A rich analytical environment for flow cytometry experimental results

Janet Siebert*

Department of Computer Science and Engineering,
University of Colorado at Denver and Health Sciences Center,
Campus Box 109, P.O. Box 173364,
Denver, CO 80217 3364, USA
E-mail: jsiebert@acm.org
*Corresponding author

Krzysztof J. Cios

Department of Computer Science and Engineering,
University of Colorado at Denver and Health Sciences Center,
Campus Box 109, P.O. Box 173364,
Denver, CO 80217 3364, USA
E-mail: krys.cios@cudenver.edu

Department of Computer Science, University of Colorado at Boulder

Department of Preventive Medicine & Biometrics (School of Medicine),
University of Colorado at Denver and Health Sciences Center

M. Karen Newell

University of Colorado at Colorado Springs,
CU Institute of Bioenergetics,
1420 Austin Bluffs Parkway,
Science Building Room 142,
Colorado Springs, CO.80918, USA
E-mail: mnewell@uccs.edu

Abstract: Existing analysis tools for flow cytometry data offer specialised but limited functionality. This work presents advantages of combining the cytometer's data with sample-specific information. Data is loaded into a relational database, where the analyst can query based on sample characteristics such as species, gender, diet type or sample stain type.

Keywords: flow cytometry; immunology; data analysis; data mining; knowledge discovery.

Reference to this paper should be made as follows: Siebert, J., Cios, K.J. and Newell, M.K. (2006) 'A rich analytical environment for flow cytometry experimental results', *Int. J. Bioinformatics Research and Applications*, Vol. 2, No. 1, pp.52–62.

Biographical notes: Janet Siebert is a data architect with 20 years of industry experience. She has worked on data conversions and data warehouses in the telecommunications, healthcare and financial services fields. She is particularly interested in knowledge transfer across disciplinary boundaries. She holds an MEd from Vanderbilt University, and an MS in Computer Science from the University of Colorado Denver.

Krzysztof J. Cios' research is in biomedical informatics, machine learning and data mining. He published two monographs, over 140 papers, and edited four journal special issues. He received his MS and PhD Degrees from the AGH University of Science and Technology, Krakow, MBA from the University of Toledo, Ohio, and DSc from the Polish Academy of Sciences. He serves on several journal editorial boards, and has been the recipient of the Norbert Wiener Outstanding Paper Award, the Neurocomputing Best Paper Award and the Fulbright Senior Scholar Award. He is a foreign member of the Polish Academy of Arts and Sciences.

M. Karen Newell research interests include immune-mediated cell death, tumor immunology, and cellular metabolism. She received her PhD in Microbiology and Immunology, University of Colorado-Health Sciences Center. She is currently Associate Professor and Markert Endowed Chair of Biology, University of Colorado at Colorado Springs, and Chief Executive Scientific Director, CU-Institute of Bioenergetics. Other academic and professional positions include Postdoctoral Fellow, McGill University, Montreal (Canada); Postdoctoral Fellow, National Jewish Center for Immunology and Respiratory Medicine, Denver, Assistant Professor, Department of Medicine, University of Vermont, Adjunct Assistant Professor, Dartmouth Medical College, and Assistant Professor, Department of Biology, UCCS.

1 Introduction

Flow cytometry is a common technique used by research biologists and immunologists. Flow cytometry processing collects data on several different attributes of each cell, and on thousands of cells per sample. This technique is used to study cell behaviour, and to investigate treatments for diseases such as cancer, HIV and sickle cell anaemia.

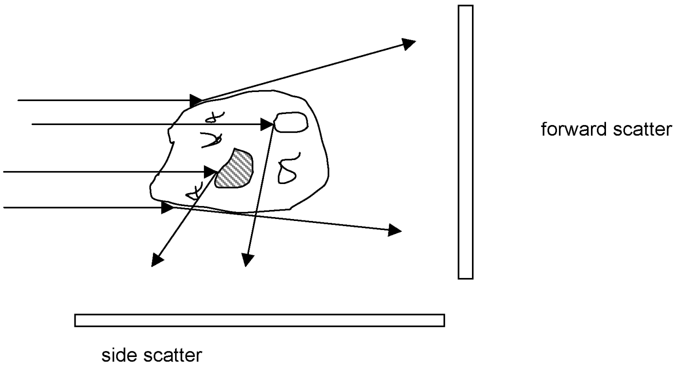
The data collected by the flow cytometer is written to a published but esoteric format. Generally, biologists use proprietary software to access the data and perform a fixed set of analyses. Unfortunately, these techniques are limited and limiting. This work provides mechanisms and methods to dramatically improve the efficiency and range of the data analysis techniques.

One of the strengths of the flow cytometer is its ability to record multiple independent and quantitative measurements on a large number of cells (Parks, 1996). A flow cytometer takes in a sample of cells or cell particles suspended in solution, sending them in a single file past a laser beam.

Figure 1 highlights the high-level physics behind the measurement of forward scatter and side scatter. Forward scatter approximates cell size. Side scatter approximates internal structure of the cell, or granularity. Taken together, forward scatter and side scatter can help identify types of cells (Parks, 1996).

Fluorescence detectors measure the presence of cells or molecules that have been dyed during the preprocessing of the sample. Each particle on which data is recorded is called an event. Data collected by the flow cytometer during the processing of a sample is written to an output file.

Figure 1 Forward and side scatter



The data is written in a standard format such as FCS2.0 or FCS3.0 where FCS stands for Flow Cytometry Standard. The formats are specified by the Data File Standards Committee of the International Society for Analytical Cytology (International Society for Analytical Cytology (ISAC), 2004). The file format includes a header section, an ASCII text section specifying parameters of the data run and a section recording the data from the events. The data section is often written in a binary encoding. Thus, the file must be processed by a utility program for the event data to be translated to a human-readable form. Table 1 shows sample event data after processing.

Table 1 Sample event data

FS	*SS*	*FL1 LOG*	*FL2 LOG*	*FS LOG*	*SS LOG*
190	274	0	0	836	877
266	206	0	0	874	846
245	265	0	0	865	873
34	43	172	0	645	672
84	206	0	0	746	846
85	72	0	0	747	729
86	124	113	0	748	789
247	252	0	0	865	868
229	206	73	0	857	846

Research biologists at the University of Colorado Institute for Bioenergetics, located at the University of Colorado at Colorado Springs, USA provided the experimental data and guidance that inspired this work. Flow cytometry is integral to their research into cellular metabolism and cellular communication. These researchers are engaged in a project that considers the links between lipid availability and cell surface expression of Major Histocompatibility (MHC) class II molecules.

Putting this work into context for the nonbiologist, lipid rafts may support the transport of MHC class II molecules to the surface of the cell. These molecules aid in resistance to certain diseases. The project considered in this paper attempts to verify the lipid raft hypothesis by showing that mice raised on a high-fat diet have more surface expression of MHC class II than those raised on a low-fat diet. The presence of MHC class II can be detected through cytometric analysis. (Personal Communication, Schweitzer et al., 2004).

The project under consideration started with 112 mice. Half of the mice were fed with a high fat diet (5% coconut oil and 5% safflower oil) and half a low fat diet (5% safflower oil). After approximately 16 weeks, the mice were killed. The spleens were removed, and a suspension of splenocytes was prepared.

Sub-samples of the splenocyte suspension were then dyed with substances designed to fluoresce in the flow cytometer. Lysosomal acidity was measured by the fluorescence of LysoSensor stain. MHC class II expression was measured by the fluorescence of a phycoerythrin conjugated rat anti-mouse I-A/I-E. In experimental data, samples stained with this substance were labelled ‘M5114’. Additionally, some samples were stained with an isotype (IgA/IgE), which binds to all nonspecific matter, thereby acting as a control.

Some of the samples were treated with CytoPerm/CytoFix processing. This process, also used by Kumar et al. (2002), permeates the cell membrane, allowing intracellular staining. This treatment is labelled ‘CPCF’ in the experimental data.

The project included four strains of mice – Balb/c, C57/Black 6, UPC2 Knockout and P6129. These species are represented in the data as B, C, U and P individuals, respectively.

CellQuest (BD Biosciences) and FlowJo (www.flowjo.com) are tools that the researchers use to analyse the flow cytometry data. These packages present the data graphically, and provide summary statistics. Summary statistics include number of events, mean and geometric mean. The software also lets users manually define subsets or clusters of data. The graphical or statistical analysis can then be performed on those clusters. Such analytical techniques are used by Desbarats et al. (1999), Huber et al. (2001) and Lee et al. (2004).

2 Methodology

This current analytical environment of flow cytometry analysis is limited, closed and sample-centric. A different environment is required to leverage more powerful analytical tools, to support analysis on sets of samples and to include essential characteristics of samples. Such a rich analytical environment (RAE) is created by parsing the FCS data and loading the resulting data into a relational database. Additionally, the parsed data is associated with information about the individual, the sample type and the experiment.

Once the data is in a standard relational database, a variety of tools and techniques can be employed. Viable tools for analysing the data include SQL; programming languages such as Java, Perl and database stored procedures; and data analysis and graphing programs. Such programs can have a mathematical or scientific focus, such as MATLAB (www.mathworks.com) and Origin (www.originlab.com). Alternatively, they can have a business intelligence focus such as Business Objects (www.businessobjects.com) and Cognos (www.cognos.com). Many powerful tools for analysing data in relational databases are available. The research biologist can leverage these tools, once the data is exposed.

Figure 2 shows the data flow for building the RAE. The Java program written for this work, FCSParser, processes the native flow cytometry into two files, events and parameters. The event data is combined with information about the sample with which it is associated. This data includes individual diet type and strain and sample type and is obtained from the sample key, which is manually created by the biologists. This sample key is manually transformed from its word-processed format into a format suitable for loading into a database. An extract from the resulting file is shown in Table 2.

Figure 2 Data flow diagram

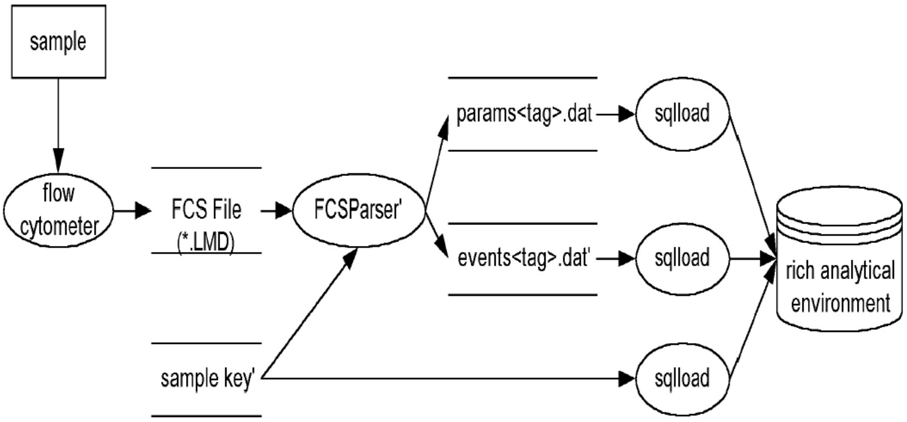


Table 2 Sample key data, reformatted for database

Experiment	*Sample number*	*Individual*	*Sample type*	*Replicate*	*Strain*	*Diet*	*Process*
1	14	BL-6	Lyso	1	B	L	NORMAL
1	15	BL-6	Lyso	2	B	L	NORMAL
1	16	BL-6	Lyso	3	B	L	NORMAL
1	17	BH-4	No stain	0	B	H	NORMAL
1	18	BH-4	Lyso	1	B	H	NORMAL
1	19	BH-4	Lyso	2	B	H	NORMAL

The data model for the RAE is based on the technique of dimensional modelling, common in data warehousing. This technique supports ease of use and high performance. The event is the fact in this dimensional model. The dimensions are INDIVIDUAL, TAG, EXPERIMENT, SAMPLE, SAMPLE_TYPE, REPLICATE, PROCESS, DIET and STRAIN. Most of these dimensions have only one attribute. As such, they are degenerate dimensions. Kimball et al. (1998) recommend collapsing the degenerate dimensions into the fact table.

An extract of the resulting data is shown in Table 3. Overall statistics on the data set are shown in Table 4. The large number of individuals (90) and the large number of samples (1419) highlight the difficulty of sample-based analysis in this project. Inspecting the data one sample at a time would be tedious and time-consuming.

Table 3 Event data, ready for loading

Tag	*Experiment*	*Sample number*	*Individual*	*Sample type*	*Replicate*	*Strain*	*Diet*	*Process*	*FS*	*SS*	*FL1LOG*	*FL2LOG*	*FSLOG*	*SSLOG*
G0036076	1	1	BL-3	No stain	0	B	L	NORMAL	75	164	387	0	734	821
G0036076	1	1	BL-3	No stain	0	B	L	NORMAL	197	203	451	0	841	844
G0036076	1	1	BL-3	No stain	0	B	L	NORMAL	173	129	432	0	826	794
G0036076	1	1	BL-3	No stain	0	B	L	NORMAL	181	116	383	0	831	782
G0036076	1	1	BL-3	No stain	0	B	L	NORMAL	181	161	344	0	831	818
G0036076	1	1	BL-3	No stain	0	B	L	NORMAL	201	212	399	0	842	848

Table 4 Selected data set statistics

	No. of samples	*No. of events*	*No. of distinct individuals*	*Avg events per sample*	*Avg events per individual*
Exp1	418	3,889,684	25	9,305	155,587
Exp2	360	2,791,145	24	7,753	116,298
Exp3	279	2,799,232	19	10,033	147,328
Exp4	362	3,119,628	22	8,618	141,484
Total	1419	12,599,689	90	–	–

3 Analysis

Once data from a flow cytometry experiment has been parsed and loaded into the RAE, a variety of analytical techniques can be employed. These techniques can be statistical or graphical. Multiple samples can be analysed with the same technique at the same time.

Summary statistics can be generated on all of the samples associated with a particular individual. These are shown in Table 5.

Table 5 Result set: study of individual BH-3

Tag	*Sample type*	*Replicate*	*Count (*)*	*Avg (FS)*	*Avg (FL2LOG)*	*Avg (FL2LOG/FS)*
G0037080	No stain	0	7352	198	2	0.0
G0037081	No stain	0	7409	195	3	0.0
G0037082	Lyso	1	7968	189	3	0.0
G0037083	Lyso	2	8791	179	4	0.0
G0037084	Lyso	3	8332	185	4	0.0
G0037184	No stain	0	6691	216	45	0.3
G0037185	Isotype	0	7127	214	59	0.4
G0037186	M5114	1	7097	211	535	3.3
G0037187	M5114	2	7002	212	544	3.3
G0037188	M5114	3	7207	212	534	3.2

The results show:

- a relatively consistent number of events in each sample
- increased fluorescence on Isotype and M5114 samples
- relative consistency across replicates of the same sample type.

Essentially, the summary statistics show expected results with a comforting level of consistency, thus providing a quick quality check.

However, in the larger data set, some samples contain an unusually high number of events. The flow cytometer is configured to process the sample until a certain number of events in a target range have been recorded. A high number of total events suggests that a particular sample is somehow different than the norm, and may require careful inspection. These samples may be suspect from a quality perspective.

In this experiment, the flow cytometer was configured to process each sample until 5,000 events were recorded within the target range. The average number of total events per sample was 8,886. The result set shown in Table 6 indicates which individuals are associated with samples that contain more than 20,000 events, and the number of such samples. A large number of such samples associated with a particular individual (e.g., CH-8, PL-1) may suggest that there is something unusual about that individual, or about the way in which samples drawn from the individual were prepared.

Table 6 Result set: suspect samples

Experiment	*Individual*	*Count* (*)
1	BL-3	3
1	BL-4	6
1	BL-5	2
1	PH-4	6
1	UL-4	1
2	CH-8	9
3	CH-2	7
3	PH-3	5
3	PL-1	10
3	UL-1	2

The RAE also supports the derivation of new measurements from the source data. One such measurement is normalised fluorescence. All other things being equal, the larger cells have more surface area for the stains to adhere to, and consequently, more fluorescence. Since forward scatter is an approximate measurement of cell size, one possible normalisation function is FL2LOG/FS. The analysis techniques discussed in the following section refer to this normalised fluorescence.

Recall that the experiment under consideration contains both normally processed samples and those processed with CytoPerm/CytoFix (CPCF). The hypothesis is that the CPCF samples will have more fluorescence, owing to the intracellular staining. The RAE allows us to select those samples treated with M5114, and compare the normal samples to the CPCF samples. A subset of the resulting data is shown in Table 7. The expected finding of more fluorescence (FL2LOG) on CPCF samples is not consistently shown. However, if fluorescence is normalised (FL2LOG/FS), a higher value is consistently shown on the CPCF samples. This result highlights the value of the normalisation technique.

Table 7 Result set (extract) – M5114 samples, normal and CPCF processing

Tag	*Individual*	*Replicate*	*Process*	*Avg (FL2LOG)*	*Avg (FL2LOG/FS)*
G0036266	BH-5	1	NORMAL	597	3.4
G0036267	BH-5	2	NORMAL	600	3.3
G0036268	BH-5	3	NORMAL	594	3.2
G0036404	BH-5	1	CPCF	442	3.5
G0036405	BH-5	2	CPCF	424	3.5
G0036406	BH-5	3	CPCF	562	4.3
G0036407	BH-5	3	CPCF	570	4.4
G0036271	BH-6	1	NORMAL	599	3.7
G0036272	BH-6	2	NORMAL	625	3.6
G0036273	BH-6	3	NORMAL	595	3.5
G0036410	BH-6	1	CPCF	568	4.7
G0036411	BH-6	2	CPCF	576	4.8
G0036412	BH-6	3	CPCF	585	4.9
G0036241	BL-3	1	NORMAL	617	3.8
G0036242	BL-3	2	NORMAL	623	4.0
G0036243	BL-3	3	NORMAL	639	3.9
G0036377	BL-3	1	CPCF	608	5.1
G0036378	BL-3	2	CPCF	611	5.0
G0036379	BL-3	3	CPCF	616	5.1

Graphical techniques also can be employed. In this work, graphs were created with a Java program incorporating JfreeChart (www.jfree.org) libraries. A SQL statement is submitted to the database, and the result set is presented graphically.

Figure 3 shows a family of histograms of Normalised FL2 for all samples for a particular individual. This technique shows the similarity or dissimilarity of the replicate samples. It also shows the similarities and the dissimilarities of different sample types.

Figure 4 contains histograms of forward scatter for all M5114 samples of a particular individual. Again, it shows the similarity of the replicate samples. It also shows that different processes have a fundamentally different distribution of forward scatter. Upon first thought, one would not expect the sample type to influence the size of the cells. However, the CPCF process perforates the cell membrane, thereby visibly altering the forward scatter. The cells essentially become smaller, reinforcing the importance of the normalisation technique discussed above.

Figure 3 Normalised FL2, individual BL-6

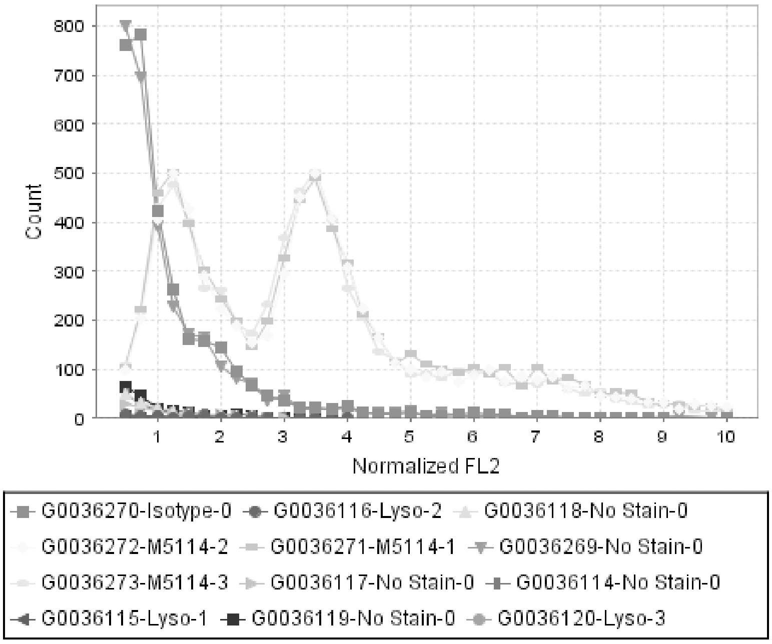
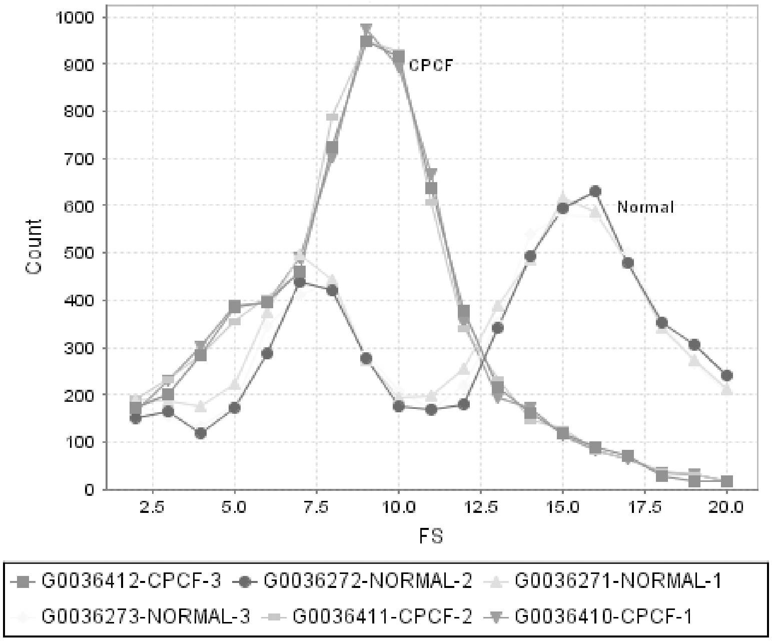


Figure 4 Forward scatter, M5114 samples



4 Discussion

The work presented in this paper demonstrates that a rich analytical environment can be created from flow cytometry data. This analytical environment allows the analyst to perform types of analysis that were not previously possible, and to gather more knowledge more quickly. The essential features of the environment are:

- an open architecture in which multiple analytical tools such as SQL and graphical libraries can be leveraged
- the combination of flow cytometry data with sample characteristics such as strain, diet and sample type
- the ability to derive new metrics from core data, such as the calculation of normalised fluorescence
- the ability to perform analyses across experiments, as opposed to on a small number of samples at a time.

Furthermore, because the environment is so rich from an analytical perspective, the possibilities for future work are significant.

5 Conclusion

The RAE empowers both the biologist and the analyst. Many types of analysis are possible when all of the data from an experiment is made available in an open environment. As biologists become more familiar with what they can accomplish with this environment, certain processes will become standard. Other processes will emerge as innovative and exciting.

References

Desbarats, J., Wade, T., Wade, W.F. and Newell, M.K. (1999) 'Dichotomy between naïve and memory CD4+ T cell responses to Fas engagement', *Proceedings of the National Academy of Sciences*, Vol. 96, No. 14, pp.8104–8109, Retrieved 25th March, 2004 from http://www.pubmedcentral.nih.gov/articlerender.fcgi ?tool=pubmed&pubmedid=10393955.

Huber, S.A., Sakkinen, P., David, C., Newell, M.K. and Tracy, R.P. (2001) 'T helper-cell phenotype regulates atherosclerosis in mice under conditions of mild hypercholesterolemia', *Circulation*, Vol. 103, pp.2610–2616, Retrieved 25th March, 2004 from http://circ.ahajournals.org/cgi/reprint/103/21/2610.pdf.

International Society for Analytical Cytology (ISAC) (2004) *Data File Standard for Flow Cytometry*, Version FCS3.0, Retrieved 25th March 2004, from http://www.isac-net.org/.

Kimball, R., Reeves, R., Ross, M., and Thornthwaite, W. (1998) *The Data Warehouse Lifecycle Toolkit: Expert Methods for Designing, Developing, and Deploying Data Warehouses*, John Wiley & Sons, Inc., New York, 771 pages.

Kumar, L., Pivniouk, V., de la Fuente, M., Laouini, D. and Geha, R. (2002) 'Differential role of SLP-76 domains in T cell development and function', *Proceedings of the National Academy of Sciences*, Vol. 99, No. 2, pp.884–889, Retrieved 17th April, 2004 from http://www.pnas.org/cgi/reprint/99/2/884.pdf.

Lee, J., Shin, J., Kim, E., Kang, H., Yim, I., Kim, J. *et al.* (2004) 'Immunomodulatory and antitumor effects in vivo by the cytoplasmic fraction of Lactobacillus casei and Bifidobacterium longum', *Journal of Veterinary Science*, Vol. 5, No. 1, pp.41–48, Retrieved 25th March, 2004, from http://www.vetsci.org/2004/pdf/41.pdf.

Parks, D.R. (1996) 'Flow cytometry instrumentation and measurement', in Herzenberg, L., Herzenberg, L., Blackwell, C. and Weir, D. (Eds.): *The Handbook of Experimental Immunology*, Blackwell Science, Boston, pp.47.1–47.12, Retrieved on 31st March, 2004 from http://herzenberg.stanford.edu/Publications/Reprints/LAH413.pdf.

Schweitzer, S.C., Reding, A.M., Ford, C.A., Villalobs-Menuey, E., Huber, S.A. and Newell, M.K. (2004) forthcoming.