Implementation and evaluation of CellScanner 1.0

Introduction

Resolving the composition of microbiome samples is an open challenge, when one considers the time and price of the available methods. For instance, 16S rRNA gene sequencing is a well-established method for this task, but it requires work-intensive steps to extract and amplify the DNA. Amplification introduces biases (1–4) and 16S rRNA gene copy number can differ from one species to another (5, 6). Due to these biases, total 16S rRNA read count per sample does not reflect total cell count and thus, read counts need to be converted to relative abundances. Absolute abundances can be obtained by multiplying these with the total cell count measured with flow cytometry or qPCR (7–10).

Alternatively, flow cytometry (FC) can be employed to investigate community composition (11, 12). In flow cytometers, particles pass a laser beam in single file. Flow cytometers then collect scattered light or, in the case of fluorescence, emitted light that provides information about cell shape and size (13). If cells are stained with a fluorochrome, they become distinguishable through fluorescence (14). FC is a promising technique to monitor microbial communities thanks to its low cost, quick acquisition of results and the fact that it gives count data instead of relative abundances.

Unsupervised techniques are commonly applied to FC data, for instance, to identify immune cells and cluster them by cell type (15, 16). Unsupervised classification also serves to identify clusters of interest in microbial communities (17–20) with tools such as FlowSom (21), FlowGrid (22), or, more recently, FlowEmmi (23) and FlowGateNIST (24). These clusters do not necessarily correspond to single species, but allow tracking changes in microbial

community composition across time or conditions. In some applications, diversity is assessed through cluster enumeration (25).

The application of supervised classification to FC data was advocated by Frankel DS et al. (26) 30 years ago. The key idea is to train classifiers on FC monoculture data that can then be applied to count the cell numbers for different species in a community without needing species-specific labels. This idea has since been implemented in several tools that count human cell types and applied a few times to microbial species, particularly phytoplankton (27–29). However, applying this concept to bacterial communities is more challenging since they usually do not have such distinctive shapes and are much smaller. Flow cytometers need a sufficiently high resolution to distinguish bacterial cells from the background (debris) (20). In practice, gates are defined by designating areas in scatter plots of FC channels to select events to keep. However, these scatter plots take only two FC measurements into account among several (i.e., 14 channels for Accuri BD C6). Gates depend on the sample studied, the analysis and the flow cytometer and thus usually have to be adjusted across samples and devices.

Despite these challenges, the composition of in vitro communities (ranging from two to five bacterial species) was, in several cases, accurately predicted with supervised classification as assessed by mock communities (30, 31). These publications are accompanied by analysis pipelines, particularly a Python script (28), referred to as ScriptP, and a MATLAB script called CellCognize (31). However, these scripts require programming knowledge and must be adapted depending on the flow cytometer and gating strategy used.

We, therefore, developed CellScanner, an open-source tool written in Python that is easy to install and use without programming skills. CellScanner supports both supervised and unsupervised classification methods. As summarised in Figure 1A, the program processes raw flow cytometer files of monocultures and then trains predictive models (classifiers) on them.

These classifiers can then be used to predict the composition of a community (*Prediction*). CellScanner can also build in silico communities of known composition from monoculture files and run classifiers on them to assess prediction accuracy for a species combination of interest (*Tool Analysis*). CellScanner can produce an average (aggregated) prediction from the individual predictions of multiple classifiers (by default 10, see Methods) through a majority vote (Figure 1B). The user can also set an unknown threshold, representing the percentage of classifiers that must agree to label an event as a species; otherwise, the event is labelled unknown (see Methods).

CellScanner offers different types of gating that are carried out before any calculation and that include both line gating and a novel automated gating strategy called *machine gating*, which applies supervised classification to monoculture data from each species and the medium (blank).

Table 1 compares CellScanner with different tools that cluster FC data of microbial communities. From the five tools listed, only CellScanner, CellCognize and ScriptP were developed for species-specific classification of microbial communities and were thus selected for comparison.

Results

Optimising CellScanner settings

First, we evaluated the effect of different CellScanner parameters on its performance using three data sets with 11 microbial species (Data sets 1 and 2, Table 2) that were combined into 30 species pairs in silico. Classifiers are trained on monoculture events that were not used for community construction. Since the correct species of each event is known, it is possible to compute the specificity, precision, F1 score, and accuracy of the classification. The classification with neural networks was significantly (p-values < 1.114e-05 for all pairs) more

accurate than the other methods, including the (unsupervised) clustering algorithm (Supplementary Figure 1) and was thus selected for parameter evaluation.

We started by comparing different gating strategies. For this, we included two gating strategies previously developed to count gut microorganisms in flow cytometry data collected with Accuri or CytoFLEX. We also added a gating method based on supervised classification, where classifiers are trained on the co-culture and blank files (containing data for the medium without inoculation) to distinguish between cells and debris. This gating method, termed machine gating, is carried out with the same classification method and settings as selected for species counting. Finally, we included the case without gating as a negative control. As expected, the absence of gating results in the highest percentage of unknown events (15.7%, Figure 2A). Because the unknown parameter labelled some background events (i.e., non-cell) as unknown, the accuracy of prediction increased even without gating. The results obtained with line gating methods (i.e., for Accuri and Cytoflex) accuracy are not significantly different. However, line gating is specific to a flow cytometer (Supplementary Table 1) and, when not adapted, can remove most cell events from a monoculture. Overall, we found that the highest accuracy is obtained with the new machine learning gating. This result highlights the importance of blank samples, which are required for machine gating as training data.

Next, we evaluated the impact of the number of runs, i.e., the number of times the prediction is carried out with a classifier trained on different subsets of the monoculture data. Accuracy significantly increases with the number of runs at first but then saturates between eight and twelve runs (Figure 2B). The number of predicted unknown events mainly causes the changes in accuracy and F1 score in the plateau. All calculations have a 70% unknown threshold. Therefore, with the number of runs increasing, classifiers needing to agree on a label (>70%) will increase, but not linearly, since the number of clusters is an integer. Majority vote results in higher accuracy than calculating the mean accuracy of ten predictions (Figure 2C) even

without unknown labelling. In addition, the accuracy, F1 score, specificity, and precision increase with the percentage of runs that must agree on an event (unknown threshold). Thus, classifying events as unknowns reduces the number of false positives.

Finally, the training set size (i.e. the number of events selected for training per monoculture) initially increases accuracy but does not further improve it (Figure 2D). Since the size of the training set also increases the calculation time (especially for neural networks), training set sizes above 2000 events are not necessary.

Comparison of tool performance for in silico and in vitro communities

Next, we compared CellScanner to CellCognize and ScriptP. We selected FC data from a collection of datasets, including different cell types (e.g., human cells, bacteria) for a fair, exhaustive comparison with enough data (e.g., 1064 in-silico communities). The collection of datasets also includes different cytometers (CytoFLEX, BD Accuri C6, FACSVerse, NovoCyte) and different cell treatments (e.g., staining). CellScanner supports logistic regression, random forest, neural network and random guessing (see Methods), ScriptP implements both random forest and linear discriminant analysis (LDA) and CellCognize is based on neural networks. For each machine learning method in CellScanner, community composition was predicted with and without the unknown parameter. Classifiers were generally trained on data from all available channels (except for the 'Time' channel).

From the six datasets selected (i.e., Datasets 1, 3-7), 1045 in silico two-species cocultures were created with a ratio of 50% for each species. Since in the real world, the species of an event is not known, we did not compute accuracy but instead assessed how close the predicted community composition is to the real one. For this, we calculated the Euclidean distance between the expected and the predicted composition (Figure 3). For random guessing, the predicted composition was expected to be close to the expected one (50:50). Most of the cells (90%) were labelled as unknown for random guessing with unknowns, resulting in small final

cell numbers and thus, large variation. CellCognize has the largest Euclidean distance to the expectation, which may be due to its embedded gating strategy. Note that for the other methods using unknowns, the number of unknowns represented less than 10% of the events for most of the predictions (Supplementary Table 2; i.e., LG:99,3%; RF:95,6%; NN:82,4%). The best-performing methods, include neural network, random forest with and without unknowns for CellScanner and random forest for ScriptP. Unless for random forest with unknown from CellScanner, these methods were not significantly different according to the paired Wilcoxon test (p-values>0.06).

Next, we evaluated method performance on four different in vitro communities (Datasets 5 and 7), which include three communities of two species and one community of three species, in different ratios. We compared the Euclidean distances between expected and predicted compositions for one or all species together, for each community, and for all communities together (Figure 4A-E). As described in Rubbens et al., 2017(30), the composition of community 1 is the hardest to predict, and except for CellCognize, all methods performed the worst on this community. According to the Wilcoxon test, no method was significantly better than all the others in all communities taken together. As seen previously, CellScanner with random forest or neural network, with and without unknowns, and ScriptP with random forest performed best overall (Figure 4F). No classification method systematically outperformed the others when all predictions of communities are taken into account.

CellScanner applied to a bi-culture of river bacteria

Finally, we applied CellScanner to an unpublished in-house data set of river bacterial species *Brevundimonas sp.* (124Z) and *Variovorax sp.* (1315Z) (Dataset 8). The species were inoculated in the same amount based on flow cytometry cell count to reach a concentration of ca. 10^4 cells/mL and grown together for 72 hours. Samples were analysed with FC at eight time points. In addition, the growth of each species was also assessed with FC in monoculture

at the same time points. We predicted the bi-culture composition with CellScanner for each time point, using all corresponding monoculture samples, merged by species, as training data for each run (Figure 5). We also compared the performance of different machine learning methods using CellScanner's *Tool Analysis* function. The highest accuracy was obtained with random forest (Figure 5A); therefore, this method was applied to predict bi-culture composition. Events were gated with machine learning (Figure 5B). The 3D plot shows clusters labelled as separate species, and only a small percentage of the events in the blank clusters are assigned to a species. The predicted ratios suggest that 1315Z quickly dominates the community (Figure 5D). In addition, the predicted growth curve shows a long lag phase for 124Z (Figure 5E). In summary, CellScanner can predict species-specific growth curves given a time series of bi-culture flow cytometry data.

Discussion

We assessed classification accuracy with in-silico as well as in-vitro communities. Although the overall community composition in vitro is known, it is not known to which species each cell belongs, and the accuracy can therefore only be assessed indirectly through the comparison of the predicted and expected community composition. For a two-species coculture, the classification can be wrong and still result in a prediction close to the correct proportion by chance. For this reason, mock communities with a range of ratios and with more than two species were evaluated (Figure 4D). However, the number of evaluated communities was small, highlighting the need for more mock community studies. We expect that increasing the number of species in the community will lower the prediction accuracy since the probability that different species generate overlapping FC values grows with the species number.

As expected, prediction accuracy is directly dependent on the dataset used as reference (i.e. the monoculture, Figure 4). We also observed that there is not a single machine learning

method that systematically outperforms the others. Logistic regression is fast but only works well on well-separated communities, whereas neural networks and random forest are slower but can deal with more challenging clustering tasks. Therefore, with three machine learning methods implemented, CellScanner enables users to find the optimal method for their dataset by carrying out an in-silico evaluation with the *Tool Analysis* function.

The unknown parameter was implemented to deal with prediction uncertainty. Such uncertainty is caused mainly by similar morphology, which leads to large overlaps in species' FC parameters. In the evaluation, this parameter was observed to reduce the number of false positives, thereby increasing the precision and specificity for each species (Figure 2C). However, in the two-species in-vitro community evaluation, not assigning unknowns led in some cases to better performance. The unknown parameter may reduce the accuracy if one species has many more unknown events than other species. In general, when CellScanner predicts a large proportion of unknowns, the community may not be well suited for supervised classification.

To our knowledge, the machine gating method implemented in CellScanner is the first to rely on supervised classification to gate FC data. Compared to other gating methods in flow cytometer programs, CellScanner machine gating brings more flexibility since it does not require defining thresholds. If a gating method relies on the same gate definition for all files, a shift in the event versus background clusters could lead to the classification of many cells as debris or vice versa. The machine gating can fail if there is a large amount of background events compared to cells and if the background in the blanks is not representative of the background found in mono- and cocultures. Previously, an alternative automated gating method, tracking cell populations, was described in OpenCyto (32), and another one based on unsupervised clustering was recently published as FlowGateNIST (24). Both eliminate the

need for blanks. However, the performance of these methods compared to gating based on supervised classification remains to be evaluated, especially on microbial data.

Finally, while supervised classification can reach high accuracies on mock communities with known composition, it is hard to assess its performance on growing communities. In such communities, cell morphology may differ from the one in monocultures (33) and may even change over time. However, since the exact composition of these communities is unknown, the result of FC analysis can only be compared to other counting methods, which have their own biases. We assessed the species composition for a growing community of selected human gut bacteria with both 16S rRNA gene sequencing and flow cytometry data analysed with CellScanner and found that the main trend agreed for both techniques (34). In general, more such benchmarks are needed for applications to determine whether supervised classification is a reliable counting method for the community of interest. Unsupervised classification avoids the need for monocultures and hence the bias that comes from changing morphology. Still, its accuracy on mock communities was too low for it to be considered a feasible alternative.

CellScanner can assess species classification accuracy on monoculture data in silico and can quickly predict community composition when given mono- and coculture data. The optional classification of events as unknown and the machine learning gating are new techniques that increase prediction accuracy. With its user-friendly GUI (graphical user interface) and informative output, CellScanner makes supervised classification of FC data available to users without programming experience.

Materials and methods

CellScanner

Overview

CellScanner is available on command line and via a graphical user interface (GUI). It wraps *fcsparser* to read flow cytometry standard (FCS) files and relies on *pyqt5* for its GUI.

The GUI version does not need Python pre-installed on Windows; for other operating systems, Python and dependencies must be installed first. CellScanner contains functions for supervised and unsupervised classification, respectively. The latter uses agglomerative clustering in scikit-learn. The GUI of CellScanner interfaces with a database that contains previously selected reference (i.e. monoculture) data and provides step-by-step guidance on tool use. CellScanner offers three supervised classification methods, including random forest, logistic regression, and neural network (implemented in scikit-learn) and also allows to assign events randomly to species as a control. CellScanner's main parameters are the number of events used for training (7/8) and testing (1/8) (defaulting to 1000 taken together) and the number of runs (defaulting to ten). The names of the flow cytometer channels to be used can be indicated if not all of them are needed; however, the '*Time'* channel is always removed from the calculation since it is not linked to cell characteristics. By default, CellScanner does the following in each run: *i)* select events randomly from the monoculture FC files with a slight overlap between data sub-sets (mean overlap below 5%, Supplementary Figure 2), *ii)* train a classifier on these events using the selected method and *iii)* apply the trained classifier to events in the community FC file. This gives ten classifications per event in the community FC file; thus, the species of an event is assigned by majority vote. In a case of a tie, the event is relabelled randomly. The unknown parameter allows CellScanner to tag an event as unknown if the percentage of the predominant species predicted for an event is lower than or equal to a user-defined threshold (*unknown threshold*). For instance, if the user keeps the default of 70% as threshold and only seven out of ten runs agree on the same species, then the event is labelled as unknown. The agreement between runs varies depending on the classification method and data set, but in most cases, nine classifiers agree on the prediction (Supplementary Figures 3).

Gating methods

Two gating methods have been implemented in CellScanner. The first is a set of equations implementing the line gating described in Vandeputte et al. (2017)(35) for the BD Accuri C6 cytometer or the Cytoflex cytometer, which distinguishes single cells from two aggregated cells (doublets). This discrimination relies on a comparison of the FSC-A and FSC-H channels, which should be proportional. The Accuri line gating is described as follow:

$$FL3A \leq 0 \; or \; FL1A \leq 0$$

$$FL3A > 0{,}0241 \times FL1A^{1.0996}$$

$$FSCA > 100000 \; \& \; SSCA > 10000$$

$$log(FSCA) > log(FSCH) + 0{,}5$$

$$log(FSCA) > log(FSCH) - 0{,}5$$

For the CytoFLEX cytometer, line gating follows this set of equations:

$$FL3A \leq 0 \; or \; FL1A \leq 0$$

$$log(FL3A) > 1{,}5 \times log(FL1A) - 2{,}8$$

$$log(FL2A) > 2{,}5 * log(FL1A) - 9$$

$$log(FSCA) > log(FSCH) + 0{,}6$$

$$\log(FSCA) > \log(FSCH) - 0{,}6 \; log(FSCA) > log(FSCH) - 0{,}6$$

Since the brand and the configuration of a cytometer affect the gating, we developed a machine learning gating that avoids arbitrary equations. For this, CellScanner compares FC data from the blank files (i.e. medium without cells) to mono- or coculture data. The tool trains six classifiers from the chosen method with 1000 events from a blank FC file and 1000 events from a species FC file. Using majority vote and an unknown threshold >70% for monocultures of the same species, the program will tag most of the overlapping non-cell

events between blank and monocultures FC files as either blank (i.e., background noise, medium debris) or unknown, and remove them. If "machine" is selected as gating technique, the user must add blank file names and tag these as 'blank'. Without indicated blank data, the program will not perform any gating.

Supervised and unsupervised classification

The *logistic regression* classifier uses the 'lbfgs' solver and L2 regularisation, which reduces the weights of predictors with a penalty term. In the multiclass case, the function automatically selects the 'multinomial' option that minimises the multinomial loss and fits across the entire probability distribution.

The *random forest* relies on 200 estimators maximum and the Gini impurity as a criterion to measure the quality of a split.

The *neural network* classifier uses 200 layers, a rectified linear unit function to activate the hidden layer and the 'lbfgs' solver. For each classifier, the parameters were optimised on a set of data and default values were set accordingly, which are only modifiable in the command-line version of CellScanner. The machine gating function uses the supervised classification method specified by the user for the prediction and associated parameter(s).

In *random guessing*, an event is randomly assigned to one of the species names indicated by the user.

The only unsupervised classification method available in CellScanner is agglomerative clustering implemented in the scikit learn package. The natural logarithm of ten is taken for every feature in the FC data. All samples with null or NaN values are discarded before applying the logarithm. To help the user labelling the clusters with their corresponding species, CellScanner takes monoculture data for each species. A distance is calculated

between the clusters and the monoculture data, and each cluster is labelled with the species of the closest monoculture.

Output

CellScanner generates the same output on command line as with the GUI. A directory named with the date and time of the start of the analysis is created in the 'Results' directory located in the installed program directory. The user can also choose where the output is saved on command line. For each step, including the prediction and the gating, if desired, the program produces excel tables with statistics, species counts, as well as 3D graphs that can be manipulated by the user to find the best angle of view and that, compare the expected to the predicted labelling if applicable. Settings are likewise saved.

For the *Tool Analysis* output, the program calculates statistics based on the true positives (TP), the false positives (FP), the false negatives (FN), the number of species in the community (n), the total number of events and the number of unknown (Nb tot event; Nb unknown). TP, FP and FN are specific to a species in the community (*i*). The calculation also considers the false negatives of a species *i* attributed to the unknown category ($FN^i_{unknown}$) and the weight defined by the number of true instances per species (w).

Accuracy is defined as $$Accuracy = \frac{\sum_i^n TP^i}{Nb\ tot\ event}$$

In the presence of unknowns, accuracy is defined as $$Accuracy = \frac{\sum_i^n TP^i}{Nb\ tot\ event - Nb\ unknown}$$

F1-score is defined by the scikit-learn package with the 'weighted' average parameter resulting in: $$F1score = \frac{\sum_i^n (w^i \cdot \frac{2TP^i}{2TP^i + FP^i + FN^i})}{\sum_i^n w^i}$$

With unknown events, the F1 score is defined as $$F1score = \frac{\sum_i^n (\frac{2TP^i}{2TP^i + FP^i + FN^i - FN^i_{unknown}})}{n}$$

The precision per community is defined as $Precision = \frac{\sum_i^n \frac{TP^i}{TP^i+FP^i}}{n}$

The sensitivity per community is defined as $Sensitivity = \frac{\sum_i^n \frac{TP^i}{P^i}}{n}$

Datasets

Data used for the evaluation of CellScanner settings

Four datasets, summarised in Table 2, were used to evaluate CellScanner settings. Each of these data sets contains blank data to perform machine gating. Dataset 1 (33) includes three microbial species from soil samples combined into three in silico communities. Dataset 2 (34) contains monoculture data collected with three different flow cytometers. The first group has seven species in 21 combinations and two blank files consistently used as blank references. The second group contains three monocultures, including *Escherichia coli* labelled with mCherry, of which three combinations were assessed. One file only was used as blank for every prediction. The last group contains three monocultures and two blank files. In total, 30 in silico combinations were obtained.

Data used for tool comparison

In silico communities

Six datasets, including one divided into three groups, composed of 111 monocultures and summarised in Table 3, were obtained from the FlowRepository (34) database and the literature to create 1045 in silico two-species cocultures. Dataset 1 (36) contains three microbial species, each divided into four time-point subgroups, from which 66 combinations were created. Dataset 3 (38) includes two microbial species from drinking water samples in six different conditions, including different time points and staining methods. From these, 66 in silico bi-cultures were created. Dataset 4 (39) contains eight human cell line monocultures, for which two were divided into three sub-groups according to the staining conditions and

channels described in their data file. Sub-groups were analysed with the same channels, ignoring the extra channels. From these, 68 in silico bi-cultures were created. Dataset 5 (30) contains 20 bacterial monocultures in two replicates from which 190 in silico communities were created. Dataset 6 (40) includes the same monocultures as Dataset 5 but analysed with a different flow cytometer. Dataset 7 (31) contains the first 31 monocultures from the *filtered_standards_32.mat* file, including beads, eukaryotic and prokaryotic species from water samples, already gated, and the logarithm (log 10) taken.

In vitro communities

Two datasets with known community composition were selected. Their composition is summarised in Table 4. Dataset 5 (30) contains three communities of two bacterial species in 13 different ratios. Dataset 7 includes the fourth community composed of three bacterial species in four ratios.

Study case

In-house dataset 8 comprises freshwater bacterial species *Brevundimonas sp*. (124Z) and *Variovorax sp*. (1315Z) grown in monoculture and bi-culture in R2 broth (R2B) medium at 20°C. Species were first grown on R2 agar separately and then transferred to R2B medium and incubated for 48h. Cell concentration in pre-cultures was estimated with flow cytometry and cells were diluted to ca. 10^4 cells/mL for inoculation in 1L Schott bottles. Species were equally mixed for the co-culture. Samples were taken at nine time points for monocultures and eight for bi-cultures in three biological replicates (see Table 5).

Evaluation

CellScanner settings

Datasets 1 and 2 were used to explore CellScanner parameters. By default, *machine gating* was performed before selecting a thousand events per species, using the *Tool Analysis*

function. All files used to create the in-silico community differed if used for the training. CellScanner was run with unknowns enabled, and neural networks were selected as the supervised classification method.

Tool comparison

All three machine learning methods from CellScanner and random guessing (for in silico communities only), ScriptP (30) with Linear Discriminant Analysis (LDA) and random forest, as well as CellCognize (29) with neural network, were applied to the same datasets to predict in silico and in vitro communities. The two latter scripts were modified to be comparable with CellScanner as follows: Script P was modified to handle two or three species and not to take into account expected statistical values. CellCognize was modified to create classifiers from two or three species and to accept diverse channel names. For each prediction, all methods build two or three species classifiers that were trained on 5000 events per species. The training sets were selected randomly for each method from the same monoculture files and thus overlap but are not composed of the same events. None of the datasets underwent any transformation before being fed to the classification methods.

In silico communities

Each of 1045 in silico bi-cultures obtained from datasets 1 and 3-7 was created with 1000 events per species, with a species ratio of 50:50. The in-silico communities were created from the same files for each tool. Still, each event was randomly selected using the python *random.randint* function. The files used for the in-silico community are technical replicates from the data file used for the training when available or monoculture from a close time point when time series were available (e.g., Table 3, dataset 1). For dataset 7 only, the same files were used for the training and the in-silico community, which may have resulted in overlapping data.

For each tool, the number of events predicted for species A and B was converted into ratios.

Unknown and gated events (labelled 'blank') were removed prior to calculation. Euclidean distances between the expected and predicted ratios were calculated for each method. A two-sided Wilcoxon paired test was carried out on the squared distances between the expected and predicted ratios of one of the two species using R version 3.5.3.

In vitro communities

The in vitro mock communities are represented in datasets 5 and 7. In dataset 7, *Escherichia coli* and *Pseudomonas veronii* were divided into two subpopulations, which were then classified into five groups with *Acinobacter johnsonii*. To calculate the ratios, the number of events was summed by species. Ratios were calculated for two or three species, and the Euclidean distances were calculated for each community. For the bi-cultures, only the squared difference of one species from the expected ratio was used in the Wilcoxon test, and each species was considered for the three-species community.

Study case

CellScanner was run with its three machine learning methods, with the unknown threshold set to >70% and machine learning gating. The program was trained on dataset 8 with 2000 events for each species, which were randomly selected from the pooled nine monoculture files. In fact, each of the ten runs used nine different time points per species as references, then simultaneously carried out the prediction for all communities from 0 h to 72 h.

Data availability

CellScanner is available at http://msysbiology.com/cellscanner.html and cloneable from GitHub at https://github.com/Clem-Jos/CellScanner.

The modified ScriptP and the modified CellCognize script, complete details on the datasets used, and result tables and all p-values are available at: https://github.com/Clem-Jos/CellScanner/tree/main/tool_comparison.

Data sets 2 and 8 are available on flowrepository.org upon acceptance of this manuscript. A temporary link for reviewers can be found below. To open the link, please paste it into a browser.

Dataset 2_FR-FCM-Z3TX:

https://flowrepository.org/id/RvFrQbe9nsqqyvGCZfsOjrPv8ksKSZWkUSWvHPZ5BdVxz64x5YsqUKRBO3LMmYXX

FR-FCM-Z4RP:

https://flowrepository.org/id/RvFrAhvGtPhjjmy5ryMYG6hr14KuzZ2b4p2aovfft58FIVNz8QGxNk0oI5GLlThd

FR-FCM-Z3TM:

https://flowrepository.org/id/RvFrzKykLCrofnoeg0fomWCf69zX2A6ntRFzRDX8qE9zyKY1CKLzPD33tzxekhjI

FR-FCM-Z3U2:

https://flowrepository.org/id/RvFreaUevyBIyxS8k1pW6elbE4YfQMUj6SqCwbM3uXEoKksJiZOLjuucpkKjUkfs

Dataset 8_FR-FCM-Z4RJ:

https://flowrepository.org/id/RvFrCGjzTqG8GX7q7eEqdyWfa7p3Bw8Nesi3orY756UNvudaPB9sdmCfSYv6Ia0B

References

1. Kennedy NA, Walker AW, Berry SH, Duncan SH, Farquarson FM, Louis P, Thomson JM, Other members not named within the manuscript author list (alphabetical by surname):, Satsangi J, Flint HJ, Parkhill J, Lees CW, Hold GL. 2014. The Impact of Different DNA Extraction Kits and Laboratories upon the Assessment of Human Gut Microbiota Composition by 16S rRNA Gene Sequencing. PLoS ONE 9:e88982.

2. Brooks JP, Edwards DJ, Harwich MD, Rivera MC, Fettweis JM, Serrano MG, Reris RA, Sheth NU, Huang B, Girerd P, Strauss JF, Jefferson KK, Buck GA. 2015. The truth about metagenomics: quantifying and counteracting bias in 16S rRNA studies. BMC Microbiol 15:66.

3. Bender JM, Li F, Adisetiyo H, Lee D, Zabih S, Hung L, Wilkinson TA, Pannaraj PS, She RC, Bard JD, Tobin NH, Aldrovandi GM. 2018. Quantification of variation and the impact of biomass in targeted 16S rRNA gene sequencing studies. Microbiome 6:155.

4. McLaren MR, Willis AD, Callahan BJ. 2019. Consistent and correctable bias in metagenomic sequencing experiments. eLife 8:e46923.

5. Větrovský T, Baldrian P. 2013. The Variability of the 16S rRNA Gene in Bacterial Genomes and Its Consequences for Bacterial Community Analyses. PLoS ONE 8:e57923.

6. Louca S, Doebeli M, Parfrey LW. 2018. Correcting for 16S rRNA gene copy numbers in microbiome surveys remains an unsolved problem. Microbiome 6:41.

7. Props R, Kerckhof F-M, Rubbens P, De Vrieze J, Hernandez Sanabria E, Waegeman W, Monsieurs P, Hammes F, Boon N. 2017. Absolute quantification of microbial taxon abundances. ISME J 11:584–587.

8. Tettamanti Boshier FA, Srinivasan S, Lopez A, Hoffman NG, Proll S, Fredricks DN, Schiffer JT. 2020. Complementing 16S rRNA Gene Amplicon Sequencing with Total Bacterial Load To Infer Absolute Species Concentrations in the Vaginal Microbiome. mSystems 5.

9. Zemb O, Achard CS, Hamelin J, De Almeida M, Gabinaud B, Cauquil L, Verschuren LMG, Godon J. 2020. Absolute quantitation of microbes using 16S rRNA gene metabarcoding: A rapid normalization of relative abundances by quantitative PCR targeting a 16S rRNA gene spike-in standard. MicrobiologyOpen 9.

10. Kim JY, Yi M, Kim M, Lee S, Moon HS, Yong D, Yong T. 2021. Measuring the absolute abundance of the microbiome by adding yeast containing 16S rRNA gene from a hyperthermophile. MicrobiologyOpen 10.

11. 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.

12. De Roy K, Clement L, Thas O, Wang Y, Boon N. 2012. Flow cytometry for fast microbial community fingerprinting. Water Research 46:907–919.

13. Picot J, Guerin CL, Le Van Kim C, Boulanger CM. 2012. Flow cytometry: retrospective, fundamentals and recent instrumentation. Cytotechnology 64:109–130.

14. Cardoso CC, Santos-Silva MC. 2019. Eight-color panel for immune phenotype monitoring by flow cytometry. Journal of Immunological Methods 468:40–48.

15. Lucchesi S, Furini S, Medaglini D, Ciabattini A. 2020. From Bivariate to Multivariate Analysis of Cytometric Data: Overview of Computational Methods and Their Application in Vaccination Studies. Vaccines 8:138.

16. Cheung M, Campbell JJ, Whitby L, Thomas RJ, Braybrook J, Petzing J. 2021. Current trends in flow cytometry automated data analysis software. Cytometry 99:1007–1021.

17. García F, López-Urrutia Á, Morán X. 2014. Automated clustering of heterotrophic bacterioplankton in flow cytometry data. Aquat Microb Ecol 72:175–185.

18. Sgier L, Freimann R, Zupanic A, Kroll A. 2016. Flow cytometry combined with viSNE for the analysis of microbial biofilms and detection of microplastics. Nat Commun 7:11587.

19. Coggins LX, Larma I, Hinchliffe A, Props R, Ghadouani A. 2020. Flow cytometry for rapid characterisation of microbial community dynamics in waste stabilisation ponds. Water Research 169:115243.

20. Rubbens P, Props R. 2021. Computational Analysis of Microbial Flow Cytometry Data. mSystems 6:e00895-20.

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

22. Ye X, Ho JWK. 2019. Ultrafast clustering of single-cell flow cytometry data using FlowGrid. BMC Syst Biol 13:35.

23. 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.

24. Ross D. 2021. Automated analysis of bacterial flow cytometry data with FlowGateNIST. PLoS ONE 16:e0250753.

25. Wanderley BMS, A. Araújo DS, Quiroga MV, Amado AM, Neto ADD, Sarmento H, Metz SD, Unrein F. 2019. flowDiv: a new pipeline for analyzing flow cytometric diversity. BMC Bioinformatics 20:274.

26. Frankel DS, Olson RJ, Frankel SL, Chisholm SW. 1989. Use of a neural net computer system for analysis of flow cytometric data of phytoplankton populations. Cytometry 10:540–550.

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

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

29. Pereira GC, Ebecken NFF. 2011. Combining in situ flow cytometry and artificial neural networks for aquatic systems monitoring. Expert Systems with Applications 38:9626–9632.

30. Rubbens P, Props R, Boon N, Waegeman W. 2017. Flow Cytometric Single-Cell Identification of Populations in Synthetic Bacterial Communities. PLoS ONE 12:e0169754.

31. Özel Duygan BD, Hadadi N, Babu AF, Seyfried M, van der Meer JR. 2020. Rapid detection of microbiota cell type diversity using machine-learned classification of flow cytometry data. Commun Biol 3:379.

32. Finak G, Frelinger J, Jiang W, Newell EW, Ramey J, Davis MM, Kalams SA, De Rosa SC, Gottardo R. 2014. OpenCyto: An Open Source Infrastructure for Scalable, Robust, Reproducible, and Automated, End-to-End Flow Cytometry Data Analysis. PLoS Comput Biol 10:e1003806.

33. Heyse J, Buysschaert B, Props R, Rubbens P, Skirtach AG, Waegeman W, Boon N. 2019. Coculturing Bacteria Leads to Reduced Phenotypic Heterogeneities. Applied and Environmental Microbiology 85:13.

34. van de Velde CC, Joseph C, Biclot A, Huys GRB, Pinheiro VB, Bernaerts K, Raes J, Faust K. 2022. Fast quantification of gut bacterial species in cocultures using flow cytometry and supervised classification. ISME COMMUN 2:40.

35. Vandeputte D, Kathagen G, D'hoe K, Vieira-Silva S, Valles-Colomer M, Sabino J, Wang J, Tito RY, De Commer L, Darzi Y, Vermeire S, Falony G, Raes J. 2017. Quantitative microbiome profiling links gut community variation to microbial load. Nature 551:507–511.

36. Fodelianakis S, Lorz A, Valenzuela-Cuevas A, Barozzi A, Booth JM, Daffonchio D. 2019. Dispersal homogenizes communities via immigration even at low rates in a simplified synthetic bacterial metacommunity. Nat Commun 10:1314.

37. Spidlen J, Breuer K, Rosenberg C, Kotecha N, Brinkman RR. 2012. FlowRepository: A resource of annotated flow cytometry datasets associated with peer-reviewed publications. Cytometry 81A:727–731.

38. Heyse J, Buysschaert B, Props R, Rubbens P, Skirtach AG, Waegeman W, Boon N. 2019. Coculturing Bacteria Leads to Reduced Phenotypic Heterogeneities. Applied and Environmental Microbiology 85:13.

39. Chung S, Kim S-H, Seo Y, Kim S-K, Lee JY. 2017. Quantitative analysis of cell proliferation by a dye dilution assay: Application to cell lines and cocultures: Simultaneous Proliferation Analysis of Cocultured Cells. Cytometry 91:704–712.

40. Rubbens P, Props R, Garcia-Timermans C, Boon N, Waegeman W. 2017. Stripping flow cytometry: How many detectors do we need for bacterial identification?: Stripping Flow Cytometry for Bacterial Identification. Cytometry 91:1184–1191.

41. FlowJo™ Software, Ashland, OR: Becton, Dickinson and Company; 2021.

Tables

Table 1: Overview of selected classification tools for microbial FC data. *LR (logistic regression), RF (random forest), NN (neural network), AC (agglomerative clustering), MGM (Multivariate Gaussian Mixture), LDA (linear discriminant analysis). CellScanner does not have prerequisites on Windows because it uses an installer that takes care of the dependencies.*

	CellScanner	*CellCognize*	*flowEMMi*	*ScriptP*	*FlowGateNIST*
Supervised classification	(LR,RF, NN)	(NN)		(FRD, LDA)	
Unsupervised classification	(AC)		(MGM)		(MGM)
User-friendly interface	X				
Prerequisite	None (Windows) Python(Mac, Linux)	Matlab	R	Python	Python
Gating integrated	X		By FlowJo (41)		X

Table 2: Summary of datasets used to assemble in silico communities for CellScanner setting evaluation. *For each prediction, one monoculture file was used for the training as a reference and another to build the in-silico community. Blank files were used as references for gating for each prediction in a group.*

DATASET			SPECIES	AS REFERENCE	AS PREDICTION	FC	CHANNELS
	Link	GROUP		CONDITION	CONDITION		
DATASET 1	FR-FCM-ZYG6	2	E111 B41 E310 BLANK	37° t4 37° t4 37° t4 LB medium H2O	37° t5 37° t5 37° t5	BD ACCURI C6	FSC-A,SSC-A,FL1-A,FL2-A,FL3-A,FL4-A,FSC-H,SSC-H,FL1-H,FL2-H,FL3-H,FL4-H,Width
DATASET 2	FR-FCM-Z3U2 + FR-FCM-Z4RP + FR-FCM-Z3TX	1	*Bacteroides thetaiotaomicron* *Bacteroides uniformis* *Blautia hydrogenotrophica* *Collinsella aerofaciens* *Escherichia coli* *Prevotella copri* *Roseburia intestinalis* BLANK	mGAM medium Anaérobic condition for 24h. All cells are in stationary phase	Techical replicates of references	BD ACCURI C6	FSC-A, FSC-H, SSC-A, SSC-H, FL1-A, FL1-H, FL2-A, FL2-H, FL3-A,FL3-H, FL4-A, FL4-H, Width
	FR-FCM-Z3MT	2	*Roseburia intestinalis* *Faecalibacterium prausnitzii* *Escherichia coli with mCherry* BLANK	mGam medium for 24h in time anaerobic condition PBS medium RMC medium	Techical replicates of references	CytoFLEX	FSC-H,FSC-A,SSC-H,SSC-A,FL1-H,FL1-A,FL4-H,FL4-A,FL3-red-H,FL3-A,APC-A750-H,PI-H,PI-A,FSC-Width
	FR-FCM-Z4RP	3	*Bacteroides thetaiotaomicron* *Bacteroides uniformis* *Blautia hydrogenotrophica* BLANK	mGam medium for 24h in time anaerobic condition mGAM medium PBS medium	Same files as reference	CytoFLEX	FSC-H,FSC-A,SSC-H,SSC-A,FITC-H,FL1-A,PerCP-H,FL3-A,APC-H,APC-A,APC-A700-H, APC-A700-A,APC-A750-H,APC-A750-A,PB450-H,PB450-A,KO525-H,KO525-A,Violet610-H,Violet610-A, Violet660-H,Violet660-A,PE-H,FL2-A,ECD-H,ECD-A,PC5.5-H,PC5.5-A,PC7-H,PC7-A,FSC-Width

Table 3: Summary of data sets used to assemble in silico communities for tool comparison. *For each prediction, one monoculture file was used for the training as a reference and another as part of the in-silico community, unless for the dataset seven for which only one file was available per species. Files used in prediction can differ from reference files by time point or technical replicate. A species subgroup is defined by flow cytometer parameters or by staining methods for species sub-populations. X2 and X3 indicate the number of sub-populations considered in the analysis for the CellCognize dataset.*

DATASET			SPECIES	AS REFERENCE	AS PREDICTION	FC	CHANNELS
Link		group		CONDITION	CONDITION		
DATASET 1	FR-FCM-ZYG6	1	E111-1	37° t1	37° t2	BD ACCURI C6	FSC-A,SSC-A,FL1-A,FL2-A,FL3-A,FL4-A,FSC-H,SSC-H,FL1-H,FL2-H,FL3-H,FL4-H,Width
			E111-2	37° t3	37° t4		
			E111-3	37° t5	37° t6		
			E111-4	37° t7	37° t8		
			B41-1	37° t1	37° t2		
			B41-2	37° t3	37° t4		
			B41-3	37° t5	37° t6		
			B41-4	37° t7	37° t8		
			E310-1	37° t1	37° t2		
			E310-2	37° t3	37° t4		
			E310-3	37° t5	37° t6		
			E310-4	37° t7	37° t8		
DATASET 3	FR-FCM-ZYWN	1	*Enterobacter sp* . T0 SG	SG t0	Technical replicates	FACSVerse	FSC-A,FSC-W,FSC-H,SSC-A,SSC-W,SSC-H,FITC-A,FITC-W,FITC-H,PerCP-Cy5.5-A,PerCP-Cy5.5-W,PerCP-Cy5.5-H,APC-A,APC-W,APC-H,AmCyan-A,AmCyan-W,AmCyan-H,APC-Cy7-A,APC-Cy7-W,APC-Cy7-H,dsRed-A,dsRed-W,dsRed-H,eCFP-A,eCFP-W,eCFP-H,PE-Cy7-A,PE-Cy7-W,PE-Cy7-H
			Enterobacter sp. T0 SGPI	SGPI t0			
			Pseudomonas sp . T0 SG	SG t0			
			Pseudomonas sp . T0 SGPI	SGPI t0			
			Enterobacter sp. T1 SG	SG t1			
			Enterobacter sp. T1 SGPI	SGPI t1			
			Pseudomonas sp. T1 SG	SG t1			
			Pseudomonas sp. T1 SGPI	SGPI t1			
			Enterobacter sp. T2 SG	SG t2			
			Enterobacter sp. T2 SGPI	SGPI t2			
			Pseudomonas sp. T2 SG	SG t2			
			Pseudomonas sp. T2 SGPI	SGPI t2			
DATASET 4	FR-FCM-ZZUZ	1	Jurkat-1	/ day 2	/ day 3	FACSVerse	FSC-A,FSC-H,FSC-W,SSC-A,SSC-H,SSC-W,FITC-A (*)
			THP1-1	/ day 2	/ day 3		
			Jurkat-2	/ day 2	/ day 3		
			THP1-2	/ day 2	/ day 3		
			Jurkat-V	Violet day 2	Violet day 3		
			THP1-R	FarRed day 2	FarRed day 3		
			HEK293	day 2	day 3		
			HuH7	day 2	day 3		
			Nb2	Violet day 2	Violet day 3		
			T98G	CFSE day 2	Technical replicate		
			U2OS	day 2	day 3		
			U937	CFSE day 2	Technical replicate		
		2	Jurkat	day 2	Day 3		*+PE-A,PerCP-Cy5.5-A,PE-Cy7-A,APC-A,APC-Cy7-A
			THP1	day 2	Day 3		
		3	Jurkat	Violet day 2	Violet day 3		*+PE-A,APC-A,V450-A
			THP1	FarRed day 2	FarRed day 3		
DATASET 5	FR-FCM-ZZSH	1	*Agribacter rhizogenes, Bacillus cubtilis, Burkholderia ambfaria, Citrobacter freundii, Cupriavidus necator, Cupriavidus pinatubonensis, Edwarsialla ictaluri, Enterobacter aerogenes, Escherichia coli, Janthinobacterium sp, Klebsiella oxytoca, Lactobacillus plantarium, micrococcus luteus, Pseudomonas fluorencens, Pseudomonas putida, Rhizobium radiobacter, Shewanella oneidensis, Sphingomonas aromaticivorans, Streptococcus salivarius, Zymonas mobilis*	A monoculture per species	Techical replicates	BD ACCURI C6	FSC-A,SSC-A,FL1-A,FL2-A,FL3-A,FL4-A,FSC-H,SSC-H,FL1-H,FL2-H,FL3-H,FL4-H,Width
DATASET 6	FR-FCM-ZY6M	1	*Agribacter rhizogenes, Bacillus cubtilis, Burkholderia ambfaria, Citrobacter freundii, Cupriavidus necator, Cupriavidus pinatubonensis, Edwarsialla ictaluri, Enterobacter aerogenes, Escherichia coli, Janthinobacterium sp, Klebsiella oxytoca, Lactobacillus plantarium, micrococcus luteus, Pseudomonas fluorencens, Pseudomonas putida, Rhizobium radiobacter, Shewanella oneidensis, Sphingomonas aromaticivorans, Streptococcus salivarius, Zymonas mobilis*	A monoculture per species	Techical replicates	FACSVerse	FSC-A,FSC-W,FSC-H,SSC-A,SSC-W,SSC-H,FITC-A,FITC-W,FITC-H,PE-A,PE-W,PE-H,PerCP-Cy5.5-A,PerCP-Cy5.5-W,PerCP-Cy5.5-H,PE-Cy7-A,PE-Cy7-W,PE-Cy7-H,APC-A,APC-W,APC-H,APC-Cy7-A,APC-Cy7-W,APC-Cy7-H,V450-A,V450-W,V450-H,V500-A,V500-W,V500-H
DATASET 7	CellScanner	1	Beads X8(sizes from 0,2 to 15µm), Acinobacter johnsonii X2, Acinobacter tjernbergiae X2, Arthrobacter chlorophenolicus X3, Bacillus subtilis X2, Caulobacter crescentus X2,Cryptococcus albidus X2, Escherichia coli MG1655 X3, Escherichia coli DH5α-λpir, Lactococcus lactis, Pseudomonas knackmussii, Pseudomonas migulae,Pseudomonas putida, Pseudomonas veronii X2, Sphingomonas wittichii	A monoculture per species	Same data file than references	NovoCyte	FSC-H,SSC-H,FITC-H,FSC-A,SSC-A,FITC-A, Width

Table 4: Summary of data sets used as in vitro communities for tool comparison. *For each prediction, monoculture files were used for the training as reference, and the in vitro community files were used for predictions.*

DATASET Link	group	SPECIES	AS REFERENCES	AS PREDICTION Ratio		FILENAME	FC	CHANNELS
DATASET 5 FR-FCM-ZZSG	1	*Pseudomonas fluorescens*	3 technical replicates	1%99% 5%95% 10%90% 20%80% 30%70% 40%60% 50%50%	60%40% 70%30% 80%20% 90%10% 95%5% 99%1	3 replicates for each ratio	BD ACCURI C6	FSC-A,SSC-A,FL1-A,FL2-A,FL3-A,FL4-A,FSC-H,SSC-H,FL1-H,FL2-H,FL3-H,FL4-H,Width
		Pseudomonas putida	3 technical replicates					
	2	*Agrobacter rhizogenes*	4 technical replicates	1%99% 5%95% 10%90% 20%80% 30%70% 40%60% 50%50%	60%40% 70%30% 80%20% 90%10% 95%5% 99%1	3 replicates for each ratio		
		Janthinobacterium sp	4 technical replicates					
	3	*Micrococcus luteus*	3 technical replicates	1%99% 5%95% 10%90% 20%80% 30%70% 40%60% 50%50%	60%40% 70%30% 80%20% 90%10% 95%5% 99%1	3 replicates for each ratio		
		Shewanella oneidensis	3 technical replicates					
DATASET 7 CellCognize	1	Acinobacter johnsonii (AJH1)	1_community.csv	40%39%21%		5 replicates for each ratio	NovoCyte	FSC-H,SSC-H,FITC-H,FSC-A,SSC-A,FITC-A,Width
		Escherichia coli 1 (ECL1)	2_community.csv	67%22%11%				
		Escherichia coli 2 (ECL2)	3_community.csv	22%66%12%				
		Pseudomonas veronii 1 (PVR1)	4_community.csv	28%28%44%				
		Pseudomonas veronii 2 (PVR2)	5_community.csv					

Table 5: Summary of the datasets used as a study case to predict the composition of a two-species coculture. *For each prediction, monoculture files were used for the training as reference, and the in vitro community files were used for predictions. All blank files were used for the gating.*

DATASET Link	AS REFERENCES SPECIES	AS REFERENCES CONDITION		COCULTURE FILES CONDITION			FC	CHANNELS
DATASET 8 FR-FCM-Z4RP	*Brevundimonas sp. (124Z)*	0h 16,5h 18h 22,5h 25,5h 30h 40,5h 48h 72h	R2 medium, at 20 °C	0h 18h 22h 26h 30h 41h 48h 72h	x 3 technical replicates	R2 medium, at 20 °C	BD ACCURI C6	FSC-A,SSC-A,FL1-A,FL2-A,FL3-A,FL4-A,FSC-H,SSC-H,FL1-H,FL2-H,FL3-H,FL4-H,Width
	Variovorax sp . (1315Z)	0h 16,5h 18h 22,5h 25,5h 30h 40,5h 48h 72h						
	Blank	10 biological replicates x 3(or 2) technical replicates						

Figures

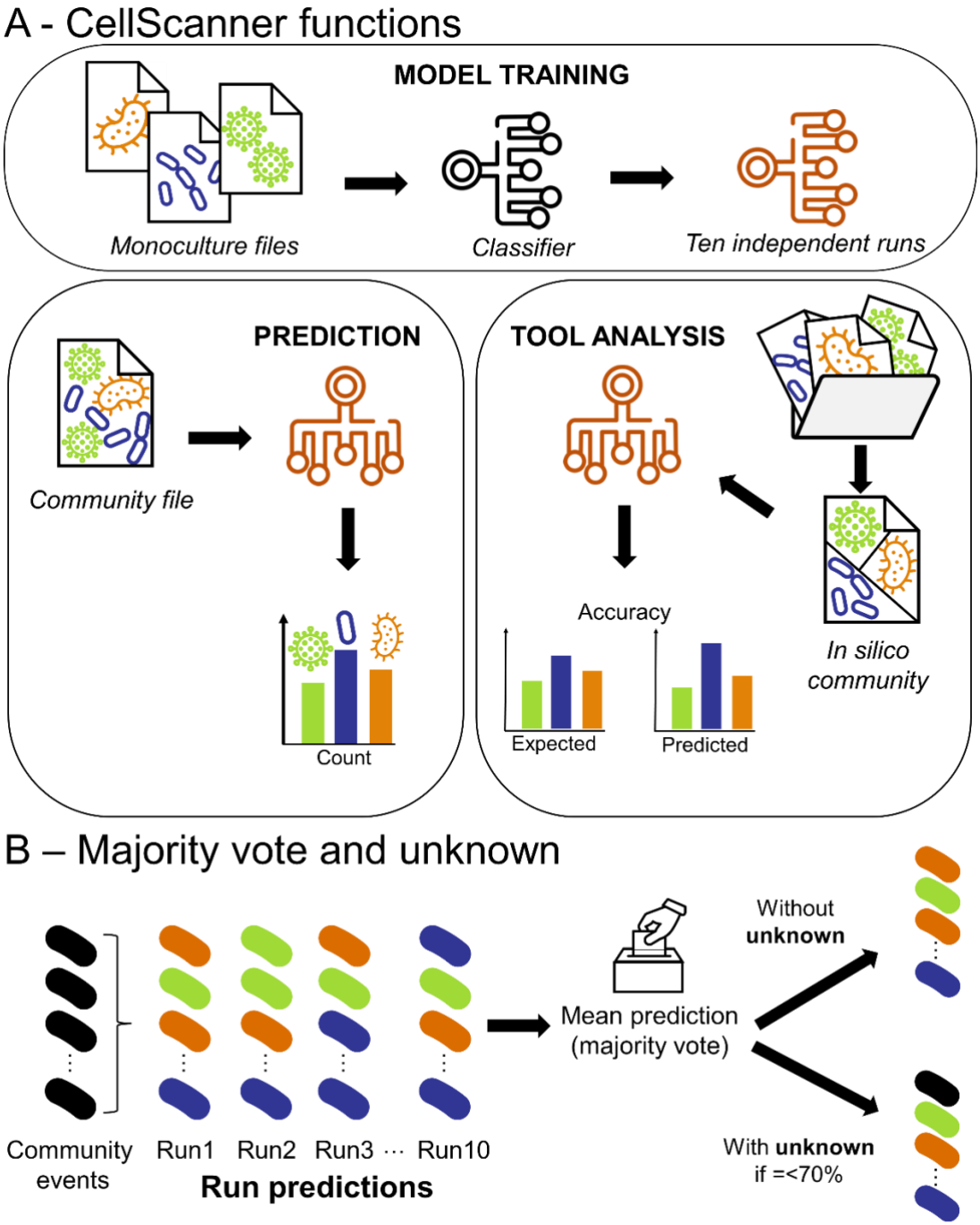


Figure 1: Overview of CellScanner. *A: CellScanner performs ten runs by default. In each run, events are randomly selected from monoculture flow cytometry (FC) files and used to train a classifier. The classifier is then applied to predict the composition of an in-vitro (prediction) or in silico (tool analysis) community FC file. B: During each run, every event in the community is classified as a species. The final species assignment is decided by majority vote across runs. In addition, CellScanner allows to label events as unknown when fewer than the specified percentage of runs agree on the species (unknown parameter).*

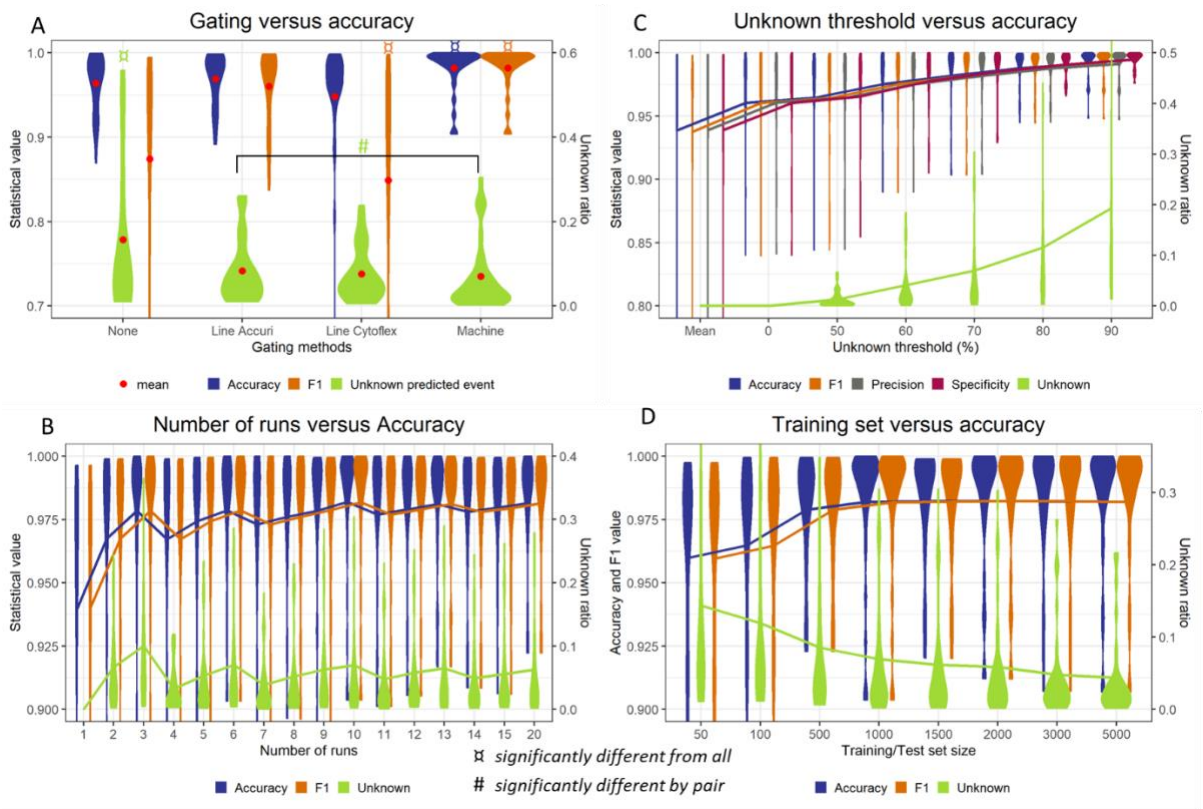


Figure 2: Accuracy of prediction is dependent on selected parameters. *Predictions were performed with different parameters for 30 species pairs combined in silico. A: Distribution of prediction accuracy for each community with three different methods of gating and no gating. Significance of differences was assessed with a two-sided Wilcoxon paired test (p-value < 0.05 shown). B: Prediction accuracy with the number of classification runs from one to 20. Wilcoxon paired test p-values are provided in CellScanner's GitHub directory. C: Prediction accuracy, specificity and precision with the unknown threshold parameter ranging from 0 to 90% for ten runs. In 'Mean', the F1, precision, specificity and accuracy values are computed as the mean over the ten classification runs without majority vote. D: Prediction accuracy obtained with a training set from 50 to 5000 events per species. For C and D, all data points are significantly different from preceding points (p-value < 0.05 for two-sided Wilcoxon paired test) unless indicated otherwise*. A-D: Distributions are depicted as violin plots.

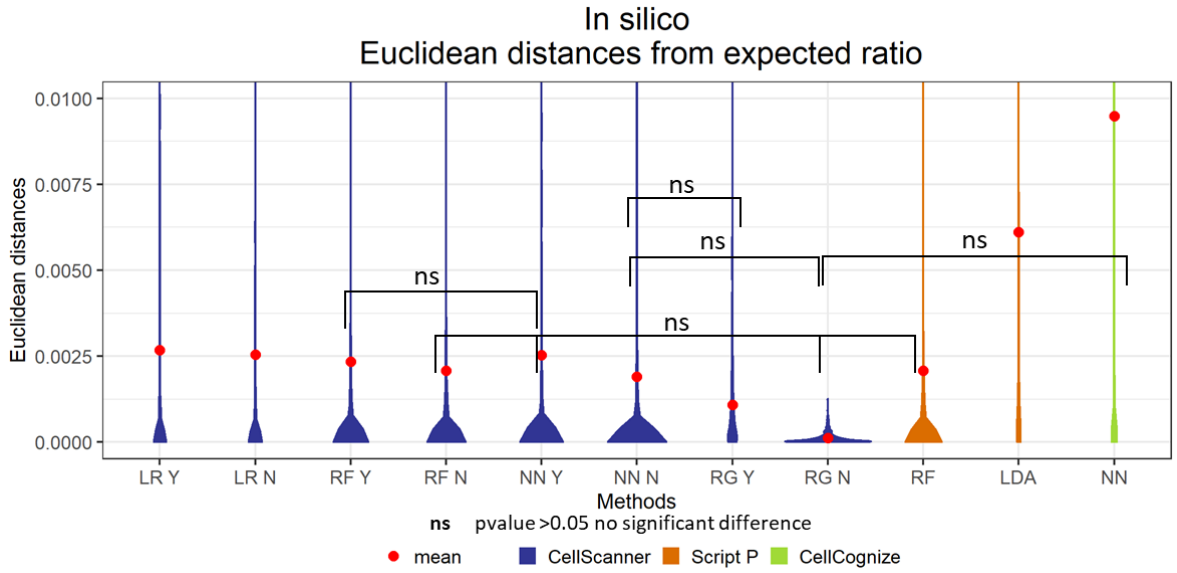
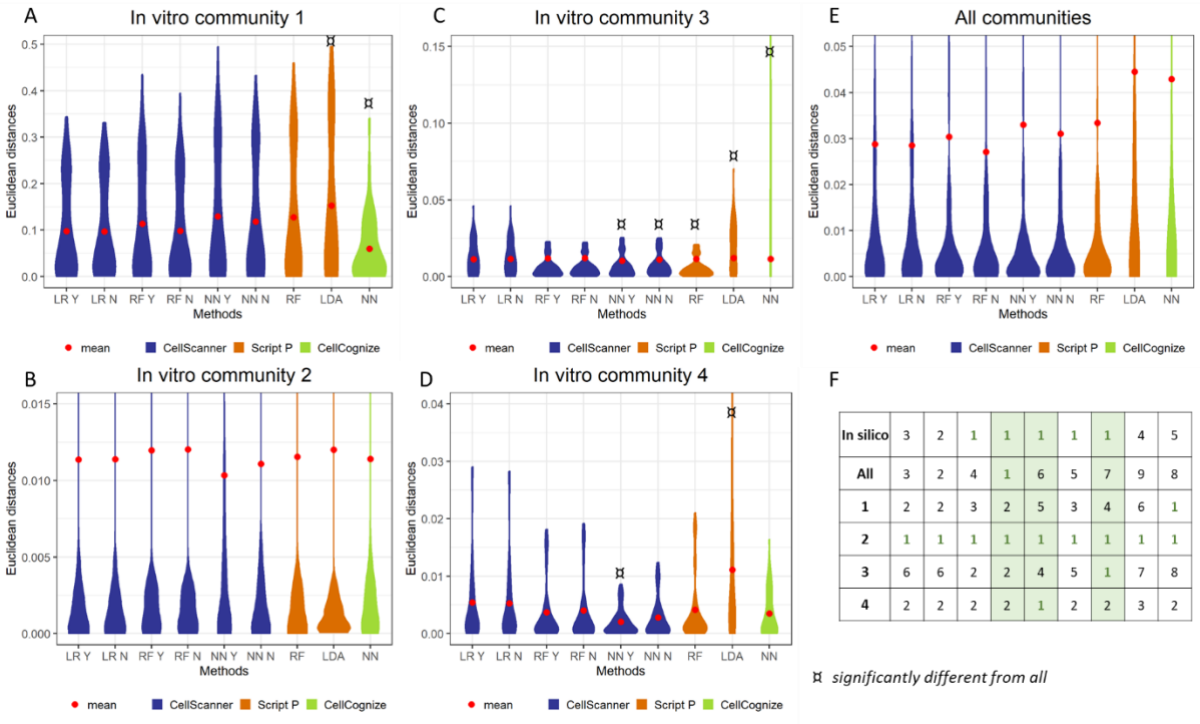


Figure 3: Comparison of classification methods in silico including CellScanner (CS) Script P (SP) and CellCognize.

The classification methods include logistic regression (LR), random forest (RF), neural network (NN) and linear discriminant analysis (LDA), with (Y) or without (N) enabling unknowns. CellScanner NN Y or Y in blue uses a different implementation than CellCognize NN in green. Euclidean distances between the expected (50:50) and predicted ratios are shown for 1045 in silico species pairs. P-values across methods per community are calculated with a two-sided Wilcoxon paired test and provided in CellScanner's GitHub directory. All values that are not annotated are significant with a p-value <0,05. Distributions are depicted as violin plots.

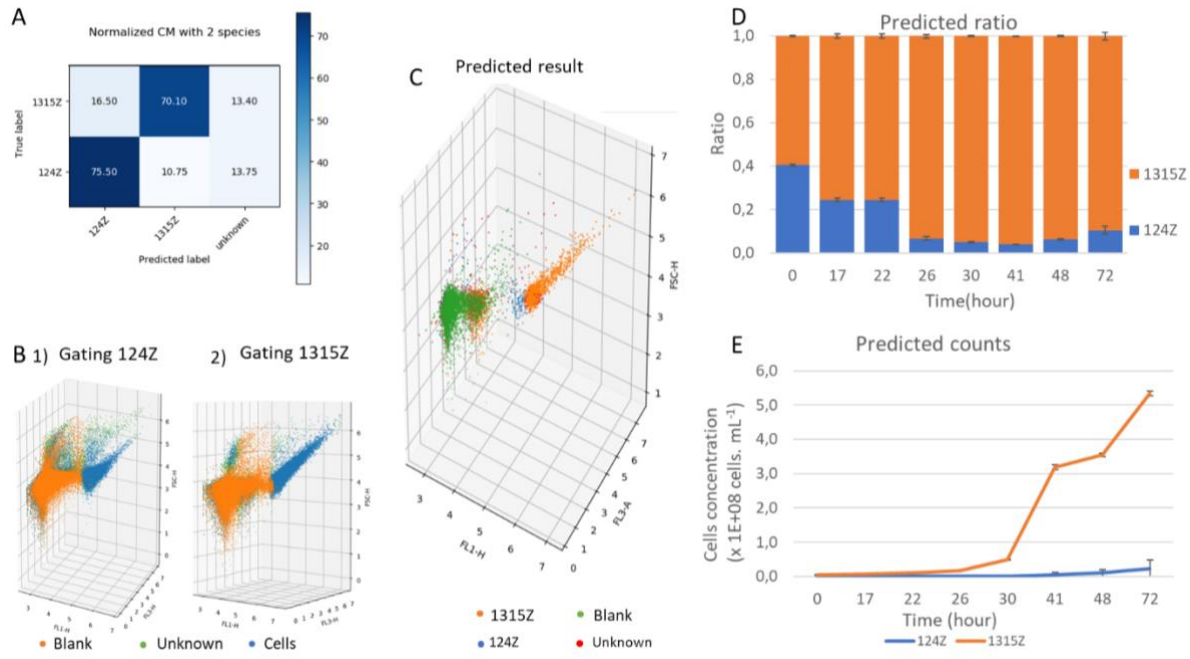


F

In silico	3	2	1	1	1	1	1	4	5
All	3	2	4	1	6	5	7	9	8
1	2	2	3	2	5	3	4	6	1
2	1	1	1	1	1	1	1	1	1
3	6	6	2	2	4	5	1	7	8
4	2	2	2	2	1	2	2	3	2

¤ *significantly different from all*

Figure 4: Comparison of classification methods in vitro, including CellScanner (CS) Script P (SP) and CellCognize and summary rank table: *A-E: Euclidean distances between the expected community compositions in vitro and the predicted compositions for three combinations of two species in 13 different ratios (A-C), one combination of three species in four different ratios (D), and for all combinations together (E). P-values across methods per community are calculated with a two-sided Wilcoxon paired test and summarised in CellScanner's GitHub directory. F: Table ranking methods for the 1064 predictions in silico, the in vitro communities 1, 2, 3 and 4 and all together (all). Distributions are depicted as violin plots.*



599

600 **Figure 5: In silico analysis and prediction for in vitro samples in a time series with unknown ratios of the bi-culture of**
601 **124Z and 1315Z.** *A: Confusion matrix for the prediction step of the in-silico community (CellScanner output). B: Gating*
602 *performed on reference files during the "Tool Analysis" of the in-silico community (CellScanner output) for 1) 124Z and 2)*
603 *1315Z. C: 3D plot for the prediction at time point 22h for the first replicate. D: Predicted ratio for time series after removal*
604 *of background and unknown cells. E: Predicted cell numbers for the two species.*

605

Supplementary Tables

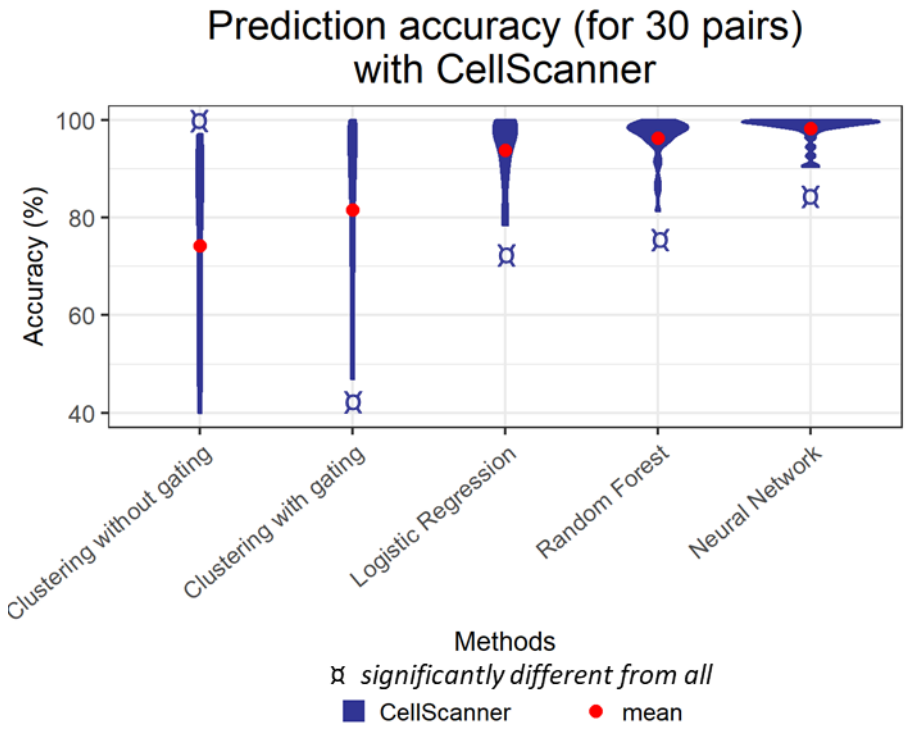
Supplementary Table 1: Accuracy and F1 score for different gating methods, grouped by cytometer. *Accuracy on top, F1 in the middle and Unknown part at the bottom for four cytometers and four gating methods. The values in bold showcase where the gating is considered adapted for the dataset, the other cases are considered too stringent (i.e. less than 100 events kept for at least one species) event though the accuracy and F1 score appears good.*

Accuracy				
	None	**Line gating for BD Accuri C6**	**Line gating for BD CytoFLEX**	**Machine gating**
CD Accuri C6(1)	0,98	0,95	0,98	1,00
CD Accuri C6(2)	0,97	**0,98**	0,94	0,98
CytoFLEX(1)	0,90	0,91	**0,95**	0,98
CytoFLEX(2)	0,94	0,99	**0,99**	0,99
F1				
CD Accuri C6(1)	0,98	0,90	0,87	1,00
CD Accuri C6(2)	0,97	**0,98**	0,81	0,98
CytoFLEX(1)	0,89	0,88	**0,95**	0,98
CytoFLEX(2)	0,94	0,99	**0,99**	0,99
Unknown				
CD Accuri C6(1)	0,10	0,09	0,03	0,01
CD Accuri C6(2)	0,08	**0,07**	0,07	0,06
CytoFLEX(1)	0,52	0,17	**0,13**	0,09
CytoFLEX(2)	0,42	0,10	**0,12**	0,14

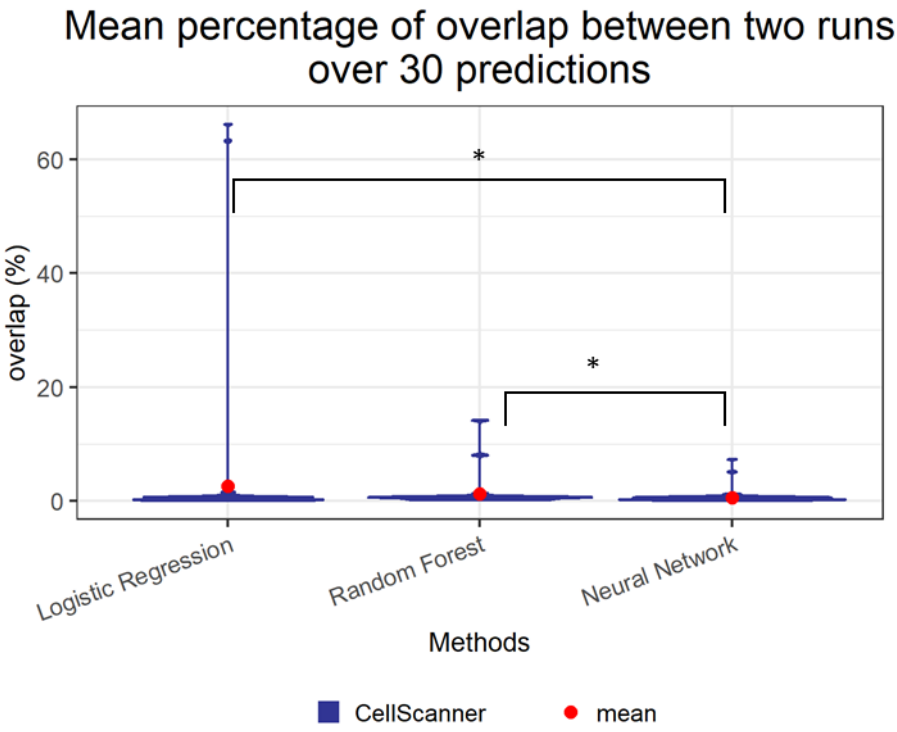
Supplementary Table 2: Prediction with a specific percentage of unknowns for all in silico and in vitro predictions. *Columns > x%: percentage or predictions with a percentage of unknown predicted events higher than x %. The mean percentage of unknown events is calculated as the mean of all percentages of unknown events for all predictions. For in vitro communities, the predictions with more than 20% of events labelled unknown are from the coculture 1 of dataset 5, which was the most difficult to predict.*

IN-SILICO COMMUNITIES							
	Percentage of predictions with a percentage of predicted unknown events over the threshold					Mean % of unknown predicted events	Standard deviation
Threshold	**>30%**	**>20%**	**>10%**	**>5%**	**>0%**		
Logistique Regression	0	0,00%	0,67%	3,73%	69,86%	12%	0,018
Random forest	0,29%	0,86%	4,40%	14,07%	67,27%	2,30%	0,042
Neural network	1,53%	5,84%	17,51%	34,07%	71,58%	5,30%	0,086
Random guessing	100%	100,00%	100,00%	100,00%	100,00%	89,10%	0,007
IN-VITRO COMMUNITIES							
	Percentage of predictions with a percentage of predicted unknown events over the threshold					Mean % of unknown predicted events	Standard deviation
Threshold	**>30%**	**>20%**	**>10%**	**>5%**	**>0%**		
Logistique Regression	0,00%	0,00%	10,95%	28,47%	99,27%	3,06%	4,20%
Random forest	0,00%	10,22%	27,01%	28,47%	100,00%	5,81%	7,59%
Neural network	10,22%	24,82%	43,07%	50,36%	100,00%	11,83%	11,26%

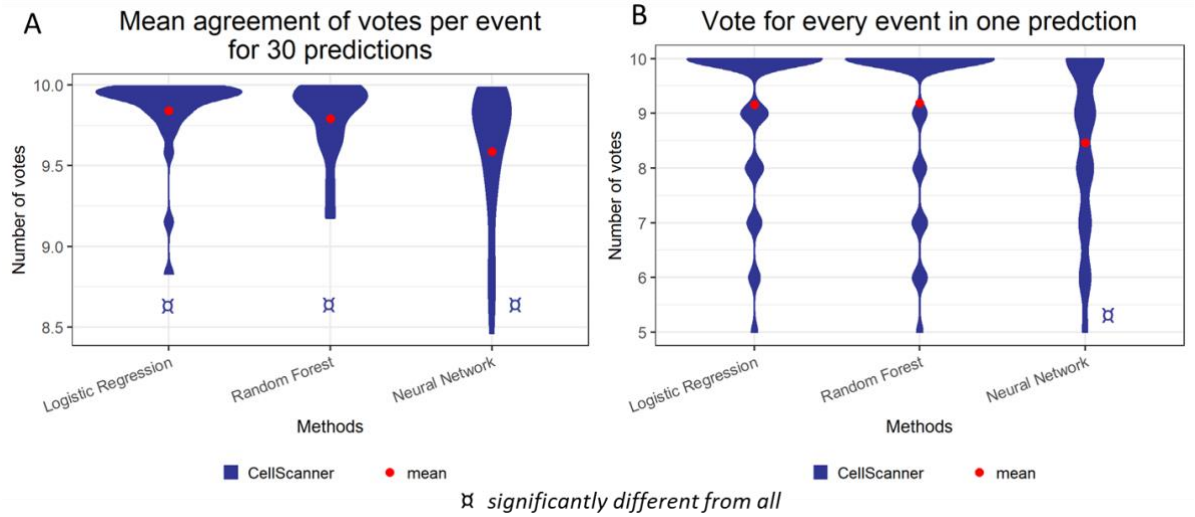
Supplementary Figures



Supplementary Figure 1: Prediction accuracy of the different classification methods of CellScanner. *Accuracy percentages are calculated for the 30 in silico coculture predictions of the CellScanner parameter analysis. The clustering method uses an agglomerative clustering function to cluster cells and then attributes each cluster to the closest species by comparing the distance of each cluster from each reference file furnished by the user. For a two-species community, CellScanner chose the number of clusters from one to three by comparing the distance between clusters in the three conditions. The tool selects the higher number of clusters with the larger distance. Three clusters can include the two expected species plus blank or contamination, and one unique cluster mimics the disappearance of one species. Here only the clustering is compared to the classification methods. If the program failed the species attribution to a cluster, the user manually labelled the cluster and modified the accuracy (only for initial accuracy under 40%). All methods give significantly different results according to the Wilcoxon test (p-value < 0.05).*



Supplementary Figure 2: Mean percentage of overlapping training data between two runs for the 30 predictions of CellScanner parameter analysis. *Dataset 1 and 2 from CellScanner parameter analysis were used to predict 30 in silico pairs where the training/testing set was extracted for comparison. Every dot is the mean of the number of events from the training set overlapping between two runs, calculated for the 45 pairs combined from the ten runs. The mean is calculated for each of the two species. * significantly different result according to Wilcoxon test (p-value < 0.05).*



640

641 **Supplementary Figure 3: Agreement of votes between runs**. *A) Mean agreement of vote per species for 30 predictions.*

642 *Each dot is the mean of the maximum number of runs agreeing on every event in a prediction. Vote extracted from prediction*

643 *of dataset 1 and 2 for CellScanner parameter setting analysis. B) Example of agreement across runs. Every dot represents*

644 *the number of runs agreeing on an event for the in-silico community of Bacteroides thetaiotaomicron and Bacteroides uniformis*

645 *of the same dataset. Identified result are significantly different according to Wilcoxon test (p-value < 0.05).*

646

647

648

649