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

MIFlowCyt

Sometimes Simpler Is Better: VLog, a General but Easy-to-Implement Log-Like Transform for Cytometry

C. Bruce Bagwell,* Beth L. Hill, Donald J. Herbert, Chris M. Bray, Benjamin C. Hunsberger

Verity Software House, Topsham, ME

Received 4 June 2016; Revised 8 September 2016; Accepted 13 October 2016

Additional Supporting Information may be found in the online version of this article.

*Correspondence to: C. Bruce Bagwell, PO Box 247, Topsham, Maine 04086. E-mail: cbb@vsh.com

Conflicting Interest Disclosure: The authors, Bagwell, Hunsberger, Herbert, Bray, and Hill, are employed by Verity Software House, a manufacturer of cytometry software.

Published online 00 Month 2016 in Wiley Online Library (wileyonlinelibrary.com)

DOI: 10.1002/cyto.a.23017

Abstract

The fundamental purpose of log and log-like transforms for cytometry is to make measured population variabilities as uniform as possible. The long-standing success of the log transform was its ability to stabilize linearly increasing gain-dependent uncertainties and the success of the log-like transforms is that they extend this notion to include zero and negative measurement values. This study derives and examines a transform called VLog that stabilizes the three general sources of variability: (1) gain-dependent variability, (2) photo-electron counting error, and (3) signal-independent sources of error. Somewhat surprisingly, this transform has a closed-form solution and therefore is relatively simple to implement. By including some quantitation elements in its formulation, the shape-dependent arguments, α and β, usually do not require optimization for different datasets. The simplicity and generality of the transform may make it a useful tool for cytometry and possibly other technologies. © 2016 International Society for Advancement of Cytometry

Key terms

Hyperlog; Logicle; hyperbolic sine; biexponential; transformations; transforms; cytometry; variance stabilization

THE general topic of creating log-like transforms that admit zero or negative numbers has been well-explored in other scientific disciplines since 1949 (1–11). In 2002 Parks et al. were the first cytometrists to investigate the general form of the hyperbolic sine function as a potential solution to the problem (12) and then later published their "Logicle" transform in 2006 (13) and revised it slightly in 2012 (14). HyperLog is also a log-like transform that accepts zero or negative valued numbers and was published in 2005 (15). Both transform implementations are functions that tend to be linear through the origin and logarithmic away from the origin. Although there has been general acceptance of log-like transforms and their application to cytometry data, the detailed implementations are often not trivial, many times involving numerical root finding routines.

A detailed analysis of cytometric measurement sources of variance is also well-described in the literature (16–21). A signal in this context is a set of detected photons emitted from a specific molecular structure that is to be quantified as a measurement. There are three basic components of measurement variability: (1) gain-dependent, (2) photo-electron counting, and (3) signal-independent. Gain-dependent variability has the general characteristic that measurement uncertainty is proportional to the gain applied to a specific signal whether it be biologic or electronic in nature. The proportionality constant, coefficient of variation (cv), typically characterizes this type of variability. The variance of this type of variability increases with the square of measurement intensity.

ISAC
International Society for Advancement of Cytometry

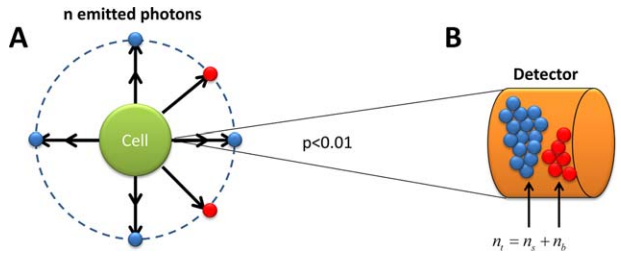

Figure 1. General Variance Formula: For fluorescence-based cytometers, while a cell is bathed in laser light, fluorochromes emit and re-emit fluorescence photons in all directions numerous times (10^3–10^9, see Panel A). A relatively small fraction of these photons (e.g., 0.01) ultimately are detected and typically produce a specific number of photo-electrons at the primary detector (see Panel B). Some of these generated photo-electrons are from the fluorochrome of interest (n_s, see blue spheres) and some are not (n_b, see red spheres). The nonsignal photo-electrons are generated from a wide variety of sources, where the major source is usually the particle's background fluorescence. [Color figure can be viewed in the online issue which is available at wileyonlinelibrary.com]

For fluorescence-based cytometers, photons that are detected and converted to photo-electrons by detectors such as photo-multiplier tubes (pmts) have a Poisson-distributed counting error. The variance of this type of variability increases linearly with measurement intensity.

The third source of measurement variability is from a number of sources. Some of this variability is due to the counting error associated with nonsignal photo-electrons, which includes sources such as ambient light, light-scatter, Raman scattering, and other signals inappropriately detected. However, the major sources of signal-independent variability are due to cellular autofluorescence and the variability associated with base-line restore peak-detection algorithms. The variance of this type of variability is constant and does not vary with measurement intensity.

The purpose of this study is to present and examine a data transform that was designed to stabilize the variability from all three sources. This transform, VLog, not only has efficient closed-form solutions, but also can be enhanced such that its parameters rarely have to be recomputed with different datasets.

Material and Methods

Mathematical Data Simulations

Mathematical analysis and presentation graphics were done using Mathcad 15.0, Parametric Technology Corporation (PTC), Needham, MA.

Normally distributed random numbers were generated with the Box–Muller equation,

$$\text{Norm}(\mu,\delta)=\mu+\delta\sqrt{-2\ln(rnd(1))}\cos(2\pi\cdot rnd(1)).$$

Gain-Dependent Variances for Synthesized Populations:

The first four gain-dependent populations were synthesized as shown below,

$$X_{i,g}=\text{Norm}(\mu,\delta)\cdot G_g,$$
$$\text{where, } \mu=50,\ \delta=5,\ n=1,000,$$
$$i=1,2,\ldots,n;\ g=1,2,\ldots,4,$$
$$G_g=g.$$

Notice that the gains have relatively low amplitude (1, 2, . . ., 4). The second four gain-dependent populations were synthesized as,

$$X_{i,4+g}=\text{Norm}(\mu,\delta)\cdot G_g,$$
$$\text{where, } g=1,2,\ldots,4,$$
$$G_g=10^g.$$

Note that these gains have high amplitudes (10^1, 10^2, . . ., 10^4). These data were used in Figure 1 to demonstrate the general utility of the log transform.

General Variances for Synthesized Populations:

The eight populations shown in Figure 3 were synthesized as shown below,

$$X2_{i,g}=\text{Norm}\left(M_g,\delta\left(M_g,cv,q_c,b\right)\right),$$
$$\text{where, } \delta(x;cv,q_c,b)=\sqrt{cv^2x^2+\frac{x}{q_c}+b^2},$$
$$g=1,2,\ldots,8,$$
$$M^T=[10\ \ 50\ \ 150\ \ 400\ \ 1000\ \ 2800\ \ 8000\ \ 17000],$$
$$cv=0.03,\ q_c=0.6,\ n=1,000,\ b=6.0,\ i=1,2,\ldots,n.$$

These data were used in Figure 3 to demonstrate the incompleteness of the Log Transform and the efficacy of the VLog Transform.

DataSets

The files represented in Figure 5 mainly come from a repository of files from a published study (22) and from Helios data described in another presentation (23).

Statistics

The formulae for the statistics mean, μ, variance, σ^2, standard deviation, σ, and coefficient of variation, cv, are shown below:

$$\mu=\frac{\sum_{i=1}^{n}x_i}{n},\ \delta^2=\frac{\sum_{i=1}^{n}(x_i-\text{mean})^2}{n-1},\quad \delta=\sqrt{\delta^2},\quad cv=\frac{\delta}{\mu}$$

Nomenclature and Abbreviations:

Typically variance is written as σ^2 and the variance function as Var. For mathematical brevity, the variance function will be shortened to capital *V*.

ADC: Analog to Digital Converter.

Derivation of the General Variance Formula:

For fluorescence-based cytometers, while a cell is bathed in laser light, fluorochromes emit and re-emit fluorescence photons in all directions numerous times (10^3–10^9, see Fig. 1, Panel A). A relatively small fraction of these photons (e.g., 0.01) ultimately are detected and typically produce a specific number of photo-electrons at the primary detector (see Panel

B). Some of these generated photo-electrons are from the fluorochrome of interest (n_s, see blue spheres in Panel B) and some are not (n_b, see red spheres in Panel B). These nonsignal photons can be grouped together and symbolized as n_b.

The variance of the total number of generated photon-electrons, V_c, is predicted by the Poisson Distribution as equal to $n_s + n_b$ and is often referred to as counting error,

$$V_c(n_t) = n_s + n_b. \tag{1}$$

These photo-electrons will eventually be amplified and digitized to form a measurement value in ADC units. If this process is linear, then there will be a proportionality constant, q_x, that converts number of generated photo-electrons to x ADC units.

$$q_x = \frac{\text{generated photo-electrons}}{\text{ADC unit}}, \tag{2}$$

Therefore, the equation for x ADC units is given as,

$$x = \frac{n_t}{q_x}. \tag{3}$$

Substituting Eq. (3) into Eq. (1) yields the variance for counting error,

$$V_c(x) = V_c\left(\frac{1}{q_x}\cdot n_t\right) = \frac{1}{q_x^2}V_c(n_t) = \frac{n_t}{q_x^2} = \frac{q_x x}{{q_x}^2} = \frac{x}{q_x}. \tag{4}$$

Note that the variance of n_t times the constant $1/q_x$ is $1/q_x$ squared times the variance of n_t. There are also a number of signal-independent sources of variability that can be grouped together as $V_b = b^2$.

However, the major source of variability in measurement systems like cytometry are gain-dependent (biological and electronic) and are generally characterized by the coefficient of variation or cv. The variance due to cv, V_{cv}, is approximately,

$$V_{cv}(x) \simeq cv^2x^2. \tag{5}$$

Thus, the total variance for the measurement value x is given as,

$$V_x(x) = V_{cv}(x) + V_c(x) + V_b = cv^2x^2 + \frac{x}{q_x} + b^2, \tag{6}$$

and the standard deviation function is the square-root of the variance or,

$$\delta(x; cv, q_x, b) = \sqrt{cv^2x^2 + \frac{x}{q_x} + b^2}. \tag{7}$$

Note that the cv variance increases with the square of x and therefore is the dominant variance, the count-dependent variance increases linearly with x, and the background variance is independent of x. The basic assumption made with Eqs. (6) and (7) is that the three types of variabilities are independent of each other and therefore the variances can be added together.

Results

Log Base Transform

The longstanding utility of the log transform is primarily due to its ability to stabilize a set of population gain-dependent variabilities. The increased dynamic range associated with the transform is really a useful side-effect of this stabilizing capability. If the overall measurement variability were only determined by the equation,

$$\delta(x; cv) = cv \cdot x.$$

$$\text{where, } \delta = \text{standard deviation}, \tag{8}$$

$$cv = \text{coefficient of variation}.$$

then, the log function would be the transform of choice since,

$$z(x; cv) = \int_1^x \frac{1}{cv \cdot t}\,dt = \frac{1}{cv}\cdot \ln(x), x > 0. \tag{9}$$

When multiple populations are synthesized over a wide dynamic range with standard deviations, sds, determined solely by Eq. (8) (see Gain-dependent populations in M&M), linear transforms are usually not adequate for complete visual inspection of all populations. Figure 2, Panels A and B show the eight population linear data in both dot-plot and histogram formats where only two or perhaps three populations can be visualized. The table in Panel C enumerates the increasing standard deviations for each of the populations.

Panels D and E show the log transformed data using Eq. (9). Each of the eight populations are now easily distinguished because the transformed sds are relatively uniform with values very close to unity (see Panel F). Transforms that convert standard deviation functions such as shown in Eq. (8) to a set of transformed standard deviations near unity will be referred to as base transforms.

The VLog Transforms

In order to account for the variabilities described by Eq. (7), we need to find a function with slopes equal to its inverse. This is done quite easily by simple integration,

$$\int \frac{1}{\sqrt{cv^2x^2 + \frac{x}{q_x} + b^2}}\,dx = \frac{1}{cv}\cdot \ln\left(\sqrt{cv^2x^2 + \frac{x}{q_x} + b^2} + cv\cdot x + \frac{1}{2cv\cdot q_x}\right) = z. \tag{10}$$

As shown above, the solution can be written in closed-form. The independent variable is the measurement value, x, and the dependent variable is the transformed value, z. The integral formula is then reparameterized to eliminate cv from the ln function argument, translocated to $z = 0$ at $x = 0$, and then made symmetric about the $z = 0$ axis yielding the base VLog transform and inverse transform,

$$z(x; cv, \alpha, \beta) = \frac{\text{sign}(x)}{cv}\cdot \ln\left(\frac{\sqrt{x^2 + \alpha|x| + \beta^2} + |x| + \frac{\alpha}{2}}{\frac{\alpha}{2} + \beta}\right), \tag{11}$$

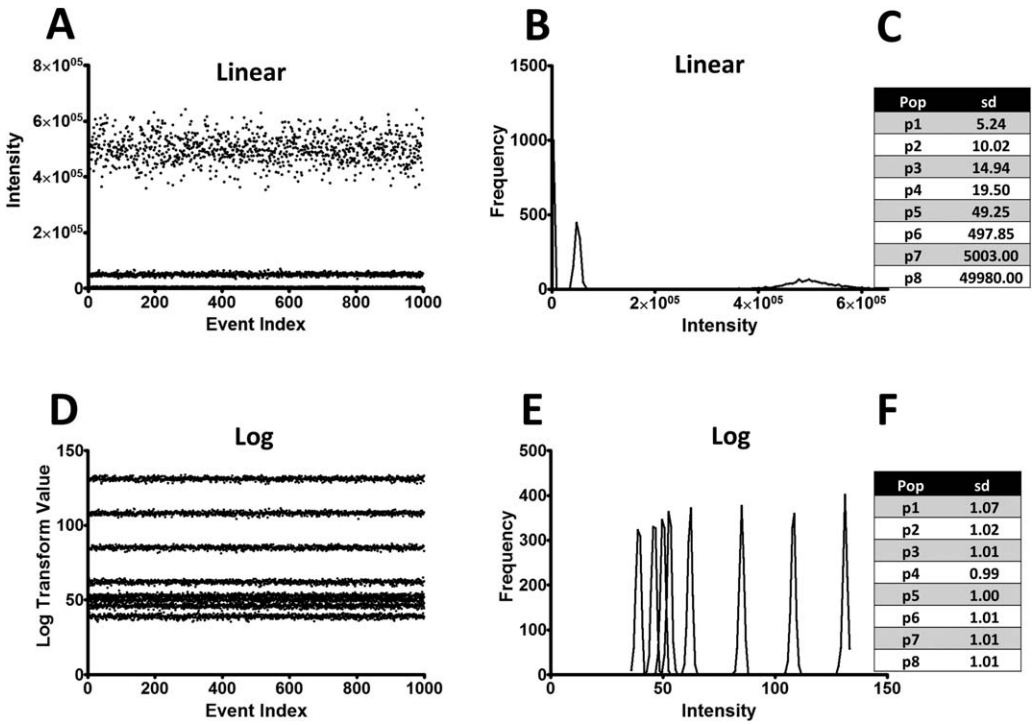

Figure 2. Log Transform: Panels A, B, and C show the linear data from populations with standard deviations (sds) determined solely by Eq. (8) (see Gain-dependent populations in M&M). Panel A shows each event's intensity value plotted against event index and Panel B shows the associated frequency histograms. Because of the linear increasing sds over a large dynamic range (see Panel C) only two populations are adequately represented in the linear domain. Panels D, E, and F show the Log transformed data using Eq. (9). Each transformed sd is relatively uniform with values very close to unity (see Panel F).

$$x(z; cv, \alpha, \beta) = \text{sign}(z) \cdot \alpha \cdot \sinh\left(\frac{cv \cdot z}{2}\right)^2 + \beta \cdot \sinh(cv \cdot z), \tag{12}$$

$$\text{where, } \alpha = \frac{1}{q_x cv^2}, \quad \beta = \frac{b}{cv}.$$

The transform parameters are cv, α, and β. Interestingly, when $\alpha = 0$, the transform behaves as the popular hyperbolic sine type of transform (see comparison with other transforms for details).

When multiple populations are distributed over a wide dynamic range with standard deviations determined solely by Eq. (7) (see general variances for synthesized populations in M&M), log transforms incompletely stabilize all the variances and as a result, the sds are higher for the lower-intensity populations (see Fig. 3, Panels A, B, and C). However, if the data are converted with the VLog base transform, the variabilities are relatively constant and have sds of near unity (see Panels D, E, and F). This characteristic also has importance for optimally determining cv, α, and β from different datasets (see Appendix Section).

VLog Normalization

Equations (11) and (12) can be normalized such that at $x = x_{\max}$, the transform is at $z = z_{\max}$ and at $x = -x_{\max}$, $z = -z_{\max}$.

$$t_{\max} = \ln\left(\frac{\sqrt{x_{\max}{}^2 + \alpha x_{\max} + \beta^2} + x_{\max} + \frac{\alpha}{2}}{\frac{\alpha}{2} + \beta}\right), \tag{13}$$

$$T_{\text{VLog}}(x; \alpha, \beta, x_{\max}, z_{\max}) = \text{sign}(x) \cdot \frac{z_{\max}}{t_{\max}} \cdot \ln\left(\frac{\sqrt{x^2 + \alpha|x| + \beta^2} + |x| + \frac{\alpha}{2}}{\frac{\alpha}{2} + \beta}\right), \tag{14}$$

$$T^{-1}_{\text{VLog}}(z; \alpha, \beta, x_{\max}, z_{\max}) = \text{sign}(z) \cdot \alpha \cdot sinh\left(\frac{z}{2} \cdot \frac{t_{\max}}{z_{\max}}\right)^2 + \beta \cdot sinh\left(z \cdot \frac{t_{\max}}{z_{\max}}\right). \tag{15}$$

The variable $t_{\max}$ only needs to be evaluated once for a specific set of parameters.

Quantitation Linear Transform

The only changes to the above formulae to support quantitation are the additions of a slope, q_s, and intercept, q_i, that convert relative intensity units, x, to absolute quantitation units, q.

$$q = q_s \cdot x + q_i, \tag{16}$$

$$T_{\text{VLog}}(q; \alpha, \beta, q_{\max}, z_{\max}), \tag{17}$$

$$T^{-1}_{\text{VLog}}(z; \alpha, \beta, q_{\max}, z_{\max}). \tag{18}$$

Although quantitation is not yet mainstream in cytometry, there is one practical utility for this type of extra transformation. As ADC resolutions of cytometers continue to increase, their relatively large magnitude is creating some display issues

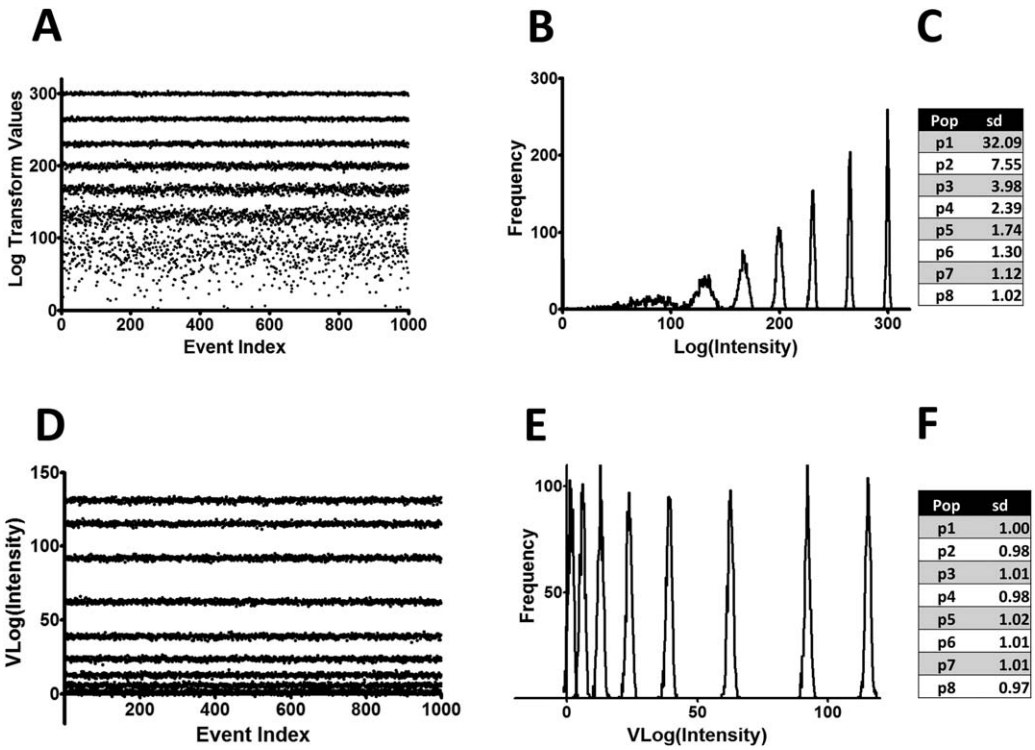

Figure 3. Log and VLog Transforms: Panels A, B, and C show the Log Transform of the linear data from populations with standard deviations (sds) determined by the general variance formula (see General Variances for Synthesized Populations in M&M). Panel A shows each event's Log transformed intensity value plotted against event index and Panel B shows the associated frequency histograms. Because the Log Transform only stabilizes the gain-dependent portion of the total variance, the sds increase with lower intensity values (see Panel C for enumerated sds). Panels D, E, and F show the Log transformed data using the VLog Transform. Each transformed sd is relatively uniform with values very close to unity (see Panel F).

for cytometrists. Some cytometers now have ADC max ranges of $>10^9$. It makes little sense to use nine decades for an axis if there is only a signal range in the data of four to five decades (see Fig. 4, Panel A). This issue can be minimized by setting the VLog's β coefficient to the background peak's location (see Fig. 4, Panel B). However, this approach will necessitate changing β for data that has different maximum ADC ranges. A better approach is to approximate the number of decades, d, implicit in the data and use the above q_s and q_i slope and intercepts as shown below,

$$q_{\max} = 10^d,\ q_s = \frac{q_{\max}}{x_{\max}},\ q_i = 0. \tag{19}$$

Also, note that the α and β parameters for VLog should be determined relative to the q quantitation domain rather than the x ADC domain. This simple procedure results in transforms that rarely need optimization even when analyzing data from cytometers that have very different ADC max ranges. Figure 4, Panel C shows the same data with number of decades, d, equal to five and $\alpha = 0$ and $\beta = 1$. In this display there is little or no wasted space by inappropriate low-level decades. Other types of transforms such as HyperLog also can benefit from allowing the number of decades to be adjusted (see Panel D). However, when counting error variability is a major source of variability as it is in this example, the α parameter can be adjusted to stabilize it (see Panels E and F, and Appendix for details).

Linear light-scatter axes for peripheral blood lymphocytes, monocytes, and granulocytes also result in increasing variabilities with increasing intensities, especially for side-scatter. Figure 5, Panel A shows a typical dot-plot of peripheral blood with linear transforms for FSC and SSC. Often by ensuring that the lymphocytes are on scale, a significant fraction of granulocytes end up being off scale for SSC (see white arrow). Panel B shows linear FSC and SSC where the granulocytes are largely on scale. A better strategy for displaying light-scatter dot-plots is to limit the number of decades to 1.5 and to use log-like transforms such as VLog. Panel C shows the same data but with VLog ($\beta = 10,\ \alpha = 0$) for FSC and VLog ($\beta = 1,\ \alpha = 0$) for SSC. If the granulocytes are on scale, the lymphocytes will still be well-defined. Even when the ADC resolutions change dramatically, the light-scatter patterns will remain very similar (not shown).

Typical cytometry markers have effective dynamic ranges between four and five decades. The second and third rows in Figure 5 show a number of examples where it was not necessary to adjust the transforms even between dramatically different cytometers like the Helios and BD FACS Diva. Panels D, E, and F show CCR7 vs. CD45RA, CD28 vs. CD45RA, and CD28 vs. CCR7 for the Helios mass cytometer and Panels G, H, and I show the same marker combinations from a BD FACS Diva fluorescence cytometer. CD8 subpopulations such

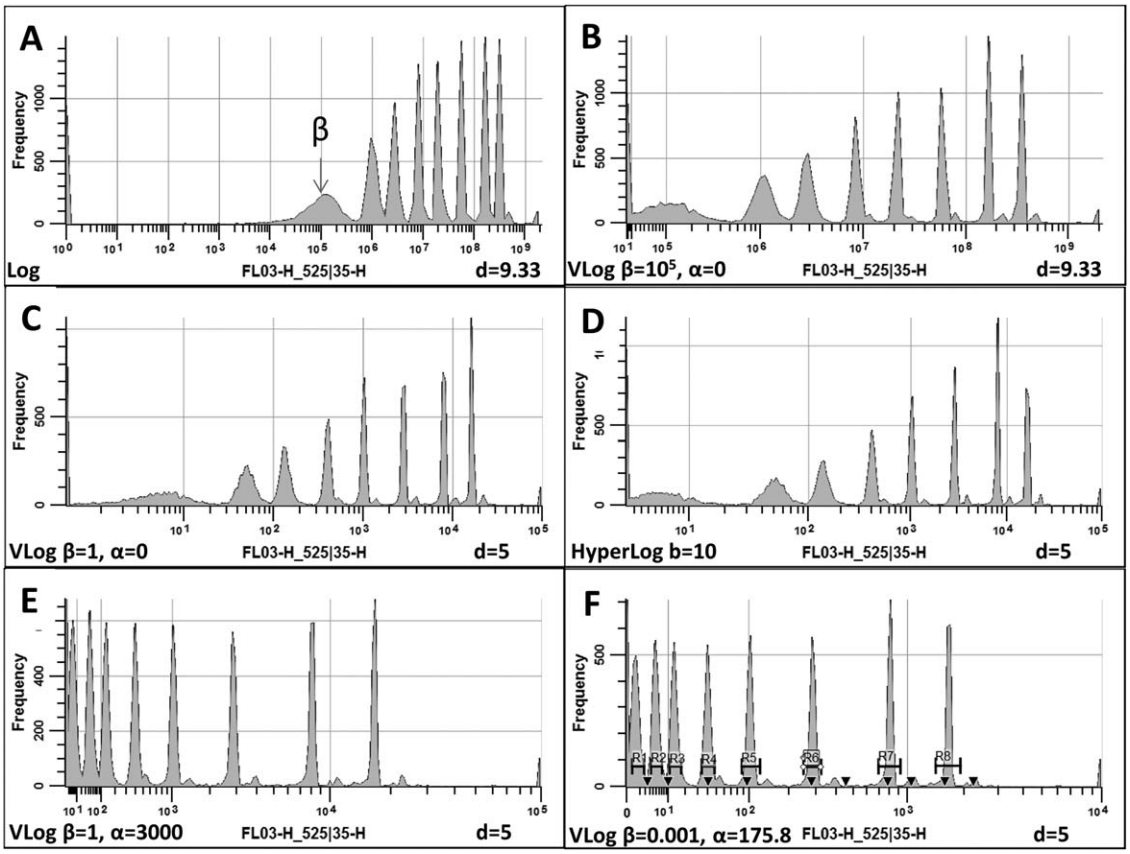

Figure 4. Transform Comparisons: Panel A shows calibration beads on a Yeti cytometer with over nine decades of ADC resolution. Notice the wasted space and the increasing uncertainties with lower-intensity populations when presented with a nine-decade Log transform display. Panel B shows the same data with a VLog transform where β is set to the background peak's approximate intensity value (10^5, see arrow in Panel A) and α is set to zero. Panel C uses the VLog transform with number of decades, d, set to 5 and $\beta = 1$ and $\alpha = 0$. The advantage of setting the number of decades to 5 is that this data will look very similar on cytometers with very different ADC max ranges. Panel D shows the HyperLog transform with similar settings, $b = 10$. Panel E shows VLog $\beta = 1$ and a manually set $\alpha = 3{,}000$ and Panel F shows the optimized VLog parameters, $\beta = 0.001$ and $\alpha = 175.85$ (see Appendix for details).

as Naïve (blue), central memory (green), and effector memory (red) are well-defined without the need of adjusting any of the transform's parameters.

Comparison with Other Transforms

As aforementioned, if the measurement variability were described by,

$$\delta(x; cv) = cv \cdot x,$$

then, the base transform that would stabilize the variabilities would be given as,

$$z(x; cv) = \frac{1}{cv} \cdot \ln(x), x > 0.$$

If this measurement variability is augmented to include a signal-independent background term,

$$\delta(x; cv, b) = cv \cdot x + b, \quad (20)$$

then the base transform is given as,

$$z(x; cv, b) = \int \frac{1}{cv \cdot x + b} dx = \ln(cv \cdot x + b), x > \frac{-b}{cv}. \quad (21)$$

Although this transform can be used for cytometry data, the transforms that follow are more popular.

If the measurement variability is slightly better posed as,

$$\delta(x; cv, b) = \sqrt{cv^2 \cdot x^2 + b^2}, \quad (22)$$

then the base transform is given as,

$$z(x; cv, b) = \int \frac{1}{\sqrt{cv^2 x^2 + b^2}} dx = \frac{1}{cv} \ln\left(\sqrt{cv^2 x^2 + b^2} + cv \cdot x\right). \quad (23)$$

This transform is also known as the GLog transform (11) which is often presented as,

$$\mathrm{GLog}(x, \lambda) = \log\left(x + \sqrt{x^2 + \lambda}\right).$$

The reason this transform is not routinely used in cytometry is probably because it is asymmetric and is not zero when x is zero. If Eq. (23) is translocated to the origin ($z = 0$, $x = 0$) and symmetrized, it becomes,

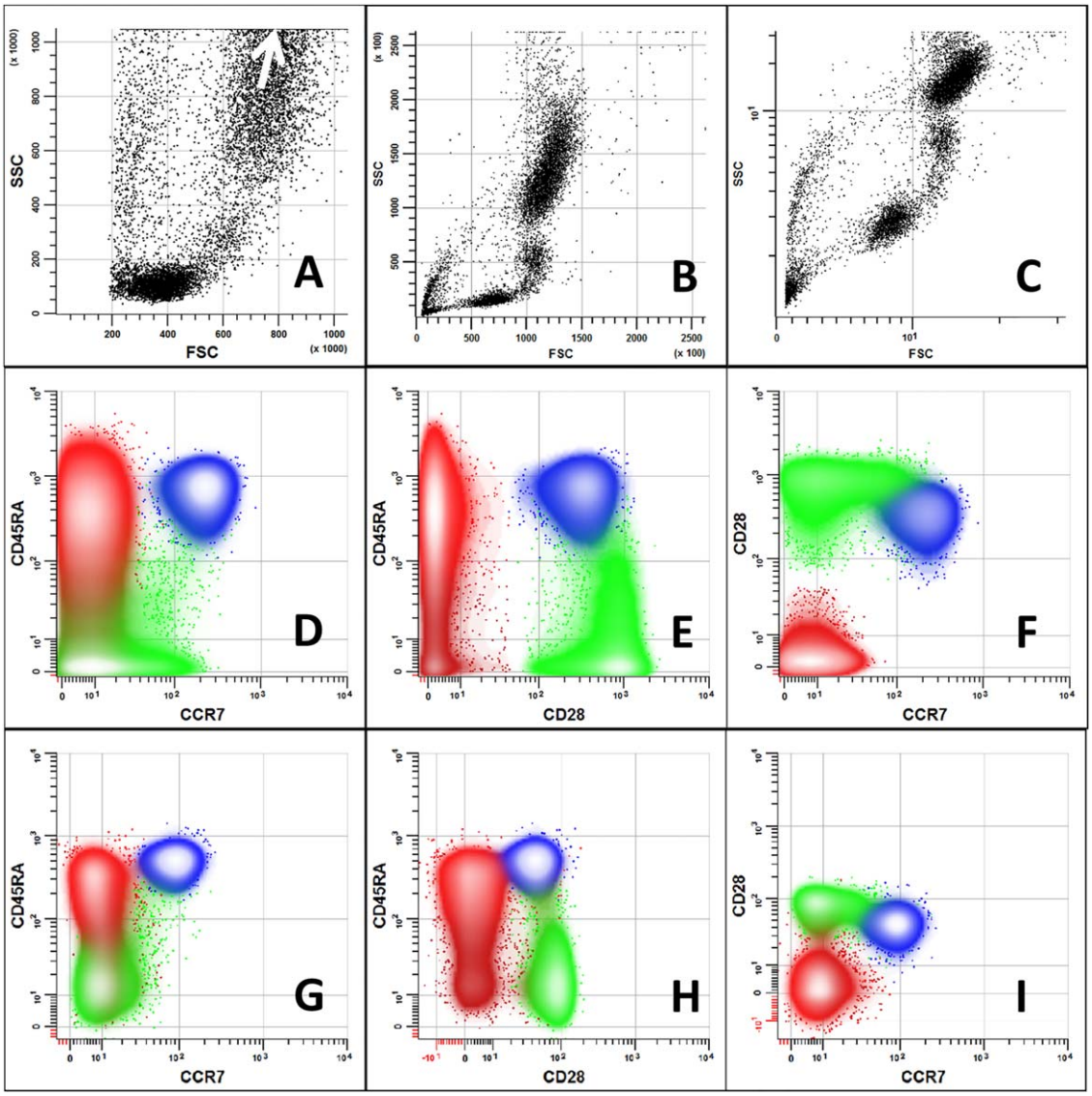

Figure 5. Typical Examples: Panel A shows a typical dot-plot of peripheral blood with linear transforms for FSC and SSC. Often by ensuring that the lymphocytes are on scale, a significant fraction of granulocytes end up being off scale for linear SSC (see white arrow). Panel B shows linear FSC and SSC where the granulocytes are largely on scale. A better strategy for displaying light-scatter dot-plots is to limit the number of decades to 1.5 and to use log-like transforms such as VLog. Panel C shows the same data but with VLog ($\beta = 10$, $\alpha = 0$) for FSC and VLog ($\beta = 1$, $\alpha = 0$) for SSC. If the granulocytes are on scale, the lymphocytes will still be well-defined. Even when the ADC resolutions change dramatically, the light-scatter patterns will remain very similar (not shown). Panels D, E, and F show CCR7 vs CD45RA, CD28 vs CD45RA, and CD28 vs CCR7 for the Helios mass cytometer. The same marker combinations in Panels G, H, and I are from a BD FACS Diva fluorescence cytometer. All transforms were set up identically (VLog β=10, α=0) even though these cytometers have very different characteristics and ADC maximum ranges. [Color figure can be viewed in the online issue which is available at wileyonlinelibrary.com]

$$z(x; cv, b) = \int \frac{1}{\sqrt{cv^2x^2 + b^2}} dx = \frac{1}{cv} \ln\left(\frac{\sqrt{cv^2x^2 + b^2} + cv \cdot |x|}{b}\right). \quad (24)$$

The inverse of this equation is given as,

$$x(z; cv, b) = \frac{b}{2cv}\left(e^{cv \cdot z} - e^{-cv \cdot z}\right) = \frac{b}{cv}\sinh(cv \cdot z), \quad (25)$$

This popular cytometry transform is known as the bi-exponential or hyperbolic sine transform. When this transform is generalized to,

$$x(z; a, b, c, d, f) = ae^{b \cdot z} - ce^{d \cdot z} + f, \quad (26)$$

it is known as the "Logicle" transform (13). Although this transform is highly flexible, it has no closed-form inverse and selecting optimal parameters is nontrivial. The HyperLog transform (see Fig. 4 Panel D) was introduced in 2005 (15) to reduce the complexity of cytometry log-like transforms by only having two parameters, a and b, to adjust,

$$x(z; a, b) = sign(z) \cdot \left(e^{a \cdot |z|} + b \cdot |z| - 1\right). \quad (27)$$

However, this simple function also does not have an inverse in closed-form which necessitated using root finding algorithms and tables for its implementation. Subsequent to the publication it was found that by leveraging Lambert *W* functions (24) the inverse could be put into a form that did not require any root-finding tables,

$$z(x;a,b)=sign(x)\cdot\frac{1}{a}\ln\left(\frac{b}{a}\cdot LambertW\left(\frac{a}{b}\cdot e^{\frac{a}{b}(|x|+1)}\right)\right) \quad (28)$$

Validation Data

Table 1 provides some validation data for those that are interested in implementing the transform.

Table 1. Validation equations

FORMULAE	VALUES
$T_{\mathrm{VLog}}(10;100,200,10000,100)$	1.132206
$T_{\mathrm{VLog}}(100;100,200,10000,100)$	10.423627
$T_{\mathrm{VLog}}(-100;100,200,10000,100)$	−10.423627
$T^{-1}_{\mathrm{VLog}}(1;100,200,10000,100)$	8.825153
$T^{-1}_{\mathrm{VLog}}(10;100,200,10000,100)$	95.473277
$T^{-1}_{\mathrm{VLog}}(-10;100,200,10000,100)$	−95.473277

Discussion

Log and Log-Like transforms for cytometry are variable-sloped functions that tend to stabilize variances to better enable population visualization and analysis. Most all these variances are described by a general variance formula [see Eq. (6)] that blends gain-dependent, counting-error, and signal-independent variabilities. The Log transform partially stabilizes the variances since it only accounts for gain-dependent variabilities, which is why cytometrically derived population variances tend to increase with lower intensities (see Fig. 3, Panel B and Fig. 4, Panel A). Also, because the Log transform is not defined for zero and less than zero values, serious distortions can occur near the origin that can lead to inappropriate conclusions about the true nature of the data. It should be noted that different populations normally have different cvs. The synthesized datasets were contrived to demonstrate how these transforms theoretically behave with their assumed sources of variance.

By numerically integrating the general variability formula's reciprocal [see Eq. (10)], a transform called VLog can be derived and mathematically represented in efficient closed-form equations [see Eqs. (11) and (12)]. Enhancing the transform to include the capability of supporting quantitative axes also solves an increasingly problematic issue with cytometry displays. As the ADC maximum ranges have increased over time, many cytometer display systems have also increased the number of decades displayed on their axes. Rather than considering the ADC max range as the source for number of decades, the quantitative system enables the user to approximate the number of measureable decades of information encoded in the data. For immunofluorescence measurements, typically four to five decades are all that are needed; however, for light-scatter measurements only 1.5 to 2 decades are normally required. By matching the number of decades to the biology rather than to the electronics, different cytometers with different ADC max ranges can be made to produce similar distributions. Figure 5 amplifies this point by demonstrating that different markers and different types of cytometers can produce very similar data patterns with exactly the same VLog or other transform's controlling parameters.

Other strategies have been adopted to deal with the high number of decades issue such as top four-decades and data zooming. Although these methods can show similar patterns to those in Figure 5, they are not equivalent. Data zooming regions will still need adjusting if ADC maximum ranges change. Also, they do not properly represent events at the extreme ends of the measurement scale. For example, if the top five decades were displayed to a user in Figure 4 Panel A, they would not be aware of the significant number of events pegged at the true origin of the axis.

One major limitation to VLog and the other mentioned transforms is the assumption that sources of error are independent of each other. This assumption is indeed an approximation and certainly not true when there is significant signal crossover between detectors. However, it is a simplification that is currently necessary in order to keep the equations from becoming overly complex.

If counting error variability is a significant source of variability in data as it is for Figure 3, Panel B and Figure 4, Panels A–D, then the VLog α parameter can be either adjusted manually (see Fig. 4, Panel E) or automatically (see Panel F) to make the affected population variances more uniform.

The relative simplicity of the VLog equations along with their generality potentially makes them a useful tool for cytometry and possibly other technologies. Software engineers interested in exploring VLog's capabilities are free to do so. Validation results have been added in Table 1 to help with these implementations.

Acknowledgments

The authors are grateful to Margaret Inokuma at BD Biosciences for generously sending the authors some of the files presented.

Literature Cited

1. Johnson NL. Systems of frequency curves generated by methods of translation. Biometrika 1949;36:149–176.
2. Bickle PJ, Doksum KA. An analysis of transformations revisited. J Am Stat Assoc 1981;76:296–311.
3. Box GEP, Cox DR. An analysis of transformations. J Royal Stat Soc 1964;26:211–243.
4. Burbidge JB, Magree L, Robb AL. Alternative transformations to handle extreme values of the dependent variable. J Am Stat Assoc 1988;83:123–127.
5. Layton DF. Alternative approaches for modeling concave willingness to pay functions in conjoint valuation. Am J Agr Econ 2001;83:1314–1320.
6. Guarnieri MD, Ortolani S, Montegriffo P, Renzini A, Barbuy B, Bica E, Moneti A. Infrared array photometry of bulge globular clusters. Astron Astrophys 1998;331:70–80.
7. Munson PA. '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 2001.
8. Tukey JW. On the comparative anatomy of transformations. Ann Math Stat 1964;28: 602–632.
9. Tukey JW. Exploratory Data Analysis. Reading, MA: Addison-Wesley; 1977.
10. Holder D, Raubertas RF, Pikounis VB, Svetnik V, Soper K. Statistical analysis of high density oligonucleotide arrays: a safer approach. GeneLogic Workshop on Low Level Analysis of Affymetrix GeneChip Data; 2001.

11. Rocke DM, Durbin B. Approximate variance-stabilizing transformations for gene-expression microarray data. Bioinformatics 2003;19:966–972.
12. Parks DR, Moore W. Cytometry Development Workshop; Pacific Grove, California; 18–21 October 2002.
13. Parks DR, Roederer M, Moore WA. A new "Logicle" display method avoids deceptive effects of logarithmic scaling for low signals and compensated data. Cytometry A 2006;69:541–551.
14. Moore WA, Parks DR. Update for the logicle data scale including operational code implementations. Cytometry A 2012;81A:273–277.
15. Bagwell C. Hyperlog-a flexible log-like transform for negative, zero, and positive valued data. Cytometry A 2005;64A:34–42.
16. Steen HB. Noise, sensitivity, and resolution of flow cytometers. Cytometry 1992;13: 822–830.
17. Wood JC, Hoffman RA. Evaluating fluorescence sensitivity on flow cytometers: An overview. Cytometry 1998;33:256–259.
18. Hoffman RA, Wood JC. Characterization of Flow Cytometer Instrument Sensitivity. Current Protocols in Cytometry, 1.20.1–1.20.18. Wiley; 2007.
19. Wood JC. Flow cytometer performance: Fluorochrome dependent sensitivity and instrument configuration. Cytometry 1995;22:331–332.
20. Wood JC. Fundamental flow cytometer properties governing sensitivity and resolution. Cytometry 1998;33:260–266.
21. Perfetto SP, Chattopadhyay PK, Wood J, Nguyen R, Ambrozak D, Hill JP, Roederer M. Q and B values are critical measurements required for inter-instrument standardization and development of multicolor flow cytometry staining panels. Cytometry A 2014;85A:1037–1048.
22. Inokuma MS, Maino VC, Bagwell CB. Probability state modeling of memory CD8(+) T-cell differentiation. J Immunol Methods 2013;397:8–17.
23. Bagwell CB, Leipold M, Maecker H, Stelzer G. High-dimensional modeling of peripheral blood mononuclear cells from a Helios Instrument. Seattle, Washington: Washington State Convention Center; 2016.
24. Corless R, Gonnet G, Hare D, Jeffrey D, Knuth D. On the Lambert W function. Adv Comput Math 1996;5:329–359.
25. Press W, Teukolsky S, Vetterling W, Flannery B. Numerical Recipes in C. The Art of Scientific Computing, 2nd ed. Cambridge University Press; 1992. pp 683–688.

Appendix

This section describes how to set VLog's parameters, α and β, manually and automatically. Using the VLog transform along with the quantitation mode where the number of decades is set to reasonable values such as 4 to 5 is quite simple. Usually setting α to 0 and β to 1 or 10 is all that is required for most files (see Fig. 4, Panel C and Fig. 5, Panels B–I). If you do wish to adjust these values manually, the β value corresponds to the approximate location of the background peak and α value can be manually increased to eliminate any low intensity peak counting error effects (see Fig. 4, Panel E), although normally this is not necessary.

Software developers who wish to optimize these values can use the approach discussed below. This approach involves first identifying the background population peak and at least two positive peaks. Figure 4, Panel F shows an example where all eight populations were initially identified with ranges on unoptimized transformed data such as shown in Figure 4, Panel C. Once the peaks are identified, the untransformed means and variances of each of the peaks are enumerated (see Table A1, columns 1 and 2). The variance equation shown in Eq. (6) can be used to estimate the cv, α, and β parameters by performing a least-squares analysis with the quadratic equation,

$$V_i = S_2 X_{i,2} + S_1 X_{i,1} + S_0 X_{i,0}, \quad \text{where,} \quad X_{i,2} = \mu_i^2,\ X_{i,1} = \mu_i,\ X_{i,0} = 1. \tag{A1}$$

The vectors, V_i and μ_i, are the observed variances and means for the peak data shown in Table A1. The solution vector, S is found by solving the least-square matrix formula,

$$S = (X^T X)^{-1} X^T V, \quad cv = \sqrt{S_2},\ \alpha = \frac{S_1}{cv^2},\ \beta = \sqrt{\frac{|S_0|}{cv}}. \tag{A2}$$

Table A1. Optimizing VLog parameters

Peak(i)	Mean(μ_i)	Var(V_i)	SD(σ_i)	SDZ(σz_i)
1	0.65	0.12	0.35	1.06
2	5.17	0.77	0.88	0.97
3	13.87	2.21	1.49	0.97
4	41.34	7.63	2.76	0.98
5	104.45	27.22	5.22	1.03
6	289.30	117.14	10.68	0.99
7	813.31	669.42	25.87	0.99
8	1,670.11	2,520.36	50.20	0.99

μ_i, V_i, and σ_i are the ith peak mean, variance, and standard deviation in quantitation units from eight peaks, $i = 1..8$, in Figure 4, Panel C. σz_i are the resultant transformed standard deviations that are near unity as described in Appendix.

The least-squares solution is fairly accurate except for the intercept, S_0. Because the peak mean data are not spaced uniformly, there is a tendency for the solution to over-determine the cv estimate at the sacrifice of the intercept, β. The interesting solution to this problem is to use the base VLog transform, Eq. (11), in an iterative nonlinear least-square analysis (25) to find an optimal set of parameters; cv, α, and β, that results in a set of transformed sds that are close to unity. If we use the least-squares solution as initial estimates and we let the function, SDZ(i;cv,α,β), return the ith peak's transformed sd given these parameters, then the objective function to minimize is given as,

$$\Phi(cv, \alpha, \beta) = \frac{\sum_{i=1}^{nPeaks} (SDZ(i; cv, \alpha, \beta) - 1)^2}{nPeaks}. \tag{A3}$$

The data in the Table A1's last column show the final transformed sds after this minimization process and Figure 4 Panel F shows the result of the transform. This procedure is not normally necessary, but might be valuable if accurate estimates of cv, α, q_x, or β are desired.