Published 2001 Wiley-Liss, Inc.† Cytometry 45:47–55 (2001)

Probability Binning Comparison: A Metric for Quantitating Multivariate Distribution Differences

Mario Roederer,[1]* Wayne Moore,[2] Adam Treister,[3] Richard R. Hardy,[4] and Leonore A. Herzenberg[2]

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

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

Background: While several algorithms for the comparison of univariate distributions arising from flow cytometric analyses have been developed and studied for many years, algorithms for comparing multivariate distributions remain elusive. Such algorithms could be useful for comparing differences between samples based on several independent measurements, rather than differences based on any single measurement. It is conceivable that distributions could be completely distinct in multivariate space, but unresolvable in any combination of univariate histograms. Multivariate comparisons could also be useful for providing feedback about instrument stability, when only subtle changes in measurements are occurring.

Methods: We apply a variant of Probability Binning, described in the accompanying article, to multidimensional data. In this approach, hyper-rectangles of *n* dimensions (where *n* is the number of measurements being compared) comprise the bins used for the chi-squared statistic. These hyper-dimensional bins are constructed such that the control sample has the same number of events in each bin; the bins are then applied to the test samples for chi-squared calculations.

Results: Using a Monte-Carlo simulation, we determined the distribution of chi-squared values obtained by comparing sets of events from the same distribution; this distribution of chi-squared values was identical as for the univariate algorithm. Hence, the same formulae can be used to construct a metric, analogous to a t-score, that estimates the probability with which distributions are distinct. As for univariate comparisons, this metric scales with the difference between two distributions, and can be used to rank samples according to similarity to a control. We apply the algorithm to multivariate immunophenotyping data, and demonstrate that it can be used to discriminate distinct samples and to rank samples according to a biologically-meaningful difference.

Conclusion: Probability binning, as shown here, provides a useful metric for determining the probability with which two or more multivariate distributions represent distinct sets of data. The metric can be used to identify the similarity or dissimilarity of samples. Finally, as demonstrated in the accompanying paper, the algorithm can be used to gate on events in one sample that are different from a control sample, even if those events cannot be distinguished on the basis of any combination of univariate or bivariate displays. Cytometry 45:47–55, 2001.
Published 2001 Wiley-Liss, Inc.†

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

A significant power of flow cytometric analysis is the collection of highly complex, multivariate data for each of thousands or millions of events for any given sample. However, the utilization of the information present in such data is typically limited by analysis tools that can operate on one or two dimensions at a time. For example, currently algorithms to compare distributions (such as Kolmogorov-Smirnoff (K-S) statistics (1, 2), Overton subtraction (3), SED (4), and even parametric models (5, 6)) are limited to univariate data. The utility of these types of algorithms is in their ability to determine whether or not given distributions are (statistically significantly) different, and/or to determine the fraction of events that are positive compared to a control distribution.

The ability to compare multivariate distributions could be important for a number of applications.

A preliminary description of this algorithm was presented in a poster at the 1994 International Society for Analytical Cytology meeting.

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

(1) Identification of biological response. It is possible that biological response to stimulation or other interaction might be identified only based on the combined measurement of several parameters simultaneously. Genetic programs that modulate gene expression often affect multiple genes in concert; however, the change in any given gene expression may not be unique to a particular experimental condition. Thus, the simultaneous evaluation of multiple response variables may be necessary for precise identification of an interaction.

(2) Identification of outlier events. Algorithms comparing univariate distributions can be used to quantitate the percent of events above a control sample (e.g., Overton histogram subtraction, Bagwell's SED algorithm (4), or Lampariello's parametric models (5, 6)). However, it is conceivable that outlying events cannot be distinguished solely on the basis of a single parameter, nor may they occur with fluorescences greater than the control; thus, these could not be accurately enumerated by these univariate approaches.

(3) Quality control feedback. Multivariate comparisons of data collected over time could identify subtle changes in the distribution of what should be identical distributions. For example, many cytometer operators recognize the importance of monitoring forward and side-scatter distributions while collecting cell samples, and implicitly believe that monitoring the bivariate display of forward and side-scattered light signals is much more informative than monitoring two univariate displays of each signal separately.

In addition, as the dimensionality of flow cytometric data increases, the demand for multivariate algorithms becomes more acute. While it is often sufficient to use univariate tools on 3- or 4-parameter data (because of the relative independence of the measurements), it is almost never sufficient when the number of parameters is greater than five or six. Because the complexity of the data (and the analysis) increases geometrically with the number of parameters, the difficulties become significant hurdles to data analysis as we migrate to routine six or more color analysis (i.e., 8+ parameter data).

In an accompanying paper (7), we describe a variation of a chi-squared statistic, termed Probability Binning (PB), that can be used to rank univariate distributions in a statistically meaningful way. In this manuscript, we describe how to extend this algorithm to multivariate data. We show that the PB algorithm behaves predictably to quantitate differences between highly artificial datasets. We then apply the PB metric to three- or four-color immunofluorescence data to demonstrate that the PB metric can be used to objectively and quantitatively rank multivariate distributions in a way that is biologically meaningful.

MATERIALS AND METHODS

Data Analysis

Artificial multivariate distributions were created as FCS files using a specially modified version of FlowJo. Distribution comparisons were performed using FlowJo version 3.3 (Tree Star, San Carlos, CA); additional analysis was performed using JMP for Macintosh (SAS Institute).

Cell Staining and Flow Cytometric Analyses

Human PBMC and mouse lymphocytes 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

Multivariate Probability Binning Algorithm

In order to carry out Probability Binning Comparison, multivariate data must first be divided into bins containing the same number of events. Thus, when binning 10,000 events into 100 bins, each bin must contain 100 events. This necessitates that the bins are of different sizes. We chose an algorithm that successively divides a multivariate dataset into bins such that each bin has the same number of events (Fig. 1). The algorithm begins by calculating the median and variance of all of the data, for each of the parameters included in the comparison. It chooses the parameter with the largest variance, and divides the events in half based on the median value of that parameter. By choosing the parameter with the largest variance, the algorithm is weighted towards assigning distinct clusters of events into distinct bins (or sets of bins).

The algorithm then repeats the process on each of the two newly-defined subsets, again determining the median and variance of all parameters for each subset. This two-fold division process continues until some specific threshold is met (see below). The result is a series of n-dimensional hyper-rectangular bins. When the original control sample is separated into these bins, each bin has roughly the same number of events. Therefore, when selecting an event at random from the control population, there is an equal probability that it will fall into any given bin.

The bins defined by the control population are then applied to a comparison sample. The number of events falling within each bin are determined, and the normalized chi-squared value (χ'^2) is calculated exactly as for the one-dimensional PBC, namely:

$$\chi'^2 = \sum_{i=1}^{\#bins} \frac{(c_i^n - s_i^n)^2}{(c_i^n + s_i^n)}$$

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

given that c_i and s_i are the number of control and test sample events falling into bin i, and E^c and E^s are the total number of events in the control and test samples. Theo-

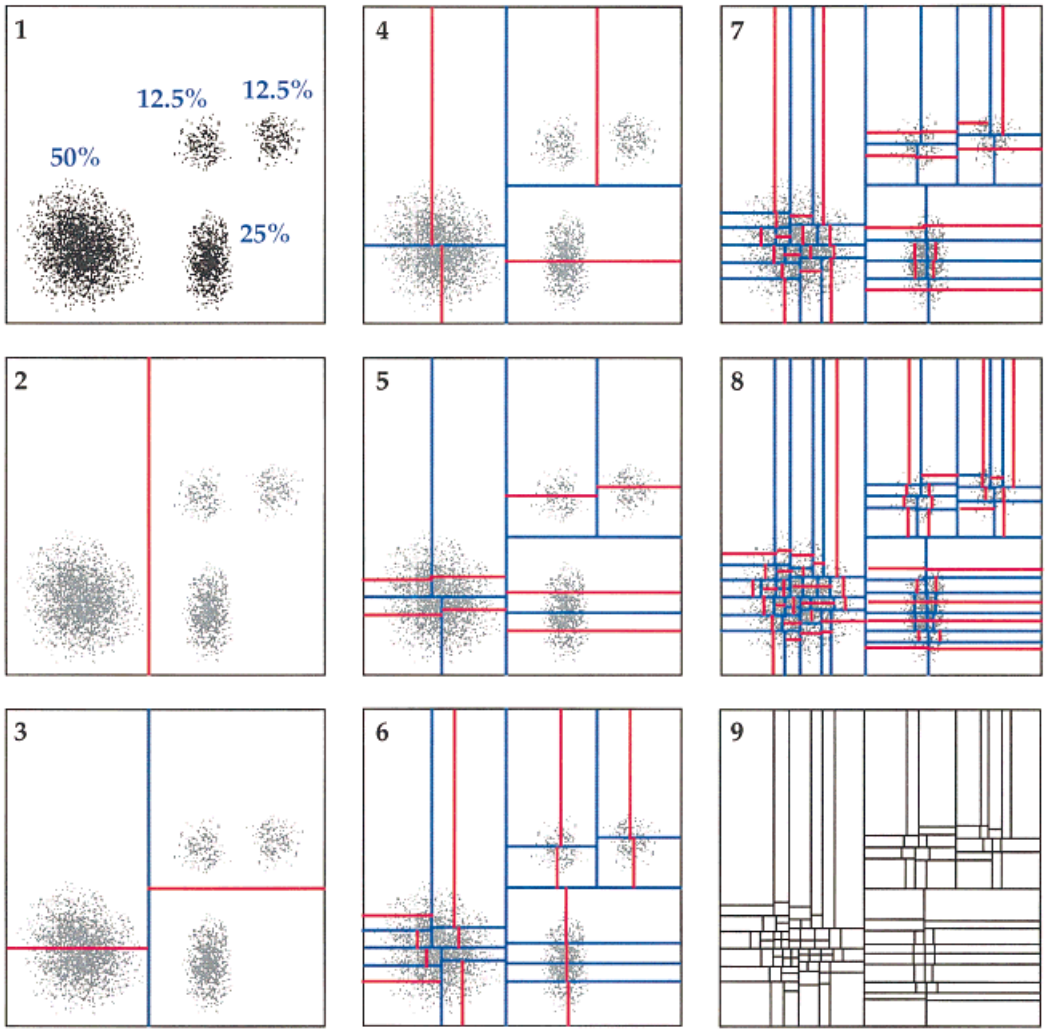

FIG. 1. Probability Binning. Shown is an example of two-dimensional probability binning on an artificial bivariate distribution. **(1)** In this distribution, there are four populations, comprising 50%, 25%, 12.5%, and 12.5% of the total, as shown in the top left. Each numbered graphic is a subsequent binning step. **(2)** The first step in binning is to divide the distribution in half along the dimension with the greatest variance (in this case, the X-axis). The red line shows this division along the median of the X parameter. **(3)** For each of the two subsets of cells (to the left and to the right of the blue line), the data is further divided in half (red lines). For each subset, the greatest variance is in the Y parameter. **(4)** For each of the four resulting subsets, the variance in both parameters is determined. For three subsets, the greatest variance is in the X axis, thus they are divided according to their respective X median values (red lines). For the lower right-hand subset, the greatest variance is in the Y axis; hence it is divided at the median for the Y values for that subset. **(5–8)** The variance and median calculations for both parameters are measured for each subset, and the data is further subdivided (for each step, previous divisions are shown in blue and the newest divisions in red). **(9)** After the division shown in (8), there are a total of 128 bins, each containing approximately $1/128^{th}$ of the total events in the control sample. Note that the bins are smallest and cluster around high event densities, and are large where events are more rare–thereby minimizing the maximum expected variance for a chi-square comparison on events in the bins. Probability binning in more than two dimensions proceeds in an analogous fashion, except that at each step, for each subset, variances and medians for all parameters are calculated to select the parameter on which that subset is divided; the resulting bins are n-dimensional hyper-rectangles.

retically, χ'^2 can range in values from a minimum of zero to a maximum of two.

Derivation of the PB metric

As demonstrated in the accompanying paper (7), a metric based on χ'^2 can be derived empirically. This metric is analogous to a t-score, i.e., a value of zero indicates no statistical significance; a value of one indicates that the χ'^2 is one standard deviation above the minimum significant value. Note that while the definition of this metric was derived based on comparison of univariate data, it holds just as well for multivariate data since fundamentally this is a comparison of bin counts, irrespective of how those bins are defined.

In order to determine the distribution of χ'^2, we generated several thousand FCS data files of artificial data. Each data file contained from 30,000 to 300,000 events, with three parameters, with randomly generated values. The values of each event for each parameter were generated using a pseudo-random number generator that creates normal (Gaussian) distributions (8). Each set of data files containing the same number of events were sub-

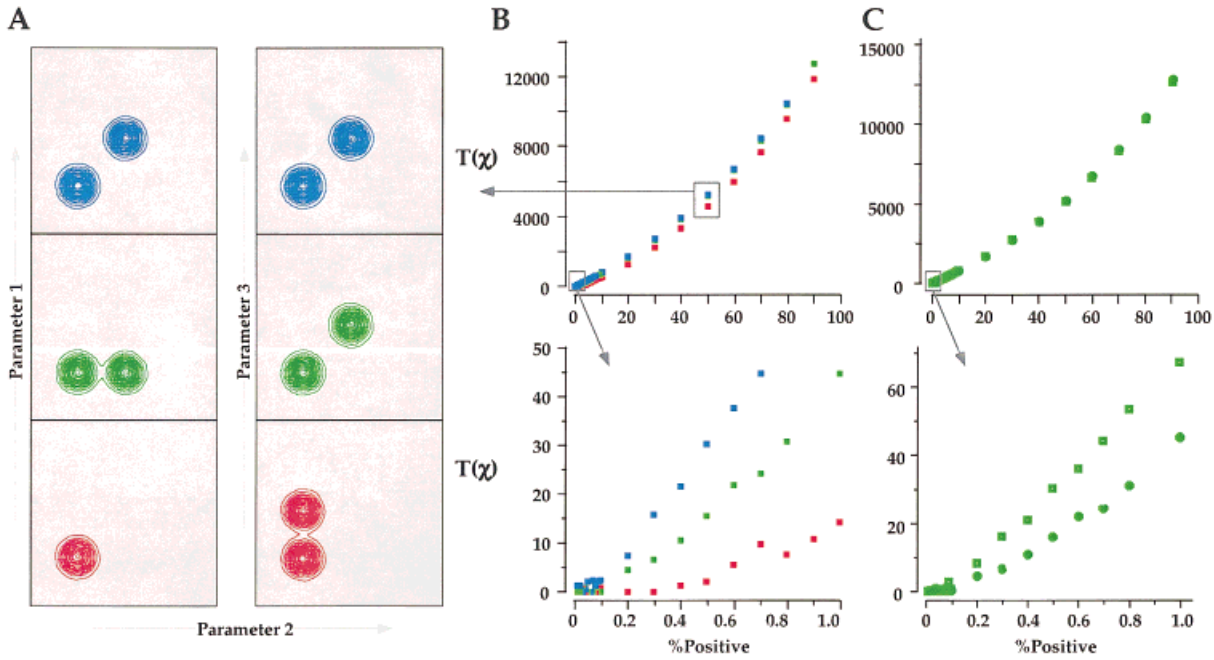

FIG. 2. Effect of the number of parameters included in the PB Comparison. Artificial distributions were constructed in which a varying percentage of well-resolved positive events were added to a negative control In this simulation, there were three fluorescence parameters for each event. The positive population was different than the negative for either one (red), two (green) or all three (blue) parameters. For parameters which were not different, the distributions for the positive and negative population were identical. **(A)** Bivariate plots of three of the three-parameter data files used in the test (the data points for these comparisons are boxed in the upper panel of **(B)**). In these three examples, the two equally-represented populations differ on the basis of only parameter three (red), both parameters two and three (green), or all three parameters (blue). The full comparisons shown in the graphs were performed on similar data sets, comprised of varying ratios of the two populations compared to a dataset comprised of only one of the two populations. (B) Each point represents the PB comparison performed on a different dataset. The PB metric could distinguish a smaller percentage of positive events when the positive events were different on the basis of more parameters. Using a cut-off of $T(\chi) = 4$, the threshold for significance was achieved at 0.6% (1/3 parameters had distinct distributions for positive and negative), 0.2% (2/3 parameters were distinct), and 0.15% (all three parameters were distinct). **(C)** In these comparisons, the positive and negative events were well-resolved in two parameters, but had identical distributions for the third parameter. The PB Comparison was performed using either all three parameters (circles) or only the two parameters that differed (squares). There is a slight decrease in the sensitivity of the PB metric to identify small populations when a parameter that is invariant is included.

jected to PB using from 300 to 5,000 bins. The resulting χ'^2 distributions were normally-distributed, and depended on the number of bins and number of events exactly as in the univariate comparison (see accompanying paper (7), Fig. 2; data not shown). Thus, we define the minimum significant normalized chi-squared value $\bar{\chi}'^2$, and the associated standard deviation for this minimum chi-square $\sigma_{\chi'^2}$ based on the number of bins B used in the comparison, and the event count E (where E is the minimum of the number of events in the control sample (E^c) or the test sample (E^s)):

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

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

As for the univariate comparison, we define the metric $T(\chi)$ as

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

The only parameter to this algorithm is the number of bins into which the control sample is to be divided. It is apparent from the above equations (since χ'^2 has a maximum value of two) that the number of bins should be kept low enough that the minimum significant value $\bar{\chi}'^2$ doesn't become so large as to preclude assigning statistical significance to any distribution. On the other hand, the number of bins should be maximized so as to most easily detect small changes in a distribution (i.e., if the entire change in a distribution were to occur within a single bin, then the statistic would not change). Thus, the effectiveness of the metric for detecting subtle changes (in terms of fluorescence intensity) may be limited by using small numbers of bins.

Unlike the case for univariate comparisons, the number of bins can quickly become limiting for this statistic (depending on the number of events collected and the number of parameters compared). The maximum reasonable number of bins is roughly 10% of the event count—leading to a minimum of about 10 events per bin. (When using more bins, the number of events per bin is smaller, and the variance associated with each bin increases dramatically). Thus, for a 30,000 event collection, the maxi-

mum number of bins is 3,000. This is obviously far too many for a univariate comparison (where the limitation is the number of channels in the histogram that contain events, typically well under 1024). However, consider the case in which a comparison of all five parameters of a 3-color sample (including the two scatter parameters) is performed. Three-thousand bins in five-dimensional space means that each parameter has been divided into approximately five divisions (on average)–i.e., $5^5 \approx 3{,}000$.

The ability of the statistic to detect differences in distributions is limited by the dimensions of the bin. Clearly, with bins that span an average of $1/5^{th}$ of the range of each parameter (although the bins are much smaller in areas with many events), the metric will be relatively insensitive to subtle changes in distribution.

However, this limitation can be overcome by reducing the number of parameters in the comparison—at a cost of losing the information provided by those parameters. In the above example, using only four parameters, each would be divided on average 7.5 times; comparing 3 parameters, each would be divided over 14 times ($14^3 \approx 3{,}000$).

Clearly, if all five parameters are needed in the comparison, then more events *may* need to be collected in order to distinguish subtle variations in the distributions. It may be necessary to use a strategy of performing comparisons using different subsets of parameters in order to identify those, which provide the least discriminating power. The final comparison would be performed using as many useful parameters as possible given the limitations on the number of events collected.

Note that the algorithm will divide parameters with greater variance (more information) more frequently than parameters with low variance. Thus, in the example above, while the average number of divisions for a parameter may only be five, it is possible that (if the cells were lymphocytes) the two scatter channels may only be divided two or three times, leaving several additional divisions for the fluorescence parameters. Some knowledge of the expected outcome can provide an indication to the user an appropriate number of events that need to be collected in order to achieve the desired level of detection.

Validation of the PB Metric for Multivariate Data

In order to validate the metric $T(\chi)$, we generated another large series of data files. In these files, we added a varying number of positive events to the negative distribution, where positive events were also distributed normally in all three parameters, but with a mean that ranged anywhere from 0.1 to 2.0 standard deviations above the negative events. In addition, the positive events differed either in only one, in only two, or in all three of the parameters. In the first case, the events for the other two parameters were distributed identically to the negative events.

We performed a thorough analysis of the dependence of $T(\chi)$ on the total number of events, the representation of positive events, the number of parameters which are different for the positive and negative populations, and the distance between positive and negative peaks. As for the univariate comparison (accompanying paper (7), Fig. 3), $T(\chi)$ depends monotonically and smoothly with both the fraction of positive events as well as the separation (Fig. 2 and data not shown). Thus, $T(\chi)$ is a statistic which not only provides an indication of the probability with which two distributions are different, but also provides a metric by which multiple distributions can be ranked.

Figure 2 demonstrates an additional feature of the PB metric applied to multivariate data, that the metric is more successful at identifying outlier events when those events differ on the basis of more than one parameter (for a given number of bins). This is expected, in that populations that differ on the basis of more parameters provide a greater amount of information with which they can be discriminated.

In Figure 3, these aspects of the multivariate PB algorithm are compared in greater detail. Here, several thousand comparisons were performed to generate each graphic. Each comparison used different distributions as described above. The contour graphs provide a visual estimate of the minimum separation and/or representation of positive events within a distribution in order for the PB metric to yield a statistically significant value. Similar to the case for univariate comparisons, the PB metric can resolve a relatively small number of well-separated cells (100–300) from a larger population, independent of the size of the larger population. In other words, the more events that are collected, the smaller the fraction of a contaminating subset can be detected.

Figure 3 confirms the more limited analyses shown in Figure 2: that the more parameters which distinguish a positive distribution, the smaller the number of events necessary for identification. In addition, inclusion of parameters that do not distinguish the subsets into the algorithm does not substantially affect the ability to resolve the subsets.

Multisample Comparison: Uniting the Control Dataset

In comparing multiple samples against each other, it is sometimes not possible (or meaningful) to assign a single sample as the control sample, against which all others are to be compared. In such a case, we construct the bins (along the algorithm exemplified in Fig. 1) on the concatenation of all test samples. Each sample is then measured against the combined dataset; thus, each sample is assigned a value that is the distance from the average of the samples. This process mitigates the potential artefact introduced by selection of a sample as a control that is actually significantly different than the expected control sample. It may also be useful to designate a set of samples to be used as the control for binning purposes, rather than including potentially distant outliers. In the end, the best approach is probably iterative: include all samples in the original binning and compute distances. After this, those samples which are most distinct can be removed from the set used as a control, and the statistic is recomputed. This

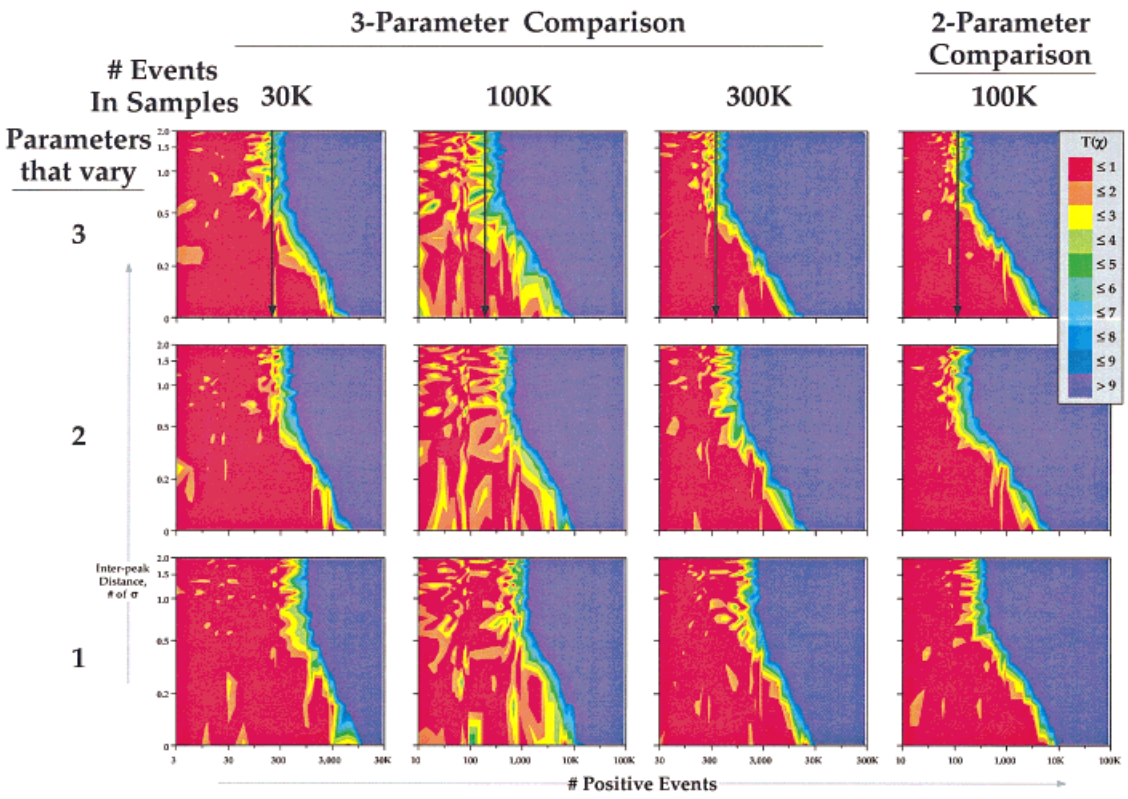

FIG. 3. Detailed analysis of the dependence of $T(\chi)$ on the fraction of positive events, the separation between positive and negative events, the number of events in the sample sets, the number of parameters that are different for the positive and negative populations, and the number of parameters entered into the statistic. As for Figure 2, artificial datasets were constructed that had different representations of two populations. The two populations could be distinguished by either one, two, or three of the three parameters in the data. In each case, 1,500 bins were chosen for the analysis; all three parameters were entered into the PB comparison. For each contour graph, 900 datasets were analyzed. A contour graph of $T(\chi)$ is shown for each condition, as a function of the separation of the two populations vs. the number of positive events. The boundary between the red and blue shaded areas (where $T(\chi)$ is approximately four) represents the minimum detectable contamination of a positive population. For example, for 30,000 events in which three parameters distinguish the positive events (top left), approximately 1,000 events that are different by 0.2 standard deviations of the negative distribution can be detected by the metric. If the distributions differ by 2.0 standard deviations (i.e., do not overlap), then the metric can identify as few as 200 events out of 30,000. These analyses confirm the conclusions in the accompanying paper on the application of the PB metric to univariate data (7): the minimum detectable number of events seems to be a relatively constant number (for a given difference in the positive and negative distributions); thus, collecting a larger number of events will allow the detection of a commensurately smaller percentage of contaminating events.

process must be done with caution, since reduction of the number of samples entered as a control can lead to sampling bias, should the chosen set of controls not be truly representative of all control samples.

Application of the PB Metric to Immunofluorescence Data

We applied the multivariate PB algorithm to immunofluorescence and light scatter data collected on a flow cytometer. Our test data was derived from a three-color immunofluorescence analysis of PBMC from 14 individuals (5 HIV^+, 9 HIV^-). Samples were stained with FITC anti-CD14, PE anti-CD16, and C y5PE anti-CD45; data from 30,000 PBMC was stored.

In this series of comparisons, all 14 data files were entered into the binning stage of the algorithm; the output is therefore the distance from the sum of all 14 collections. As shown in Figure 4, various combinations of the five parameters were entered into the comparison; the PB metric (distance) is graphed, separated by HIV-status.

It is evident that there is a common difference in the distribution of staining between HIV-infected adults and healthy HIV-uninfected adults. In univariate comparisons, this difference is most evident in the CD14 stain; other parameters also carry additional statistical weight to the difference. Interestingly, parameters which by themselves carry no information as to the HIV-infection status (for example, forward and side scattered light) can increase the fidelity of identification of HIV-infected adults when added to the comparison. This simply underscores the multivariate nature of the differences in these distributions, suggesting that it would be nearly impossible on the basis of only univariate comparisons to achieve the desired distinction.

Application of the PB Metric to Identify Genetically-Controlled Staining Patterns

We tested the hypothesis that the PB metric could quantitate differences between lymphocyte staining patterns of cells derived from different tissues and/or strains

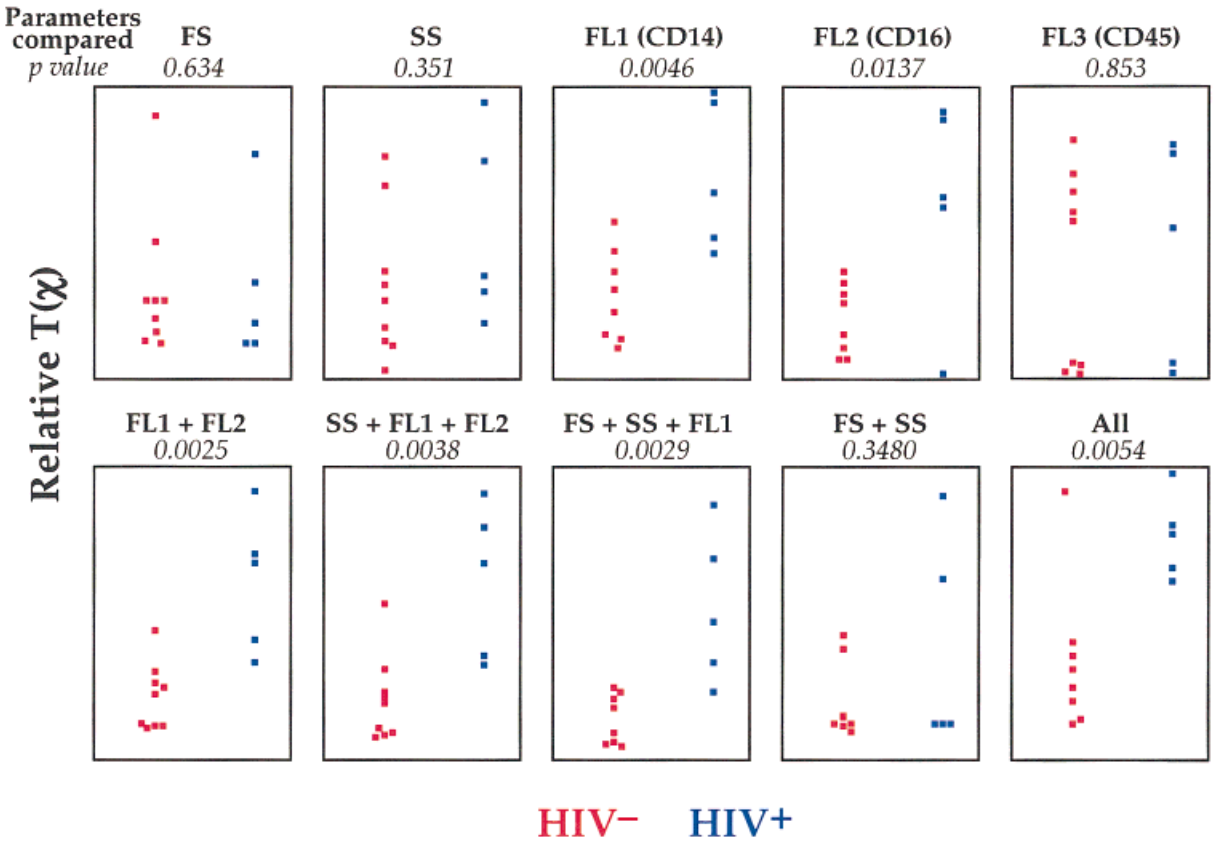

FIG. 4. Application of the PB metric to three-color, five-parameter immunofluorescence data. PBMC from 5 HIV^+ and 9 HIV^- individuals was compared using the PB metric. Cells were stained with FITC-anti CD14, PE anti-CD16, and Cy5PE anti-CD45. Data was broadly gated for CD45+ events and then entered into a multi-sample PB comparison. Each sample was compared against a control comprised of the aggregate of all 14 samples. The relative $T(\chi)$ is shown for comparisons that included different combinations of the 5 parameters. The $T(\chi)$ distributions for HIV^+ and HIV^- were compared using a t-test; the p-value for this comparison is shown above each graph.

of mice. In particular, we analyzed two four-color immunophenotyping panels (principally identifying B cell subsets) applied to bone marrow, lymph nodes, and splenocytes obtained either from inbred Balb/c or C57bl mice, or mice derived from an F1 cross of these two strains. The PB metric was applied to data gated only for live cells (by propidium iodide exclusion) and by scatter gating.

As shown in Figure 5, the PB metric ranks these four-color distributions in a manner that has significant biological meaning. For example, whenever samples were compared to a single control mouse, a litter mate was always much more similar than samples from the same tissues of other strains. Interestingly, the F1 mice were always ranked closer to a parental strain than the other strain–demonstrating that the complex immunophenotype of the F1 hybrid is likely a mixture of phenotypes of each parent. Furthermore, staining patterns of mice differing only by age were much more closely related than mice differing genetically.

Figure 5 allows the ranking of factors which contribute to differences in immunophenotyping patterns of lymphocytes. For example, the most important factor is tissue location: i.e., patterns from different tissues even in the same mouse are more distantly-related than are patterns from the same tissues of genetically-disparate mice. Of the factors evaluated in this study, the relationship is as follows: tissue > background genetics (strain) > age > litter.

Finally, Figure 5C illustrates the reproducibility of the PB metric. Comparison of samples obtained from different genetically identical litter mates to a control sample gave an average coefficient of variation of 11%. This variation is considerably less than the actual differences observed between experimental conditions (i.e., strain, tissue, age).

DISCUSSION

Fundamentally, the multivariate PB algorithm is identical to the univariate PB algorithm (7). The unique aspect of the multivariate algorithm is how the data is divided into bins on which the comparison is performed. As for the univariate comparison, the algorithm we chose results in a binning that equally divides the control distribution. In other words, any single event from the control distribution, selected at random, as the same probability of falling into any of the given bins.

The advantage of this method over existing binning methods is most evident for multiparametric data. Typical

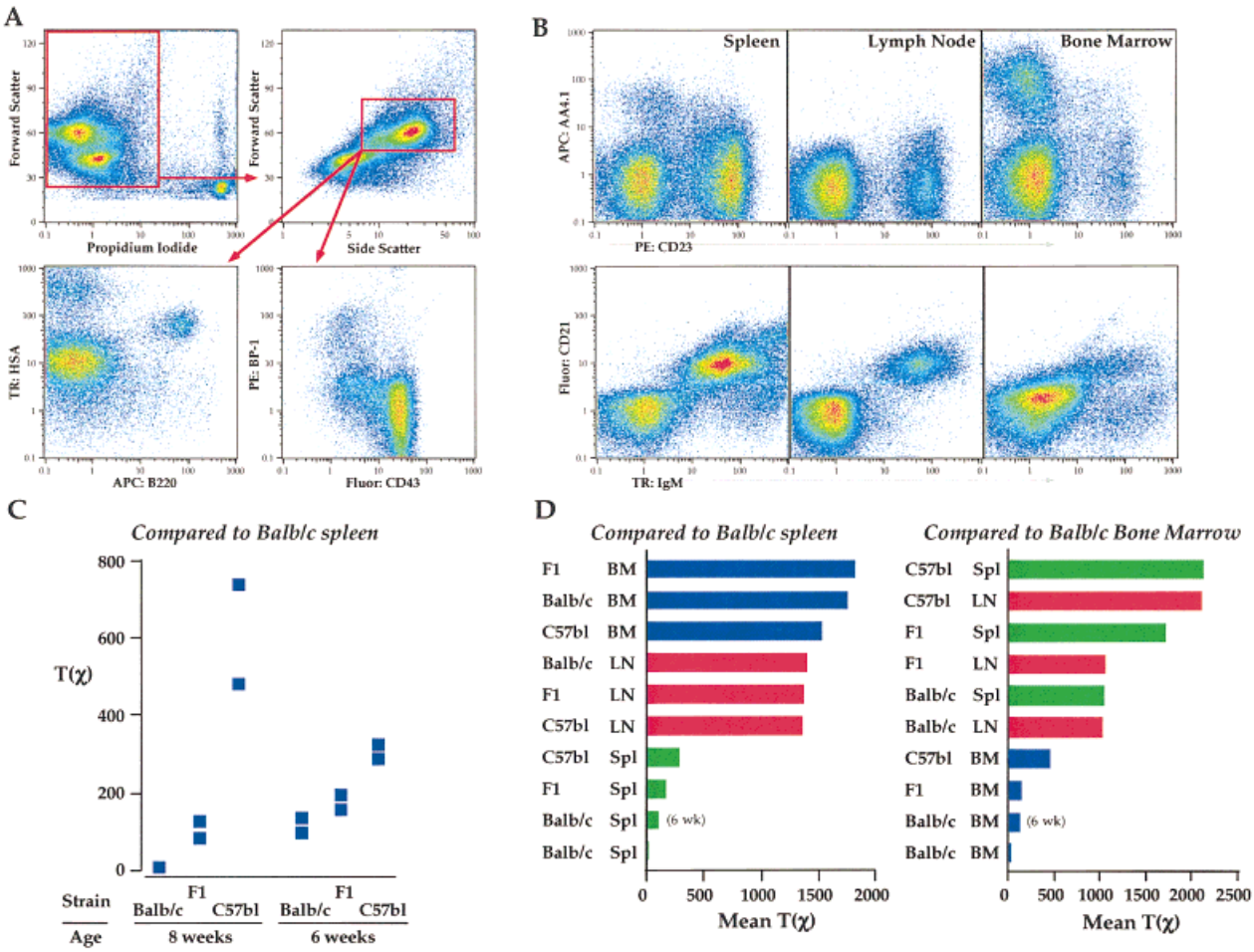

FIG. 5. Application of the multiparameter PB comparison algorithm to identify genetically-controlled murine lymphocyte staining patterns. Mouse lymphocytes were isolated from bone marrow (BM), lymph nodes (LN), or spleen (Spl) from Balb/c, C57bl, or Balb/c x C56bl (F1) mice. Mice were sacrificed at eight weeks of age (except, where noted, at six weeks). **(A)** Balb/c bone marrow cells were stained either with PI and a combination of antibodies to CD43, CD45/B220, HSA, and BP-1. Large live cells were selected based on Forward scatter and PI fluorescence (top left) as well as forward vs side scatter (top right). The bottom two panels are bivariate plots showing the staining patterns obtained with the four antibodies. **(B)** Cells were stained with PI and a combination of antibodies to CD21, CD23, IgM and AA4.1. Live cells were selected as in (A). The panels are bivariate plots showing the staining patterns obtained with the four antibodies for cells from a Balb/c spleen, lymph node, or bone marrow. **(C)** The gating shown in (A) was applied to bone marrow cells from two littermates from each of three strains at two different ages. One eight-week old Balb/c mouse was used as a control sample; all others were compared to it using the PB algorithm. All four immunofluorescence stains were entered into the comparison; approximately 1000 bins were used for the 100,000 event data files. **(D)** The gating similar to that shown in (B) was applied to cells from spleen, lymph nodes, or bone marrow from the same mice as in (C). All four immunofluorescence stains were entered into a PB comparison using 1000 bins on 100,000 events; a single Balb/c spleen (left) or bone marrow (right) was used as the control.

flow cytometry data is collected with 1,024 channel resolution. Thus, the maximum number of bins for a univariate distribution is 1,024 –many of which will be empty for a typical measurement. With two parameter data, however, the maximum number of bins is $1{,}024^2$, or greater than 10^6. With four-color, six parameter data, the number of bins is over 10^{18}. This is much too large to handle with today's computational limits. Thus, the standard approach is to only divide each channel into, for example, 10 bins. Even so, the six-parameter data results in 10^6 bins, a vast majority of which are empty.

Rather than using equally-sized bins, our algorithm generates a large number of very small bins where events are densely packed, and much larger bins where events are spread out (Fig. 1). The limit on the number of bins is primarily driven by the number of events. Given the dependence of the smallest meaningful chi-squared value on event counts, we choose to divide the data into at most $n/10$ bins, where n is the number of events (i.e., resulting in at least 10 events per bin). The division of the data equally into bins is important, because it minimizes the maximum expected variance for the statistic.

The algorithm we chose to divide the distribution chooses the parameter with the greatest variance and divides it in half. A possible downside of this approach is the tendency to divide uniform populations right down the middle of the population. This may tend to emphasize subtle shifts in the position of the bulk population rather

than subtle shifts in outlying events. The other possible disadvantage of using the variance to decide on which parameter to divide is the inherent assumption that all of the parameters are scaled similarly. With typical log-scaled flow cytometric data, this is a reasonable assumption. However, it may be useful to scale the variance of each parameter by the range of the data (or some other normalization factor) in order to weight parameters which have high information content in a relatively small number of channels.

Perhaps a more powerful approach to each division step would be to evaluate the distributions parametrically and decide how many divisions to make at that particular step. For example, if the distribution of the parameter to be divided was uniform or normal, then it might be divided into three tertiles, so as to have bins that emphasize the tails of the distribution and have no division in the middle of the distribution. The choice of which parameter to divide might include criteria not only related to variance, but, for example, how Gaussian the distribution is—i.e., preferentially splitting non-normal distributions with the presumption that they carry more information than normal distributions. These types of algorithmic refinements may become more important for the analysis of 8+ parameter files, where the number of divisions of each parameter is highly limited by the number of events that were collected.

This limitation was discussed in the Results section. While inclusion of parameters that are not different between samples does not necessarily impede the ability of the algorithm to detect differences (Fig. 2), it significantly reduces the number of times other parameters can be divided. The more divisions that are applied to any given parameter, the more subtle the differences in distributions of that parameter can be detected. In any case, it will always be true that collection of more events in each sample will result in greater fidelity for identifying subtle changes or appearances of rare subsets.

The application of the PB algorithm to immunofluorescence data demonstrated that it is potentially very useful for identifying differences in multivariate distributions—even in quantitating those differences to achieve biologically meaningful results. The PB comparison is nonparametric and does not rely on user interaction (beyond possibly setting criteria for the desired number of bins), and thus represents an objective way to compare these distributions. As shown in Figure 4, this algorithm could be used to distinguish HIV+ from HIV– samples. It is important to note that this distinction was possible *without* the prior identification of which samples were HIV+ or HIV–. Each sample was compared to a combined control comprised of the union of all samples. Of course, by comparing the samples to a union of only known HIV– samples, the algorithm is likely to be even more powerful at distinguishing unknown samples.

The application of the algorithm to the analysis of B cells from different strains, ages, and tissues of mice proved to be very interesting. Based on the PB output, we can rank how different is distribution of expression of these markers (in four-dimensional space). Not unexpectedly, we found that cells from the same tissue of a litter mate were much more closely related that cells from a different mouse; different mice were more or less different depending on their genetic backgrounds. Cells from different tissues were even more different than this. These types of analyses may prove highly informative in exploring the genetic control immunophenotype in mouse and even in man.

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 the accompanying paper (9).

In summary, Probability Binning is a novel and unique statistic for comparing multivariate distributions of event data. It provides a reasonable and reproducible probability that two or more distributions are different, it can be used to determine the percent contamination by an second population of events, and it can be used to rank differences between test samples to identify those most or least like the control sample.

LITERATURE CITED

1. Finch PD. Substantive difference and the analysis of histograms from very large samples [letter]. J Histochem Cytochem 1979;27:800.
2. 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.
3. Overton WR. Modified histogram subtraction technique for analysis of flow cytometry data [published erratum appears in Cytometry 1989 Jul;10(4):492–494]. Cytometry 1988;9:619–626.
4. Bagwell C. A journey through flow cytometric immunofluorescence analyses. Clin Immunol Newsletter 1996;16:33–37.
5. Lampariello F. Evaluation of the number of positive cells from flow cytometric immunoassays by mathematical modeling of cellular autofluorescence. Cytometry 1994; 15:294–301.
6. 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.
7. Roederer M, Treister A, Moore W, Herzenberg LA. Probability Binning Comparison: a metric for quantitating univariate distribution differences. Cytometry 2001;45:37–46.
8. Leva JL. A fast normal random number generator. Transactions in mathematical software 1992;18:449–453.
9. Roederer M, Hardy RR. Frequency difference gating: a multivariate method for identifying subsets that differ between samples. Cytometry 2001;45:56–64.