Bioinformatics, 31(10), 2015, 1623–1631
doi: 10.1093/bioinformatics/btv008
Advance Access Publication Date: 18 January 2015
Original Paper

Systems biology

Deep profiling of multitube flow cytometry data

Kieran O'Neill[1,2], Nima Aghaeepour[1], Jeremy Parker[1], Donna Hogge[1], Aly Karsan[1], Bakul Dalal[3] and Ryan R. Brinkman[1,4,*]

[1]Terry Fox Laboratory, BC Cancer Agency, [2]Bioinformatics Graduate Program, University of British Columbia, [3]Department of Hematopathology, Vancouver General Hospital and [4]Faculty of Medical Genetics, University of British Columbia, Vancouver, Canada

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on August 27, 2014; revised on December 19, 2014; accepted on January 2, 2015

Abstract

Motivation: Deep profiling the phenotypic landscape of tissues using high-throughput flow cytometry (FCM) can provide important new insights into the interplay of cells in both healthy and diseased tissue. But often, especially in clinical settings, the cytometer cannot measure all the desired markers in a single aliquot. In these cases, tissue is separated into independently analysed samples, leaving a need to electronically recombine these to increase dimensionality. Nearest-neighbour (NN) based imputation fulfils this need but can produce artificial subpopulations. Clustering-based NNs can reduce these, but requires prior domain knowledge to be able to parameterize the clustering, so is unsuited to discovery settings.

Results: We present flowBin, a parameterization-free method for combining multitube FCM data into a higher-dimensional form suitable for deep profiling and discovery. FlowBin allocates cells to bins defined by the common markers across tubes in a multitube experiment, then computes aggregate expression for each bin within each tube, to create a matrix of expression of all markers assayed in each tube. We show, using simulated multitube data, that flowType analysis of flowBin output reproduces the results of that same analysis on the original data for cell types of >10% abundance. We used flowBin in conjunction with classifiers to distinguish normal from cancerous cells. We used flowBin together with flowType and RchyOptimyx to profile the immunophenotypic landscape of NPM1-mutated acute myeloid leukemia, and present a series of novel cell types associated with that mutation.

Availability and implementation: FlowBin is available in Bioconductor under the Artistic 2.0 free open source license. All data used are available in FlowRepository under accessions: FR-FCM-ZZYA, FR-FCM-ZZZK and FR-FCM-ZZES.

Contact: rbrinkman@bccrc.ca.

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

1 Introduction

Flow cytometry (FCM) immunophenotyping is a powerful and high-throughput analytical technique allowing the rapid quantification of proteins on cells in suspension on a per-cell basis (Craig and Foon, 2008). Today, it is a critical step in both research and clinical decision making for leukemias (Craig and Foon, 2008; Swerdlow *et al.*, 2008; Wood *et al.*, 2007), human immunodeficiency virus (HIV; De Rosa *et al.*, 2001) and a host of other diseases. However, a major limitation is that data are typically analysed manually. Although the combinatorial space in FCM data are often vast, investigators rely primarily on intuition to guide their analysis, which may be error-prone and difficult to reproduce (Aghaeepour *et al.*, 2013; O'Neill *et al.*, 2013). Many prominent users of the technology have called for improved techniques and software for automating FCM data analysis (Chattopadhyay *et al.*, 2008; Finn, 2009; Robinson *et al.*, 2012).

In answer, the bioinformatics community has developed tools for computational deep profiling of high-dimensional FCM data (Bendall and Nolan, 2012; Bendall *et al.*, 2012). One example is flowType (Aghaeepour *et al.*, 2012a; O'Neill *et al.*, 2014), which exhaustively stratifies all combinations of markers, and its sister package RchyOptimyx (Aghaeepour *et al.*, 2012b; O'Neill *et al.*, 2014), which finds those cell types most important to external outcomes such as disease state or patient survival, and distills these to their simplest possible form. Another example is Spanning-tree Progression Analysis of Density-normalized Events (SPADE), which clusters cells multidimensionally, then maps those clusters into a 2D representation using a minimum spanning tree algorithm (Qiu *et al.*, 2011). These and other approaches have recently been compared and found to be extremely effective as part of a classification pipeline predicting acute myeloid leukemia (AML) from healthy patients based on FCM data (Aghaeepour *et al.*, 2013).

However, in many cases, the number of proteins needing to be assayed exceeds the number that the cytometer available can measure in a single run. Furthermore, it is often essential, especially in clinical testing, to use negative controls (either unstained or isotype) to counteract technical variation across samples (Maecker and Trotter, 2006). As a solution to this problem, standard practice is to aliquot a sample into multiple tubes, each of which is run to assay overlapping subsets of the total set of desired of proteins. This process is common for modern clinical diagnostic FCM data; especially, when immunophenotyping leukemias, where the standard method is to include the pan-leukocyte marker (CD45) in each tube, and use this in combination with right angle scattered light (side-scattered light; SSC) to identify leukemic blasts in each tube separately (Lacombe *et al.*, 1997). For example, the current standard for leukemia diagnosis established by the EuroFlow consortium recommends an 8-colour multitube panel of overlapping reagents (van Dongen *et al.*, 2012).

Without some means of combining tubes together, existing techniques for deep profiling can only be applied serially to each tube in a multitube FCM assay, which results in a substantial loss in depth. This can be illustrated with an example: consider an assay with six tubes containing six markers each, with two of those markers overlapping (being present) in every tube. The complete number of distinct markers will be $2 + 4 \times 6 = 26$. When examining a binary division of each marker into positive and negative expression, the total number of possible cell types present is $3^{26} \approx 2.5 \times 10^{12}$ (Aghaeepour *et al.*, 2012a). However, working one tube at a time, only 3^6 cell types can be elucidated in each tube, for a total of $3^6 \times 6 = 4374$. For this example, serial analysis can only explore approximately one hundred millionth of the complexity of the phenotype space.

Per-cell nearest-neighbour (NN) merging of tubes attempts to address this. This method is founded on making the assumption that a cell in one tube is identical to its NN in another tube in terms of the common population markers (Pedreira *et al.*, 2008b). The expression vectors of all the NNs across tubes are merged, creating a single, high-colour matrix of cellular expression across all tubes. NN merging has proven effective as part of classification pipelines (da Costa *et al.*, 2010; Pedreira *et al.*, 2008a; van Dongen *et al.*, 2012). However, as others have shown (Lee *et al.*, 2011), and we show later in this article, populations defined in terms of population markers are frequently made up of a mixture of cell types, and NN consequently tends to produce spurious combinations of markers. This makes the merged output from NN poorly suited for applying the deep profiling techniques, such as flowType and SPADE, as it tends to skew the counts of cell types. One proposed solution to this is to constrain the NNs mapping with clustering incorporating domain knowledge (Lee *et al.*, 2011). However, this latter method requires that all cell types expected to be present be prespecified in order to parameterize the clustering step. So, although well suited to diagnostic pipelines where the goal is to quantify known cell types, it is poorly suited for discovery of new cell types. There remains a need for a multitube combination method for FCM data that produces conservative, non-imputed data suitable for deep profiling. In this article, we describe flowBin, an R/Bioconductor package that we developed to fulfil this need.

2 Approach

FlowBin is designed to accept multiple FCM assays from the same multitube assay and combine these into a complete matrix of measurements for all the markers. To this end, flowBin consists of four stages: (i) normalization, (ii) binning, (iii) bin matching across tubes and (iv) expression measurement (Supplementary Fig. S1).

2.1 Population marker normalization

A consideration in combining multitube FCM is that variations between staining patterns across the aliquots need to be minimized (Pedreira *et al.*, 2008b). In opposition to this are a host of sources of technical variation, ranging from slight differences in sample handling and preparation to instrument drift between runs. Although great pains are taken by operators to reduce these, small variations may still exist. To counteract this, we included a feature in flowBin to quantile normalize overlapping markers across tubes.

Beacause tubes contain physical samples drawn from a common population, their true distributions in terms of overlapping markers are expected to be identical, and any deviations to represent technical variation. Quantile normalization transforms two samples so that they have identical distributions and has been used extensively in gene expression analysis (Bolstad *et al.*, 2003). FlowBin uses quantile normalization to bring similar cells into good registration, using the quantile normalization implementation from the Limma Bioconductor package (Smyth, 2005). The Limma implementation is capable of normalizing in the presence of missing values, and hence can normalize data where the number of cells per tube varies.

2.2 Binning of population markers

In order to bin cells in terms of the overlapping markers present in all tubes, flowBin provides two methods: *K*-means clustering and probability binning. *K*-means clustering with a high value for *K* and NN joining has been used successfully in the past for identifying cell populations in FCM data (Aghaeepour *et al.*, 2011). Probability binning is a binary space partitioning method for FCM data in which each partitioning step maintains equal probability density within both partitions created (Roederer *et al.*, 2001). Probability binning has been developed for use as a 'micro-gating' algorithm in the form of the flowFP Bioconductor package (Rogers *et al.*, 2008). Either method is typically used to partition the cells in a sample into bins containing enough cells to be able to extract average expression values; in our examples, we chose this to be around 200 cells per bin.

2.3 Bin matching across tubes

To enable expression to be combined accurately across tubes, the bins must be spatially as close to identical as possible in each tube. This is easier for flowFP, because the partitions have linear edges, which can easily be mapped directly to individual tubes. *K*-means, in contrast, produces approximately spherical clusters, with more difficult to describe boundaries. Rather than attempt to extract the boundaries of *K*-means clusters, flowBin draws on the idea of NN mapping, except that bin membership is mapped, rather than cellular identity. FlowBin's *K*-means clusters the first tube in the set, and then for each subsequent tube, labels each cell according to the label of the closest cell in the first tube. Once all the cells in each tube have been assigned to cross-tube bins, flowBin moves on to calculating the expression of each bin in terms of the tube-specific expression markers.

2.4 Expression measurement

FlowBin provides three methods for determining expression in each bin, modelled on common practice by FCM analysts. Two make use of a negative control tube (i.e. one that has been stained for the overlapping markers but has no antibodies or non-reactant antibodies in the expression channels). First, normalized median fluorescent intensity (MFI) can be computed (by subtracting the untransformed MFI in terms of negative control from that of the expression marker). Second, a threshold may be set at the 98th percentile of the negative control, and the proportion of cells exceeding that threshold in the expression marker channel reported, as is common practice in many studies (Colburn *et al.*, 2009; Garrido *et al.*, 2001; Hensor *et al.*, 2014). Finally, if negative controls are not available, simple median fluorescent intensity can be computed as the expression measure.

2.5 Downstream analysis

The final output of flowBin is a high-dimensional matrix of expression values for each bin (Supplementary Fig. S1). This can provide a useful overview of the makeup of a sample, for example by plotting a heatmap of bin expression values. However, far greater utility comes in downstream analysis. FlowBin output can be treated as though it were FCM data with a low number of events but a high number of markers. Then, methods for deep profiling, such as flowType and RchyOptimyx, can be applied.

3 Validation

3.1 Validation of quantile normalization

To validate flowBin's quantile normalization, we used an AML dataset (Flow Repository:FR-FCM-ZZYA) used in FCM: critical assessment of population identification methods (FlowCAP; Aghaeepour *et al.*, 2013). FlowCAP is a set of challenges in which automated FCM analysis methods are compared. FlowCAP-II compared classification pipelines, and one dataset included multitube data for AML. This dataset contains flow cytometry standard (FCS) files for 359 patients (normal = 316, AML = 43) with eight tubes each, with six markers assayed in each tube. Every tube had an assay for CD45 as well as forward-scattered light (FSC) and SSC. FlowBin's normalization was evaluated by applying it to each of these markers.

An example result is shown in Supplementary Figure S3. In a single dimension, all tubes are made to have identical cumulative distribution functions. This is as expected for quantile normalization. To achieve a quantitative, *n*-dimensional assessment of registration, we used cytometric fingerprinting (Rogers *et al.*, 2008). In the example case shown, there was substantial deviation across tubes before normalization, which was almost completely removed after.

To measure this objectively over the whole dataset, we applied flowFP to measure the standard deviation (SD) before and after normalization for each tube of all 359 samples. Of a total 2513 tubes, 2207 (88%) showed improvement, 39 (1.6%) showed no change and 267 (11%) showed a wider (worse) SD following normalization.

3.2 Comparison of binning methods

To aid users in choosing between the two binning methods flowBin provides, we compared them on a representative sample from the same AML dataset as for the quantile normalization (FlowRepository: FR-FCM-ZZYA). This is shown in Supplementary Figure S4.

K-means produces spheroid bins that are more likely to follow the contours of the underlying data. FlowFP, in contrast, produces bins with strictly horizontal–vertical borders, which are grid-like and less likely to follow the underlying contours of the data. In terms of variation in number of cells per bin, flowFP produces bins with little variation (SD of means = 0.07). In contrast, *K*-means produces bins with a much wider variation (SD of means = 255), with the largest bins containing up to 20-fold as many cells as the smallest.

3.3 Comparison to NNs and Choice of *k*

To compare flowBin to the per-cell NN merging of Pedreira *et al.*, we created a small, synthetic example using real data containing peripheral blood mononuclear cells stained for CD3, CD4 and CD8. For the source data, we used data from US Military HIV Natural History Study (FlowRepository: FR-FCM-ZZZK). We first removed doublets, debris, dead cells and monocytes, as per Aghaeepour *et al.* (2012a), leaving only CD14^{-} live cells. We then created two artificial tubes by randomly sampling two sets of 5000 cells from the original sample. Both tubes contained CD3 as the overlapping marker through which they were recombined, while one tube contained CD4, and the other CD8.

We repeated this resampling of the cells 100 times each for flowBin (using *K*-means clustering and median fluorescent intensity without negative controls), with $k \in \{2^3 \ldots 2^{11}\}$ and for NNs. To evaluate performance, we set quadrant gates at the thresholds of CD4^{+} and CD8^{+} based on the raw data. For each sampling, we computed the root mean square deviation (RMSD) of the proportion of cells (or flowBin bins) falling within each quadrant compared with the raw data.

The results are shown in Figure 1 and Supplementary Figure S5. The RMSD for flowBin formed a curve, decreasing from low values of *k*, to a trough at $k = 128$ and $k = 256$, then increasing as *k* approached the number of cells. For higher values of *k*, and for NNs, a spurious CD4+CD8+ population was produced (Fig. 1c and d), which is absent from the original data, and occurs only rarely in nature (Parel and Chizzolini, 2004). FlowBin, with $k = 128$, produced a spectrum of values in a hyperbolic curve between the two 'true' populations.

3.4 Validation on simulated multitube data from polychromatic FCM

To assess the abilities and limitations of flowBin, we again took data from US Military HIV Natural History Study. We preprocessed the data as per Aghaeepour *et al.* (2012a), screening out debris, doublets and non-viable cells, then finally gating for CD3^{+} cells (T cells). Patients with fewer than 3000 events remaining were removed, leaving 426 patients, with 12 fluorescent and two scatter channels.

To create simulated tubes, we chose CD3, CD4 and CD8 to use as common markers, then divided the remaining nine among three

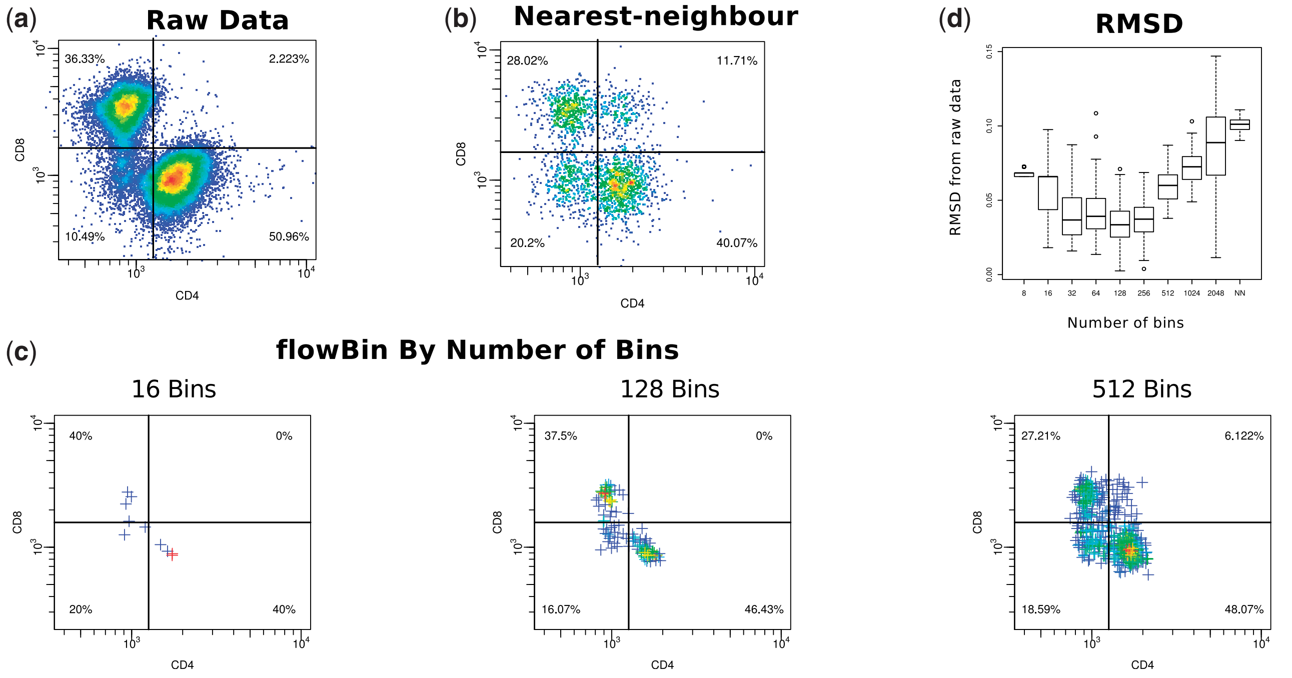

Fig. 1. Comparison between NNs merging and flowBin for two tubes computationally sampled from a real dataset. (a) Raw data (compensated, transformed and filtered for debris), gated for CD3^{+} cells, and showing the true CD4 and CD8 distribution. (b) Example of merging by NNs. (c) Examples of merging by flowBin, with varying bin size. (d) Results of merging 100 times each for NNs and flowBin. The best result (lowest RMSD) was for 128 bins, whereafter increasing bin number caused RMSD to tend towards that of NN

tubes. We divided the events for each patient randomly into three, and discarded all the markers for each that were not to be included in that tube. A summary of all the markers present in each tube is shown in Supplementary Table S1.

We then ran flowBin on each patient's three tubes, using FSC, SSC, CD3, CD4 and CD8 as binning markers, with 128 bins and flowFP as the binning method. We ran flowType on the flowBin output (excluding CD3), and carried out survival analysis (Cox-PH and the log-rank test) on the flowType data as per Aghaeepour *et al.* (2012a). We also ran flowType and the subsequent survival analysis on the original, full-colour FCM data, again as per Aghaeepour *et al.* (2012a).

We compared the cell counts of individual cell types between the true counts from the flowType run on the original high-colour data, and the flowType run on the flowBin data, in terms of their Pearson correlation. We also compared the *P*-values of the log-rank test for each cell type.

Running flowType with 11 markers and two partitions per marker gives a total of $3^{11} = 177, 147$ possible cell types. We excluded those with 0% abundance across all patients, leaving 119 479. Examining the three characteristic survival-associated cell types found in Aghaeepour *et al.* (2012a), more abundant cell types (especially KI-67^{+}CD127^{-}) appear to have better correlation, while rarer cell types (especially CD45RO^{+}CD8^{+}CCR5^{-} CD27^{+}CCR7^{-} CD127^{-}) have much poorer correlations (Fig. 2a). Importantly, the flowBin results for KI-67^{+}CD127^{-} show a strong correlation with the true data, despite KI-67 and CD127 being in separate tubes. Based on Pearson's *r* for all cell types, this pattern holds for those with high abundance (Fig. 2b). Although some low-abundance cell types show strong correlations, it is likely that this was by chance, due to their having very low values in all patients. Because the flowBin results for the majority of cell types with a median abundance of 10% or more had a strong Pearson correlation with the true data, we chose to only do further analysis on those, leaving 1896 cell types.

Comparing *P*-values, these showed a relatively good correlation ($R^2 = 0.65$), with the *P*-values resulting from flowBin being slightly higher than the true *P*-values (Fig. 2c). Following Bonferroni correction, the cell types that were called as significant were matched between the true high-colour analysis and flowBin. FlowBin called only 58 of the 592 (9.8%) the cell types that would be significant in the true data. However, of the 66 flowBin called significant, 58 (88%) were correctly called.

4 Applications

4.1 Separation of AML and normal cells

It is frequently desirable to isolate dysplastic cells from healthy tissue in order to characterize the dysplasia; we demonstrate here how flowBin can be used to achieve this using the same multitube AML dataset (Flow Repository:FR-FCM-ZZYA). Theoretically, the samples from the AML patients should contain a mixture of normal and leukemic blast cells, while the healthy patient samples should only contain normal cells. The problem of separating abnormal from normal cells is thus one of novelty detection, for which techniques, such as single-class support vector machines (SVMs), are available (Chen *et al.*, 2001). We applied flowBin to both the normal and AML patients using *K*-means binning. We pooled all bins from the normal samples, and trained a single-class SVM on these using the kernlab package (Karatzoglou *et al.*, 2004). We then applied the trained SVM to the bins from the AML patients to predict which were normal and which dysplased.

The bins predicted to be normal fall mainly into well-clustered populations showing expression patterns typical of healthy myeloid,

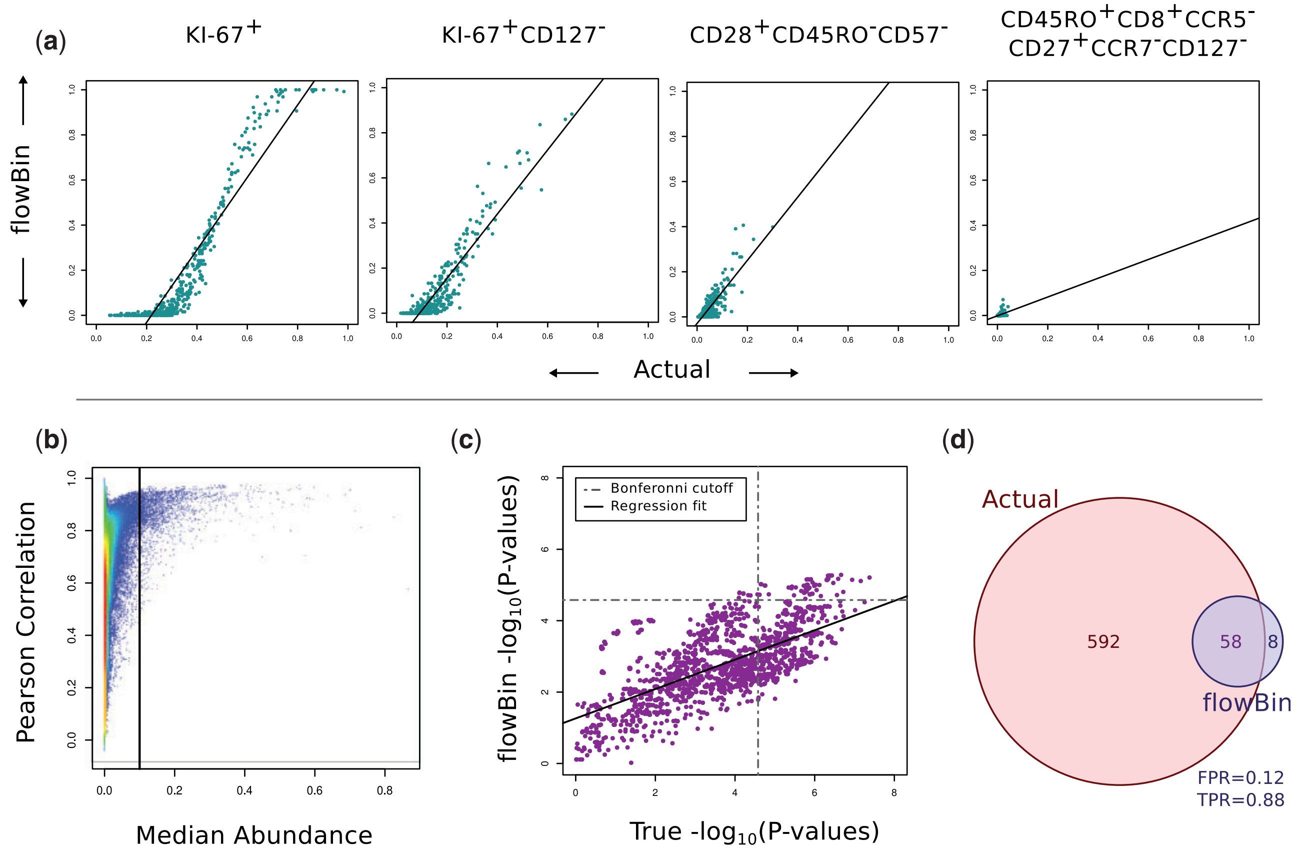

Fig. 2. Performance on flowBin in reproducing a high-colour FCM analysis on simulated FCM **data.** (a) Comparison of counts of selected phenotypes between actual data and simulated multitube data recombined by flowBin, with linear regression fit lines. Ki-67$^+$ was selected as being representative of a phenotype with the full range of abundance across patients. The remaining three phenotypes are the representative phenotypes of the three classes found in the original study (Aghaeepour *et al.*, 2012a). More abundant phenotypes show a good, though imperfect fit. Less abundant phenotypes show a poorer fit. (b) Pearson correlations for all phenotypes between actual values and flowBin-recombined values, versus median abundance of the phenotype. Below an abundance of ~0.1 (10% of all cells in the sample), the correlation becomes decoupled from abundance. (c) Comparison of *P*-values between actual and flowBin-recombined data, for only those phenotypes with >10% abundance. The *P*-values show a good correlation ($R^2 = 0.65$). (d) Cell types called as being significant for d. Cell types called as being significant on flowBin-recombined simulated multitube data (blue) and the raw data (purple). Just fewer than 10% of the phenotypes that were called as significant in the actual data were also called as such in the flowBin-processed data (high Type II error). Eight phenotypes were inappropriately called as significant (very low Type I error)

lymphoid and erythroid cells, respectively (Supplementary Fig. S6). In contrast, the bins predicted to be abnormal show much greater variance, and expression patterns typical of AML such as CD34 and CD117 expression, and co-expression of the lymphoid markers CD4 and CD7 with myeloid markers.

4.2 FlowCAP-II

We entered a pipeline using flowBin for feature extraction along with a voting classifier method. Binning within patients raised the problem of linking features across patients for classification. To solve this, we took each bin from each sample as a separate training instance, labelled with the sample label, and then trained a SVM classifier. For class prediction, we took the majority vote of the predicted labels for a given sample's bins. Classification with parameter optimization and 3-fold cross-validation was implemented using the ksvm R package, but could in theory be made to work with any modern classification method (Supplementary Fig. S7).

In the original FlowCAP-II competition, this pipeline performed poorly in comparison with other algorithms, with an F-measure of 0.46 (Aghaeepour *et al.*, 2013). In dissecting this performance, it emerged that the classifier was vulnerable to class imbalance (43 of 359 patients had AML). To address this issue, we added bagging with down sampling to the classifier (Breiman, 1996). This improved classifier, produced an F-measure of 0.96 when evaluated using the same methodology as was used in FlowCAP (Supplementary Fig. S8).

4.3 FlowBin with flowType and RchyOptimyx to find cell types in AML correlated with *NPM1* mutation

To demonstrate the utility of flowBin, we applied it to a novel dataset of 129 *de novo* AML cases (FlowRepository: FR-FCM-ZZES). Each of these cases had multitube FCM data available from the time of diagnosis. In addition, each had been genotyped for clinically relevant frameshift mutations in the 12th exon of the *NPM1* gene. These mutations (hereafter referred to as *NPM1-mt*) indicate a good prognosis and have a marked correlation with the absence of CD34 on the AML blast cells (Grisendi and Pandolfi, 2005; Schnittger *et al.*, 2005; Thiede *et al.*, 2006; Verhaak *et al.*, 2005) and, in certain cases, HLA-DR (Syampurnawati *et al.*, 2008).

Although high-level, single marker studies have been performed to find other recurrent immunophenotypic characteristics of *NPM1-mt* AML (Dalal *et al.*, 2012; Syampurnawati *et al.*, 2008), deep

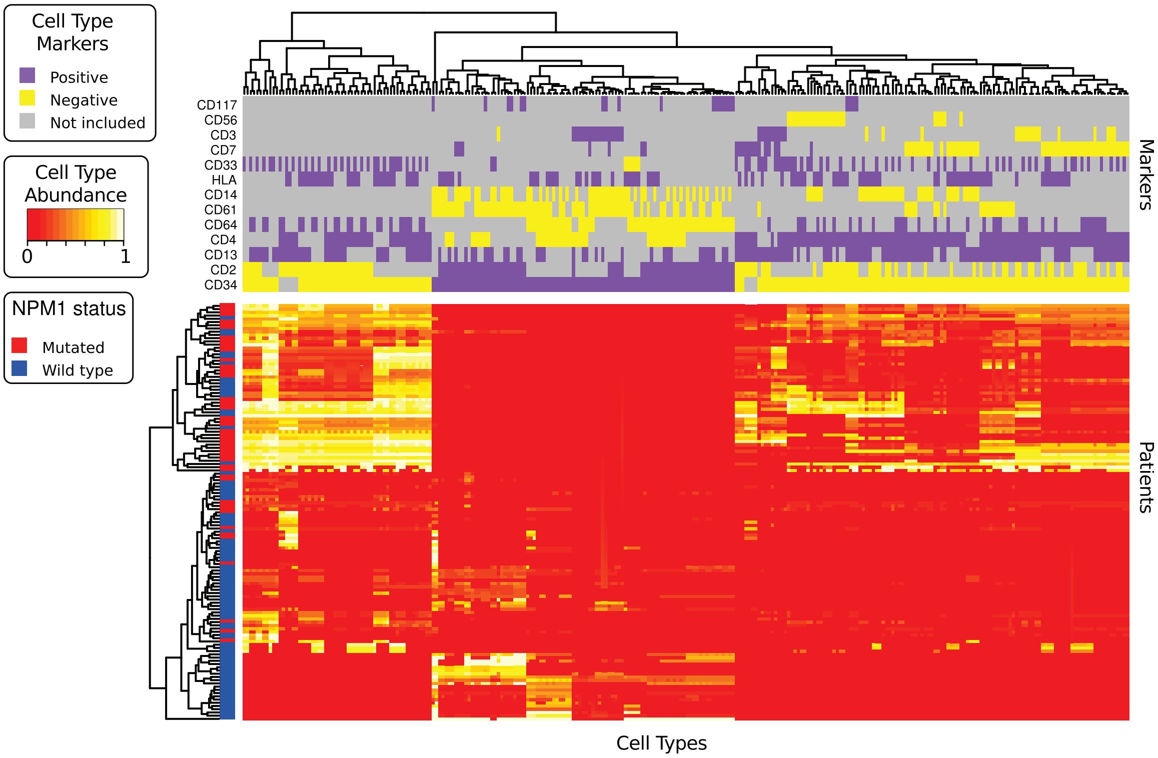

Fig. 3. Overview of cell types which showed significant differences in abundance between *NPM1-mt* and *NPM1-wt*. The cell types cluster into four main groups, two of which are characteristically CD34+ and associated with the *NPM1-wt* patient group, and two of which are characteristically CD34− and associated with the *NPM1-mt* patient group

profiling of the condition's immunophenotypic landscape has yet to be undertaken. We performed such an analysis using flowBin and flowType, with the hypothesis that within the immunophenotypic landscape of AML, there may be additional cell types that are more strongly correlated with *NPM1* mutation than CD34+/− alone.

To this end, we first used flowBin to combine tubes for each sample and measure the expression of the mapped bins. We then used flowType to delineate and count all cell types present, defined over all combinations of up to six markers. We filtered out all cell types not present in any patient, leaving 616 285. We then tested for differences in abundance of each of these cell types between *NPM1-mt* and wild-type patients using the Mann–Whitney *U* test with Bonferroni–Holm correction for multiple testing (Supplementary Fig. S2).

We then performed exploratory analysis using RchyOptimyx to visualize the hierarchies within the 801 significant cell types. We found that adding CD19−, CD20− and CD10− had little to no effect on the *P*-value of a cell type (Supplementary Fig. S9). This is likely due to these (B-lymphoid) markers being extremely rare in AML (Swerdlow *et al.*, 2008), and hence not being expressed on the important cell types at all, so that a negative gate for any of them did not change which cells had that cell type. We consequently excluded all cell types involving CD19−, CD20− or CD10−, bringing the total cell types to 272.

These 272 cell *NPM1* mutation-associated types are illustrated in Figure 3. The cell types cluster into four main groups, two of which are characteristically CD34+ and associated with the *NPM1-wt* patient group, and two of which are characteristically CD34− and associated with the *NPM1-mt* patient group.

Relative abundance of four groups of cell types with stronger *P*-values, chosen by further exploration using RchyOptimyx, are shown in Supplementary Figure S10. Gating for the presence of myeloid lineage markers CD13 and CD33 within the CD34− compartment yields much stronger differences in abundance between *NPM1-wt* and *NPM1-mt* than CD34− alone. Gating for CD2− within the CD34− compartment yields a slightly better separation than CD34− alone, but gating down further to CD4− and CD13+ is a cell type that, while present in most *NPM1-mt*, is absent or below 20% abundance in nearly all *NPM1-wt*. Gating for CD61− and CD14− within the CD34+ compartment leads to a cell type which is common in *NPM1-wt* but almost entirely absent in *NPM1-mt*. Gating for HLA-DR+ and CD64− within the CD34+ compartment leads to a cell type that occurs in a subset of *NPM1-wt* but is entirely absent in *NPM1-mt*.

5 Discussion

5.1 Validation of quantile normalization

The results of the comparison between non-normalized and normalized data suggest that quantile normalization can improve the registration of cells in terms of their population markers. This suggests that it should be considered in general before applying tube combination methods, including flowBin.

5.2 Comparison of binning methods

For binning, flowFP gives much less numerically dispersed results than *K*-means, which is essential for accurate flowType counts (Supplementary Fig. S4). FlowFP binning may thus be a better choice for downstream applications that depend on accurate cell counts. For example, if flowFP is used for binning, the assumption can be made that each data point in the flowBin results has the same number of cells contained within it. If flowType is then applied to that data, the counts of cell types which flowType produces can be considered to be relatively accurate representations of the true counts.

K-means, in contrast, gives better fitted bins, but with greater variation. *K*-means binning thus may be a more attractive choice if later back gating of interesting populations is desired. *K*-means may also be the only feasible choice in cases where there are many population markers, as flowFP's binary space partitioning runs into combinatorial difficulties.

5.3 Comparison to NNs and choice of *k*

The comparison (Fig. 1) showed that flowBin with 128 or 256 bins produced a distribution across quadrants and overall that more closely approximated the true distribution, while NNs and flowBin with higher bin numbers produced an artificial $CD4^{+}CD8^{+}$ population. For all cases, including NNs, the deviation was relatively small, never exceeding 0.15. This is likely due to the $CD4^{+}$ and $CD8^{+}$ populations being slightly separated in terms of their CD3 expression (Supplementary Fig. S5b), allowing them to be separated in a way better than random by both methods. The appearance of the spurious population as bin number increases is likely due to cells with similar values of CD3 expression but opposite CD4 and CD8 expression being mismatched. This is likely due to flowBin, especially at lower bin numbers, having a smoothing effect, causing expression levels to be weighted towards the centres of each population, and thus better separated in terms of CD3.

These results suggest that flowBin, with 128 or 256 bins, is a better choice for situations where it is desirable to recover the underlying cell types accurately, such as cell type discovery. NNs merging, due to its tendency to produce spurious combinations of marker expression, is poorly suited to techniques which require precise counts of particular cell types, such as flowType, or manual analysis. However, NNs has proven extremely effective when used as part of a classification pipeline (da Costa *et al.*, 2010; van Dongen *et al.*, 2012), whereas flowBin loses some information as a result of averaging. As such, the two methods can be complementary to each other.

5.4 Validation on simulated multitube data from polychromatic FCM

When compared with simulated multitube data in a complete analysis pipeline with flowType, flowBin reproduced the true underlying trends in the data, but with lowered statistical power and sensitivity. The individual correlations (Fig. 2a and b) show that cell type counts based on flowBin recombination of tubes reproduce the true counts for most cell types of >10% abundance. This suggests that while flowBin may not be suitable for analysis efforts examining rare cell populations (such as minimal residual disease in leukemia or T cell subsets), it is a useful tool for examining more abundant cell types (e.g. the gross heterogeneity among tumour types).

For reproducing the results over the entire pipeline, flowBin performed well once cell types of <10% average abundance were filtered out. *P*-values of the survival analysis were close to those of the pipeline run on the true data, but slightly raised. This resulted in increased Type II error, but minimal Type I error. This suggests that while flowBin introduces some noise, the final effect is only to lower the statistical sensitivity of analyses performed on flowBin expression data, but to produce few false positives.

Interestingly, KI-67^{+} showed a slightly sigmoidal distribution in Figure 2a. The likely reason for this is because flowBin produces averages, while flowType uses a threshold (Fig. 1c). Thus, bins with fewer positive cells will tend to have an average (flowBin) expression which falls below the flowType threshold, it may be an avenue for future investigation to attempt to mitigate this, for example by applying a logit transformation to flowBin expression data before passing it to flowType. However, fitting logit transforms would be challenging for most cell types, as their distributions would be a mixture of the sigmoidal distributions of their component markers. Indeed, for KI-$67^{+}CD127^{-}$ in Figure 2a, the sigmoidal curve is partially cancelled out, most likely by such mixing.

Also importantly, in this simulation, all three of the cell types of interest found in the original study (Aghaeepour *et al.*, 2012a), Figure 2a, were defined by markers spanning multiple tubes. Thus, none of those cell types could be found using flowType on the uncombined single tubes. For multitube FCM data, there is a definite benefit in combining tubes using flowBin and then running flowType, rather than just running flowType alone. However, it may still be useful to apply flowType to the single tubes for a lower dimensional but higher sensitivity search in parallel with the flowBin–flowType analysis.

5.5 Separation of AML and normal cells

We have shown how flowBin in combination with one-class SVM can be useful for separating normal from aberrant cells where a good training set of normal cells is available. This can have applications for later AML studies where the ratio of normal cells to leukemic is a confounding factor.

5.6 Identifying AML patients (FlowCAP 2)

We have shown that, despite poor performance in the initial FlowCAP-II competition, a voting classifier used in conjunction with flowBin can separate AML from normal cells, when balanced bagging is added. On the AML dataset from FlowCAP, balanced bagging improved performance substantially, with the F-measure increasing from 0.46 to 0.96, a number roughly in the middle of the field compared with the other pipelines entered. This could be improved further by incorporating other machine learning best practices, such as feature selection. However, there are two levels of recursion already: the voting classifier and the bagging. Adding another layer of recursion would likely increase the computational complexity prohibitively. Instead, it would be better to recommend that flowBin be used in conjunction with the best-performing techniques from FlowCAP, most notably flowType and SPADE (Aghaeepour *et al.*, 2013).

5.7 FlowBin with flowType and RchyOptimyx to find cell types in AML correlated with *NPM1* mutation

The overall pattern of association between CD34 expression and *NPM1* mutation shown in Figure 3 fits with previous reports (Dohner *et al.*, 2005; Thiede *et al.*, 2006; Verhaak *et al.*, 2005). However, flowBin with flowType was able to find a multitude of immunophenotypes within those classes ($CD34^{+}$ and $CD34^{-}$).

Referring to the groups examined in more detail in Supplementary Figure S10, the first group, CD34−CD13+CD33+, fits with observations that blasts often express CD13 and CD33 in *NPM1-mt* AML (Swerdlow *et al.*, 2008). The second and third groups, involving CD34−CD2− and CD34+CD2+, are both cell types which have been reported before in acute promyelocytic leukemia (Albano *et al.*, 2006), but not associated with *NPM1*. For the cell types in the second and third groups, CD4 has been recently reported to be associated with *NPM1-mt*, along with *t*(9;11) and monocytoid AML (van Dongen *et al.*, 2012). In the third group, CD61, a marker of megakaryoblasts, may fit with the observation of dysmegakaryopoeisis in *NPM1-mt* AML (Falini *et al.*, 2010).

6 Conclusion

FlowBin is a complete pipeline for combining multitube FCM data via markers shared across tubes. Quantile normalization of those markers to reduce technical variation is included. Of the two forms of binning included, flowFP is most suitable for later flowType analysis, while *K*-means fits the contours of the data better. Compared with NNs merging of tubes, flowBin produces cleaner data, with far fewer false double-positive marker combinations. FlowBin with flowType can reproduce true data for more abundant cell types, and this data are suitable for statistical testing, albeit with lowered statistical power (increased Type II error). Also, importantly, this is only true for cell types of >10% abundance, suggesting that flowBin, and potentially recombining of multitube FCM in general, is not suitable for analyses involving rarer cell types.

We have found a series of cell types associated with *NPM1* mutation in AML. Although many of these cell types fit with previously reported trends, most represent new, previously unreported cell types associated with *NPM1* mutation. Importantly, we have also demonstrated that flowBin output can be used for downstream discovery analysis using tools such as flowType and RchyOptimyx.

Acknowledgements

This work was supported by Natural Sciences and Engineering Research Council, National Institutes of Health/R01EB008400, the Canadian Cancer Society, the Michael Smith Foundation for Health Research, the International Society for the Advancement of Cytometry scholars program (NA), the Canadian Institutes of Health Research/Michael Smith Foundation for Health Research scholarship for strategic training in bioinformatics (NA) and UBCs 4YF scholarship (NA).

References

Aghaeepour,N. *et al.* (2011) Rapid Cell Population Identification in Flow Cytometry Data. *Cytometry. Part A: the journal of the International Society for Analytical Cytology,* **79**, 6–13.

Aghaeepour,N. *et al.* (2012a) Early immunologic correlates of HIV protection can be identified from computational analysis of complex multivariate T-cell flow cytometry assays. *Bioinformatics,* **28**, 1009–1016.

Aghaeepour,N. *et al.* (2012b) RchyOptimyx: cellular hierarchy optimization for flow cytometry. *Cytometry A,* **81**, 1022–1030.

Aghaeepour,N. *et al.* (2013) Critical assessment of automated flow cytometry data analysis techniques. *Nat. Methods,* **10**, 228–238.

Albano,F. *et al.* (2006) The biological characteristics of CD34+ CD2+ adult acute promyelocytic leukemia and the CD34 CD2 hypergranular (M3) and microgranular (M3v) phenotypes. *Haematologica,* **91**, 311–316.

Bendall,S.C. and Nolan,G.P. (2012) From single cells to deep phenotypes in cancer. *Nat. Biotechnol.,* **30**, 639–647.

Bendall,S.C. *et al.* (2012) A deep profiler's guide to cytometry. *Trends Immunol.,* **33**, 323–332.

Bolstad,B.M. *et al.* (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics,* **19**, 185.

Breiman,L. (1996) Bagging predictors. *Mach. Learn.,* **24**, 123–140.

Chattopadhyay,P.K. *et al.* (2008) A chromatic explosion: the development and future of multiparameter flow cytometry. *Immunology,* **125**, 441.

Chen,Y. *et al.* (2001) One-class SVM for learning in image retrieval. In *International Conference on Image Processing,* Institute of Electrical and Electronics Engineers (IEEE), **Vol. 1**, pp. 34–37.

Colburn,N.T. *et al.* (2009) A role for γ/δ T cells in a mouse model of fracture healing. *Arthritis Rheum.,* **60**, 1694–1703.

Craig,F.E. and Foon,K.A. (2008) Flow cytometric immunophenotyping for hematologic neoplasms. *Blood,* **111**, 3941.

da Costa,E.S. *et al.* (2010) Automated pattern-guided principal component analysis vs expert-based immunophenotypic classification of B-cell chronic lymphoproliferative disorders: a step forward in the standardization of clinical immunophenotyping. *Leukemia,* **24**, 1927–1933.

Dalal,B. *et al.* (2012) Detection of CD34, TdT, CD56, CD2, CD4, and CD14 by flow cytometry is associated with NPM1 and FLT3 mutation status in cytogenetically normal acute myeloid leukemia. *Clin. Lymphoma Myeloma Leuk,* **12**, 274–279.

De Rosa,S.C. *et al.* (2001) 11-color, 13-parameter flow cytometry: identification of human naive T cells by phenotype, function, and T-cell receptor diversity. *Nat. Med.,* **7**, 245–248.

Dohner,K. *et al.* (2005) Mutant nucleophosmin (NPM1) predicts favorable prognosis in younger adults with acute myeloid leukemia and normal cytogenetics: interaction with other gene mutations. *Blood,* **106**, 3740.

Falini,B. *et al.* (2010) Multilineage dysplasia has no impact on biologic, clinicopathologic, and prognostic features of AML with mutated nucleophosmin (NPM1). *Blood,* **115**, 3776–3786.

Finn,W.G. (2009) Beyond gating: capturing the power of flow cytometry. *Am. J. Clin. Pathol.,* **131**, 313.

Garrido,S.M. *et al.* (2001) Acute myeloid leukemia cells are protected from spontaneous and drug-induced apoptosis by direct contact with a human bone marrow stromal cell line (HS-5). *Exp. Hematol.,* **29**, 448–457.

Grisendi,S. and Pandolfi,P.P. (2005) NPM mutations in acute myelogenous leukemia. *N. Engl. J. Med.,* **352**, 291.

Hensor,E.M.A. *et al.* (2014) A1.33 predicting the evolution of inflammatory arthritis in ACPA-positive individuals: can T-cell subsets help? *Ann. Rheum. Dis.,* **73**(Suppl 1), A14–A14.

Karatzoglou,A. *et al.* (2004) Kernlab—an S4 package for kernel methods in R. *J. Stat. Softw.,* **11**, 1–20.

Lacombe,F. *et al.* (1997) Flow cytometry CD45 gating for immunophenotyping of acute myeloid leukemia. *Leukemia,* **11**, 1878–1886.

Lee,G. *et al.* (2011) Statistical file matching of flow cytometry data. *J. Biomed. Inform,* **44**, 663–676.

Maecker,H.T. and Trotter,J. (2006) Flow cytometry controls, instrument setup, and the determination of positivity. *Cytometry A,* **69A**, 1037–1042.

O'Neill,K. *et al.* (2013) Flow cytometry bioinformatics. *PLoS Comput. Biol.,* **9**, e1003365.

O'Neill,K. *et al.* (2014) Enhanced flowType/RchyOptimyx: a bioconductor pipeline for discovery in high-dimensional cytometry data. *Bioinformatics,* **30**, 1329–1330.

Parel,Y. and Chizzolini,C. (2004) CD4+ CD8+ double positive (DP) T cells in health and disease. *Autoimmun. Rev.,* **3**, 215–220.

Pedreira,C.E. *et al.* (2008a) A multidimensional classification approach for the automated analysis of flow cytometry data. *IEEE Trans. Biomed. Eng.,* **55**, 1155–1162.

Pedreira,C.E. *et al.* (2008b) Generation of flow cytometry data files with a potentially infinite number of dimensions. *Cytometry A,* **73**, 834–846.

Qiu,P. *et al.* (2011) Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE. *Nat. Biotechnol,* **29**, 886–891.

Robinson,J.P. *et al.* (2012) Computational analysis of high-throughput flow cytometry data. *Expert Opin. Drug Discov.,* **7**, 679–693.

Roederer,M. *et al.* (2001) Probability binning comparison: a metric for quantitating multivariate distribution differences. *Cytometry A,* **45**, 47–55.

Rogers,W.T. *et al.* (2008) Cytometric fingerprinting: quantitative characterization of multivariate distributions. *Cytometry A,* **73**, 430–441.

Schnittger,S. *et al.* (2005) Nucleophosmin gene mutations are predictors of favorable prognosis in acute myelogenous leukemia with a normal karyotype. *Blood,* **106**, 3733.

Smyth,G.K. (2005) Limma: linear models for microarray data. In: R., Gentleman *et al.* (eds.) *Bioinformatics and Computational Biology Solutions Using R and Bioconductor. Statistics for Biology and Health.* Springer, New York, pp. 397–420.

Swerdlow,S.H. *et al.*, eds. (2008) *WHO Classification of Tumours of Haematopoietic and Lymphoid Tissues*. International Agency for Research on Cancer, France.

Syampurnawati,M. *et al.* (2008) DR negativity is a distinctive feature of M1/M2 AML cases with NPM1 mutation. *Leuk. Res.*, **32**, 1141–1143.

Thiede,C. *et al.* (2006) Prevalence and prognostic impact of NPM1 mutations in 1485 adult patients with acute myeloid leukemia (AML). *Blood*, **107**, 4011.

van Dongen,J.J.M. *et al.* (2012) EuroFlow antibody panels for standardized n-dimensional flow cytometric immunophenotyping of normal, reactive and malignant leukocytes. *Leukemia*, **26**, 1908–1975.

Verhaak,R.G.W. *et al.* (2005) Mutations in nucleophosmin (NPM1) in acute myeloid leukemia (AML): association with other gene abnormalities and previously established gene expression signatures and their favorable prognostic significance. *Blood*, **106**, 3747.

Wood,B.L. *et al.* (2007) 2006 Bethesda international consensus recommendations on the immunophenotypic analysis of hematolymphoid neoplasia by flow cytometry: optimal reagents and reporting for the flow cytometric diagnosis of hematopoietic neoplasia. *Cytometry B Clin. Cytom.*, **72**, S14–S22.