Flow cytometry (FCM) is a powerful analytical tool that is widely used worldwide, as it allows the depiction of the innate complexity of a vast range of biological systems in few seconds. It is a technique based on the spectroscopic properties of suspended particles that allows data to be graphically summarized by biplots, known as cytograms. Such versatility got raises to different analytical protocols which are commonly not interchangeable among expertise fields. In this sense, environmental sciences, in particular, faces major concerns when dealing with the adoption of non-specific protocols - a particularity essentially driven by the highly heterogeneous nature of environmental samples. Such intrinsic variety makes it difficult to adjust formal analytical protocols that both keep standardized mathematical rationales and retain a clear ecological meaning, namely when the focus of the analysis rely on the cytometric diversity - the quantitative evaluation of the differences among cytograms. Despite of the availability of promising tools conceived or adapted to approach cytometric diversity, most of them face common technical challenges, as perspective adjustment, dilution correction, resolution setup and enlightenment on the role of cytograms subregions to global diversity. To address such questions and harmonize formal mathematical rationales with coherent biological interpretation, we have developed flowDiv - a pipeline designed for environmental flow cytometry data analysis that handles data through consolidated macroecological methods to offer biologically apprehensive outputs. flowDiv was implemented using R language and has been published on CRAN (https://cran.r-project.org/web/packages/flowDiv/) with source code also available on GitHub (https://github.com/bmsw/flowDiv). Applied to a dataset from 31 freshwater bodies in Argentinian Patagonia, flowDiv uncovered significant aspects regrading environmental cytometric diversity, as its relation with taxonomic diversity and the role of environmental variables on cytometric diversity.
**Introduction**

Flow cytometry (FCM) is an extremely versatile analytical technique that has been broadly and successfully applied in multiple fields, from industries to medical and environmental researches\[1, 2, 3\]. Fundamentally, it is based upon the spectroscopic study of hydrodynamically focused particles: oriented by a stream of fluids, suspended particles are systematically and individually passed across lasers beams to have their optical properties depicted. Once acquired, this characteristics are stored in feature matrices from which graphical outputs are generated and used to conduct data exploration, which are generally two dimensions dot-plots known as cytograms, where each dot represents a particle (cell or debris).

The powerfulness of FCM is supported by its level of detailed descriptions (up to 20 optical features) of tens of thousands (up to 40,000) of particles per second, a feature that propitiate a fast and robust way of accessing the quali-quantitative properties of a specific biological system. Naturally, such a versatility resulted in a vast range of applications and as a consequence, conceiving a vast number of different FCM analytical protocols, each one focused on very specific features of the biological system under study, and commonly not interchangeable.

In this sense, environmental sciences, in particular, faces major obstacles when dealing with the adoption of non-specific protocols - a particularity essentially driven by the highly heterogeneous nature of environmental samples\[4, 5\]. Notwithstanding, this heterogeneity plays a central role on ecology, as it tells much about the morphophysiological features of a particular population or microbial community \[4, 5\]. For this reason, many efforts on FCM applications have been directed to the development and adequacy of methods focused on the depiction of cytograms resemblances - a notion currently explored under the closely related concepts of cytometric pattern \[6\], cytometric fingerprint \[6\] and cytometric diversity \[7, 8\].

Worth to note that cytometric resemblance studies have been driven commendable efforts both with respect to implementation \[9, 10, 11, 12\] and its critical assessment \[6\], but there are some crucial aspects on the available approaches that are still moot, as the most suitable model to handle FCM environmental data and the exact biological meaning of the analysis' outputs. In this sense, one could reasonably argue that a suitable choice would be for the procedure that rightly balances the necessary mathematical formalism with a clear biological interpretation - or, in other words, one could choose for the method that properly applies metrics commonly used in ecology (as they commonly fit both requirements \[13\]).

Notably, although available tools do incorporate, in some sense, ecological methods to handle FCM data, the perception of an explicit application of such methods to access cytometric similarity remains underexploited. Intriguingly, since this paradigm was pioneered and successfully applied by Li, under the name of "cytometric diversity" more than twenty years ago \[7\], only few works have been dedicated to delve into this line \[14, 15, 8, 16\], notwithstanding its astounding implications.
In his work, Li conceived an elegant approach to access cytometric diversity [7]: it consisted on comparing the ecological properties of minor pieces of data (i.e. bins), by applying a 16x16 cartesian grid to 2-dimensions cytograms and, in sequence, constructing a 2-D contingency table from it. Each entry of the contingency table represented a bin, which were then abstracted as special cases of taxonomic units (in a pure pragmatic analogy to a biological species but totally regardless their phylogeny or functional properties) and further managed according to some ecological method - in the seminal case, evenness and alpha diversity.

Albeit a very promising approach, there were some important methodological gaps that kept unclear, namely: i) the issue of low dimensionality; ii) resolution ; iii) throughput; iv) data resemblance; and v) bins’ explicit role on cytometric diversity.

The issue of the low dimensionality refers to the impossibility of analyzing more than two channels at once - although suffice in many scenarios [14], this select subset of information of two channels per run impairs a deeper scrutiny of data as it does allow an efficient control on the effect of additional features and, therefore, fails to resolve more efficiently for multicolor assays.

Resolution, on the other hand, deals with making a formal definition of the suitable number of bins, prior to analysis of a cytogram. Ideally one could work with the highest possible image resolution, limited only to the dynamic range of flow cytometry assay, as to achieve maximum discriminative power for analysis. Such choice, however, has the potential to lead to computation overhead issues, notably when dealing with big data sets. A reasonable balance point would be to narrow image resolution to limits where there would be preservation of the highest amount of data information as to maintain computation less intensive - although self-evident, that denouement holds no closed-form approaches yet.

Throughput mainly concerns on the integration of differently acquired datasets: for any scenario, a proper comparison amongst different cytograms requires data to be settle to common perspectives in order to address the events of interest correctly, a constraint that demands all files to be acquired under the same machine settings or, more usually, embedded some sort of common standards (i.e. internal standards such as reference beads) [17]. From a computational point of view, different machine settings output strictly different images (cytograms), even if standards are used. Hence, comparing cytograms would be quite impaired (almost impossible) unless some control of perspective is applied to data prior to analysis, enabling a proper management of differently acquired assays.

Additionnally, there are two closely connected aspects, present but not explicitly covered in Li’s work: data resemblance and bins’ explicit role on cytometric diversity. Data resemblance concerns originate from the fact that, once individual cytograms could be depicted by its alpha diversity and evenness, it could reasonably be possible to access their resemblance by intertwined metrics, specifically $\beta$-diversity. In fact, alpha diversity is just one aspect of a triad of interrelated phenomena, $\alpha$, $\beta$ and $\gamma$ diversities, that holds, in its fundamental form, a pretty simple, but profound, relation that can be mathematically expressed either in an additive ($\gamma = \alpha + \beta$) or multiplicative ($\gamma = \alpha \cdot \beta$) and used to access not only...
the individual properties of a system (as for $\alpha$ diversity) but how these properties might account for differences between those same systems (i.e. $\beta$ diversity) when contemplated under a global (pooled) scenario ($\gamma$ diversity).

Therefore, as an immediate adaptation of traditional community ecology concepts, the idea of data resemblance raises questions about bins’ direct contribution to cytometric diversity: how could exactly bins lead to differences between cytograms? And more importantly: how could bins correlate to analyst’s prior defined cytometric populations? These are very crucial questions that were not explicitly addressed in Li’s seminal work and without which, diversity measures gives limited information [18].

In this article, we address solutions for these fundamental questions as we debate on the implementation of flowDiv - a pipeline for analyzing environmental flow cytometry data using multidimensional contingency tables, devised as an improvement and extension of Li’s ideas. In the final section, we illustrate one use case successfully built on that approach.

Design and Implementation

flowDiv is implemented in R language and is structured in nineteen stages of processing and eleven stages of oriented decision (Fig. 1). Here we describe the rationale behind each stage in detail.

Data Read

First step of the pipeline consists of reading and parsing preprocessed (i.e. compensated, normalized or transformed) [19] FCS data. Input may be structured either as FlowJo® workspaces or, equivalently, GatingSet R objects.

This process is wrapper for some flowWorkspace [20] and flowCore [21] subroutines and is meant to reduce the complexity of the global analysis by reducing the number of required softwares to, at most, a couple. By that, we bring the assay to a rather manageable and more reproductive execution.

Gate Selection

Once imported, next action consists on the extraction of user defined regions of interest: the gates. Gates are regions defined by either their channels and respective borders (limits) that must be supplied to the algorithm: while borders are internally and automatically parsed, information about which channels to use should be defined, on the fly, by the annalist.

This step is one of the keystones of the algorithm as it expands data analysis to higher dimensions, allowing the setting of more than two channels per analysis.
Ranges Definitions

For any selected channel, a histogram with equal-sized numbers of bins is generated. This task requires two parameters to be previously outlined: channels ranges and bins width.

The ranges by which channels will be binned can be defined by both the relative maximum and minimum values of the pooled set of channels (i.e. dynamic ranges) or by setting absolute limits for each channel separately (i.e. fixed ranges).

Fixed ranges defines static frontiers for the histograms, propitiating a global model for comparative analysis between different runs of the algorithm. Dynamic ranges, on the other hand, guarantees that only the limits spanned by data are considered on the binning process, maximizing the information gain of the analysis.

Normalization

To fit specific scenarios where data embeds any control standards (beads) but are acquired under different machine settings - more specifically for scenarios where operator accounts for changes on data mean

Figure 1: Schematic view of flowDiv’s workflow.
while control for its variance - we provide an approach to settle data to a common perspective through a translational transformation of data (called, in our pipeline, normalization).

A translation is a rigid transformation in which every point is moved the same distance in the same direction as to conserve the size and the shape of the original data. Formally, for any vector $v = (a_1, a_2, ..., a_n)$, it is possible to apply a transformation $T$ on $v$ such that:

$$T(v) = (a_1 + \Delta b_1, a_2 + \Delta b_2, ..., a_n + \Delta b_n)$$  \hspace{1cm} (1)

Where $b = (\Delta b_1, \Delta b_2, ..., \Delta b_n)$ represents the coordinates of displacement for each point.

In our context, $v$ represents the channels vectors of a particular cytogram and $b$ is the vector of the difference computed between beads mean values of each channel and a grand mean, calculated from the pooled beads mean values for each channel of all cytograms on the set, such that:

$$\Delta b_{ij} = \sum_j \frac{w_{ij}}{n} - \bar{w}_{ij}$$  \hspace{1cm} (2)

Here, $\bar{w}_{ij}$ represents the arithmetic mean of beads values from channel $i$ of cytogram $j$, and $n$ corresponds to the absolute number of samples (cytograms).

Following translation, flowDiv runs the variance stabilization of data with basis on the approach implemented by Azada et al. (2015), on flowVS package [2]. Briefly, this steps proceeds to a inverse hyperbolic since (asinh) transformation of data with the form:

$$T(v_i) = asinh\left(\frac{v_i}{c_i}\right)$$  \hspace{1cm} (3)

Being $c_i$ a normalization factor calculated for each channel $i$ individually [2].

**Binning**

After ranges being defined and data centralized, the algorithm proceeds to data binning: here, analyst will be inquired about how many bins should be used in histograms constructions.

Defining a suitable number of bins is not a trivial task, though: that is a data-driven and highly subjective process that must be reasonably addressed, on the penalty of impaired analysis[22].

In view of the innate highly variability of natural environments, it is not reasonable to define a deep-seated number of bins that represents any sort of data; conversely, binning should be changeable, attending the nature of data at hand. To deal with this, we have implemented a subroutine for inferring the best number of bins, which is based on Freedman–Diaconis rule [23]:

$$bins_{ij} = \left\lceil \frac{\max(x_{ij}) - \min(x_{ij})}{2 \cdot IQR(x_{ij}) \cdot n_j^{1/3}} \right\rceil$$  \hspace{1cm} (4)
Here $bins_{ij}$ represents the ceiling number of bins for channel $i$ of sample $j$; $n$ the number of observations for the sample $j$; $IQR$ stands for interquartile range and $x_{ij}$ represents the channel vector $i$ of sample $j$.

The best number of bins $bins_b$ is calculated simply by the arithmetic mean of all suggested bins pooled, as follows:

$$bins_b = \frac{\sum_i \sum_j bins_{ij}}{\max(i) \cdot \max(j)}$$

### Contingency Tables

Binning process results on creation of common, mutually exclusive, exhaustive and ordered classes (bins) which are then cross-tabulated and used on the construction of multi-way contingency tables.

By definition, a multi-way contingency table [24] or, equivalently, a $n$-dimensional contingency table, is a set of natural numbers $S$ such that:

$$S = \{x_{ik} \mid i = 1, 2, \ldots, m \text{ and } k = 1, 2, \ldots, n\}$$

Where $x_{ik}$ corresponds to the number of counts referring to bin $i$ of channel $k$.

### Vectorization

Every $n$-dimensional contingency table $S$ is further linearly transformed into column vectors, in a process known as vectorization, creating a one-to-one correspondence between elements of the multidimensional space and elements of its transformed form, as following:

$$V_j = vec(S_j) = \{x_{11}, \ldots, x_{i1}, \ldots, x_{1k}, \ldots, x_{ik}\}$$

The rationale behind this step is to make data more manageable for subsequent manipulation, by reducing its dimensionality while keeping its information unchanged.

### Volume Correction

There are circumstances where environmental samples are previously diluted before running a flow cytometer experiment: such dilutions may occur as a direct consequence of stain, fixative or beads solution affix or, additively, as a requirement to keep events counting within a protocol-specified range [2].

Any of these situations must be properly considered at the final calculations, with the aim to correctly determine the real abundance of any targeted event. In our pipeline, we deal with dilution bias by applying a user-defined correction factor to any individual sample, such as:
\[ F = W \cdot D_{cf} \]  \hspace{1cm} (8)

Where \( W \) is an \( n \times j \) matrix composed of all column vectors \( V_j \) and, \( D_{cf} \) is a diagonal matrix whose element \( d_{ij} \) corresponds to the ratio between the minimum true volume passed (i.e. the real volume analysed, considered after correction for dilutions of any nature) of all samples pooled and the true volume passed for sample \( j \). The choice for the minimum value, to the detriment of the maximum one, was considered as to downweight eventual background noises generated on relative long-term runnings.

### Diversity Analysis

Each element of the contingency table can be reasonably thought as special cases of operational taxonomic units (OTU’s), or, in a more straight nomenclature, an OBU - Operational Bin Unit. Hence, after vectorization, any cytogram can be abstracted as a unique species-like vector and the concatenation of such vectors generates a "site \( \times \) species"-like matrix that could be meaningfully applied for further traditional analysis in community ecology.

To make this step feasible and as adjustable as possible, we take advantage of another celebrated suite of tools provide by vegan package [25], as to access three important indicators of biological diversity: \( \alpha \)-diversity, species evenness and \( \beta \)-diversity.

On the hand of the great multitude of ways of mathematically defining such indicators [26, 27], virtually any definition could be adapted to the very recent problem of cytometric diversity of natural environments. Still, there are some guidelines that could be reliably contemplated and, as starting points, are used to define our pipeline defaults.

### Alpha Diversity

Many different indices of alpha diversity can be derived from an unifying notation developed by Hill (1973) [28], collectively know as "Hill numbers" [29]:

\[ qN = \left( \sum_{i=1}^{S} p_i q \right)^{1/q} - q \]  \hspace{1cm} (9)

In that notation, \( qN \) corresponds to diversity index of order \( q \), \( p_i \) is the proportional abundance of specie (bin, in our context) \( i \) and \( q \) is any number on the interval \( ]-\infty, +\infty[ \).

Decision about which value of \( q \) to set will rely on how species may be weighted according to its commonness: the highest the \( q \) the more common species will contribute to diversity; conversely, the lowest the \( q \) the more rare species will contribute to it.

Without any stronger argument suggesting species (bins) to be differently weighted, one may choose for an index that treats common and rare species equally. Such an index can be achieve by setting \( q = 1 \), which gives rises to exponential Shannon entropy, the default \( \alpha \)-diversity index of our pipeline:
$1^N = H = \exp(-\sum_{i=1}^{S} p_i \ln p_i)$

### Bins Evenness

From equation 9 one may see that, irrespective to the order $q$ of an specific index, the concept of diversity is mainly driven by two factors: the total number of species $S$ ("richness") and their relative proportion $p_i$ ("evenness").

For our practical purpose, such an outcome strictly means that once we get sufficient information about any cytogram, by direct accessing its diversity, we can further evaluate how that diversity is partitioned between non-zeroed-valued bins ("species") and their respective entropy ("evenness").

A rather intuitive way of accessing evenness, here called $E$, is through the ratio of the measured value of diversity, $H$ (equation 10), and the total number of species $S$[30]:

$$E = \frac{H}{S}$$

Sheldon (1969) [30], had analyzed some properties of evenness measures derived from equation 11 and suggested the following formula (aftermath known as Pielou’s evenness[31]) as one suited for general use:

$$E_{\text{pielou}} = \frac{\ln H}{\ln S}$$

Even though Sheldon’s concerns where strictly focused on very lower values of $S$ ($\leq 20$), a scenario very unlike to occur on recent assays, we maintain Pielou’s evenness as our default choice both for historical (Li’s first approach were based on that index) and practical (Pielou’s index is one of the best know evenness index amongst ecology researches) reasons.

### Beta Diversity

As with the aforementioned indices, there is an even more vast range of $\beta$-diversity indices that could be applied to cytometric diversity [32].

On respect to $\beta$-diversity, choice may be guided by the tendency many data have to generate sparse (i.e. many-zero) vector - an inherited drawback caused by our algorithm in attempting to define sharped frontiers (i.e. hypercubes) for smoother data (i.e. hyperellipsoids). Thereupon, double zeros problem may be tenacious in such approach, favoring the choice for asymmetrical metrics to the detriment of symmetrical ones, like euclidean.

Amongst most known asymmetrical distance measures, Bray-Curtis [33] semimetrics seems very appropriate as it is both suitable for raw count data and makes no a priori constraints about weights on low-valued and high-valued bins (i.e. treats them equally) [34].
Bray-Curtis distance is defined as above:

\[
d_{ij} = \frac{\sum_{b=1}^{n} | b_{ik} - b_{jk} |}{\sum_{b=1}^{n} (b_{ik} + b_{jk})}
\]  

(13)

In the equation \(d_{ij}\) is the Bray-Curtis dissimilarity between the cytograms \(i\) and \(j\), \(k\) is the index of a given bin and \(n\) is the total number of bins in a cytogram.

As for \(\alpha\)-diversity, it is possible to partition \(\beta\)-diversity into two different phenomena: nestedness and turnover [35]. In short, these two components would correspond, respectively, to differences accounted for AND and XOR relations between two sets of species (e.g. Baselga, 2009 [35]). For Bray-Curtis distance, partition is like follows [36]:

\[
d_{BC} = \frac{\min(B,C)}{A + \min(B,C)} + \frac{|B - C|}{2A + B + C} \cdot \frac{A}{A + \min(B,C)}
\]  

(14)

Where \(A = \sum_{b=1}^{n} \min(x_{ij}, x_{ik})\), \(B = \sum_{b=1}^{n} x_{ij} - \min(x_{ij}, x_{ik})\) and \(C = \sum_{b=1}^{n} x_{ik} - \min(x_{ij}, x_{ik})\). In the formula, the first term corresponds to turnover and the second one is the nestdness.

Applied to our context, these two elements serve as suitable proxies to depict how differences in cytograms might be partitioned between their superimposition ("nestedness") and differences in their bin contents ("turnover") and are also implemented in our pipeline as a wrapper of betapart:bray.part() function.

Finally, so to accommodate for other ecologically meaningful distance measures (see [37] and [25] for details), we have additionally incorporated another optional step: transformation. Internally, this process is just a wrapper for decostand\{vegan\} function.

**Ordination Analysis**

Once \(\beta\)-diversity indices are acquired, next step consists on a ordination and biplot of the results (cytograms and bins) as to provide both a suitable visual summary for analysis and a further investigation on contribution of bins for observed differences.

Similar to choice of diversity index, restrictions about which gradient analysis to use will rely on technical aspects of the experiment design, specially on the distance measure elected for ordination [13].

Non-Metric Multidimensional Scaling (nMDS) has some convenient properties, in special the potential to accommodate any (dis)similarity measure for ordination [34].

Considering its versatility and the circumstance we have selected Bray-Curtis distance as our default distance measure, nMDS is applied in our pipeline.

---

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26934v1 | CC BY 4.0 Open Access | rec: 16 May 2018, publ: 16 May 2018
Cluster Analysis on Bins Ordination

Direct interpretation of bins significance is seldom restricted to its size: the wider the bins are, the bigger the regions represented by them on the cytograms become and, consequently, less information they hold. On the other hand, increasing resolution (i.e. narrowing bins width) increment information gain but make bins difficult to interpret individually.

That standstill shall be addressed by an strategy that simultaneously permits more flexible settings on bins widths and a straightforward visual interpretation of bins on the context of any cytogram.

Clusterization of bins ordination scores based on lower dimensions (up to three; two in our implementation) is an elegant strategy for this, whereas it endows a practical purpose of keeping track of broader regions of the contingency tables while allowing further inspection of plots using traditional visual approaches.

Here we use a broadly used method for cluster analysis: $K$-means clustering.

Shortly, $K$-means clustering goal is to partition $n$ observations into $k$ mutually exclusive clusters. More formally, $K$-means aims at minimizing a squared error function $J$, such that:

$$\arg\min_{\mu} J = \arg\min_{\mu} \sum_{i=1}^{k} \sum_{j=1}^{n} \| x_{ij} - \mu_i \|^2$$

In that equation, $\| x_{ij} - \mu_i \|^2$ is the Euclidian distance between a data point $x_j$, belonging to cluster $i$, and the cluster centre $\mu_i$.

In our context, the set of observations $x = (x_1, x_2, ..., x_n)$ represents the set of 2-dimensional real vectors, defined by each one of the $n$ bins’ ordination scores, got on previous step.

Choice of $K$

Determining the ideal number of clusters $K$ is not a trivial task: unless analysts have some reasonable practical assumptions about how many clusters there should be (what could be guided by the number of cytometric populations found on pre-analysis, for example), there is no gold standard way of defining it. To solve for this, questioner should appeal to some of many criteria available to help taking such decision (cf.[38]).

Here, we adopt Calinski-Harabasz [39] criterion to guide our definition on the best number of clusters. Calinski-Harabasz criterion, $C$, is defined as:

$$C = \frac{n-K}{K-1} \cdot \frac{BG_{SS}}{WG_{SS}}$$

Here, $n$ is the number of bins, $K$ is the number of clusters, $WG_{SS}$ is the sum of squares within the clusters and $BG_{SS}$ is the sum of squares between the clusters. In our pipeline, $K$ is tested iteratively - within a pragmatically-defined range - from one to ten, and the lower $C$ is set as a suggestion of the
suitable number of clusters.

**Mapping**

Clustering of bins scores results on a vector of group membership that are used to create a single masking image, which is further required to map clusters regions onto each cytogram individually. That final step provides a novel straightforward way of visually interpreting the bins ordination directly on cytograms, hence improving analysis.

**Use Case**

**Introduction**

To evaluate flowDiv, we have analyzed bacterioplankton data from 31 freshwater bodies in Argentinian Patagonia, collected from provinces of Chubut, Santa Cruz and Tierra del Fuego. These water systems seem to be a very appropriate benchmark for our pipeline as they endow a clear geospatial gradient as well as encompass a multitude of different ecological characteristics already proved to reflect in their bacterial community structure [40, 41, 42]. In order to assess flowDiv’s consistency we have briefly contrasted it with other five available cytometric fingerprint computational tools: Dalmatian Plot [11], Cytometric Histogram Image Comparison (CHIC) [10], Cytometric Barcoding (CyBar) [12], FlowFP [9] and PhenoFlow [16].

**Material and Methods**

**Data sets**

We focused this study case on three distinct datasets, concerning each water system individually: (1) twelve environmental morphometric and physicochemical variables; (2) flow cytometry FCS files manually gated for bacterioplankton populations; and (3) bacterial polymerase chain reaction-denaturing gradient gel electrophoresis (PCR-DGGE) bands’ relative intensities. Detailed information about study sites, protocols, sampling design and environmental parameters must be referred to Schiaffino et al. [40, 41, 42].

**Environmental parameters**

Samples were collected from the euphotic zone, during springs of years 2007 (Chubut and Santa Cruz) and 2008 (Tierra del Fuego) along a latitudinal gradient from 45°55’S to 54°36’S with recording of the following parameters: latitude, longitude, area, temperature, pH, conductivity, dissolved oxygen (DO), dissolved nitrogen (DN), diffuse attenuation coefficient ($K_d$), chlorophyll a (Chla), phosphate and dissolved organic carbon (DOC).
Flow cytometric data

Flow cytometry data were acquired by a FACSCalibur (Becton Dickinson) flow cytometer equipped with a standard 15 mW blue argon-ion (488 nm emission) laser and a red laser diode (635 nm), using 1 µm fluorescent beads as internal controls and SYTO 13 as nucleic acid stain. Bacterioplankton populations were manually gated by their cytometric signature on detection channels for 90° light scatter (SSC), green fluorescence (GF) and red fluorescence (RF) (e.g. Gasol et. al 2015 [2]). Gating strategy was performed with FlowJo® v.10 software. Cytograms were binned through channels SSC-H, FL1-H and FL3-H, using 75 bins per feature.

Pipelines comparison

Pairwise comparisons were accomplished on Bray-Curtis distance matrix of all pipelines by using adaptations built on the implementations proposed by Koch et al. (2014) [6] and Props et al. (2016) [16]. All the five pipelines were benchmarked against DGGE’s Bray-Curtis distance matrix.

Statistics

All statistical analysis was performed with R version 3.3.2 (2016), using the following additional packages: vegan [25], RVAideMemoire [43], gvlma [44], corrplot [45], gplots [46] and ggplot2 [47]. Graphical summaries were created with corrplot, gplots and ggplot2.

Principal component analysis (PCA), non-metric multidimensional scaling (NMDS) and regression of environmental vectors onto NMDS ordination were based on stats::prcomp(), vegan::metaMDS() and vegan::envfit() functions.

Tests on ordination scores’ centroids were conducted with permutational multivariate analysis of variance (PERMANOVA) while controlling lakes spatial variation and analysis of multivariate homogeneity of group dispersions done with vegan::adonis() and vegan::betadisper().

All linear models, including one-way ANOVA tests, were conducted after checking for model assumptions by gvlma::gvlma(). Following this rationale, we have applied a monotonic transformation $T(x) = \ln(x)$ to our chosen alpha index (i.e. the exponential of Shannon index) as to fulfill such assumptions. Additionally, in order to correct for imbalanced factors, we have merged mesotrophic (n=13) and eutrophic (n=4) groups (cf. Schiaffino et al. (2013)[41]) into a single class, referred hereafter as "meso-eutrophic".

Distances matrices for pairwise comparison of methods and Mantel’s test for dissimilarity matrices were ran with vegan::vegdist() and vegan::mantel() based on 26 out of the original 31 lakes. Reason for this lower sample size is the fact that not all samples used in comparisons are common to all methods (cf. "Supplementary Code" above).

All permutational hypothesis tests’ pseudo-F were based on 999 replicates and the significance level $\alpha$ for all tests was set to be 0.05. Complete code for statistical analysis is available on-line and can be
Results and Discussion

Alpha Diversity and Evenness

Principal components analysis (PCA) on three cytometric indices - Shannon index, evenness and richness - revealed a smoothed separation pattern amongst samples (Fig. 2a), suggesting that differences in cytometric diversity could be explained by water body trophic status, fertilization level (increasing from oligotrophic to meso-eutrophic status) and consequent ecosystem primary productivity increase (for review see [48]) in a similar way to what is observed for taxonomic diversity of cyanobacteria (e.g. [49]). Additionally, it also indicated that differences in trophic status could be driven by cytometric richmesses rather than by cytometric evenness or Shannon cytometric diversity. To test these hypotheses, we performed a Wilcoxon rank sum test under the null hypothesis that average cytometric diversity indexes (i.e. Pielou, Shannon and Richness) are not dependent on water body trophic status. Only cytometric richmesses had support for those hypotheses, disclosing significant differences between groups.

Furthermore, cytometric alpha diversity was proved to be significantly more influenced by richness than by evenness (Fig. 3), notably obeying a different pattern as observed by Li (1997), who observed a significant correlation not with richness correlation but with evenness [7]. Additionally, Spearman’s rank correlation unveiled that eight out of thirteen environmental variables pose a significant relation with those cytometric indicators (Fig. 3).

![Figure 2](https://github.com/bmsw/Supplementary-Code/blob/master/Statistical_Analysis.R)

Figure 2: (a) PCA correlation biplot, boxplots (b, c and d) and density plot (e) computed from 31 patagonian water systems using three cytometric indices: richness, Pielou’s evenness and Shannon’s index. Principal components 1 and 2 explained more than 90% of data variation (65.59% and 32.85%, respectively). Shaded areas on PCA biplot represent 95% confidence ellipses.
Figure 3: Correlation matrix based on Spearman’s rank correlation coefficient (a) of four cytometric indices (total count of events, richness, Pielou’s evenness and Shannon’s index) and environmental variables. Black crosses indicates non-significant correlations. Linear regression between cytometric Shannon index and $\log_{10}$ cytometric richness (b), pH (c), $\log_{10}$ DOC and $\log_{10}$ $K_d$. Point sizes reflect $\log_{10}$ cytometric richness values.

In particular, pH, $K_d$ and DOC are variables directly associated with the trophic status. It is was demonstrated that at low DOC concentrations, only some bacterial specialists were able to actively incorporate the various types of organic matter effectively [50], as a consequence, the bacterial diversity would be low. Accordingly, the positive relationship observed between alfa diversity and DOC is in line with the idea that higher concentration of DOC, that is associated to more diverse DOC composition, would result in a higher diversity of bacterial that use these variety of compounds.

**Beta Diversity**

Ordination of Bray-Curtis distances reported an apparent location effect (i.e. differences in groups means) defined by trophic status of water systems (Fig. 4a). Indeed, although tests for homogeneity of groups variances did not reveal any statistical difference, PERMANOVA results flagged out significant divergences between all centroids distances. Yet, partitioning of distances revealed that nestedness is responsible for most of beta diversity among systems (Fig. 4b). Additionally, ordination scores seemed to have a higher significant linear correlation with nine environmental variables that are again related to trophic status and geografical location: DOC, chlorophyll a, pH, $K_d$, latitude, longitude, area, altitude and temperature.

**Mapping**

Biplot of samples and bins – based on channels FL1-H and SSC-H – shows a broadly common area shared by most of cytograms (blue and green clusters, Fig. 5a), as could be anticipated by nestedness
patterns from previous sections (Fig 4b). Samples are distinctly associated with specific image regions (clusters of bins), which subsequent visual inspection revealed to correspond, partially or totally, to known cytometric subpopulations (figures 5c-f and S6).

Figure 5: NMDS biplot (a) and mask of bins onto channels FL1-H and SSC-H (b). Cytograms from Ponds 7 (c) and 13 (d) are overplayed by (b) to reveal how known gated populations relates to ordination clusters (e and f). Red dotted arrows indicate logical pathway through figures.
Pairwise Comparisons

flowDiv and FlowFP were the only pipelines that significantly and positively correlated with DGGE information (Mantel statistic \( r \) equals 0.20 and 0.19, respectively). The significant correlation between molecular and cytometric diversity was also reported by García et al. 2015 [51] and Props et al. 2016 [16]. Those techniques were also highly correlated (Mantel statistic \( r = 0.65 \)), probably due their common principles (i.e. binning-based techniques).

Table 1: Mantel statistics based on Bray-Curtis distance matrix calculated for pairwise comparisons of pipelines. Asterisks (*) represent significant results at \( p < 0.05 \).

|        | DGGE  | CHIC  | Dalmation plot | CyBar  | flowFP | PhenoFlow | flowDiv |
|--------|-------|-------|----------------|--------|--------|-----------|--------|
| DGGE   | -     |       |                |        |        |           |        |
| CHIC   | 0.05  | -     |                |        |        |           |        |
| Dalmation plot | -0.05 | -0.06 | -              |        |        |           |        |
| CyBar  | -0.07 | -0.07 | -0.11          | -      |        |           |        |
| flowFP | 0.18* | 0.13  | -0.34          | 0.42*  | -      |           |        |
| PhenoFlow | 0.10  | 0.08  | -0.35          | 0.15   | 0.37*  |           |        |
| flowDiv | 0.20* | 0.12  | -0.20          | 0.12   | 0.65*  | 0.22*     | -      |

Although flowDiv did not significantly correlate with the remaining techniques, that discrepancies could be merely interpreted as a matter of tuning, caused by differences in their default working principles [6, 16].

Availability

flowDiv is freely available on CRAN (https://cran.r-project.org/web/packages/flowDiv/) and its source code can be found on GitHub (https://github.com/bmsw/flowDiv).

Conclusions

The need for both the reduction of analytical subjectivity and emphasis on more practical aspects of environmental flow cytometry studies evokes a paradigm shift as to harmonize objectivity with applicability. flowDiv provides a straightforward and rather intuitive way of proceeding such sort of analysis as it concatenates formal mathematical solutions and biological rationale into a intuitive framework especially designed to explore cytometric diversity.

Beyond solving for relevant technical issues, as analytical protocols discrepancies, flowDiv supplies an intelligible foundation to the use of multidimensional contingency tables on environmental FCM analysis. On one hand, multidimensional contingency tables resolves quite efficiently for multicolor assays as it keeps a epistemological relation to the fairly known ecological tables, thus endowing a straightforward biological interpretation of diversity indices derived from the pipeline; on the other hand, as flow cytometry data analysis mainly lays on graphical scrutiny of populations, ordinating...
the cells of those tables (i.e. the bins) and keeping track of them onto the cytograms provides an elegant strategy of understanding the global and local behavior of cytometric populations, allowing the study of patterns commonly not detected by traditional and often time-consuming visual methods, like microscopy. Finally, cytometric diversity provides a fast and cheap way to analyze diversity, as both molecular and cytometric diversity are correlated, being a promising screening tools for limnological surveys as well as for monitoring.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

We thank Romina Schiaffino and Irina Izaguirre for sharing data on Patagonian lakes and both Francisco Paulo Freire Neto and Ng Haig They for technical assistance. This work was supported by São Paulo Research Foundation (FAPESP) within processes 2014/14139-3 and 2016/50494-8. . .

References

[1] J. Comas-Riu and N. Rius, “Flow cytometry applications in the food industry,” *Journal of industrial microbiology & biotechnology*, vol. 36, no. 8, pp. 999–1011, 2009.

[2] J. M. Gasol and X. A. G. Moran, “Flow Cytometric Determination of Microbial Abundances and Its Use to Obtain Indices of Community Structure and Relative Activity,” *Hydrocarbon and Lipid Microbiology Protocols - Springer Protocols Handbooks*, pp. 1–29, 2015.

[3] A. Adan, G. Alizada, Y. Kiraz, Y. Baran, and A. Nalbant, “Flow cytometry: basic principles and applications,” *Critical reviews in biotechnology*, vol. 37, no. 2, pp. 163–176, 2017.

[4] J. Vives-Rego, P. Lebaron, and G. Nebe-von Caron, “Current and future applications of flow cytometry in aquatic microbiology,” *FEMS Microbiology Reviews*, vol. 24, no. 4, pp. 429–448, 2000.

[5] Y. Wang, F. Hammes, K. De Roy, W. Verstraete, and N. Boon, “Past, present and future applications of flow cytometry in aquatic microbiology,” *Trends in biotechnology*, vol. 28, no. 8, pp. 416–424, 2010.

[6] C. Koch, F. Harnisch, U. Schröder, and S. Müller, “Cytometric fingerprints: Evaluation of new tools for analyzing microbial community dynamics,” *Frontiers in Microbiology*, vol. 5, no. JUN, pp. 1–12, 2014.
[7] W. Li, “Cytometric diversity in marine ultraphytoplankton,” *Limnology and Oceanography*, vol. 42, no. 5, pp. 874–880, 1997.

[8] M. V. Quiroga, G. Mataloni, B. M. Wanderley, A. M. Amado, and F. Unrein, “Bacterioplankton morphotypes structure and cytometric fingerprint rely on environmental conditions in a sub-antarctic peatland,” *Hydrobiologia*, vol. 787, no. 1, pp. 255–268, 2017.

[9] H. Holyst and W. Rogers, *flowFP: Fingerprinting for Flow Cytometry*, 2009. R package version 1.30.0.

[10] C. Koch, I. Fetzer, H. Harms, and S. Müller, “Chic—an automated approach for the detection of dynamic variations in complex microbial communities,” *Cytometry Part A*, vol. 83A, no. 6, pp. 561–567, 2013.

[11] P. Bombach, T. Hübschmann, I. Fetzer, S. Kleinsteuber, R. Geyer, H. Harms, and S. Müller, “Resolution of natural microbial community dynamics by community fingerprinting, flow cytometry, and trend interpretation analysis.,” *Advances in biochemical engineering/biotechnology*, vol. 124, p. 151, 2011.

[12] J. Schumann, C. Koch, S. Günther, I. Fetzer, and S. Müller, *flowCyBar: Analyze flow cytometric data using gate information*, 2015. R package version 1.10.0.

[13] P. Legendre and L. Legendre, “Chapter 7 - ecological resemblance,” in *Numerical Ecology* (P. Legendre and L. Legendre, eds.), vol. 24 of *Developments in Environmental Modelling*, pp. 265 – 335, Elsevier, 2012.

[14] W. K. W. Li, “Macroecological patterns of phytoplankton in the northwestern north atlantic ocean,” *Nature*, vol. 419, pp. 154–157, Sep 2002.

[15] F. Ribalet, *cytoDiv: Cytometric diversity indices*, 2012. R package version 0.5-3.

[16] R. Props, P. Monsieurs, M. Mysara, L. Clement, N. Boon, and D. Hodgson, “Measuring the biodiversity of microbial communities by flow cytometry,” *Methods in Ecology and Evolution*, vol. 7, no. 11, pp. 1376–1385, 2016.

[17] L. A. Herzenberg, J. Tung, W. A. Moore, L. A. Herzenberg, and D. R. Parks, “Interpreting flow cytometry data: a guide for the perplexed,” *Nature immunology*, vol. 7, no. 7, pp. 681–685, 2006.

[18] C. J. ter Braak, “Principal components biplots and alpha and beta diversity,” *Ecology*, vol. 64, no. 3, pp. 454–462, 1983.

[19] K. O’Neill, N. Aghaeepour, J. Špidlen, and R. Brinkman, “Flow Cytometry Bioinformatics,” *PLoS Computational Biology*, vol. 9, no. 12, 2013.
[20] G. Finak and M. Jiang, *flowWorkspace: Infrastructure for representing and interacting with the gated cytometry*, 2011. R package version 3.18.10.

[21] B. Ellis, P. Haaland, F. Hahne, N. Le Meur, N. Gopalakrishnan, J. Spidlen, and M. Jiang, *flowCore: flowCore: Basic structures for flow cytometry data*, 2016. R package version 1.38.2.

[22] L. Birgé and Y. Rozenholc, “How many bins should be put in a regular histogram,” *ESAIM: Probability and Statistics*, vol. 10, no. November, pp. 24–45, 2006.

[23] D. Freedman and P. Diaconis, “On the histogram as a density estimator: L2 theory,” *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, vol. 57, no. 4, pp. 453–476, 1981.

[24] A. H. Andersen, “Multidimensional contingency tables,” *Scandinavian Journal of Statistics*, vol. 1, no. 3, pp. 115–127, 1974.

[25] J. Oksanen, F. G. Blanchet, M. Friendly, R. Kindt, P. Legendre, D. McGlinn, P. R. Minchin, R. B. O’Hara, G. L. Simpson, P. Solymos, M. H. H. Stevens, E. Szoecs, and H. Wagner, *vegan: Community Ecology Package*, 2017. R package version 2.4-3.

[26] H. Tuomisto, “A consistent terminology for quantifying species diversity? Yes, it does exist,” *Oecologia*, vol. 164, no. 4, pp. 853–860, 2010.

[27] A. Chao, N. J. Gotelli, T. C. Hsieh, E. L. Sander, K. H. Ma, R. K. Colwell, and A. M. Ellison, “Rarefaction and extrapolation with Hill numbers: A framework for sampling and estimation in species diversity studies,” *Ecological Monographs*, vol. 84, no. 1, pp. 45–67, 2014.

[28] M. O. Hill, “Diversity and evenness: A unifying notation and its consequences,” *Ecology*, vol. 54, no. 2, pp. 427–432, 1973.

[29] L. Jost, “Entropy and diversity,” *Oikos*, vol. 113, no. 2, pp. 363–375, 2006.

[30] A. L. Sheldon, “Equitability indices: dependence on the species count,” *Ecology*, vol. 50, no. 3, pp. 466–467, 1969.

[31] E. Pielou, *An introduction to mathematical ecology*. Wiley-Interscience, 1969.

[32] P. Koleff, K. J. Gaston, and J. J. Lennon, “Measuring beta diversity for presence – absence data,” *Journal of Animal Ecology*, vol. 72, pp. 367–382, 2003.

[33] J. R. Bray and J. T. Curtis, “An Ordination of the Upland Forest Communities of Southern Wisconsin,” *Ecological Monographs*, vol. 27, no. 4, pp. 325–349, 1957.

[34] P. L. Buttigieg and A. Ramette, “A guide to statistical analysis in microbial ecology: a community-focused, living review of multivariate data analyses,” *FEMS microbiology ecology*, vol. 90, no. 3, pp. 543–550, 2014.
[35] A. Baselga, “Partitioning the turnover and nestedness components of beta diversity,” *Global Ecology and Biogeography*, vol. 19, no. 1, pp. 134–143, 2010.

[36] A. Baselga, “Separating the two components of abundance-based dissimilarity: balanced changes in abundance vs. abundance gradients,” *Methods in Ecology and Evolution*, vol. 4, no. 6, pp. 552–557, 2013.

[37] P. Legendre and E. D. Gallagher, “Ecologically meaningful transformations for ordination of species data,” *Oecologia*, vol. 129, no. 2, pp. 271–280, 2001.

[38] M. Charrad, N. Ghazzali, V. Boiteau, and A. Niknafs, “NbClust: An R package for determining the relevant number of clusters in a data set,” *Journal of Statistical Software*, vol. 61, no. 6, pp. 1–36, 2014.

[39] T. Caliński and J. Harabasz, “A dendrite method for cluster analysis,” *Communications in Statistics-theory and Methods*, vol. 3, no. 1, pp. 1–27, 1974.

[40] M. Romina Schiaffino, F. Unrein, J. M. Gasol, R. Massana, V. Balague, and I. Izaguirre, “Bacterial community structure in a latitudinal gradient of lakes: the roles of spatial versus environmental factors,” *Freshwater Biology*, vol. 56, no. 10, pp. 1973–1991, 2011.

[41] M. R. Schiaffino, J. M. Gasol, I. Izaguirre, and F. Unrein, “Picoplankton abundance and cytometric group diversity along a trophic and latitudinal lake gradient,” *Aquatic Microbial Ecology*, vol. 68, no. 3, pp. 231–250, 2013.

[42] M. R. Schiaffino, M. L. Sánchez, M. Gerea, F. Unrein, V. Balagué, J. M. Gasol, and I. Izaguirre, “Distribution patterns of the abundance of major bacterial and archaeal groups in patagonian lakes,” *Journal of Plankton Research*, vol. 38, no. 1, pp. 64–82, 2015.

[43] M. Hervé, *RVAideMemoire: Diverse Basic Statistical and Graphical Functions*, 2017. R package version 0.9-65.

[44] E. A. Pena and E. H. Slate, *gvlma: Global Validation of Linear Models Assumptions*, 2014. R package version 1.0.0.2.

[45] T. Wei and V. Simko, *corrplot: Visualization of a Correlation Matrix*, 2016. R package version 0.77.

[46] G. R. Warnes, B. Bolker, L. Bonebakker, R. Gentleman, W. H. A. Liaw, T. Lumley, M. Maechler, A. Magnusson, S. Moeller, M. Schwartz, and B. Venables, *gplots: Various R Programming Tools for Plotting Data*, 2016. R package version 3.0.1.

[47] H. Wickham, *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2009.
[48] W. K. Dodds and J. J. Cole, “Expanding the concept of trophic state in aquatic ecosystems: it’s not just the autotrophs,” *Aquatic Sciences*, vol. 69, no. 4, pp. 427–439, 2007.

[49] S. A. Wood, M. Y. Maier, J. Puddick, X. Pochon, A. Zaiko, D. R. Dietrich, and D. P. Hamilton, “Trophic state and geographic gradients influence planktonic cyanobacterial diversity and distribution in new zealand lakes,” *FEMS Microbiology Ecology*, vol. 93, no. 2, 2017.

[50] H. Sarmento, C. Morana, and J. M. Gasol, “Bacterioplankton niche partitioning in the use of phytoplankton-derived dissolved organic carbon: quantity is more important than quality,” *The ISME journal*, vol. 10, no. 11, pp. 2582–2592, 2016.

[51] F. C. García, L. Alonso-Sáez, X. A. G. Morán, and Á. López-Urrutia, “Seasonality in molecular and cytometric diversity of marine bacterioplankton: the re-shuffling of bacterial taxa by vertical mixing,” *Environmental microbiology*, vol. 17, no. 10, pp. 4133–4142, 2015.

**Additional Files**

Additional file 1 — Cytograms and overlaid masks of bins onto channels FL1-H and SSC-H for all 31 Patagonian lakes used in this study.

Additional file 2 — Heatmaps based on distance matrices (Bray–Curtis distance) from 26 Patagonian lakes used in this study.
Figure 6: Cytograms and overlaid masks of bins onto channels FL1-H and SSC-H for all 31 Patagonian lakes used in this study.
Figure 7: Heatmaps based on distance matrices (Bray-Curtis distance) from 26 Patagonian lakes used in this study. Data are from: (a) DGGE, (b) CHIC, (c) flowCyBar, (d) Dalmation Plot, (e) FlowFP, (f) PhenoFlow and (g) flowDiv pipelines. Dendrograms were based on Ward’s hierarchical agglomerative clustering method.