Van Gassen S, Callebaut B, Van Helden MJ, Lambrecht BN, Demeester P, Dhaene T, Saeys Y: **FlowSOM: using self-organizing maps for visualization and interpretation of cytometry data**. *Cytometry A* 2015, **87**:636-645.

39. Dhoble AS, Lahiri P, Bhalerao KD: **Machine learning analysis of microbial flow cytometry data from nanoparticles, antibiotics and carbon sources perturbed anaerobic microbiomes**. *J Biol Eng* 2018, **12**:19.

40. • Heyse J, Schattenberg F, Rubbens P, Müller S, Waegeman W, Boon N, Props R: **Predicting the presence and abundance of bacterial taxa in environmental communities through flow cytometric fingerprinting**. *mSystems* 2021, **6**:e0055121 http://dx.doi.org/10.1128/mSystems.00551-21
A two stage flow cytometric pipeline using supervised machine learning to predict the presence/absence of bacterial taxa and their relative abundance based on cytometric fingerprints of sorted clusters with taxonomically different patterns.

41. Locey KJ, Lennon JT: **Scaling laws predict global microbial diversity**. *Proc Natl Acad Sci U S A* 2016, **113**:5970-5975.

42. Zimmermann J, Hübschmann T, Schattenberg F, Schumann J, Durek P, Riedel R, Friedrich M, Glauben R, Siegmund B, Radbruch A *et al.*: **High-resolution microbiota flow cytometry reveals dynamic colitis-associated changes in fecal bacterial composition**. *Eur J Immunol* 2016, **46**:1300-1303.

43. Boddy L, Morris CW, Wilkins MF, Al-Haddad L, Tarran GA, Jonker RR, Burkill PH: **Identification of 72 phytoplankton species by radial basis function neural network analysis of flow cytometric data**. *Mar Ecol Prog Ser* 2000, **195**:47-59.

44. Boddy L, Morris CW, Wilkins MF, Tarran GA, Burkill PH: **Neural network analysis of flow cytometric data for 40 marine phytoplankton species**. *Cytometry* 1994, **15**:283-293.

45. Nguyen B, Rubbens P, Kerckhof FM, Boon N, De Baets B, Waegeman W: **Learning single-cell distances from cytometry data**. *Cytometry A* 2019, **95**:782-791.

46. •• Özel Duygan BD, Hadadi N, Babu AF, Seyfried M, van der Meer JR: **Rapid detection of microbiota cell type diversity using machine-learned classification of flow cytometry data**. *Commun Biol* 2020, **3**:379
Development of a supervised machine learning based pipeline for flow cytometry data recognizing focal strains with different growth physiology within background of diverse microbial communities and differentiating various cell types in microbial communities to monitor community dynamics.

47. Rubbens P, Props R, Boon N, Waegeman W: **Flow cytometric single-cell identification of populations in synthetic bacterial communities**. *PLoS One* 2017, **12**:e0169754.

48. Rajwa B, Venkatapathi M, Ragheb K, Banada PP, Hirleman ED, Lary T, Robinson JP: **Automated classification of bacterial particles in flow by multiangle scatter measurement and support vector machine classifier**. *Cytometry A* 2008, **73**:369-379.

49. van der Meer JR, Özel Duygan BD: *CellCognize: a Neural Network Pipeline for Cell Type Classification from Flow Cytometry Data*. . Edited by https://Zenodo.org 2020 http://dx.doi.org/10.5281/zenodo.3822094.

50. Özel Duygan BD, Gaille C, Fenner K, van der Meer JR: **Assessing antibiotics biodegradation and effects at sub-inhibitory concentrations by quantitative microbial community deconvolution**. *Front Environ Sci* 2021, **9**:407.

51. • Özel Duygan BD, Rey S, Leocata S, Baroux L, Seyfried M, van der Meer JR: **Assessing biodegradability of chemical compounds from microbial community growth using flow cytometry**. *mSystems* 2021, **6** e01143-01120
This article highlights how supervised machine learning of flow cytometry community data unveiled specific subpopulation changes in freshwater microbiota exposed to low concentrations of fragrances.

52. De Vrieze J, Heyer R, Props R, Van Meulebroek L, Gille K, Vanhaecke L, Benndorf D, Boon N: **Triangulation of microbial fingerprinting in anaerobic digestion reveals consistent fingerprinting profiles**. *Water Res* 2021, **202**:117422.

53. Koch C, Harms H, Müller S: **Dynamics in the microbial cytome-single cell analytics in natural systems**. *Curr Opin Biotechnol* 2014, **27**:134-141.

54. Melzer S, Winter G, Jäger K, Hübschmann T, Hause G, Syrowatka F, Harms H, Tárnok A, Müller S: **Cytometric patterns reveal growth states of *Shewanella putrefaciens***. *Microb Biotechnol* 2015, **8**:379-391.

55. • Rubbens P, Props R, Kerckhof FM, Boon N, Waegeman W: **Cytometric fingerprints of gut microbiota predict Crohn's disease state**. *ISME J* 2021, **15**:354-358
A nice example of microbial community structure analysis based on flow cytometric fingerprints and 16S rRNA gene amplicon sequencing, showing highly correlated outcomes of stool microbiota indicative of Crohn's disease.

56. Bombach P, Hübschmann T, Fetzer I, Kleinsteuber S, Geyer R, Harms H, Müller S: **Resolution of natural microbial community dynamics by community fingerprinting, flow cytometry, and trend interpretation analysis**. *Adv Biochem Eng Biotechnol* 2011, **124**:151-181.

57. Liu Z, Cichocki N, Bonk F, Günther S, Schattenberg F, Harms H, Centler F, Müller S: **Ecological stability properties of microbial communities assessed by flow cytometry**. *mSphere* 2018, **3**.

58. •• Liu Z, Cichocki N, Hübschmann T, Süring C, Ofiteru ID, Sloan WT, Grimm V, Müller S: **Neutral mechanisms and niche differentiation in steady-state insular microbial communities revealed by single cell analysis**. *Environ Microbiol* 2019, **21**:164-181

An excellent illustration of interpretation of ecological principles based on flow cytometry community fingerprinting analysis of wastewater-derived laboratory communities.

59. Liu Z, Müller S: **Bacterial community diversity dynamics highlight degrees of nestedness and turnover patterns**. *Cytometry A* 2020, **97**:742-748.