Reviewer's report
Title: Flow: Statistics, visualization and informatics for flow cytometry
Version: 1 **Date:** 10 May 2008
Reviewer number: 1
Reviewer's report:
Major Compulsory Revisions:

1) The authors frequently use the terms "clustering" and "classification" in different contexts. For the mixture modeling and KDE approaches the term clustering is appropriate. As far as I can tell, classification in the machine learning sense is not performed at all (this would involve some form of supervised learning). A term better suited for assignment of OBO terms to subsets of the data might be "annotation".

We agree with the reviewer that “annotation” is a more appropriate term and have changed the manuscript in the appropriate places.

2) In flow cytometry data analysis, the term "filtering" is used for the process of applying gates on the data. In this manuscript, filtering describes a potentially random subsetting operations. The authors should try to disambiguate the term.

We have now used the term “sub-sampling” to avoid confusion.

3) The manuscript does not address how to treat groups of flow measurements in a complex experiment. There doesn't seem to be a concept of sample annotation or coordinated analysis of common samples within a sample group, or specific data structures do deal with that. Does the hierarchical tree structure allow for sample groups? And can meta data be edited?. Since annotation and keeping track of experimental meta data is one crucial point in high throughput experiments, the authors should point out how this can be handled within their software.

We have implemented new features to address the deficiencies in managing grouped data in the software. For annotation purposes, we treat any non-terminal node in the control tree as a group, and have added the ability to annotate all such nodes with key-value pairs, which can be edited by selecting the “Annotate” option in the contextual (right click) menu. We suggest the convention that such annotations apply to all the children of this node. Such annotations are viewable in the informational pane on the right side of the Control panel. It is also possible to change the name of any node in the tree by left clicking on its label, create new empty groups, as well as to move nodes to different locations on the tree using the Cut and Paste options on the Tree or context menus, which provides some flexibility in data organization. These annotation options are now described in the manuscript on Page xxx.

Apart from annotation, another common use of groupings is to allow “batch” operations. We have added a new batch option to the context menu, which will apply the last operation performed on the selected node to other groups that can be selected from the presented dialog box. An example of doing a batch gating operation has been added to the manuscript to illustrate this new feature on Page xxx.

4) The tree doesn't make use of inheritance, that is, features are only available at a certain level in the tree and not for all downstream leaves. Is this a design decision or a shortcoming of the underlying data structure?

We have implemented limited inheritance for groups (non-terminal nodes), but generally avoided doing so for leaves where there may be more than a single interpretation of the user action. For example, an operation on a group that does not have an associated data leaf will use the nearest data found on recursing up the tree from that group. We have also added inheritance for the clustering labels “z” so that children of a node will have the same z-labels as their parent unless over-ridden.

5) I was not able to make the projection feature work on my installation. Neither could I install the software from the source code provided in the supplementary materials. When creating 3D plots for transformed data, the software doesn't seem to pick up the scale changes as all transformed channels are stuck on the axes.

The projection feature makes use of the fastICA library of the R statistical package using the bindings provided by the open source Rpy module. Unfortunately, the Rpy module is typically tied to a particular version of R, and will not work with other versions of R. We have addressed this in several ways – 1) we have provided a Debian binary package which will attempt to install compatible R and Rpy versions, 2) we have provided a standalone version of the PCA projection as a python plugin as an example of how to write such projection plugins, and 3) we have documented the installation steps in detail for specific versions of R and RPy on several platforms that work for us in the package README file and website documentation.

We have fixed the 3D scale change bug.

6) The separation between basic features and plug in functionality is not very clear. E.g., adding ellipse outlines to a plot only makes sense for normal or at least symmetric mixture components. While the focus is obviously on automated gating, there still seems to be a need for the classical flow data analysis tools. The very rudimentary tools provided with the base annotation might not be sufficient.

We have revised the software so that only the operations applicable to a selection are enabled – so the option to plot ellipses will only be enable when the selected group has components that has the necessary mean and covariance matrix entries.

We have also added or improved the following classical flow data analysis tools 1) data compensation in which changes to the compensation matrix are simultaneously reflected in an associated dot plot, 2) dot plots by default now color events by the local density to provide more information, 3) the option of overlays for histogram plots and 4) quadrant gates with the percentage of events in each quadrant. We look forward to user feedback from the flow community as to what other features would be useful, or even better, to contributors providing plug-ins for needed functionality!

Minor Essential Revisions:
1) Page 2, first paragraph: "The basic ideas is that if a subset..."
This might be true in theory, however in practise many gate selections rely either on expert knowledge by the investigator or information that is borrowed from other samples like negative or positive controls. Either way, the selection is not exclusively based on the features of the underlying data.

On reviewing the statement highlighted by the reviewer, we agree that it glosses over the complexity of gating in practice and have rewritten the sentence to emphasize that while the goal of automating flow cytometric analysis is our motivation, it is certainly not a reality with the current tools at our disposal. The revised statement on Page 2 is reproduced here

"The motivating idea is that if a subset of cells forms a distinct visual subgroup that can be demarcated visually with sequential gating, it should also be possible in principle to extract the subset with the appropriate statistical model. In practice, accurate gating is often only possible with extensive expert knowledge and by comparison with both negative and positive controls, and formidable statistical challenges exist. Nevertheless, we believe that statistical modeling can ameliorate much of the tedium and subjectivity inherent in gating on PFC data, and that *Flow* is a useful medium to increase awareness of such methods among the flow community."

2) Page 6, second paragraph: "All the clustering routines..." delete one of the two "alls"

Done.

3)Page 9, fourth paragraph: "One limitation of flow is..." The authors should mention the available markup languages here, which are the recommended way to interface between programs on a file level. The FCS format is strictly limited to the raw measurement data and mostly instrument specific meta data.

We thank the reviewer for bringing to our attention the existence of markup languages like FlowML, which would be an ideal format for exchanging flow data and more suitable than the FCS format. We have therefore amended the sentence to read

An attractive solution would be to write a routine that translates the HDF5 data into the FlowML markup language (citation needed), which provides an standard well-documented format for flow data exchange and likely to be supported by other flow cytometry software in the near future. Users are encouraged to develop such plugins for their own needs and contribute to the further development of the software.

Discretionary Revisions:
1) The installation process is very tedious since there are many dependencies on other software. The documentation on the project web page is helpful, but providing binaries would be crucial for the intended target audience to make use of the software. A debian meta package containing all dependencies would speed up the installation process for Linux users.

We have provided Debian and Ubuntu packages in an apt repository located at xxx. Instructions for installation etc needed. Universal binaries for OS X 10.4 and 10.5 are also available for download from xxx.

2) QA/QC is not mentioned at all in the manuscript, but seems to be a major deal for high throughput data analysis. The authors' views on how to implement that using their software would be helpful.

There are many aspects of QA/QC in flow cytometry, one of the most critical of which is rigorous instrument and reagent calibration, and the use of well-defined standard protocols for sample preparation. While this is largely beyond the scope of software, we are considering the incorporation of annotation features for the Minimal Standards for Flow Cytometry (MiFlowcyt) in a future release of the software. Some simple “identities”, for example, that the sum of total CD4+ and CD8+ event should be close to total CD3+ events) are also useful in checking for egregious errors, and it should be possible to incorporate such validation in the software that come into play once the user has labeled the cell clusters found. Finally and more ambitiously, we are investigating the possibility of using a database of flow cytometric phenotype statistics to automatically flag statistically anomalous data sets.

3) There seems to be an issue with the security certificate of the
project web page. The latest Firefox 3.0 did not display the page before setting a security check exception.

This has been fixed.

Level of interest: An article of importance in its field
Quality of written English: Acceptable
Statistical review: Yes, and I have assessed the statistics in my report.

Reviewer's report
Title: Flow: Statistics, visualization and informatics for flow cytometry
Version: 1 **Date:** 13 May 2008
Reviewer number: 2
Reviewer's report:

1. The challenging in the analysis of high-throughput data is one the bottlenecks in the current post-genomic era. The need of tools that allow the easy application of statistical methods is increasing. Although the Bioconductor project provides the packages rflowcyt and flowCore (and derivatives) for the analysis of Flow Cytometry data, the nature of the R graphic display makes difficult the interaction with the user. Therefore, the user needs to know how to do the analysis using the command line interface. The existence of free tools for the analysis of FCS data from an interactive point of view is very welcome, and this reviewer finds that this software will be undoubtedly very useful for experimentalist.

# Major Compulsory Revisions (which the author must respond to before a decision on publication can be reached)
None found.

# Minor Essential Revisions (such as missing labels on figures, or the wrong use of a term, which the author can be trusted to correct)

2. In page 2, Background, first paragraph. The sentences: “Polychromatic flow cytometry (PFC) is the only currently available assay that can track individual functional responses in different cell subsets simultaneously...” needs a citation.

We have added a citation for this claim (Seder et al 2008), and amended the phrase to be closer to the statement found in paragraph two of the cited review, although to the best of our knowledge the original statement is correct.

Polychromatic flow cytometry (PFC) is the only currently available assay that can track multiple functional responses simultaneously, on both single cell and population levels.

3. In page 9, Discussion and Conclusions, first paragraph. The correct case for Rflowcyt is rflowcyt and for FlowCore is flowCore, according to the description in the Bioconductor web page.

This has been corrected.

# Discretionary Revisions (which are recommendations for improvement but which the author can choose to ignore)
4. Although this software might be very useful for people working with FCS data, the drawback this reviewer sees is that the installation of all the dependencies

needed for the program to run is not trivial. Consequently, a regular user will find a lot of problems with it, limiting the number of potential users or requiring the advice of an expert. Therefore, my advice is that as much as possible, the authors buddle some of the requirements in the distribution. This will have the benefit of reducing the dependencies and including compatible versions of them.

We have now provided Debian/Ubuntu binary packages, which is currently the most widely used Linux distribution, as well as universal binaries for Mac OS X 10.4 and 10.5 blah blah blah …

Level of interest: An article of importance in its field
Quality of written English: Acceptable
Statistical review: No, the manuscript does not need to be seen by a statistician.