Cytometry PART A
Journal of the International Society for Advancement of Cytometry

ISAC
International Society for Advancement of Cytometry

How Many Events Is Enough? Are You Positive?

ONE of the more perplexing problems in flow cytometry revolves around the issue of: "Is it real"? If a sample has one event in a particular gate, is that event real? Is it significant? Should it be believed? Is the sample "positive"? If the answer to these is "No", then what is the threshold number of events above which the answer becomes "Yes"?

Flow cytometry is unique among biological technologies by providing an enormous number of measurements on which to base conclusions. Not only can more than a dozen measurements be made on each cell, millions of cells can be analyzed in the context of a single sample (tube); dozens of samples may be analyzed in the context of a single biological specimen. Hence, there is the potential for enormous precision on measurements; distributions arising from these measurements can have exceedingly small standard errors of the mean.

The precision of a subset frequency is easily defined. The standard deviation for relatively rare populations is simply $n^{-1/2}$ where "n" is the number of events comprising the subset. For a gate with a single event, the relative precision on its frequency is $\pm 100\%$; for a gate with 1,000 events, it is $\pm 3\%$. However, assay variation (biology, experimental, operator, etc.) is typically greater than 30%. Thus, once the number of events in a gate exceeds 10, the precision of the frequency measurement is dominated by assay errors, not the paucity of events analyzed.

The overwhelming amount of data available by flow cytometry has led to some confusion about the statistical significance (or lack thereof) in the precision of the subset frequency measurements. Indeed, it might be tempting to conclude that when one collects a million events, finding a single event in a gate is not "meaningful".

This question is not solely the domain of extremely rare event detection; it has become common as we characterize small subsets with additional measurements. For example, when we assess the quality of a T cell response, we often measure five different functions simultaneously (1). The quality of the response is defined by the pattern of the co-expression of the five functions—a pattern comprised of 32 possible combinations (2^5).

Vaccine-induced T cell responses are often as low as 0.1%. With a sample of one million stimulated peripheral blood mononuclear cells, after applying gates to define singlets, viable lymphocytes, and T cells, the total number of cells that are in the cytokine gate can be only a few hundred. Dividing that into 32 fractions means that many of these functionally defined subfractions will have very few events (2). Are they "real"?

The question of whether events are "real" or not is fundamentally inappropriate. Of course they are "real". The appropriate question is: do the events represent what the researcher claims they are—in this case, a set of antigen-specific cells with a given functional response. To answer that, we must first determine what the alternative explanations for any given event are: 1) it is "noise" of some sort—e.g., a dead cell or cell fragment that had unusual fluorescent properties, putting it in the gates; 2) it is "experimental background"—e.g., a real cell with the appropriate fluorescent markers, but is not a cell being quantified by the assay (in this example, a T cell that is not specific for the tested antigen, perhaps having been preactivated in vivo); 3) it is an antigen-specific cell with the appropriate properties. Only in the last case do we want to report the event in our results; unfortunately, for any given event there is no way that we can distinguish between the possibilities.

This has led to some discomfort with very low event numbers, leading to the temptation to use an arbitrary minimum number of events below which a frequency measurement is deemed zero: For example, at least 10 events must be in a gate to define the sample as positive (irrespective of the frequency). Conversely, it is tempting to conclude that, upon seeing a cloud of a thousand events that are closely distributed in the desired gate, a sample is clearly positive.

Both of these temptations must be avoided, as neither is based on sound principles.

Consider a study in which the T cell response to a vaccine is being measured through typical intracellular cytokine staining (ICS) assay. In assessing two vaccinees, the gating strategy reveals in one subject a single event (out of one million collected) in the cytokine-positive gate. In the second vaccinee, there were one thousand positive events. Is either sample "positive"?

This question cannot be answered. The reason is that "positive" has a contextual meaning that is far deeper than

Received 23 January 2008; Accepted 30 January 2008

*Correspondence to: Mario Roederer, Vaccine Research Center, NIAID, NIH, Bethesda, MD 20892.

E-mail: roederer@nih.gov

Published online 28 February 2008 in Wiley InterScience (www.interscience.wiley.com)

DOI: 10.1002/cyto.a.20549

what is evident on the surface. In fact, both samples are positive, in that both have measurable cytokine-expressing events. But this is not what we are really asking when we want to know if the samples are "positive".

To know if these samples represent responses to the vaccine (i.e., they are "different" from unvaccinated responses and thereby "positive"), we must first determine the distribution of "negatives". Consider the case in which 50 unvaccinated subjects were analyzed by the same assay, and all 50 had zero events in the gated distribution. There is now an extremely high probability that even the sample with a single event in the gate is "positive" (indeed, even compared to 10 controls subjects that all have zero events, the single event sample is significantly different, $p < 0.002$).

On the other hand, if the same 50 unvaccinated subjects generated a distribution of gated events that ranged from 0 to 1500, then we may be unable to conclude that either of the two vaccine subjects was "positive".

In the case of measuring the quality of the T cell response, recall that there may be a few hundred events divided into 32 categories (subsets). Validation may have shown that a few hundred cytokine-positive events is well above the bounds of "negatives", so the overall response is "positive". But is the pattern resulting from the division into 32 gates meaningful? The answer only comes from reproducibility measurements on known positive specimens. These positive controls will be used to determine how much variability there is in the definition of the quality. Only once this variability has been defined can one determine if differences in the patterns between subjects is significant.

These examples serve to illustrate a basic principle: ascribing a quality such as "positivity" or "real" for an experimental sample cannot be done without knowing the expected distributions for "negative" or control samples; only a rigorous validation of the specific assay will suffice (3).

Note that we can take advantage of all the information present in a flow cytometry experiment to help determine whether responses are likely to be real. The distribution of fluorescence on cell is usually characteristic; if the events falling in a particular gate have an unusual fluorescence distribution, then that would be evidence that perhaps those events are not representative of a "positive" response. But be careful: this assessment needs to be made irrespective of the number of events collected (don't examine *only* low frequency results), and with the realization that the distribution of a very small number of events could easily be biased by the limited sampling – in the extreme case, with one event, virtually no conclusion could be drawn based on *where* that event appeared in the gates. And, of course, any such qualifications of analyses must be done transparently and objectively, and should be validated.

In conclusion, there is no theoretical reason to employ an artificial threshold number of events, below which a frequency is deemed "negative". The assessment of "positivity" can only be made by comparison of the measurement against a set of control samples, using standard statistical tools to compare the frequencies.

Mario Roederer*
Vaccine Research Center
NIAID, NIH
Bethesda
Maryland 20892

Literature Cited

1. Betts MR, Nason MC, West SM, De Rosa SC, Migueles SA, Abraham J, Lederman MM, Benito JM, Goepfert PA, Connors M, Roederer M, Koup RA. HIV nonprogressors preferentially maintain highly functional HIV-specific CD8+ T cells. Blood 2006;107:4781–4789.
2. Precopio ML, Betts MR, Parrino J, Price DA, Gostick E, Ambrozak DR, Asher TE, Douek DC, Harari A, Pantaleo G, Bailer R, Graham BS, Roederer M, Koup RA. Immunization with vaccinia virus induces polyfunctional and phenotypically distinctive CD8(+) T cell responses. J Exp Med 2007;204:1405–1416.
3. 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.