Published in final edited form as:

J Immunol Methods. 2011 January 5; 363(2): 143–157. doi:10.1016/j.jim.2010.08.004.

QUALITY ASSURANCE OF INTRACELLULAR CYTOKINE STAINING ASSAYS: ANALYSIS OF MULTIPLE ROUNDS OF PROFICIENCY TESTING

Maria C. Jaimes^{a,*}, Holden T. Maecker^b, Ming Yan^a, Vernon C. Maino^a, Mary Beth Hanley^a, Angela Greer^c, Janice M. Darden^{d,1}, and M. Patricia D'Souza^d

^a BD Biosciences, 2350 Qume Drive, San Jose, CA 95131

^b Stanford University School of Medicine, 300 Pasteur Drive, Stanford, CA 94305

^c PPD Inc., 929 North Front Street, Wilmington, NC 28401

^d Division of AIDS, National Institute of Allergy and Infectious Diseases, National Institute of Health, 6700-B Rockledge Drive, Bethesda, MD 20892

Abstract

When evaluating candidate prophylactic HIV and cancer vaccines, intracellular cytokine staining (ICS) assays that measure the frequency and magnitude of antigen-specific T-cell subsets are one tool to monitor immunogen performance and make product advancement decisions. To assess the inter-laboratory assay variation among multiple laboratories testing vaccine candidates, the NIH/NIAID/DAIDS in collaboration with BD Biosciences implemented an ICS Quality Assurance Program (QAP). Seven rounds of testing have been conducted in which 16 laboratories worldwide participated. In each round, IFN- γ , IL-2 and/or TNF- α responses in CD4+ and CD8+ T-cells to CEF or CMV pp65 peptide mixes were tested using cryopreserved peripheral blood mononuclear cells (PBMC) from CMV seropositive donors. We found that for responses measured above 0.2%, inter-laboratory %CVs were, on average, 35%. No differences in inter-laboratory variation were observed if a 4-color antibody cocktail or a 7-color combination were used. Moreover, the data allowed identification of important sources of variability for flow cytometry-based assays, including: number of collected events, gating strategy and instrument setup and performance. As a consequence, in this multi-site study we were able to define pass and fail criteria for ICS assays, which will be adopted in the subsequent rounds of testing and could be easily extrapolated to QAP for other flow cytometry-based assays.

Keywords

Flow Cytometry; Intracellular Cytokine Staining; Quality Assurance

*Corresponding author. Tel: 408-954-6396. Fax: 408-954-2156. maria_jaimes@bd.com.

¹Present address: Social and Scientific Systems, Inc., 8757 Georgia Avenue, 12th Floor, Silver Spring, MD 20910

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

1. INTRODUCTION

The intra-cellular cytokine staining (ICS), enzyme-linked Immunospot (ELISpot) assay and staining with HLA-peptide multimers are technologies commonly used for the monitoring of antigen-specific immune responses. ICS has the advantage over these other techniques in that this flow-based application simultaneously permits functional and phenotypic assessment of the responding T-cell populations.

In humans, adaptive cellular immune responses play a vital role in the containment of HIV-1 replication. During primary infection, the appearance of HIV-specific cytotoxic T-lymphocytes (CTL) is correlated with decline from peak viremia (Goonetilleke et al., 2009). Moreover, the long-term, non-progressor status is associated with robust CTL responses (Rinaldo et al., 1995; Harrer et al., 1996; Betts et al., 1999), and the loss of HIV-specific T-cells is associated with rapid progression to AIDS (Klein et al., 1995). Because control of infection is required to prevent disease, and as the best licensed vaccines against other pathogens do not necessarily prevent these infections completely, a successful HIV vaccine will probably also need to elicit cell-mediated immune (CMI) responses capable of controlling HIV infection. Consequently, utilizing validated assays of CMI responses would enhance comparisons among various vaccine developers and enable data-driven prioritization of candidate vaccines.

Numerous vaccine clinical trials, conducted at many sites simultaneously, are currently testing candidate prophylactic HIV vaccines and use ICS to monitor immunogen performance and make product advancement decisions (Cheng et al.; Koup et al.; De Rosa and McElrath, 2008; McElrath et al., 2008). The interpretation of the results obtained from these ICS assays across different vaccine developers is a difficult task, due to the variety of methods, protocols and statistical criteria available to detect vaccine-specific T-cell responses. To make product advancement decisions, it is necessary to compare data across different trials; consequently, a standardization and Quality Assurance of ICS assay is critical. Moreover, such a Quality Assurance Program (QAP) would provide ongoing proficiency data for participating institutions to meet Good Clinical Laboratory Practice (Ezzelle et al., 2008; Sarzotti-Kelsoe et al., 2009). Benefits of the QAP include: opportunity for participants to monitor their own performance over time; use of the QAP as an internal competency test for staff once trained and qualified; and an ability to compare performance with peers running the same assay.

Published studies have addressed the intra- and inter-assay precision of ICS in whole blood and peripheral blood mononuclear cells (PBMC) (Nomura et al., 2000; Horton et al., 2007; Maecker et al., 2008; Nomura et al., 2008). A recent study by our group on standardization and precision of ICS between laboratories (Maecker et al., 2005) revealed that ICS could be performed by multiple laboratories using a common protocol with good inter-laboratory precision (18–24%). This precision improves as the frequency of responding cells increases. In an effort to standardize the assays across laboratories, in 2005, we created a QAP for ICS assays. This program was developed to assess the inter-laboratory variability when sharing a common standardized protocol and reagents. Here, we present the data from seven consecutive rounds of testing. A total of 16 laboratories from seven different countries participated in the study in which pre-tested PBMC, along with lyophilized antigens and antibodies, were distributed. The laboratories were requested to determine the percentage of cytokine+, CD4+ and CD8+ cells in each sample. The analysis of the data generated in this program has allowed us to identify factors responsible for ICS variability among laboratories that need to be taken into consideration when performing Quality Assurance of flow cytometry assays and reporting data for vaccine clinical trials.

2. MATERIALS AND METHODS

2.1. Participating institutions

In the first round of testing, ten laboratories worldwide participated. This number increased to 16 by Round 5. A list of the participants is provided in Table 1. All participants have agreed on the content of this publication. Of note, most of these laboratories were involved in a previous study aimed at standardizing the protocol used in this ICS QAP (Maecker et al., 2005).

2.2. PBMC preparation and cryopreservation

Concentrated leukocytes were prepared by machine leukopheresis with anticoagulant ACD-A by BRT Laboratory (Baltimore, MD). PBMC were isolated within eight hours post-collection using a ficoll gradient. Briefly, an average of 11ml of leukocytes were diluted with phosphate buffered saline (PBS) to 35ml and underlaid with 12 ml of ficoll. Following centrifugation for 30 minutes at 450g (at room temperature), the cell layer was collected; the cells were washed three times with PBS and re-suspended in RPMI-1640 media, supplemented with 10% heat-inactivated fetal bovine serum (FBS) (cRPMI-10). Cell concentration was determined using the Guava ViaCount assay (Guava Technologies, Inc., Hayward, CA), and PBMC were frozen at 15×10^6 cells/mL in freezing media (22% FCS, 7.5% DMSO and 70.5% RPMI).

Pre-screening of the PBMC donors for CMV responses was initially done at SeraCare Life Sciences (Gaithersburg, MD) using an IFN- γ and IL-2 ELISpot assay. Later, the central laboratory (BD Biosciences, San Jose, CA) performed an ICS assay and selected the donors for each round.

Two vials of the cryopreserved PBMC for each donor were shipped to participant laboratories using a liquid nitrogen dry shipper. A recommended thawing procedure was provided to the participants.

2.3. Lyophilized stimulation and staining reagents

We used preconfigured lyophilized stimulation and staining plates, which had been previously validated (Dunne 2004) to provide simplified assay setup.

Peptide stimuli, together with brefeldin A (Sigma-Aldrich Corp, St. Louis, MO), were provided in lyophilized form within appropriate wells of a polypropylene V-bottom 96-well plate. The peptide stimuli used consisted of mixtures of peptides for the CMV pp65 protein (138 peptides consisting of 15 amino acid residues, overlapping by 11 aa residues each, BD Biosciences, San Jose, CA; used at a final concentration of 1.7 μ g/ml/peptide) and CEF (containing 32 peptides of 8–11mers, epitopes from CMV, EBV NA, and Influenza NP proteins used at a final concentration of 1.0 μ g/ml/peptide, SynPep, Dublin, CA). Of note, CEF was expected to induce CD8, rather than CD4 responses (Currier et al., 2002).

Lyophilized staining antibody cocktails were provided in the corresponding wells of a second plate. All antibodies were obtained from BD Biosciences. The following antibody mixture was used for all rounds: CD4 FITC/IFN- γ + IL-2 PE/CD8 PerCP-Cy5.5/CD3 APC. In addition, during Rounds 1 and 2, the following cocktails were also tested: IFN- γ FITC/CD69 PE/CD4 PerCP-Cy5.5/CD3 APC and IFN- γ FITC/CD69 PE/CD8 PerCP-Cy5.5/CD3 APC. Finally, starting in Round 6, a 7-color cocktail was introduced (TNF FITC/IFN- γ PE/CD8 PerCP-Cy5.5/CD4 PE-Cy7/IL-2 APC/CD3 V450/viability dye) and tested in parallel to the 4-color cocktail described above. The viability marker used was Live Dead Fixable Aqua fluorescent reactive dye (Molecular Probes, Inc., Eugene, OR).

Lastly, starting in Round 6 of testing, BD™ Cytometer Setup & Tracking (CS&T) beads and Spherotec 8-peak beads (BD Biosciences, San Jose, CA) were provided in the staining lyoplate, in order to assess instrument performance.

2.4. Assay guidelines

Participants were provided with a recommended protocol previously standardized in a concerted effort with some of the participating laboratories (Maecker et al., 2005). In Rounds 1 and 2, four donors were tested against three stimuli (BFA only, CEF and CMVpp65), and the cytokine responses were measured by using two different antibody cocktails (described above). From Rounds 3 to 5, only one antibody cocktail was used. This allowed a triplicate test of the responses of three donors to each stimuli. Beginning in Round 6, a 4- and 7-color cocktail were tested in parallel. In order to preserve the possibility of testing each condition in triplicate, the number of stimuli was decreased from three to two (BFA only and CMVpp65). Despite the changes introduced between proficiency testing rounds, at least one identical condition was always maintained between rounds (i.e: same donor, same stimulus and same antibody cocktail) thus allowing comparisons from one round to the other.

Briefly, after thawing, PBMC were rested overnight in cRPMI-10 at 37°C 5% CO₂. The following day, cells were counted, viability was determined and cells were washed and re-suspended at a concentration of five million cells/ml. Cells were added to the appropriate wells of the lyophilized stimulation plate and mixed in order to reconstitute the lyophilized pellets. The plate was then incubated for six hours at 37°C. After the stimulation, participants were given the choice to either store the plate overnight at 4°C or 18°C, or immediately proceed to the processing and staining of the cells. Of note, this was one of the rare steps in the protocol where participants were given flexibility since from previous optimization studies it was determined that time and temperature intervals did not have an impact on detected responses.

Activated cells were treated with 2mM EDTA for 15 minutes at room temperature and washed using wash buffer (PBS 0.5% BSA 0.1% NaN₃, or equivalent buffer). It is important to note that participants followed two different strategies to aspirate the wells after the required washing steps; some used vacuum manifold aspiration (recommended in the protocol) and others drained and flicked the plate by hand. Cells were then lysed using FACS Lysing Solution (BD Biosciences) and were incubated for ten minutes at RT. At this point, participants were also given the choice of freezing the cells at -80°C or continuing with the permeabilization and staining steps. Following fixation, cells were washed and permeabilized for ten minutes at RT using FACS Permeabilizing Solution 2 (BD Biosciences). Cells were then washed twice. During the second wash, the wells of the lyophilized staining plate were hydrated with wash buffer and then transferred to the appropriate well in the antigen plate, mixed and incubated for 60 minutes at RT in the dark. Cells were then washed twice and re-suspended in PBS + 1% paraformaldehyde. Laboratories were provided with common stocks of 20mM EDTA solution, FACS Lysing Solution and FACS Permeabilizing Solution 2.

2.5. Collection and analysis of results

The participating institutions had four weeks to complete the assay and return their results after each round of testing.

Participating laboratories uploaded a spreadsheet with their calculated results (percentage of CD4+, CD8+ and cytokine+ cells detected for each sample) via a website, along with their raw data in the form of FCS files. They were also required to fill out an online questionnaire,

which tracked protocol variables such as cell viability and recovery, resting of cells, stimulation time, storage of stimulated samples, staining variables, flow cytometer used, acquisition criteria, and instrument setup procedure. The website allowed us to efficiently collect the data and helped to enforce conventions in data labeling and spreadsheet formatting; thus, allowing reliable retrieval of data.

Once data were analyzed, a report summarizing the results from each round was distributed to all the participant laboratories. Each laboratory was identified in the report with a number, and those numbers were kept confidential and changed from one round to the other. There is no correlation between the numeric order of laboratories in Table 1 and in the Figures.

2.6. Statistical analysis

The percentage coefficient of variation (% + CV) was calculated as $100 \times \text{Standard Deviation (SD)} / \text{mean}$ for each sample, from the percentage of cytokine-positive cells reported by each laboratory, or derived from centralized analysis of that sample. The mean CV for each experiment was taken as the average of the entire individual sample CVs. Statistical significance of the differences in the averages of CVs between experiments was calculated using a Kruskal-Wallis test, followed by a Dunn's Multiple Comparison test. The significance of the difference between data from the 4- versus 7-color cocktail was calculated by comparing the CVs of these experiments using a Wilcoxon signed rank test for matched pairs. Correlation coefficients were calculated for comparisons between central and individual analysis, as well as gold standard versus laboratories cytokine responses. Statistical analyses were done using GraphPad Prism software (San Diego, CA).

3. RESULTS

3.1. Inter-laboratory variation detected using a 4-color ICS assay

Cytokine responses from the different donors to each stimuli tested were compared across laboratories in each round. The variability across laboratories was calculated as %CV or SD. As noted before (Maecker et al., 2008), %CV were not informative for low cytokine responses. Based on all the data generated in the current study, we established that %CV would stabilize and could only be used when assessing responses higher than 0.2%; hence, SD was used for cytokine responses lower than 0.19%.

The 4-color cocktail CD4 FITC/IFN- γ + IL-2 PE/CD8 PerCP-Cy5.5/CD3 APC was used in all rounds of testing. An example of the staining obtained using this combination is presented in Figure 1.

The initial rounds of testing included donors with high, intermediate and low responses to CMVpp65 and/or CEF. Beginning in Round 6, only donors with intermediate and low responses were included, since detection of these levels of antigen-specific responses was the most frequent and challenging scenario in the analysis of unknown clinical specimens.

Examples of the different levels of responses obtained across donors and stimuli are presented in Figure 2A. In these plots, the responses detected in each replicate for each laboratory are represented, and the variation detected across laboratories is indicated at the top of each graph. A summary of all %CVs or SDs calculated for each response in each round is presented in Figure 2B. For cytokine responses higher than 0.2%, the %CVs detected across rounds were overall very similar (mean of 32.8%, 26.4%, 27.6%, 34.2%, and 36.3% in Rounds 2, 3, 4, 5, and 6, respectively). However, in Rounds 1 and 7, the %CV were higher (mean of 57.4% and 64.4% for Rounds 1 and 7, respectively) compared to the ones in other rounds; although there was not a statistically significant difference ($p > 0.05$). For lower responses, the SD followed a similar trend with the smaller variations detected in

Rounds 2, 3, 5, and 6 (range of mean of all SD from 0.01 to 0.08) and higher variations in Rounds 1, 4 and 7 (0.59, 0.25 and 0.11, respectively; $p < 0.0001$ for Round 1 versus Rounds 2, 3, 5, and 6).

Finally, it is also significant to specify that when analyzing the responses to the different stimuli, background subtraction was not performed. The rationale behind this decision was that if an error had occurred in the negative control wells, such an error will be carried over to the stimulated ones by doing the subtraction. However, we did compare the variation across laboratories with and without background subtraction (data not shown), and no significant differences were observed.

3.2. Inter-laboratory variation detected using a 7-color ICS assay

In order to evaluate an ICS assay that would mimic the daily practice in most of the participating laboratories, a 7-color cocktail was included to begin in Round 6. This cocktail was designed so that poly-functional T-cells (i.e. T-cells producing more than one cytokine) could be identified. A survey conducted across the participating institutions revealed that most laboratories were interested in detecting IFN- γ , IL-2 and TNF- α in separate channels. Since all laboratories utilized frozen specimens, the inclusion of a viability dye was suitable (Horton et al., 2007). The resulting 7-color cocktail, unlike the 4-color cocktail, was designed in the central laboratory and has not been formally validated. The detection of cytokine responses using this combination is presented in Figure 3A. Using this 7-color combination, the cytokine responses that were consistently detected across donors after CMVpp65 stimulation were CD4+ and CD8+ T-cells expressing all three cytokines (IFN- γ +, IL-2+, TNF- α +) or IFN- γ and TNF- α . These specific subsets were easily identified by all participants, across all donors in Rounds 6 and 7. The variation observed in the quantification of poly-functional T-cells was similar to the one obtained when only looking at a single cytokine or at the combination of cytokines in one channel, as done in the 4-color cocktail (data not shown).

In order to directly compare responses obtained with the 4-color versus 7-color cocktails, the frequencies of all subsets identified as producing IFN- γ and/or IL-2 with the 7-color cocktail were summed for both CD4+ and CD8+ T-cells. Interestingly, increasing the number of colors in the assay, and hence its complexity, did not result in an increase in the variability observed across laboratories. The responses detected by the different participants, using both combinations, is shown as an example in Figure 3B. Moreover, the mean of the %CVs calculated across responses detected with both cocktails were very similar (50.0% versus 53.4% in Round 6 and 62.1% versus 59.8% in Round 7), as well as the mean of the SD for low responses (0.09 for both cocktails in Round 6 and 0.17 versus 0.19 in Round 7; Figure 3C).

3.3. Definition of optimal range and outlier measurements

Because of the variability across laboratories, we needed to create a robust method for defining outlier measurements. Therefore, at the start of Round 3, the central laboratory performed five to six assays, with at least two different operators on different days, using the same batch of reagents distributed to the laboratories and acquiring the cells in different instruments (FacsCalibur, Canto II and LSR II). We considered that the cumulative data from those experiments reflected the variability inherent to this assay and provided a range of where the responses should fall. We refer to the results obtained from these series of experiments in each round as the Gold Standard (GS). In most of the measurements, the range defined by $\pm 2SD$ of the mean of the GS was considerably narrower than the one defined by the mean $\pm 2SD$ of the measurements from all laboratories. In addition, the mean $\pm 2SD$ of the GS also included most of the measurements that fell between the 25th to 75th

percentiles of all measurements. Moreover, for a given response, the mean and median of the GS were almost identical (correlation coefficient ≥ 0.9 , data not shown) and highly correlated with the median of the responses reported by all participants (Figure 4A). Taking this into consideration, we determined that if a response reported by a participant was not within $\pm 2SD$ of the GS, it would be considered an outlier. Using this strategy, we were then able to quantify the number of outlier measurements for each laboratory in each round. As shown in Figure 4B for Rounds 3, 4 and 5 in which we tested the same donors and used the same stimuli, there was considerable variation in this number from one laboratory to another. In general, if a laboratory had a high number of outlier measurements in a given round, this number decreased in subsequent rounds (clear examples are Laboratories 3, 11 and 13). After completion of every round, the central laboratory generated a detailed report summarizing the performance of each laboratory and provided feedback regarding any issues encountered during that particular round. This response very likely contributed to improved performance for successive rounds, if the operator was consistent between rounds. Variation in the GS measurement from one round to the next would reflect variation across different days, different operators and different lots of reagents. We found (data not shown) that the GS was consistent in detecting almost the same range of responses between rounds. In addition, the %CVs or SDs were similar and significantly lower than the ones calculated with the data from all laboratories.

After careful analysis of the data generated by all the laboratories in the seven rounds of the ICS QAP, we have identified critical factors to consider when evaluating laboratory performance of ICS assays. We will describe these factors and present data supporting the most effective factors contributing to intra-laboratory variation and generation of outlier measurements.

3.4. Viability and recovery

When performing the assays, participants were instructed to determine the percentage of viability and recovery using their preferred technique (trypan blue and hemocytometer, Guava, etc).

Importantly, the viability of the cells across rounds was very good and in general, higher than 85% (Table 2). The protocol distributed to participants contained a recommended thawing procedure. However, isolated cases of low viability were reported and always related to inappropriate thawing technique, such as using non-optimal media or thawing too many sample vials at a time. As expected, low viability was consistently associated with suboptimal responses (data not shown).

Notably, the viability of the cells obtained in laboratories within the US compared to laboratories overseas (Asia and Africa, for example) were almost identical (Supplemental Figure 2S), suggesting that shipping frozen PBMC using liquid nitrogen containers is suitable for these types of studies. It is important to point out that the requirement of this shipping methodology had been established in previous studies (Bull et al., 2007), and it was determined that despite the high cost associated with this shipping strategy, it was critical to adequately preserve the specimens.

In initial rounds of this study, we did not include a viability marker in the staining combination, but a viability marker was included at the beginning of Round 6 in the 7-color cocktail. Using this staining, we were able to determine that across institutions the percentage of viable cells recovered after the assay was comparable (data not shown).

In contrast to the consistent viability values, we detected considerable variation in the number of cells recovered in each laboratory (Table 2). Recovery could range from 40 to

140% for a given donor. We could not find a correlation between % recovery and outcome of the assay (as determined by the number of outlier measurements for that donor; data not shown). We hypothesize that this could indicate discrepancies in cell counting methodologies used across laboratories. Although ICS assays, in contrast to other cell-based assays such as ELISpot or proliferation, are not sensitive to variation in the cell input (similar results are obtained if one or two million cells are stimulated; data not shown), it is necessary that a minimum number of cells are stimulated to generate enough events for acquisition and analysis as discussed below.

3.5. Number of collected events

When analyzing the individual FCS files generated in initial rounds, it appeared that the participating laboratories had acquired a wide range of total cell numbers per well, and that within the same laboratory, different numbers were acquired across different wells as shown in Figure 5A. Although the protocol for Rounds 1 through 5 stated that a minimum of 80,000 CD3+ lymphocytes were needed to be acquired per well, some laboratories were not in compliance, acquiring too few events (20,000 or fewer CD3+ lymphocytes). Starting in Round 3, the central analysis of the data included evaluation of total numbers of CD3+ lymphocytes. The protocol was modified, and it was emphasized that a minimum of 80,000 and a maximum 100,000 cells (this top limit was introduced in order to narrow the range of acquired events across laboratories) of interest needed to be acquired per well. Moreover, feedback was given to the laboratories that were not able to acquire sufficient events. Possible causes for this problem were inaccurate cell counting methodology, inadequate centrifugation speed after fixation and permeabilization, or cell loss during aspiration of the wells after wash (either by incorrect use of the manifold or poor plate flicking technique). There was no correlation between different cell counting techniques (Guava, hemocytometer, etc.) or aspiration methodology (flicking versus vacuum manifold) and cell numbers acquired. The feedback given to the laboratories resulted in improvements for certain laboratories, but others still struggled to reach the minimum number of required events. In addition, since Round 6, the number of events to be acquired was further increased to 120,000 to 150,000 CD3+ lymphocytes. Due to the need for analyzing smaller cell subsets identified with the 7-color cocktail, the cell recovery requirement at the end of the assay became more challenging for some laboratories.

As expected, acquisition of a low number of events had a clear impact on the accuracy and precision of the data, and could contribute to decrease in inter-laboratory reproducibility. In the example shown in Figure 5B, acquisition of a low number of events led to imprecision by under-estimating cytokine responses, especially when looking at low frequency responses like the CD4 cytokine response in this case.

3.6. Intra-laboratory variation

Beginning with Round 3, we were able to provide enough cells to the participants so that each response was tested in triplicate. This allowed us to evaluate intra-laboratory variability. For each laboratory, the %CV or SD, depending on the level of response, was calculated for each set of triplicates (Figure 6). Moreover, this variability was compared to the intra-laboratory variability of the GS, since that set of data was considered optimal. The variability within triplicates for the GS was low with a mean CV lower than 7% and a mean SD lower than 0.02 in Rounds 3 through 7.

In general, the majority of the laboratories generated reproducibly tight data, with %CVs and SDs, which were equal to or lower than the ones from the GS (represented by small boxes in Figure 6; for example, Laboratories 8 and 9 for both %CV and SD). Variable data represented as large boxes (i.e. Laboratories 5 and 13 CVs) or high %CVs and SDs (above

the GS mean) usually was attributed to technical problems during the assay (cell loss in a given well due to inadequate aspiration method, cross-contamination, etc.). Interestingly, although there was no correlation between data from different laboratories and/or rounds of testing, high variability within triplicates usually indicated a high number of outlier measurements for a given laboratory.

3.7. Instrument performance and setup

Since central analysis of individual FCS files was performed in each round, we were able to detect issues related to instrument performance and setup that had an impact on the quality of the data, which could also account for inter-laboratory variability.

For instance, compensation was examined by looking at all the possible color combinations in un-gated plots. Over or under-compensation was seen in some cases, especially during initial rounds. As expected, inadequate compensation yielded outlier measurements (data not shown).

In addition, we observed that different populations (FSC, SSC, CD3, CD4, CD8, cytokine +) would fall in diverse locations. This variation could contribute to inaccurate measurement of the responses, if the populations of interest were off-scale or difficult to discriminate. Furthermore, to conduct a centralized analysis, this variability made it extremely difficult to generate an accurate generalized analysis template to fit the majority of the data.

Beginning in Round 3, single stained cells or pre-stained compensation beads were lyophilized and included in the staining plates so all participants could use exactly the same reagents for instrument setup. In addition, the protocol distributed included target values for each channel (generated using the pre-stained lyophilized cells or beads in the central laboratory) specific for the different instruments, along with instructions for manual or automatic compensation. This strategy led to overall better instrument setup and more homogenous data across sites (data not shown) that significantly facilitated central analysis.

In addition to providing detailed guidelines for instrument setup, we evaluated and compared the performance of flow cytometers across institutions, starting at Round 6. This assessment was done both quantitatively and qualitatively in each participating laboratory by providing calibration beads (CS&T and fluorescence calibration 8-peak beads), pre-stained compensation beads and a detailed protocol. For each detector, the following parameters were calculated: Qr (a measure of fluorescence detection efficiency); Br (optical background); SDen (standard deviation of electronic noise); and Stain Index (SI) of the compensation beads. SI provides an estimate of the sensitivity/resolution of each channel and is calculated using the following formula: $SI = (\text{median } [+ \text{ peak}] - \text{median } [- \text{ peak}]) / 2 * \text{robust SD}$. A high SI, high Qr and low Br are desirable, as they reflect good resolution and low background for a given channel. For each detector, all the calculated parameters listed here varied widely across institutions. For example, the SI of the compensation beads ranged from 28.5 to 151, 46.1 to 347.3, and 6.8 to 36.1 for FITC, APC and V450 channels, respectively. Of note, SI calculated from beads was highly correlated with SI calculated from stained cells (Supplemental Figure 1). More importantly, the combination of the instrument performance information obtained from the SI and the 8-peak beads, together with the measurements Qr and Br for each detector, allowed us to dissect the specific causes for low sensitivity in a given detector. For example, if high Br was calculated, poor resolution was very likely due to high optical background, suggesting that the filter and/or the flow cell needed to be checked for that particular instrument. If low Qr was calculated, then low sensitivity was linked to an issue with the optics for that given detector. In the example shown in Supplemental Figure 1S, the difference in resolution in the FITC channel

between the instruments in Laboratories 1 and 2 was due to a combination of high background and low detection efficiency of the FITC detector in Laboratory 2.

The results of the instrument performance assessment were summarized in the individual reports distributed at the end of each round, and suggestions were made in order to correct any documented issues.

3.8. Gating strategy

All participating laboratories were requested to provide raw data (FCS files) at the end of each round. This enabled the central laboratory to analyze all of the data generated by each participating laboratory. The responses obtained from this centralized analysis were compared to those from the individual analysis performed in each laboratory.

In the majority of cases, good correlation was observed between the two analyses, with correlation coefficients higher than 0.9 (data not shown). When a poor correlation was obtained, the central laboratory carefully analyzed the gating strategy used by a particular laboratory. One of the most common problems encountered is represented in Figure 7. Non-inclusion of down-modulated cell surface antigen populations (CD3 dims, CD4 dims or CD8 dims) led to underestimation of the gated cytokine responses, since some of the cytokine+ cells fall into these dim subsets as it has been previously shown (Maecker et al., 2001; Bitmansour et al., 2002).

When these problems or any other non-optimal gating issues were observed, feedback was given to the relevant participants. In addition, a suggested gating strategy was provided, and it was highly recommended to back-gate on the cytokine+ cells, in order to visualize the location of those cells within the CD3, CD4 and CD8 parameters. When it was possible to follow up on these recommendations with the same operator, usually high correlation was achieved after one or two rounds of testing between the centralized and the laboratory analysis (data not shown).

Finally, a common problem accounting for low correlation between centralized versus individual analysis was defining how conservative to set the gates to discriminate cytokine positive and negative cells.

3.9. Pass/fail criteria for ICS assays

After analysis of the data generated during seven rounds of this ICS QAP, we have identified key factors responsible for inter-laboratory variability that reflect the optimal or suboptimal performance of an ICS assay. Those factors cover all aspects of this assay, from cell processing (cell viability, cell recovery and intra-laboratory variability), instrument setup and acquisition (compensation and number of collected events) to data analysis.

During all rounds of testing, the importance of these factors was extensively discussed with all participants, and data analysis regarding each element was provided within the individual report generated for every laboratory. As shown in Figure 8, when these different factors were taken into consideration and only optimal data is considered, along with a centralized analysis, the inter-laboratory variability decreased significantly.

The participants of this program agreed to adopt these factors, along with the optimal range determined by the GS, as pass/fail criteria (Table 3). For each parameter, we established a target value that we consider optimal. These key indicators of performance were based either on the data provided by all laboratories (e.g., for cell viability and recovery in each round) or data from the GS (e.g., for intra-laboratory variability), and were set so excellent versus good versus fair performances could be discriminated. Additionally, a scoring system

was designed that allows one to accurately grade the performance of a given laboratory in each specific aspect of the assay. For example, in regard to the number of collected events, 10 points are allotted if the average of relevant events collected across wells is equal to the target number provided in the protocol; 8 points are allotted if the average is between 80–99% of the target; 6 points if the average is only 60–79% and so on. Zero or no points are allotted if the average number of collected events is less than 19% of the target. When this scoring system was applied to the data generated in Round 7, the scores assigned to each laboratory highly correlated with the overall impression assigned by the central laboratory and the quality of the data (data not shown) based on its experience.

4. DISCUSSION

Currently, participation in the ICS QAP described in this study is a NIH/NIAID/DAIDS requirement for those laboratories supporting trials where NIAID holds the IND application. The implementation of this QAP has been a useful tool to determine whether a specific laboratory is capable of providing reliable data for evaluation of T-cell responses to HIV vaccine candidates.

From the data collected during seven rounds of proficiency testing, we were able to evaluate the reproducibility of 4- and 7-color ICS assays across laboratories and to define pass and fail criteria for future use.

As stated in previous studies (Maecker et al., 2005; Maecker et al., 2008), the degree of variation detected depends on the frequency of the cytokine responses evaluated. For responses above 0.2%, inter-laboratory %CVs of less than 35% seemed suitable based on our data. For cytokine responses below 0.2%, SD lower than 0.1 should be expected. It will be important to establish these values as a reference point for other organizations wishing to conduct similar ICS proficiency testing. Although in the present study we only evaluated antigen-specific responses to CMVpp65 and CEF, we were able to study donors with a wide range of responses; hence these results could be generalized to a variety of antigens eliciting similar frequencies of responses. With the current panels, we were able to evaluate IFN- γ , IL-2 and TFN- α production. It remains to be determined if similar levels of variation are observed for assays evaluating other cytokine responses.

The level of inter-laboratory variation observed was, as expected, higher compared to intra-assay and inter-assay variations (%CV of 10% and 25%, respectively) previously reported (Nomura et al., 2000). However, it was significantly lower than the inter-laboratory variation reported for other types of functional assays, such as ELISpot (Cox et al., 2005; Janetzki et al., 2008) and tetramer detection of antigen-specific T-cells (Britten et al., 2009). This was likely due to the fact that in contrast to these previous studies, all laboratories in this study performed the ICS assay following a common SOP and using common reagents. The requirement for a validated common protocol and common set of reagents is mandatory if factors accounting for variability are to be dissected. Variations in protocol design can yield very different cytokine responses and may include: pre-incubation of freshly thawed PBMC prior to activation, use of co-stimulatory antibodies during the activation step, variation in use of different fixation and permeabilization buffers and protocols, and variation in selection of antibody clones and panel combinations. It is important to point out that there were only few steps in the detailed protocol distributed where the participants had choices on the procedure to follow and that previous studies had demonstrated that those alternatives did not impact the ICS data. In addition to following a standardized protocol, we believe that the use of lyophilized reagents provide a clear advantage for this kind of program, as this format minimizes variation due to reagent handling and pipetting.

Furthermore, lyophilized reagents pre-configured in 96-well plates provide convenience of assay setup and offer long-term reagent stability.

Interestingly, increasing the complexity of the assay by using a multicolor cocktail that allowed the detection of low frequency populations did not have an impact on the inter-laboratory variation. This indicates once a SOP and common reagents are used, the variation is due to factors that equally affect a 4-color or a 7-color assay. Moreover, one could expect that following the guidelines established in this study, going beyond 7-color assay, will not lead to increased variation.

The main factors identified in this study for inter-laboratory variation, either independently or in combination, were number of collected events, analysis/gating strategy and instrument setup, as well as performance. It is critical to establish a minimum number of events to be collected, using statistical methods that take into account the frequencies of responses to be measured and the level of precision, as this directly impacts the precision of the measurements. Despite efforts in troubleshooting, low recoveries were consistently detected in some laboratories. Even after modifying the centrifugation speed and aspiration technique at the various washing steps, we did not see a significant improvement, and the minimum number of events specified in our protocol was not consistently reached across laboratories. With the introduction of a multi-color cocktail and the possibility of detecting low frequencies of poly-functional T-cells, the necessity of acquiring a minimum number of needed events becomes even more critical. We are currently evaluating modifications in the protocol, which could enhance cell recovery. These elements include decreasing the number of washing steps and/or using alternative fixation and permeabilization buffers.

As pointed out in previous studies (Maecker et al., 2005; Britten et al., 2009), data analysis is a significant contributor to inter-laboratory variation for flow cytometry-based assays, and centralized data analysis is a strategy that can be implemented to minimize laboratory-to-laboratory variation in reported results. Although if as in the present study, the various institutions performing the assays have flow cytometers with distinct configurations (i.e. different lasers and filters), the use of a standardized analysis template is not always feasible with the current software available. Development of new, more portable analysis tools at different institutions is ongoing and will very likely facilitate centralized analysis. Also, we found that providing a suggested analysis strategy was important, as all participants could visualize a proposed way to set up the gates. The strategy followed in this study was to set the cytokine positive and negative gates based on the biological control (unstimulated samples for each donor), rather than using fluorescence minus one (FMOs) or isotype control, but specific rules on how “high” or “low” to set those gates were not included in the protocol. We are currently working with all participants to define such guidelines, as it is evident this plays a vital role in discrepancies between the centralized versus the individual analysis and in the inter-laboratory variation observed.

Another factor that we have started assessing during the last two rounds is the flow cytometer performance. The data obtained demonstrated that measured performance varied widely across institutions; more importantly, we have been able to correlate instrument performance with assay performance. For instance, poor resolution of different cell subsets was predictable when looking at non-optimal Qr and Br values. As a result of this evaluation, we were able to provide laboratories with data to compare their instrument performance with others and address any nonconforming issues. In the future, we expect the evaluation of instrument performance will lead to better instrument quality control (QC) in all participating institutions, and will also allow us to more easily establish minimum instrument performance reference values (mainly Qr and Br for each detector) needed to

generate optimal data when using the specific antibody combinations provided in this program.

It is important to identify that although we are certain the above discussed factors impact the quality of the data, they are not exclusive, and perhaps additional unknown elements may explain the sporadic assay variation within a round. One of the factors that is difficult to quantify is the experience of the operators performing the assay and their familiarity with the protocol. Consequentially, during the first round, several new operators unfamiliar with the protocol performed the assay. After feedback was provided and experience gained with the protocol and reagents, variability significantly decreased in the next rounds. In Round 7, however, several laboratories had new personnel performing the assay, which may potentially account for the increased variation compared to previous rounds. This may also explain why some laboratories did not improve significantly between rounds, while others did. To increase operator competency, a laboratory may request “trainee panels” prior to participating in a proficiency program. It would also be expected that within each laboratory, internal training and competency testing of staff is completed before performing the external proficiency testing for any assay.

The investment required to achieve good inter-laboratory precision in a functional assay is high. With minimal effort, a single laboratory might be able to achieve consistent results and even demonstrate sufficient sensitivity for a given purpose. However, this is not sufficient to assure consistency across laboratories, which is a prerequisite if multiple laboratories contribute to a study, or if data from multiple laboratories are to be compared across studies. For this purpose, we attempted to create acceptance criteria around a “gold standard”, which, while not a true test of accuracy, allowed for measurement of inter-laboratory precision relative to a presumably optimal result.

A significant contribution of this multi-site study is our ability to define pass and fail criteria for ICS assays, which will be adopted in the subsequent rounds of testing. We expect these criteria will help the central laboratory to objectively and accurately qualify the performance of each participant and identify areas to enhance performance, if necessary. We anticipate that the adoption of this strategy, combined with regular participation in a QAP, will provide the participant laboratories with the necessary tools to generate optimal and accurate data and expect other organizations will benefit from the lessons learned through this program. Considering the nature of different flow cytometry-based T-cell assays, such as proliferation measurements using carboxylfluorescein diacetate succinimidyl ester (CFSE) and antigen-specific T-cell detection employing tetramer staining among others, we believe the elements of this QAP could be easily extrapolated to other flow cytometry-based assays that assess antigen-specific cell frequencies.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This project has been funded in whole or in part with Federal funds from the National Institute of Allergy and Infectious Disease, National Institutes of Health, Department of Health and Human Services, under Contract No. HHSN26620050022C.

References

Betts MR, Krowka JF, Kepler TB, Davidian M, Christopherson C, Kwok S, Louie L, Eron J, Sheppard H, Frelinger JA. Human immunodeficiency virus type 1-specific cytotoxic T lymphocyte activity is

- inversely correlated with HIV type 1 viral load in HIV type 1-infected long-term survivors. *AIDS Res Hum Retroviruses*. 1999; 15:1219–28. [PubMed: 10480635]
- Bitmansour AD, Douek DC, Maino VC, Picker LJ. Direct ex vivo analysis of human CD4(+) memory T cell activation requirements at the single clonotype level. *J Immunol*. 2002; 169:1207–18. [PubMed: 12133941]
- Britten CM, Janetzki S, Ben-Porat L, Clay TM, Kalos M, Maecker H, Odunsi K, Pride M, Old L, Hoos A, Romero P. Harmonization guidelines for HLA-peptide multimer assays derived from results of a large scale international proficiency panel of the Cancer Vaccine Consortium. *Cancer Immunol Immunother*. 2009; 58:1701–13. [PubMed: 19259668]
- Bull M, Lee D, Stucky J, Chiu YL, Rubin A, Horton H, McElrath MJ. Defining blood processing parameters for optimal detection of cryopreserved antigen-specific responses for HIV vaccine trials. *J Immunol Methods*. 2007; 322:57–69. [PubMed: 17382342]
- Cheng C, Gall JG, Nason M, King CR, Koup RA, Roederer M, McElrath MJ, Morgan CA, Churchyard G, Baden LR, Duerr AC, Keefer MC, Graham BS, Nabel GJ. Differential specificity and immunogenicity of adenovirus type 5 neutralizing antibodies elicited by natural infection or immunization. *J Virol*. 84:630–8. [PubMed: 19846512]
- Cox JH, Ferrari G, Kalams SA, Lopaczynski W, Oden N, D'Souza MP. Results of an ELISPOT proficiency panel conducted in 11 laboratories participating in international human immunodeficiency virus type 1 vaccine trials. *AIDS Res Hum Retroviruses*. 2005; 21:68–81. [PubMed: 15665646]
- Currier JR, Kuta EG, Turk E, Earhart LB, Loomis-Price L, Janetzki S, Ferrari G, Bix DL, Cox JH. A panel of MHC class I restricted viral peptides for use as a quality control for vaccine trial ELISPOT assays. *J Immunol Methods*. 2002; 260:157–72. [PubMed: 11792386]
- De Rosa SD, McElrath MJ. T cell responses generated by HIV vaccines in clinical trials. *Curr Opin HIV AIDS*. 2008; 3:375–9. [PubMed: 19372993]
- Dunne JMH. Automation of Cytokine Flow Cytometry Assays. *J Assoc Lab Automation*. 2004; 9:5–9.
- Ezzelle J, Rodriguez-Chavez IR, Darden JM, Stirewalt M, Kunwar N, Hitchcock R, Walter T, D'Souza MP. Guidelines on good clinical laboratory practice: bridging operations between research and clinical research laboratories. *J Pharm Biomed Anal*. 2008; 46:18–29. [PubMed: 18037599]
- Goonetilleke N, Liu MK, Salazar-Gonzalez JF, Ferrari G, Giorgi E, Ganusov VV, Keele BF, Learn GH, Turnbull EL, Salazar MG, Weinhold KJ, Moore S, Letvin N, Haynes BF, Cohen MS, Hraber P, Bhattacharya T, Borrow P, Perelson AS, Hahn BH, Shaw GM, Korber BT, McMichael AJ. The first T cell response to transmitted/founder virus contributes to the control of acute viremia in HIV-1 infection. *J Exp Med*. 2009; 206:1253–72. [PubMed: 19487423]
- Harrer T, Harrer E, Kalams SA, Elbeik T, Staprans SI, Feinberg MB, Cao Y, Ho DD, Yilma T, Caliendo AM, Johnson RP, Buchbinder SP, Walker BD. Strong cytotoxic T cell and weak neutralizing antibody responses in a subset of persons with stable nonprogressing HIV type 1 infection. *AIDS Res Hum Retroviruses*. 1996; 12:585–92. [PubMed: 8743084]
- Horton H, Thomas EP, Stucky JA, Frank I, Moodie Z, Huang Y, Chiu YL, McElrath MJ, De Rosa SC. Optimization and validation of an 8-color intracellular cytokine staining (ICS) assay to quantify antigen-specific T cells induced by vaccination. *J Immunol Methods*. 2007; 323:39–54. [PubMed: 17451739]
- Janetzki S, Panageas KS, Ben-Porat L, Boyer J, Britten CM, Clay TM, Kalos M, Maecker HT, Romero P, Yuan J, Kast WM, Hoos A. Results and harmonization guidelines from two large-scale international Elispot proficiency panels conducted by the Cancer Vaccine Consortium (CVC/SVI). *Cancer Immunol Immunother*. 2008; 57:303–15. [PubMed: 17721781]
- Klein MR, van Baalen CA, Holwerda AM, Kerkhof Garde SR, Bende RJ, Keet IP, Eeftink-Schattenkerk JK, Osterhaus AD, Schuitemaker H, Miedema F. Kinetics of Gag-specific cytotoxic T lymphocyte responses during the clinical course of HIV-1 infection: a longitudinal analysis of rapid progressors and long-term asymptomatics. *J Exp Med*. 1995; 181:1365–72. [PubMed: 7699324]
- Koup RA, Roederer M, Lamoreaux L, Fischer J, Novik L, Nason MC, Larkin BD, Enama ME, Ledgerwood JE, Bailer RT, Mascola JR, Nabel GJ, Graham BS. Priming immunization with DNA

- augments immunogenicity of recombinant adenoviral vectors for both HIV-1 specific antibody and T-cell responses. *PLoS One*. 5:e9015. [PubMed: 20126394]
- Maecker HT, Dunn HS, Suni MA, Khatamzas E, Pitcher CJ, Bunde T, Persaud N, Trigona W, Fu TM, Sinclair E, Brecht BM, McCune JM, Maino VC, Kern F, Picker LJ. Use of overlapping peptide mixtures as antigens for cytokine flow cytometry. *J Immunol Methods*. 2001; 255:27–40. [PubMed: 11470284]
- Maecker HT, Hassler J, Payne JK, Summers A, Comatas K, Ghanayem M, Morse MA, Clay TM, Lysterly HK, Bhatia S, Ghanekar SA, Maino VC, Delarosa C, Disis ML. Precision and linearity targets for validation of an IFN γ ELISPOT, cytokine flow cytometry, and tetramer assay using CMV peptides. *BMC Immunol*. 2008; 9:9. [PubMed: 18366814]
- Maecker HT, Rinfret A, D'Souza P, Darden J, Roig E, Landry C, Hayes P, Birungi J, Anzala O, Garcia M, Harari A, Frank I, Baydo R, Baker M, Holbrook J, Ottinger J, Lamoreaux L, Epling CL, Sinclair E, Suni MA, Punt K, Calarota S, El-Bahi S, Alter G, Maila H, Kuta E, Cox J, Gray C, Altfeld M, Nougarede N, Boyer J, Tussey L, Tobery T, Brecht B, Roederer M, Koup R, Maino VC, Weinhold K, Pantaleo G, Gilmour J, Horton H, Sekaly RP. Standardization of cytokine flow cytometry assays. *BMC Immunol*. 2005; 6:13. [PubMed: 15978127]
- McElrath MJ, De Rosa SC, Moodie Z, Dubey S, Kierstead L, Janes H, Defawe OD, Carter DK, Hural J, Akondy R, Buchbinder SP, Robertson MN, Mehrotra DV, Self SG, Corey L, Shiver JW, Casimiro DR. HIV-1 vaccine-induced immunity in the test-of-concept Step Study: a case-cohort analysis. *Lancet*. 2008; 372:1894–905. [PubMed: 19012957]
- Nomura L, Maino VC, Maecker HT. Standardization and optimization of multiparameter intracellular cytokine staining. *Cytometry A*. 2008; 73:984–91. [PubMed: 18612990]
- Nomura LE, Walker JM, Maecker HT. Optimization of whole blood antigen-specific cytokine assays for CD4(+) T cells. *Cytometry*. 2000; 40:60–8. [PubMed: 10754518]
- Rinaldo C, Huang XL, Fan ZF, Ding M, Beltz L, Logar A, Panicali D, Mazzara G, Liebmann J, Cottrill M, et al. High levels of anti-human immunodeficiency virus type 1 (HIV-1) memory cytotoxic T-lymphocyte activity and low viral load are associated with lack of disease in HIV-1-infected long-term nonprogressors. *J Virol*. 1995; 69:5838–42. [PubMed: 7637030]
- Sarzotti-Kelsoe M, Cox J, Cleland N, Denny T, Hural J, Needham L, Ozaki D, Rodriguez-Chavez IR, Stevens G, Stiles T, Tarragona-Fiol T, Simkins A. Evaluation and recommendations on good clinical laboratory practice guidelines for phase I-III clinical trials. *PLoS Med*. 2009; 6:e1000067. [PubMed: 19536325]

Figure 1. ICS 4-color cocktail

Dot plots representative of the 4-color staining used throughout all rounds of the program. In this example, and following the program's suggested gating strategy, cells were first gated on CD3+ (APC), then on lymphocytes and from there, CD4+ (FITC) and CD8+ (PerCP Cy55) subsets are identified. For each of those two subsets, cells producing IFN- γ and IL-2 were subsequently identified in the PE channel.

Figure 2. Inter-laboratory variability of cytokine responses across seven rounds of the ICS QAP using a 4-color cocktail

(a) High, intermediate and low CMVpp65 responders were tested by the participants in each round of the ICS QAP. Cytokine+ (IFN- γ + IL-2), CD4+ and CD8+ responses were determined by each laboratory. Each horizontal line represents a response obtained within an individual laboratory and replicates are linked together with a vertical line. The numbers on top of each plot represent the %CV (for responses >0.2%) or SD (for responses <0.2%) of the data across all laboratories. (b) The %CVs or SDs were calculated for all measurements performed in a given round (each dot is the CV or SD for a given response). The horizontal lines represent the mean.

Figure 3. Inter-laboratory variability of cytokine responses across seven rounds of the ICS QAP using a 7-color cocktail

(a) Dot plots representative of cytokine responses detected using the 7-color cocktail. Plots on the top row are gated on live lymphocytes CD3+, CD4+ (left) or CD8+ (right). For each subset, IFN- γ or – cells were identified, and from there, cells producing all possible combinations of the three cytokines, IFN- γ IL-2 and TNF (bottom row), were identified. (b) Inter-laboratory variation was compared between the 4-(left column plots) and the 7-color (right column plots) cocktails. In this example, the CD4+ response to CMVpp65 for Donor 150 (top row) and the CD8+ response to CMVpp65 for Donor 177 (bottom row) are shown. (c) The %CVs or SDs were calculated for all measurements performed in Rounds 6 and 7 (each dot is the CV or SD for a given response) using the 4- and 7-color cocktails. The horizontal lines represent the mean.

Figure 4. Gold Standard

(a) Correlation between the median of the data reported by the laboratories (x axis) and the mean of the GS (y axis) for each cytokine+, CD4+ and CD8+ responses evaluated in Rounds 3, 4 and 5 of the ICS QAP. (b) Quantification of outlier measurements using the GS to define the optimal range across Rounds 3 through 5 for laboratories that participated in these three rounds.

Figure 5. Number of collected events

(a) Absolute number of CD3+ lymphocytes collected in each individual well tested. Data shown was generated in Round 3. (b) Effect of acquisition of low cell numbers. Plots represent CD8+ (upper row) and CD4+ (lower row) responses to CMVpp65 in a given donor when acquiring 20,000 (left plots) or 80,000 (right plots) CD3+ lymphocytes. Numbers in the upper quadrant represent absolute number of events in quadrant/% cytokine +, CD4+ or CD8+ response.

Figure 6. Intra-laboratory variation

Variability within each set of triplicates was calculated for each laboratory. The %CV was calculated for cytokine responses $>0.2\%$, SD was calculated for responses $<0.2\%$. The horizontal line represents the mean of the CV or SD of the GS. Data presented is from Round 4.

Figure 7. Effect of gating on variability

(a) CD8 response to CMVpp65, excluding (top dot plots) or including (bottom dot plots) CD8 dim cells. (b) CD4 response to CMVpp65, excluding (top) or including (bottom) CD4+ CD8 dim cells. CD4 versus CD8 dot plots are gated on CD3+ lymphocytes. Numbers in the upper right quadrant correspond to the % cytokine+, CD4+ or CD8+ cells.

Figure 8. Effect of acceptance criteria and gating strategy on inter-laboratory variability

The %CV from data reported by all laboratories (a) data filtered using acceptance criteria (uncompensated, low number of events and data from triplicates with high SD or CVs were excluded). (b) Data from central analysis of filtered data. (c) The horizontal line represents the median. Data on the left plot corresponds to CD8+ response to CEF from Donor 153 in Round 5, and data on the right plot corresponds to CD4+ response to CMVpp65 from Donor 150 in Round 6.

Table 1**Institutions and Study Participants**

Institution	Participants
Fred Hutchinson Cancer Research Center, HVTN Seattle, WA	Steve de Rosa
US Military HIV Research Program (USMHRP) Immunology Laboratory Rockville, MD	Jeffrey Currier, Josephine Cox, Anais Valencia
Duke Center for AIDS Research/HVTN Durham, NC	Janet Staats, Jennifer Enzor
University of California San Francisco San Francisco, CA	Elizabeth Sinclair, Lorrie Epling
Massachusetts General Hospital Boston, MA	Marcus Altfeld, Galit Alter
Alpha Vax Research Triangle Park, NC	Elizabeth Reap, Kevin Katen
Vaccine Research Center Bethesda, MD	Robert Bailer, Ellen Turk, Kim Davis
Rush Medical Center Chicago, IL	Alan Landay, Betty Donoval
Université de Montréal-CANVAC/National Immune Monitoring Laboratory (NIML) Montreal, Canada	Rafick-Pierre Sekaly, Claire Landry, Dominique Sauvé
Imperial College, IAVI London, UK	Peter Hayes, Dilbinder Gill
National Institute Communicable Diseases (NICD) Johannesburg, South Africa	Clive Gray, Patti Kay, Stephina Nyoka
US Army Medical Component of the Armed Forces Research Institute for the Medical Sciences (USAMC- AFRIMS), Bangkok, Thailand	Silvia Kim, Alexandra Schuetz, Suchada Sukhumvittaya
Makerere University Walter Reed Project (MUWRP) Kampala, Uganda	Mike Eller, Leigh Anne Eller, Prossy Naluyima -
Mbeya Research Programme, Mbeya, Tanzania	Alexandra Schuetz
BD Pharmingen., La Jolla, CA	John Hodgson, Susan Saiget
BD Biosciences, San Jose, CA	Holden Maecker, Maria Jaimes, Mary Beth Hanley

Table 2

Viability and Recovery of PBMC tested across rounds

Donor	Round	% viability ($\pm 2SD$)	% recovery ($\pm 2SD$)
150	6	94.36 (± 2.29)	48.58 (± 12.46)
153	3	90.40 (± 4.74)	72.79 (± 20.11)
	4	91.16 (± 4.58)	72.50 (± 15.52)
	5	91.92 (± 5.14)	68.39 (± 12.50)
157	1	91.8% (± 6.3)	N/D
	2	94.22 (± 2.87)	93.89 (± 20.11)
165	1	89.1% (± 13.8)	N/D
	2	92.37 (± 3.53)	83.14 (± 16.74)
	3	91.38 (± 4.01)	86.86 (± 16.74)
	4	91.74 (± 4.55)	84.83 (± 20.30)
	5	94.32 (± 3.60)	78.90 (± 16.49)
166	1	95.6% (± 11.2)	N/D
	2	93.14 (± 3.98)	75.06 (± 26.80)
171	1	92.2% (± 4.9)	N/D
	2	90.57 (± 5.92)	81.03 (± 17.95)
173	3	94.54 (± 2.54)	103.83 (± 26.80)
	4	93.49 (± 3.42)	105.00 (± 24.70)
	5	93.98 (± 4.60)	94.15 (± 18.74)
176	6	92.27 (± 3.77)	77.68 (± 15.19)
	7	92.32 (± 4.68)	88.81 (± 19.2)
177	6	92.81 (± 2.97)	85.58 (± 18.64)
	7	92.07 (± 3.86)	94.29 (± 21.9)
185	7	92.65 (± 4.36)	103.33 (± 19.6)

Table 3

ICS Assay Acceptability Criteria

Criteria	Target	Weight as % of total
Cell Viability	$\pm 2SD$ from group's mean	10 %
Cell Recovery	$\pm 2SD$ from group's mean	10 %
Number of Collected Events	Within target defined per protocol	10 %
Intra-laboratory Variability (compared to the GS for triplicates)	%CV: $\leq 7\%$	10 %
Instrument Setup	MFI for all channels in range per protocol	10 %
	Adequate compensation	
Analysis (correlation between participating laboratory and central analysis)	$R^2 \geq 0.9$ CD4 cyt+ (4-color panel)	20%
	$R^2 \geq 0.9$ CD8+ cyt+ (4-color panel)	
	$R^2 \geq 0.9$ CD4+ cyt+ (7-color panel)	
	$R^2 \geq 0.9$ CD8+ cyt+ (7-color panel)	
Total Measurements within Optimal Range (based on CENTRAL ANALYSIS)	4-color panel	30%
	7-color panel	
Total		100%