A standardized and adaptable data format for flow cytometric data transmission for statistical analyses

Donald K. Carter[1], Cindy Molitor[2], Cristine Cooper[2], Matthew Bellew[3], Kevin Krouse[3], Steven G. Self[2], and Stephen C. De Rosa[1,4]

[1]HIV Vaccine Trials Network Laboratory Program, Vaccine and Infectious Disease Institute (VIDI), Fred Hutchinson Cancer Research Center (FHCRC); [2]Statistical Center for HIV/AIDS Research and Prevention, VIDI, FHCRC; [3]LabKey Software Foundation (www.labkey.org); [4]Department Laboratory Medicine, University of Washington School of Medicine, Seattle, WA.

Introduction

Although previously restricted to selected research laboratories, advanced multi-parameter flow cytometric assays are becoming more commonly used in the clinical setting. In our clinical trials laboratory, we perform multiple 10- to 12-color assays in high-throughput fashion to monitor immune responses to candidate HIV vaccines. A challenge has been developing a format for data transmission to statisticians that is both flexible to allow for different staining panels and yet standardized to allow for shared programming to extract the appropriate statistics.

Methods

All of our assays characterize lymphocytes and therefore the gating hierarchy to identify lymphocytes is shared between assays. Further gating to identify T, B and NK cell subpopulations is then assay-specific. Our data format "stacks" the statistics for each of these additional gates using LabKey Flow. Each gate has two associated fields, one for the name of the gate or population and the other for the statistic such as the event count. These additional gates/populations are hierarchical so that subsequent gates represent a subpopulation of the prior gate. The file format is generic with fields named numerically as "Subset 1 Name", "Subset 1 Count", etc. Another field also defines the specific assay performed and a separate document, referred to as the analysis plan, details the specific output that the assay will be generating. These plans are used to construct metadata tables that define the variables generated for each analysis. The data in the generic file format are programmatically parsed into separate datasets for each analysis. Another feature of our data format are field values specifying control samples and the type of control. For example, we include a sample control in each experiment that is later extracted separately for trend analysis.

File Format Specification

The specification includes predefined keywords that uniquely identify the person/animal providing the tissue/blood, the experimental conditions applied to the sample, and derived statistics from flow analysis.

Table 1. Examples of fields for identification of the sample.

Field	Filed Name	Type	Notes
Protocol Number	PROTOCOL	Numeric	For clinical trial samples
Volunteer ID	PTID	Numeric OR Character if PTIDTYPE is O	
PTID Type	PTIDTYPE	Character	S - SCHARP, M – Merck, O – Other
Visit Number	VISITNO	Numeric	
Sample Draw Date	DRAWDT	Date	Format: MM/DD/YYYY or DDMONYYYY
Sample Type	SAMPLE_TYPE	Character	Blood, PBMC, Mucosal-sigmoid, Mucosal-rectal
Cryostatus	CRYOSTATUS	Character	Fresh, Cryopreserved

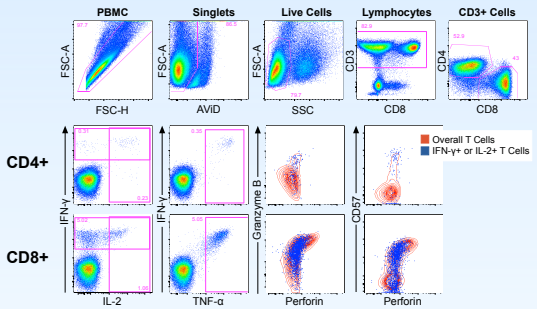
Table 2. Examples of fields for experiment details.

Field	Filed Name	Type	Notes
Lab ID	LABID	Character	DU, CA, FH, EM, VR
Experiment Assay ID	EXP_ASSAY_ID	Numeric	Code for type of assay (ICS, CFSE, B cell)
Test Date	TESTDT	Date	Format: MM/DD/YYYY or DDMONYYYY
Assay ID	ASSAYID	Character	Experiment name such as batch number
Antigen Identifier	ANTIGEN	Character	Values from antigen list
Replicate Number	NREPL	Numeric	
Purpose of well	WELLROLE	Character	To identify experiment controls, compensation controls, etc. E.g., NegCtrl-specimen PosCtrl-specimen TestAg-specimen NegCtrl-ExptCtrl PosCtrl-ExptCtrl

Table 3. Other fields are for derived statistics and are named generically to allow for different analysis schemes.

Field	Filed Name	Type	Notes
Total number of cells collected	COLLECTCT	Numeric	
Type of first cell subset	SUBSET1	Character	The analysis plan associated with each experiment assay will define acceptable values for each subset
Number of events in first cell type gate	SUBSET1_NUM	Numeric	All flow statistics are event counts rather than frequency
Type of second cell subset	SUBSET2	Character	
Number of events in second cell type gate	SUBSET2_NUM	Numeric	
	Etc., up to 10		

Figure 1. Example of intracellular cytokine staining.



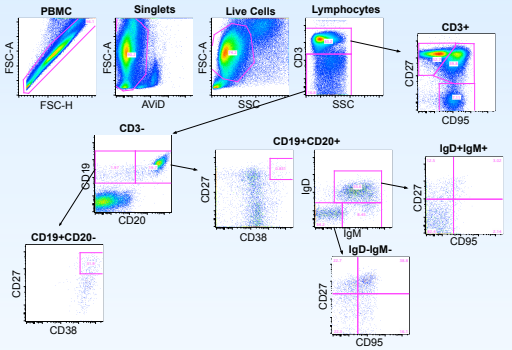
Intracellular cytokine staining assay gating scheme.

Example staining for PBMC from one healthy adult in response to CMV stimulation. The progressive gating scheme is as shown: singlets, live cells, lymphocytes, CD3+ T cells. T cells are then gated as CD4+ or CD8+. For each of these, three gates identify the cytokine-producing cells. This provides marginal frequencies for these, i.e., cells producing that cytokine alone or in combination with the other cytokines. Not shown are the gates for granzyme B and CD57. Boolean operators are used to derive any combinations of these cytokines/markers. For example IFN-γ OR IL-2 OR TNF-α identifies any cytokine producing cell.

Table 4. Field "mapping" for intracellular cytokine staining.

Subset 1	Subset 2	Subset 3	Subset 4	Subset 5	Subset 6
Singlets	Live	Lymphocytes	CD3+	CD4+	IFNg+
					IL2+
					TNFa+
					GzB+
					IFNg_OR_IL2_OR_TNFa
					IFNg+IL2+TNFa+
					IFNg+IL2+TNFa-
					IFNg+IL2-TNFa-
					IFNg+IL2+TNFa+GzB+57+
					Etc.
				CD8+	Same as for CD4+

Figure 2. Example of B cell phenotyping assay requiring multiple "branches" in the analysis scheme.



B cell phenotyping assay gating scheme.

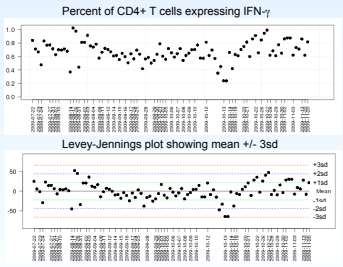
Example staining for PBMC from one healthy adult. The progressive gating scheme begins as shown: singlets, live cells, and lymphocytes, then branches to CD3+ T cells and CD3- NK/B cells. CD19+ B cells are subdivided into CD20- or CD20+. Plasmablasts (CD27+CD38^{br}) are restricted to the CD20- subset. The CD20+ B cells are further gated on the markers as shown.

Table 5. Field "mapping" for B cell phenotyping assay.

Subset 1	Subset 2	Subset 3	Subset 4	Subset 5	Subset 6	Subset 7
Singlets	Live	Lymph	CD3+	27+95+		
				27+95-		
				27-95+		
			CD3-	19+20-	27+38br	
				19+20+	27+38br	
					IgD+IgM+	27+
						95+
					IgD-IgM-	27+
						95+

Figure 3. Assay quality control and trend analysis.

A useful control to include in each experiment is a test sample with a known response to an antigen and/or predetermined frequencies of populations of interest. This can be PBMC from an individual who has been leukapheresed and multiple vials cryopreserved. Trending can be monitored to ensure reproducible assay performance over time. Appropriate use of keywords (see "WELLROLE") to identify these "experiment" controls allows for automated trending as shown in the example below from the LabKey Flow software used in our laboratory.



Trend analysis over time.

Data for a PBMC control sample included in each experiment. Shown is the CD4 response by ICS for CMV stimulation. Below is the Levey-Jennings plot for these data. Note that the two lowest data points were investigated and found to be due to an error restricted only to the control sample and not the test samples.

Conclusions

High-throughput performance of flow cytometric assays also requires high-throughput data analysis. A practical requirement in the analyses of these complex data sets is data transmission for statistical analysis. The data format we developed standardizes this step and yet is highly adaptable to new assays and revised gating of existing data sets. A major advantage is that by "stacking" the data and limiting the number of fields, new assays and new analyses only require addition of new field values rather than new fields. A disadvantage is the requirement for specialized data formatting, therefore this is most appropriate for standardized assays and analyses where the data format can be pre-programmed.