Supplementary File 1: Data, Software, and Methods

We have used both simulated data and real data in our study. In order to study the behavior of different transformation methods, we have developed an FCS data simulator to write values uniformly selected from 2^{-32} to 2^{32} into binary FCS format. The simulated data was fed into the transformation system being tested. By observing the output, we can plot the curve of an transformation function and identify its problems and limitations. It also allows us to solve the equations to identify the transformation parameters used in the system.

The real FCS3.0 files used in our experiments were provided by four independent labs from three institutions (See our Acknowledgement in main text). The BD FCS files were generated with BD LSR-II flow cytometers while the Accuri FCS files with Accuri C6 cytometers. All FCS files we collected are not gated or preprocessed. Specifically, the three BD FCS3.0 files with transformation results shown in this paper are: FCS3example.fcs (90745 events with 11 channels), Abcam.fcs (249421 events with 6 channels), and s1986.fcs (1159355 events with 12 channels). The three FCS file headers containing FCS3.0 keywords (with operator names and original file names hidden) can be found in Supplementary File 2. Names and versions of the transformation systems we have tested include: FlowJo 8.8.6 [1], FCS2CSV 1.0 [2], flowTrans 1.4.0 [3], and flowCore 1.18.0 [4]. The software running environment is: Bioconductor 2.8 in R 2.13.1 (for R programs) with MacOSX 10.6.8.

If we regard the transformation of FCS data as a function f, which maps a raw value x to a transformed value $f(x)$ that can be plotted on the user's screen, we can describe the transformation function in *FCSTrans* in three components:
1) a linear transformation function $li(x) = a_i x + b_i$
2) a logarithmic transformation function $lo(x) = a_o \times log_{10}(x) + b_o$
3) a logicle transformation function $lc(x) = a_c \times logicle(x) + b_c$
Besides the constants a_i, b_i, a_o, b_o, a_c, and b_c, the *logicle* function also has several parameters that need to be identified. The procedure of parameter identification is described in the Sections below.

Linear transformation

The linear transformation used in *FCSTrans* is a simple linear rescaling of the input range ($\$PnR$ keyword in FCS header) to the output range (resolution).

$$li(x) = x \times \frac{resolution}{\$PnR} \tag{1}$$

For example, when $\$PnR = 2^{18} = 262144$ for resolution 4096, we have $li(x) = x/64$. This is consistent with the BD FACS Diva manual [5], which describes how the linear transformation can be done for BD 18-bit data to a target resolution:

$$X_{Lin} = X_{FCS} \times \frac{2^{18}}{resolution} \tag{2}$$

where X_{Lin} is the original value (i.e., x) and X_{FCS} (i.e., $li(x)$) the linear transformed value.

Logarithmic transformation

Logarithmic transformation is usually applied to the fluorescent parameters in FCS2.0 files whose data can be 18-bit, 4-decade, and sometimes 12-bit. For 18-bit input, the logarithmic transformation according to BD FACS Diva manual [5] is:

$$X_{Lin} = 10^{\frac{X_{FCS} \times num\_of\_decades}{resolution}} \times \frac{2^{18}}{10^{num\_of\_decades}} \tag{3}$$

where X_{Lin} is the original value (i.e., x) and X_{FCS} (i.e., $lo(x)$) the logarithmic transformed value. Suppose that resolution is 4096 and number of decades is 4, we can rewrite the Eq.(3) as

$$x = 10^{\frac{lo(x) \times 4}{4096}} \times \frac{2^{18}}{10^4} \tag{4}$$

i.e.,

$$lo(x) = 1024 \times log_{10}(x) - 1452.6 \tag{5}$$

In order to study the boundary conditions of Eq.(5), we let $lo(x) = 0$ in Eq.(5) and get $x = 26.2$. When $x = 2^{18} = 262144, lo(x) = 4095$, which is the maximum resolution. So we know Eq.(5) maps values from $26.2 \sim 262144$ to $0 \sim 4095$. A problem we can see in Eq. (5) is that the lower bound 26.2 can sometimes be too large. Many events after compensation can be close to zero and even negative values. It may be necessary to decrease the lower bound to accommodate more events into the meaningful input range.

There are different ways to decrease the lower bound. Through simulating a full range of input data to feed into FlowJo and observing the output, we have found that FlowJo chooses a meaningful input range $3 \sim 29923$ instead of $26 \sim 262144$, which seems to fit the common range in existing FCS2.0 files very well.

Knowing the meaningful range of the logarithmic transformation in FlowJo, we solved the equations (i.e., $lo(3) = 0$ and $lo(29923) = 4095$) to figure out its logarithmic transformation formula:

$$lo(x) = \begin{cases} 1024 \times log_{10}(x) - 488.6 & \text{if} \quad 3 \le x \le 29923 \\ 4095 & \text{if} \quad x > 29923 \\ 0 & \text{if} \quad x < 3 \end{cases} \tag{6}$$

In *FCSTrans*, Eq.(6) is generally applied to 18-bit FCS2.0 fluorescent data. For 4-decade data ($\$PnR = 10^4 = 10000$) and 12-bit data ($\$PnR = 2^{12} = 4096$) with traditional output resolution 1024, we have figured out that the generic logarithmic transformation formula is straightforward, just mapping the range of $log_{10}(\$PnR)$ to $resolution$:

$$lo(x) = log_{10}(x) \times \frac{resolution}{log_{10}(\$PnR)} \tag{7}$$

i.e., for 4-decade data ($10^4 = 10000$):

$$lo(x) = log_{10}(x) \times \frac{1024}{log_{10}(10000)} = log_{10}(x) \times 256 \tag{8}$$

Eq.(8) can be derived from Eq.(3) when $\$PnR = 2^{18}$ is replaced by 10000 and the resolution 4096 of 18-bit data replaced by 1024 of 4-decade data.

Similarly, for 12-bit data with resolution of 1024:

$$lo(x) = log_{10}(x) \times \frac{1024}{log_{10}(4096)} = log_{10}(x) \times 283.5 \tag{9}$$

We have found that Eqs.(8) and (9) are also used in Verity software [6].

Logicle transformation

The logicle transformation is a biexponential method that combines the desirable attributes of the log scale for large signals with those of the linear scale for unstained or background signals [7, 8]. The general format of a biexponential function is described in Eq.(10). In terms of FCM data transformation, $S(x)$ should be used inversely because the biexponential method provides a mapping from a transformed value x to an untransformed value $S(x)$. For easy understanding, we follow the tradition to describe the logicle transformation using it inverse form, which can be written in a biexponential format, as shown in Eq.(11).

$$S(x) = ae^{bx} - ce^{-dx} + f \tag{10}$$

$$logicle^{-1}(x) = \begin{cases} T \times e^{-(m-w)}(e^{x-w} - p^2 e^{-\frac{x-w}{p}} + p^2 - 1) & \text{if} \quad x \geq w \\ -T \times e^{-(m-w)}(e^{w-x} - p^2 e^{-\frac{w-x}{p}} + p^2 - 1) & \text{if} \quad x < w \end{cases} \tag{11}$$

While the set of parameters in the logicle transformation seems different from that of the biexponential function, the latter is determined once the former has been specified and both transformation functions generate the same output. Having noticed that calculating the logicle function directly is difficult, we implemented the logicle transformation through using the biexponential function in three steps: 1) identify parameters for the logicle transformation; 2) determine the parameters for the biexponential function based on the logicle parameters determined in Step 1; 3) transform the data using the biexponential equation with the parameters determined in Step 2.

Parks et al [7] has extensively studied the general strategy of choosing parameters for logicle transformation: T is the maximum data value in the displayed scale, m the range of the display in relation to the width of high data value decades, w the strength and range of linearization around zero, p a constant so that $w = 2p\frac{ln(p)}{p+1}$. In terms of FCM, T is the upperbound of the data range specified by the $PnR keyword in the FCS3.0 file header and m can be derived from the number of display decades; the only unknown parameter is w (p can be decided by w). As suggested by Parks et al. [7], the way of choosing an appropriate value for w is to set a negative reference value marking the lower end of the distribution to be displayed. By observing the inputs and the outputs from FlowJo and flowCore, we have identified that -111 has been traditionally used as the negative cut-off. Then we solved the equation $S(0) = -111$ with known parameters for the value of w, and finally obtained the following values for the parameters used in the logicle transformation:

$$w = 1.161, \quad T = \$PnR = 2^{18} = 262144, \quad \text{and} \quad m = 10.36 \tag{12}$$

where we found $m = num\_of\_decades \times ln(10) = 4.5 \times ln(10) = 10.36$ is also recommended by Parks et al [7].

Knowing the logicle parameters, it is straightforward to complete Steps 2 and 3 in R using the biexponential routines. The final step in our implementation of the logicle transformation is to linearly rescale the output by a scaling factor so that the output range is consistent with existing systems:

$$scale = \frac{resolution}{m} = \frac{4096}{10.36} = 395.37 \tag{13}$$

The logicle transformation is an increasing function and generates non-negative outputs when inputs are larger than -111. Inputs smaller than -111 are truncated to zero. With parameter setting in Eqs.(12) and (13), the logicle transformation in *FCSTrans* can be summarized as follows.

$$lc(x) = \begin{cases} 395.37 \times logicle(x) & \text{if} \quad -111 < x < 262144 \\ 4095 & \text{if} \quad x \geq 262144 \\ 0 & \text{if} \quad x \leq -111 \end{cases} \tag{14}$$

Eq.(14) is derived based on the assumption that most FCS3.0 files are 18-bit. Recent flow cytometers like Accuri C6 have been able to generate 24-bit data [9], which can be transformed with the logicle method by changing the parameter setting of T and m with: $T = 2^{24} = 16777216$ and $m = log_{10}T \times ln(10) = 16.64$. For scatter parameters in Accuri FCS files, while the linear transformation function in Eq.(1) is still valid, it is usually necessary to zoom in the data plot to visually identify the cell populations because the plotting range is dramatically extended from 2^{18} to 2^{24}.

References

[1] http://flowjo.com, Accessed February 02, 2012

[2] http://sourceforge.net/projects/flowcyt/files/GenePattern Flow Cytometry Suite/FCS2CSV/, Accessed February 02, 2012

[3] Greg Finak, Juan-Manuel Perez, Andrew Weng, and Raphael Gottardo, Optimizing transformations for automated, high throughput analysis of flow cytometry data, BMC Bioinformatics 11:546 (2010)

[4] http://bioconductor.org/packages/2.6/bioc/manuals/flowCore/man/flowCore.pdf, Accessed February 02, 2012

[5] BD FACS Diva Software 6.0 Reference Manual,Beckton and Dickinson, http://facs.stanford.edu/sff/doc/BDFACSDivaV6Manual.pdf, Accessed March 06, 2012

[6] Verity software house Inc., A Discussion of Linear-to-Log Data Conversion in Flow Cytometry, http://www.vsh.com/publication/LinLog.pdf, Accessed February 02, 2012

[7] David R. Parks, Mario Roederer, and Wayne A. Moore, A New Logicle Display Method Avoids Deceptive Effects of Logarithmic Scaling for Low Signals and Compensated Data, Cytometry Part A 69A:541-551 (2006)

[8] James W. Tung, Kartoosh Heydari, Rabin Tirouvanziam, Bita Sahaf, David R. Parks, Leonard, A. Herzenberg, and Leonore A. Herzenberg, Modern Flow Cytometry: A Practical Approach, Clin Lab Med. 27(3): 453-v (2007)

[9] http://accuricytometers.com/files/Accuri_Revolutionizes_Flow_Cytometry.pdf, Accessed February 02, 2012