Published 2001 Wiley-Liss, Inc.† Cytometry 45:37–46 (2001)

Probability Binning Comparison: A Metric for Quantitating Univariate Distribution Differences

Mario Roederer,[1]* Adam Treister,[2] Wayne Moore,[3] and Leonore A. Herzenberg[3]

[1]Vaccine Research Center, NIH, Bethesda, Maryland
[2]Tree Star, Inc., San Carlos, California
[3]Department of Genetics, Stanford University, Stanford, California

Received 21 December 2000; Revision Received 2 May 2001; Accepted 7 June 2001

Background: Comparing distributions of data is an important goal in many applications. For example, determining whether two samples (e.g., a control and test sample) are statistically significantly different is useful to detect a response, or to provide feedback regarding instrument stability by detecting when collected data varies significantly over time.

Methods: We apply a variant of the chi-squared statistic to comparing univariate distributions. In this variant, a control distribution is divided such that an equal number of events fall into each of the divisions, or bins. This approach is thereby a mini-max algorithm, in that it minimizes the maximum expected variance for the control distribution. The control-derived bins are then applied to test sample distributions, and a normalized chi-squared value is computed. We term this algorithm Probability Binning.

Results: Using a Monte-Carlo simulation, we determined the distribution of chi-squared values obtained by comparing sets of events derived from the same distribution. Based on this distribution, we derive a conversion of any given chi-squared value into a metric that is analogous to a t-score, i.e., it can be used to estimate the probability that a test distribution is different from a control distribution. We demonstrate that this metric scales with the difference between two distributions, and can be used to rank samples according to similarity to a control. Finally, we demonstrate the applicability of this metric to ranking immunophenotyping distributions to suggest that it indeed can be used to objectively determine the relative distance of distributions compared to a single control.

Conclusion: Probability Binning, as shown here, provides a useful metric for determining the probability that two or more flow cytometric data distributions are different. This metric can also be used to rank distributions to identify which are most similar or dissimilar. In addition, the algorithm can be used to quantitate contamination of even highly-overlapping populations. Finally, as demonstrated in an accompanying paper, Probability Binning can be used to gate on events that represent significantly different subsets from a control sample. Cytometry 45: 37–46, 2001. Published 2001 Wiley-Liss, Inc.†

Key words: flow cytometry; data analysis; K-S statistics; histogram comparisons

Comparing univariate distributions is a common task in the analysis of flow cytometric data. Such comparisons are useful for quality control (during or after sample acquisition) to ensure that sample measurements, particularly light scatter, are not drifting due to undetected changes in the fluidics (particularly, nozzle clogs), excitation intensity, or other instrument components. Automated comparisons that are nonparametric and do not require user intervention would be particularly well-suited for such feedback.

Comparison of distributions is also useful for analyzing biological responses. For example, examination of a marker of activation to determine the fraction of responding cells (and the extent to which they respond) is often used to quantify immune activation, cellular responses to stimulation, etc. The principal problem with such assays is that the distribution of responding cells often overlaps significantly with the nonresponding cells, making it difficult to determine quantitative measures of the response.

Several algorithms are currently in use to address these problems. The first described nonparametric test for flow cytometric data was based on Bayes' theorem applied to channel-by-channel determination of means and standard deviations (1). Bagwell has since developed a much more sophisticated comparison algorithm, now termed "SED" (2). The Kolmogorov-Smirnoff (K-S) statistic provides a probability that two flow cytometric univariate histograms are different (3, 4). However, the K-S statistic has a signif-

*Correspondence to: Mario Roederer, Vaccine Research Center, NIH, 40 Convent Dr., Room 5509, Bethesda, MD 20892-3015.
E-mail: Roederer@drmr.com

icant drawback, including that while it likely *underestimates* the probability with which discrete data sets (such as flow histograms) are unique, it is far *too sensitive* to provide meaningful values (5). For example, even collecting histograms for the same cells twice in succession often results in distributions that are statistically significantly different.

Cox applied a chi-square statistic to compare histograms (5). In this algorithm, the number of events in any given bin is used to calculate a channel-by-channel confidence interval. The difference between two distributions is essentially normalized by these confidence intervals to determine where the variation is greater than expected. This method works well when large numbers of events (relative to the number of bins into which each histogram is divided) are available; with more limited event numbers, the predicted variation for any given channel becomes very large. The authors noted that the number of events per channel should be at least 20. Nonetheless, the expected variance is highly variable across a distribution, meaning that the statistic is weighted towards those portions of the distribution containing more events—making it less sensitive to changes in outlier populations if the bulk of the data is unchanged.

Neither the K-S probability nor the chi-square statistics provide a measure of the percent positive cells, i.e., those above the control. Overton published an algorithm that essentially subtracts histograms on a channel-by-channel basis (6). This algorithm accurately quantitates responding populations; the accuracy depends on the frequency of responders as well as the separation between the distributions of responders and non-responders. The Overton algorithm does not provide an estimate of the probability with which the distributions are distinct.

Finally, Lampariello developed a parametric statistic for determining percent positive cells based on a model of the distribution of cellular autofluorescence (7, 8). This method is distinct from the others in that it is parametric, i.e., it fits an expected distribution to the control (unstained) sample. It is useful only when the control sample is indeed unstained, and cannot be used as a general method for comparing any distributions.

We developed a novel algorithm for the comparison of distributions, which we term Probability Binning (PB) Comparison. The PB comparison is a related to the Cox chi-square approach, but with modified binning such that it minimizes the maximal expected variance. In an accompanying manuscript (9), we extend PB comparison algorithm to compare multivariate data. In this manuscript, we describe the application of the algorithm to univariate data, and show that it not only detects small differences between histograms, it does so in a quantitative way. This means that the algorithm can be used to rank distributions in terms of how similar they are to a control. The output of the algorithm, a single value, can be used to determine the statistical significance of the difference between distributions, to estimate the relative distance between the distributions, and can even be used to estimate the representation of a highly overlapping populations within a test sample, irrespective of the actual distributions of the two populations.

MATERIALS AND METHODS

Data Analysis

Artificial univariate distributions were created as FCS files using a specially modified version of FlowJo (Tree Star, San Carlos, CA). Distribution comparisons, including Probability Binning, Kolmogorov-Smirnoff (4), Overton cumulative subtraction (6), and SED (2), were performed using the standard FlowJo version 3.4; additional analyses were performed using JMP for Macintosh (SAS Institute).

Cell Staining and Flow Cytometric Analyses

Human PBMC were obtained by standard methods; at least 10^6 cells were used for each stain. Cells were stained on ice for 15 min with fluorescently-conjugated antibodies and then washed three times with staining medium (biotin, flavin-deficient RPMI supplemented with 4% newborn calf serum and 0.02% sodium azide). Data were collected on a FACStarPlus (Becton Dickinson, San Jose, CA).

RESULTS

Probability Binning Algorithm

To compare different distributions, we took the approach of binning the distributions–i.e., dividing the distributions into a relatively small number of bins. The number of events falling into these bins is compared for a test and control sample, and a chi-squared computation is performed on the counts (i.e., the square of the differences divided by the sum). Rather than the standard binning algorithm, which selects bins of equal width, our binning algorithm selects bins such that each bin contains the same number of events. The result of this algorithm is that a randomly-selected event from the control sample has an equal probability of falling into any of the bins. This process results in bins of unequal width (Fig. 1), with the property that each bin carries equal weighting when used for further statistical tests. Importantly, this process minimizes the maximum expected variance for the bins. We refer to this process as PB.

In order to compare a test distribution, the number of events for that test distribution that fall into each of the bins generated from the control is calculated. It is apparent that the total number of bins, B, into which the control distribution is divided should affect the comparison. For example, if $B = 1$, then no distributions would be different; higher values of B are predicted to resolve distributions with greater fidelity. The number of events that falls into bin i for the control sample is c_i and for the test sample, s_i. Given that the total number of events in the control sample is E^c and the test sample is E^s, we define the normalized bin counts as:

$$c_i' = \frac{c_i}{E^c} \quad \text{and} \quad s_i' = \frac{s_i}{E^s}$$

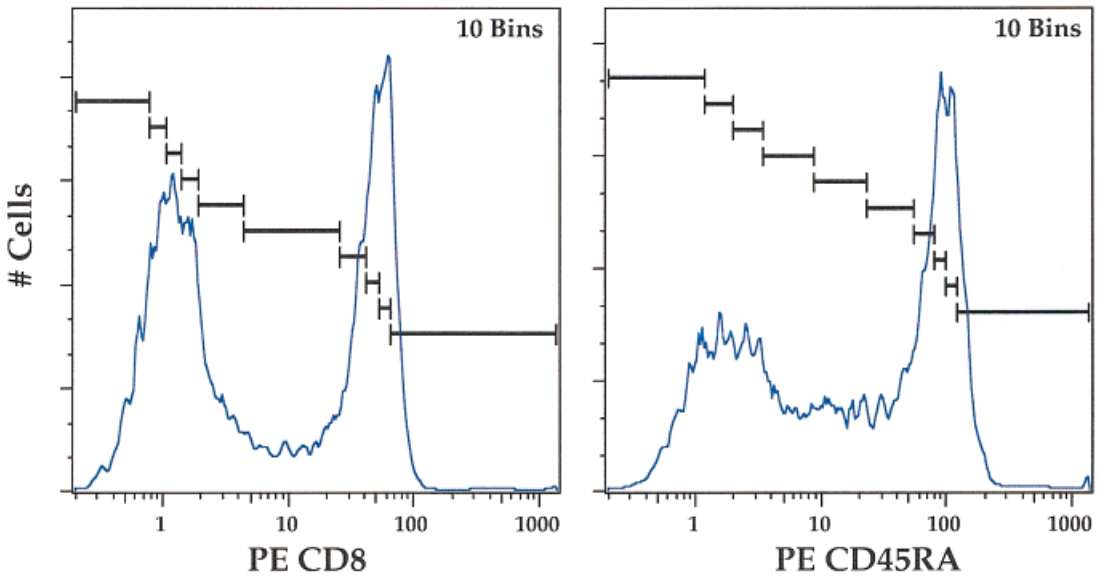

FIG. 1. Probability Binning. Shown are two examples of univariate distributions of putative control samples. Each distribution is divided into ten bins, containing equal numbers of cells. The first bin ranges from the lowest possible value up to the point where 10% of the events in the control sample have been included. The second bin ranges from the 10^{th} to the 20^{th} percentiles, and so forth. Thus, if an event is taken at random from this distribution, it has an equal probability of falling within any of the ten bins. Note that the bins are much narrower around high density clusters (i.e., where there is considerably more information per unit intensity), and much wider where events are much more disperse (and correspondingly lower information density per unit intensity).

Thus, s_i is the fraction of the total events in the test sample that fall within bin i.

We then define a normalized χ^2 value (χ'^2):

$$\chi'^2 = \sum_{i=1}^{B} \frac{(c'_i - s'_i)^2}{(c'_i + s'_i)}$$

Theoretically, χ'^2 can range in values from a minimum of zero to a maximum of 2, irrespective of the number of events in the test or control samples.

Derivation of the PB Metric

When comparing two sets of values derived from the same distribution, χ'^2 will in general be greater than zero

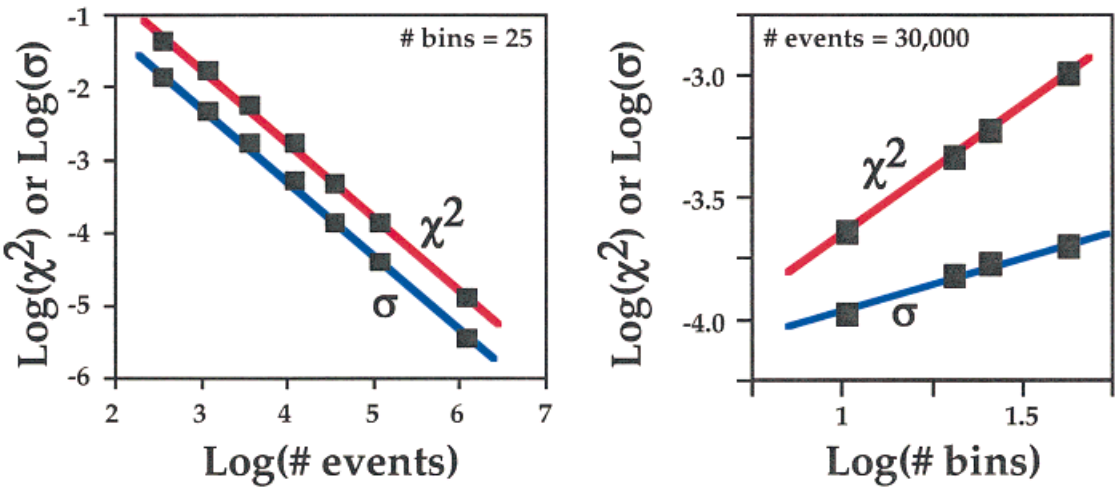

FIG. 2. Dependence of the minimum observed χ'^2 (see Results) on binning and event counts. Data files were generated containing events randomly generated from the same original distribution; each file had between 100 and 10^6 events. For each data point in the graphs above, 1000 files were generated; the first was compared against all others. The distribution of χ'^2 for each of these 999 comparisons was roughly Gaussian; plotted above is the mean and standard deviation (σ) of the χ'^2 values vs. event count (left) where each distribution was divided into 25 bins, or vs. bin count (right) for distributions with 30,000 events.

due to random variation (sampling error). We determined what distribution of χ'^2 results for comparing identical distributions, in order to empirically derive the minimum statistically significant value of χ'^2 (i.e., the minimum value from which a confident decision of histogram difference can be made). To do this, we generated seven sets of 1,000 FCS data files; each set had the same number of events, ranging from 100 to 10^6 per file. The value for each event was generated using a pseudo-random number generator that creates normal (Gaussian) distributions (10).

Each set of data files was then subjected to the PB using either 10, 20, 25, or 50 bins, and the mean and standard deviation of 1,000χ'^2 values was calculated. The resulting χ'^2 distributions for each case were nearly normally-distributed (data not shown), making the standard deviation of the distribution an appropriate measure of the variance of χ'^2.

The results of these analyses are summarized in Figure 2. We found that the mean χ'^2 for a comparison of events with the same distribution (mean minimum observed χ'^2, or $\overline{\chi'^2}$) was proportional to the number of bins and inversely proportional to the number of events. In addition, we found that the standard deviation of the distribution of χ'^2, $\sigma_{\chi'^2}$, was inversely proportional to the number of events and inversely proportional to the square root of the number of bins.

By least-squares multivariate modelling to the 28 measured values of the mean minimum observed χ'^2 and their corresponding standard deviations, we determined the constants of proportionality to derive equations to predict the distribution of χ'^2 based only on the number of bins used in the PB algorithm and the number of events E (where E is the lesser of E^s and E^c):

$$\overline{\chi'^2} = \frac{B}{E}$$

$$\sigma_{\chi'^2} = \frac{\sqrt{B}}{E}$$

The value $\overline{\chi'^2}$ is the minimum potentially meaningful value for a comparison. In other words, this is the value that is obtained for comparing equivalent data sets; any χ'^2 equal to or less than this indicates that the two compared data sets have the same distribution.

Because the distribution of the PB χ'^2 for comparison of equivalent sets of data is normal, we can define a metric that is analogous to the t-score for any measured χ'^2_{m}. This relates to the significance of the value of χ'^2_{m}. Values less than zero are set to zero, since they cannot arise from statistically-different distributions.

$$T(\chi) = \max\left(0, \frac{(\chi'^2_m - \overline{\chi'^2})}{\sigma_{\chi'^2}}\right)$$

Given $\chi_{\mathrm{m}}'^2$, this metric, $T(\chi)$, is the number of standard deviations above the minimum meaningful value for that comparison. Therefore, a value $T(\chi) = 0$ implies that the two distributions are indistinguishable ($p = 0.5$); a value $T(\chi) = 1$ means that $\chi_{\mathrm{m}}'^2$ is one standard deviation above the minimum value and that the two distributions are the same with a probability $p < 0.17$. A value $T(\chi) > 4$ implies that the two distributions are the same with a $p < 0.01$ (i.e., 99% confidence that the distributions are *different*).

Validation of the PB Metric

In order to validate the metric $T(\chi)$, we generated another series of data files. In this series, we added a varying number of positive events to the negative distribution, where the positive events were also distributed normally but with a mean that ranged anywhere from 0.1 to 4.0 standard deviations above the negative events. These bimodal distributions were compared against a file containing only the negative distribution. As for Figure 2, the number of events and number of PB bins used in the statistic was also varied to determine the influence of these values on $T(\chi)$.

Shown in Figure 3 is the dependence of $T(\chi)$ on the fraction (% positive) and separation (ΔPeak) of the events in the test distribution. (See Fig. 4 for examples of these distributions). $T(\chi)$ scales monotonically and smoothly with both the fraction of positive events as well as the separation between the positive and negative events. Once the separation is such that there is no more overlap between the positive and negative events (i.e., more than two standard deviations apart), $T(\chi)$ no longer increases with increasing separation. This is expected, because the $T(\chi)$ does not depend on the *shape* of the distribution (i.e., is nonparametric with regard to the distribution of events).

Therefore, $T(\chi)$ is a statistic which not only provides an indication of the probability with which two distributions are different, but simultaneously provides a metric by which multiple distributions can be ranked. The higher the value of $T(\chi)$, the less like the control sample.

A thorough analysis of the dependence of $T(\chi)$ on the total number of events, the number of PB bins (B) used, the representation of positive events, and the distance between the positive and negative peaks is shown in Figure 4. These contour plots illustrate several features about the $T(\chi)$ metric. First, the minimum number of positive events for a well-separated ($>2\sigma$) population that results in a $T(\chi)$ value with a 99% confidence of difference (i.e., $T(\chi) > 4$), is about 100 (when $B = 25$). Interestingly, this minimum detectable event count does not depend on the number of negative events. Therefore, the algorithm can detect 100 events out of 10^4 (1% positive) with the same precision as 100 events out of 10^6 (0.01% positive). Second, the minimum number of detectable positive events depends only slightly on the bin count B. Higher values of B can detect lower % positives. (However, because $\overline{\chi'^2}$, the minimum statistically significant value, is proportional to B, it is possible that increasing the number

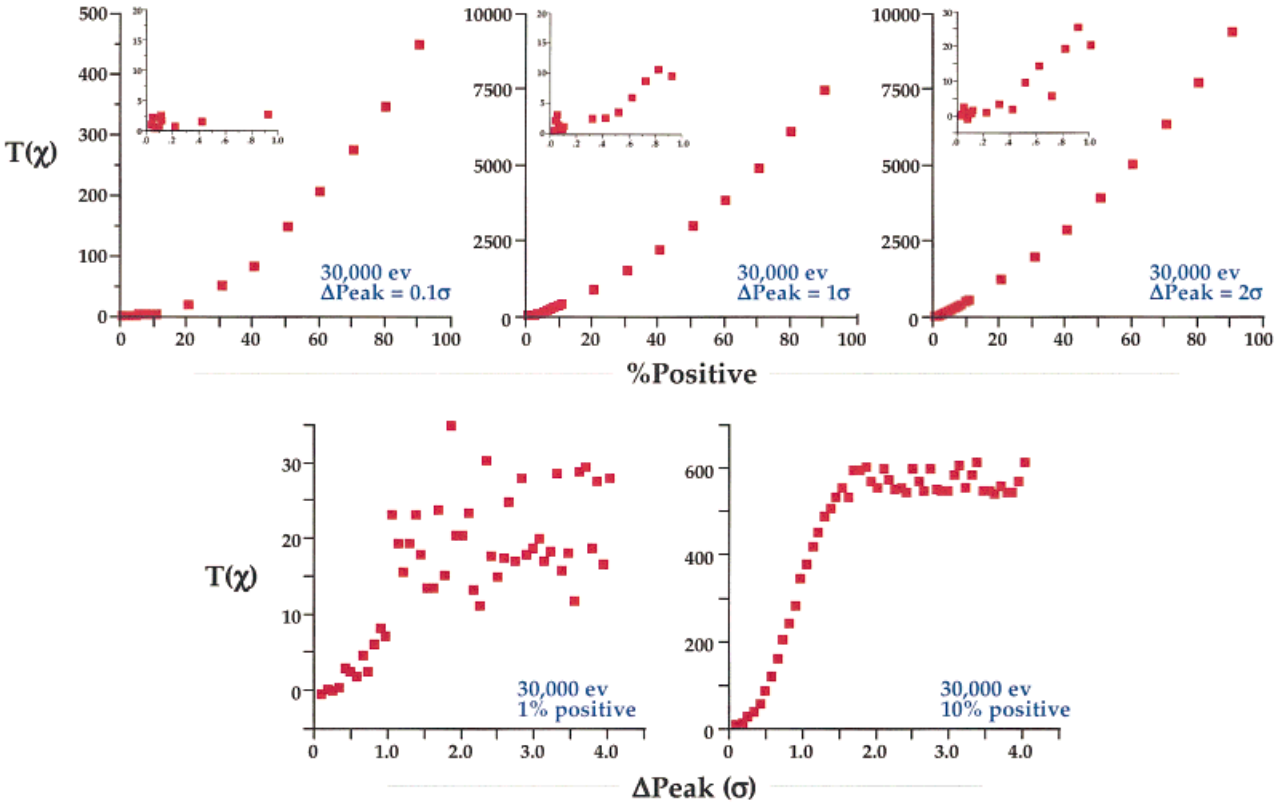

FIG. 3. $T(\chi)$ as a metric to quantitate differences in univariate distributions. Data files were generated containing events randomly selected from two different Gaussian distributions, one a control distribution, the second a positive distribution centered above the control distribution by a distance measured in units of standard deviations (each Gaussian had the same width = 1σ). Each data file had a defined proportion of positive events. Each point represents the $T(\chi)$ calculated from a comparison of such a data file with a control distribution (0% positive). (Upper panels) $T(\chi)$ vs. % positive, where the positive distribution is nearly overlapping the control (the distributions are only 0.1σ apart), slightly overlapping (1σ), and nearly non-overlapping (2σ). Insets show graph at lower values. Note that $T(\chi)$ varies monotonically and smoothly with increasing proportions of positive events, although for lower separation, greater numbers of events are required to achieve statistical significance. (Lower panels) $T(\chi)$ vs. the inter-peak distance for distributions with 1% or 10% positive events. With one 1% positive, $T(\chi)$ becomes statistically significant only when positive events are at least 0.5σ above the control; at 10%, $T(\chi)$ is significant event when the peaks are separated by only 0.1σ.

of bins too high will lead to a decrease in the resolving power). Third, as expected, the less the separation between the positive and negative events, the greater the % positive must be in order for the algorithm to detect the presence of the positive events. The black curve on the main contour plot in Figure 4 illustrates the boundary between indistinguishable and distinguishable distributions. For example, if the difference between a positive and negative population is only 0.25 standard deviations, then the algorithm requires 1000 positive events (irrespective of the number of negative events!) to generate a statistically significant value.

Application of the PB Metric to Immunofluorescence Data

We next applied this algorithm to immunofluorescence and light scatter data collected on by flow cytometry. While two immunofluorescence (or light scatter) distributions may be statistically significantly different by algorithms (including the PB comparison), we wished to determine if the PB comparison could still be used to rank distributions. The goal is to determine the minimum value of $T(\chi)$ that has *biological* significance. Certainly, this minimum value would be different depending on the nature of the data being analyzed and needs to be determined empirically.

Our test data was derived from a three-color immunofluorescence analysis of PBMC from 18 individuals (7 HIV^+, 11 HIV^-), four collected on one day, and 14 on another day. A panel of six different three-color stains was collected on each individual; data from 30,000 PBMC was stored. For the analyses in Figures 5 and 6, data was first gated on forward and side scatter for lymphocytes; only the data for lymphocytes was included in the PB comparison. In this case, a single tube from an HIV^- sample served as the control tube for PB comparison.

Figure 5 illustrates that even gated Side Scatter distributions (which have very low variance) still yield information when compared by PB. The six side scatter distributions derived from the same individual are much more closely related than the distributions from other individuals. Interestingly, four of the five HIV^+ individuals had the greatest difference from the HIV^- control.

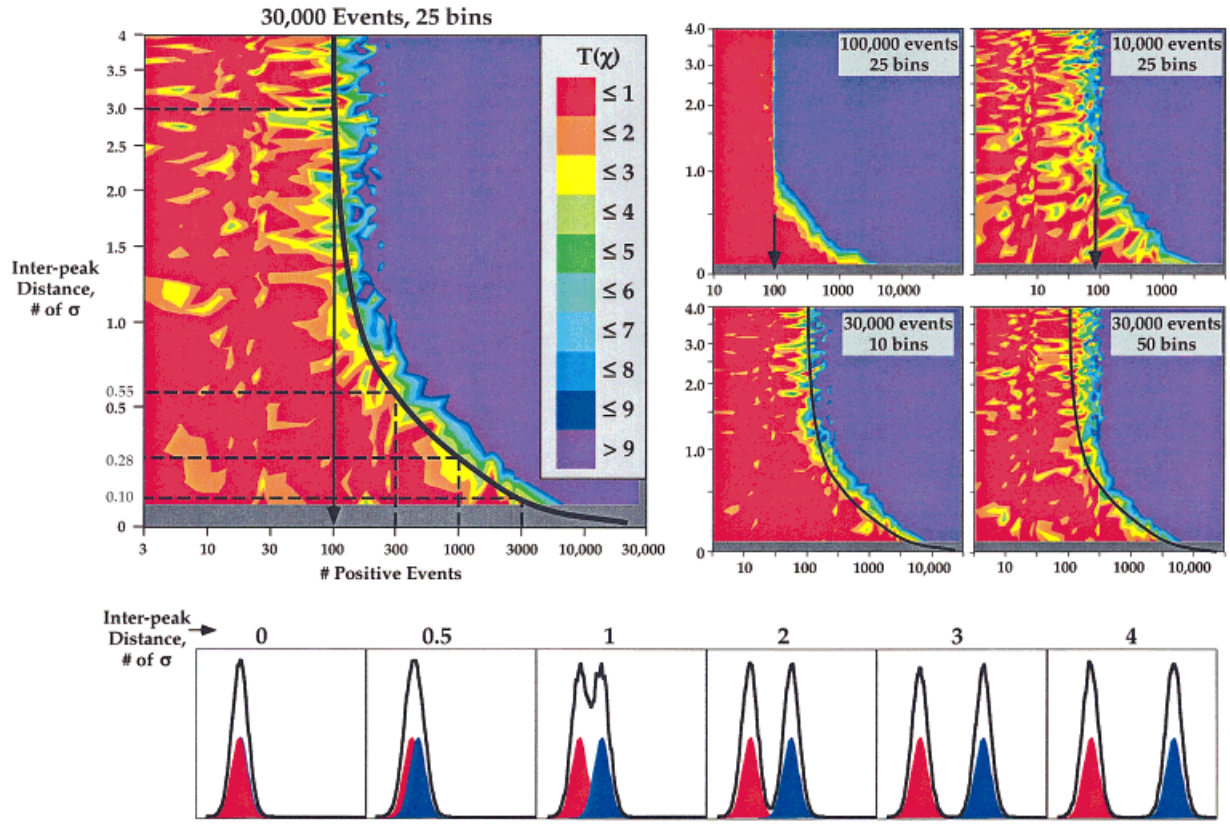

FIG. 4. Detailed analysis of the dependence of $T(\chi)$ on the separation and proportion of a positive population. For each contour plot, 1,800 data files were generated and compared to the control distribution. Each data file had a different proportion (0.01 to 90%) of a positive population that was centered above the control population from between 0.1 and 4.0 standard deviations of the control distribution. The histograms across the bottom show examples of the control distribution, followed by the distributions (black) containing equal mixtures (50%) of control (red) and positive (blue) events separated by different distances. The 1,800 $T(\chi)$ values were contoured and colored according to the magnitude of $T(\chi)$ (first panel, inset). Since a $T(\chi)$ of greater than four is associated with a >99% probability that the distribution is different than the control, this value is taken as the threshold for statistical significance. The curved black line in the large contour plot is drawn approximately along this threshold; thus, distributions to the left of this line are not statistically significantly different from the control distribution; to the right, they are. The analysis was repeated for four other combinations of event counts and binning; the curved black lines in the smaller graphs are position at the *same* location as for the 30,000 event/25 bin analysis for comparison. The vertical arrows point to the minimum detectable positive event count for distributions that are non-overlapping: note that this value is nearly 100 positive events, irrespective of the number of events in the total population.

Figure 5 also illustrates the dependence of the metric on the representation of a well-resolved population (in this case, CD8 expression on lymphocytes). As expected from Figure 3, $T(\chi)$ increases as the percentage of CD8 T cells among lymphocytes moves away (either higher or lower) from the control sample's percentage.

Perhaps the most important reason for using a statistic to compare univariate distributions is to determine whether or not highly overlapping distributions are different. Giorgi and colleagues (11) have shown that the expression of CD38 on CD8 T cells is a powerful predictor of subsequent progression of HIV disease. However, CD38 expression is continuous from negative to positive, and often only a fraction of the cells express it. Typically, the extent of CD38 expression is quantified by the mean (or median) CD38 fluorescence intensity of the cells. Because of the low and variable expression of CD38, we used it as a test for the PB comparison.

Figure 6 illustrates that the $T(\chi)$ values for these comparisons can distinguish those subjects with a significant expression of CD38. Indeed, for these comparisons, the value of $T(\chi)$ is proportional to the median CD38 expression, confirming the analyses of Figures 2-4 that the $T(\chi)$ metric scales with the degree of difference of univariate distributions. In addition, Figure 6 illustrates that selecting a minimum value of 4 for $T(\chi)$ (i.e., 4 standard deviations above noise, or a p value > 99%) is a reasonable criterion for asserting that two distributions are indeed different.

Using the PB Metric to Quantify Representation of an Unresolved Population

The smooth and monotonic dependence of $T(\chi)$ on the fraction of a contaminating population (Fig. 3) suggests that $T(\chi)$ could be used not just to quantitate distribution differences but to estimate the fraction of a poorly-resolved contaminating population. (Fig. 5 shows that it can be used to estimate the fraction of a well-resolved population, CD8 T cells, based on the univariate CD8 histogram). To test this hypothesis, we compared the frequency of monocytes within PBMC (a known value based

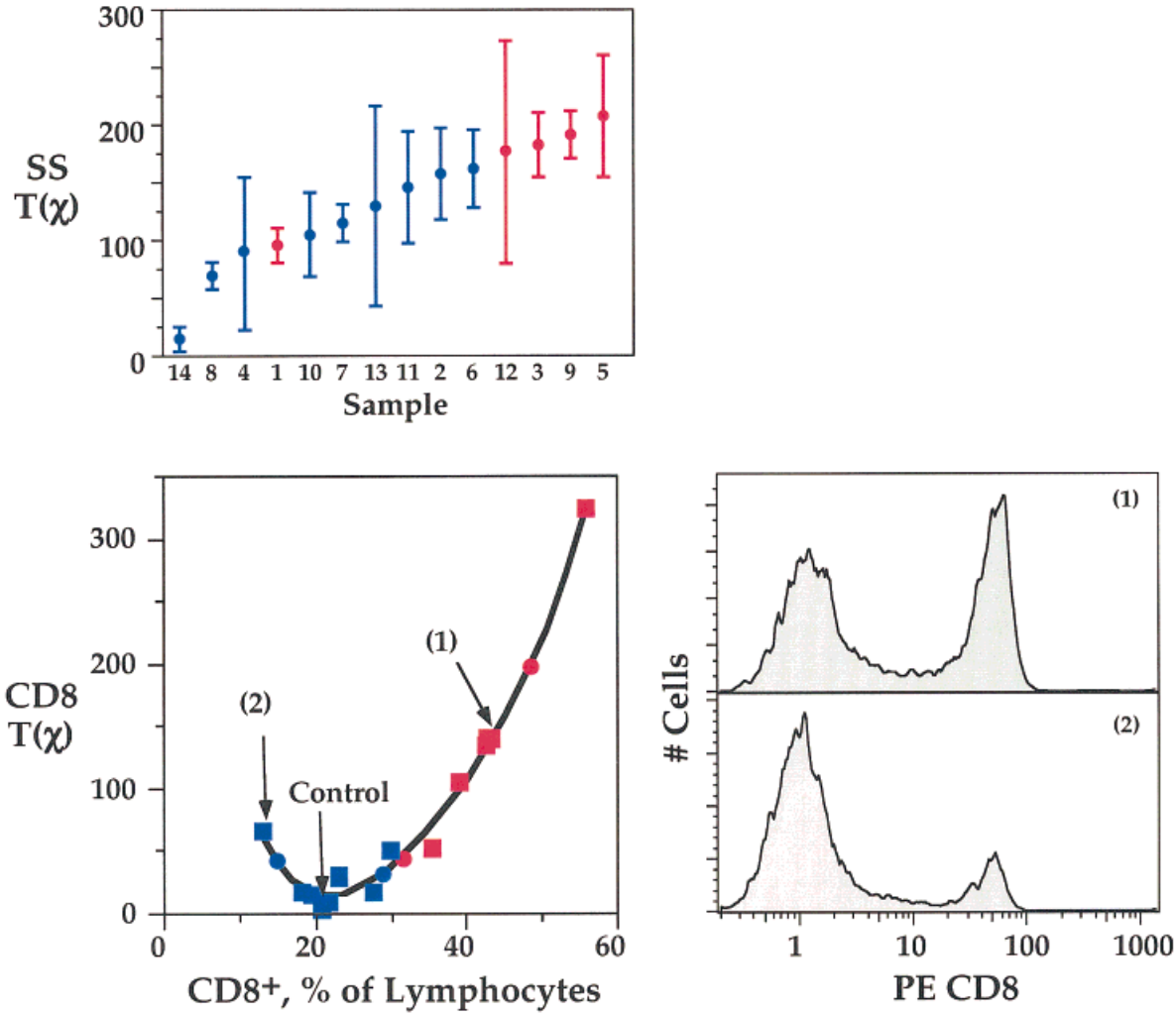

FIG. 5. Evaluation of $T(\chi)$ in clinical specimens. (Top) PBMC from 14 individuals (13 are HIV^+) were stained with six different panels of antibodies and analyzed on the same day. Lymphocytes were gated by forward and side scatter (SS); the side scatter distribution of the lymphocytes was compared to the HIV^- control (#14). Because the distributions were already gated for side scatter, there is limitation on the variation that could be encountered. For each of the six samples from each patient, the $T(\chi)$ distribution for SS is shown (mean $\pm$ 1σ).

on CD14 staining) with the $T(\chi)$ obtained by comparing the Forward Scatter distributions for total PBMC with a Forward Scatter distribution for a pure lymphocyte population (Fig. 7). These analyses demonstrate that the $T(\chi)$ metric can indeed be used to quantitate a contaminating population, even when the distribution of that population overlaps significantly with the control (although a three to five point calibration curve will be necessary to effect such quantitation accurately).

For distributions where the contaminating population in a test distribution was to the right (greater fluorescence) than the test sample, the PB metric, K-S statistic, SED value, and Overton methods performed equally well (Fig. 8). In fact, the minimum detectable difference in these distributions occurred at the same threshold of contamination. As shown here, all of these statistics have a threshold for significance that is based on the absolute number of contaminating cells, rather than a percentage of the events. This is contrary to what was previously published for K-S (12); the difference may be due to the much larger sampling of distributions performed in our analysis compared to the previously published analysis (several thousand vs. five).

DISCUSSION

Many have pointed out the caveats to performing statistics to compare univariate binned data such as that from flow cytometry. Nonetheless, there is a great need for such comparisons, for both quality control during (or after) sample acquisition, and for identifying outlying samples based on the measurement of a single response.

A few methods have been applied to this problem. The most successful for quantitating responsiveness is the Overton cumulative histogram subtraction. This algorithm predicts % positive quite well, given sufficient positive events and/or sufficient separation of the positive and

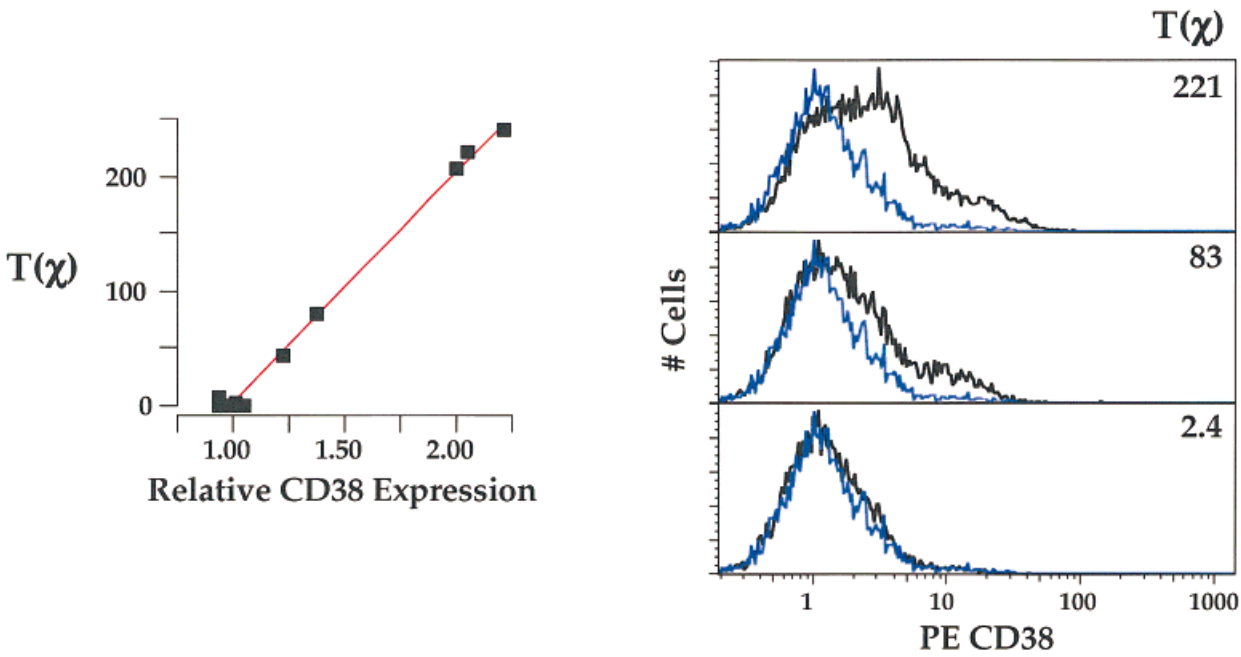

FIG. 6. Discrimination of CD38 expression profiles on CD8 T cells. PBMC from 14 individuals were stained to evaluate CD38 expression on T cells. (Right) Histograms for three HIV^+ individuals (black) are compared to the HIV^- (blue) control. The $T(\chi)$ value for each comparison is shown in the corner. The K-S statistic was also calculated on these distributions; it gave a statistical significance of >99.9% for all three, whereas the $T(\chi)$ of 2.4 is not considered significant for the last histogram. (Left) The $T(\chi)$ value for the comparison is plotted against median CD38 expression. Note that the statistic scales linearly with increasing median intensity of the sample distributions.

negative. The PB algorithm performed no better than the Overton method at enumerating positive events above a control distribution (Fig. 7). However, the Overton method does not provide an indication of the probability with which two distributions are different (i.e., a probability to assign the percent positive calculation); nor does it provide confidence intervals.

Lampariello developed a parametric statistic for determining percent positive cells based on a model of the distribution of cellular autofluorescence (7, 8). This statistic is expected to perform extremely well (is highly sensitive to small proportions of negative populations) under the conditions for which it was designed—use of unstained control samples. However, it will fail should the control distribution not fit the model of autofluorescence distribution–for example, when comparing distributions where the control sample has stained cells. In addition, the model depends on accurate autofluorescence modeling, which is not possible when significant fluorescence compensation is necessary because of the significant broadening of the distribution (13).

The K-S statistic can sensitively detect statistically significant differences between distributions. However, the p value on the K-S statistic is not a metric; it has not been demonstrated that smaller p values correspond to distributions that are less like the control. Additionally, it has been demonstrated that K-S *underestimates* the difference of discrete (binned) distributions such those from flow cytometry.

The PB comparison is as sensitive as the K-S to detecting statistically significant differences (Fig. 8). However, the PB comparison provides a metric, $T(\chi)$, which *can* be used to quantitate differences between distributions. Therefore, for any given data set (or type of data), a biologically-meaningful minimum $T(\chi)$ can be empirically determined. Only histograms which have $T(\chi)$ values larger than this empirical minimum can then be considered to be different.

While the PB comparison works well with univariate distributions, the original impetus for its development was the comparison of multivariate distributions. No other statistical test has been applied to two- or more parameter flow cytometric distributions. As described in Roederer et al. (9), the PB comparison statistics as derived here work well to identify differences in multiparametric distributions. Thus, the PB comparison can serve as a general solution for the comparison of flow cytometric data, either as quality control feedback or to measure biological differences, irrespective of the number of parameters that may vary.

One difficulty in performing these types of analyses (especially, population distance metrics) is identifying a suitable control population. One approach to this problem is to combine all samples to be compared into a single concatenated control population, or, alternatively, to specify a set of control populations to be combined. Probability binning is then performed on this combined dataset. Each individual sample is then compared against

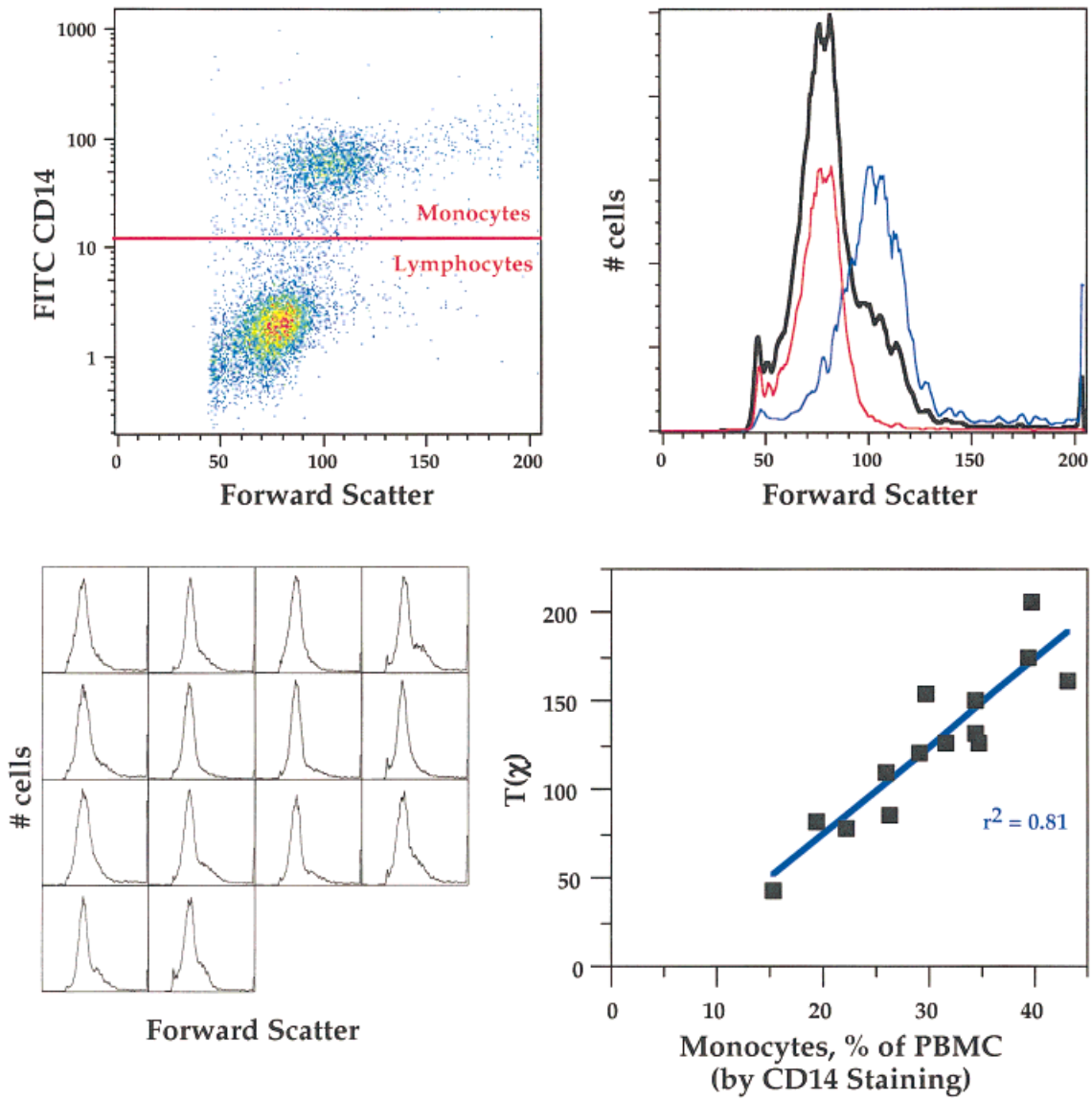

FIG. 7. PB Statistics accurately estimate contaminating population frequencies. (Top Left): Forward scatter vs. CD14 for PBMC. Monocytes are defined as $CD14^+$; Lymphocytes as $CD14^-$. (Top Right): Histograms of Forward Scatter for total PBMC (black), $CD14^-$ Lymphocytes (red), and $CD14^+$ Monocytes (blue). The red and blue histograms were scaled to equal heights; in this sample, the Monocytes comprise 30% of the PBMC. (Bottom left): The Forward Scatter histograms for 14 samples collected on the same day. The overlap between the Monocytes and Lymphocytes precludes accurate estimation of Monocyte representation based solely by gating on Forward Scatter. (Bottom right): The PBMC Forward Scatter distribution was compared to a single $CD14^-$ pure lymphocyte population from one sample. The resulting $T(\chi)$ values are plotted against the known Monocyte representation (based on CD14 staining). The highly-correlated linear relationship demonstrates that the $T(\chi)$ can be used to quantitate the representation of Monocytes among PBMC based solely on the Forward Scatter histograms and a representative Forward Scatter distribution of pure Lymphocytes. By extension, $T(\chi)$ could be used to estimate the representation of highly overlapping contaminating cells for any parameter, given a pure control histogram of that parameter.

this combined dataset in order to provide a metric that is the distance from the average of the populations. This method can alleviate the problem of identifying a suitable control population. The utility of such averaging will be greatest in comparing real-life samples, wherein sample-to-sample variation amongst the controls could lead to an over-estimation of differences should an atypical sample be used for a control.

Finally, one of the powers of Probability Binning is that the algorithm can be used to identify which parts of a (multivariate) distribution are different from a control sample: i.e., it can generate a gate comprising of the events which are different in the sample compared to a control. This is explored in detail in Roederer et al. (14). As such, it is relevant to note that Overton's (6), Bagwell's (2), and Lampariello's (8, 12) statistics are all tuned to identifying contaminating populations that have a greater fluorescence distribution than the control (i.e., positives). The PB algorithm can identify contaminating populations no matter where the occur rel-

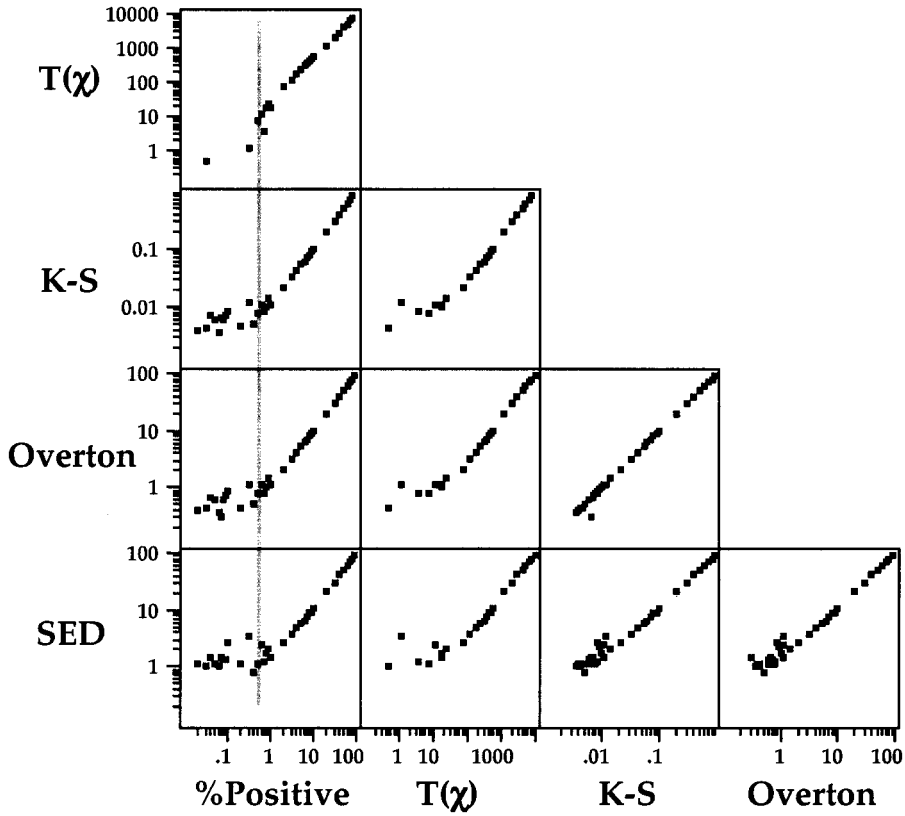

FIG. 8. Comparison of various univariate analysis methods. Distributions selected from Figures 3 and 4 were used to compare the statistical methods PB (described here), Kolmogorov-Smirnoff D-value (4), Overton's modified histogram subtraction (6), and Bagwell's SED method (2). For this purpose, we compared distributions that had varying fractions of positive cells (% positive) out of 30,000 events to a distribution that had no positive cells. The positive cells were separated by 2σ from the negatives (i.e., well-separated). The methods also performed nearly identically for distributions that were highly overlapping distributions separated by only 0.06σ (data not shown). Each data point shows the value for one test distribution comparison. The vertical light gray line indicates the minimum fraction of cells that is accurately identifiable by the methods. The highly close agreement of these methods is expected for these test distributions, and indicates that all provide roughly the same information. However, the SED and Overton methods are specifically designed to identify positive cells, i.e., cells to the right of control distribution, and do not accurately identify contaminating populations if their distribution is not to the right of the control. The PB method can identify a contaminating population no matter where in the distribution it occurs: to the left, within, or to the right.

ative to the control distribution– even *within* a complex distribution.

In summary, Probability Binning is a novel statistic for comparing distributions of event data. It has several advantages over existing algorithms, in that it provides a reasonable probability (and confidence interval) of two distributions being different, it can be used to determine the percent contamination by an second population of events, and, unlike any other metric, it can be used to rank differences between test samples to identify those most or least like the control sample. Most importantly, however, the Probability Binning algorithm can be applied to data comprising any number of distinct parameters.

ACKNOWLEDGMENTS

We thank Dr Robert Murphy for critical reading of the manuscript, Dr. Bruce Bagwell for sharing the SED algorithm, and Dr. David Parks, Dr. Stephen De Rosa, and Steve Perfetto for useful discussions and ideas.

LITERATURE CITED

1. Bagwell CB, Hudson JL, Irvin GL, III. Nonparametric flow cytometry analysis. J Histochem Cytochem 1979;27:293–296.
2. Bagwell C. A journey through flow cytometric immunofluorescence analyses. Clin Immunol Newsletter 1996;16:33–37.
3. Finch PD. Substantive difference and the analysis of histograms from very large samples (letter). J Histochem Cytochem 1979;27:800.
4. Young IT. Proof without prejudice: use of the Kolmogorov-Smirnov test for the analysis of histograms from flow systems and other sources. J Histochem Cytochem 1977;25:935–941.
5. Cox C, Reeder JE, Robinson RD, Suppes SB, Wheeless LL. Comparison of frequency distributions in flow cytometry. Cytometry 1988;9:291–298.
6. Overton WR. Modified histogram subtraction technique for analysis of flow cytometry data [published erratum appears in Cytometry 1989;10(4):492–494]. Cytometry 1988;9:619–626.
7. Lampariello F. Evaluation of the number of positive cells from flow cytometric immunoassays by mathematical modeling of cellular autofluorescence. Cytometry 1994; 15:294–301.
8. Lampariello F, Aiello A. Complete mathematical modeling method for the analysis of immunofluorescence distributions composed of negative and weakly positive cells. Cytometry 1998;32:241–254.
9. Roederer M, Moore W, Treister AS, Hardy RR, Herzenberg LA. Probability Binning Comparison: a metric for quantitating multivariate distribution differences. Cytometry 2001;45:47–55.
10. Leva JL. A fast normal random number generator. Transactions in mathematical software 1992;18:449–453.
11. Giorgi JV, Liu Z, Hultin LE, Cumberland WG, Hennessey K, Detels R. Elevated levels of CD38+ CD8+ T cells in HIV infection add to the prognostic value of low CD4+ T cell levels: results of 6 years of follow-up. The Los Angeles Center, multicenter AIDS cohort study. J Acquir Immune Defic Syndr 1993;6:904–912.
12. Lampariello F. On the use of the Kolmogorov-Smirnov statistical test for immunofluorescence histogram comparison. Cytometry 2000;39:179–188.
13. Roederer M. Spectral compensation for flow cytometry: visualization artifacts, limitations, and caveats. Cytometry 2001; in press.
14. Roederer M, Hardy RR. Frequency difference gating: a multivariate method for identifying subsets that differ between samples. Cytometry 2001;45:56–64.