Cytometry Part A 69A:541–551 (2006)

A New "Logicle" Display Method Avoids Deceptive Effects of Logarithmic Scaling for Low Signals and Compensated Data

David R. Parks,[1]* Mario Roederer,[2] and Wayne A. Moore[1]
[1]Department of Genetics, Stanford University, Stanford, California 94305
[2]Vaccine Research Center, National Institutes of Health, Bethesda, Maryland 20892

Received 26 September 2005; Revision Received 18 January 2006; Accepted 20 January 2006

Background: In immunofluorescence measurements and most other flow cytometry applications, fluorescence signals of interest can range down to essentially zero. After fluorescence compensation, some cell populations will have low means and include events with negative data values. Logarithmic presentation has been very useful in providing informative displays of wide-ranging flow cytometry data, but it fails to adequately display cell populations with low means and high variances and, in particular, offers no way to include negative data values. This has led to a great deal of difficulty in interpreting and understanding flow cytometry data, has often resulted in incorrect delineation of cell populations, and has led many people to question the correctness of compensation computations that were, in fact, correct.
Results: We identified a set of criteria for creating data visualization methods that accommodate the scaling difficulties presented by flow cytometry data. On the basis of these, we developed a new data visualization method that provides important advantages over linear or logarithmic scaling for display of flow cytometry data, a scaling we refer to as "Logicle" scaling. Logicle functions represent a particular generalization of the hyperbolic sine function with one more adjustable parameter than linear or logarithmic functions. Finally, we developed methods for objectively and automatically selecting an appropriate value for this parameter.
Conclusions: The Logicle display method provides more complete, appropriate, and readily interpretable representations of data that includes populations with low-to-zero means, including distributions resulting from fluorescence compensation procedures, than can be produced using either logarithmic or linear displays. The method includes a specific algorithm for evaluating actual data distributions and deriving parameters of the Logicle scaling function appropriate for optimal display of that data. It is critical to note that Logicle visualization does not change the data values or the descriptive statistics computed from them.

Key terms: data display; flow cytometry; fluorescence compensation; data scaling; data transformation

Practical experience has demonstrated that marker distributions measured by flow cytometry are often more-or-less log-normal or are composed of mixtures of log-normal distributions. Logarithmic data scales, which show log-normal distributions as symmetrical peaks, are widely used and accepted as those facilitating analysis of fluorescence measurements in biological systems (1).

On the other hand, cell populations with low mean, high variance, and approximately normally distributed fluorescence values occur commonly in various kinds of flow cytometry data. In particular, data values for cell populations that are essentially unstained or are negative for a particular dye, after fluorescence compensation, should be distributed more-or-less normally around a low value representing the autofluorescence of the cells in that data dimension. Data sets resulting from computed compensation commonly (and properly) include populations whose distributions extend below zero. (When analog compensation is used, such distributions should also appear, but the electronic implementations distort or truncate the distributions, so that negative values are suppressed.)

Logarithmic displays, however, cannot accommodate zero or negative values and often show a peak above the actual mean or median of the population with a pileup of events on the baseline (see Fig. 1). This effect has been the source of considerable confusion and has been commonly referred to as the "log artifact." Linear scaling is more appropriate and more easily interpreted for display

*Correspondence to: David R. Parks, Stanford University, Beckman Center B007, Stanford, CA 94305-5318, USA.
E-mail: drparks@stanford.edu
Published online 7 April 2006 in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/cyto.20258

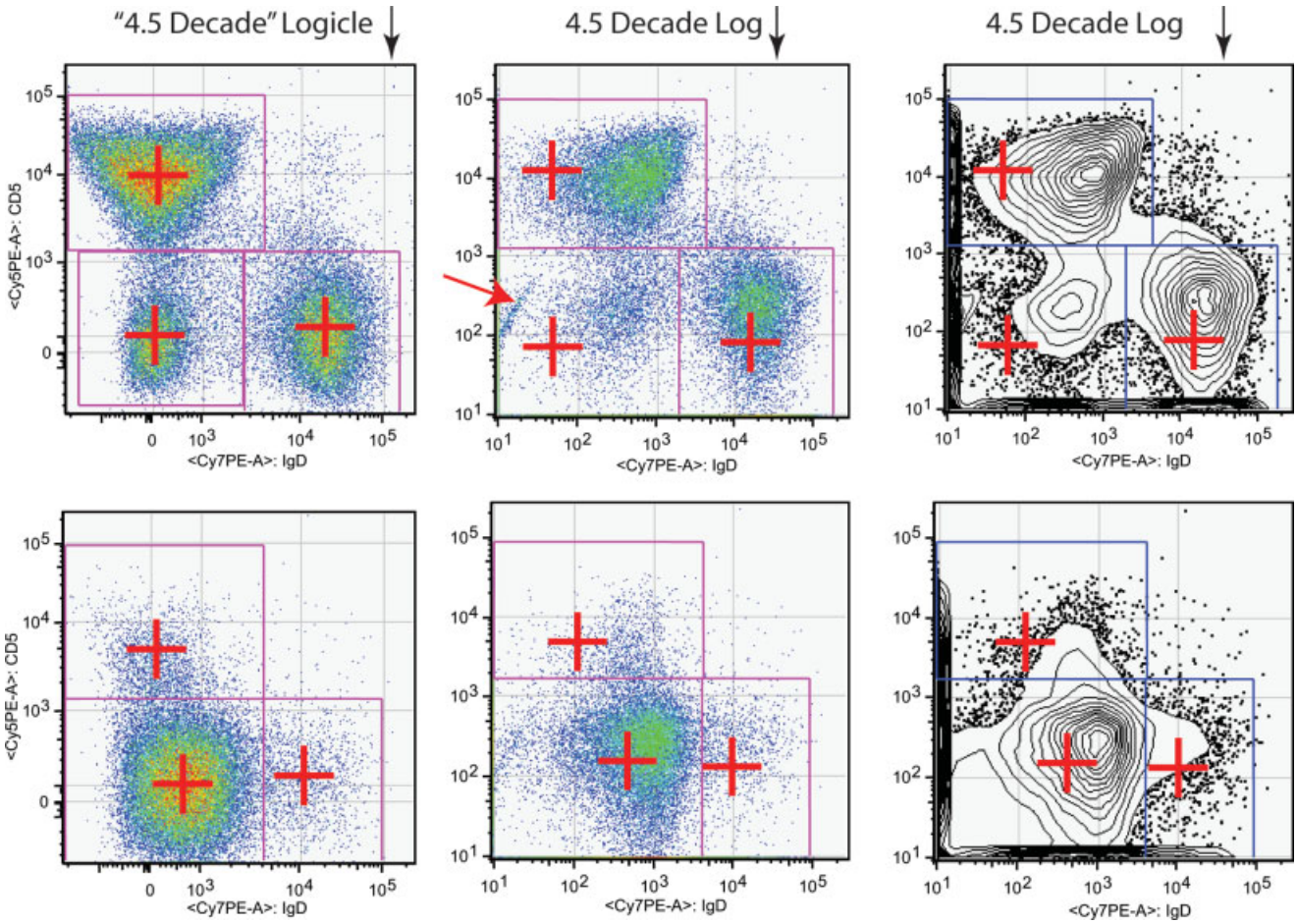

FIG. 1. Two samples of mouse spleen cells stained for CD5 and IgD and gated in light scatter and other fluorescent markers for viable lymphocytes (upper panels) or viable non-T-cells (lower panels). The measurements were made on a FACSAria and compensated and analyzed in FlowJo. The arrow in the upper–center logarithmic display panel points out a data artifact resulting from applying the compensation calculation to log data, in which low or negative data values have been truncated. This does not occur when the full data is retained.

of fluorescence compensated data on cell populations that are low to negative for a particular dye.

Thus, there is a need for display scales that combine the desirable attributes of the log scale for large real signals with those of the linear scale for unstained and near-background signals. The Logicle method presented here solves this problem by plotting data on axes that are asymptotically linear in the region, around a data value of zero and asymptotically logarithmic at higher (positive and negative) values.

Figure 1 illustrates the utility of Logicle scaling in facilitating accurate interpretation of flow cytometry data. The Logicle displays in the left panels show well-defined cell populations in the gated regions. There are very few events (well under 1%) on the baselines, and the medians of events in each region (marked with crosses) are appropriately central to the visual data. In contrast, logarithmic presentation of the same data sets (center and right panels) makes the actually compact cell populations look split into above-baseline and on-baseline "populations." In each data set, about 45% of the data events are on the baselines (and in the dot displays almost invisible). The medians of the populations are nowhere near their visual centers. Logarithmic scaling, therefore, produces unintuitive data displays, and can lead to incorrect data evaluations and attempts to define separate populations that are not in reality separate. Additional benefits of Logicle data display are discussed later in the Results section.

Background on Multicolor Fluorescence and Compensation

In a flow cytometer, each fluorescence detector accepts light from a particular laser excitation and in a particular range of emission wavelengths optimized to detect a particular dye. However, each dye whose excitation is nonzero at that laser wavelength and whose emission is not zero in the detector's emission band will contribute signal on that detector. Therefore, although fluorescent dye combinations used in flow cytometry are selected to minimize spectral overlaps in multicolor measurements, each dye will typically contribute signal on several detectors, and each detector will receive some signal from several dyes.

For each cell in a biological analysis, we generally want to separate the signal contributions from the different dyes, so that an estimate of the amount of each fluorescent

reagent is obtained. The process of converting from fluorescence color measurements to dye estimates is commonly called fluorescence compensation. Although the technique was originally developed for analysis of two-color single laser measurements (2), it is particularly critical in multicolor work. By evaluating the response of each of the detectors to a series of compensation control samples, each of which is labeled with only one dye, we construct a matrix of relative spectral overlaps. For each cell, we multiply a set of detector color measurements by the inverse of the spectral overlap matrix to obtain the corresponding set of dye estimates for the cell. This calculation is based on simple linear algebra, so any particular set of color measurement values yields a specific set of dye estimates. The estimated dye amounts are exactly those whose total signal on each detector would yield the color measurements actually observed.

Statistical Uncertainties in Dye Estimates

As is so often the case, this algebraic analysis is not complete in the real world. The fundamental deviation comes from the quantum nature of light and the finite amount of light detected. Thus, the detected signal is subject to what is commonly called counting statistics, governed by the Poisson distribution. In practice, the limiting step is the number of photoelectrons emitted at the cathode of the photomultiplier tube. The standard deviation of actual measurements in relation to their theoretical expectation scales with the square root of the number of photoelectrons detected.

For cells with just autofluorescence or very low dye levels, the effects of photon statistics, possible electronic noise, and real differences in low-level fluorescence among cells in a particular population often result in signal distributions with low means and high relative variances. This problem becomes magnified after fluorescence compensation, since the compensated value is subject to error contributions from multiple measurements. This can readily lead to a standard deviation of the dye estimate which is greater than the mean for that estimate. This phenomenon has been discussed and illustrated by Roederer (3). The end result is that distributions of compensated dye estimates for cells that are unstained by a given dye are often nearly normal and centered near zero, and may have large variances compared to the corresponding distributions for totally unstained cells. In particular, this process can properly result in negative dye estimates for some cells even though, of course, negative dye amounts are not possible. These negative values must not be disregarded, since truncating them will deform the data distributions and result in incorrect computation of signal means.

The overall result is that cell samples measured by flow cytometry often contain cell populations whose signal distributions are appropriately represented in logarithmic displays along with populations whose distributions cannot be properly shown in a logarithmic display. Logicle functions and methods were developed to provide unified displays in which these different populations can all be represented in a clear and intuitive way.

MATERIALS AND METHODS

Test Particles and Cell Samples

Spherotech Rainbow multidye particles (Spherotech, Libertyville, IL) were used for the data in Figure 5. Reagent capture beads carrying a monoclonal rat-anti-mouse κ antibody and matched blank beads (both from BD Biosciences, San Jose, CA), were used to produce the data in Figure 6. All of these particles are about 3 μm in diameter.

Cell samples, used to generate the illustrations, are described in the figure captions.

Instrumentation

Illustrative data were obtained using a FACSAria (BD Biosciences, San Jose, CA), which employs linear digital data acquisition with 14 bit sampling at 10 MHz rate. Area signals are produced as sums, over a range of ~50–100 samples, and are presented as 18-bit linear values. Since background subtraction is included in the evaluation, zero and negative data values can occur. These are preserved in the floating point FCS files.

Data Analysis and Modeling

Analysis of data from the flow cytometer was carried out in FlowJo (Tree Star, Ashland, OR) version 4.3 or later. This package includes the ability to import floating point FCS files into Logicle scaling, so that negative data values are retained. FlowJo provides support for computed fluorescence compensation, including automatic selection of appropriate Logicle scale functions for each compensated data dimension. In producing Figure 2, the spectral matrices used in computing fluorescence compensation were edited externally and imported into FlowJo to generate illustrative data distributions.

Modeling and plotting of Logicle functions and various other functions for comparisons were carried out in Microsoft Excel.

RESULTS

Criteria for a New Data Display Method

We developed the following criteria for defining a new scaling function that would yield better displays for much flow cytometry data than can be produced using traditional logarithmic or linear scaling.

- The display formula supports a family of functions that can be optimized for viewing different data sets.
- The function becomes logarithmic for large data values, to ensure a wide dynamic range and to provide good visualizations of the often log-normal distributions, at high fluorescence intensities.
- The function becomes linear near zero, and extends to negative data values and is symmetrical around zero, providing near-linear visualization, appropriate for linear–normal distributions at low fluorescence intensities.
- The transition between the linear to logarithmic regions is as smooth as possible, to avoid introducing artifacts in the display.

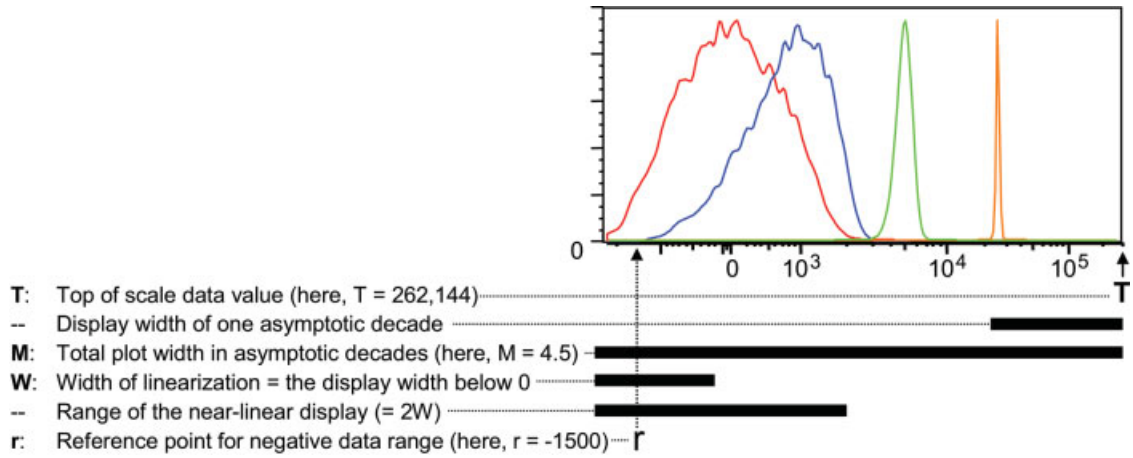

FIG. 2. How the Logicle parameters relate to the resulting Logicle scale and data display. M and W are expressed in decades, i.e., base 10 log units. Their natural log forms are $m = M \ln(10)$ and $w = W \ln(10)$. The data curves are only for illustration here but are used and described in Figure 5. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

- As the linearization strength is increased to accommodate a wider range of linearized data values, the reasonably linear region of the data values grows faster than the size of the linearized region in the display. Thus, the user has a visual indication that a greater degree of linearization is in use, but the display space is balanced between more linear and more logarithmic regions.

Specification of Logicle Functions

By considering these criteria and examining the behavior of a number of functions, we concluded that particular generalizations of the hyperbolic sine function (sinh), which we came to call Logicle functions, can best meet the criteria. The hyperbolic sine function itself has the desirable properties of being essentially linear near zero, becoming exponential for large values (leading to a logarithmic display scale there), and making a very smooth transition between these regions (i.e., it is continuous in all derivatives), but it does not provide enough flexibility to meet the display needs encountered in flow cytometry.[1]

The hyperbolic sine function itself is given as follows:

$$\sinh(x) = (e^{x} - e^{-x})/2 \qquad (1)$$

This can be generalized to what we call biexponential functions,

$$S(x; a, b, c, d, f) = ae^{bx} - ce^{-dx} + f \qquad (2)$$

Interpreting the condition of maximal linearity around data value zero to mean that the second derivative of the function should be zero therein, we identified a subset of biexponential functions with this property and call them Logicle scaling functions.

Besides the constraint just specified, there are four further choices that need to be made to fix the five parameters in Eq. (2) (a, b, c, d, and f), and thereby define a specific display. How these choices appear in an actual Logicle display is illustrated in Figure 2. The parameters described later and in Figure 2 are not simply a, b, c, d, and f, but, once specified, they uniquely determine the function in Eq. (2). The first choice is the maximum data value in the displayed scale (T). The second is the range of the display in relation to the width of high data value decades (M or m in decade or natural log formulations, respectively). If this is held constant among plots optimized to different data sets, the nearly logarithmic area at the upper end of each display will be essentially the same, while the region near data zero is adjusted to optimize for different data sets. We have found that a total plot width of 4.5 "decades" is usually a good choice for displaying flow cytometry data.

The third choice is the strength and range of linearization around zero (W or w). The linear slope at zero (in, for example, data units per pixel or data units per mm in a printout) and the range of data values in the nearly linear zone are determined by this selection. In displaying a particular data set, the linearized range must be adequate to cover broad population distributions that do not display well on log scales. This, in particular, is the selection that is critical in matching displays to particular data sets and in ensuring that the linearized zone covers the range of statistical spread in the data. If the transition toward log behavior occurs in too low data values, the artifacts seen in logarithmic displays will not be suppressed.

The fourth choice is to specify the range of negative values to be included in the display (which also defines the position of the data zero in the plot). This range must be great enough to avoid truncating populations of interest. In practice, as shown in Figure 2, we find that it is desirable to link the third and fourth choices as a single

[1] *NOTE:* In cytometry, we normally label "logarithmic" axes with values from the corresponding exponential function, rather than with the logarithm itself, e.g., decade labels like 10, 100, 1,000 and not 1, 2, and 3. The Logicle functions defined in the equations given later are data value functions. Their inverses provide Logicle display functions in the same way that exponential scaling functions provide logarithmic data displays.

value. This assures that the lowest negative data values in view correspond to the approximate edge of the linearized zone. As discussed earlier under Statistical Uncertainties, negative values should occur only as a result of statistical spreading, and, therefore, they should be displayed within the near-linear zone.

Assuming that the top-of-scale value and the nominal "decade" width of the display have been selected, linking the third and fourth choices results in a family of functions with only one parameter to be adjusted to match the particular data set being displayed.

Using natural log units, an expression for the Logicle scaling function that embodies all of the constraints and choices described earlier is given as follows:

$$S(x;w) = Te^{-(m-w)}(e^{x-w} - p^2e^{-(x-w)/p} + p^2 - 1) \quad \text{for } x \geq w \qquad (3)$$

In Eq. (3), T is the top of scale data value (e.g., 10,000 for common 4 decade data or 262,144 for an 18 bit data range).

$w = 2p\ \ln(p)/(p + 1)$ is the width of the negative data range and the range of linearized data in natural log units. p is introduced for compactness in presenting the Logicle function, but p and w together represent a single adjustable parameter.

m is the breadth of the display in natural log units. For a 4.5 decade, display range $m = 4.5\ \ln(10) = 10.36$.

The display is defined for x in the range from 0 to m. Negative data values appear in the space from $x = 0$ to $x = w$, and positive data values are plotted between $x = w$ and $x = m$ (where the top data value T occurs). The form shown as Eq. (3) is for the positive data zone, where $x \geq w$. For the negative zone where $x < w$, we enforce symmetry by computing the Logicle function for the corresponding positive value $(w - x)$ and changing the sign. The data zero at $x = w$ is where the second derivative is zero, i.e., the most linear area.

To select an appropriate value for w to generate a good display for a particular data set, we obtain a reference value marking the low end of the distribution to be displayed. As described later, we typically select the data value at the fifth percentile of all events that are below zero as this reference value. Designating this (negative) value as "r," and using its absolute value abs(r), w is computed as follows:

$$w = (m - \ln(T/\text{abs}(r)))/2 \qquad (4)$$

Equations (3) and (4) can be rewritten using base 10 representation in order to express the parameters in terms of "decades" of signal level or display:

$$S(X;W) = T * 10^{-(M-W)}(10^{X-W} - p^2 * 10^{-(X-W)/p} + p^2 - 1) \quad \text{for } X \geq W \qquad (5)$$

In Eq. (5), $W = 2p\ \log(p)/(p + 1)$ is the width of the negative data range and the range of linearized data in "decades" and M is the breadth of the display in "decades." For a 4.5 decade, display range $M = 4.5$.

We obtain W from the negative range reference value "r" as follows:

$$W = (M - \log(T/\text{abs}(r)))/2 \qquad (6)$$

Figure 2 illustrates the relationship between these parameters and the resulting Logicle display.

Specifying a logarithmic display requires two values corresponding to T and M, and the scaling near the upper end of a Logicle plot approximates that of a logarithmic display with the same values of T and M. The additional linearization width, W, adapts the Logicle scale to the characteristics of different data sets.

Logicle functions with different values of W are plotted in Figure 3 along with the linear and exponential functions that match them around data zero and at high data values, respectively. Use of an exponential function for scaling is what results in a logarithmic scale. Note that each Logicle curve closely follows its matched linear function at low signal values, confirming good linearity in the region around data zero. At middle signal values that vary depending on the value of W, the Logicle functions depart from linearity and move smoothly toward the exponential line. At high signal levels, the Logicle curves become indistinguishable from the exponential line.

Figure 4 shows a Logicle curve for $W = 1.0$ and its matched linear and exponential curves displayed with a linear signal level scale. The signal level scale is expanded (top of scale is 300 rather than 10,000) to show in detail the matching of the Logicle and linear curves at low signal levels and the divergence of the Logicle curve at higher levels and the beginning of its approach to the exponential curve.

Strategy for Selecting the Width Parameter

As we have discussed, proper estimates of dye signals using measurements on individual cells may be negative, but actual negative dye amounts are impossible. Therefore, any negative values present in the compensated data must be due to purely statistical effects. This is true despite the presence of essentially arbitrary positive staining distributions. Thus, for a population with near zero mean and significant statistical spread, the most negative values indicate the necessary range of the negative part of the scale, and they also indicate the range of linearization needed to ensure that the population will be displayed in a compact and unimodal form. The positive part of the population is less helpful, since it may overlap with other populations in the data set and may not provide a clear upper end with which to define a suitable range for linearization.

A simple strategy of choosing the fifth percentile of the negative data values to set this scale seems to work well and combines adequate sensitivity to extreme values with reasonable sampling stability. Using this strategy (the one currently implemented in FlowJo and illustrated in Figure 2

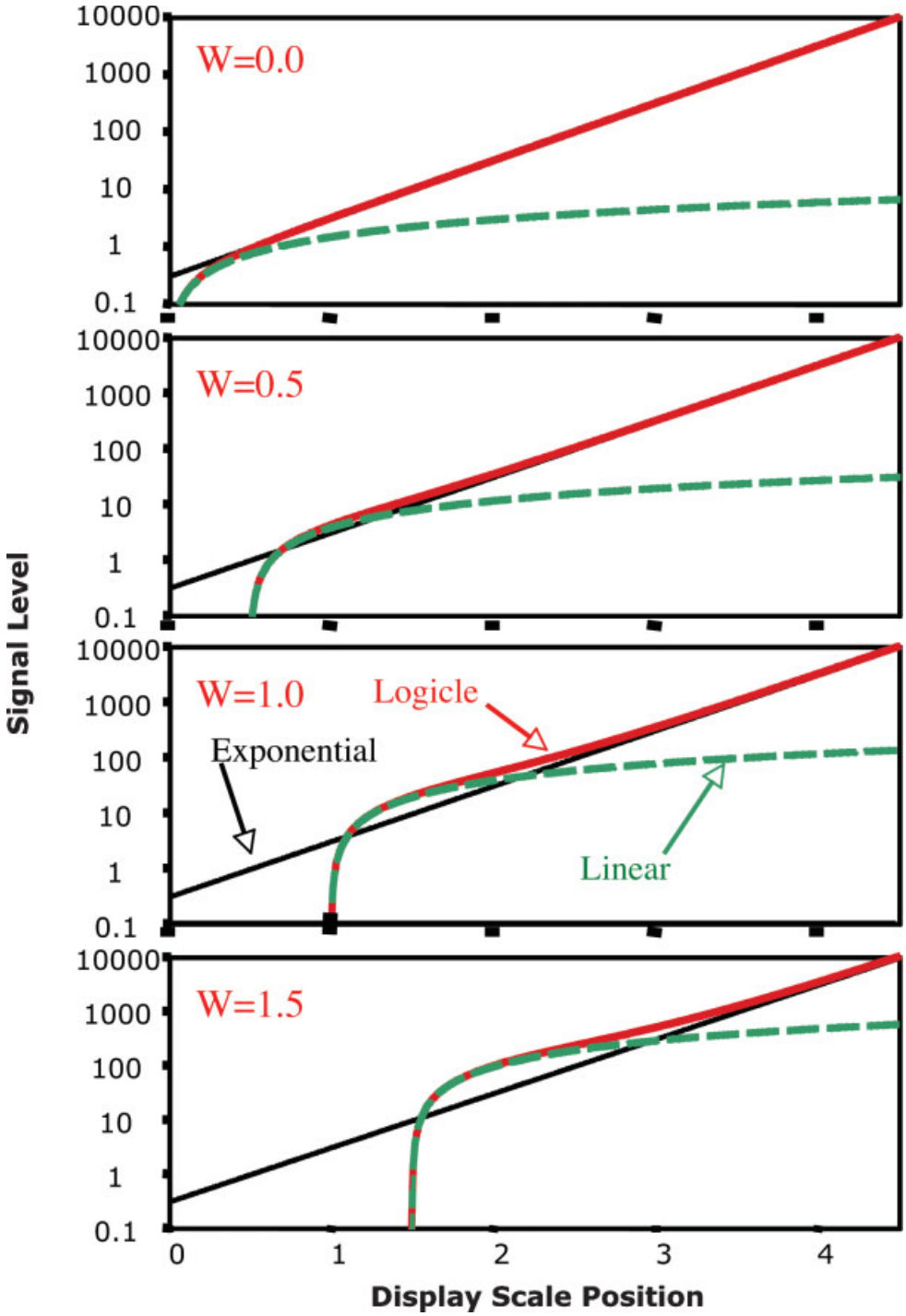

FIG. 3. Logicle, linear, and exponential scaling functions. The Logicle functions are plotted for $W = 0$, $W = 0.5$, $W = 1.0$, and $W = 1.5$. The display range covers 4.5 "decades," and the signal level scale is logarithmic, so only the positive data values can be represented. The black diagonal line in each panel is a pure exponential, i.e., the scaling function for a standard logarithmic display. The green broken lines are pure linear functions with zero crossings, and slopes matched to the corresponding Logicle curves (red). [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

based on the leftmost of the four data distributions), the visible negative data range extends somewhat below the fifth percentile of negatives reference data value, so that almost all the negative data (out to roughly 1.5 times the negative reference data value) is actually seen in the plot.

In cases where no negative data values occur or the negative values are all close to zero, our experience indicates that a minimal Logicle scale sufficient to linearize data in the range of cell autofluorescence provides a more readily interpreted view of the data than does a purely logarithmic scale.

In some data sets, there are few negative data values, but some aberrant events yielding extreme negative values also occur. In such cases, the fifth percentile of negatives value may lead to a value of W, too high for optimal display of the main data set. Gating out the unrepresentative negative data points and reapplying the automatic scale selection to the gated data cures this problem.

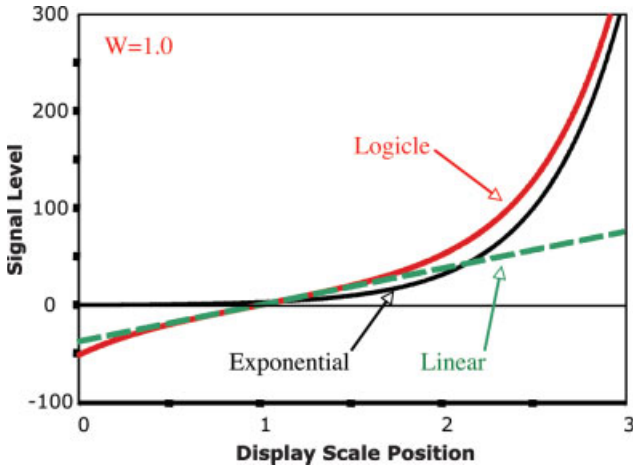

FIG. 4. Logicle, linear, and exponential scaling functions. The same Logicle function, shown in the $W = 1.0$ panel of Figure, 3 is presented with a linear signal level scale. To visualize the relationships between the different functions clearly, the signal level is shown only from −100 to +300, and the display range is shown only from 0 to 3. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

To achieve consistency in data display when analyzing experiments that include a number of samples to be compared, it is appropriate to fix the Logicle scale (for each dimension) based on the most extreme sample present (usually one with the maximum number of labels in use) and use these fixed scales to analyze all similarly stained samples in the experiment. The current implementation in FlowJo bases the scale selection on a single user-specified (gated) data set. A simple and probably desirable variant of this method which has not yet been implemented in user software would operate on a group of data sets designated to be analyzed together. The Logicle width parameter would be evaluated for each dimension in each data set, and the largest resulting width in each dimension would be selected for the common displays. In general, when there are multiple populations in a single sample or multiple samples to be viewed on the same display scale, the population or sample with the greatest negative extent should drive the selection of W.

The method we have chosen for defining the negative end of the display scale in relation to the linearization width makes it possible to evaluate the appropriateness of a particular scaling for a specific data set, by examining the negative data region. If a substantial fraction of the negative data values pile up at the low end of the scale, the value of W is too low to properly display this data, and a higher value of W should be used. If there is a lot of empty negative data space below the lowest population of interest, the linearized region around zero is more compressed than necessary. The population will be properly compact and unimodal, but it would be advantageous to lower W and obtain a more expanded view.

The Effective Dynamic Range of a Logicle Display

We can give a precise expression for the range of variation in scale across a Logicle plot in a form analogous to the "dynamic range" of a logarithmic plot. An ordinary logarithmic scale is often characterized by the number of "decades", i.e., by the common logarithm of the ratio of the maximum to the minimum data values. Clearly, with Logicle scales that extend through zero, such a formula cannot work. However, if we consider the variation in the number of data units corresponding to a given width on the display, we get a relevant and useful ratio corresponding to the range of expansion or compression of the data across the plot. Mathematically, this is the ratio of the highest and lowest values of the slope or derivative of the scale function within the plot. For an ordinary logarithmic scale, this method yields exactly the same results as the usual procedure, i.e., the common logarithm of this ratio of slopes is the same as the number of decades, as defined earlier. For a Logicle scale, the ratio of maximum to minimum derivatives (at the top of scale and data zero, respectively) varies as a function of the linearization width W.

Working from the expression in Eq. (3), the derivative is given as follows:

$$S'(X;W) = Te^{-(m-w)}(e^{x-w} + pe^{-(x-w)/p}) \quad \text{for } x \geq w \qquad (7)$$

The effective dynamic range discussed earlier is $S'(m;w)/S'(w;w)$, i.e., the ratio of derivatives at $x = m$ and $x = w$.

For the Logicle curves illustrated in Figure 3 with $M = 4.5$ decades, the effective dynamic ranges are 4.2, 3.5, 2.8, and 2.1 decades for width values $W = 0.0$, 0.5, 1.0, and 1.5, respectively. (The dynamic range of the logarithmic plot with comparable scaling in the upper range would be 4.5 decades.)

Illustrations and Interpretation of Logicle Displays

Figure 5 shows a comparison of logarithmic and Logicle displays of four signal level distributions that have different means but the same real width of about 2,000 signal level units. Note that the two higher level curves look essentially the same in the two displays, since they occur at signal levels where the Logicle scale is nearly logarithmic. However, the lowest curve is shown very differently in the two graphs. In the Logicle plot, the mean data value occurs at the visual center of the peak and very few data events (less than 1%) fall at the low edge of the scale. In contrast, the logarithmic display for this data set fails to convey an accurate view of the data, in that the mean of the data appears in a highly counter-intuitive location far from the apparent peak of the plot. Also, of course, 49% of very low and negative data values are piled up in an uninterpretable spike at the left edge of the display. This kind of behavior constitutes what may be referred to as a "log artifact" or, more colorfully, the "valley of death." The second curve from the bottom is intermediate in that it is well represented in the Logicle display, but shows a moderate amount of "log artifact" in the logarithmic display.

Figure 6 illustrates the value of Logicle displays for intuitive and accurate interpretation of fluorescence compensated data and their particular value in the analysis of data acquired in high resolution linear data systems. A mixture

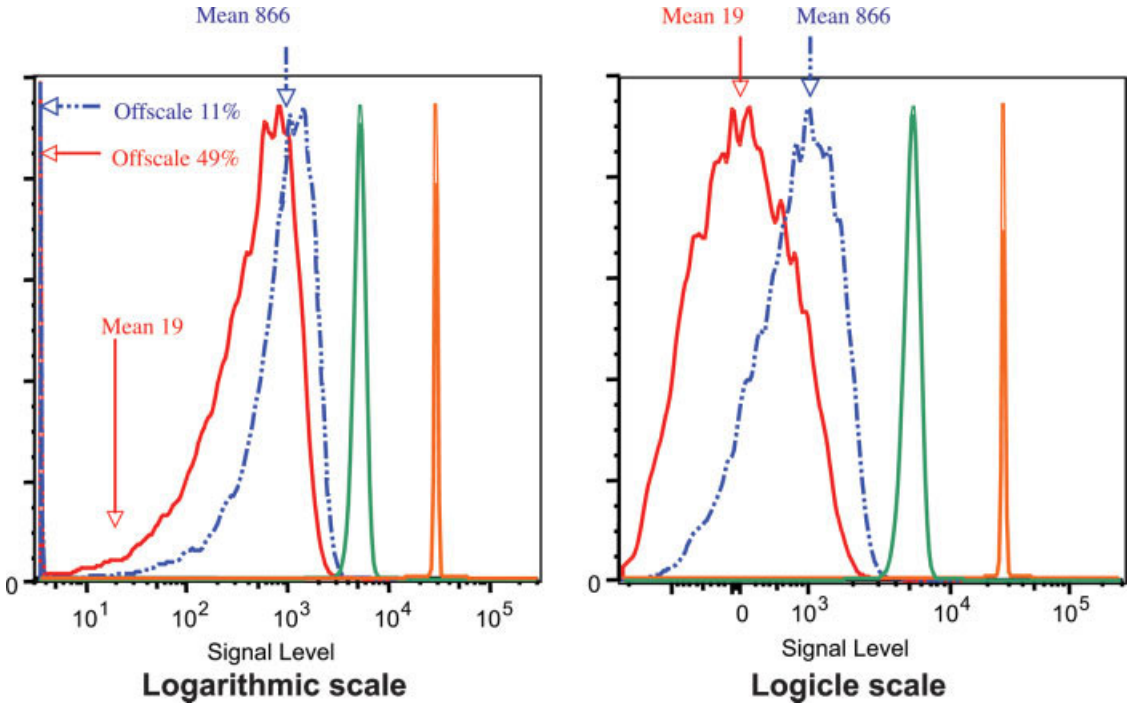

FIG. 5. Logarithmic and Logicle presentations of four data distributions whose means are different but whose real widths are the same. The distributions were generated by applying different "compensation" amounts to a distribution for single test particles. The highest curve at about 30,000 units represents zero compensation. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

of unlabeled microspheres and antibody capture microspheres (BD Biosciences) loaded with FITC antibody were analyzed on a FACSAria cytometer (BD Biosciences), which produces floating point data with values up to 2^{18} or 262,144 and may include (background subtracted) data values below zero. In the upper panels, uncompensated data is shown in Logicle, 4-decade log, 5.5 decade log pseudocolor dot display, and 5.5 decade log contour display. Computed compensation based partly on this sample itself leads to the matching compensated data set shown in the lower panels. In the uncompensated Logicle display, the population of unlabeled particles forms a compact two-dimensional peak centered near zero and includes some negative values for events whose measured signal was below the average background. The 4-decade log display piles up all data values below 26 (= 262,144/10,000) at 26, and the 5.5-decade displays pile up zero and negative data values at 1. The arrow in the 5.5 decade log pseudocolor dot display points out the distracting but otherwise harmless "picket fencing" in the low region, where display pixels are denser than actual data values.

In Logicle displays of compensation control samples, it is easy to confirm that compensation is correct. The compensated Logicle display (lower left) shows clearly that, as expected for an FITC compensation control, the centers of the distributions for unlabeled particles and FITC-labeled particles match at a value near zero in the <PE-A> dimension. It is obvious that the FITC high population has greater spread in the <PE-A> dimension (as would be expected from the discussion earlier under Statistical Uncertainties in Dye Estimates) and that the threshold amount of real PE needed for identification of PE positive events would be greater on the FITC high population than on the FITC negative population. In the logarithmic displays of the compensated data, the apparent center of the FITC labeled population looks higher in the <PE-A> dimension than does the center of the unlabeled population. This is another manifestation of the "log artifact." In fact, the PE dimension medians of the two populations are equal. Adjusting compensation by eye using logarithmic displays is unlikely to lead to correct compensation.

An appropriate selection of the Logicle width parameter assures that almost all data events will be displayed on scale. Less than 1% of the events in the compensated Logicle display in Figure 6 fall on the baselines. In the logarithmic displays, 45–80% of the events in the two populations fall on the baselines where their frequencies and actual measurement values cannot be interpreted visually. The events piled up on the low margins of the logarithmic color dot plots are almost invisible, while the pileup contours on the margins of the logarithmic contour plot make it look like that there may be separate populations there.

DISCUSSION

Additional Benefits of Logicle Methods

Full range Logicle displays of fluorescence compensated data are useful in detecting errors, in avoiding erroneous interpretations and in monitoring for quality control purposes. Since negative data values should be generated

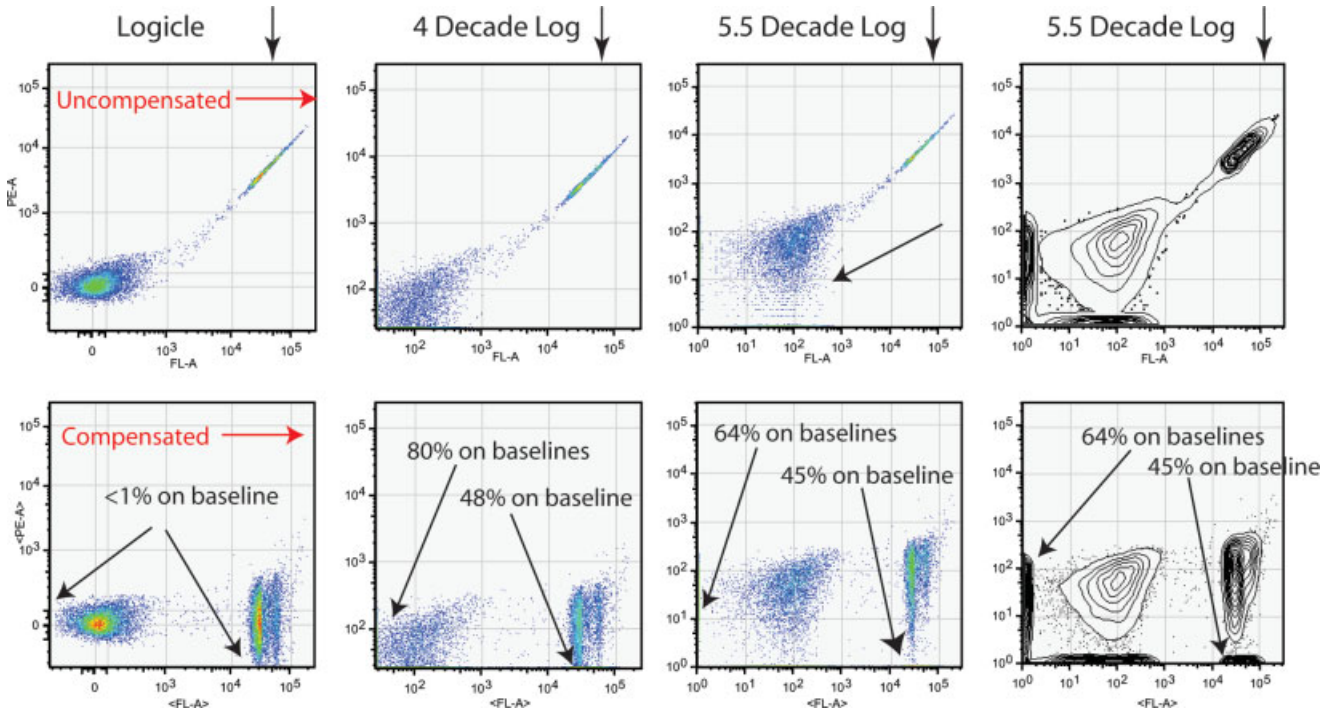

FIG. 6. Comparison of Logicle, 4-decade log, and 5.5-decade log displays for uncompensated and compensated versions of a single stain compensation control sample. The sample consists of a mixture of unlabeled microspheres and reagent capture microspheres (BD Biosciences) loaded with FITC antibody. The arrow in the third upper panel points out "picket fencing" in the 5.5-decade log display. The 80% and 64% on baselines designation includes events on the lower part of either the horizontal or vertical axis. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

purely by statistical processes producing more-or-less normal distributions, data distributions in the negative zone should reflect this and not include peaks or other additional structure. Any such structure points to a problem in the data itself or in the data processing which should be corrected before proceeding with the analysis. In particular, errors in defining the compensation matrix or applying the wrong matrix for the data will frequently produce clear visual artifacts in the negative data range.

Logicle coding could provide a compact way to store and transfer high dynamic range data of the types appropriate for Logicle display while retaining appropriate resolution over the whole data range. For example, recent instruments from BD Biosciences produce data values from 2^{18} down through zero to negative values, presented as 32 bit real numbers. Logicle coding at 10–12 bits could retain all the relevant resolution in most data acquired on such instruments.

Alternative Approaches and Proposals

As described later, several possible data display methods and functions occurred to us or have been suggested by the work of others. In the course of investigating these and evaluating their limitations, we framed the list of criteria presented at the beginning of the Results section that led us to define Logicle functions. None of the other methods fulfills these criteria, and no proposal for alternative displays that we are aware of has adequately addressed the issue of how to choose the scale parameter(s) optimally to match particular data. Bagwell (4) discusses the factors involved in making adequate choices among his Hyperlog functions, but recommends generic scale choice rather than optimization to particular data.

We considered the method of adding a constant to all data values, thus making all or nearly all of the negative values positive and then taking the logarithm, but found that, while it mitigates the distortions of populations with high variance and small mean that occur in logarithmic displays, it still produces the "log artifact." It also does not have good linearity in the near zero region.

Another fairly obvious approach is to simply pick a transition point and use the logarithm for higher data values and a linear scale for smaller values. If the splice is made so that the resulting function is smooth, i.e., continuous in the first derivative so as to minimize distortion of distributions at this boundary, then the function is completely determined by the choice of splice point. We found that the derivative matching requirement in a linear-log splice leads to functions with too little flexibility or adjustability to meet the criteria we set. Splice functions that do not match at least the first derivative at the splice would tend to generate significant artifacts in the display.

One application area where functions close to linear around zero and logarithmic for high data values have been developed is in coding and compression of audio signals where the process is called "companding" (5). Such audio is, of course, bipolar, so negative values must be

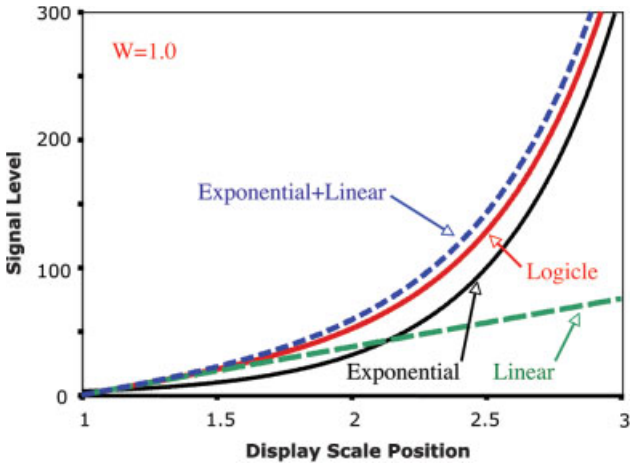

FIG. 7. Detailed comparison of Logicle and exponential-plus-linear display functions. The Logicle, exponential, and linear curves in this plot are drawn from the same data shown in Figure 4, but only the display scale range from 1 to 3 is shown here. The exponential + linear function was chosen to match the Logicle function as well as possible, near zero and at high values. The scale expansion allows the small differences between the Logicle and Exp + Lin curves to be seen clearly. Note that both the Logicle and exponential + linear functions have the same slope as the linear line at display scale = 1 (signal level zero) and that the Logicle curve stays somewhat closer to both the linear and log curves. Bagwell's hyperlog display function (4) is equivalent to the exponential + linear formulation illustrated here. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

handled, and human hearing has a more-or-less logarithmic response to high signal values, so recording such values to high resolution is not important. There are two versions in use. The American one is the same as the offset log described earlier. The European version uses the log-linear splice approach, which is also discussed earlier. These techniques, as defined, are not flexible enough to deal adequately with flow cytometry data.

Hyperbolic sine functions with zero offset and scale adjustment, but without the generalization and constraints that yield the Logicle formulation, have been used to provide a variance stabilizing transformation for microarray expression level data (6,7,8). This transformation (called "glog" by Munson) was found to be useful in making valid tests for significance of gene expression changes, but it was not investigated as a method for data display and visual interpretation. When used for data display, this transformation behaves similarly to the log-linear splice, and does not provide the kind of adjustment needed to optimize visual interpretability.

We also investigated and rejected an approach that combines the linear and logarithmic properties by adding together a linear function, an exponential function and a constant, and then using the inverse function as a scale. This functional form has subsequently been promoted by Bagwell using the name Hyperlog (4). In regions where the exponential term is large, the linear term is essentially irrelevant and, conversely, when the exponential term is small, the linear term dominates. This turns out to closely approximate the behavior of the Logicle functions, and a version similar to any given Logicle function can be obtained by replacing the e^{-x} term in the Logicle function in Eq. (3) with a truncated power series expansion. The expansion is given as follows:

$$e^{-x} = 1 - x + x^2/2! - x^3/3! + \ldots,$$

using just the $1 - x$ terms, we replace $e^{-(x-w)/p}$ with $1 - (x - w)/p$ in Eq. 3 and obtain

$$\begin{aligned} S_1(x; w) &= T e^{-(m-w)} \\ &\quad \times (e^{x-w} - p^2(1 - (x - w)/p) + p^2 - 1) \\ \text{or} \quad S_1(x; w) &= T e^{-(m-w)}(e^{x-w} + p(x - w) - 1) \\ &\quad \text{for } x \geq w \end{aligned} \tag{8}$$

In Bagwell's presentation (4), the corresponding equation (unnumbered) uses b for p and y for $x - w$. Figure 7 compares this function with the corresponding exponential, linear, and Logicle functions. At $x = w$, it has the same data value of zero and the same slope as the corresponding Logicle function. However, it does not fulfill our criterion that the second derivative should be zero at the data zero, so that near zero it departs from linearity more quickly than does the corresponding Logicle function. Also, at the high end, it approaches true log more slowly than the corresponding Logicle function. Therefore, we did not consider it further.

CONCLUSIONS

We have reached several conclusions regarding the effects of Logicle display on the quality of data interpretation and accuracy of statistical results:

1. Logicle display per se has no effect on statistical results, since these are computed on the underlying data—not on the position of displayed events in plots.
2. Similarly, use of Logicle displays cannot change the overlap (or lack thereof) of different cell populations.
3. In many cases, use of Logicle transformation will improve the validity of statistical results compared to data analysis software which truncates low and negative values outside displayed log or linear scale ranges and, therefore, cannot compute correct statistics for populations including such data values.
4. Logicle displays help to confirm correct compensation in that the visual centers of positive and negative populations in single stain compensation controls line up when compensation is correct. This is not true in logarithmic displays.
5. Logicle displays may lead to better selection of population boundaries (gates), and therefore improve validity of results. Logarithmic displays distort broad, low-mean populations, to give a peak above the true center of the distribution and pileup of low to negative events at the scale minimum (baseline). This can lead to improper, or at least suboptimal, gate boundary selection. Logicle dis-

plays avoid this tendency by being nearly linear in the region near zero.
6. Since Logicle scales go smoothly from linear to logarithmic, they do not introduce artifacts that might obscure real distinctions between populations or give the impression of population distinctions that are not real.
7. Logical transformed data is quite likely to be more suitable than plain log or linear scaling for automated analysis, such as peak finding and cluster analysis, since local distortions and edge pileups are avoided or at least minimized.
8. The methods described here for automatically selecting the Logicle width parameter to match particular data generally work well, but further work is needed in this area to provide more flexible user control of the transformation.

The Logicle scaling functions and Logicle display methods provide visualizations of flow cytometric data that are readily interpreted by viewers and convey full and accurate information regarding the underlying distributions of the data and patterns of expression.

ACKNOWLEDGMENTS

The authors acknowledge the help of Adam Treister of Tree Star Inc., Richard R. Hardy of the Fox Chase Cancer Center, Martin Bigos of the Gladstone Institute for Virology at the University of California, San Francisco and Joseph Trotter of BD Biosciences. They also thank John Mantovani and Leonore A. Herzenberg of Stanford University for help in producing the figures and the manuscript.

LITERATURE CITED

1. Parks DR, Bigos M. Analysis of flow cytometry data. In: Herzenberg L, Blackwell C, Weir D, editors. The Handbook of Experimental Immunology, 5th ed. Boston: Blackwell; 1996.
2. Loken MR, Parks DR, Herzenberg LA. Two-color immunofluorescence using a fluorescence-activated cell sorter. J Histochem Cytochem 1977; 25:899–907.
3. Roederer M. Spectral compensation for flow cytometry: Visualization artifacts, limitations, and caveats. Cytometry 2001;45:194–205.
4. Bagwell CB. Hyperlog—a flexible log-like transform for negative, zero, and positive valued data. Cytometry A, 2005;64:34–42.
5. Smith SW. The Scientist and Engineer's Guide to Digital Signal Processing, 2nd ed. San Diego, CA: California Technical Publishing; 1999. Chapter 22, p 362–364.
6. Durbin BP, Hardin JS, Hawkins DM, Rocke DM. A variance-stabilizing transformation for gene-expression microarray data. Bioinformatics 2002; 18 (Suppl. 1):S105–S110.
7. Huber W, von Heydebreck A, Sultmann H, Poustka A, Vingron M. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. Bioinformatics 2002;18 (Suppl. 1): S96–S104.
8. Munson P. J. "Consistency" Test for Determining the Significance of Gene Expression Changes on Replicate Samples and Two Convenient Variance-Stabilizing Transformations. Genelogic Workshop on Low Level Analysis of Affymetrix Genechip data. Bethesda, MD; 2001.