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

SPICE: Exploration and Analysis of Post-Cytometric Complex Multivariate Datasets

Mario Roederer,[1]* Joshua L. Nozzi,[2] Martha C. Nason[3]

[1]ImmunoTechnology Section, Vaccine Research Center, NIAID, NIH, Bethesda, Maryland

[2]Bioinformatics and Computational Biosciences Branch, NIAID, NIH, Bethesda, Maryland

[3]Biostatistics Research Branch, NIAID, NIH, Bethesda, Maryland

Received 17 September 2010; Revision Received 1 December 2010; Accepted 3 December 2010

Grant sponsors: NIAID, NIH

*Correspondence to: Mario Roederer, Vaccine Research Center, NIH, 40 Convent Dr., Room 5509, Bethesda, MD 20892-3015, USA

Email: Roederer@nih.gov

Published online 12 January 2011 in Wiley Online Library (wileyonlinelibrary.com)

DOI: 10.1002/cyto.a.21015

• Abstract

Polychromatic flow cytometry results in complex, multivariate datasets. To date, tools for the aggregate analysis of these datasets across multiple specimens grouped by different categorical variables, such as demographic information, have not been optimized. Often, the exploration of such datasets is accomplished by visualization of patterns with pie charts or bar charts, without easy access to statistical comparisons of measurements that comprise multiple components. Here we report on algorithms and a graphical interface we developed for these purposes. In particular, we discuss thresholding necessary for accurate representation of data in pie charts, the implications for display and comparison of normalized versus unnormalized data, and the effects of averaging when samples with significant background noise are present. Finally, we define a statistic for the nonparametric comparison of complex distributions to test for difference between groups of samples based on multi-component measurements. While originally developed to support the analysis of T cell functional profiles, these techniques are amenable to a broad range of datatypes. Published 2011 Wiley-Liss, Inc.†

• Key terms

immunophenotyping; data analysis; subsets; multivariate; polyfunctional

THE proliferation of polychromatic flow cytometry, in terms of instrumentation (1,2), reagents (3,4), data analysis techniques (5,6), and applications (1,7), has led to the generation of highly complex datasets on a routine basis. The dimensionality of these datasets is high, providing enormous challenges for analysis and data reduction to interpret results. Flow cytometry data analysis software has been designed to help with this, and large-scale efforts toward automation are underway. However, these efforts have been primarily directed at the single-sample analysis arena; the post-processing of complex datasets remains an area requiring innovation.

When analyzing T cell responses, many laboratories routinely measure multiple different functional components on a cell-by-cell basis, e.g., production of IFNγ, IL2, and/or TNF. There is evidence that the pattern of production of these cytokines, termed "quality" (8), rather than the magnitude of the response, may be an important correlate of protection against pathogens (9–13). A single sample measurement may be thought of as a vector of responses; in this example, it would be a seven-element vector that comprises the percentage of T cells that made each unique combination of the three cytokines. The complexity of this analysis (and size of the measurement vector) grows geometrically with each additional measurement, such as CD4 vs. CD8, restriction to particular differentiation stages (14), or inclusion of additional functional outcomes. The goal of the analyses is often to define the element (or combination of elements) within this vector, for which magnitude correlates with a given biological result. As an example, protection afforded by a vaccine against Leishmania major was correlated with the magnitude of only those CD4 cells that simultaneously produced three cytokines—a fraction of the total CD4 response (11).

ISAC
International Society for Advancement of Cytometry

To do this comparison, it becomes necessary to simultaneously analyze many measurement vectors, grouped by various categorical variables that describe each sample: e.g., treatment, gender, age group, or other experimental conditions. Researchers require graphical interfaces to easily display measurement vectors in forms like bar charts or pie charts, where different subsets of individuals can be grouped on the basis of any (combination) of categories. In some cases, averaging (or other mathematical operations) across subsets of individuals is also desired.

To support this mode of data exploration and statistical analysis, we developed a set of algorithms implemented in an Apple Mac™-based software application named SPICE ("Simplified Presentation of Incredibly Complex Evaluations"). SPICE is supported and distributed by the National Institute of Allergy and Infectious Diseases, NIH, and is freely available (http://exon.niaid.nih.gov/spice). Currently, SPICE supports the analysis and display of a single measurement type (e.g., frequency or MFI); development to support multivariate analysis is underway.

Here we report on the algorithms and techniques we used in developing this application and analysis platform, including a unique implementation of a statistical test to compare measurement vectors (distributions) between two groups of samples, so that developers can implement similar tests and displays in other software applications. In addition, we highlight important features of the analysis and presentation of this type of data. While original implementation and examples shown here are based on the analysis of antigen-specific T cells, none of the algorithms are specific to that domain; we routinely use SPICE to analyze and present any complex datasets that are described by multiple categorical variables, including demographic data.

Methods

Data

Data used in this manuscript are either artificial (Fig. 2), or from studies of HIV-specific T cell representation in infected subjects collected in our laboratory. Standard intracellular cytokine staining assays were used. As all data are purely for illustration of algorithms and displays, thus no information about the subjects or assay results is provided. All human samples were collected under NIAID IRB approval. Flow cytometry data was analyzed using FlowJo v9.1 (TreeStar, Ashland, OR). Background subtraction and formatting of exported data from FlowJo was performed with Pestle v1.6.2 (see below). Statistical analysis and display of multicomponent distributions was performed with SPICE v5.1 (freely available from http://exon.niaid.nih.gov/spice/).

Preprocessing of Data for SPICE

A significant power of SPICE is the ability to easily navigate complex data to show distributions and calculate statistics on subsets of measurements grouped, overlaid, or compared on the basis of different categorical descriptors. Many such descriptors are not necessarily part of the exported flow cytometry data, such as demographic or other patient information. To facilitate the analysis of such datasets, we created a data pre-processing program. Pestle handles such functions as background subtraction (e.g., for functional data), editing of the primary dataset, creation of additional categorical variables, and merging with other databases to provide additional categorical descriptors for each sample. Pestle is freely available by request from MR.

Data Computation for Graphical Display

In general, SPICE displays measurement values across a number of categories, for a group of subjects. If the number of categories is n and the number of subjects is m, define a measurement value as V_{ij} where $1 \leq i \leq n$ and $1 \leq j \leq m$. Further define a normalized measurement value as the proportion out of all measurements for a given subject (i.e., each value becomes the fraction of the total response within a subject), as V'_{ij}:

$$V'_{ij} = \frac{V_{ij}}{\sum_{k=1}^{n} V_{kj}}$$

One visualization of this data is a point-chart, where a single point is drawn for each of the m subjects, vertically aligned for each of the n categories. These could be represented as absolute values (V_{ij}), or normalized values (V'_{ij}). A bar chart could also be shown, where each bar illustrates a summary of the distribution of the m values for each category: e.g., interquartile range, min-max, or a bar drawn from zero to the average. The average values for the absolute or normalized distributions are defined as the measurement vectors X and X':

$$X_i = \frac{\sum_{k=1}^{m} V_{ik}}{m} \qquad X'_i = \frac{\sum_{k=1}^{m} V'_{ik}}{m}$$

Another visualization of the data values is a pie chart. In general, the size of each pie slice P_i is the average value for a given category, normalized to the total of the average measurements across all categories (such that the sum of all P'_i is 1). It is important to note that the vector P_i will not in general be the same as the vector P'_i; these two representations convey different information as discussed in Results.

$$P_i = \frac{X_i}{\sum_{k=1}^{n} X_k} \qquad P'_i = \frac{X'_i}{\sum_{k=1}^{n} X'_k}$$

Results

Thresholding of Data for Pie Charts

Pie charts are often used to represent measurement vectors; they can quickly convey patterns of distributions (albeit with shortcomings such as the inability to convey the total magnitude or underlying variability). However, negative values cannot be represented in a pie chart. Negative values arise in measurement distributions as a consequence of background subtraction. Negative values result because of measurement and natural errors (i.e., sometimes the measurement for a stimulated sample is smaller than that for the unstimulated control; a background subtraction yields a value less than

zero). To represent these graphically, the negative values must be set to zero. However, doing so only for negative values will systematically bias the overall dataset, by only increasing some values—thus, the average across all measurements, after thresholding negative values to zero, is greater than it was before thresholding. It is notable that there will be small positive values that are also essentially equivalent to background (and should be zero). The same error distribution leading to negative values will lead as often to positive values for measurements that are nominally zero (Fig. 1).

A reasonable approach to analyzing data where thresholding is required is to set all measurements below some small positive value to be zero. This approach has two significant benefits: first, it removes the systematic bias introduced by zeroing only negative values, by reducing as many values (from small positive values to zero) as increasing (from negative to zero), thus leaving the total and average unchanged for the distribution. Second, this process removes error introduced by including small positive values in representational analyses; i.e., by setting values too small to be distinguished from background to zero, they will not contribute to distributions being displayed.

The difficulty is the determination of what value to use for this thresholding. This value should be at the upper limit of the distribution of measurements that are nominally zero, i.e., the upper limit for background-corrected "negative" measurements. Most of the time, however, such distributions are not measured and this limit is undefined. One approach is to assume that the distribution of background values is symmetric around zero: an estimate of the upper range can then be made by examination of the range of negative values and choosing a value near the lower extent of their range (purple arrow, Fig. 1); a threshold is chosen as the same absolute value (green line, Fig. 1). In this example, the 75th percentile of the values below zero is chosen as a fairly robust measure of the extent of the distribution. With a large number of measurements, a more extreme percentile could be used, such as the 90th. In order for the 90th percentile to be reasonably robust, there should be at least 100 measurement values in the distribution (ensuring that there are at least five measurements below the 90th percentile of the fifty negative values). In the example, choosing the 75th percentile results in 87.5% of nominally-negative values being set to zero: the 50% that are negative and 75% of those that are positive. A more extreme threshold would also zero more "true" (but low) positive values.

The importance of this threshold should not be underestimated. By zeroing all measurements below this small positive value, their effect on summary statistics (such as a mean) is moderated. This reduces noise and possible artefacts in the visualization. Overall, this ensures that the statistics and graphs more accurately reflect the positive response distribution.

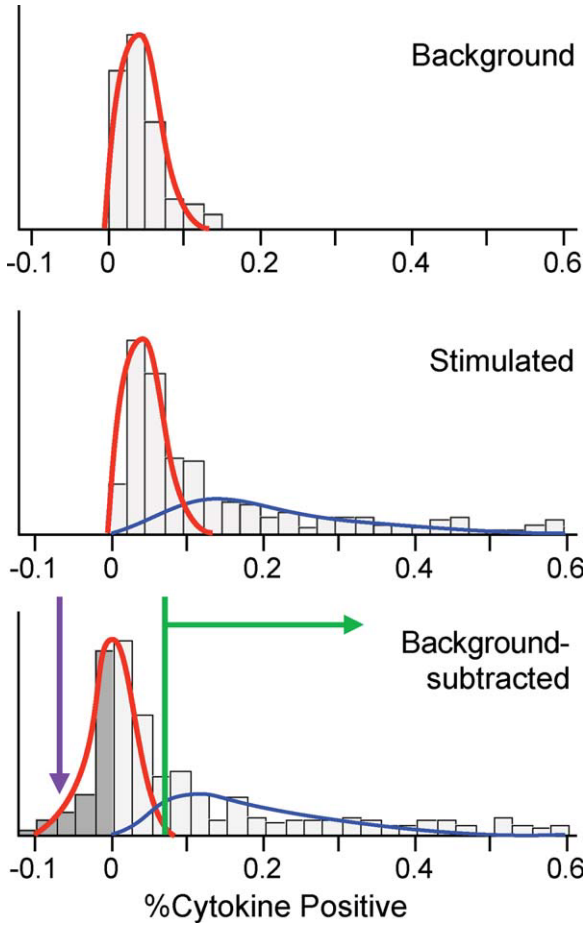

Figure 1. Thresholding distributions to eliminate negative values. PBMC were stimulated with (''Stimulated'') or without (''Background'') antigen; the proportion of T cells producing cytokine is shown as histograms. The top panel shows the results for the control, unstimulated cultures; small levels of background (up to 0.1% of T cells) are evident. The red line is an approximation of the distribution. The middle panel shows the results for the stimulated cultures. A low proportion of positives is evident, the distribution of which is represented by the blue line. To obtain a distribution of the positive event magnitudes, the background value for each culture is subtracted from this measurement; the resulting distribution is shown at bottom. The major set of negative cultures is now centered on zero, with a symmetric spread arising from measurement and experimental errors. A small positive threshold (green) is chosen based on an assumption that the negative cultures are symmetrically distributed and estimating the extent of that distribution from the values below zero (i.e., purple line); for further analysis or display of this distribution, those values can be set to zero. This does not introduce a systematic bias, and statistics will reflect values principally from true positive cultures.

Statistical Comparison of Multicomponent Distributions

Typically, distributions of measurements are compared by parametric (e.g., Students *t* test) or nonparametric (e.g., Wilcoxon rank test) algorithms. However, there is a need for simple, widely-applicable (i.e., not too many assumptions) tests that can be used to compare multicomponent distributions (distributions of measurement vectors). For example, we would like to determine if the pattern of representation of subsets of T cells for one group of individuals is the same or different from the pattern for another group. Each pattern is represented by a measurement vector. The null hypothesis is that the distribution of vectors for each group come from the same distribution and are not distinct.

A metric that can be used for such a test is based on a chi-squared analysis. For each component of a distribution ("slice" in a pie chart), a chi-squared-like value is computed. These values are summed over all categories; this sum ($\overline{\chi}^2$) is a metric by which the distributions can then be compared.

Specifically, we are given two measurement vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$ (see also Methods). Each individual measurement for each of $\boldsymbol{n}$ categories is given by $\boldsymbol{X_i}$ and $\boldsymbol{Y_i}$ ($1 \leq \boldsymbol{i} \leq \boldsymbol{n}$). These might represent the background-subtracted frequencies of antigen-specific cells making different combinations of cytokines: $\boldsymbol{X_1}$ and $\boldsymbol{Y_1}$ would be the average frequencies of $IL2^+IFN\gamma^-TNF^-$ cells for subjects in the two groups, respectively; $\boldsymbol{X_2}$ and $\boldsymbol{Y_2}$ could be $IL2^+IFN\gamma^+TNF^-$, and so forth for all seven combinations of the three cytokines. Define the total for each distribution ($\boldsymbol{T}^X$ and $\boldsymbol{T}^Y$):

$$T^X = \sum_{i=1}^{n} X_i \quad T^Y = \sum_{i=1}^{n} Y_i$$

For each category, we define the expected measurement values $\boldsymbol{E}^X$ and $\boldsymbol{E}^Y$ based on the null hypothesis (no difference between $\boldsymbol{X}$ and $\boldsymbol{Y}$). Then, the differences between the actual measured values and the expected values are $\boldsymbol{D}^x$ and $\boldsymbol{D}^y$:

$$E_i^X = (X_i + Y_i) \times \frac{T^X}{T^X + T^Y} \quad E_i^Y = (X_i + Y_i) \times \frac{T^Y}{T^X + T^Y}$$

$$D_i^X = X_i - E_i^X \quad D_i^Y = Y_i - E_i^Y$$

The normalized chi-squared sum is then defined as follows, setting to zero any term where $\boldsymbol{E_i} = 0$:

$$\overline{\chi}^2 = \sum_{i=1}^{n} \left[\frac{D_i^X \times D_i^X}{E_i^X} + \frac{D_i^Y \times D_i^Y}{E_i^Y} \right]$$

A statistical comparison of the two distributions is accomplished nonparametrically by a partial permutation test. (Note that for distributions with large numbers of subjects, a complete permutation test is impractical; thus, a partial permutation test is performed by Monte Carlo simulation). In this test, all samples in the two groups are aggregated into a single list. For a single iteration, each sample is randomly assigned to one of two test groups in the same proportions as the original groups. The $\overline{\chi}^2$ for the test groups is compared to the original measurement groups. After thousands of iterations, the fraction of comparisons which resulted in a larger (more extreme) $\overline{\chi}^2$ than the measured comparison is determined, and defined as the "$\boldsymbol{P}$" value. This value represents the probability of achieving a difference more extreme than the measured difference.

Because this is a nonparametric permutation test, it can only be performed on distributions with more than one sample per group. The number of samples and the number of iterations determines the minimum observable $\boldsymbol{P}$ value. Specifically, $\boldsymbol{P}$ values less than or equal to 0.05 can only be attained when there are at least three samples in each group; the minimum $\boldsymbol{P}$ value observable is inversely related to the number of permutations; with 1,000 iterations, the lowest $\boldsymbol{P}$ value is $\boldsymbol{P} = 10^{-3}$ (1/1,000). Note that the number of iterations does not affect the magnitude of the estimated $\boldsymbol{P}$ value, only the precision with which it is determined.

The proposed test statistic is derived from a simple analogy with a Chi-squared statistic; this makes it appealing and fairly intuitive, and we believe useful for many different types of analyses. The permutation-based testing algorithm means that no specific form of the distribution of the statistic needs to be assumed, making it applicable across a wide variety of problems. A few characteristics of this test are important to note, however. First, this is a global test of whether the distribution of the vectors differs between the two groups. Therefore, any differences between the groups may be picked up by this test, including inter-group differences in the measurement variability. This is important to consider when the two groups have vectors of proportions based on systematically and dramatically different numbers of events. Another limitation of the global test is that it does not allow for a listing of the individual categories which are or are not different between the two groups, but rather for a judgment of whether there is evidence that the groups differ in any way. In addition, the test is based on an unpaired, two group comparison. In the case of paired data, a test that took advantage of that structure in building a permutation distribution would be more appropriate. Finally, there are many other possible test statistics one could consider here; it would certainly be possible to choose a test statistic that puts more or less weight on certain combinations (for example, those that are rare), and might be possible to increase the efficiency of the test by implementing a different statistic or algorithm that more explicitly took into account the within-individual correlation structure.

Multiple Comparisons Adjustments

SPICE provides the ability to compare two subsets of data for any given category using a Student's *t* test or a Wilcoxon rank test. Given the complexity of the datasets, this can result in significant Type 1 errors due to the sheer number of different categories and subsets that are simultaneously compared. Simple adjustment for multiple comparisons is problematic, since the number of comparisons that are actually performed may be difficult to ascertain. We advocate a two-step approach: first, use the overall distribution comparison statistic described above. In those cases where the distributions are statistically significantly different, the individual categories can be inspected to determine which contribute to the overall difference. The implementation of multiple comparisons adjustment is best left to the researcher, who can evaluate the necessity in the context of the overall experiment.

Impact of Normalization

Normalization can lead to very different interpretations and statistical comparisons of a distribution; it is important to be able to analyze both normalized and un-normalized data. For analysis of cytokine responses, we refer to the normalized distribution as the "quality" of the response (8). In this context, "normalization" means representing each measurement value as its relative contribution to the total of all measure-

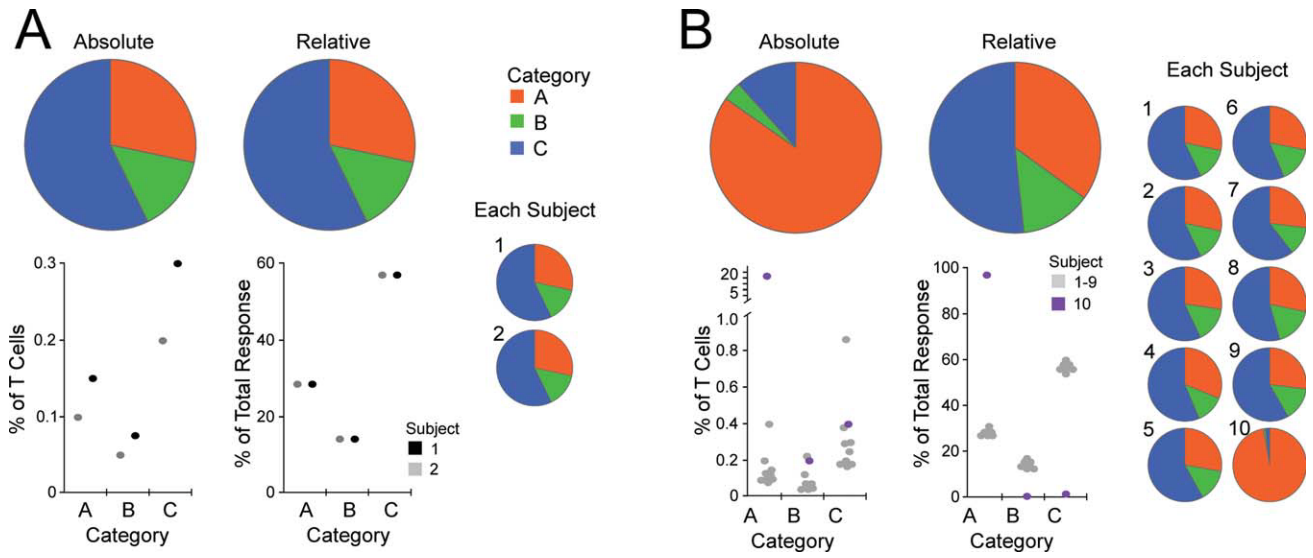

Figure 2. Impact of normalization on visual displays. These are artificial datasets for illustration; they represent dividing up a T cell response into three categories. (**A**) Data for two individuals is shown. The two individuals have different overall magnitude of the representation of the three categories of cells, but the response distributes identically into the categories. Hence, the pie charts that average the results (top) are identical whether they use absolute, unnormalized data (left) or relative, normalized data (right). (**B**) Here, 10 individuals are analyzed, one of whom is a sharp outlier from the other 9. The pie charts averaging the 10 individuals are very different when using absolute numbers (left) vs. relative numbers (right). This is because the absolute representation weights each individual according to the absolute representation of the subsets thus, Subject no. 10 is weighted almost 100-fold greater than the other subjects. The resulting averaged pie chart looks similar to the pie chart for Subject no. 10, and not like the other nine. When averaging the normalized values (right), Subject no. 10 is weighted equally to all others, and thereby contributes only 10% of the information. Hence, the average pie chart looks very similar to the majority of subjects.

ments for that sample—in other words, converting absolute values to "% of total." For the graphics, the normalized measurement vectors $\boldsymbol{X}'$ are used in place of the raw measurement vectors $\boldsymbol{X}$ (see methods).

As an example, see Figure 2. In Figure 2A, the distribution of three subsets is given for two individuals. The overall representation of the three subsets in Subject 2 is 50% greater than for Subject 1, leading to disparate points in the chart of the absolute unnormalized values (left panel). After normalizing to the total representation of only these three subsets, it is evident that the relative (normalized) distributions are identical in composition (right panel), despite differing in magnitude. A contrasting result can be seen in Figure 2B. Here, data from 10 individuals is analyzed; nine are very similar and one is an outlier. A pie chart visualization of the normalized data is quite different than the un-normalized data, because of the relative weighting. For the normalized data, the outlier sample is weighted by its representation in the sample set (i.e., one-tenth); in the unnormalized display, the outlier sample is weighted by the magnitude of the measurement. The unnormalized distribution is what an analysis of a mixture of an equal number of cells from each of the 10 individuals would look like, but it does not necessarily represent accurately any (or perhaps even most) of the individual samples.

Impact of Noise on Normalized Distribution

The differential weighting of sample data described above is particularly important to take into account when there is significant measurement noise in (some) samples. For example, consider the problem of determining the phenotype of antigen-responsive T cells (i.e., those that are cytokine-positive following stimulation). Background subtraction is necessary to determine the magnitude of this response. However, it is not possible to use background subtraction when determining the phenotype of the response (i.e., both stimulated and background cytokine-positive cells may be 100% CD4-positive; subtracting this would result in 0%, which is clearly nonsensical). Furthermore, it is usually the case that the phenotype of background-responding cells is different than antigen-responding cells. In a sample where a majority of the responding cells are antigen-specific, the overall phenotype reflects those cells since the contribution of background-cells is low. But for samples with a low response magnitude, this is not the case: in a sample where the magnitude of the positive response is of the same magnitude as the background, half of the cells in the phenotype analysis would be antigen-specific and half would be background, yielding a mixed phenotype.

When constructing a representation of the phenotype of antigen-responding cells from multiple subjects, inclusion of low-response subjects will skew this phenotype to look more like background cells. Here, perhaps, it is more desirable to display un-normalized data, so those samples with "true" responses are weighted more heavily and the average phenotype is more reflective of those responses. Another reasonable approach would be to use normalized distributions, and eliminate samples from the distribution where the magnitude of the responding cells was less than (for example) three-fold that of the background.

An example of this is shown in Figure 3. In this experiment, the phenotype (naïve, central memory, transitional

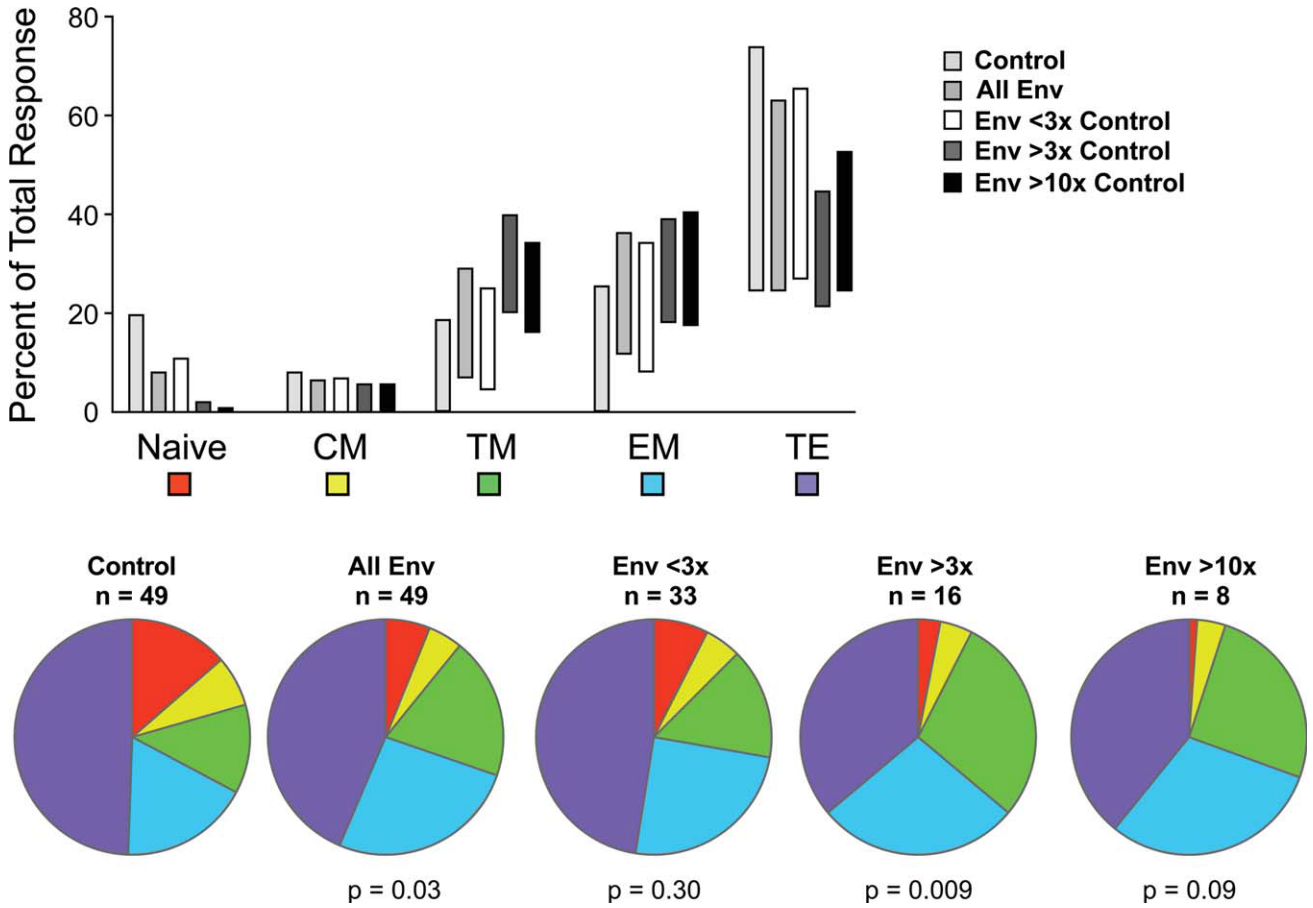

Figure 3. Elimination of low-responders improves discrimination. The phenotype of all cytokine positive cells was determined; the fraction that falls into a naïve, central memory (CM), transitional memory (TM), effector memory (EM), or terminal effector (TE) memory subset is shown in the bar chart and pie charts. Distributions are shown for the background (unstimulated) control, for all stimulated samples (stimulated with HIV envelope peptides, Env), for those samples with Env responses less than three-fold above background, and for those with responses greater than 3-fold or 10-fold above background. Note that the phenotype for a response that is three-fold above background would represent a mixture of cells that is 25% background and 75% antigen-specific. Bars show interquartile ranges for each measurement. p values were computed using the permutation test described in the text, comparing each distribution against the control.

memory, or terminal effector) of antigen-specific cells was determined. Samples comprised PBMC from 49 individuals stimulated with a variety of peptide pools, or left unstimulated (background control). The phenotype of the background-population is different from the antigen-specific; notably, there is a significant representation of naïve T cells that spontaneously produce cytokine. The phenotype of all responders is different from this, with a P value of 0.03; however, there are still a proportion of cells with a naïve phenotype. Because the responses are quite low in these samples, this phenotype arises from the contribution of samples with few or no antigen-specific cells that are still being averaged into the entire mixture. Indeed, by separating the cohort into those individuals with low responses (less than three-fold above background) vs. high responses, this effect becomes clear. In the high responders group, there are no naïve T cells contributing to the distribution, and the statistical comparison (compared to control) becomes much more significant. While not surprising, it is critical to recognize that the averaging required to compute the size of each pie slice $\boldsymbol{P_i}$ (see methods) gives equal weighting to low vs. high responders, and that this skews the average phenotype toward that of the background events.

The selection of samples with responders that are three-fold above control is somewhat arbitrary; at this level, it guarantees that the contribution of background events to the phenotype is less than 25% (i.e., three-fold implies 75% specific events, 25% background). Figure 3 also illustrates the profiles for only those samples with responses that are 10-fold above background. For these samples, the contribution of naïve T cells to the memory response is virtually eliminated, as predicted. Overall, the phenotype of these highly-enriched env-specific cells is essentially identical to that seen with the samples that are three-fold above background, lending credibility to this threshold as reasonable. Notably, the use of the far more restrictive threshold reduces the number of samples to 8; the smaller sample size results in less significance when comparing the phenotyping profiles.

Overall, Figure 3 illustrates that it is important to compare the phenotype of the "positive" response with the "background" control, as well as to evaluate the effect of varying the threshold of positivity. Finally, comparison of the profiles obtained for samples that are above background but below the threshold for "high positives" should be done to consider whether the selected "high positive" responders are representative of the whole population.

Discussion

Polychromatic flow cytometry gives us the ability to divide the immune system into a large number of discrete subsets. For example, measuring three cytokine responses in T cells immediately results in 16 subsets (eight combinations of all three cytokines, for both CD4 and CD8). Correlates analysis attempts to identify whether any one (or a combination of these) subsets is related to a biological or clinical measure. During data exploration, this process is accomplished most efficiently by pattern recognition—relying on the human brain to pick out associations using graphical interfaces that aggregate all of this information in some simplified forms.

As the complexity of the data set increases geometrically by addition of additional measurements, such as additional functions or phenotypic markers, we discovered that there were no software tools available to easily display and compare the patterns of distributions across different sample sets. In addition, there was a need to easily reduce the complexity of the data by collapsing different dimensions at will: for example, to reduce the eight cytokine combinations (from three functions) to four by disregarding the contributions of one of the cytokines.

For this purpose, we developed the program SPICE. The user interface in SPICE has been designed to easily select any subset of categorical criteria for display, overlay, and/or statistical analysis. These operations are termed "pivot" operations in spreadsheet parlance; SPICE performs them in a user-friendly, highly optimized fashion.

During the development of SPICE for the presentation and analysis of data from our experiments, we discovered the need to address how distributions are thresholded for display in pie charts (a common display type for multicomponent distributions). Specifically, background-corrected assays common in biology will result in negative values that cannot be properly represented in pie charts. We describe an approach to select a threshold for zeroing measurement values that minimizes systematic bias and maximizes the information content from positive measurements. The magnitude of this threshold should be reported when using this type of approach.

The comparison of distributions from individual categories in these complex datasets is fraught with danger of Type 1 statistical errors—i.e., mistakenly identifying a difference as significant. This comes from the large number of comparisons that can be easily performed. Researchers must be particularly careful to correct for multiple comparisons when pre-specified criteria were not set. To help overcome such bias, we developed a statistical test that compares all of the distributions at once.

This algorithm is based on a chi-squared metric, and uses a nonparametric partial permutation (Monte Carlo simulation) to define how extreme the difference between two sample sets is. Currently, this comparison is not paired; we are considering algorithms for implementing a paired version of this test. In exploring datasets, we advocate only using comparisons on individual categories once an overall significant difference based on the total distributions has been found.

A number of recent examples illustrate the power of this statistical comparison. To date, most of the use has been to distinguish the quality (functional repertoire) of HIV- or SIV-specific T cells depending on therapy (15), clinical status (16–20), or vaccine strategy (21,22). The statistic was also used to show that alloreactive T cells in graft-vs.-host disease have a unique functional profile (23), that tuberculosis-specific T cell function changes following therapy (24), and that HIV-1 and HIV-2 specific T cells differ in function and phenotype (25). These examples illustrate the power of the ability to compare multi-component measurements, particularly, by reducing the comparison to a single test rather than comparing individual components and requiring a correction for multiple comparisons.

In summary, we outline a number of algorithmic considerations when analyzing complex data described by multiple categories. We report on a freely-available, US Government-supported application, SPICE, which implements these algorithms in a user-friendly graphical interface. A supporting program, Pestle, is also available to assist with data pre-processing, database merging, and data formatting. These programs represent a first step in the aggregation and analysis of data from multiple flow cytometric analyses, i.e., post-cytometric data analysis.

Acknowledgments

SPICE was originally conceived by MR through extensive discussions of concepts with Dr. Pratip Chattopadhyay. The authors thank Drs. Dean Follman, Steve De Rosa, Pratip Chattopadhyay, Michael Betts, Melody Duvall, Melissa Precopio, Patricia Darrah, Yolanda Mahnke, and other members of the VRC Laboratory of Immunology for detailed discussions, suggestions, and testing of the algorithms and programs. Development of SPICE version 5 was assisted by Yasmin Mohamoud, Yentram Huyen, and other members of the Bioinformatics and Computational Biosciences Branch of the NIAID, NIH.

Literature Cited

1. Chattopadhyay PK, Hogerkorp CM, Roederer M. A chromatic explosion: The development and future of multiparameter flow cytometry. Immunology 2008;125: 441–449.
2. Perfetto SP, Chattopadhyay PK, Roederer M. Seventeen-color flow cytometry: Unraveling the immune system. Nat Rev Immunol 2004;4:648–655.
3. Chattopadhyay PK, Price DA, Harper TF, Betts MR, Yu J, Gostick E, Perfetto SP, Goepfert P, Koup RA, De Rosa SC, Bruchez MP, Roederer M. Quantum dot semiconductor nanocrystals for immunophenotyping by polychromatic flow cytometry. Nat Med 2006;12: 972–977.
4. Roederer M, Kantor AB, Parks DR, Herzenberg LA. Cy7PE and Cy7APC: Bright new probes for immunofluorescence. Cytometry 1996;24:191–197.
5. Lugli E, Roederer M, Cossarizza A. Data analysis in flow cytometry: The future just started. Cytometry A 2010;77A:705–713.
6. Roederer M, Moody MA. Polychromatic plots: Graphical display of multidimensional data. Cytometry A 2008;73A:868–874.

7. Chattopadhyay PK, Roederer M. Good cell, bad cell: Flow cytometry reveals T-cell subsets important in HIV disease. Cytometry A 2010;77A:614–622.
8. Seder RA, Darrah PA, Roederer M. T-cell quality in memory and protection: Implications for vaccine design. Nat Rev Immunol 2008;8:247–258.
9. 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.
10. Boaz MJ, Waters A, Murad S, Easterbrook PJ, Vyakarnam A. Presence of HIV-1 Gag-specific IFN-gamma+IL-2+ and CD28+IL-2+ CD4 T cell responses is associated with nonprogression in HIV-1 infection. J Immunol 2002;169:6376–6385.
11. Darrah PA, Patel DT, De Luca PM, Lindsay RW, Davey DF, Flynn BJ, Hoff ST, Andersen P, Reed SG, Morris SL, Roederer M, Seder RA. Multifunctional TH1 cells define a correlate of vaccine-mediated protection against Leishmania major. Nat Med 2007; 13:843–850.
12. Harari A, Petitpierre S, Vallelian F, Pantaleo G. Skewed representation of functionally distinct populations of virus-specific CD4 T cells in HIV-1-infected subjects with progressive disease: Changes after antiretroviral therapy. Blood 2004;103:966–972.
13. Sutherland JS, Young JM, Peterson KL, Sanneh B, Whittle HC, Rowland-Jones SL, Adegbola RA, Jaye A, Ota MO. Polyfunctional CD4(+) and CD8(+) T cell responses to tuberculosis antigens in HIV-1-infected patients before and after anti-retroviral treatment. J Immunol 2010;184:6537–6544.
14. Appay V, van Lier RA, Sallusto F, Roederer M. Phenotype and function of human T lymphocyte subsets: Consensus and issues. Cytometry A 2008;73A:975–983.
15. Macatangay BJ, Szajnik ME, Whiteside TL, Riddler SA, Rinaldo CR. Regulatory T cell suppression of Gag-specific CD8 T cell polyfunctional response after therapeutic vaccination of HIV-1-infected patients on ART. PLoS One 2010;5:e9852.
16. Almeida JR, Price DA, Papagno L, Arkoub ZA, Sauce D, Bornstein E, Asher TE, Samri A, Schnuriger A, Theodorou I, Costagliola D, Rouzioux C, Agut H, Marcelin AG, Douek D, Autran B, Appay V. Superior control of HIV-1 replication by CD8+ T cells is reflected by their avidity, polyfunctionality, and clonal turnover. J Exp Med 2007;204:2473–2485.
17. Ferre AL, Hunt PW, Critchfield JW, Young DH, Morris MM, Garcia JC, Pollard RB, Yee HF Jr, Martin JN, Deeks SG, Shacklett BL. Mucosal immune responses to HIV-1 in elite controllers: A potential correlate of immune control. Blood 2009;113:3978–3989.
18. Maenetje P, Riou C, Casazza JP, Ambrozak D, Hill B, Gray G, Koup RA, de Bruyn G, Gray CM. A steady state of CD4+ T cell memory maturation and activation is established during primary subtype C HIV-1 infection. J Immunol 2010;184:4926–4935.
19. Nemes E, Bertoncelli L, Lugli E, Pinti M, Nasi M, Manzini L, Manzini S, Prati F, Borghi V, Cossarizza A, Mussini C. Cytotoxic granule release dominates gag-specific CD4+ T-cell response in different phases of HIV infection. AIDS 2010;24:947–957.
20. Favre D, Lederer S, Kanwar B, Ma ZM, Proll S, Kasakow Z, Mold J, Swainson L, Barbour JD, Baskin CR, Palermo R, Pandrea I, Miller CJ, Katze MG, McCune JM. Critical loss of the balance between Th17 and T regulatory cell populations in pathogenic SIV infection. PLoS Pathog 2009;5:e1000295.
21. 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 2010;5:e9015.
22. Sun Y, Santra S, Schmitz JE, Roederer M, Letvin NL. Magnitude and quality of vaccine-elicited T-cell responses in the control of immunodeficiency virus replication in rhesus monkeys. J Virol 2008;82:8812–8819.
23. Melenhorst JJ, Scheinberg P, Chattopadhyay PK, Gostick E, Ladell K, Roederer M, Hensel NF, Douek DC, Barrett AJ, Price DA. High avidity myeloid leukemia-associated antigen-specific CD8+ T cells preferentially reside in the bone marrow. Blood 2009;113:2238–2244.
24. Young JM, Adetifa IM, Ota MO, Sutherland JS. Expanded polyfunctional T cell response to mycobacterial antigens in TB disease and contraction post-treatment. PLoS One 2010;5:e11237.
25. Duvall MG, Precopio ML, Ambrozak DA, Jaye A, McMichael AJ, Whittle HC, Roederer M, Rowland-Jones SL, Koup RA. Polyfunctional T cell responses are a hallmark of HIV-2 infection. Eur J Immunol 2008;38:350–363.