Deconvolution model for cytometric microbial subgroups along a freshwater hydrologic continuum

Authors

Stefano Amalfitano[1*], Stefano Fazi[1], Anna M. Romaní[2], Butturini Andrea[3]

Institutions

[1] Water Research Institute, National Research Council of Italy (IRSA-CNR), Rome, Italy

[2]Institute of Aquatic Ecology, University of Girona, Spain

[3] Department of Ecology, University of Barcelona, Spain

*Corresponding author

Tel.: +39 06 90672 854

Fax: +39 06 90672 787

E-mail address: amalfitano@irsa.cnr.it

Keywords

Flow Cytometry; River Continuum; Prokaryotes; Bacteria.

Abstract

Flow cytometry is suitable to discriminate and quantify aquatic microbial cells within a spectrum of fluorescence and light scatter signals. Using fixed operational and gating settings, a mixture model, coupled to Laplacian operator and Nelder-Mead optimization algorithm, allowed deconvolving bivariate cytometric profiles into single cell subgroups. This procedure was applied to outline recurrent patterns and quantitative changes of the aquatic microbial community along a river hydrologic continuum. We found five major persistent subgroups within each of the commonly retrieved populations of cells with Low and High content of Nucleic Acids (namely, LNA and HNA cells). Moreover, we assessed changes of the cytometric community profile over-imposed by water inputs from a wastewater treatment plant. Our approach for multiparametric data deconvolution confirmed that flow cytometry could represent a prime candidate technology for assessing microbial community patterns in flowing waters.

Introduction

Flow cytometry has been used in combination with statistical tools for dredging multiparametric representations of single cells within microbial communities from different aquatic environments (Glüge et al., 2014; Koch et al., 2014; Li, 2002). Because sample acquisition offers snapshots of single cells by delivering a multivariate dataset exportable for post-hoc analysis (Davey and Davey, 2011), different bioinformatics approaches were proposed to discriminate cytometric subgroups based on specific light scatter and fluorescence signals (Andreatta et al., 2004; Le Meur, 2013; De Roy et al., 2012). The basic cytometric cell detection combines: (i) light signals of the laser beam, scattered off at small and large angles form the cell interrogation point and related to cell size and morphology (i.e., forward and side light scatters); (ii) primary fluorescence signals related to type and content of endocellular autofluorescent pigments (i.e., autofluorescence); (iii) secondary fluorescence signals, owing to type and content of cell constitutive compounds detected upon specific staining procedures (e.g., nucleic acids) (Shapiro, 2005).

Numerous studies provided algorithms that automatically generate approximated gates to distinguish two or more cytometric groups in univariate and bivariate cytograms for a smoother interpretation of cytometric datasets (Aghaeepour et al., 2013; Hahne et al., 2009; Pyne et al., 2009; Verschoor et al., 2015). The cytometric-fingerprinting similarity among samples can be assessed and indicated through specific deviation plots and heat maps (Hsiao et al., 2016; Rogers and Holyst, 2009). However, gating and deconvolution procedures have found standardized procedures mainly for clinical diagnostic applications (Chattopadhyay and Roederer, 2012; Mittag and Tárnok, 2009; Perfetto et al., 2006).

Such procedures are fraught with failure when exploring cytometric profiles of environmental microbial communities, since the cytometric description of a natural system can be far more puzzled than that of a clinical specimen (Hyrkas et al., 2015; Koch et al., 2013b). The aquatic microbial communities comprise large populations of phylogenetically and phenotypically

74 dissimilar cells, whose structural and cytometric dynamics depend on their specific metabolic
75 preferences and abilities to cope with local environmental conditions (Koch et al., 2014).
76 Moreover, dispersal of microorganisms among communities (e.g., passive movements),
77 species sorting (e.g., selection of species within the local pool) and biotic interactions (e.g.,
78 resource competition, grazing activity, prey-predator balance) may fundamentally affect the
79 community structure and assembly processes (Shade et al., 2012).
80 Natural abiotic ranges and gradients were reported to determine the cytometric fingerprinting
81 of local communities within a given water mass at different temporal and spatial scale
82 (Schiaffino et al., 2013; Van Wambeke et al., 2011). Moreover, a large body of literature
83 reported a high level of analysis at log-scales to deal with the cytometric complexity of
84 environmental samples, such as those from marine, freshwater and groundwater systems
85 (Amalfitano et al., 2014; Boi et al., 2016; Vila-costa et al., 2012).
86 Given such structural and functional complexity, cytometric fingerprints may provide
87 information on structural dynamics of microbial communities, by detecting the modifications
88 of scatter and fluorescence signals of recurrent localized cytometric subgroups. Although
89 disregarding a direct taxonomic sense, localization and signal intensities of cytometric
90 subgroups have been used to define the cytometric profile of samples, which can be then
91 compared to others from the same system. This approach can be especially effective to assess
92 structural dynamics of microbial communities in highly dynamic systems such as flowing
93 waters, which receive inputs from tributaries of distinct characteristics to that of the main
94 stem. When waters flows directionally, each section along the hydrologic continuum acts as
95 both recipient and source of waters with definite physical, chemical and biological
96 characteristics (Nelson et al., 2009). The mass transport and the rate of external inputs to the
97 recipient volume determine the intensity of the mass effects and the residence time available
98 for microbial life processes (Niño-García et al., 2016).

By disregarding external inputs and stressors, it is assumable that the water network in a river system behaves as a passive corridor, particularly at high flow velocities and over short distances (Butturini et al., 2016), with the aquatic microbial community showing preserved structural traits. In such conditions, the cytometric fingerprinting might be also recurrent.

Here, we provide a methodological procedure suitable to deconvolve cytometric bivariate datasets into *n* subjacent subgroups. Specifically, we tested a method for processing microbiological patterns in the headwaters of a Mediterranean river, assuming that significant external water inputs may potentially affect community structure and, thus, the cytometric profiles along the river continuum.

Methods

Study site and sampling

River waters were sampled during a morning survey from the upstream area of the River Tordera (Barcelona, Spain), approximately every 3 km from its natural spring to the coastline. The anthropic impact is relatively low since only small urban settlements are located within the study area. Thus, the river flows almost unimpacted for about 20 Km, until the outflow of a small Waste Water Treatment Plant (WWTP) reaches the main stem (Freixa et al., 2016). We collected five samples from the river before the WWTP outflow (T1, T2, T3, T4, T5), the WWTP waters before the conjunction with the river (A1) and the river waters after the WWTP outflow (T6). All samples were immediately fixed (2% formaldehyde, final concentration).

Flow cytometry and cytograms

The aquatic microbial community was characterized within one week from sampling by using the Flow Cytometer A50-micro (Apogee Flow System, Hertfordshire, England) equipped with a solid-state laser set at 20 mV and tuned to an excitation wavelength of 488 nm. The

125 volumetric absolute counting was carried out on fixed samples, stained with SYBR Green I
126 (1:10000 dilution; Molecular Probes, Invitrogen) for 10 min in the dark at room temperature.
127 The light scattering signals (forward and side scatters) and the green fluorescence (530/30
128 nm) were acquired for the single cell characterization. Thresholding was carried out using the
129 green channel. Samples were run at low flow rates to keep the number of events below 1000
130 events/s (Gasol and Moran, 2015). The total number of prokaryotes was determined by their
131 signatures in a plot of the side scatter vs the green fluorescence. The intensity of green
132 fluorescence emitted by SYBR-positive cells allowed for the discrimination among cell
133 groups exhibiting different nucleic acid content and morphology. The instrumental settings
134 were kept the same for all samples in order to achieve comparable data (Prest et al., 2013).
135 The Apogee Histogram Software (v89.0) was used for data handling and visualization. A
136 preliminary gating was applied to distinguish the single-celled prokaryotic community from
137 background caused by suspended abiotic particulate, cells in aggregates and electronic noise.
138 This step was applied to all samples during the data acquisition in a cytogram of SSC versus
139 Green fluorescence, in accordance with previous published protocols (Gasol and Moran,
140 2015). The microbial community at each site was then represented in a plot of FSC vs Green
141 fluorescence at the image resolution of 1024x1024 pixels. For each sample, the data matrix of
142 the two variables was exported (.csv files) and deconvolved by the methodological approach
143 described below.
144
145 **Deconvolution approach and model description**
146 According to the Finite Distribution mixture Modelling (FDM), the complex surface $f_{(x,y)}$ of a
147 bivariate cytogram (i.e., Sybr Green fluorescence vs forward scatter) is described as the sum
148 of *n* subjacent peaks (eq. 1):

149 $$f_{(x,y)} = \sum_{i=1}^{n} c_{(x,y)_i} \qquad (1)$$

Each peak represents a subgroup that fits a predefined probabilistic density functions ($c_{(x,y)}$). Here, an asymmetric parameter (r) was incorporated into the Gaussian PDF probability model (Fruhwirth-Schnatter, 2006) in order to cope with asymmetries and long tails (Kato et al., 2002) (eq. 2):

$$c_{(x,y)_i} = a_i e^{-\left(\begin{cases} \frac{(\mu-\mu_i)^2}{2\sigma_i^2} & \text{if } \mu>\mu_i \\ \frac{(\mu-\mu_i)^2}{2r_i^2\sigma_i^2} & \text{otherwise} \end{cases}\right)} \quad (2)$$

If the skewness r_i (r_{ix}, r_{iy}) is equal to the unity, equation 2 is equivalent to a Gaussian distribution defined by its mean μ_i (μ_{ix}, μ_{iy}), deviation σ_i (σ_{ix}, σ_{iy}), and height a_i (a_{ix}, a_{iy}).

A two-steps procedure was performed to estimate the unknown parameters of recurrent peaks (μ_i, σ_i, a_i *and* r_i), and to cluster those peaks into subgroups of events with a direct quantification of their density.

In step A, a surface analysis was performed according to the Nelder-Mead optimisation algorithm to detect and locate the position of local maxima ($L_n = \{\mu_1, \mu_2, \mu_3, \mu_n\}$) in the $f_{(xy)}$, and the position of local minima of the differential Laplacian operator of $f_{(xy)}$ ($\nabla^2 f$) (Ganzha and Vorozhtsov, 1996; Horst and Pardalos, 2013). To avoid overestimating the number (n) of potential subjacent peaks (L_n), we first extracted all i distinct subsets of L_n, (given $i=2^n-1$), then we run equations 1 and 2 for each subset i.

In step B, the optimal number of peaks (L_n) was found at the lowest value of the Bayesian Information Criterion (*BIC*) (Schwarz, 1978). To avoid meaningless results, all selected peaks must have a positive height ($a_i > 0$).

Steps A and B are detailed below.

The analysis of $f_{(xy)}$ was performed to detect the position of potential peaks in a density plot. This step combines two search strategies:

a) Detection of global and local maxima in the $f_{(xy)}$:

$$\text{Max} f_{(xy)} = \{\mu_a, , \mu_b, , \mu_c, \ldots, , \mu_n\} \quad (3)$$

b) Detection of local minima (μ'_i) of the differential Laplacian operator of $f_{(xy)}$ ($\nabla^2 f$):

$$\text{Min } \nabla^2 f = \{\mu'_a, \mu'_b, \mu'_c, \ldots, \mu'_n\} \quad (4)$$

$\nabla^2 f$ describes the sum of the second derivative of $f_{(xy)}$ with respect to *x* and *y* (Ganzha and Vorozhtsov, 1996). It was used to detect shoulders, edges and non-evident peaks in complex surfaces (Butturini and Ejarque, 2013).

The search for maxima in $f_{(xy)}$ and minima in $\nabla^2 f$ was performed with the Nelder-Mead optimization algorithm under constrained conditions (Horst and Pardalos, 2013). The sensitivity of this algorithm can be increased or reduced by modifying selected parameters (namely: the contraction ratio, the expansion ratio, the reflection ratio and the shrink ratio). In our application, we used the standard values for these parameters (0.5, 2, 1, and 0.5 respectively) (Nelder et al., 1964) as they guaranteed an exhaustive search of main local minima in $\nabla^2 f$ into a relatively short computational time. The $\nabla^2 f$ operator is sensible to edges. Therefore, the minimum in $\nabla^2 f$ surface found in the proximity perimeter of the cytogram were omitted. Once Max $f_{(xy)}$ and Min $\nabla^2 f$ were obtained, results were joined to sort all distinct coordinates that appear in the two lists:

$$L_n = \text{Max } f_{(xy)} \text{ U Min } \nabla^2 f \quad (5)$$

In which L_n is the list of the putative *n* peaks in $f_{(x,y)}$.

In complex surfaces such those of cytograms, the Nelder-Mead algorithm can be trapped in local minima (or maxima), which are very close to each other and presumably identify the

same peak. From a statistical perspective, it was assumed that these neighbour peaks fall into the same cluster. In this case, it is necessary to merge them into a single coordinate. The search for clusters was performed according to the *fixed radius near neighbour* approach (Bentley et al., 1977). At each detected coordinate (μ_i), a circular influence area (IA_i) of radius R is associated ($IA_i = \pi R^2$), centred at the point μ_i. The value of the radius R was the same for all detected μ_i and fixed to set the IA value to 7.5% of the planar area of the surface matrix. The coordinates within the area IA_i of μ_i were automatically grouped into the same cluster. Two criteria were established to assign a coordinate to each cluster:

Criterion # 1 (applicable for the equations 3 and 5): the coordinate with the highest maxima was selected, the rest was discarded.

Criterion # 2 (applicable for the equation 4): the coordinate with the lowest $\nabla^2 f$ was selected, the rest was discarded.

Each cytogram was converted into an n x n array (bins of 5% of width) to obtain the surface $f_{(x,y)}$ and its $\nabla^2 f$. When ignoring *a priori* the number cytometric subgroups, Step A identifies a set of n potential peaks and their coordinates (L_n). In order to obtain the optimal number of subgroups, we adopted the Bayesian Information Criterion (*BIC*) descriptor:

$$BIC_i = -2 \ln (ML_i) + k_i \ln (O) \qquad (6)$$

Where ML_i is the maximized likelihood of the model associated to the subset i, k_i the number of input parameters (i.e. number of element in the subset i), and O the sample size. The model with the smallest *BIC* value was selected as the most representative one (Schwarz, 1978). This procedure first requires the identification of all i distinct subsets of L_n:

$$P(L_n) = \{\{\mu_1\}_1, \{\mu_2\}_2, \{\mu_3\}_3, \{\mu_1, \mu_2\}_4 \ldots\ldots \{\mu_1, \mu_2, \mu_{3,\ldots,} \mu_n\}_i\} \qquad (7)$$

Where $i=2^n-1$

Successively, FDM (see equations 1 and 2) is run for each *i* subset, and the optimal model (i.e., the one with the optimal number of subgroups) was selected by relying on the lowest *BIC* value.

Figures 2 and 3 describes the entire process for the cytogram T6. In this cytogram the Nelder-Mead algorithm detected 3 local maxima in the $f_{(x,y)}$ (Fig. 1a) and 7 local minima in the $\nabla^2 f$ (Fig. 1b). According to eq. 5, L_n represented a list of 10 potential peaks (n=10) and $2^{10}-1$ = 1023 subsets were generated (Fig. 2). The model with the lowest *BIC* values was the one with 8 peaks (*BIC*=5469, r^2=0.978, Fig. 2). The model output with the higher r^2 (0.979) was discarded because of the higher *BIC* value (5522). This process was executed for all cytograms in the data set.

A hyper-scatterplot was created by including all *BIC* selected peaks (μ_i) which were retrieved from the bivariate cytograms in the dataset.

The Voronoi diagram tessellation approach was adopted to cluster all *BIC* selected peaks into adjacent polygons with boundaries outlined by the Delaunay triangulation algorithm (Aurenhammer and Klein, 2000). All events that lie within a polygon are assigned to the centre of that polygon, and the number of events lying within each polygon was converted into cell concentration values.

Statistical analyses

A hierarchical clustering produces, based on Ward's method and Euclidean distance, was used to show how sampling points were clustered according to percentages of cytometric groups over the total events in the cytogram. The overall significance of such difference was tested by the non-parametric Kolmogorov-Smirnov test to give information if the densities of

different cytometric groups change differently along the river system (difference in mean and if this depend on site). The multi-group SIMilarity PERcentage test (SIMPER), using the Bray-Curtis similarity measure (multiplied by 100), was run to assess which cytometric clusters were primarily responsible for the observe difference between groups of sample and the average dissimilarity among sites and (Clarke, 1993).

Results and discussion

In line with the consolidated approach for analyzing planktonic prokaryotes across a wide range of natural and engineered water systems (Bouvier et al., 2007), two major cytometric populations, namely cells with low nucleic acid content (LNA cells) and cells with high nucleic acid content (HNA cells), could be discriminated from our river continuum samples without further data processing and counted by using fixed polygons. As expected, HNA cells were relatively brighter in fluorescence and bigger in size than LNA cells (Fig. 3). In a previous study, drinking waters were distinguishable from one another based on the percentage of HNA cells and the direct comparison of their green fluorescence histograms (Prest et al., 2013).

Our methodological approach allowed deconvolving five major recurrent subgroups within either LNA (clusters 1-4 and 10) or HNA (clusters 5-9) cytometric populations. Samples form T2 and T4 retained the lowest and larger number of subgroups, respectively (2 LNA+ 4 HNA and 4 LNA + 5 HNA). The identified peaks within the Voronoi tessellation poligons, applied for cell counting, are shown in fig. 4.

The dynamic of each identified cytometric subgroup and the evolution of the microbial community structure as a whole were assessable along the hydrologic continuum at such finer scale analysis level to provide a basis for a variety of existing statistical analysis. Hierarchical clustering offered an indication of the cytometric community similarities among sampling sites (Fig. 5). The cytometric community profiles were cross-compared (Table 1) and the

276 community changes over-imposed by an external water input (i.e., the outlet of a wastewater
277 treatment plant) were statistically endorsed. Moreover, clusters #3 and #8, belonging to LNA
278 and HNA respectively, were recognized as those groups mostly contributing to the overall
279 average dissimilarity among sites (i.e., SIMPER test). In this study, the cytometric fingerprint
280 appeared sensitive to detect the complex microbial community dynamics of flowing waters,
281 since expected changes due to tributary inputs were clearly detected.
282 Deconvolution models represent a way to deal with the complexity of water samples from
283 natural and engineered water systems, allowing for an understanding of microbial interactions
284 and structuring dynamics not described by the traditional approaches (Koch et al., 2013a).
285 A key perspective of the proposed deconvolution approach is the ability to discern recurrent
286 subgroups of cells within complex mixtures, without an a priori knowledge of which cells are
287 which. It is noteworthy that a further advanced step of cell sorting can be potentially
288 performed to provide gate-specific phylogenetic information according to selected
289 fluorescence properties and phenotypes of major cytometric populations (Schattenhofer et al.,
290 2011; Vila-Costa et al., 2012).
291 Thousands of particles and microbial cells within a wide size range (from virus-like particles
292 to prokaryotes and small protists) can be analyzed per second by flow cytometry, thus
293 providing a direct quantification of their abundance and morphological traits within minutes
294 from sampling (Van Nevel et al., 2013). A prototype machine for the automatic and
295 programmable staining of aquatic bacteria was successfully tested on-line and in real time to
296 monitor the quality of drinking water at the household tap (Besmer et al., 2014).
297 Owing to high versatility and potential for rapid analysis of large numbers of cells
298 individually, specific benefits of flow cytometry will also accrue from novel bioinformatics
299 and statistical approaches to analyze the multiparametric dataset, thus leading to significant
300 and promising technological advancements in innovating the field of real time control of
301 water quality.

302

303 **Conclusions**

304 The application of flow cytometry to freshwater samples collected along a river continuum
305 allowed discriminating diverse and recurrent subgroups of aquatic microorganisms by their
306 constitutive traits at the single-cell level. Our data suggest that a flow cytometric approach
307 could be suitable to detect changes of single-cell subgroups, thus serving as a candidate tool
308 for water quality assessments in complex environmental settings.

309

310 **Acknowledgments**

311 This research was partly funded by the Spanish Ministry of Education and Science (MEC)
312 through the project FLUMED-HOTSPOTS (CGL2011-30151-C02).

313

314

315 **References**

316 Aghaeepour, N., Finak, G., FlowCAP Consortium, DREAM Consortium, Hoos, H.,
317 Mosmann, T. R., et al. (2013). Critical assessment of automated flow cytometry data
318 analysis techniques. *Nat. Methods* 10, 228–238. doi:10.1038/nmeth.2365.

319 Amalfitano, S., Del Bon, A., Zoppini, A., Ghergo, S., Fazi, S., Parrone, D., et al. (2014).
320 Groundwater geochemistry and microbial community structure in the aquifer transition
321 from volcanic to alluvial areas. *Water Res.* 65, 384–394.
322 doi:10.1016/j.watres.2014.08.004.

323 Andreatta, S., Wallinger, M. M., Piera, J., Catalan, J., Psenner, R., Hofer, J. S., et al. (2004).
324 Tools for discrimination and analysis of lake bacterioplankton subgroups measured by
325 flow cytometry in a high-resolution depth profile. *Aquat. Microb. Ecol.* 36, 107–115.
326 Available at: http://www.int-res.com/abstracts/ame/v36/n2/p107-115/.

327 Aurenhammer, F., and Klein, R. (2000). Voronoi diagrams. *Handb. Comput. Geom.* 5, 201–
328 290.

329 Bentley, J. L., Stanat, D. F., and Williams, E. H. (1977). The complexity of finding fixed-
330 radius near neighbors. *Inf. Process. Lett.* 6, 209–212. doi:http://dx.doi.org/10.1016/0020-
331 0190(77)90070-9.

332 Besmer, M. D., Weissbrodt, D. G., Kratochvil, B. E., Sigrist, J. A., Weyland, M. S., and
333 Hammes, F. (2014). The feasibility of automated online flow cytometry for In-situ
334 monitoring of microbial dynamics in aquatic ecosystems. *Front. Microbiol.* 5, 1–12.
335 doi:10.3389/fmicb.2014.00265.

336 Boi, P., Amalfitano, S., Manti, A., Semprucci, F., Sisti, D., Rocchi, M. B., et al. (2016).
337 Strategies for Water Quality Assessment: A Multiparametric Analysis of
338 Microbiological Changes in River Waters. *River Res. Appl.* 32, 490–500.
339 doi:10.1002/rra.2872.

340 Bouvier, T., Del Giorgio, P. A., and Gasol, J. M. (2007). A comparative study of the

341 cytometric characteristics of High and Low nucleic-acid bacterioplankton cells from
342 different aquatic ecosystems. *Environ. Microbiol.* 9, 2050–2066. doi:10.1111/j.1462-
343 2920.2007.01321.x.
344 Butturini, A., and Ejarque, E. (2013). Technical Note: Dissolved organic matter fluorescence -
345 A finite mixture approach to deconvolve excitation-emission matrices. *Biogeosciences*
346 10, 5875–5887. doi:10.5194/bg-10-5875-2013.
347 Butturini, A., Guarch, A., Roman??, A. M., Freixa, A., Amalfitano, S., Fazi, S., et al. (2016).
348 Hydrological conditions control in situ DOM retention and release along a
349 Mediterranean river. *Water Res.* 99, 33–45. doi:10.1016/j.watres.2016.04.036.
350 Chattopadhyay, P. K., and Roederer, M. (2012). Cytometry : Today ' s technology and
351 tomorrow ' s horizons. 57, 251–258. doi:10.1016/j.ymeth.2012.02.009.
352 Clarke, K. R. (1993). Non-parametric multivariate analyses of changes in community
353 structure. *Aust. J. Ecol.* 18, 117–143. doi:10.1111/j.1442-9993.1993.tb00438.x.
354 Davey, H. M., and Davey, C. L. (2011). "Multivariate Data Analysis Methods for the
355 Interpretation of Microbial Flow Cytometric Data," in *High Resolution Microbial Single*
356 *Cell Analytics*, eds. S. Müller and T. Bley (Berlin, Heidelberg: Springer Berlin
357 Heidelberg), 183–209. doi:10.1007/10_2010_80.
358 Freixa, A., Ejarque, E., Crognale, S., Amalfitano, S., Fazi, S., Butturini, A., et al. (2016).
359 Sediment microbial communities rely on different dissolved organic matter sources
360 along a Mediterranean river continuum. *Limnol. Oceanogr.* doi:10.1002/lno.10308.
361 Fruhwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models (Springer*
362 *Series in Statistics)*. 1st ed. Springer Available at:
363 http://gen.lib.rus.ec/book/index.php?md5=D7BBD1C1E3558E77F042CB2FB5AF337E.
364 Ganzha, V. G., and Vorozhtsov, E. V (1996). *Numerical Solutions for Partial Differential*
365 *Equations: Problem Solving Using Mathematica*. Taylor & Francis Available at:
366 https://books.google.it/books?id=W8o0pI5l-WwC.

367 Gasol, J. M., and Moran, X. A. G. (2015). Flow Cytometric Determination of Microbial
368 Abundances and Its Use to Obtain Indices of Community Structure and Relative
369 Activity. *Hydrocarb. Lipid Microbiol. Protoc. - Springer Protoc. Handbooks*, 1–29.
370 doi:10.1007/8623.
371 Glüge, S., Pomati, F., Albert, C., Kauf, P., and Ott, T. (2014). The challenge of clustering
372 flow cytometry data from phytoplankton in lakes. *Nonlinear Dyn. Electron. Syst.*
373 *Commun. Comput. Inf. Sci. (volume 438)*, 379–386. doi:10.1007/978-3-319-08672-
374 9_45.
375 Grabowski, R.C.; Gurnell, A. M. (2016). Hydrogeomorphology- Ecology Interactions in
376 River Systems. *River Res. Appl.* 22, 1085–1095. doi:10.1002/rra.
377 Hahne, F., LeMeur, N., Brinkman, R. R., Ellis, B., Haaland, P., Sarkar, D., et al. (2009).
378 flowCore: a Bioconductor package for high throughput flow cytometry. *BMC*
379 *Bioinformatics* 10, 106. doi:10.1186/1471-2105-10-106.
380 Horst, R., and Pardalos, P. M. (2013). *Handbook of Global Optimization*. Springer US
381 Available at: https://books.google.it/books?id=yBDaBwAAQBAJ.
382 Hsiao, C., Liu, M., Stanton, R., Mcgee, M., Qian, Y., and Scheuermann, R. H. (2016).
383 Mapping cell populations in flow cytometry data for cross-sample comparison using the
384 Friedman-Rafsky test statistic as a distance measure. *Cytom. Part A* 89, 71–88.
385 doi:10.1002/cyto.a.22735.
386 Hyrkas, J., Clayton, S., Ribalet, F., Halperin, D., Virginia Armbrust, E., and Howe, B. (2015).
387 Scalable clustering algorithms for continuous environmental flow cytometry.
388 *Bioinformatics* 32, 417–423. doi:10.1093/bioinformatics/btv594.
389 Kato, T., Omachi, S., and Aso, H. (2002). "Asymmetric Gaussian and Its Application to
390 Pattern Recognition," in *Structural, Syntactic, and Statistical Pattern Recognition: Joint*
391 *IAPR International Workshops SSPR 2002 and SPR 2002 Windsor, Ontario, Canada,*
392 *August 6--9, 2002 Proceedings*, eds. T. Caelli, A. Amin, R. P. W. Duin, D. de Ridder,

393 and M. Kamel (Berlin, Heidelberg: Springer Berlin Heidelberg), 405–413.
394 doi:10.1007/3-540-70659-3_42.
395 Koch, C., Fetzer, I., Schmidt, T., Harms, H., and Müller, S. (2013a). Monitoring functions in
396 managed microbial systems by cytometric bar coding. *Environ. Sci. Technol.* 47, 1753–
397 1760. doi:10.1021/es3041048.
398 Koch, C., Günther, S., Desta, A. F., Hübschmann, T., and Müller, S. (2013b). Cytometric
399 fingerprinting for analyzing microbial intracommunity structure variation and identifying
400 subcommunity function. *Nat. Protoc.* 8, 190–202. doi:10.1038/nprot.2012.149.
401 Koch, C., Harms, H., and Müller, S. (2014). Dynamics in the microbial cytome-single cell
402 analytics in natural systems. *Curr. Opin. Biotechnol.* 27, 134–141.
403 doi:10.1016/j.copbio.2014.01.011.
404 Li, W. K. W. (2002). Macroecological patterns of phytoplankton in th enorthwestern North
405 Atlantic Ocean. *Nature* 419, 154–157. doi:10.1038/nature00983.1.
406 Le Meur, N. (2013). Computational methods for evaluation of cell-based data assessment-
407 Bioconductor. *Curr. Opin. Biotechnol.* 24, 105–111. doi:10.1016/j.copbio.2012.09.003.
408 Mittag, A., and Tárnok, A. (2009). Basics of standardization and calibration in cytometry - A
409 review. *J. Biophotonics* 2, 470–481. doi:10.1002/jbio.200910033.
410 Nelder, J. A., Mead, R., Nelder, B. J. a, and Mead, R. (1964). A simplex method for function
411 minimization. *Comput. J.* 7, 308–313. doi:10.1093/comjnl/7.4.308.
412 Nelson, C. E., Sadro, S., and Melack, J. M. (2009). Contrasting the influences of stream
413 inputs and landscape position on bacterioplankton community structure and dissolved
414 organic matter composition in high-elevation lake chains. *Limnol. Oceanogr.* 54, 1292–
415 1305. doi:10.4319/lo.2009.54.4.1292.
416 Van Nevel, S., Koetzsch, S., Weilenmann, H. U., Boon, N., and Hammes, F. (2013). Routine
417 bacterial analysis with automated flow cytometry. *J. Microbiol. Methods* 94, 73–76.
418 doi:10.1016/j.mimet.2013.05.007.

419 Niño-García, J. P., Ruiz-González, C., and del Giorgio, P. A. (2016). Interactions between
420 hydrology and water chemistry shape bacterioplankton biogeography across boreal
421 freshwater networks. *ISME J.*, 1–12. doi:10.1038/ismej.2015.226.
422 Perfetto, S. P., Ambrozak, D., Nguyen, R., Chattopadhyay, P., and Roederer, M. (2006).
423 Quality assurance for polychromatic flow cytometry. *Nat. Protoc.* 1, 1522–1530.
424 doi:10.1038/nprot.2006.250.
425 Prest, E. I., Hammes, F., K??tzsch, S., van Loosdrecht, M. C. M., and Vrouwenvelder, J. S.
426 (2013). Monitoring microbiological changes in drinking water systems using a fast and
427 reproducible flow cytometric method. *Water Res.* 47, 7131–7142.
428 doi:10.1016/j.watres.2013.07.051.
429 Pyne, S., Hu, X., Wang, K., Rossin, E., Lin, T.-I., Maier, L. M., et al. (2009). Automated
430 high-dimensional flow cytometric data analysis. *Proc. Natl. Acad. Sci. U. S. A.* 106,
431 8519–8524. doi:10.1073/pnas.0903028106.
432 Rogers, W. T., and Holyst, H. A. (2009). FlowFP: A Bioconductor Package for Fingerprinting
433 Flow Cytometric Data. *Adv. Bioinformatics* 2009, 193947. doi:10.1155/2009/193947.
434 De Roy, K., Clement, L., Thas, O., Wang, Y., and Boon, N. (2012). Flow cytometry for fast
435 microbial community fingerprinting. *Water Res.* 46, 907–919.
436 doi:10.1016/j.watres.2011.11.076.
437 Schattenhofer, M., Wulf, J., Kostadinov, I., Glockner, F. O., Zubkov, M. V., and Fuchs, B. M.
438 (2011). Phylogenetic characterisation of picoplanktonic populations with high and low
439 nucleic acid content in the North Atlantic Ocean. *Syst. Appl. Microbiol.* 34, 470–475.
440 doi:10.1016/j.syapm.2011.01.008.
441 Schiaffino, M. R., Gasol, J. M., Izaguirre, I., and Unrein, F. (2013). Picoplankton abundance
442 and cytometric group diversity along a trophic and latitudinal lake gradient. *Aquat.*
443 *Microb. Ecol.* 68, 231–250. doi:10.3354/ame01612.
444 Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Stat.* 6, 461–464.

doi:10.1214/aos/1176344136.

Shade, A., Peter, H., Allison, S. D., Baho, D. L., Berga, M., B??rgmann, H., et al. (2012). Fundamentals of microbial community resistance and resilience. *Front. Microbiol.* 3, 1–19. doi:10.3389/fmicb.2012.00417.

Shapiro, H. M. (2005). Using Flow Cytometry. *Pract. Flow Cytom.*, 736. doi:10.1002/0471722731.fmatter.

Verschoor, C. P., Lelic, A., Bramson, J. L., and Bowdish, D. M. E. (2015). An introduction to automated flow cytometry gating tools and their implementation. *Front. Immunol.* 6, 1–9. doi:10.3389/fimmu.2015.00380.

Vila-Costa, M., Gasol, J. M., Sharma, S., and Moran, M. A. (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. doi:10.1111/j.1462-2920.2012.02720.x.

Van Wambeke, F., Catala, P., Pujo-Pay, M., and Lebaron, P. (2011). Vertical and longitudinal gradients in HNA-LNA cell abundances and cytometric characteristics in the Mediterranean Sea. *Biogeosciences* 8, 1853–1863. doi:10.5194/bg-8-1853-2011.

462 **Table 1**.

463 Average dissimilarity between sites computed by SIMPER test (lower part of the matrix), and
464 related statistical diversity assessed by the non-parametric Kolmogorov-Smirnov test (upper
465 part of the matrix).

466

SIMPER \ KS p	T1	T2	T3	T4	T5	T6	A1
T1	-	0.975	0.313	0.975	0.313	0.031	0.000
T2	16.4	-	0.975	0.675	0.675	0.031	0.000
T3	23.3	8.2	-	0.313	0.675	0.031	0.000
T4	12.0	20.0	23.3	-	0.111	0.007	0.000
T5	29.4	15.2	7.8	29.6	-	0.031	0.000
T6	62.2	52.2	46.9	63.0	42.2	-	0.002
A1	95.4	93.9	93.0	95.5	92.1	81.7	-

467

468

469 **Figure legends**

470 **Figure 1**.

471 Quantile contour plot ($f_{(x,y)}$) of the model cytogram (T6) used to describe the deconvolution
472 process (a) and its associate Laplacian ($\nabla^2 f$) (b)**.** Black arrows indicate the position of the
473 local maxima in $f_{(x,y)}$. Red arrows indicate the position of the local minima in ∇^2f . According
474 to eq. 5, ten relevant peaks were detected in this cytogram from sample T6 (see fig. 3).

475

476 **Figure 2**. Visual example of the optimal model selection process. This example refers to the
477 sample T6 with ten potential peaks (n=10) (see fig 3). Panel *a* shows the relationship between
478 BIC values and number of peaks obtained executing the FDM $z_{(x,y)}$ (eqs. 1 and 2) for all
479 possible subsets i of the ten potential peaks, where $i=2^{10}-1=1023$. Gray disks and black dots
480 discern modeled cytograms adjust with r^2 lower and higher than 0.95 respectively. Panel *b*
481 shows the contour plots of four modelled cytograms with a "poor" adjust (*i*); a good not
482 "optimal" adjust (*ii*); the "optimal" adjust (i.e., lower BIC values, *iii*) and overfitted adjust
483 (i.e. larger number of peaks, *iv*). Large white and small black dots in contours plots show
484 location of potential and selected peaks respectively.

485

486 **Figure 3**.

487 Representative cytograms of freshwaters sampled form the upstream area of the River
488 Tordera (Barcelona, Spain). Curved arrows indicate the directional connections between
489 sampling sites along the hydrologic continuum. The green fluorescent signals (Sybr Green I)
490 were used to discriminate two major populations of cells with low and high content of nucleic
491 acids (LNA and HNA, respectively). Small and large sized cells were distinguished according
492 to forward scatter signals.

493

494 **Figure 4**. The Voronoi tessellation mask was calculated considering all recurrent peaks and
495 applied back to each cytogram. The number of events lying within each polygon was
496 converted into cell concentration values.
497
498 **Figure 5**.
499 Microbial community structure as assessed by the proposed deconvolution model. a) Cell
500 abundance within each polygon identified by the Voronoi tessellation mask. b) Relative
501 percentages and subgroup distribution within the LNA and HNA cytometric populations.
502 Subgroups were ordered according to their average green fluorescence. c) Hydrologically
503 connected samples were joined according to their cytometric profile by the Ward's clustering
504 method such that increase in within-group variance is minimized.
505

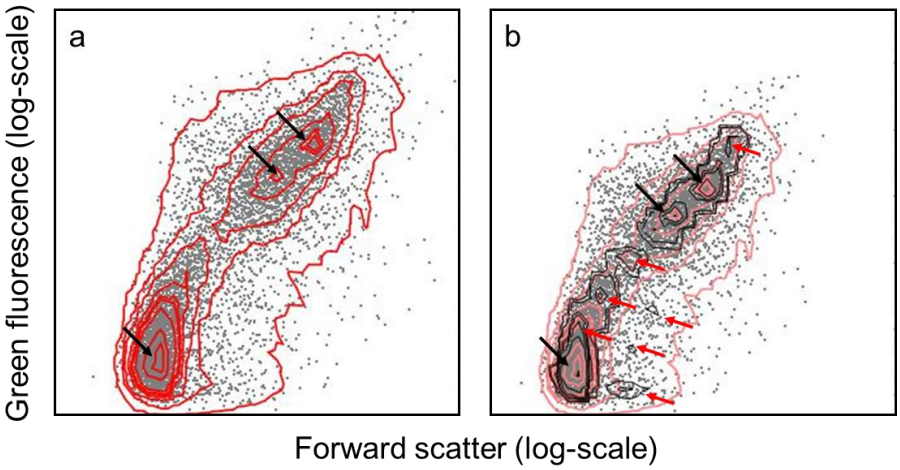


Figure 1.

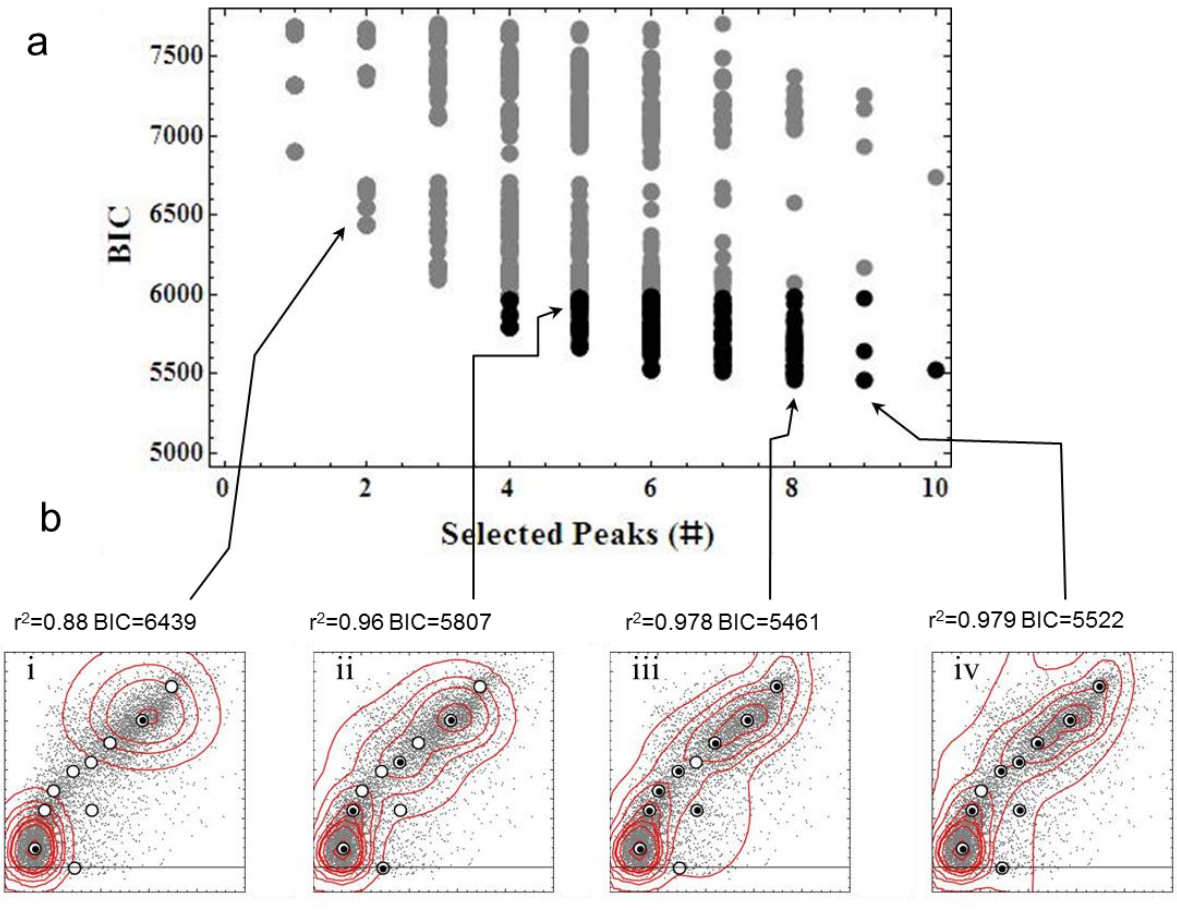
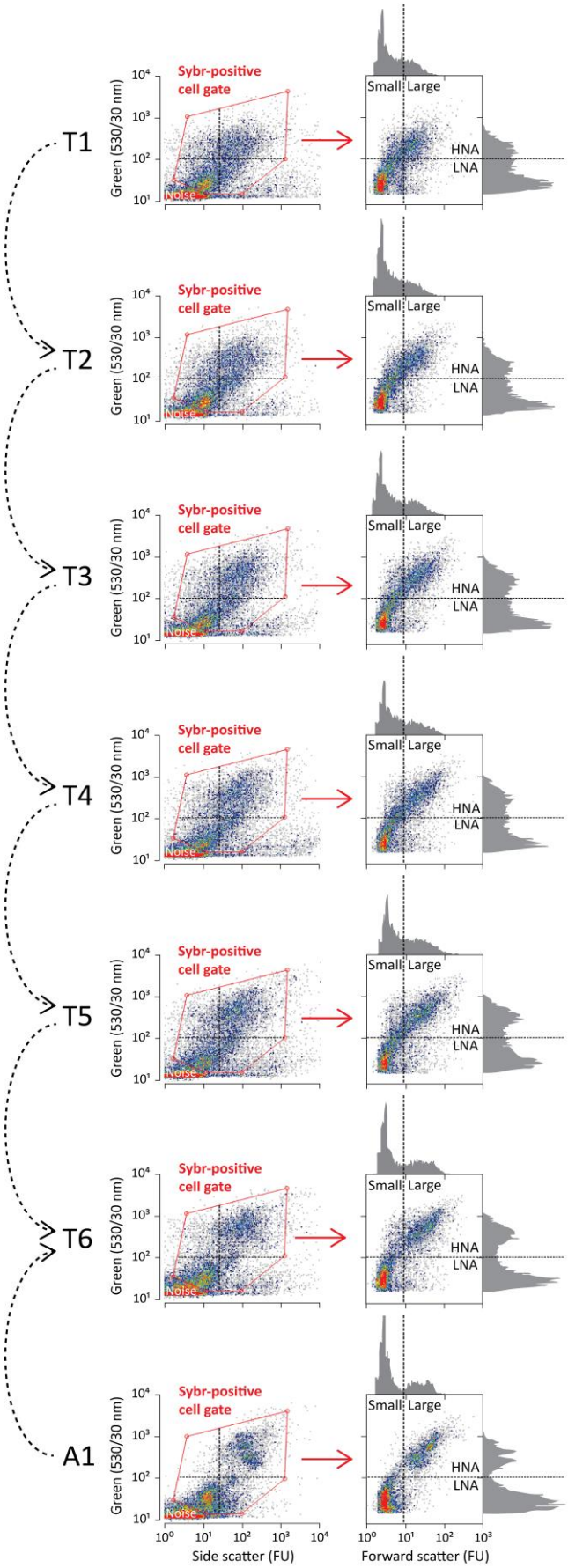
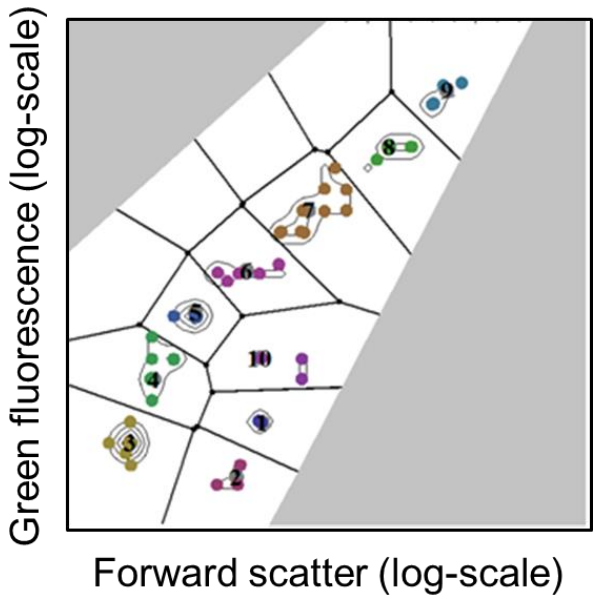


Figure 2.



513

514 **Figure 3**.

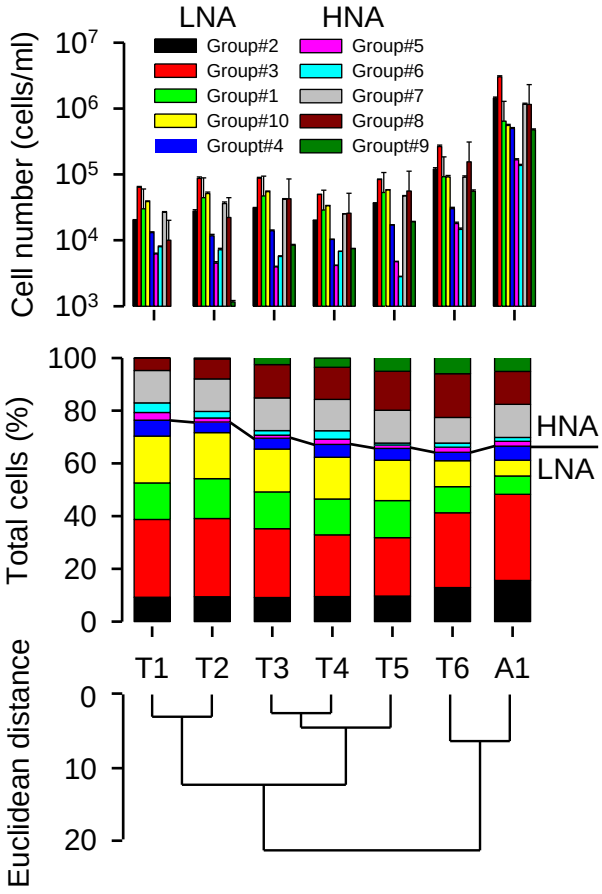


515

516 **Figure 4**.

517

518



519

520 **Figure 5**.