Vol. 72: 175–185, 2014
doi: 10.3354/ame01691

AQUATIC MICROBIAL ECOLOGY
Aquat Microb Ecol

Published online May 7

Automated clustering of heterotrophic bacterioplankton in flow cytometry data

Francisca C. García*, Ángel López-Urrutia, Xosé Anxelu G. Morán

Centro Oceanográfico de Gijón, Instituto Español de Oceanografía, 33212 Gijón (Asturias), Spain

ABSTRACT: Flow cytometry has become a standard method to analyze bacterioplankton. Analysis of samples by flow cytometry is automatic, but it is followed by manual classification of the bacterioplankton groups in flow cytometry standard (FCS) files. This classification is a time consuming and subjective task performed by manually drawing the limits of the groups present in cytograms, a process referred to as gating. The automation of flow cytometry data processing based on pattern recognition techniques could provide an efficient tool to overcome some of these disadvantages. Here, we propose the use of model-based clustering techniques for the automatic detection of low (LNA) and high (HNA) nucleic acid bacterioplankton groups in FCS files. To validate our method, we compared the automatic classification with a flow cytometry database from a 9 yr time series collected in the central Cantabrian Sea that had been manually analyzed. The correlation between automatic and manual gating methods was >0.9 for cell counts and 0.7 to 0.95 for side scatter values, a proxy of cell size. In addition, no significant differences were found in the mean annual cycle of LNA and HNA cell abundance depicted by both methods. We also quantified the subjectivity of manual gating. The coefficient of variation for heterotrophic bacteria counts obtained by different analysts was around 10 to 20%. Our results suggest that the combination of flow cytometry and automatic gating provides a valuable tool to analyze microbial communities objectively and accurately, allowing us to safely compare bacterioplankton samples from different environments.

KEY WORDS: Bacteria · Flow cytometry · Automatic clustering · Aquatic sciences

Resale or republication not permitted without written consent of the publisher

INTRODUCTION

Heterotrophic prokaryotes make up the largest living biomass of aquatic ecosystems, playing a key role in the carbon cycle by incorporating dissolved organic matter and recycling nutrients in the oceans (Azam et al. 1983, Hansell & Carlson 1998). The estimation of total bacterioplankton abundance has evolved from the tedious estimation by microscopy techniques to automatic counting using flow cytometry. Flow cytometry allows counting at a rate of 200 to 2000 cells s^{-1}. Fast sample processing by flow cytometry has proven very useful for large-scale studies of plankton communities (Gasol & del Giorgio 2000). In addition to abundance estimates, interest in automated flow cytometric properties arises from the increasing use of side scatter (SSC) signals as a surrogate for cell size (Calvo-Díaz & Morán 2006, Felip et al. 2007), thus allowing for precise estimation of the biomass contributed by each subgroup.

The analysis of bacterial population abundance by flow cytometry consists of 2 separate steps. The first step, which is truly automated, is the processing of the sample under the flow cytometer. The second, not automated step, is the posterior analysis of the flow cytometer output. Flow cytometry uses the cell properties of light scattering and fluorescence, and records the information of each analyzed cell in a flow cytometry standard (FCS) file. The analysis of

*Corresponding author: paqui.garcia@gi.ieo.es

© Inter-Research 2014 · www.int-res.com

the samples by flow cytometry is automatic, but the clustering of the data stored in the FCS files is usually done manually. Two widespread groups of heterotrophic bacteria separated by their relative nucleic acid (NA) content, after appropriate NA staining, and commonly referred to as low (LNA) and high (HNA) are found in almost any aqueous sample (Li et al. 1995, Gasol et al. 1999, Bouvier et al. 2007), from freshwater, to estuarine, to open-ocean waters. After some debate, a consensus is emerging that LNA and HNA subgroups comprise fundamentally different phylotypes (Schattenhofer et al. 2011, Vila-Costa et al. 2012). These and other groups usually have to be identified by means of drawing the limits between populations in scatterplots, a process referred to as gating. Manual gating is done in 2-dimensional graphical representations of the flow cytometry variables of SSC and fluorescence (most frequently red and green). However, the limit between populations is not always clear, and significant subjectivity is introduced based on the analyst's criterion. In addition, it is often hard to visually discriminate subgroups in flow cytometry samples (Andreatta et al. 2004, Finak et al. 2009), introducing an important error in the bacterial population estimates. The magnitude of this error is even more important in large-scale studies where flow cytometry files are processed by different analysts.

In recent years, many automatic techniques for gating populations have been developed to minimize the time-consuming step of manual processing. Different clustering methods have been proposed (Rajwa et al. 2008, Bashashati & Brinkman 2009, Scheuermann et al. 2009, Lahesmaa-Korpinen et al. 2011, Aghaeepour et al. 2013). A clustering technique that has given good results in medical and health sciences is model-based clustering. Model-based clustering is an unsupervised clustering technique, meaning that it attempts to classify the data in a given number of homogeneous groups without the need for training by the user with a set of classified examples. This technique tries to find the best model that describes the structure in the data. Lo et al. (2009) proposed a model-based clustering approach based on *t*-mixture models with a Box-Cox transformation. Contrary to previous methods, their algorithm can detect groups with an elliptical shape and is unbiased to the presence of outliers. An additional advantage of automatic clustering techniques is that the number of groups detected is variable, so, for example, it is possible to easily detect subgroups within the HNA and LNA categories usually detected by manual gating.

Our objective was to evaluate the performance of *t*-mixture model-based clustering for HNA and LNA bacterioplankton from a coastal environment. We developed a methodology for automated clustering of bacterioplankton flow cytometry data and tested it with a long-term database that had also been analyzed manually. Our method is an adaptation of the flowClust function (Lo et al. 2009) for the identification of bacterioplankton groups. In addition, we provide an evaluation of the errors introduced by the subjectivity in the manual gating of bacterioplankton by comparing the analyses of the same flow cytometry plots by different analysts.

MATERIALS AND METHODS

Sample collection and analysis using traditional manual gating

A monthly 9 yr time series of flow cytometry bacterioplankton samples collected under the Radiales programme (Spanish Institute of Oceanography, IEO) at 3 stations off Gijón in the southern Bay of Biscay was used to validate the methodology developed. Samples were collected at 8 depths (from the surface to 150 m depth). Bacterioplankton samples were preserved with 1 % paraformaldehyde + 0.05 % glutaraldehyde and frozen at −80°C. To analyze heterotrophic bacterioplankton, an aliquot of 0.4 ml was stained with 2.5 µmol l^{-1} SYTO-13 NA fluorochrome (Molecular Probes) and analyzed using a FACSCalibur flow cytometer (BD/Becton, Dickinson and Company) equipped with a laser emitting at 488 nm.

We used a total of 2050 files in this intercomparison. Flow cytometry scatterplots are routinely analyzed manually as part of the ongoing time-series programme. Manual gating has been performed by different analysts (although they all received the same training), so a certain degree of variability can be expected in the manually gated data. Fluorescent latex beads (1.0 µm diameter, Molecular Probes) were added as an internal standard to relate the measured SSC and fluorescence signals, which might change slightly from sample to sample due to laser drift, to the constant SSC and fluorescence of the beads. LNA and HNA bacteria were easily distinguished within bacterioplankton based on their relative green fluorescence (FL1, a variable related to NA content, (Marie et al. 1997)). LNA cells were almost always smaller (lower SSC values) than the HNA counterparts (Calvo-Díaz & Morán 2006).

Automatic gating

Detection of beads

Because beads are used as the standard, they need to be identified very precisely. In many samples, the beads population is a mixture of single beads, doublets, triplets and higher associations. Beads usually lay in an area of the cytogram where no other population exists, making it easier to separate these subpopulations. Hence, we have developed an additional methodology to analyze the beads first, so that they could be used later to correct bacterioplankton data.

For the automatic analysis of beads, we first selected a window of SSC and FL1 containing the beads population for all samples. A window wide enough is needed to ensure that the beads do not lay outside of the selected window if slight changes in the laser occur.

In this window, the population of single beads is a bi-normal distribution (i.e. the beads have normal SSC and FL1 distributions). This bi-normal distribution of single beads sometimes has 2 or 3 adjacent subpopulations (doublets, triplets, etc.). To automatically gate the subpopulation of single beads within this wide window, and since the single bead population is the most abundant, we first located the SSC and FL1 corresponding to the most abundant histogram class. Then, for the specific brand of beads used in our samples, a range of ±0.2 log SSC and FL1 units around the mode encompassed all the single beads but removed doublets and triplets and was therefore used to select a reduced window containing all single beads.

Once this reduced window was automatically selected, we fitted a normal distribution to the SSC data. For a large number of samples, the beads population was superimposed over background noise. Therefore, the beads population was defined as those particles falling within the 99% confidence intervals of the fitted normal distribution. The same analysis was performed automatically for all selected FCS files. If different beads were used, the ranges of SSC and FL1 for the window selection would need to be readjusted.

Detection of bacterioplankton groups

Although bacterioplankton tend to show marked bimodal distributions of FL1, distinct groups may be occasionally hard to detect because they are composed of different subgroups with similar SSC and FL1. In addition, not all groups of bacterioplankton are distributed following a normal distribution. *t*-mixture model-based clustering is supposedly able to cope with these peculiarities. We used the R package flowClust (Clustering for Flow Cytometry) (Lo et al. 2009) to evaluate the ability of this technique to cluster natural bacterial populations.

Software

FlowClust uses a model-based clustering approach based on *t*-mixture models with a Box-Cox transformation. In *t*-mixture models, the number of subpopulations to detect needs to be specified *a priori*. The flowClust package uses a expectation-maximization algorithm to calculate the parameters when fitting *t*-distributions to the *n* subpopulations (Lo et al. 2008), selecting the best fit. To test the goodness of the fitted *t*-distributions, flowClust compares the error and tries to minimize it. A modified Box-Cox transformation is needed to minimize the effects of outliers present in the flow cytometry files. The resulting classification allows each cell to be assigned to one of the *n* subpopulations. Lo et al. (2008) provides a more detailed description of the flowClust algorithm.

File analysis

FCS files were read into R using the Bioconductor package flowCore (Hahne et al. 2009) providing flow cytometric signals for each cell. The variables used in the flowClust function were SSC, FL1 and red fluorescence (FL3), characteristic of autotrophic cells. Once the flow cytometry file was read, we had a multiparameter matrix with all values of each of the cells of that file. Data were log-transformed for subsequent analyses.

As we have explained above, to apply a classification using *t*-mixture models, we need to specify the number of clusters (*K*) to be fitted. But the number of bacterioplankton populations in a cytogram changes, and there can be several groups of noise (e.g. viruses, sample contamination). To select the best number of clusters in a sample, and given that the number of subpopulations was rarely >10, we fitted 10 *t*-mixture models to each cytogram, with *K* from 1 to 10. For each *K* we calculated the Bayesian information criterion (BIC) (Lo et al. 2009) relative to the maximum value of BIC for the 10 fitted models. The most appropriate value of *K* for each cytogram was selected as the lower *K* with BIC >95%.

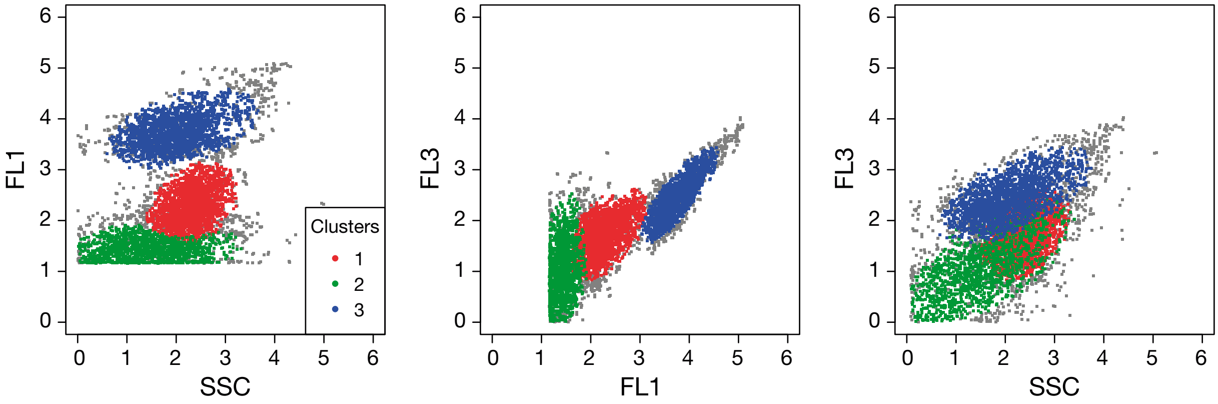


Fig. 1. Clustering output obtained by the flowClust function for an example file. The axes show the flow cytometer variables used for clustering: green fluorescence (FL1), red fluorescence (FL3) and side scatter (SSC)

We filtered the flow cytometry data before the fit of the 10 *t*-distribution models. Filtering data prior to analysis is a common practice in automatic clustering methodologies (Andreatta et al. 2001, Zare et al. 2010, Ribalet et al. 2011) that reduces the effect of noise and speeds up the analysis. However, this pretreatment has to be carefully applied to not affect the biologically important information. The filtering procedure is based on a binning of the data, so an image representation of the data is produced. Each pixel of the image represents the number of cells present in the given combination of SSC and FL1. To filter the data, if <2 cells fall in a pixel, they are removed. Three sequential filters were performed on each FCS file. First, a 3-dimensional filter (SSC, FL1, FL3 with 40 bins in each axis), then 2-dimensional filters (first SSC vs. FL1, then FL1 vs. FL3 with 80 bins in each axis). This pre-filtering is an optional step prior to the clustering function; therefore, it could be used or not depending on the expert criterion.

Once we had the optimal number of clusters for each file, the flowClust function returned each of the *K* groups and each cell was assigned to a given group. The flowClust is performed on the filtered data and hence, even if the threshold to filter the data is carefully selected, the number of cells belonging to each group could be underestimated. To correct this underestimation, we first calculated the boundaries of the cloud of points in the 3-dimensional (SSC, FL1 and FL3) space for each of the *K* groups. Then, with the unfiltered data, we used these boundaries to assign each cell to a group. If a cell was not within any group, it was labelled as an outlier. To calculate the boundaries of a group, we used a convolution algorithm that calculates the smallest polygon that includes the input points (using function convhulln of the package 'geometry' available in R).

Fig. 1 shows an example summary graphic of the flowClust results for one file. FlowClust returns a numbered list of groups, but we do not know which cluster represents each subgroup of bacteria (HNA and LNA in this case) or whether the cluster is in fact a group of noise. To assign each cluster to a category, we designed a method to label each group when all the FCS files had been analyzed. We first calculated the mean value of the flow cytometry parameters for each cluster. Then we corrected these data using mean values for the beads for the corresponding FCS file. We plotted the corrected mean value for each group using the SSC, FL1 and FL3 signals. We obtained a plot of all the centres of the groups of all the FCS files analyzed. We gated this plot manually to select which groups were assigned as HNA or LNA (Fig. 2). The purpose of this step was to label the groups with names instead of numbers.

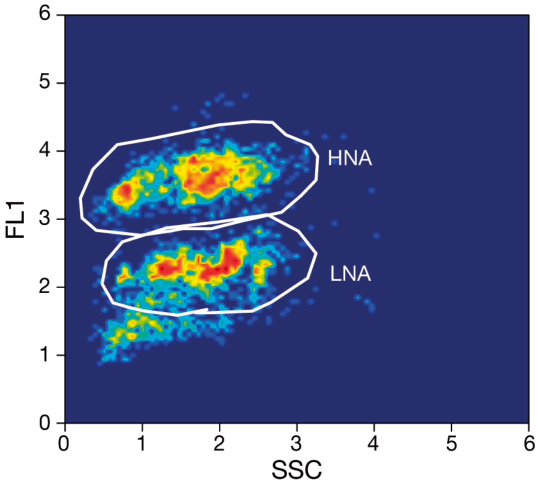


Fig. 2. Manual gating in the labelling step of flowClust output groups using the parameters side scatter (SSC, *x*-axis) and green fluorescence (FL1, *y*-axis). HNA: high nucleic acid content; LNA: low nucleic acid content

The R code to automatically analyze bacterioplankton FCS files is provided in the Supplement at www.int-res.com/articles/suppl/a072p175_supp/.

Validation

We compared our automatic clustering with the manual analysis of 2050 files. The automated analysis of the beads was also compared with the manual analysis. Bacterioplankton groups were divided into 2 categories, HNA and LNA, to match the groups detected in the manual-gated database processing. To determine the inherent variability in the manual analysis, 10 FCS files randomly selected from the Radiales database were analyzed by 6 different experts. For the 2050 files analyzed by the ongoing Radiales monitoring programme, we had information on the cytometric properties for each group, and mean size values calculated using the arithmetic mean. Since the distribution of flow cytometric properties is frequently not normal, other measures of central tendency, such as the median, might be more appropriate. To evaluate which measure of central tendency is more adequate, we manually analyzed 50 files selected to have maximum variability in the data and including a wide range of depths and dates. Files were manually analyzed by the same person. We also provide error estimations between different methods and experts.

RESULTS

The correlation between the automatic clustering and the manual gating data was high for the counts (Fig. 3A–C; r = 1 for beads, r = 0.96 for LNA and r = 0.97 for HNA groups), but rather low for the average SSC (Fig. 3E,F; r = 0.37 for LNA bacteria and r = 0.56 for HNA bacteria). These estimates of average SSC represent the arithmetic mean of the SSC of the particles within each group, but many bacterioplankton groups have SSC distributions that are not normally or even log-normally distributed. Hence, the median of the distribution is a more reliable estimate of the average SSC as it is less sensitive to data distribution. Because the Radiales dataset was analyzed using the arithmetic mean and it was not feasible to manually reanalyze the 2050 FCS files, we selected 50 files and manually gated

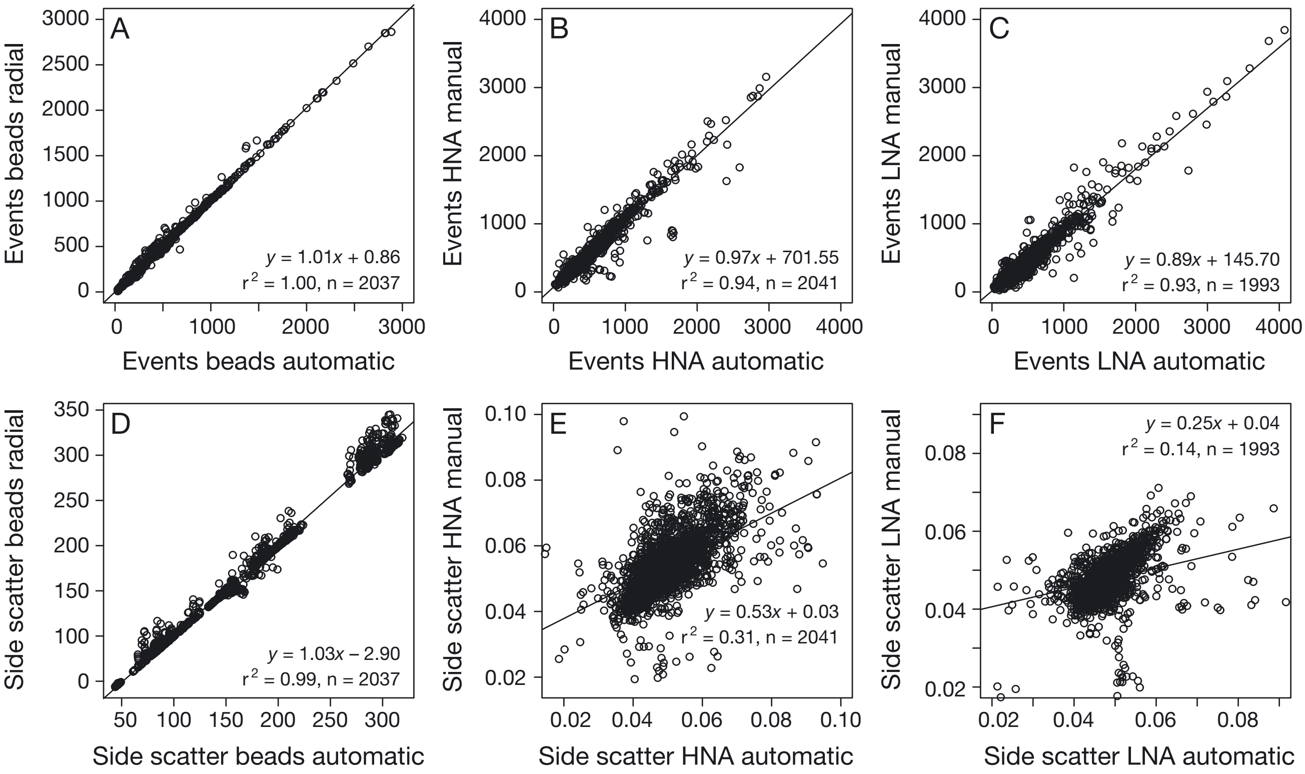


Fig. 3. Correlation between the output of manual (*y*-axis) and automatic (*x*-axis) methods. Upper panels show the counts and lower panels the side scatter for (A,D) beads, (B,E) high nucleic acid content (HNA), and (C,F) low nucleic acid content (LNA) groups

the data. We calculated the median instead of the arithmetic mean for both automatic and manual methods and a better correlation between the median SSC estimates was found (see Appendix; r = 0.76 for LNA bacteria and r = 0.99 for HNA bacteria).

However, a good correlation is not a measure of the correspondence between 2 methodologies. Usually, when comparing 2 analyses, a test on whether the slope and intercept of the fit are significantly different than 1 and 0 is provided. For the data shown in Fig. 3, all slopes and intercepts are significantly different than 1 and 0, respectively. Even for the beads where the differences between the 2 methodologies are irrelevant (i.e. when manual counts were 1000 beads the automatic method counted 1011), the ANOVA revealed significant differences (test for slope different to 1: $F_{1,2036} = 85$, $p < 0.001$). This suggested that in our database of 2050 FCS files, the ANOVA had so much power that even the slightest differences were detected as statistically significant and, therefore, we can not rely just on the p-value to study the goodness of fit of our method.

Despite this problem with the exceedingly high power of the ANOVA, the significance of the tests points to dissimilarities between the methodologies. However, it is also necessary to study the magnitude and whether these differences are biologically important (Peters 1991).

To detect the disagreement or bias between the methodologies, we estimated the error, which represents the ratio between the automatic and the manual methods (Fig. 4). A positive error hence indicates an overestimation by the automatic method while a negative error implies an underestimation by the automatic method. When this error was calculated, we found important differences between the manual and automatic methods both in the HNA and the LNA bacteria groups, especially at the lowest counts (Fig. 4).

To understand the origin of the error, we evaluated the influence of the error introduced by the analyst who manually gated the data. We therefore asked an external analyst who was not involved in the processing of the Radiales database to manually gate the same 50 files (Fig. 5C,D). Although the error between these 2 manual gating analyses (Fig. 5C,D) was lower than the comparison between automatic and manual (Fig. 5A,B), the differences observed suggest that even when the gating was done manually by 2 different experts important errors were found.

Interestingly, the errors of the latest comparison, and in particular the errors in Fig. 5C, suggest that the external analyst underestimates the abundance at the lowest counts, very similarly to the automatic method. We hypothesized that this underestimation was due to the fact that when an analyst is asked to gate 50 files for an intercomparison study, he/she pays special attention to adjusting the gates as much as possible, trying to avoid the inclusion of adjacent noise or outliers. To test this hypothesis, we asked an analyst involved in the routine processing of the Radiales database to gate these 50 files (we refer to this analysis as re-analysis data). The counts were quite similar to the external analyst (Fig. 5E,F) and different to the routine database counts (data not shown). The error was higher for HNA at the lowest counts. We tested for significant differences in the counts for each intercomparison in Fig. 5 using paired *t*-tests. For HNA bacteria, counts obtained by the external analyst and the re-analysis counts were not significantly different (Fig. 5E: *t*-test, $t = -1.47$, df = 49, p = 0.15). However, for the other 2 cases, significant differences were found (Fig. 5A: *t*-test, $t = -20.51$, df = 48, $p < 0.001$; Fig. 5C: *t*-test, $t = -10.59$, df = 49, $p < 0.001$). A different tendency was found for the LNA subgroup with significant differences between counts by the external analyst and the re-analysis (Fig. 5D: *t*-test, $t = 2.7$, df = 49, p = 0.009; Fig. 5F: *t*-test, $t = -7.43$, df = 49, $p < 0.001$), and no differences between automatic and manual gating (Fig. 5B: *t*-test, $t = 1.88$, df = 47, p = 0.07).

To obtain a quantitative estimate of the influence of the analyst, we tested the variation in the counts obtained by 6 experts. When the same flow cytom-

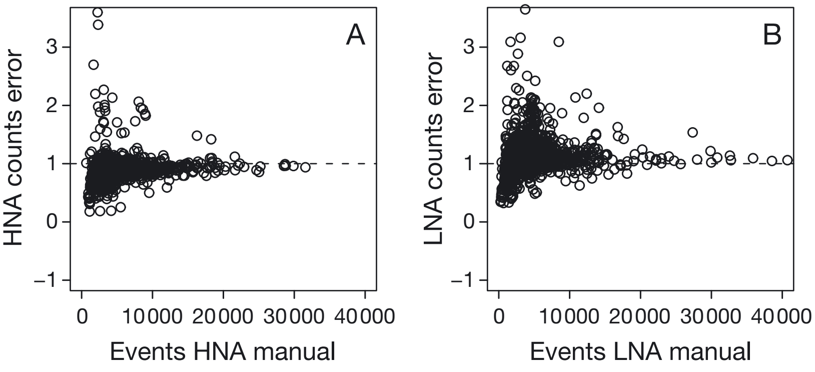


Fig. 4. Relationship between error (defined as the ratio between the automatic and manual counts) and the number of events analyzed for (A) high nucleic acid content (HNA) and (B) low nucleic acid content (LNA) bacteria using the whole Radiales database

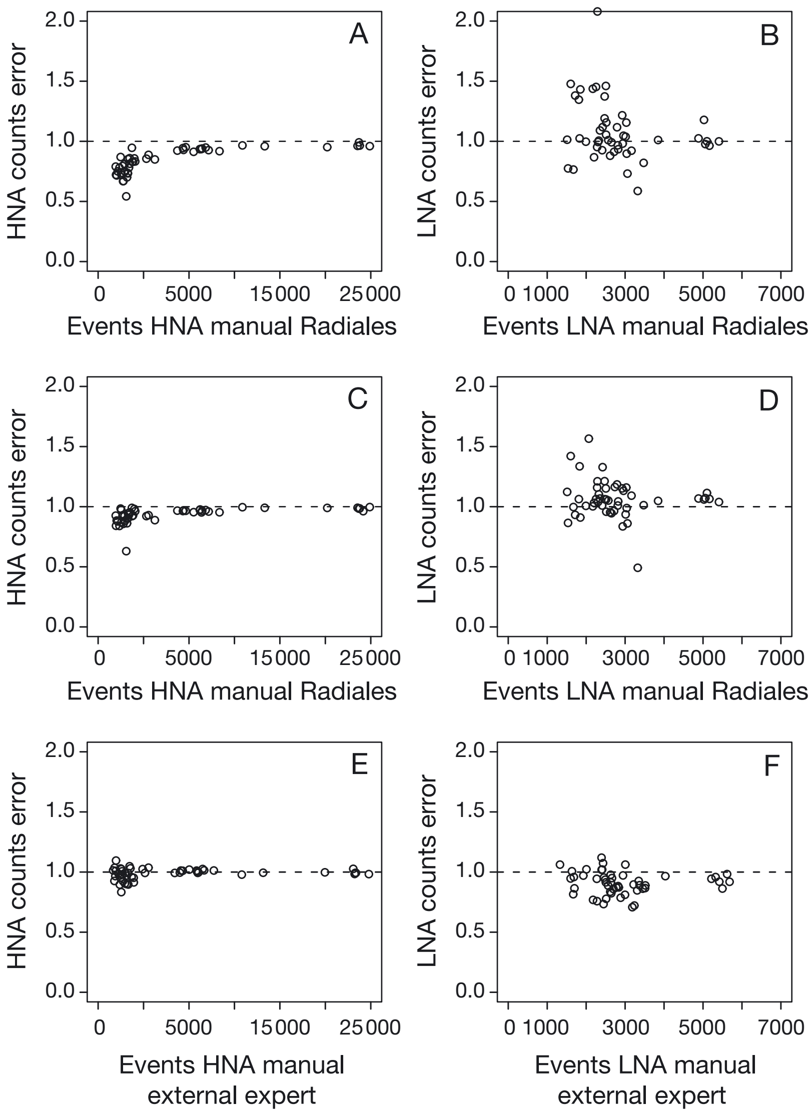


Fig. 5. Error between different analysts and the automatic and manual methods for bacteria counts using 50 files. (A,B) Ratio between automatic and manual gating (*y*-axis) against the manual counts (*x*-axis). (C,D) Ratio between the counts manually obtained by an external expert and the manual counts of the Radiales database. (E,F) Ratio between the gating of an external expert and another expert who collaborated in the processing of the Radiales database

etry file was analyzed by different experts the coefficient of variation between the HNA and LNA bacterial groups counts was relatively high, with values around 10 to 20%. For the beads counts, the variability dropped to around 5% (Fig. 6). Similar to Fig. 5, the variation between experts was higher for samples with lower abundance of cells for beads and HNA groups; however, we did not observe this effect for the LNA subgroup.

Finally, to study whether the differences between methods were ecologically significant, we calculated a mean annual cycle using all available surface data at station 3. Seasonal cycles of total bacterial abundance obtained by manual and automatic methods were similar (Fig. 7A). However, when the contribution of HNA cells (% HNA) was examined, some differences were found (Fig. 7B) yet showing a similar pattern. Fig. 7C,D shows the seasonal cycles of mean LNA–HNA cell biovolume. Differences were higher for the HNA group (13%) than for LNA (1.4%), but the pattern was quite similar.

DISCUSSION

Recently, several automatic techniques for flow cytometry data processing have been widely applied in medical fields (Le Meur 2013, Robinson et al. 2012, Aghaeepour et al. 2013). However, in microbial ecology, experts still rely largely on manual gating of the FCS files. One peculiarity of flow cytometry samples of bacterioplankton when we process samples is the high variability between samples hampering automated analysis. Although some of these automated methods have been applied to analyze planktonic groups (Andreatta et al. 2004, Ribalet et al. 2011), these methodologies are not able to cope with such dynamic features, due to the presence of noise or variability of the community composition, usually encountered in large-scale studies. For example, Ribalet et al. (2011) used a set of pre-defined windows where picoplankton populations should lay and the method detected the spatial dynamics of each group of bacterioplankton within each window. Therefore, their method is not readily applicable whenever the position in the cytogram of each group changes. This is frequently the case in large datasets. Andreatta et al. (2001) used image analysis to identify subgroups of bacterial populations in FCS files, but again, manual gates were required to distinguish each bacterial population from the background noise.

Another special feature of bacterioplankton FCS files is that the number of subgroups and the boundaries between them are not always clearly defined. Frequently, it is even hard to identify them manually. Usually, it is easy to differentiate between the 2 widespread groups of bacterioplankton assessed here (HNA and LNA), but often other subgroups are also present. In addition, the presence of outliers in the cytogram is quite common and not all automatic methods respond adequately to the presence of a significant contribution of them (Luta 2011). The model-based clustering used in the flowClust algorithm is robust to the presence of outliers (Lo et al. 2009, Finak et al. 2010). Another interesting feature of this method is the ability to detect non-elliptical population shapes, such as the HNA cell distribution (e.g. cluster 3 in Fig. 1) in contrast to the more circular LNA bacteria cell distribution (e.g. cluster 1 in Fig. 1), as is usually the case in aquatic samples (Bouvier et al. 2007). Although other automated techniques are unbiased for population shape, they are not able to detect overlapping groups (Naumann & Wand 2009, Naumann et al. 2010, Sugár & Sealfon 2010, Ge & Sealfon 2012). Finally, other methods cannot be applied to large datasets due to computational efficiency (Zare et al. 2010). In addition, the flowClust algorithm is provided as open software and it is thus free to use and modify.

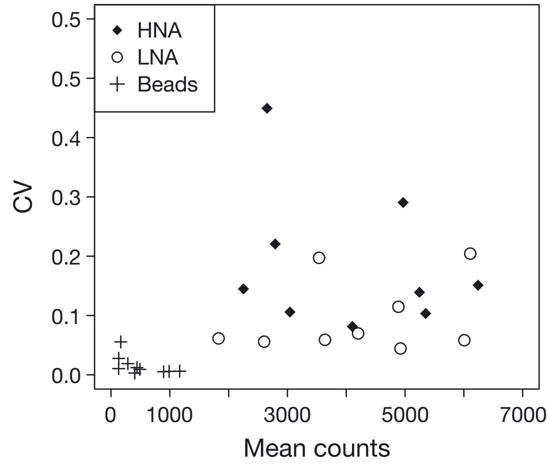


Fig. 6. Relationship between the coefficient of variation (CV) for 10 flow cytometry standard (FCS) files analyzed by 6 different experts and the mean counts. HNA: high nucleic acid content, LNA: low nucleic acid content

The main aim of this work was to apply the flowClust algorithm to develop an automatic and standardized method for processing flow cytometry analysis of heterotrophic bacterioplankton groups that could be routinely adopted as an alternative to manual processing. We tested it under a real-case scenario using a large dataset and compared the results with the traditional, manual gating technique. We used a database consisting of 9 yr of monthly sampling of continental shelf bacteria from the surface down to 150 m, characterized by a wide range of natural variability at the seasonal and spatial (inshore–offshore and vertical gradients) scales (Calvo-Díaz & Morán 2006, Morán & Calvo-Díaz 2009).

As we have explained in the 'Results' section, automatically clustered bacterioplankton groups were aggregated into 2 categories, HNA and LNA, to be able to compare them with the manual method. Manual gating is more limited for identifying bacterioplankton subgroups or even cyanobacteria, especially when groups such as *Prochcloroccus* overlap with the HNA group in natural bacterioplankton samples. Nevertheless, the methodology we propose is able to detect a higher number of bacterioplankton groups and subgroups as there is no *a priori* restriction on the number of groups that can be detected. However, we recommend

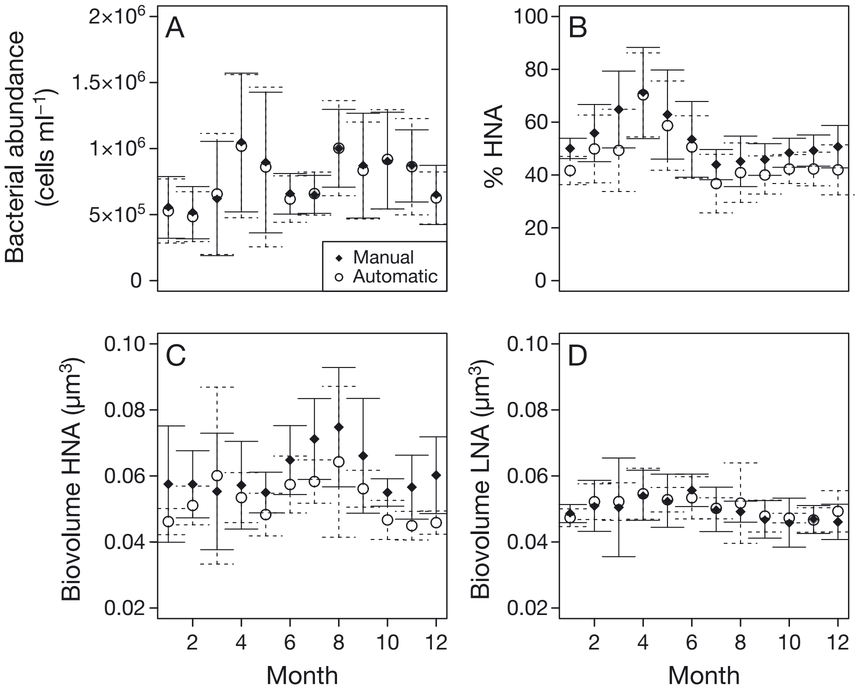


Fig. 7. Mean annual cycle at the surface of the RADIALES station 3 off Gijón, Spain, for the 2002–2010 period obtained by the manual and automatic methods. (A) Monthly mean bacterial abundance counts, (B) monthly mean contribution of high nucleic acid content (HNA) cells, (C) mean biovolume for HNA and (D) mean biovolumes of low nucleic acid content (LNA) cells. Solid error bar: manual SD; dashed error bar: automatic SD

visual inspection of the flowClust output to understand the results of the automatic clustering. For example, in open-ocean surface samples where *Prochcloroccous* populations frequently partially fall within the HNA cell cluster, our methods detected them quite accurately (data not shown). Nevertheless, in these types of conditions both *in vivo* and stained samples are usually analyzed, so it would be easier to automatically gate both samples in a similar way to how it is done with manual gating to differentiate autotrophic and heterotrophic cells.

We have shown that the subjectivity of the analyst can introduce around 10 to 20% of variability in the manual gating of bacterioplankton samples (Fig. 6). The consequences of this variability are important, especially in large-scale studies where data from different analysts are combined. An automatic approach such as the one we have developed certainly does better at comparing different datasets, producing more consistent counts. Despite its importance, the subjectivity of manual gating has not been quantified previously. We have shown that not only different analysts reach different counts due to this subjectivity error but also the same expert can introduce some bias due to fatigue when analyzing a large number of samples.

The quality of the samples analyzed by flow cytometry is important for the results of the automated analysis. Similarly, when gating is done manually, samples are analyzed more effectively if all the groups are centred with the adequate number of cells per sample and without much noise. In testing the method, major differences between automated and manual counts were due to the presence of these problematic files. It is therefore important to follow recommendations (Marie et al. 1997, Gasol & del Giorgio 2000) on using bacterioplankton optimal flow rates and number of cells per sample. At higher rates, populations begin to overlap and it is more difficult to set limits between them, even automatically.

Despite previous filtering being applied to the data, the method is relatively fast (<5 min per sample). The processing time for our method is similar to the time spent processing FCS files manually, but the clustering is unsupervised and the time required is only computing time and not analyst time. However, we strongly recommend a visual check of the analysis output. This method has the advantage of being more objective and reproducible. It is possible to reduce the analysis time using a server with multiple processors, as we did. Moreover, the fact that the FCS files are read using the flowCore package allows access to the information on each cell in the file. This information allows us to know the distribution of any group and thus improves the subsequent statistical treatments. We propose that the beads should be analyzed before the clustering technique using another method because beads measurements are required to correct for the deviations that cytometer lasers can experience with time (Shapiro 1995).

FlowClust provides an effective tool for gating the populations, mainly for counts. Significant differences were still found in the biovolume estimates (with an empirically determined calibration between SSC and cell diameter (Calvo-Díaz & Morán 2006)), especially for the HNA group (Fig. 7B), which was 13% higher. While the LNA group has a more or less spherical population shape, the HNA group is usually more irregular with a non-spherical shape. Consequently, the measure of central tendency that is used to calculate the mean size introduces more variability for the HNA case (Fig. 7B–D). % HNA values obtained by the automatic method (Fig. 7C) are usually lower than the % HNA obtained by manual gating, although the distinct seasonal pattern of maxima in April and minima in July (Calvo-Díaz & Morán 2006, Morán & Calvo-Díaz 2009) was well reproduced. This is due to both an underestimation of the abundance of HNA cells and an overestimation of LNA cells by the automatic method (Fig. 4). These methodological deviations between the manual and automatic analyses can result in different ecological patterns in cell size and biomass estimates. These differences become more important in large-scale studies or when we compare different databases. In these type of studies, the importance of using objective and standardized methods of clustering in microbial ecology becomes critical.

In summary, our methodology is a powerful tool to analyze groups and subgroups of heterotrophic bacterioplankton, allowing the processing of thousands of files, quickly and with reduced error. The computer-based processing of the FCS files results in the full automation of sample analysis by flow cytometry. It increases the efficiency and quality of the results and makes them comparable with other data from large-scale studies. The technique could be easily adapted to the analysis of phytoplankton samples or even extended to the analysis of viruses.

Acknowledgements. This work was supported by METabolic OCean Analysis (METOCA) project funded by Spanish National Investigation+Development+Innovation (I+D+I) Plan (METOCA, CTM2009-13882-MAR) and by the time-series project RADIALES from the Instituto Español de Oceanografía (IEO). F.C.G. was funded by a Formación de Personal Investigador (FPI) grant programme from the Spanish Ministry of Economy and Competitivity (MINECO). We thank J. Bueno and S. Sal for their useful comments.

LITERATURE CITED

Aghaeepour N, Finak G, Hoos H, Mosmann TR, Brinkman R, Gottardo R, Scheuermann RH (2013) Critical assessment of automated flow cytometry data analysis techniques. Nat Methods 10:228–238

Andreatta S, Wallinger MM, Posch T, Psenner R (2001) Detection of subgroups from flow cytometry measurements of heterotrophic bacterioplankton by image analysis. Cytometry 44:218–225

Andreatta S, Wallinger MM, Piera J, Catalan J, Psenner R, Hofer JS, Sommaruga R (2004) Tools for discrimination and analysis of lake bacterioplankton subgroups measured by flow cytometry in a high-resolution depth profile. Aquat Microb Ecol 36:107–115

Azam F, Fenchel T, Field JG, Gray JS, Meyerreil LA, Thingstad F (1983) The ecological role of water-column microbes in the sea. Mar Ecol Prog Ser 10:257–263

Bashashati, A, Brinkman RR (2009) A survey of flow cytometry data analysis methods. Adv Bioinformatics 2009: 584603

Bouvier T, del Giorgio PA, Gasol JM (2007) A comparative study of the cytometric characteristics of high and low nucleic-acid bacterioplankton cells from different aquatic ecosystems. Environ Microbiol 9:2050–2066

Calvo-Díaz A, Morán XAG (2006) Seasonal dynamics of picoplankton in shelf waters of the southern Bay of Biscay. Aquat Microb Ecol 42:159–174

Felip M, Andreatta S, Sommaruga R, Straskrabova V, Catalan J (2007) Suitability of flow cytometry for estimating bacterial biovolume in natural plankton samples: comparison with microscopy data. Appl Environ Microbiol 73:4508–4514

Finak G, Bashashati A, Brinkman R, Gottardo R (2009) Merging mixture components for cell population identification in flow cytometry. Adv Bioinformatics 2009:247646

Finak G, Perez JM, Weng A, Gottardo R (2010) Optimizing transformations for automated, high throughput analysis of flow cytometry data. BMC Bioinformatics 11:546

Gasol JM, del Giorgio PA (2000) Using flow cytometry for counting natural planktonic bacteria and understanding the structure of planktonic bacterial communities. Sci Mar 64:197–224

Gasol JM, Zweifel UL, Peters F, Fuhrman JA, Hagstrom A (1999) Significance of size and nucleic acid content heterogeneity as measured by flow cytometry in natural planktonic bacteria. Appl Environ Microbiol 65:4475–4483

Ge Y, Sealfon SC (2012) flowPeaks: a fast unsupervised clustering for flow cytometry data via *K*-means and density peak finding. Bioinformatics 28:2052–2058

Hahne F, LeMeur N, Brinkman RR, Ellis B and others (2009) flowCore: a Bioconductor package for high throughput flow cytometry. BMC Bioinformatics 10:106

Hansell DA, Carlson CA (1998) Deep-ocean gradients in the concentration of dissolved organic carbon. Nature 395: 263–266

Lahesmaa-Korpinen, AM, Jalkanen SE, Chen P, Valo E and others (2011) FlowAnd: comprehensive computational framework for flow cytometry data analysis. J Proteomics Bioinformatics 4:245–249

Le Meur N (2013) Computational methods for evaluation of cell-based data assessment — Bioconductor. Curr Opin Biotechnol 24:105–111

Li WKW, Jellett JF, Dickie PM (1995) DNA distributions in planktonic bacteria stained with TOTO or TO-PRO. Limnol Oceanogr 40:1485–1495

Lo K, Brinkman RR, Gottardo R (2008) Automated gating of flow cytometry data via robust model-based clustering. Cytometry A 73:321–332

Lo K, Hahne F, Brinkman RR, Gottardo R (2009) flowClust: a Bioconductor package for automated gating of flow cytometry data. BMC Bioinformatics 10:145

Luta G (2011) On extensions of *k*-means clustering for automated gating of flow cytometry data. Cytometry A 79: 3–5

Marie D, Partensky F, Jacquet S, Vaulot D (1997) Enumeration and cell cycle analysis of natural populations of marine picoplankton by flow cytometry using the nucleic acid stain SYBR Green I. Appl Environ Microbiol 63: 186–193

Morán XAG, Calvo-Díaz A (2009) Single-cell vs. bulk activity properties of coastal bacterioplankton over an annual cycle in a temperate ecosystem. FEMS Microbiol Ecol 67: 43–56

Naumann U, Wand MP (2009) Automation in high-content flow cytometry screening. Cytometry A 75:789–797

Naumann U, Luta G, Wand MP (2010) The curvHDR method for gating flow cytometry samples. BMC Bioinformatics 11:44

Peters RH (1991) A critique for ecology. Cambridge University Press, Cambridge

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 A 73:369–379

Ribalet F, Schruth DM, Armbrust EV (2011) flowPhyto: enabling automated analysis of microscopic algae from continuous flow cytometric data. Bioinformatics 27: 732–733

Robinson JP, Rajwa B, Patsekin V, Davisson V (2012) Computational analyses of highthroughput flow cytometry data. Exp Opin Drug Discov 7:679–693

Schattenhofer M, Wulf J, Kostadinov I, Gloeckner FO, Zubkov MV, Fuchs BM (2011) Phylogenetic characterisation of picoplanktonic populations with high and low nucleic acid content in the North Atlantic Ocean. Syst Appl Microbiol 34:470–475

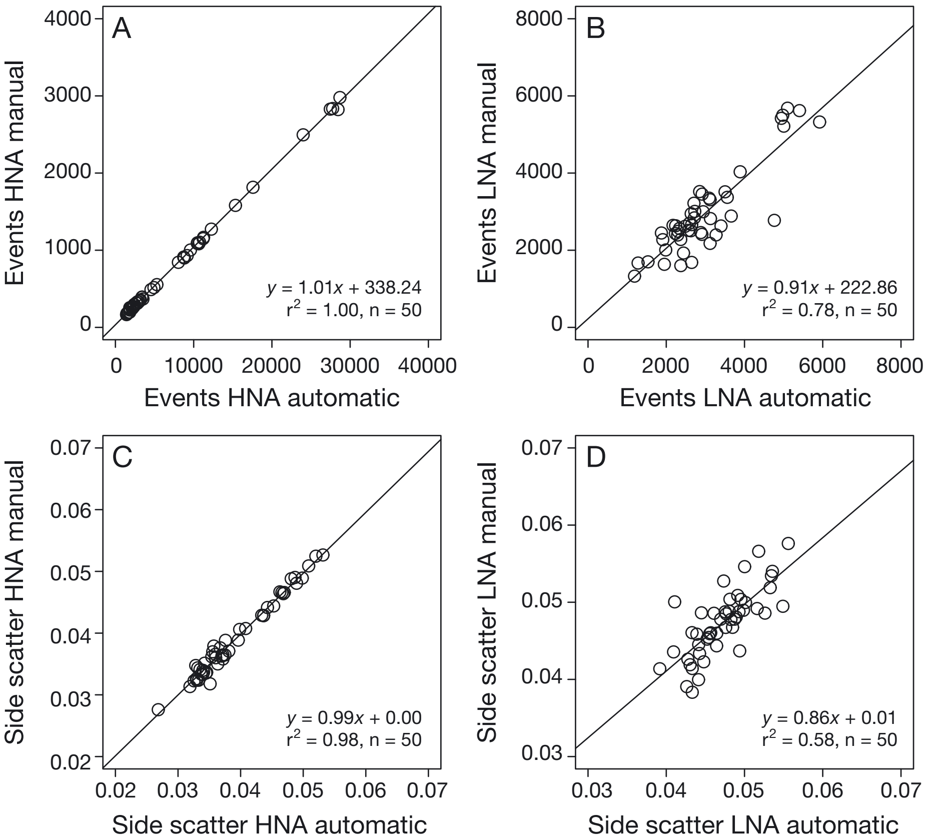
Scheuermann R, Quian Y, Wei C, Sanz I (2009) ImmPort FLOCK: automated cell population identification in high dimensional flow cytometry data. J Immunol 182 (Meeting Abstract Suppl):42.17

Shapiro H (1995) Practical flow cytometry, 3rd edn. Wiley-Liss, New York, NY

Sugár IP, Sealfon SC (2010) Misty Mountain clustering: application to fast unsupervised flow cytometry gating. BMC Bioinformatics 11:502

Vila-Costa M, Gasol JM, Sharma S, Moran MA (2012) Community analysis of high- and low-nucleic acid-containing bacteria in NW Mediterranean coastal waters using 16S rDNA pyrosequencing. Environ Microbiol 14:1390–1402

Zare H, Shooshtari P, Gupta A, Brinkman RR (2010) Data reduction for spectral clustering to analyze high throughput flow cytometry data. BMC Bioinformatics 11:403



Appendix. Correlation between the output of manual (y-axis) and automatic (x-axis) methods using 50 files from the Radiales database. The upper panels show the relationship for counts of (A) high nucleic acid content (HNA) bacteria and (B) low nucleic acid content (LNA) bacteria, while the lower panels show the comparison between median side scatter of (C) HNA bacteria and (D) LNA bacteria

Editorial responsibility: Karel Šimek,
Ceské Budějovice, Czech Republic

Submitted: July 29, 2013; Accepted: February 18, 2014
Proofs received from author(s): April 10, 2014