BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 14 2014, pages 2060–2065
doi:10.1093/bioinformatics/btu127

Bioimage informatics

Advance Access publication March 27, 2014

flowFit: a Bioconductor package to estimate proliferation in cell-tracking dye studies

Davide Rambaldi[1,*], Salvatore Pece[1,2] and Pier Paolo Di Fiore[1,2,*]

[1]Department of Experimental Oncology, Istituto Europeo di Oncologia, 20141 Milan and [2]Dipartimento di Scienze della Salute, Università degli Studi di Milano, 20122 Milan, Italy

Associate Editor: Jonathan Wren

ABSTRACT

Summary: Herein we introduce flowFit, a Bioconductor package designed to perform quantitative analysis of cell proliferation in tracking dye-based experiments. The software, distributed as an R Bioconductor library, is based on a mathematical model that takes into account the height of each peak, the size and position of the parental population (labeled but not proliferating) and the estimated distance between the brightness of a cell and the brightness of its daughter (in which the dye is assumed to undergo a 2-fold dilution). Although the algorithm does not make any inference on cell types, rates of cell divisions or rates of cell death, it deconvolutes the actual collected data into a set of peaks, whereby each peak corresponds to a subpopulation of cells that have divided N times. We validated flowFit by retrospective analysis of published proliferation-tracking experiments and demonstrated that the algorithm predicts the same percentage of cells/generation either in samples with discernible peaks (in which the peaks are visible in the collected raw data) or in samples with non-discernible peaks (in which the peaks are fused together). To the best of our knowledge, flowFit represents the first open-source algorithm in its category and might be applied to numerous areas of cell biology in which quantitative deconvolution of tracking dye-based experiments is desired, including stem cell research.

Availability and implementation: http://www.bioconductor.org/packages/devel/bioc/html/flowFit.html (Bioconductor software page). http://www.bioconductor.org/packages/2.13/bioc/vignettes/flowFit/inst/doc/HowTo-flowFit.pdf (package vignette). http://rpubs.com/tucano/flowFit (online tutorial).

Contact: pierpaolo.difiore@ifom.eu or davide.rambaldi@gmail.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on August 1, 2013; revised on February 14, 2014; accepted on February 28, 2014

1 INTRODUCTION

In cell proliferation-tracking experiments, cells are stained—before being cultured under various conditions—with a fluorescent dye to follow proliferation kinetics. At each mitosis, the dye is roughly equally divided between daughter cells, so that each daughter is about half as fluorescent as the mother. At any given moment, therefore, the fluorescence intensity of a cell bears witness to its divisional history when compared with the intensity of the parental population.

In a typical experimental pipeline, cells are first labeled with a tracking dye, then allowed to proliferate (frequently in the presence of specific stimuli) and finally analyzed by flow cytometry by collecting data for (i) unlabeled cells (used as a negative control), (ii) unstimulated labeled cells (the non-proliferating population) and (iii) sample(s) of labeled and stimulated cells (the proliferating populations). If the characteristics of the logarithmic amplifier on the flow cytometer are known, it is possible to derive, from a histogram of fluorescence intensity, the proportion of cells that have undergone any particular number of divisions (Givan et al., 1999; 2004). Generally, a flow cytometer with 1024 channels (channels range) represents a nominal range of 4 log decades.

Several fluorochromes have been used to track cell proliferation (Parish, 1999), including DNA-binding fluorescent dyes (Hoechst 33342, thiazole orange), cytoplasmically distributed fluorescent dyes (calcein, 2′,7′-bis-(2-carboxyethyl)-5(6)-carboxyfluorescein), covalent coupling fluorescent dyes (Carboxyfluorescein succinimidyl ester, fluorescein isothiocyanate) and membrane-inserting fluorescent dyes (PKH26, CellVue Lavander). Although each dye displays advantages and disadvantages for individual experimental purposes [reviewed in (Parish, 1999)], a common problem is represented by the quantification of the fluorescence signal: under optimal conditions some dyes [e.g. carboxyfluorescein diacetate succinimidyl ester (CFSE)] form separate peaks defining different proliferative generations, whereas others (e.g. PKH26) do not (Givan *et al.*, 2004). A good proliferation-tracking algorithm should be able to give similar results in both cases.

Here we introduce flowFit, an R (R Development Core Team, 2012) Bioconductor library (Gentleman *et al.*, 2004) that fits a set of peaks (corresponding to different generations of cells) to the histogram of fluorescence intensity acquired during a fluorescence-activated cell sorting (FACS) experiment. The package is integrated with the Bioconductor libraries used for the analysis of flow cytometry datasets: flowCore (Hahne *et al.*, 2009) and flowViz (Sarkar *et al.*, 2008). The performance of flowFit was evaluated by the retrospective analysis of a published dataset of dye-tracking experiments (Quah and Parish, 2012), in which lymphocytes were stained with three different proliferation-tracking dyes: CFSE, cell proliferation dye eFluor 670 (CPD) and CellTrace Violet (CTV). In particular, we meta-analyzed a subpopulation of CD4+ lymphocytes labeled with the three different staining reagents. The three samples were allowed to proliferate under identical growth conditions; however, differences in the characteristics of the dyes yielded rather different

*To whom correspondence should be addressed.

fluorescence histograms with different peak definitions (Quah and Parish, 2012). FlowFit was able to deconvolute equally well in all situations and to estimate the same percentage of cells per generation under all conditions.

2 METHODS

2.1 Model

The proliferation-fitting algorithm uses an R implementation of the Levenberg–Marquardt algorithm (Elzhov *et al.*, 2012) to fit a set of N peaks to the histogram of the fluorescence intensity from the acquired flow cytometry data file (FCS file). The Levenberg–Marquardt algorithm provides a numerical solution to the problem of minimizing a function over a space of parameters. It is an iterative technique that locates the minimum of a multivariate function. The 'minimum' is expressed as the sum of the squares of non-linear real-valued functions (Levenberg, 1944). The algorithm adjusts the function parameters to reduce the sum of squares (residuals) between the real data and the model. When the current solution is far from the initial guess of parameters, the algorithm behaves as a steepest descendent method. When the current solution is close to the correct solution, it becomes a Gauss–Newton method (Björck, 1996). The proliferation-fitting algorithm fits an initial single peak on the parental population (cells labeled and unstimulated), according to this formula:

$$a^2 e \frac{(x-\mu)^2}{2s^2} \tag{1}$$

where a^2 is proportional to the height of the peak, μ is the peak position on the FACS scale and s is proportional to the peak size (for the parental population, it corresponds to the variance in the initial staining). The formula for the next peak (corresponding to the cells that have divided once) will be:

$$b^2 e \frac{(x-(\mu-D))^2}{2s^2} \tag{2}$$

where b^2 is proportional to the height of the peak, and D is the distance between two generations of cells.

The distance between two cell generations is defined as the distance between a mother cell and its progeny (that contains half of the amount of dye present in the mother). This distance is constant on a logarithmic scale and depends on the number of data points analyzed by the FACS instrument and on the range of log decades.

It is possible to convert the FACS fluorescence intensity (FFI) recorded by the instrument into the relative fluorescence intensity (RFI) expressed in molecules of equivalent fluorochrome (Spherotech, 2012) using the following formula:

$$RFI = 10^{\frac{FFI*l}{c}} \tag{3}$$

The inverse formula can be used to convert RFI to FFI:

$$FFI = \frac{c^* log(RFI)}{(l^* log(10))} \tag{4}$$

where:

(1) RFI is the relative fluorescent intensity
(2) FFI is the fluorescence intensity on the FACS scale
(3) l is the number of log decades in the FACS instrument
(4) c is the number of data points (channels) in the instrument

Using these formulas, it is possible to estimate the spacing between generations of cells on the FACS instrument scale: we first convert the peak position of the parental population (unstimulated population) into RFI, then we calculate the RFI of the first-generation cells (RFI/2) and, finally, we convert back the RFI into FFI; the difference between FFI of the parental population and FFI of the first generation of cells represents the distance between generations (D). In the flowFit library, this spacing is automatically computed with the function *generationsDistance*. The number of log decades on the instrument can be estimated by converting the linear scale of the detector into a log scale:

$$l = log(R) \tag{5}$$

where R is the acquisition resolution (data range for the detector).

The *proliferationFitting* and *parentFitting* functions automatically compute the log decades from the keywords in the FCS file or using the logarithm of $R[log(R)]$. If the functions find a 'log decades' keyword for the current detector in the FCS file, they use the value in the keyword; otherwise, log decades are estimated from the detector acquisition resolution.

The algorithm automatically computes the number of peaks to be fitted on the proliferating population using the data range for the detector: first, the distance between two generations of cells is estimated using the *generationsDistance* function; then the maximum number of peaks that can be fitted on the FACS instrument scale is estimated with the following formula:

$$N = \frac{D}{c} \tag{6}$$

where:

(1) N is the number of peaks to fit on the dataset
(2) D is the distance between two generations of cells
(3) c is the number of data points (channels) in the instrument

The formula for a model with three peaks will be:

$$M = a^2 e \frac{(x-\mu)^2}{2s^2} + b^2 e \frac{(x-(\mu-D))^2}{2s^2} + c^2 e \frac{(x-(\mu-2D))^2}{2s^2} \tag{7}$$

where the parameters a, b and c are proportional to the height of the peaks.

In the 'fixed model' (options fixedModel = TRUE), the Levenberg–Marquardt algorithm estimates the height of each peak but allows the user to keep one or more of these variables constant: parental peak position (μ), parental peak size (s) and the distance between two generations (D); the variables added to the list of *fixedPars* will be kept constant during the algorithm iterations. In the 'dynamic model' (options fixedModel = FALSE), the Levenberg–Marquardt algorithm uses as parameters all the variables in the model. In general terms, the dynamic model best suits the analysis of samples showing discernible peaks. Conversely, when the sample features poorly defined or no discernible peaks and the model is unable to accommodate the peaks, it is preferable to keep those variables fixed.

Once the height of each peak is estimated, it can be used to estimate the number of cells per generation through a numerical integration of (7):

$$I_{all} = \int_v^W M \tag{8}$$

where v and W are the lower and upper limits of the FACS instrument scale, respectively. For each generation, the number of cells in the peak can be computed through another numerical integration:

$$I_i = \int_v^w \left(h_i^2 e \frac{(x-(\mu-(i-1)D))^2}{2s^2} \right) \tag{9}$$

and the percentage of cells in a peak can be estimated as:

$$P_{g_i} = \frac{I_i}{I_{all}} * 100 \tag{10}$$

To evaluate the proliferation in the sample, flowFit uses the *proliferationIndex* function: the proliferation index is calculated as the sum of the cells in all generations—including the parental—divided by

the computed number of original parental cells theoretically present at the beginning of the experiment (Munson, 2010). The proliferation index is a measure of the fold increase in cell number in culture, over the course of the experiment:

$$\frac{\sum_{i=0}^{I} N_i}{\sum_{i=0}^{I} N_i / 2^i} \tag{11}$$

where i is the generation number (parental generation = 0). In the absence of proliferation, when all cells are in the parental generation, the formula becomes

$$\frac{N_0}{N_0/2^0} = 1 \tag{12}$$

defining the lower limit of the proliferation index.

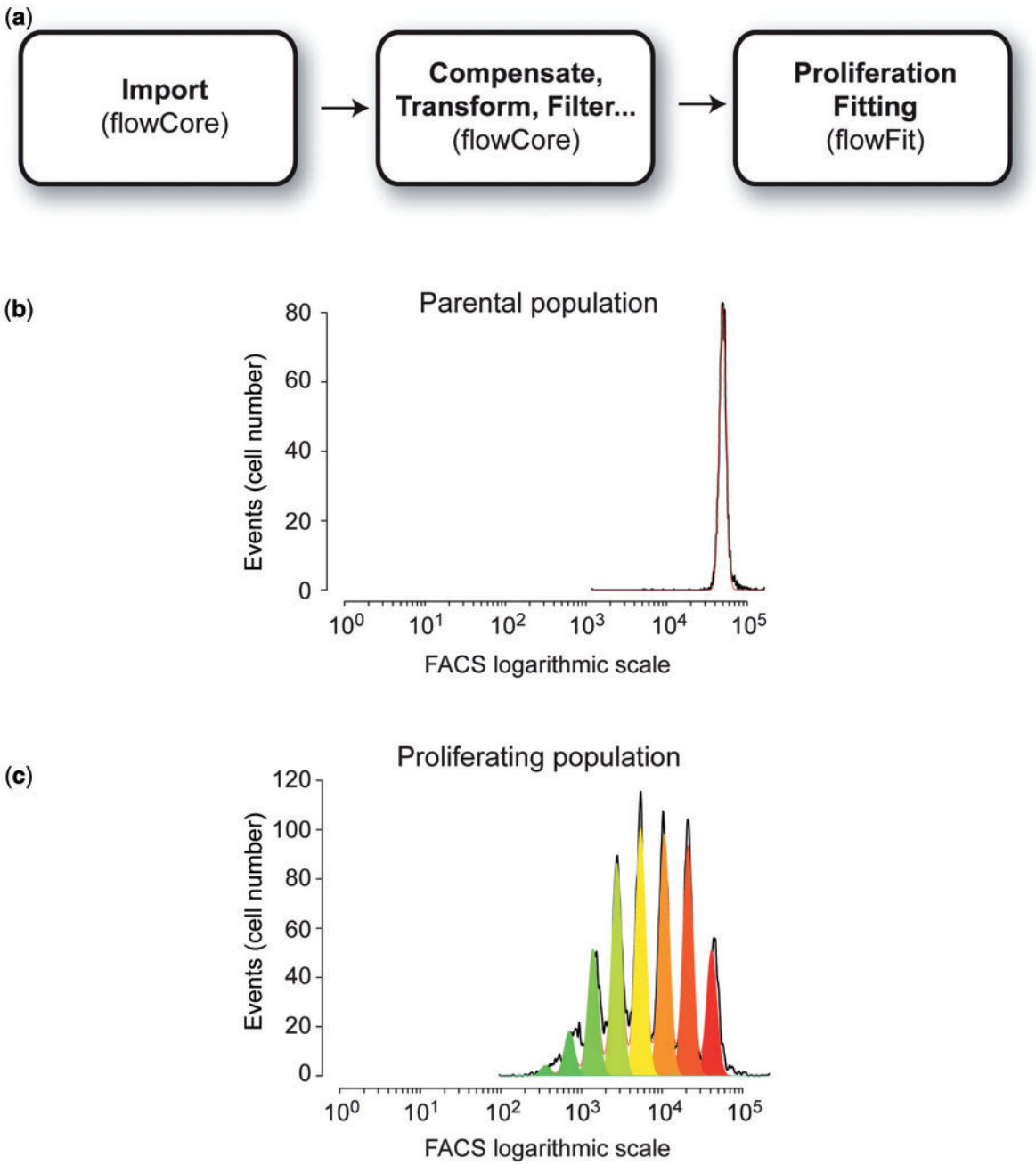

Fig. 1. The flowFit pipeline: (**a**) The flowCore methods are used to import and transform the FACS data. When the raw data are gated and transformed, it is possible to apply the proliferation-fitting algorithm. (**b**) Graphical output example for the function *parentFitting*: the position and size of the first peak are established. (**c**) Graphical output example for the function *proliferationFitting*: the estimated parental peak position and size are used to perform a fit on the proliferating populations

2.2 Implementation

The flowFit is distributed in the R software language (R Development Core Team, 2012) as a Bioconductor (Gentleman *et al.*, 2004) package. The package is freely available and distributed under the Artistic 2.0 License (Open Source Initiative, 2013). The flowFit package is composed of two main functions (*parentFitting* and *proliferationFitting*) and five secondary functions (*logTicks*, *proliferationIndex*, *generationsDistance*, *proliferationGrid* and *getGenerations*).

The typical pipeline involves a first step (Fig. 1a) where the FACS raw data are imported, transformed and gated with the flowCore library (Hahne *et al.*, 2009). By applying the function *parentFitting* to the sample representing the parental population, the position and size of the first peak are then established (Fig. 1b). Finally, the estimated parental peak position and size are used to perform a fit on the proliferating population with the function *proliferationFitting* (Fig. 1c). The position and size of the parental population can be used as an initial guess for the Levenberg–Marquardt algorithm (dynamic model); in this case, the algorithm will adjust these variables, in the proliferating population, as needed to achieve optimal fitting. Alternatively, position and size of the parental population can be used as constants (fixed model). In general, when the input data have poorly defined cell division peaks, the fixed model is preferable.

Because tracking dyes are partitioned roughly equally between daughter cells, it is possible to assume that the size of cell division peaks in the data is mainly due to the variance in the initial staining. For this reason, the proliferation-fitting algorithm uses the width of the parental peak to calculate the size of each peak in the final population [s in Equation (7)].

FlowFit depends on flowCore, a Bioconductor package that provides data structures and basic functions to deal with FCS data. The user will load the data and gate them with the flowCore functions to select the correct population and to remove aggregates and debris. After these preprocessing steps, the user will use the flowFit library to perform the proliferation fitting. The algorithm is not able to detect ploidy patterns.

3 RESULTS

The proliferation-fitting algorithm was tested by retrospective analysis of a published dataset consisting of three samples of mouse lymphocytes stained with CFSE, CTV and CPD (Quah and Parish, 2012). We focused on the raw data used to generate Figure 2a of the original paper. This subset of raw data contains four samples:

(1) CD4+ non-stimulated lymphocytes labeled with CFSE, CPD and CTV (parental population)
(2) CD4+ lymphocytes labeled with CFSE and stimulated
(3) CD4+ lymphocytes labeled with CPD and stimulated
(4) CD4+ lymphocytes labeled with CTV and stimulated

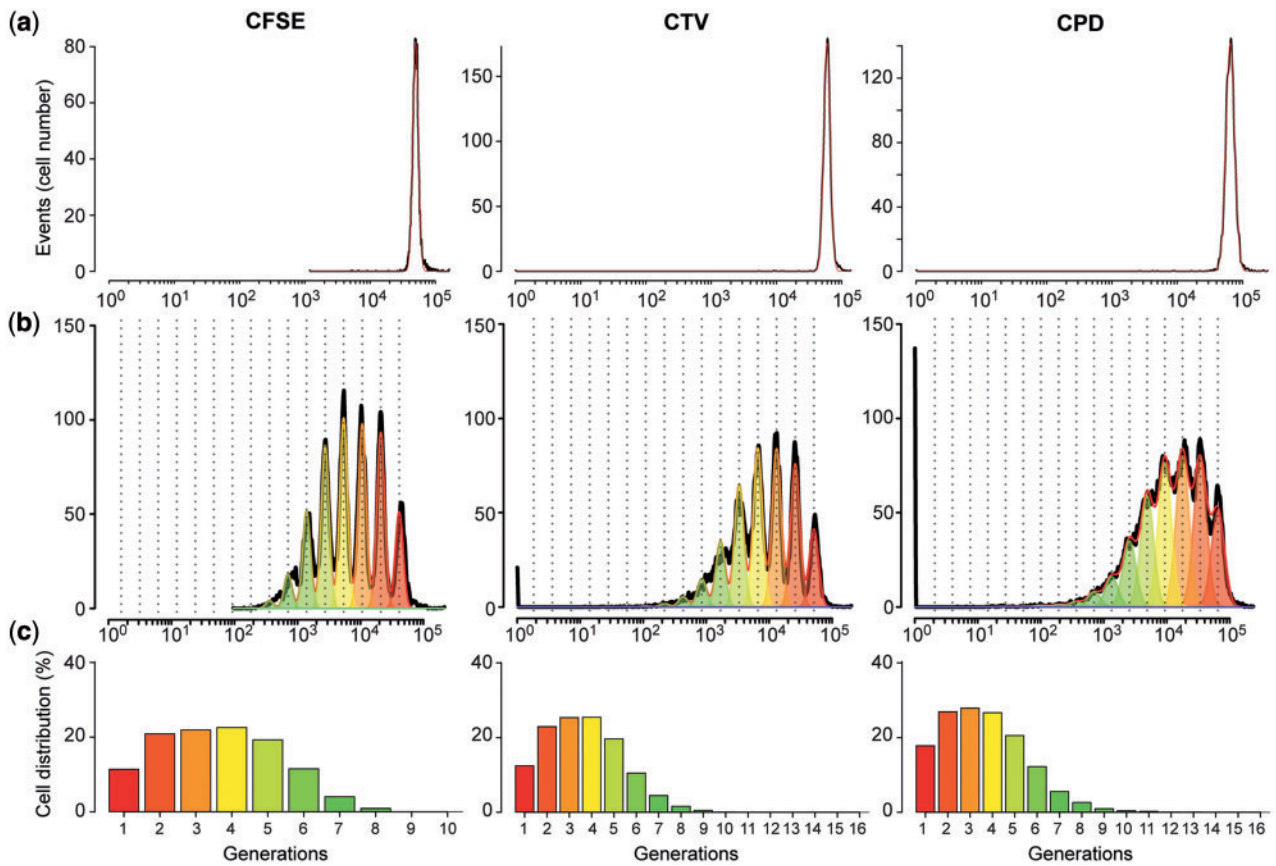

Fig. 2. The flowFit on the Quah and Parish dataset. (**a**) Histogram plots and fittings for the unstimulated (parental) populations. (**b**) Histogram plots for stimulated populations (stained with CFSE, CTV and CPD, as indicated on the top of the figure). (**c**) Bar graphs representing the distribution of cells/generation in the three samples

The datasets, already gated for CD4+ lymphocytes, were log10 transformed after truncation at 1. Then, we performed three *parentFitting* on the 'parental sample' to establish the position and size of the parental population for the three dyes (Fig. 2a). Using the parental peak position and size as a best guess, we performed a *proliferationFitting* with the dynamic model on the CFSE and CTV samples. In the CPD sample, the stimulated population did not display well-separated peaks, at variance with the CTV and CFSE samples. Thus, for the CPD sample, we used a fixed model (*proliferationFitting* with fixedModel = TRUE), keeping the parental peak position fixed (Fig. 2b and c). Finally, we compared the percentage of cells per generation in the three samples and evaluated the concordance among the fittings. Using the chi-squared test, we did not observe any significant difference in the distribution of cells per generation among the three estimates ($p = 1$), arguing that the algorithm is correctly estimating the percentage of cells per generation in the three samples (Supplementary Fig. S1). Altogether, the *proliferationFitting* analysis performed on the datasets published by Quah and Parish, which yielded similar results compared with those of the original study, argues for the suitability of flowFit to the analysis of cell proliferation in tracking dye experiments with either poorly or well-discernible peaks, based on the dual possibility to fit them, respectively, with the dynamic model or with the fixed model.

We also performed an *in silico* analysis to benchmark the performance of the algorithm with random samples (Table 1 and Supplementary Materials Section 3, including Supplementary Figs. S7–S21) for the dynamic and fixed models. In general, the dynamic model performed well with samples where the parental peak size (that corresponds to the variance in the initial staining) is small and the proliferating population forms well-separated peaks. The fixed model, on the other hand, performs well also with samples with a poorly defined cell division peaks (cf: Supplementary Figs. S8 and S14).

4 DISCUSSION

The objective of this work was to provide an open-source algorithm to quantify the proliferation of cells in tracking dye experiments. Of note, although the proliferation-fitting algorithms currently available for this type of analyses (ModFit LT^{TM}, FCS Express and FlowJO) are commercial software, flowFit is already publicly available and published with the Artistic 2.0 License in the Bioconductor R framework. We provided statistical demonstration of the quality of the fitting with both *in silico* analysis and real dataset analysis. In particular, we demonstrated the intrinsic versatility of flowFit in providing reliable results in the analysis of samples with both discernible and non-discernible peaks. At variance with ModFit LT^{TM}, which uses a statistical analysis of the fluorescence histogram to identify potential peaks and a Gaussian model for the single peak, flowFit uses a Levenberg–Marquadt algorithm to minimize the difference between the observed data and the model, and uses a non-Gaussian formula (1) to define a peak. In addition, flowFit is fully integrated in the Bioconductor R framework.

In recent years, several approaches, based on dye tracking, have been developed to identify and isolate adult stem cells (SCs) from normal and tumor tissues (e.g. Cicalese *et al.*, 2009; Pece *et al.*, 2010). These methodologies rely on the fact that—in the normal adult SC compartment—SCs divide asymmetrically to generate two daughter cells with opposite proliferative and differentiative fates: one of the two daughter cells retains the SC fate and withdraws into quiescence, whereas the other becomes a progenitor and enters a tumultuous phase of repeated symmetric divisions, required to expand the progenitor compartment, followed by terminal lineage differentiation. One characteristic of the SC compartment in cancer is that the cancer SCs tend to skip rounds of asymmetric division, in favor of symmetric divisions, after which both daughters divide again (Morrison and Kimble, 2006). This leads to an expansion of the SC compartment. The 'rate of skipping' seems to be higher in poorly differentiated more aggressive tumors (Cicalese *et al.*, 2009; Pece *et al.*, 2010). Methodologies to follow the kinetics of SC division *in vitro* and to compare the rate of symmetric versus asymmetric divisions are needed if we want to understand the biological bases of homeostasis of normal SC compartments and of how expansion of the SC compartment drives cancer growth and aggressiveness. We envision this as one of the major areas of application for flowFit. To this aim, the integration of flowFit with a stochastic model of cell proliferation based on a Gillespie algorithm will likely be required to analyze in real tumor tissues the proliferation kinetics of cancer SCs compared with that of their tumor progeny, and to compare proliferation data from these populations with relevant theoretical models of proliferation (for instance, models of asymmetric versus symmetric mode of division in the cancer SC versus the transient-amplifying compartment of progenitors). This approach will probably be instrumental to identifying parameters in the stochastic model able to yield the best overlap between real data and simulations.

Table 1. *In silico* benchmark results

Benchmark	Varying peak size	Random samples (%)
Dynamic model	0.17	80.5
Fixed model	0.4	100

Note: Summary of *in silico* analysis: random samples with different levels of division peak quality (Supplementary Section 2.1). The maximum peak width that gives reliable results (Column: *varying peak size, 017*) correspond to a parent population with 75% of the events distributed across 25.7 channels (in a FACS with 1024 channels and 4 log decades), a 'peak size' of 0.4 corresponds to a parent population peak with 75% of the events distributed across 61.7 channels (in a FACS with 1024 channels and 4 log decades); random samples with different distribution of cells per generation (Supplementary Section 2.2): the percentage of samples that gives reliable results is depicted (Column: *random samples*).

ACKNOWLEDGEMENTS

The authors thank B.J.C. Quah and C.R. Parish for providing the original datasets from the paper: 'new and improved methods for measuring lymphocyte proliferation *in vitro* and *in vivo* using CFSE-like fluorescent dyes'.

They also thank A.Cocito, B.Alvarez-Moya and M.Ferrero-Gimeno for useful discussions.

Funding: The Associazione Italiana per la Ricerca sul Cancro (AIRC) (to S.P. and P.P.D.F.); from the Italian Ministries of Education-University-Research (MIUR) and of Health (to S.P. and P.P.D.F.); from the Monzino Foundation and the European Research Council (to P.P.D.F.); from the CARIPLO Foundation (to S.P. and P.P.D.F.); and from the G. Vollaro Foundation (to S.P.).

Conflict of Interest: none declared.

REFERENCES

Björck,Å. (1996) *Numerical Methods for Least Squares Problems*. SIAM.

Cicalese,A. *et al*. (2009) The tumor suppressor p53 regulates polarity of self-renewing divisions in mammary stem cells. *Cell*, **138**, 1083–1095.

Elzhov,T.V. *et al*. (2012) minpack.lm: R interface to the Levenberg-Marquardt nonlinear least-squares algorithm found in MINPACK, plus support for bounds.

Gentleman,R.C. *et al*. (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*., **5**, R80.

Givan,A.L. *et al*. (1999) A flow cytometric method to estimate the precursor frequencies of cells proliferating in response to specific antigens. *J. Immunol. Methods*, **230**, 99–112.

Givan,A.L. *et al*. (2004) Use of cell-tracking dyes to determine proliferation precursor frequencies of antigen-specific T cells. *Methods Mol. Biol*., **263**, 109–124.

Hahne,F. *et al*. (2009) flowCore: a Bioconductor package for high throughput flow cytometry. *BMC Bioinformatics*, **10**, 106.

Levenberg,K. (1944) A method for the solution of certain non-linear problems in least squares. *Quarterly Appl. Math*., **2**, 164–168.

Morrison,S.J. and Kimble,J. (2006) Asymmetric and symmetric stem-cell divisions in development and cancer. *Nature*, **441**, 1068–1074.

Munson,M.E. (2010) An improved technique for calculating relative response in cellular proliferation experiments. *Cytometry A*, **77**, 909–910.

Open Source Initiative. (2013) Artistic Licence 2.0.

Parish,C.R. (1999) Fluorescent dyes for lymphocyte migration and proliferation studies. *Immunol. Cell Biol*., **77**, 499–508.

Pece,S. *et al*. (2010) Biological and molecular heterogeneity of breast cancers correlates with their cancer stem cell content. *Cell*, **140**, 62–73.

Quah,B.J.C. and Parish,C.R. (2012) New and improved methods for measuring lymphocyte proliferation *in vitro* and *in vivo* using CFSE-like fluorescent dyes. *J. Immunol. Methods*, **379**, 1–14.

R Development Core Team. (2012) *R: A Language and Environment for Statistical Computing*.

Sarkar,D. *et al*. (2008) Using flowViz to visualize flow cytometry data. *Bioinformatics*, **24**, 878–879.

Spherotech. (2012) Measuring molecules of equivalent fluorochrome (MEF) using spherotm rainbow and ultra rainbow calibration particles.