OMICS A Journal of Integrative Biology
Volume 10, Number 2, 2006

Data Standards for Flow Cytometry

JOSEF SPIDLEN,[1] ROBERT C. GENTLEMAN,[2] PERRY D. HAALAND,[3] MORGAN LANGILLE,[1] NOLWENN LE MEUR,[2] MICHAEL F. OCHS,[4] CHARLES SCHMITT,[3] CLAYTON A. SMITH,[1] ADAM S. TREISTER,[5] and RYAN R. BRINKMAN[1]

ABSTRACT

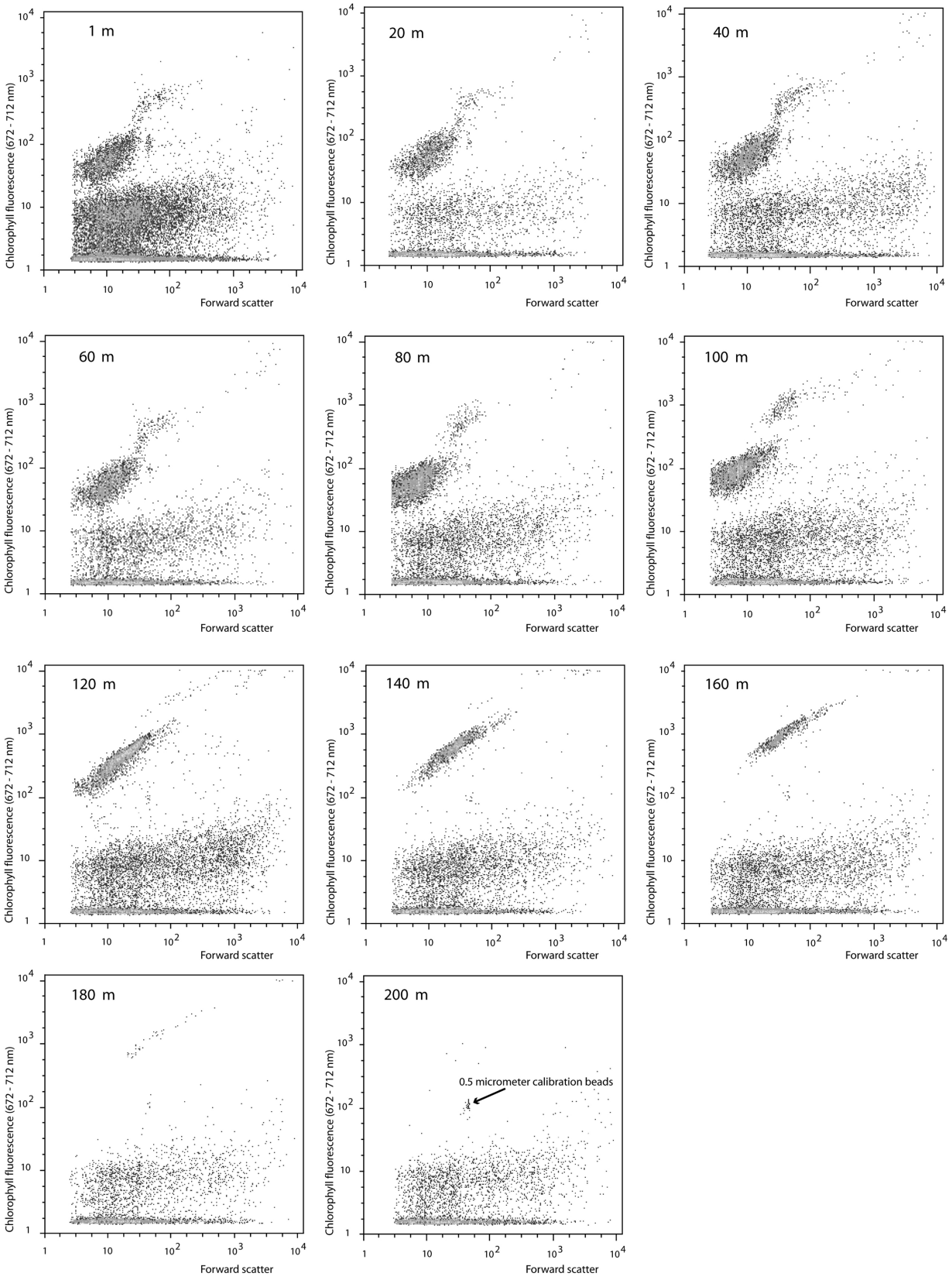
Flow cytometry (FCM) is an analytical tool widely used for cancer and HIV/AIDS research, and treatment, stem cell manipulation and detecting microorganisms in environmental samples. Current data standards do not capture the full scope of FCM experiments and there is a demand for software tools that can assist in the exploration and analysis of large FCM datasets. We are implementing a standardized approach to capturing, analyzing, and disseminating FCM data that will facilitate both more complex analyses and analysis of datasets that could not previously be efficiently studied. Initial work has focused on developing a community-based guideline for recording and reporting the details of FCM experiments. Open source software tools that implement this standard are being created, with an emphasis on facilitating reproducible and extensible data analyses. As well, tools for electronic collaboration will assist the integrated access and comprehension of experiments to empower users to collaborate on FCM analyses. This coordinated, joint development of bioinformatics standards and software tools for FCM data analysis has the potential to greatly facilitate both basic and clinical research—impacting a notably diverse range of medical and environmental research areas.

This paper is part of the special issue of OMICS on data standards.

INTRODUCTION

FLOW CYTOMETRY (FCM) is a technique used in basic and clinical research for studying the immunological status of patients treated with vaccines or other immunotherapies, for characterizing cancer, HIV/AIDS infection, and other diseases, as well as for research and therapy involving stem cell manipulation (Braylan, 2004; Hengel et al., 2001). It is also used for studying environmental samples, such as detecting specific mocroorganisms in soil or water samples (Lomas, 2004) (Fig. 1). In FCM, intact cells and their constituent components are tagged with fluorescently conjugated monoclonal antibodies and/or stained with fluorescent reagents and then analyzed individually. In the instrument, hydrodynamic forces line cells

[1]Terry Fox Laboratory, British Columbia Cancer Research Center, Vancouver, Canada.
[2]Fred Hutchinson Cancer Research Center, Seattle, Washington.
[3]BD Technologies, Research Triangle Park, North Carolina.
[4]Fox Chase Cancer Center, Philadelphia, Pennsylvania.
[5]Tree Star, Inc., San Carlos, California.


1 m
20 m
40 m
60 m
80 m
100 m
120 m
140 m
160 m
180 m
200 m
0.5 micrometer calibration beads
Chlorophyll fluorescence (672 - 712 nm)
Forward scatter
1
10
10^2
10^3
10^4

up in single file and the fluorescent molecules in/on each cell are excited by laser light at speeds that can exceed 70,000 cells per second (Bonetta, 2005). The fluorescence emission from each cell is collected by a series of photomultiplier tubes, and the subsequent electrical events are collected and analyzed on a computer that assigns a fluorescence intensity value to each signal in Flow Cytometry Standard (FCS) data files (Seamer et al., 1997). FCM analysis involves identifying intersections or unions of polygonal regions in hyperspace that are used to filter or "gate" data and define a subset or sub-population of events (or exclude, for example, cell debris) for further analysis or sorting.

The International Society for Analytical Cytology (ISAC) has adopted the FCS Data File Standard for the common representation of FCM data. This standard is supported by all of the major analytical instruments to record the measurements from a sample run through a cytometer, and scientists can choose among instruments and software with no major data compatibility issues. However, this standard stops short of describing the protocol used or the computational post-processing and data analysis performed in an FCM experiment. It also does not cover data interpretation, one of the most difficult and time-consuming aspects of the entire analytical process (Braylan, 2004). FCM has traditionally been a manually intensive technique; however, automated high-throughput FCM techniques have been recently developed that can rapidly collect large data sets with complexities similar to gene microarrays (Gasparetto et al., 2004). The huge amount of information generated by high-throughput technologies need to be transformed into executive summaries that are brief enough for creative studies by a human researcher (Brazma, 2001). One of the most insidious problems in accomplishing this goal is the lack of standard data formats for information exchange (Chicurel, 2002). One basic challenge for FCM is to greatly simplify, from the end user's viewpoint, data analysis and extraction of statistical information (Boddy et al., 2001; Herzenberg et al., 2002). This needs to happen in a highly systematic, automated, and traceable way that retains flexibility and is consistent with current visually implemented tools. Further requirements include organizing data in such a way that raw data can be combined from multiple centers and scientists and clinicians at remote locations can collaborate on data interpretation. Such needs are currently lagging behind the ability to actually collect the samples and run the FCM analyses (De Rosa et al., 2003).

DATA STANDARDS METHODOLOGY AND GOALS

To address these shortcomings, we responded to the NIH Program Announcement "Innovations in Biomedical Computational Science and Technology" (PAR-03-10), which solicited proposals for the development of new informatics, computational and mathematical tools and technologies including platform-independent translational tools for data exchange and for promoting interoperability. We have brought together a cross-disciplinary international collaborative group of bioinformaticists, computational statisticians, software developers and clinician scientists, from both academia and industry (including both software and hardware suppliers) to collaborate on the development of data standards for flow cytometry. In conjunction with the ISAC data standards committee and an Institute of Electrical and Electronics Engineers (IEEE) Working Group (Bioinformatics Standards for Flow Cytometry WG, P1943.2), our goal is to provide consistency in the electronic recording of flow cytometry data analysis. We aim to create universal solutions for representing, collecting, annotating, archiving, analyzing and disseminating flow cytometry data and analyses.

FIG. 1. An example of flow cytometry analysis: Sargasso Sea samples at different depth. Chlorophyll fluorescence versus forward scatter of marine microbes (mainly the cyanobacterium *Prochlorococcus*—a microorganism approximately 0.2 μm in diameter) from samples taken at different depths at the Bermuda sampling station (BATS program; directed by Michael Lomas of the Bermuda Biological Station for Research). Chlorophyll fluorescence corresponding to the relative chlorophyll content was detected using a filter for 672–712 nm. Forward-scattered light (FSC) is proportional to cell-surface area or size. FSC is a measurement of mostly diffracted light and is detected just off the axis of the incident laser beam in the forward direction. Measurements were obtained using a Cytopeia Influx jet in air sorter, 200 mWatt excitation by a 488 nm laser. The plots indicate that chlorophyll content and size of these marine organisms remain constant, with increasing the depth up to 100 m where they both increase and up to 200 m where the populations disappear. The small tight cluster that appears in some of the panels (indicated by arrow in 200-m sample) represents 0.5-μm calibration beads.

First, we are establishing a guideline that outlines the minimum information required to unambiguously record, report, interpret, and reproduce FCM experiments, the Minimum Information for a Fluorescent Activated Cell Experiment (MI-FACE), to promote the standardized documentation of experimental details (Fig. 2). In developing this guideline, we are, in part, capitalizing on previous bioinformatics standards developments, most notably the development of the Minimum Information About a Microarray Experiment (MIAME) standard (Brazma et al., 2001; Ball and Brazma, *this issue*), a process that is being successfully adopted by other communities (Le Novere et al., 2005; Fiehn et al., 2006, Taylor et al., *this issue*). This guideline will be encapsulated using the Unified Modeling Language (UML) and a model will be created to enable its application within various software components using the Extensible Markup Language (XML) standard. It is also necessary to develop platform independent extensions of the model for describing compensation and gating to enable the cross-platform comparison of gating results and analytical pipelines. In order to ensure collaboration, it is also essential to provide the standardized use of the data attributes though documentation and controlled vocabularies (Taylor et al., *this issue*; Orchard et al., 2005).

CURRENT STATUS

Though we are still within the first year of this project, we have developed the first draft of a Fluorescent Activated Cell Experiment Ontology (FACE Ontology), including controlled vocabulary for referring to terms. We built on the knowledge encapsulated in CytometryML (Leif et al., 2003), an XML encapsulation of FCM metadata, transforming CytometryML into the World Wide Web Consortium (W3C) standardized Web Ontology Language (OWL files). Furthermore, we can capitalize on ontologies being developed for other functional genomics platforms, for example, the Functional Genomics Ontology (FuGO) (Whetzel et al., *this issue*). We are working with the FuGO development group to build on and extend aspects of FuGO as an upper ontology for FCM development purposes.

Also, we have already developed the first draft of MI-FACE, including a detailed specification of the gating process, its documentation, and corresponding standards for XML-based technologies. The stan-

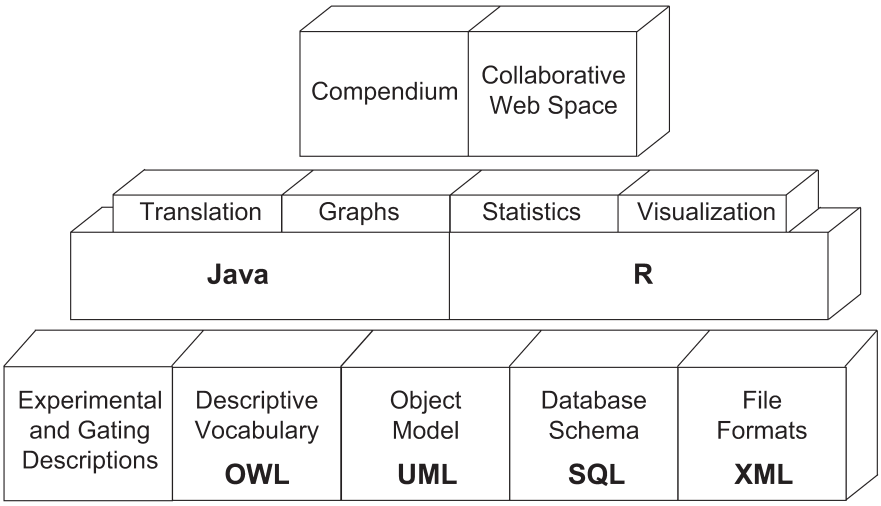


FIG. 2. Data standards for flow cytometry, project methodology. Figure shows the stepwise methodology selected to achieve projects goals. The basic corner stones are being developed first, for example, a guideline that outlines the minimum information required for a FCM experiment, encapsulated within a data-centric modeling language, standardized platform-independent file formats, standardized database representation, and a controlled terminology system. Software tools corresponding to these standards are being created in order to provide reference implementations. Finally, a collaborative web space and experiments' compendiums will support reproducible research analysis, including possibilities of verification by independent researchers.

dardization of the gating process is presently our most well-developed component, reflecting the high need for standardized gating procedures. The lack of a shared representation of gates in FCM prevents a variety of collaborative opportunities to recreate experimental methods and results. We have developed a detailed description of the gating specification, a W3C Schema usable to validate gating XML files, user documentation, and a set of examples of gating XML files. The first version of a reference platform independent software tool is being developed (named FACE-Java) which can read an FCS data file, along with an accompanying XML file describing gates according to the specification we have developed, and process this information to provide descriptive statistics. The release version of the FACE-Java software package will be useful to validate compliance of alternative software tools implementing the gating standard.

However, Java is not the only platform that we are focused on. In order to support statistical analysis in FCM an R package (RFlowCyt) is under development. The R Project for Statistical Computing is a very popular open-source research platform for evaluating and implementing statistical methods. RFlowCyt provides a platform for rapid prototyping of statistical methods, report generation, and as a sophisticated cross-platform (operating systems) data analysis tool. It currently can import data from FCS 2.0 and 3.0 files, provides a preliminary interface for gating, and computes post-gating distributional tests for two sample comparisons.

CHALLENGES

When any data standard is developed, the biggest challenge is insuring acceptance in the wider community. To cover the widespread acceptance of our work, it is critical that development of the standards takes place in a open and collaborative manner that involves the entire FCM community. Therefore, we are collaborating with the international standards body for FCM (ISAC), which will be solicited for input and approval as the standards mature. We are also involving biologists from various fields of FCM application, bioinformaticists, and members of the FCM software and hardware industry. Moreover, we believe that providing further motivation besides being standard-compliant can increase the likelihood of the acceptance. We are therefore developing a software tool that enables and fosters the exchange, re-exploration and re-interpretation of data and analyses by scientists. The fundamental tenet of scientific research is that the published results of any study have to be open to independent validation or refutation (Quackenbush, 2004). With journal publications authors are trapped within a format and a language that is not conductive to the complete description of software manipulations performed on data. The audience is separated from the actions and details of the algorithms used and are often forced to make assumptions regarding computational details which can result in completely different results. Our compendium will not only be a traditional comprehensive compilation of a body of knowledge and experiment results, but also it will contain a kind of active experiment result document (Gentleman et al., 2003). Such an active document will be linked to the FCM raw data (FCS files) and experiment and analysis details including gating descriptions, and it will be created automatically and dynamically based on the linked information. This feature not only motivates researchers but it also significantly facilitates reproducible and extensible FCM data analyses.

CONCLUSION

In short, as FCM-based analyses expand in their size and complexity, there is now a critical need for the development of high quality FCM data standards and tools to facilitate FCM-based research. Through the standards we develop, along with associated tools that foster the exchange and re-exploration of experiments by scientists, we hope to significantly accelerate more sophisticated and collaborative research involving FCM data. This work has the potential to impact on diverse research fields, from manipulation of stem cells, to microbiological analyses of our environment.

To obtain current information about this project, to download the proposed specifications, or to join the discussion concerning the standards under development, please visit our web site (<www.flowcyt.org>).

ACKNOWLEDGMENTS

We thank Dr. Ger van den Engh for kindly providing the Sargasso Sea sampling figure. R.R.B. is supported by the Michael Smith Foundation for Health Research. The project is supported by the grant number NIH/NIBIB R01 EB-5034.

REFERENCES

BALL, C.A., and BRAZMA, A. (2006). MGED standards: work in progress. OMICS (*this issue*).

BODDY, L., WILKINS, M.F., and MORRIS, C.W. (2001). Pattern recognition in flow cytometry. Cytometry **44,** 195–209.

BONETTA, L. (2005). Flow cytometry smaller and better. Nat Methods **2,** 785–795.

BRAYLAN, R.C. (2004). Impact of flow cytometry on the diagnosis and characterization of lymphomas, chronic lymphoproliferative disorders and plasma cell neoplasias. Cytometry A **58,** 57–61.

BRAZMA, A. (2001). On the importance of standardisation in life sciences. Bioinformatics **17,** 113–114.

BRAZMA, A., HINGAMP, P., QUACKENBUSH, J., et al. (2001). Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. Nat Genet **29,** 365–371.

CHICUREL, M. (2002). Bioinformatics: bringing it all together. Nature **419,** 751, 753, 755 passim.

DE ROSA, S.C., BRENCHLEY, J.M., and ROEDERER, M. (2003). Beyond six colors: a new era in flow cytometry. Nat Med **9,** 112–117.

FIEHN, O., KRISTAL, B., VAN OMMEN, B., et al. (2006). Establishing reporting standards for metabolomic and metabonomic studies: a call for participation. OMICS (*this issue*).

GASPARETTO, M., GENTRY, T., SEBTI, S., et al. (2004). Identification of compounds that enhance the anti-lymphoma activity of rituximab using flow cytometric high-content screening. J Immunol Methods **292,** 59–71.

GENTLEMAN, R., and LANG, D.T. (2003). Statistical analyses and reproducible research. Available at: ⟨www.biostat.harvard.edu/~rgentlem/Pdf/RR.pdf.⟩

HENGEL, R.L., and NICHOLSON, J.K. (2001). An update on the use of flow cytometry in HIV infection and AIDS. Clin Lab Med **21,** 841–856.

HERZENBERG, L.A., PARKS, D., SAHAF, B., et al. (2002). The history and future of the fluorescence activated cell sorter and flow cytometry: a view from Stanford. Clin Chem **48,** 1819–1827.

LEIF, R.C., LEIF, S.B., and LEIF, S.H. (2003). Cytometry ML, an XML format based on DICOM and FCS for analytical cytology data. Cytometry A **54,** 56–65.

LE NOVERE, N., FINNEY, A., HUCKA, M., et al. (2005). Minimum information requested in the annotation of biochemical models (MIRIAM). Nat Biotechnol **23,** 1509–1515.

LOMAS, M. (2004). Taking a closer look at the ocean. Bermuda Biological Station for Research, Inc., annual report. Available at: ⟨www.bbsr.edu/ar04.pdf⟩.

ORCHARD, S., MONTECCHI-PALAZZI, L., HERMJAKOB, H., et al. (2005). The use of common ontologies and controlled vocabularies to enable data exchange and deposition for complex proteomic experiments. Pac Symp Biocomput **10,** 186–196.

QUACKENBUSH, J. (2004). Data standards for "omic" science. Nat Biotechnol **22,** 613–614.

SEAMER, L.C., BAGWELL, C.B., BARDEN, L., et al. (1997). Proposed new data file standard for flow cytometry, version FCS 3.0. Cytometry **28,** 118–122.

TAYLOR, C.F., HERMJAKOB, H., JULIAN, JR., R.K., et al. (2006). The work of the Human Proteome Organisation's Proteomics Standards Initiative (HUPO PSI). OMICS (*this issue*).

WHETZEL, P.L., BRINKMAN, R.R., CAUSTON, H., et al. (2006). Development of FuGO: an ontology for functional genomics investigations. OMICS (*this issue*).

Address reprint requests to:
Dr. Ryan R. Brinkman
Terry Fox Laboratory
British Columbia Cancer Research Center
675 West 10th Ave.
Vancouver, BC, V5Z 1L3 Canada

E-mail: rbrinkman@bccrc.ca