Automated Quadratic Characterization of Flow Cytometer Instrument Sensitivity * (flowQB Package: Introductory Processing Using Data NIH))

April 12, 2014

1 Licensing

Under the Artistic License, you are free to use and redistribute this software.

2 Background

Flow cytometer sensitivity has been well defined in the community and relates to two functions: (1) How well a dim staining population is resolved from an unstained population; (2) How well various dim staining populations can be distinguished from each other. The most common terms for characterizing flow cytometer sensitivity are known as Q (detector efficiency) and B (background light level), which are calculated using beads acquired on a cytometer. Any routine method to measure Q and B must be rapid and sufficiently accurate to provide useful results, see the works in [1, 2, 3].

Manually gating multi-peak bead data to generate the MFIs and SDs from the beads is an extremely time-consuming process. We showed the feasibility of an automated approach based on the clustering algorithm Kmeans to detect the bead sub-populations and to generate the MFIs and SDs in an easy and fully automatic rapid fashion. Furthermore, we extended the standard linear formulation, used to derive coefficients for Q and B calculation, with a quadratic term that is taking into account intrinsic variance from both the instrument and the bead product, see the work in [4].

Consequently we have developed a fully automated R Bioconductor package that we call flowQB. This document is intended to provide full access to

*This project was supported by the Terry Fox Foundation and the Terry Fox Research Institute and by grant 700374 from the Canadian Cancer Society, Toronto.

the R implementation of the theoretical descriptions of Q and B calculation in our manuscript [4] and to explain the generic functions that enable R as an informatics research platform for flow cytometer sensitivity:

- To calculate automatically the detector efficiency (Q), optical background (B), and electronic noise.
- To determine the optimal voltages for each fluorescence parameter when a series of voltages are applied to the photomultiplier tube (PMT) to setup the optimal separation and sensitivity, see the work in [4].
- **Methods** We propose a collection of R generic functions for automatic Q and B calculation. We have implemented these functions in the Bioconductor package flowQB. We illustrate their use with a number of case studies.
- **Results** We hope that these proposed R generic functions will become the base for the development of many tools for the Q and B calculation.
- **keywords** Flow Cytometry, High Throughput, Doublet, Instrument Sensitivity, Kmeans, Mean Fluorescence Intensity (MFI), Molecules of Equivalent Soluble Fluorochrome (MESF), linear and quadratic regressions, Q (detector efficiency) , B (background light level).

3 R QB Generic Functions

We define generic functions for our automatic Q and B calculations. A generic function is a standard R function with a special body to do an analytical calculation. We have four global generic functions to conduct an automatic Q and B calculation with the following utilities:

- Function ReadDD reads a given FCS file and remove doublet events for a given channel (ChanGiven), see section 4.
- Function KmeansMeanSD takes a given 2D array and generates a number of clusters and output their Means and SDs, see section 5.
- Function MFI2MESF converts the MFI means to MESF means. For instance, the MFI output of KmeansMeanSD are converted to MESF values. The SDs are also corrected for Illumination-Correction, see section 6.
- Functions lrMESF and qrMESF use the Means and the variances SD^2 in terms of the MESF values to conduct a linear and quadratic regression, see section 7.
- Function DiscriminantExamination uses the values in the output of the function qrMESF to estimate the discriminant of the resulting quadratic equation and can be used as an additional interpretation tool to aid in understanding cytometer sensitivity, see section 8.

- An advanced processing uses logicle transformation to transform the data as previously described in [**?**]. To identify bead singlets, it used the approach of gating on FSC & SSC (see [3, 2] for more information on doublet discrimination effects on Q and B calculations). It performed Kmeans clustering as an automated gating procedure in order to measure the bead MFIs and SDs. Two strategies are used: (1) 2D Kmeans processing, the selected marker with SSC or (2) Multi-level gating of a set of selected channels. The functions "gPS" and "rPS" are for 2D Kmeans processing. "gPS" function output Gaussian distribution fitting for peak statistics extraction, the fitting parameters means and standard deviations are the estimations of MFIs and SDs values. Estimations of MFIs and SDs values using Robust Statistics [**?**] is implemented in the function "rPS" in flowQB. Estimations of MFIs and SDs values using multi level Kmeans is implemented in the functions "MultilevelgPS" and "MultilevelrPS" in flowQB. The function "MultilevelrPS" uses robust statsitics and "MultilevelgPS" uses Gaussian distribution fitting R software "MASS" package to extract the statistics for the peak found by multi-Kmeans gating. We implemented the weighted least squares regression by using the function "qrWEIGHTEDMESF".

  The vignette *AdvancedflowQB.Rnw* has a detailed illustration how to use these functions: "gPS", "rPS" "MultilevelgPS" , "MultilevelrPS" and "qrWEIGHTEDMESF".

4 Doublet Discrimination Function

The function ReadDD reads a given FCS file and remove doublet events for a given channel (ChanGiven). When spherical beads are run through the cytometer, sometimes two pass too close together which results in a doublet reading. To exclude doublets from further analysis, a step of doublet discrimination is applied before moving onto the auto-gating step. A standard approach for doublet discrimination is taken here, see our manuscript [4] for more information on doublet discrimination effects on Q and B calculations.

Two channels Chan1DD and Chan2DD are used to detect the doublet events. For instance, the ratio of forward scatter height to forward scatter area is used $\left(\frac{\text{FSC.H}}{\text{FSC.A}}\right)$. The further this ratio is from 1, the more likely it is that the event considered is a doublet.

Let $P \in [50, 100]$ be the percentage of the singlet events to keep; i.e, only events with forward scattering between $\frac{P}{100} \leq \frac{FS.H}{FS.A} \leq 2 - \frac{P}{100}$ are extracted by the function ReadDD. The output is a 2D array $(ChanGiven, ChanCompanion)$ having the positive fluorescent intensities of channel of interest $ChanGiven$ and the companion channel $ChanCompanion$ which will be used to facilitate the 2D Kmeans clustering, for instance side scattering.

As the function ReadDD reads only a FCS file, we need to extract the FCS file from extdata folder:

```
> rm(list=ls(all=TRUE))
> library(flowQB)
> File= system.file("extdata","NIH.fcs",package="flowQB")
```

The function ReadDD reads first the FCS file $NIH.fcs$, using the read.FCS function in flowCore package. Second, the Forward Scattering Area index 1 and Forward Scattering Height index 2 are used to obtain singlet events with a Doublet Discrimination value equals to P=96. Finally, the processing returns a 2D singlet events for the channel of interest index 5 (B515), with the companion channel Side Scattering, index 3 (SS).

```
> P <- 96
> MFI2D <- ReadDD(File,1,2,P,5,3)
```

Note that $MFI2D$ is now a 2D array.

5 Kmeans to Extract Means and SDs

The function KmeansMeanSD takes a given 2D array, generates a number of clusters and outputs thr Means and SDs. For instance, the output of ReadDD is used by KmeansMeanSD to generate the required statistics for Q & B calcultaion.

K-means clustering in 2D is used as an automated gating procedure in order to measure the bead MFIs and SDs. K-means clustering is a method which partitions a set of data points (or events) into K clusters. Intuitively, each data point is assigned to the cluster whose mean value is closest to that of the data point in question. More formally, it aims to find the partition which minimizes the within-cluster sum of errors.

Function KmeansMeanSD returns the MFI Means and SDs of the 8 clusters.

```
> MFIMeansSDs=KmeansMeanSD(MFI2D,8,300,300,1)
```

Note that $MFIMeansSDs$ is now a 2D array having MFI Means and SDs. The first dimension is associated to the MFI Means:

```
> MFIMeansSDs[,1]
```

[1] 9.125472e+00 4.241918e+01 9.177818e+02 4.071844e+03 2.074090e+04 [6] 5.064784e+04 1.106388e+05 2.097744e+05

The second dimension is associated to the MFI SDs:

```
> MFIMeansSDs[,2]
```

[1] 4.757743 19.834325 107.087508 233.277625 599.271541 1160.629743 [7] 2074.637101 3903.530374

6 Mean Fluorescence Intensity (MFI) to Molecules of Equivalent Soluble Fluorochrome (MESF)

The function MFI2MESF converts the MFI means to MESF means. For instance, the MFI output of KmeansMeanSD are converted to MESF values. The SDs are also corrected for Illumination-Correction.

The MFI associated with each micro-sphere population (cluster) is a function of the assigned MESF value for that micro-sphere population. Since the cytometer is a linear device, the recorded intensity of the fluorescence pulses (and hence the MFIs) should be correlated linearly with the MESF values, which in turn are proportional to the number of fluorophores on the micro-sphere, and hence to the fluorescence signal, see our maniscript [4]. This gives the relation $MESF = p \times MFI$, where p is a constant.

The Q & B values will be defined in terms of the MESF values.

For MESF calculation, the constant conversion between MFI and MESF is set to $p = \frac{357217.00}{7102}$. MFIs are converted to MESFs without correcting the SDs as we set $IllCorrCV = 0$.

```
> p <- 357217.00/7102
> MFI2MSEF=MFI2MESF(MFIMeansSDs,p,0)
```

Note that MESF: MESF Mean and MESFV: MESF Variance (SD^2 or σ^2) and $MFI2MSEF$ is now a 2D array having MESF Means and SDs. The first dimension is associated to the MESF Means:

```
> MFI2MSEF[,1]
```

[1] 4.589938e+02 2.133604e+03 4.616267e+04 2.048060e+05 1.043228e+06 [6] 2.547489e+06 5.564920e+06 1.055125e+07

The second dimension is associated to the MESF Variances:

```
> MFI2MSEF[,2]
```

[1] 4.651414e+04 9.499463e+05 2.894907e+07 1.376055e+08 9.084653e+08 [6] 3.407793e+09 1.088878e+10 3.854907e+10

7 Linear and Quadratic Regressions

Let $(c_0, c_1, c_2) = (\sigma_E^2, \frac{1}{K}, \sigma_S^2)$ where σ_E^2 summarizes electronic sources of noise, K is the photon transfer constant for the PMTs and σ_S^2 incorporates intrinsic variance from both the cytometer instrument and the bead product, and any source of noise in general that depends linearly on the light signal.

We need to estimate the coefficients (c_0, c_1, c_2) in the following quadratic model:

$$\mathrm{SD}^2_{\mathrm{mesf}} = c_2 \times \mathrm{MESF}^2 + c_1 \times \mathrm{MESF} + c_0. \quad (1)$$

The coefficients of this quadratic model form two quantities of interest: Q $= \frac{1}{c_1}$ (detection efficiency) and B $= \frac{c_0}{c_1}$ (optical background or background noise), see the work in [4].

The R function for regression analysis *lm* is used to fit linear regression in the function lrMESF and the quadratic regression in the function qrMESF.

The functions lrMESF and qrMESF use the Means and SD_s^2 in terms of the MESF values to generate estimate (c_0^e, c_1^e, c_2^e).

The Q & B are estimated as $Q = \frac{1}{c_1^e}$ and $B = \frac{c_0^e}{c_1^e}$[1].

The MESF Means and SDs values in $MFI2MSEF$ are used to compute the Q value using the two regression approaches:

- Linear Regression: Only bead categories Dim1 (Peak1), Dim2 (Peak2) and Dim3 (Peak3), their MESF Means and SDs values will be extracted from $MFI2MSEF$, are used to estimate Q & B values. SDs in MFI2MSEF should be correct for illumination.
- Quadratic regression: the bead categories Dim1 (Peak1), Dim2 (Peak2), Dim3 (Peak3) and Dim4 (Peak4), their MESF Means and SDs values will be extracted from $MFI2MSEF$, are used to estimate Q & B values. Note the Quadratic regression is processed without correcting the SDs.

For quadratic Q and B calculation, the peaks of the cluster 3 to cluster 6 are extracted from $MFI2MSEF$ and used in the quadratic regression:

```
> QQB <- qrMESF(MFI2MSEF,3,6)
> OV <- c(Q=as.numeric(QQB[1]), B=as.numeric(QQB[2]), Rsquared=as.numeric(round(QQB[3], 2)))
> OV[1]
```

Q 0.001857138

```
> OV[2]
```

B 15194.25

```
> OV[3]
```

Rsquared 1

Note that $c1 = \frac{1}{Q}$, $c0 = \frac{B}{Q}$ and $c2 = sigmaS2$.

For linear Q and B calculation, the peaks of the cluster 3 to cluster 5 are extracted from $MFI2MSEF$ and used to compute the linear regression coefficients. As we need the illumination correction MFIs are converted to MESFs and SDs are corrected using the beads in cluster 8.

```
> IllCorrCV <- MFIMeansSDs[8,2]/MFIMeansSDs[8,1]
> MFI2MSEF <- MFI2MESF(MFIMeansSDs,p,IllCorrCV)
> LQB <- lrMESF(MFI2MSEF,3,5)
> OV <- c(Q=as.numeric(LQB[1]), B=as.numeric(LQB[2]), Rsquared=as.numeric(round(LQB[3], 2)))
> OV[1]
```

[1] If $c_1^e = 0$, the functions will generate a stop(*"No linear term to calculate Q and B"*)

Q 0.002003786

```
> OV[2]
```

B 24735.08

```
> OV[3]
```

Rsquared 1

```
>
```

Note that $c1 = \frac{1}{Q}$, $c0 = \frac{B}{Q}$.

8 Discriminant Examination

We examine the discriminant of the quadratic equation:

$$\mathrm{SD}^2_{\mathrm{mesf}} = c_2^e \times \mathrm{MESF}^2 + c_1^e \times \mathrm{MESF} + c_0^e. \tag{2}$$

Let $\Delta = (c_1^e)^2 - 4c_0^e c_2^e$, there are two possible scenarios: $\Delta \geq 0$ and $\Delta < 0$.

The calculation of Δ is implemented in the flowQB package, so that its estimate can be used as an additional interpretation tool to aid in understanding cytometer sensitivity:

- If $\Delta \geq 0$, the larger the variation product of (σ_E) and (σ_S), the lower the upper bound on the detection efficiency Q.
- If $\Delta < 0$, the lower the variation product of (σ_E) and (σ_S), the greater the upper bound on the detection efficiency Q.

For more details about the physical interpretation of Δ values, see the work in [4].

Discriminant of the Quadratic Equation:

```
> Coefs <- DiscriminantExamination(as.numeric(QQB[1]),as.numeric(QQB[2]),as.numeric(QQB[4]))
> Coefs[1]
```

c0 8181544

```
> Coefs[2]
```

c1 538.4631

```
> Coefs[3]
```

c2 0.0003124441

```
> Coefs[4]
```

Delta 279717.4

```
> Delta <- Coefs[4]
> if(Delta >= 0)
+ {
+ cat(paste("The sign of the discriminant is positive with the value", round(Delta,2), ". ")
+ cat("The larger the variation product of (sigmaE2) and (sigmaS2), ")
+ cat("the lower the upper bound on the detection efficiency Q. ")
+ }
```

The sign of the discriminant is positive with the value 279717.38 . The larger the variation product of (sigmaE2) and (sigmaS2), the lower the upper bound on the detection efficiency Q.

```
> if(Delta < 0)
+ {
+ cat(paste("The sign of the discriminant is negative with the value", round(Delta,2), ". ")
+ cat("The lower the variation product of (sigmaE2) and (sigmaS2), ")
+ cat("the greater the upper bound on the detection efficiency Q. ")
+ }
>
```

References

[1] J. Wood, *Fundamental Flow Cytometer Properties Governing Sensitivity and Resolution*, Cytometry 33, (1998), p. 260 - 6.

[2] R. Hoffman and J. Wood, *Characterization of Flow Cytometer Instrument Sensitivity*, Current Protocols in Cytometry, Chapter 1: Unit 1.20 (2007).

[3] E. Chase and R. Hoffman, *Resolution of Dimly Fluorescent Particles: a Practical Measure of Fluorescence Sensitivity*, Cytometry 33 (1998), p. 267-279.

[4] F. El Khettabi et al. 2013, *Automated Quadratic Characterization of Flow Cytometer Instrument Sensitivit*, to be submitted.