Cytometry
PART A

Per-Channel Basis Normalization Methods for Flow Cytometry Data

Florian Hahne,[1†] Alireza Hadj Khodabakhshi,[2†] Ali Bashashati,[2] Chao-Jen Wong,[1] Randy D. Gascoyne,[3] Andrew P. Weng,[2] Vicky Seyfert-Margolis,[4] Katarzyna Bourcier,[4] Adam Asare,[4] Thomas Lumley,[5] Robert Gentleman,[1] Ryan R. Brinkman[2*]

[1]Fred Hutchinson Cancer Research Center, Seattle, Washington

[2]Terry Fox Laboratory, BC Cancer Agency, Vancouver, British Columbia, Canada

[3]BC Cancer Agency, Vancouver, British Columbia, Canada

[4]Immune Tolerance Network, UCSF, San Francisco, California

[5]Department of Biostatistics, University of Washington, Seattle, Washington

V. Seyfert-Margolis' present address is Food and Drug Administration, Silver Spring, Maryland.

The views presented in this article do not necessarily represent those of the Food and Drug Administration.

Received 21 May 2009; Revision Received 28 September 2009; Accepted 12 October 2009

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

Grant sponsor: NIH; Grant numbers: 1R01EB008400, NO1-AI-15416 (PI: Bluestone), 1R01EB005034; Grant sponsor: Canadian Institutes of Health Research; Grant number: IHP 94132; Grant sponsors: National Institute of Allergy and Infectious Diseases and the Juvenile Diabetes Research Foundation, the Michael Smith Foundation for Health Research

[†]The first two authors contributed equally to this work.

Abstract

Between-sample variation in high-throughput flow cytometry data poses a significant challenge for analysis of large-scale data sets, such as those derived from multicenter clinical trials. It is often hard to match biologically relevant cell populations across samples because of technical variation in sample acquisition and instrumentation differences. Thus, normalization of data is a critical step before analysis, particularly in large-scale data sets from clinical trials, where group-specific differences may be subtle and patient-to-patient variation common. We have developed two normalization methods that remove technical between-sample variation by aligning prominent features (landmarks) in the raw data on a per-channel basis. These algorithms were tested on two independent flow cytometry data sets by comparing manually gated data, either individually for each sample or using static gating templates, before and after normalization. Our results show a marked improvement in the overlap between manual and static gating when the data are normalized, thereby facilitating the use of automated analyses on large flow cytometry data sets. Such automated analyses are essential for high-throughput flow cytometry.

Key terms

flow cytometry; data analysis; statistical analysis; normalization

Improvements in instrumentation, throughput, and automation have made multiparameter flow cytometry an important tool in the clinic. FCM use is increasingly used in the study of human subjects participating in clinical trials as many new therapies in development target immune system cells or factors. Along with the rise in numbers of tests performed on clinical samples, the size of FCM data sets is expanding as more parameters are measured routinely for discretely defined subpopulations of immune cells. Quality control procedures carried out in clinical FCM laboratories aim to minimize electronic drift and error during data acquisition (1). However, variation both within and amongst laboratories can change the absolute position of cell populations across samples when displayed on a common axis, thus impacting interpretation of the findings, which are often subtle in nature and affected by variability among humans. In addition to biological variation, technical variations are also problematic and must be minimized to assess true biological signal. They are often introduced by changes in the instrumentation channel voltages or by switching antibody variants and impose substantial complications for automated analysis of FCM data, especially for the interpretation of cell populations and marker extraction.

Common practice is to deal with technical variation by manually inspecting each multigraph generated by the flow cytometer in one or two dimensions, a process which is both subjective and time intensive, thus creating a bottleneck in the ability

ISAC
International Society for Advancement of Cytometry

*Correspondence to: Ryan R. Brinkman, Terry Fox Laboratory, BC Cancer Agency, Vancouver, BC, Canada
Email: rbrinkman@bccrc.ca

Published online 6 November 2009 in Wiley InterScience (www.interscience.wiley.com)
DOI: 10.1002/cyto.a.20823

to analyze large data sets and potentially skewing data interpretation. Furthermore, higher dimensional approaches that aim to exploit interactions between all available channels can only poorly cope with such variations. Given the pressing need for analysis automation of high-throughput FCM data (2) and to eliminate the per-case investigation of samples in manual analysis, it is important that the cross-sample and cross-clinic variation problems be resolved computationally. A critical component of automated analysis is the normalization of the FCM data, which carried out before any data analysis or interpretation. Normalization is a preprocessing step performed on a set of samples with the aim of removing effects that arise from technical variation rather than from biological differences. It is carried out by computing and subsequently removing population shifts, either on a multidimensional basis or by treating each channel separately. We are actively researching multidimensional approaches, however significant challenges remain unresolved at this time. Here, we only consider the somewhat simpler problem of one-dimensional normalization. Many of the factors that influence the absolute fluorescence intensity of a particular stain are not controllable or are simply unknown, so explicit adjustment is not practicable. In this article, we describe two methods, which normalize a single channel by identifying and matching common prominent features, which we refer to as landmarks, across samples. We use areas of high local density (i.e., density peaks) in the individual data channels as landmarks. Under ideal conditions we expect these landmarks to be well aligned among all samples, and hence use them to determine and remove the technical variation by shifting the underlying data in a way that the distance between all matching landmarks is minimized.

Data normalization also helps in matching (labeling) biologically relevant cell populations across a set of samples. This matching process is a major step in high-throughput FCM data analysis. It is often performed by clustering of the median of clusters (meta-clustering) (3). However, if there is significant technical variation in the data, the meta-clustering step will assign similar cell populations with dissimilar absolute positions to different meta-clusters. This results in different labeling for these biologically similar cell populations.

Methods

We developed two different normalization methods – gaussNorm and fdaNorm. Both methods are freely available as open source software in the flowStats (4) package of Bioconductor (5). The methods differ in the implementation of three major steps (Fig. 1):

1. Landmark identification: For each channel we assume that there is a fixed set of landmarks that are of interest. One can either think of the landmarks as being known a priori or they can be estimated from the data. In each sample, we estimate these landmarks by identifying peaks in the density. In other cases, landmarks might also represent other features, such as valleys, and those can easily be included. In Figure 1, left panel, we assume that there are two landmarks, a peak corresponding to negative staining (L_1) and the second peak corresponding to positive staining (L_2). However, the total number of peaks identified for each sample may vary (Fig. 1, left panel, samples 3 and 4). To correctly compute the maximum number of biologically relevant landmarks in a sample, we need to determine whether a change in peak locations is due to the variety of the corresponding populations or whether it is caused by a variation in experimental settings. For instance, we might encounter two samples, each with single peaks that are quite far apart. The distance between them may reflect two distinct subpopulations of cells, or alternatively it may be technical variation.
2. Landmark registration: For each sample we have identified candidate landmarks, and the process of landmark registration establishes the correspondence between the peaks identified in that sample and the landmarks of interest (Fig. 1, middle panel). In other words, we assign a landmark label (L_1 or L_2) to each peak in each sample. In the simplest case, the number of peaks identified is the same as the number of landmarks. However, cell populations can be depleted in particular samples, or there may be samples for which spurious peaks have been identified, in which case an algorithm to register peaks to landmarks is needed.
3. Landmark alignment: In the final step, the data for each sample are transformed independently, so that the landmarks in the sample are aligned with each other (Fig. 1, right panel). The transformation functions are order preserving and should not significantly alter the distribution of individual peaks.

The gaussNorm Method

Landmark identification. We assume that the maximum number of the landmarks m is a predetermined parameter and identify all local maxima in the kernel density estimate of the input data. Many of these local maxima are due to noise and do not correspond to true populations of interest. These spurious peaks mostly occur around the end of the spectrum, and they tend to have low-density values. Moreover, we may encounter cell populations that consist of several close peaks, especially when the kernel density estimate has small bandwidth. Despite these challenges, we recommend using small bandwidth kernel density estimates for detecting peaks because oversmoothing increases the risk of missing the smaller peaks. To deal with spurious peaks, we only select the ones that most likely correspond to distinct cell

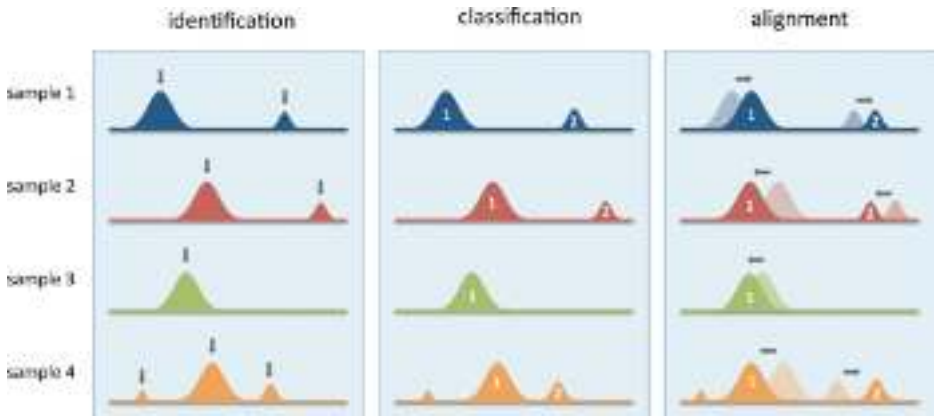

Figure 1. Schematics of the proposed normalization methods. Identification: for each sample, landmarks (indicated by arrows) are estimated from density estimates of the underlying data. The total number of landmarks for each sample can vary. Classification: corresponding landmarks (indicated by 1, 2) are matched across samples. The total number of landmarks has to be fixed at this stage. Alignment: the underlying data are transformed to minimize the distance between corresponding landmarks. [Color figure can be viewed online at www.interscience.wiley.com]

populations. More precisely, for each peak p we define a confidence score $s(p)$ as follows:

$$s(p) = \left(\sum_i \kappa(p) - \kappa(i)\right) \cdot \kappa(p), i \in [(p - \mathrm{bw_s})/2, (p + \mathrm{bw_s})/2], \quad (1)$$

where bw_s is a bandwidth constant and $\kappa(.)$ is the kernel density estimate. In general, this confidence score reflects both the peaks sharpness and height. Subsequently, we cluster the peaks into groups as follows. Assuming p_i and p_{i+1} are the locations of two consecutive peaks, if $p_{i+1} - p_i$ is less than a threshold then these peaks belong to the same group. The default value of this threshold is 5% of the range of the data in the implementation of the method. For each group of peaks, we retain only the peak with the highest confidence score. Finally, we select at most m landmarks from the set of peaks that have the highest confidence score.

Landmark registration. The aim of this step is to classify the landmarks into m classes. If the data have exactly m landmarks, we label them with numbers from 1 to m consecutively with respect to their locations. For samples with less than m landmarks, let $B[i]$ be the median of all the landmarks with label i. We call vector B the base landmarks and we say $B[i]$ has label i. Let P be the vector of e landmarks ($e < m$) for an unlabeled sample. We find a match between the elements of P and B with the minimum sum of the distance between the matching landmarks. Note that in a match, each element in B is paired with at most one element in P and each element in P is paired with exactly one element in B. This way each element in P gets the same label as its matching landmark in B. Once the labeling in this step is done, we recalculate the base landmark vector B.

Landmark alignment. Here, we transform the data such that the landmark with label i is moved to the fixed position $B[i]$. The gaussNorm method uses Algorithm 1 to normalize the data vector D with the landmarks vector P; label($P[i]$) is the label of $P[i]$ in the best match between P and B. In this algorithm, the amount by which the data points are shifted is exponentially decreased as the points move away from the landmark locations. This makes it possible to move the landmarks independent of each other, and therefore, the distance between the landmarks and their absolute positions can be easily adjusted.

Algorithm 1: This algorithm is used by the gaussNorm method to align the landmarks. Here $\varphi(x, \sigma^2) = e^{-x^2/2\sigma^2}$, where σ^2 is the standard variation of the vector `D` and `label(P[i])` is the label of landmark i.

```
for i = 1 to |P| do
    for k = 1 to |D| do
        D[k] += φ(|D[k] - P[i]|, σ²) · (P[i] - B[label(P[i])])
    end
end
```

The fdaNorm Method

Landmark identification. Kernel density estimates contain essential properties of the data's features such as gradient and curvature. We use a robust statistical testing framework, proposed by Duong et al. (6), to infer significant landmark based on the gradient and curvature derivatives of modal regions with practical bandwidth selection. The landmark estimates are thus identified as maxima in the high-density regions where the derivatives differ significantly from zero (6). This approach typically yields fewer spurious peaks. The total number of landmarks m is determined from the data as the mode (i.e., the most frequent) of the number of landmarks identified in the samples. For example, if for 9 out of the 10 samples, we identified two landmarks, m is set to 2.

Landmark registration. Using the *k*-means clustering method, we cluster all landmark locations into m clusters, independently of samples. Subsequently, the landmark locations for each sample are labeled by these cluster assignments. In cases where more than m landmarks are identified for a particular sample or when multiple landmarks share the same classification label, only the landmark with the smallest distance to the cluster centroid is used for a given class.

Landmark alignment. The kernel density estimate for each sample i is represented by a B-spline interpoland x_i. For each x_i, the landmark alignment process requires the identification of argument, t_{ij}, $j = 1,\ldots,m$, associated with each of the landmark j (7). The fact that the set of x_i functions exhibits location variation of the landmarks makes autogating more challenging. To overcome this difficulty, we align landmarks across samples at fixed locations t_{0j} by transforming curves x_i for all i.

Let y be a fixed function in the same class as x_i (8). The alignment proceeds by transforming x_i by a strictly monotone function h_i on the argument of x_i, i.e., $x^*(t) = x_i\{h_i(t)\}$ such that the distance between y and the transformed curves x^* is minimized. The problem of transforming the arguments of the curves is referred to as curve registration (8,9). The monotone function h_i is known as a warping function in the engineering literature (8) with properties (7):

- $h_i(T_{\text{start}}) = T_{\text{start}}$, where T_{start} is the starting point of the domain.
- $h_i(T_{\text{end}}) = T_{\text{end}}$, where T_{end} is the right end point of the domain.
- $h_i(t_{0j}) = t_{ij}$, $j = 1,\ldots,m$.
- h_i is strictly increasing (i.e., $h(t_2) > h(t_1)$ for $t_2 > t_1$). That is, h is invertible such that h and $x\{h(t)\}$ have an one-to-one correspondence.

The estimation of h_i relies on minimizing the penalized squared error criterion (8)

$$F_\lambda(y, x|h) = \int \|y(t) - x\{h(t)\}\|^2 dt + \lambda \int \omega^2(t) dt, \quad (2)$$

where λ is a fixed smoothing parameter, and $\omega = D^2/Dh$ is the relative curvature of h. It is usually sufficient to estimate y by the cross-sample average $\tilde{x}$. See (7,8,10) for details on calculation of ω and h. The warping calculation and curve registration utilities are available from the fda package (11) in the statistical programing language R (12) and MATLAB.

Evaluation

We have used two independent data sets to test and compare our normalization method. All data are available as Supporting Information. The first data set (lymphoma) consists of 30 randomly selected lymph node biopsies from patients treated at the British Columbia Cancer Agency between 2003 and 2008. These patients were histologically confirmed to have diffuse large B-cell lymphoma (DLBCL). Cells derived from the lymph nodes were stained for three cell surface markers, CD3, CD19, and CD5. In 2005, there was a voltage change in the instrument settings that caused subpopulations for the CD3 and CD5 channels to shift in an absolute manner while maintaining their relative distribution. The second data set, obtained from a renal transplant retrospective study conducted by the Immune Tolerance Network (ITN), included analysis of peripheral blood cells stained using antibodies to the CD3, CD4, CD8, CD69, and HLADr markers. The between-sample variation in this set was much smaller than the lymphoma data set, but within the typical range expected for most high-throughput clinical studies. Both data sets were initially gated on total lymphocytes to remove artifactual events such as cell debris and doublets.

To test the success of our normalization strategies, we compared analyses of the data using static versus manual gating, where we assumed the manual gating to be the gold standard. Manual gates were adjusted for between-sample variability on a sample-to-sample basis by an expert analyst. Static gates were defined to be constant across all samples, based on a manual gate placed by an expert on one sample chosen at random. By assuming that most of the experimental variation would be removed using our algorithms, we attempted to correctly define subpopulations across all the samples using the static gates and compare those results with the manually gated data to assess our success. This allowed us to evaluate whether our normalization strategies could in fact reliably assign high-density areas to the appropriate subpopulation that would be defined using manual gating. If the normalization did remove experimental variation, then we expected that the statically gated data would coincide with the manually gated data. More precisely, after successful normalization, we expected statically gated data to be highly correlated with the manually gated data.

To determine the degree of agreement between manual and static gating, we computed the Jaccard index J (a statistic used for comparing the similarity and diversity of sample sets) as a measure of overlap. Let D_{m} and D_{s} be the subset of data points returned by a manual gate and its corresponding static gate. J is the size of the intersection of D_{m} and D_{s} divided by the size of their union. This measure takes into account both the sensitivity and specificity of a static gate when compared with its corresponding manual gate. The value of J is 0 if there is no overlap, and 1 if there is perfect agreement between the two gates.

We also assessed the effectiveness of our data normalization approach to facilitate the ability of meta-clustering to correctly match (i.e., label) cell populations across samples. The DLBCL data set was first analyzed with flowClust (13) to automatically gate cell populations. The cluster median of each population was then selected, and these values were then clustered using flowClust in a "meta-clustering" step to effectively group the same populations together across samples. The same procedure was then carried out, but subsequent to an initial data normalization step, to test the hypothesis that normalization would facilitate matching of like cell populations in the final meta-clustering step by aligning like populations to a common general location. The optimum number of clusters in each case was found using the BIC measure provided by the flowClust tool.

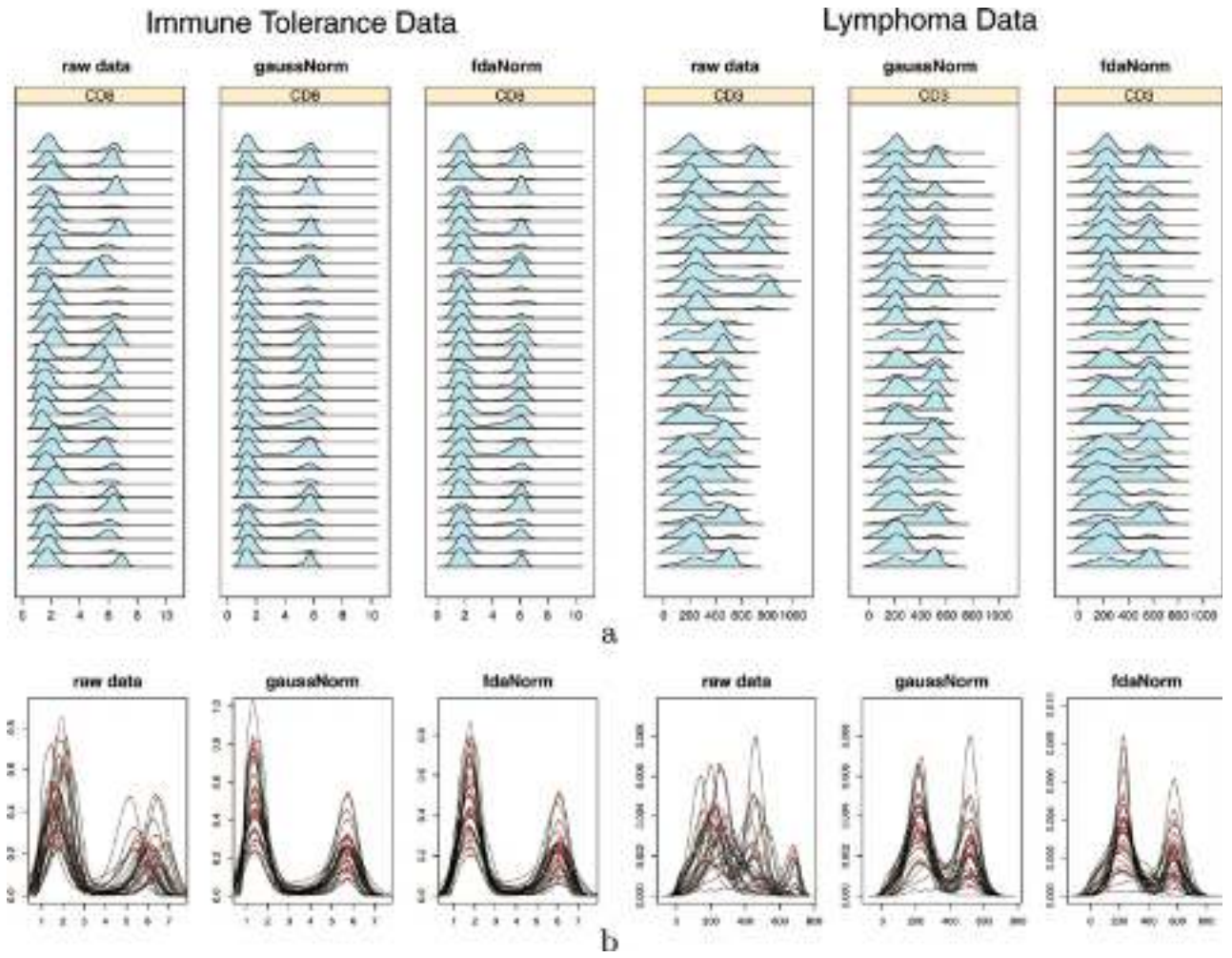

Figure 2. (**a**) Stacked density plots of a single FCM channel before and after normalization with the two algorithms for two different data sets. Each row in the plot corresponds to data for one sample. (**b**) Overlay of the same density plots. The dots represent landmarks. As seen in the leftmost panels (raw data), there is considerable variation in the locations of high-density areas for the unnormalized data, which is mostly removed after applying either one of the normalization methods (middle and rightmost panels). [Color figure can be viewed online at www.interscience.wiley.com]

Finally, we assessed the performance of two existing autogating techniques, the magnetic gates in FlowJo (Tree Star) and autopositioning gates in WinList (Verity Software House) on the DLBCL dataset.

Results

Density plots from before, Figure 2a left panel, and after normalization with the gaussNorm and fdaNorm methods, Figure 2a middle and right panels, respectively, are shown for a single channel for sample set data from the ITN and lymphoma groups. Specifically, the top left plot in Figure 2a depicts the stacked density curves of the CD8 channel for all the ITN samples, and the top middle and top right plots show the density curves of the same data after they have been normalized using the gaussNorm and fdaNorm methods, respectively. Note that the peaks in the normalized data have been shifted to be in much closer proximity and are well aligned. This is more evident in the overlay of the same density curves in Figure 2b, where the landmark estimates are indicated by dots. The right panels of Figures 2a and 2b depict similar results for the CD3 channel of the lymphoma data set. In both examples, the two peaks can only be reasonably separated by a single line for all samples after data normalization. The two methods appear to perform comparably for both data sets as indicated by the *P*-values of a *t*-test comparing the values of *J*. Results for the additional channels can be found in the Supporting Information of this manuscript.

Manual gates were defined separately for each sample, identifying a number of different biologically relevant cell populations. For each of these populations, we also defined two static gates; one for the raw data and the other for the normalized data. For the lymphoma data, we defined these static gates based on sample number 9 as the reference sample because its populations were representative of the populations

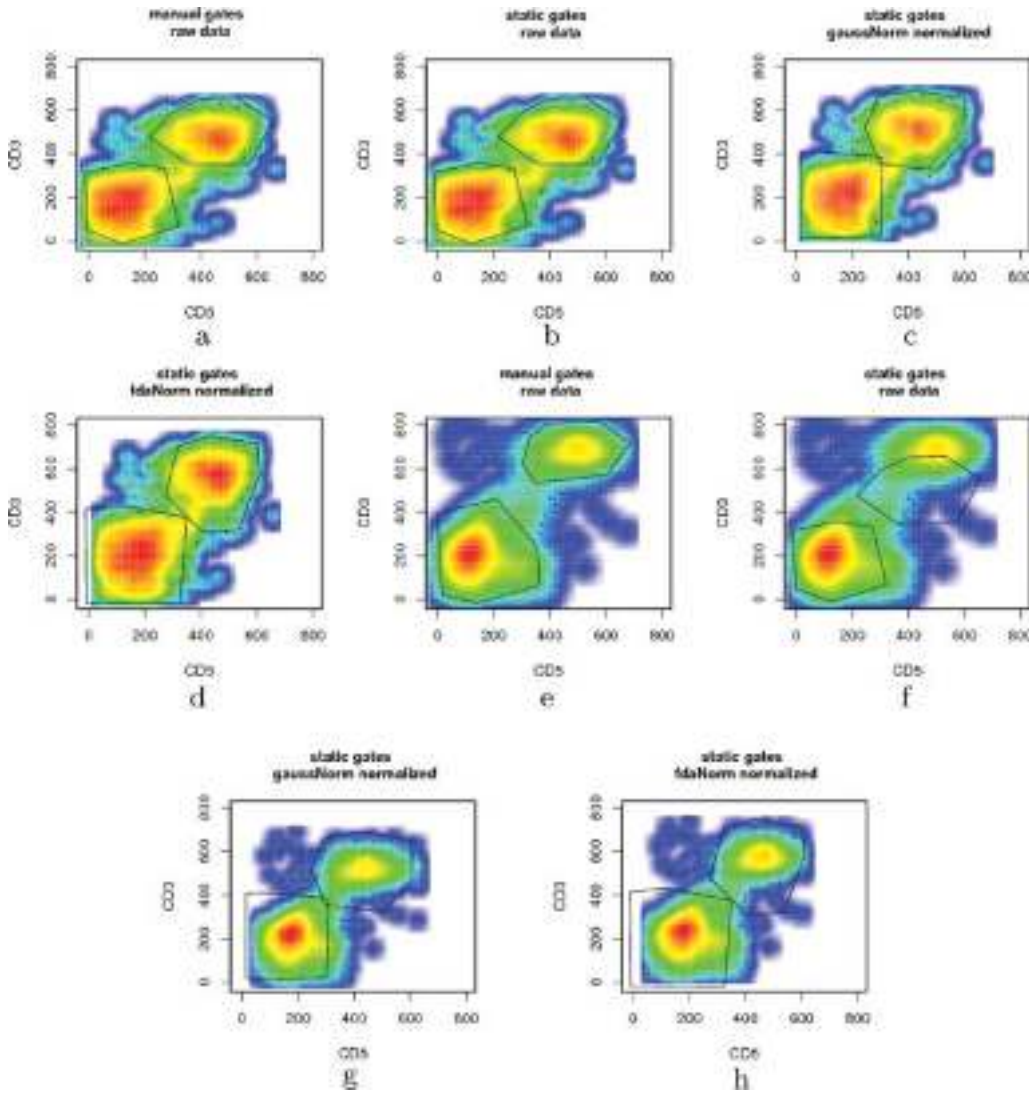

Figure 3. (**a**) The manual gate definitions on CD3 vs. CD5 panel of sample number 9, the reference sample of the lymphoma data. The static gates defined on prenormalized (**b**), gaussNorm-normalized (**c**), and fdaNorm-normalized (**d**) data of the reference sample. (**e**) The manual gate definitions on CD3 vs. CD5 panel of sample number 30. The static gates applied on pre-normalized (**f**), gaussNorm-normalized (**g**), and fdaNorm-normalized (**h**) data of sample number 30. [Color figure can be viewed online at www.interscience.wiley.com]

of interest in the total data set. Figure 3 depicts the manual and static gates defined on the CD3 vs. CD5 panel of the reference sample and the corresponding gates on an arbitrary sample, sample number 30. Shown in Figure 3a are two manual gates defined in the CD3 vs. CD5 parameter space of the reference sample. Figure 3b depicts the corresponding static gates defined on the reference sample before it is normalized, and Figures 3c and 3d depict the static gates defined on the reference sample after normalization by either the gaussNorm or the fdaNorm method, respectively. Note that because the static gates are defined on a representative sample from each group, it is very natural that they are different between the groups. Similarly, the plot in Figure 3e shows the equivalent gates defined for sample number 30, and Figures 3f–3h depict the application of static gates on sample 30 pre-, post-gaussNorm-, and post-fdaNorm- normalization, respectively. As shown in Figure 3f, the applied static gate for the CD3+CD5+ population did not capture this population accurately because of a variation in channel CD3. The plots in Figures 3g and 3h clearly show that normalization has fixed this problem by shifting the CD3+CD5+ population back into the boundaries of the static gates. See Supporting Information for similar results on panels SSC vs. CD5 and SSC vs. CD3. The static gates for the ITN data set were chosen in a similar fashion.

Figure 4 shows box plots of the values of the Jaccard index J between static gates and the customized manual gates for both prenormalized and postnormalized data. Ideally, we expect to have complete overlap between static gates and manual gates after normalization. For all gate definitions, the degree of overlap was similar or higher after normalization. Overall, both methods performed similarly, within reasonable

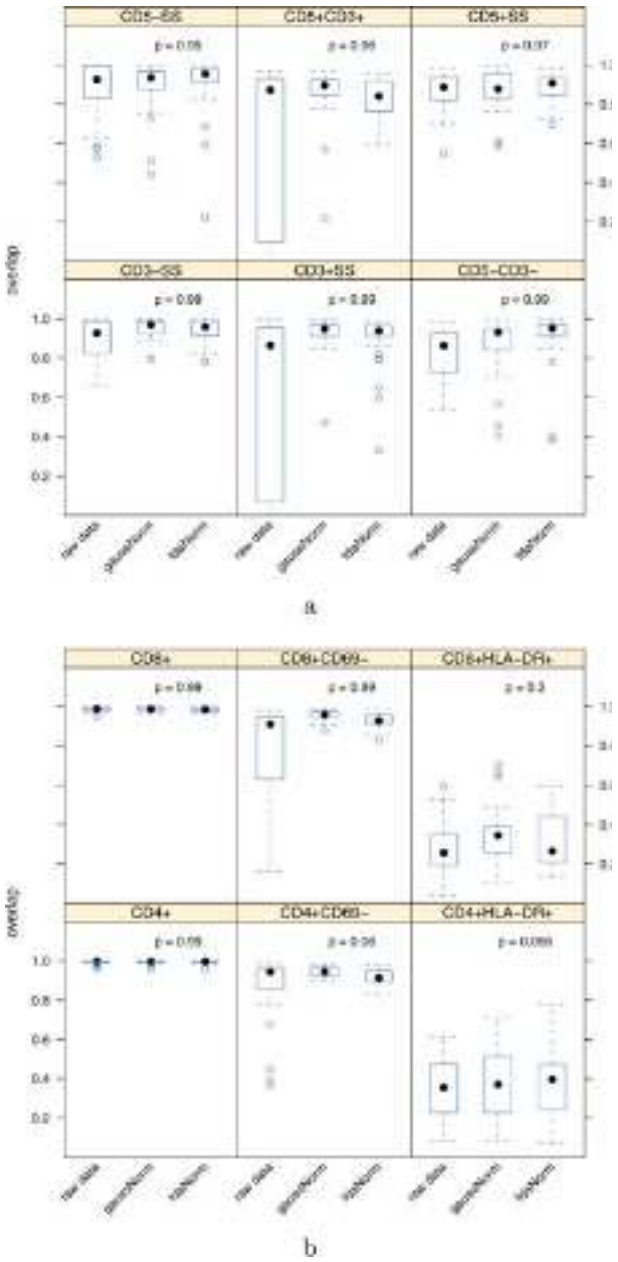

Figure 4. Box plots of calculated percentage overlap between gold standard manual gating and static gating before and after normalization for (**a**) the lymphoma data set and (**b**) the ITN data set. Each panel represents the results for a different gate region. In all cases, the average overlap is equal or better after normalization, and both methods perform comparably, as indicated by the *P*-values from a *t*-test comparing the results for the two methods. [Color figure can be viewed online at www.interscience.wiley.com]

computation times. Some outlier samples with very low overlap remained. In particular, we identified 4 gates out of 180 gates in the lymphoma data set with overlap values less than 50%. These gates belong to samples number 10 and 28 for which the normalization method missed one of the landmarks in CD3 channel. To help assist in identifying such problematic cases, we provide the capability to produce two groups of diagnostic plots. Shown in Figures 5a and 5b are the first group diagnostic plots highlighting the amount of normalization that was necessary to adjust the landmarks for channels CD3 and CD8 of lymphoma and ITN data sets, respectively. On the *y*-axis are the 30 samples, and each dot represents a landmark location. The vertical lines connecting the dots represent the common base landmark, i.e., the location of all commonly registered landmark after normalization. In Figure 5a, a predefined sample grouping which reflects the altered instrument settings is indicated by different colors, and it is apparent that the landmark locations differ systematically between the two groups. Cases where landmarks are missing in a particular sample are indicated by "x" plotting symbols. Using these diagnostic plots, one can identify samples with missing landmark values or those with a landmark location that does not belong to any group of landmarks, in other words those that are outliers with respect to the landmark locations. Such samples have higher probability of having normalization errors and therefore, should be inspected manually. It is worth noting that the two samples *V*10 and *V*28 that scored poorly in the overlap score as mentioned earlier both fall into the category of samples with missing landmarks.

The second group of diagnostic plots visualize the results of backgating analysis. Shown in Figure 6 are the results of a back-gating analysis of the events under the different peaks into the forward- and side-scatter dimensions. The assumption here is that, if events under a certain peak indeed represent similar biological entities, they might also form tight populations when projected back into the morphological channels. Plotted are ellipses which represent the location and covariance of the population in the forward- and side-scatter dimension for the respective samples. Small ellipses indicate low variance, and hence a more tightly clustered population. Again color is used to indicate the sample grouping. There is a clear separation between the samples from the different groups, and within a given group there are occasional outliers.

For the ITN data, the normalization had almost no effect on the CD4 and CD8 gating, most likely because the respective cell populations separated well enough in the raw data for a static gate to work. The CD69 gating, however, improved dramatically. Only small improvement was observed for the HLADr gating, and in general, the overlap was quite low for this marker. This observation can be explained either by the presence of biological variation, which is not modeled by the static gating, or by inaccuracies in the manual gating step. Because the HLADr-negative populations were very small and diffuse, it was difficult to precisely gate this data manually as well (Fig. 7b). In addition, the Jaccard index is not robust toward small numbers, and minor deviations in these rare populations quickly lead to low values of *J*.

Both the CD69 and the HLADr gates of the ITN data set were gated on a single parameter, i.e., they represented a single line in the parameter space for each sample. Thus, we could use the same warping functions computed by the fdaNorm method for the underlying data to transform the manual gate locations. Assuming that the normalization procedure indeed captured most of the technical variation that was adjusted for

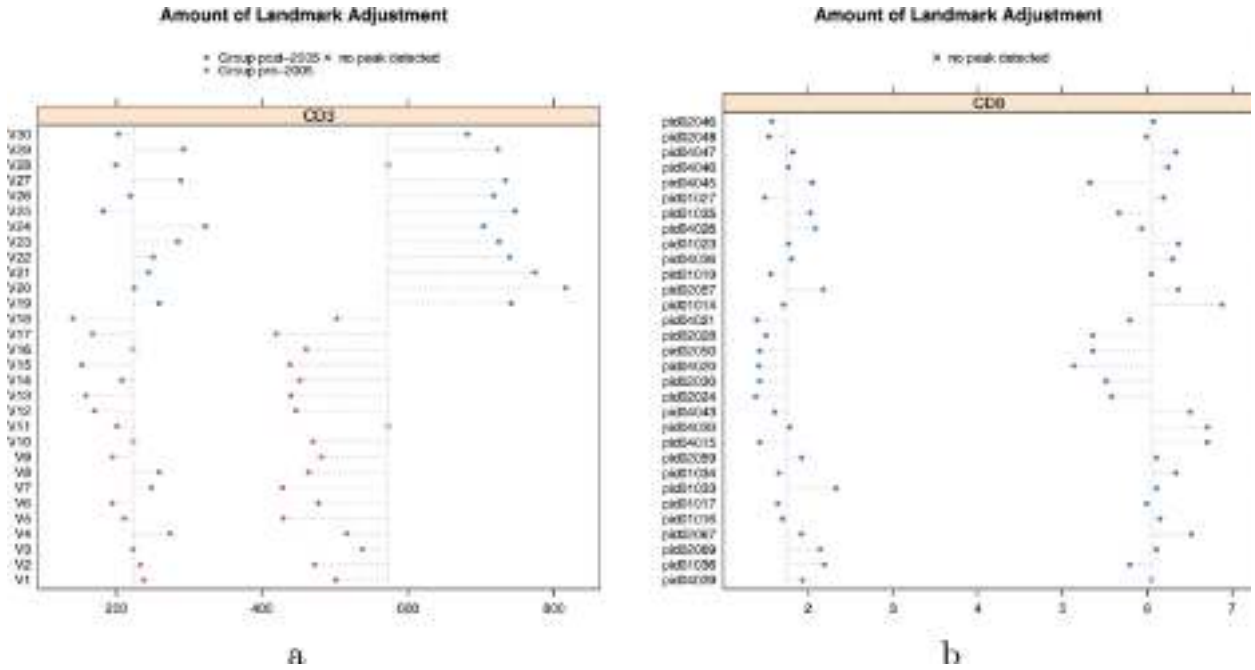

Figure 5. Diagnostic plots highlighting the amount of normalization that was necessary to adjust the landmarks for a particular channel for the lymphoma 5**a** and ITN 5**b** data sets. On the *y*-axis are the samples, and each dot represents a landmark location. The vertical lines connecting the dots represent the common base landmark, i.e., the location of all commonly registered landmark after normalization. The data from these two patient groups indicated by color in Figure 5a were expected to be significantly different in MFI because of a change in laser voltage settings. This is reflected by the differing amounts of adjustment that was necessary for alignment.

in the manual gating step, the distribution of gate location should be much more uniform after normalization. In Figure 7a, we plotted the manual gate locations before and after applying the transformation. For the CD69 channel, the transformed locations were much more uniform, suggesting that the automated normalization procedure was able to remove technical variation to a similar extent as we were able to address manually. However, not much improvement was seen for the HLADr channel. This could mean that our method has missed one or several relevant biological attributes that have

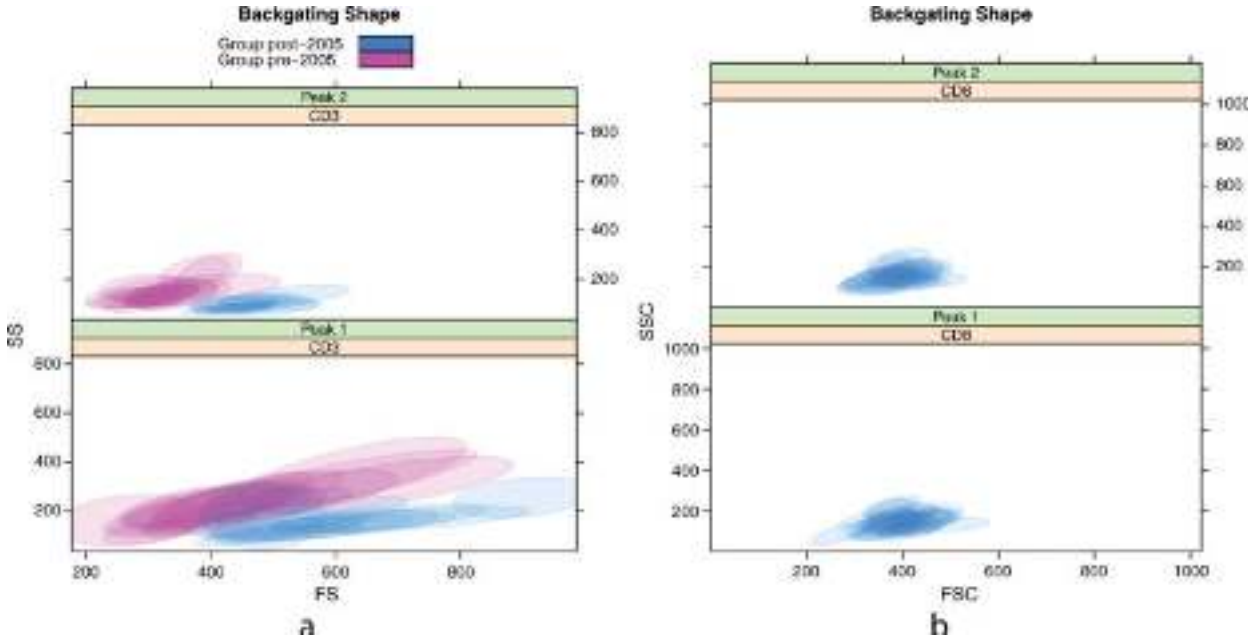

Figure 6. Diagnostic plots showing the results of a backgating analysis of the events under the different peaks into the forward- and side-scatter dimensions for lymphoma 6**a** and ITN 6**b** data sets. In (a) a predefined sample grouping which reflects the altered instrument settings is indicated by different colors, and it is apparent that the backgated populations differ systematically between the two groups.

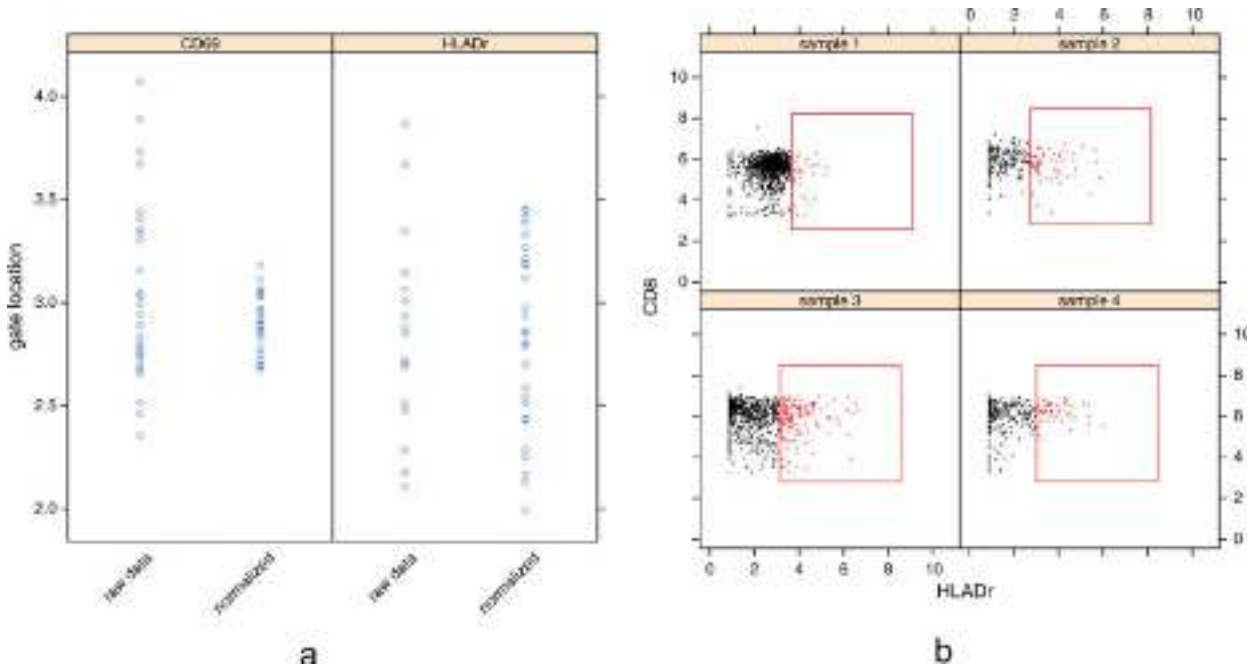

Figure 7. (**a**) Locations of the manual *CD69* and *HLADr* range gates for all samples in the ITN data set, before and after transformation using the warping function computed by the fdaNorm method for the underlying data. The locations of the *CD69* gate are much more uniform after transformation, whereas transformation of the *HLADr* gate locations had little effect. (**b**) Manual gates of the *HLADr* channel for the first four samples. The selection of the cutoff is rather subjective and often ambiguous. [Color figure can be viewed online at www.interscience.wiley.com]

been accounted for in the manual gating. Alternatively, this could reflect the uncertainty in the gating of some cell subpopulations, and the remaining variation reflects the ambiguity of manual selections (Fig. 7b).

Figure 8 shows an example of the case where populations were mislabeled after the meta-clustering step when data were not normalized. The subfigures represent the density plots of the pooled data of the 30 samples in the

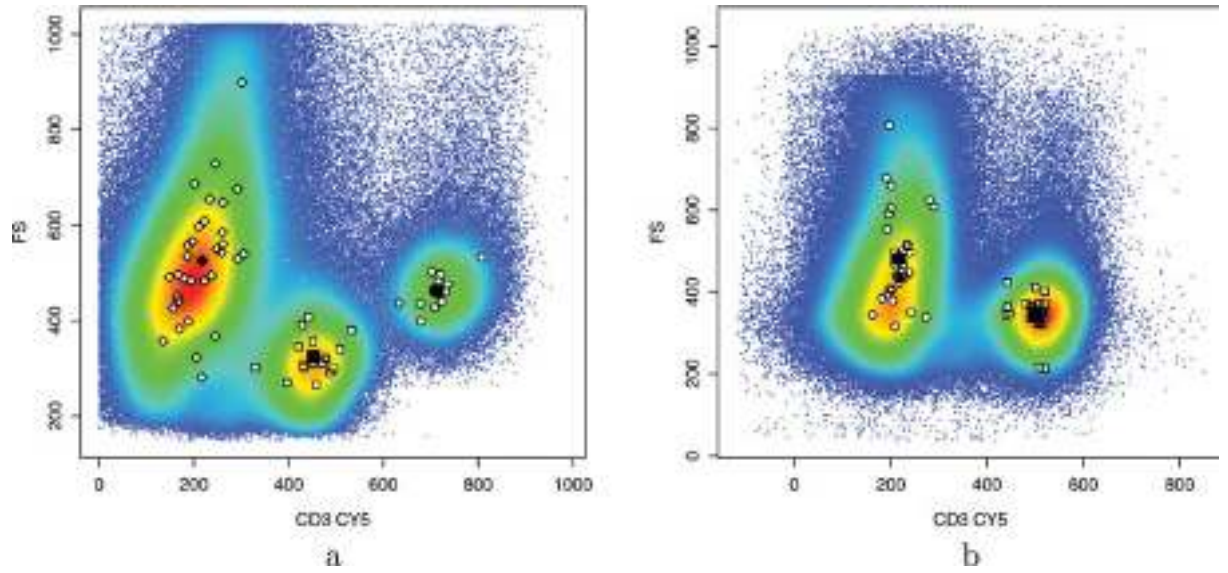

Figure 8. Clustering of the median of clusters (i.e., meta-clustering) on DLBCL data. The background color density plots were generated by pooling the 30 samples in the DLBCL dataset pre- (**a**) and post- (**b**) normalization. The superimposed white squares, diamonds, and circles represent the median of the clusters and the black squares, diamonds, and circles represent the median of the meta-clusters identified by flowClust. (a) The optimum number of meta clusters is incorrectly three based on the BIC measure when the data are not normalized. In this case, the square and circle meta-clusters are all CD3+ cell populations that are labeled differently. (b) This problem is resolved when the meta-clustering is performed on normalized data. Here the optimum number of clusters based on the BIC measure is two which corresponds to CD3− and CD3+ cell populations. [Color figure can be viewed online at www.interscience.wiley.com]

DLBCL dataset pre- (Fig. 8a) and post- (Fig. 8b) normalization. In these plots, white symbols represent the median of the clusters, and the black symbols represent the median of meta-clusters. When the data are not normalized, the meta-clustering results in three clusters (using the BIC measure), which are represented as squares, triangles, and circles in Figure 8a. However, notice that the square and circle clusters are all CD3+ cell populations that are labeled differently. Performing the same analysis on the normalized data leads to two distinct meta-clusters, which is the correct number of biologically relevant cell populations in the data (Fig. 8b).

We have also assessed the performance of two conventional autogating tools; the magnetic gates in FlowJo and autopositioning gates in Winlist on the DLBCL dataset. Our aim for this experiments was to see whether the current autogating tools can produce similar overlapping results between manual gates and static gates when the data are not normalized. We defined the CD3+CD5+ gate on a chosen reference sample and applied this gate with the autogating option enabled on the rest of the samples in the DLBCL dataset.

Supporting Information Figure 5a shows the reference sample and the static CD3+CD5+ gate defined. The rest of the plots in Supporting Information Figures 5–8 show the application of the static CD3+CD5+ gate on the other 29 DLBCL samples with the magnetic gate option enabled in FlowJo software. The magnetic gates only worked for the samples with small population shifts. In samples with drastic population shifts (i.e., 12 of 29 samples in Supporting Information Figs. 5–8) the magnetic gates failed to adjust to the correct population. In these samples the magnetic gates surrounded the CD3-CD5- populations instead. This is because the CD3-CD5- populations in these 12 samples were closer to the static gate location than the CD3+CD5+ populations. Similarly, we defined a static CD3+CD5+ gate on the reference sample (Supporting Information Fig. 9a) and applied it to eight samples with the autopositioning option enabled in WinList (Supporting Information Figs. 9a–9i). As seen in these Supporting Information Figures in five of eight samples the static gate was not correctly adjusted to the CD3+CD5+ populations.

Discussion

We developed two novel methods for the normalization of single-channel FCM data, based on the alignment of prominent landmarks. Both of our methods performed equally well when evaluating data from two independent FCM data sets. More importantly, we demonstrated that normalization facilitates subsequent subpopulation analysis as it allowed for the use of semiautomated static gating for many of the major cell subpopulations. Comparisons of manually gated data to the same data normalized with our algorithms and statically gated showed significant correlations for many subpopulations. In particular, we were able to increase the overlap between manual and static gating to greater than 95% for all of the large and medium-sized populations. In some cases, of rarer populations, our algorithms failed to allow for static gate subpopulation definitions. We believe this is due, in part, to the fact that density estimation is difficult to perform when the data are dispersed in a high-dimensional space and when the number of events is not sufficient to define an area of high density. This is often the case with small subpopulations that are distributed throughout a quadrant or with subpopulations in which there are undefined margins (tails) as in the case of CD25 staining.

Importantly, our data normalization approach is also useful for fully automated data analysis (14,15). Conventional "auto-gating" tools such as magnetic gates in FlowJo (Tree Star) and autopositioning gates in WinList (Verity Software House) were developed to provide dynamic adjustment of static gates to identify target cell populations across a set of samples, based on manual identification of population of interest. However, because these methods are restricted to 2D gates and they do not modify the underlying data, they cannot be used in high-throughput analysis such as those mentioned earlier. More importantly, they require the a priori identification of all populations of interest for adjustment, some of which may be unknown ahead of time for many research investigations. Finally, our experiments show that the autogating tools in FlowJo and WinList fail to automatically adjust the gates in samples with large population shifts and hence are not able to produce similar results presented in this manuscript.

In contrast, the data normalization methods we developed facilitate cell population labeling in high-throughput automated analysis pipelines. By adjusting populations in an automated fashion (i.e., without gating a representative of every cell population), the median positions of similar cell populations are in close proximity when these are later identified using automated gating tools such are flowClust (13) or FLAME (3). As a result matched cell populations can be given the same label during a meta-clustering step where the centers are grouped together based on their position.

The proposed normalization methods are based on four major assumptions: (i) with proper compensation, the individual fluorescence channels can be transformed individually; (ii) after debris removal, high density areas in the data represent biologically similar cell populations; (iii) multivariate high density areas show up in at least one univariate data projection as landmarks; and (iv) the median fluorescence intensity (MFI) of these high density regions is largely irrelevant and can be rescaled by an order-preserving transformation. Although there is broad consensus on (i) and (ii), the third and fourth assumptions depend on the design of the experiment and on the nature of the markers of interest. Many markers (e.g., CD3, CD4, and CD8) can be treated as binary (i.e., a particular cell either stains positive or negative for a given marker). In these settings, fluctuations in the actual MFI are not informative and basically reflect unintended between-sample variation. The signal of interest is the proportions of the negative and positive cell populations. Hence, removing the shifts in MFI, and thereby aligning the cell population of interest, does not result in a loss of signal. However, in some applications (16), the classification of cell populations is based on the actual MFI values as expression is tightly regulated as a

response to a certain stimulus. Currently, it is impossible to disambiguate the technical and the biological contributions to the between-sample variation in such cases. Clearly, the normalization methods proposed here cannot be applied to FCM data in such applications.

The results of the HLADr channel for the Immune Tolerance data present a more general problem. Rare populations are of increasing interest in many FCM experiments; however, they impose challenges to data analysts. As the number of cells in these populations is very small, they are not represented by high-density areas, and hence will not be targeted by our methods. The reliable identification of rare populations is problematic even in careful manual analyses, and the variation introduced by the subjectivity of the human investigator can be considerable. Clearly, more research is needed to provide computational solutions for these sorts of problems.

Acknowledgments

This research was performed as a project in collaboration with the Immune Tolerance Network, an international clinical research consortium headquartered at the University of California San Francisco.

Literature Cited

1. Oldaker TA. Quality control in clinical flow cytometry. Clin Lab Med 2007;27:671–685.
2. Bocsi J, Tarnok A. Toward automation of flow data analysis. Cytometry Part A 2008;73A:679–680.
3. Pyne S, Hu X, Wang K, Rossin E, Lin TI, Maier LM, Baecher-Allan C, McLachlan GJ, Tamayo P, Hafler DA, De Jager PL, Mesirov JP. Automated high-dimensional flow cytometric data analysis. Proc Nat Acad Sci USA 2009;106:8519–8524.
4. Hahne F. flowStats: Getting Started With FlowStats. 2009; note: R package version 1.0.13. URL: http://www.bioconductor.org/packages/2.5/bioc/vignettes/flowStats/inst/doc/GettingStartedWithFlowStats.pdf
5. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JY, Zhang J. Bioconductor: Open software development for computational biology and bioinformatics. Genome Biol 2004;5:R80.
6. Duong T, Cowling A, Koch I, Wand MP. Feature significance for multivariate kernel density estimation. Comput Stat Data Anal 2008;52:4225–4242.
7. Ramsay JO, Silverman BW. Functional Data Analysis. Springer; 1997; ISBN: 0-387-94956-9.
8. Ramsay JO, Li XC. Curve registration. J R Stat Soc Ser B 1998;60:351–363.
9. Silverman BW. Incorporating parametric effects into functional principal components analysis. J R Stat Soc Ser B 1995;57:673–689.
10. Ramsay JO. Estimating smooth monotone functions. J R Stat Soc Ser B 1998;60:365–375.
11. Ramsay JO, Wickham H, Graves S, Hooker G. fda: Functional Data Analysis. 2008; note: R package version 2.0.4. URL: http://cran.r-project.org/web/packages/fda/fda.pdf
12. R Development Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. 2009; URL: http://cran.r-project.org/doc/manuals/fullrefman.pdf
13. Lo K, Hahne F, Brinkman R, Gottardo R. flowclust: A bioconductor package for automated gating of flow cytometry data. BMC Bioinformatics 2009;10:145.
14. Hahne F, Le Meur N, Brinkman RR, Ellis B, Haaland P, Sarkar D, Spidlen J, Strain E, Gentleman R. flowcore: A bioconductor package for high throughput flow cytometry. BMC Bioinformatics 2009;10:106.
15. Lo K, Brinkman RR, Gottardo R. Automated gating of flow cytometry data via robust model-based clustering. Cytometry Part A 2008;73:321–332.
16. Zsuzsa H, Béla N JR, Flora K, Attila K, Janos KA. Mean fluorescence intensity rate is a useful marker in the detection of paroxysmal nocturnal hemoglobinuria clones. Clin Chem Lab Med 2005;43:919–923.