

Chapter 2

Breaking the Dimensionality Barrier

C. Bruce Bagwell

Abstract

Recent advances in biotechnology have resulted in cytometers capable of performing numerous correlated measurements of cells, often exceeding ten. In the near future, it is likely that this number will increase by fivefold and perhaps even higher. Traditional analysis strategies based on examining one measurement versus another are not suitable for high-dimensional data analysis because the number of measurement combinations expands geometrically with dimension, forming a kind of complexity barrier. This dimensionality barrier limits cytometry and other technologies from reaching their maximum potential in visualizing and analyzing important information embedded in high-dimensional data.

This chapter describes efforts to break through this barrier and allow the visualization and analysis of any number of measurements with a new paradigm called Probability State Modeling (PSM). This new system creates a virtual progression variable based on probability that relates all measurements. PSM can produce a single graph that conveys more information about a sample than hundreds of traditional histograms. These PSM overlays reveal the rich interplay of phenotypic changes in cells as they differentiate. The end result is a deeper appreciation of the molecular genetic underpinnings of ontological processes in complex populations such as found in bone marrow and peripheral blood.

Eventually these models will help investigators better understand normal and abnormal cellular progressions and will be a valuable general tool for the analysis and visualization of high-dimensional data.

Key words: Flow cytometry, Multidimensional analysis, Dimensionality barrier, Probability state model

1. Introduction

1.1. Measurements, Parameters, Markers, Colors, Particles, and Events

Cytometers perform measurements on cells. Whether a cytometer is an image or flow system, these measurement signals are normally digitized, processed in some manner, and stored. Unfortunately, these particle measurements have historically been referred to as “parameters.” Since the term “parameter” has a very different statistical meaning, especially in the context of modeling, it will only be used in a modeling context for this chapter.

The word “measurement” or “variable” will be used to mean a specific digitized attribute of a particle. Measurements derived from fluorescence photons are commonly referred to as “colors.” For example, six-color data means that the data contains six correlated measurements from different fluorescence colors. Usually other measurements such as time, forward scatter, and side scatter accompany these colors. For both flow and image cytometry, most color detectors can generate numerous types of measurements by means of pulse or statistical processing.

The objects that are measured by a cytometer are referred to as “particles.” These particles are often cells, but they don’t need to be. Once the cytometer’s pulse processing algorithms decide that a modulation in a signal is probably a real particle and not noise, they digitize all the measurements and the particle becomes an electronically identified “event.” All measurements derived from events are generally stored on computer storage media as a “listmode” file.

1.2. Difficulties with Visualization and Analysis of High-Dimensional Data

Cells are largely composed of interdependent chemical machines moving from state to state along thermodynamic gradients. Technologies that perform numerous measurements on single cells have the potential of uncovering the specifics of how these machines perform their various functions. Recently the technology of cytometry has evolved to making 17 correlated measurements from cells (1, 2) and is on the precipice of dramatically increasing this number to over 50 (3–5).

The rate-limiting issue that has impeded this evolution toward high-dimensional displays and analyses is what has commonly been termed the dimensionality barrier or curse (6, 7). This barrier is one of overwhelming complexity in both viewing and analyzing large numbers of correlated measurements. Visualization of multiple-dimensional data is traditionally done by viewing numerous bivariate dot or contour plots. As the number of correlated measurements, n , increases, the number of bivariate plots expands geometrically by the relation, $n \times (n - 1) / 2$. For seventeen-dimensional data, a total of 136 biviates are necessary to view all correlations in data, which severely limits visual interpretation of the embedded information.

Analysis methods that involve measurement partitioning schemes (8, 9), also run into this barrier because the number of partitions can also increase geometrically with measurement number, ultimately increasing the required number of events to unattainable values or requiring unreasonably large partition sizes. Nevertheless, these methods do provide valuable information on quantitative differences between high-dimensional data files.

1.3. Other Approaches to the Dimensionality Barrier Problem

There have been numerous schemes that have been proposed to ameliorate this dimensionality problem (10–16). Many of these methods work quite well, but most will eventually fail when the number of measurements becomes too high. A review of current

approaches to high-dimensional cytometry data visualization has recently been published (17). In general, any approach to high-dimensional data visualization that attempts to examine one measurement or combination of measurements versus another eventually runs into this barrier.

Numerous clustering methods have been used to successfully classify cytometric data into relevant populations (14, 18–23). Since clustering methods generally use distance or proximity functions to find neighboring events, they obviate many of the complexity issues associated with the dimensionality barrier. Unfortunately, most clustering routines have two weaknesses. The first is that they usually are not good at finding small continuums between clusters. These continuums are many times far more important than the clusters themselves. The second problem is that after the clustering routine is finished, visual inspection of clusters is generally done by examining bivariate with color-coded dots representing found clusters. As already discussed, as soon as bivariate involving measurements are used to inspect data, the dimensionality barrier interferes with the complete interpretation of the data's embedded information.

1.4. Why Write this Chapter

This chapter provides a brief historical account of how this dimensionality curse, as it relates to complex processes, was solved by a radically different approach to both visualization and analysis of high-dimensional data. The technique is termed Probability State Modeling and its specifics have been described elsewhere (24–26).

The rest of this chapter will be devoted to an exploration of some PSM applications. The main reason for writing this chapter is to introduce the reader to this new and powerful visualization and analysis tool and show how it leads to a more thorough understanding of complex cellular systems.

2. Probability State Modeling: A Radically Different Approach

2.1. The Ginger Root View of Bone Marrow

Work on this project began in earnest early in 2001. The general idea was to develop new visualization methods that could reveal the underlying complex measurement relations found in tissues like bone marrow. At the time, the commonly held image of populations in bone marrow, proposed by Howard Shapiro in the early 1980s (27), was that they were similar to the variegated bulges of a ginger root. Populations that were close to each other from an ontological point of view would end up being close to each other in the ginger root three-dimensional solid structure. This concept suggested that the proper approach to the problem of high-dimensional display and analysis would involve some kind of nearest-neighbor clustering routine that

2.2. Creating a Sophisticated Model Representing Complex Cytometry Data

could produce this kind of solid structure from the analysis of correlated measurements.

Investigation of new clustering and visualization algorithms is facilitated by software that can simulate realistic data. It's always important to have data sets where important truths about the data are known in order to understand how well computer algorithms approximate these truths. For this project, it was necessary to design a computer simulator that could produce high-dimensional data that was similar if not identical to that produced by a cytometer.

At the core of the simulator is a model. A model is a mathematical construct that normally simulates some real-world process or processes. The model needs to be complex enough to represent all lineages in bone marrow and general enough to handle other types of cytometry-derived populations. It needs to be able to define measurement uncertainty as well as population heterogeneity. It also has to represent complex continuums between developmental stages of multiple lineages in bone marrow.

Building a complex model is much like writing a book. In order to deal with the complexity, a book is normally divided into logical chapters. Within each chapter, there may be a number of subheadings. These subheadings may be further nested. This nested approach takes a complex system and breaks it into simpler and more manageable parts. A model that can represent something like bone marrow has the following nested format.

```
Model {bone marrow}
  B-Cell Lineage {Type}
    CD19 Parameter Profile {Constant}
      Control Definition Points
      Means (r states)
      SD's (r states)
    CD34 Parameter Profile {One Step-Down}
    CD20 Parameter Profile {One Step-Up}
    CD10 Parameter Profile {Three Step-Down's}
    CD45 Parameter Profile {Three Step-Up's}
    ...
  Common Progression Scale {0–100}
  Frequency {r states}
  T-Cell Lineages {Type}
  Monocytic Lineage {Type}
  Myeloid Lineages {Type}
  Erythroid Lineage {Type}
  ...
```

Unfortunately, it is necessary to develop some syntactical terms to describe the various parts of this model design. The model chapters are the major lineages in bone marrow and are called “Types.” For simplicity, only the B-Cell Lineage section is expanded above. Each lineage or Type is further divided into subsection’s or “Parameter Profiles” that describe how particular measurements change as a function of lineage progression. These changes are well-known for normal bone marrow development (28–30).

The CD19 Parameter Profile, for example, is relatively constant in B-cell maturation; whereas, CD34 has an early step-down in intensity. A critical element in the model is the common progression axis or scale that relates all the Parameter Profiles to each other. In this particular model, progression starts at zero and ends at 100%, and is divided into r separate bins or states. The reason for this construction will become clearer when the model is used for analysis instead of simulation.

The shape of a Parameter Profile is controlled by a set of important inflection points called Control Definition Points (CDP). The CDP’s are the real parameters for this model. A CDP contains a position along the progression axis, a relative measurement intensity value, and a standard deviation (SD). The SD defines the measurement uncertainty and heterogeneity at a particular progression value. Each Parameter Profile also had a set of r means and SD’s for each state along the progression.

Finally, the model Type contains a frequency value for each state. We won’t directly need these frequencies for simulation, but they will become important later during analysis.

2.3. Simulating Realistic Data

Now that the model is defined, simulating realistic data is straightforward. The first step is to randomly pick a Type from the model. In order to do this, the approximate percentage of each Type in the model needs to be specified. The Type needs to be randomly chosen in order that the data appear as realistic as possible. An example best describes how this stochastic selection is done. Suppose the percentages for B-cells (2.5%), T-cells (10.0%), Monocytes (25%), Myelocytes (50%), and Erythroids (12.5%) are stored in the following list,

{2.5, 10.0, 25.0, 50.0, 12.5}

A stochastic selection function (31) takes this list and randomly chooses one of the elements based on these weights. It will pick the first element approximately 2.5% of the time, the second element 10.0% of the time, the third element 25.0% of the time, and so on. A good analogy to this type of selection is “The Wheel of Fortune” game show where a wheel is spun by a contestant and then randomly stops at various points along the wheel. If the wheel segments are proportional to the above Type percentages, then a Type will be stochastically selected with each spin of the wheel.

Once a Type is stochastically selected, the system randomly picks one of the states along the selected Type's progression. This type of selection is uniform which means that each state has an equal chance of being chosen. Once the algorithm knows the state, it can examine each of the Parameter Profiles and estimate a measurement value and SD by interpolation between the appropriate CDP's. The last step is to use a normally distributed random number generator defined by the SD and add this uncertainty value to the interpolated measurement value. This final number is stored on disk in a listmode type of structure. This process is continued for the remainder of the Parameter Profiles and then repeated over and over again until the desired number of simulated events is achieved. The end result is the creation of a listmode file that looks very similar to those produced by cytometers.

2.4. Three Insights

There is a direct relationship between the simulator's model and synthesized data. The information contained in the model is reflected in the synthesized data. These seemingly obvious statements are at the heart of this new technology. A plot of all the model Type's Parameter Profile's vs. progression represents all measurement correlations that are in the synthesized data file. This fact means that if there is a way to reverse directions and find the model from the data, then we can represent all the correlations in the data without running into the dimensionality barrier.

Since a uniform random number generator picked the progression states within a Type, the progression axis represents cumulative fraction or probability. This type of axis allows subpopulation percentages to be directly read from the axis scale. This insight was very important since it precisely defined the concept of progression without the use of time. Progression in this context is defined probabilistically.

The third insight was that any combination of Parameter Profiles could be synthesized and the frequency of all the states would always be uniform. The probability of an event representing a state is equal for all states since that is how the generator was designed. In many ways, this last insight was the most important since it meant that it might be possible to simplify an analysis strategy by examining each measurement in succession rather than looking at all measurements concurrently.

2.5. Synthesizer to Classifier

What these three insights really meant was that if it were possible to create a model from observed data, then the problem of representing any number of correlated measurements in one graph was solved. The question at this point was whether it was feasible to use the model to accurately classify observed events into progression states. More specifically, was there a way of classifying all synthesized events such that all the frequencies of the chosen states were as uniform as possible? When a model is in this uniform frequency condition, it can synthesize data that is indistinguishable from the observed

data and therefore can be conceived of as representing the data. This type of model is called a Probability State Model (PSM).

2.6. Choosing a State to Associate with an Event

How is an event classified into a specific state? The key to answering this question is to associate a probability for each state that represents the likelihood that the event belongs to that state. Once this list of probability weights is known for an event, a stochastic selection process picks the state to associate with the event. Once the state is known, the algorithm keeps track of how many events are in each state. If the model represents the data, then these frequencies will be uniform.

2.7. Non-uniform Frequency Gradient and the Minimization Searching Algorithm

At this point, our probability state model will have uniform state frequencies if it classifies data that it has synthesized. How does this relate to finding a new model with unknown data? Consider what happens if one or more of the CDP's are moved from the positions they were in when the data was synthesized. When the data is reclassified with this different model, the uniform nature of the state frequencies will be perturbed. The further the CDP's are moved from their original positions, the greater the perturbation.

If the degree of non-uniformity of the frequencies is quantified by something like reduced chi-square (32), standard minimization routines (33, 34) can find the optimal positions of the CDP's that minimize this non-uniformity. Thus, by finding the optimal positions of the CDP's that minimize the non-uniformity of the frequencies, a model can be determined from the data. As already stated, when the model represents the data, all the measurement correlations can be presented graphically.

2.8. The Big Surprise

The ability to show all measurement correlations is a very important step in fully understanding high-dimensional data, but there is more to the story. It turns out that if the data is partitioned by the positions of the CDP's, the resultant percentages account for population overlap due to measurement error (26). Thus, the overlay graphs that show state means versus progression also show percentages that account for population overlap.

2.9. Multiple Types

Since each Type generates a list of probability weights, the maximum weight from each vector can then be used in a stochastic selection routine to pick the Type. Once the Type is chosen, then another stochastic selection method picks a specific state. Thus, two cascading stochastic selection routines are used to pick a Type and state within that Type for every event or cell in a listmode file. This means that a single model can represent every lineage in a sample from normal bone marrow.

2.10. Overlay Graphics

As the system classifies events into specific Types and states, it keeps a running mean and standard deviation for each state. These means and standard deviations can then form bands for each

measurement showing all measurement correlations and variances along the progression axis. Since the progression axis is in %events units, this type of format also shows all the percentages of contained subpopulations. As mentioned earlier, these percentages account for population overlap due to measurement error. This single easy-to-understand graph can represent the same information as hundreds of bivariate plots.

2.11. Modeling Strategy

In order to synthesize a listmode file representing complex populations found in bone marrow, every one of the Parameter Profiles needs to be defined. However, when reversing directions and modeling data, very little a priori knowledge is necessary for a solution. A good analogy is the crossword puzzle. When a crossword puzzle is created, the author must know the answer to everyone of the clues. When solving a crossword puzzle, the answer to the first clue needs to be known, but later in the solution, the extra information from previous answers enables intelligent guesses for the other clues. Thus, less and less information is necessary during the process of solving the puzzle. In fact, at a certain point, the final solution becomes apparent without the need of any of the remaining clues. Probability State Modeling works in the same way.

2.12. Simple to Complex, Known to Unknown

The strategy in modeling data is to define the simplest Parameter Profiles that are known and move toward more complex Parameter Profiles that are less known. In many cases, once a few of these profiles are defined, every one of the remaining will be obvious from the classified data.

A subtle, but important characteristic of this system is that each Parameter Profile will theoretically distribute the events equally across the states whether it is considered alone or in combination with others. Therefore, a solution to a very complex set of measurements can be approached one measurement at a time. It is not necessary to consider all the measurements at one time. This characteristic makes the analysis normally quite simple and straightforward.

2.13. PSM Results are Understandable to Everyone

Another important point is that the a priori knowledge necessary for the modeling process is identical to the knowledge necessary for interpretation of traditional gating analyses. The only difference between the two approaches is that information is used early when modeling and later for traditional gating interpretation. The advantage of using the knowledge early is that after the modeling is represented as an overlay graph, very little further interpretation is necessary. This attribute is important because it enables non-cytometrists to directly appreciate the results generated by cytometry.

2.14. Compensation is a Requirement for PSM

If the data are not properly compensated (35, 36) for signal cross-over, then the reasons behind coordinated changes in measurements can be ambiguous. Changes in measurement intensities due to

important biological processes can be indistinguishable from simple signal cross-over. The general idea behind PSM is to better understand high-dimensional data. Lack of compensation tends to confuse rather than illuminate and therefore compensated data is a requirement for PSM.

3. PSM Applications

3.1. B-Cell Lineage

3.1.1. B-Cell Progression in Human Bone Marrow

Figure 1 shows a PSM overlay from two four-color listmode files derived from “uninvolved” human bone marrow that were provided by Michael Loken at Hematologies, Inc., Seattle, WA. The files’ measurements are File 1: CD19 APC, CD34 PE, CD22 FITC, CD45 PerCP, SSC, and FSC; and File 2: CD19 APC, CD10 PE, CD20 FITC, CD45 PerCP, SSC, and FSC. The colors of the fluorescence Parameter Profiles are the emission colors of the above fluorochromes. FSC and SSC are light and dark gray respectively. The common measurements among the files are CD19, SSC, CD45, and FSC.

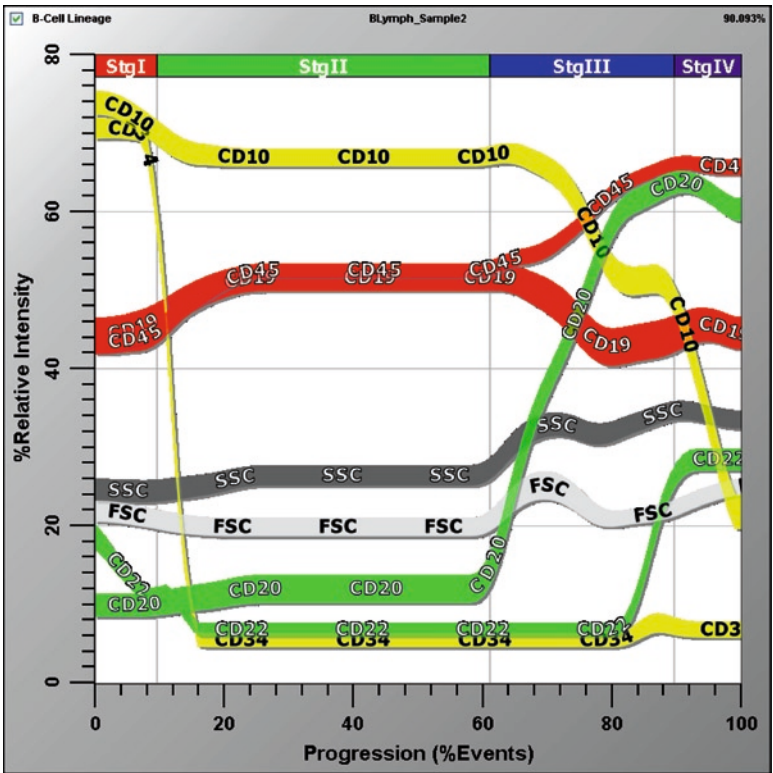


Fig. 1. Eight-dimensional model of B-cell lineage in human bone marrow from two 4-color listmode files. File 1 contained the measurements: CD19, CD34, CD22, CD45, SSC, and FSC. File 2 contained CD19, CD10, CD20, CD45, SSC, and FSC. Four stages of B-cell development are observed characterized by high coordinated changes in marker expression.

CD19 and SSC are used as simple selection measurements to classify all the B-cells in the marrow. For the first file, CD34, CD22, and CD45 stratify many of the B-cell events along the progression axis. For the second file, CD19, SSC, and CD45 select and stratify the events as in file 1, but then CD10 and CD20 further stratify the events to form four apparent stages.

The end of Stage 1 (red in electronic version) is characterized by a downregulation of CD34, CD10, and CD22 and an upregulation of CD45 and CD19. Arrows can be used to more compactly denote this kind of stage boundary definition: CD34↓↓, CD10↓, CD22↓, CD45↑, and CD19↑. A double arrow denotes a dramatic change in marker intensity. The underlined marker indicates that it was solely used for defining the boundary position. The end of Stage 2 (green in electronic version) is characterized by CD20↑↑, CD45↑, SSC↑, and FSC↑. The end of Stage III (blue in electronic version) and beginning of Stage IV (purple in electronic version) are defined by CD10↓↓, CD19↓, and CD22↑.

3.1.2. Common Measurement Scaffolds and High-dimensional Modeling

If common measurements modulate uniquely along a progression, they form a kind of scaffold that all other measurements can relate to whether they are in the same file or different files. With properly designed common measurements, multiple files can be analyzed to form a single model that is representative of the measurement correlations in all files. The CD19, SSC, and CD45 common measurements are good enough to create this scaffold and thus a composite model representing six colors is possible from the analysis of two four-color files (Fig. 1). As evidence of this integration, Fig. 2 shows a bivariate display of CD20 versus CD22; where each measurement is derived from a separate file. In this figure, the CD20 values come from raw measurement values and CD22 values are calculated from the model. The arrows in the figure show the model's predicted progression and the event dot colors match the stage colors.

The implication of this common measurement scaffold is that cytometry is no longer limited to the confines of detector or fluorochrome availability. High-dimensional models are possible from low-dimensional listmode files. The key, of course, is to choose correct common measurements to tie everything together. As discussed later (see Subheading 3.1.6), this modeling system can also help in the selection of the best common measurements.

3.1.3. High-Dimensional B-Cell Progression

Figure 3 shows an eleven-dimensional PSM analysis of “uninvolved” human bone marrow data from Brent Wood at University of Washington, Seattle, WA. The measurements and fluorochromes are kappa FITC, lambda PE, CD19 PE-Texas Red, CD34 PerCP-Cy5.5, CD20 PE-Cy7, CD45 Pacific Blue, CD38 A594, CD10 APC, CD5 APC-Cy7, SSC, and FSC. The colors of the fluorescence Parameter Profiles are the emission colors of the above fluorochromes. FSC and SSC are light and dark gray respectively.

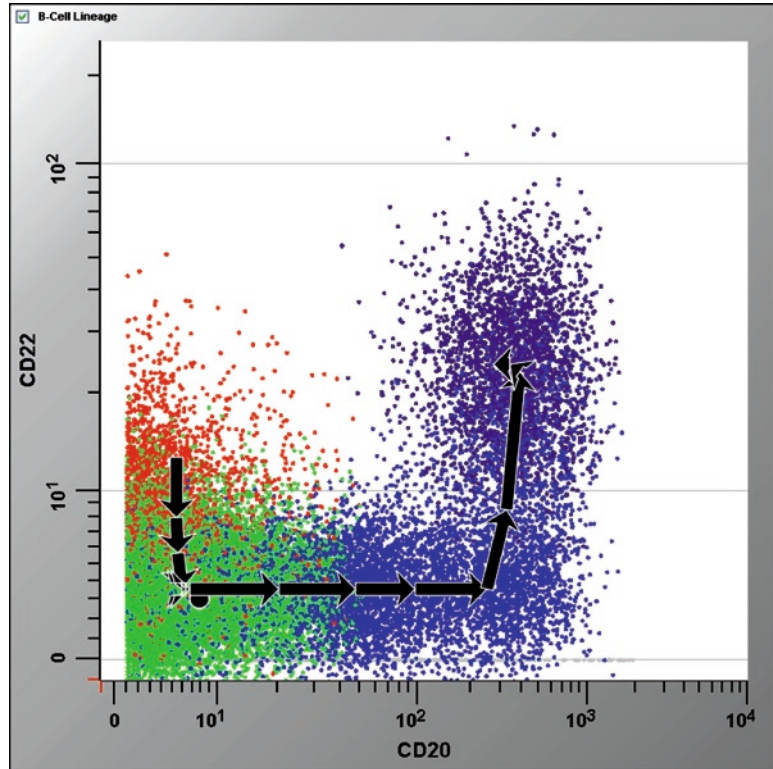


Fig. 2. CD20 versus CD22 bivariate where each measurement is derived from a separate file. CD20 values come from raw measurement values and CD22 values are calculated from the model. The arrows show the model's predicted progression and the event dot colors match the stage colors shown in Fig. 1 (in electronic version).

CD19 and SSC are again used as selection measurements. CD34 and CD38 are modeled as step-down Parameter Profiles and CD20, a step-up. Both CD45 and CD10 are three-level Parameter Profiles. The rest of the Parameter Profiles are not modeled and are distributed along the progression axis by the modeled measurements.

The variations of the additional markers in this file appear tightly coordinated with the B-cell stages defined earlier. Specifically, CD38 elevates slightly while CD34 downregulates and then decreases dramatically while CD10 downregulates. Kappa, lambda, and CD5 begin to be expressed when CD45 and CD20 are upregulated. CD5 is partially downregulated while CD10 and CD38 down regulate.

3.1.4. Importance of the Transitions

There are a number of comments to make concerning the above PSM analyses of B-cell data from bone marrow. The most important point is that when all the measurement correlations are represented in a PSM overlay, B-cell development appears to occur in distinct steps or transitions (Figs. 1 and 3). Each transition involves numerous synchronized changes in specific markers.

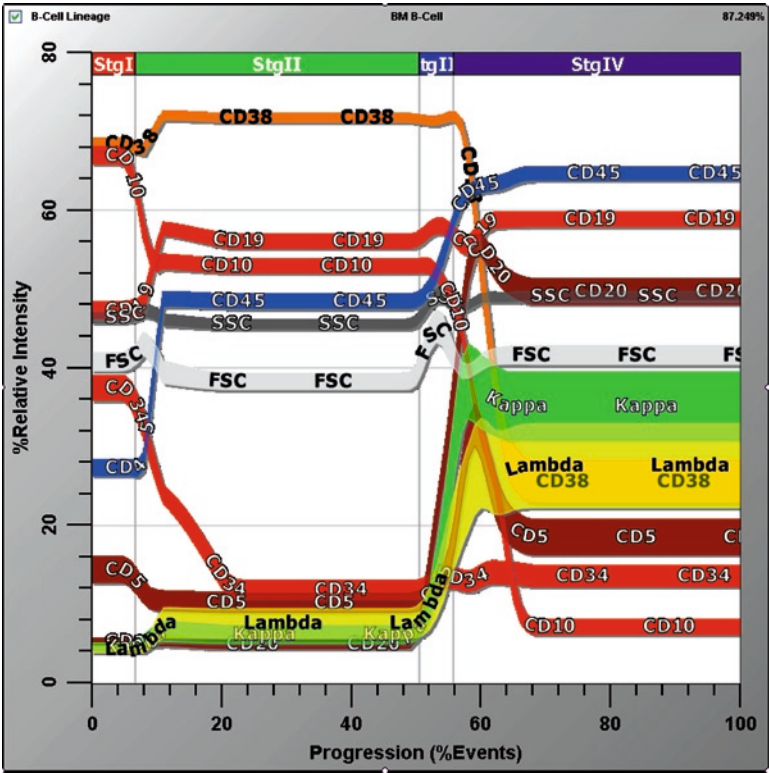


Fig. 3. Model of B-cell lineage in human bone marrow from eleven-dimensional listmode file. The file contains the measurements: kappa, lambda, CD19, CD34, CD20, CD45, CD38, CD10, CD5, SSC, and FSC. The variations of these markers appear to be tightly coordinated with the stages shown in Fig. 1. Specifically, CD38 elevates slightly while CD34 downregulates and then decreases dramatically, while CD10 downregulates. Kappa, lambda, and CD5 begin to be expressed when CD45 and CD20 are upregulated. CD5 is partially downregulated while CD10 and CD38 downregulate.

These observations are consistent with the hypothesis that B-cell ontogeny is controlled by master regulatory genes (37). The ability to clearly visualize these transitions and sort them may eventually give molecular biologists the tools they need to determine which master regulatory genes are involved in the different stages of B-cell ontogeny. Other lineages may also have similar control mechanisms (38–40).

3.1.5. The Ideal B-Cell Panel

Identifying the important transitions in B-cell lineage progression also helps in the prediction of which markers are likely to be most important in staging B-cells. CD34 and TdT (not shown) are important in identifying the first transition (Stage I to II). CD10, CD45, and CD20 help identify all three transitions. In general, markers that have relatively small line-spreads and that dramatically change intensity during one or more of these transitions are ideal candidates for common measurements in a multi-tube B-cell panel.

3.1.6. Empirical
Determination of the Best
Antibody Panel

PSM provides another approach to the determination of an optimal set of common measurements. When an event is stochastically selected into a specific state along the model’s progression axis, the program estimates the probability of the decision being correct. The average of this probability for all classified events can be expressed as a percent and labeled as %Fidelity Index (%FI). This percentage quantifies how well all measurements in the model work together to categorize events along the progression axis.

Since each modeled measurement can be including or excluding from this classification process, it is possible to examine all combinations of relevant measurements and rank them according to their %FI. Table 1 shows the result of such an analysis on the above eleven-dimensional bone marrow B-cell data.

Selection measurements such as CD19 and SSC are not part of this analysis since they don’t stratify events along the progression axis. This analysis suggests that along with CD19, the markers CD34 and CD45 would be the best three-color choice to stage bone marrow B-cells (%FI=4.17). The best four-color

Table 1
Ranking of Antibody Panels

Number non-constant markers	Markers	%Fidelity index
0		1.00
1	CD20	2.10
1	CD34	2.22
1	CD38	2.50
1	CD45	3.07
2	CD34 CD20	3.44
2	CD20 CD45	3.58
2	CD38 CD20	3.59
2	CD38 CD34	3.70
2	CD38 CD45	4.07
2	CD34 CD45	4.17
3	CD34 CD20 CD45	4.54
3	CD38 CD20 CD45	4.66
3	CD38 CD34 CD20	4.68
3	CD38 CD34 CD45	4.95
4	CD38 CD34 CD20 CD45	5.40

choice would be CD19, CD38, CD34, and CD45 (%FI=4.95). These choices are only relevant to this particular sample. If a number of samples were analyzed in this manner, it would be possible to engineer an ideal panel of antibodies based on the objective %FI measurements.

3.2. Other Lineages

The advantages of PSM analyses of B-cell lineages are also applicable to other lineages or progressions in bone marrow, peripheral blood, and other tissue sites. The remaining portion of this chapter will show a sampling of these PSM analyses.

3.2.1. Myeloid Progression in Human Bone Marrow

Figure 4a, b show a seventeen-dimensional PSM analysis of “uninvolved” human bone marrow data from Brent Wood at University of Washington. The measurements and fluorochromes from the first twelve-dimensional file are CD15 FITC, CD33 PE,

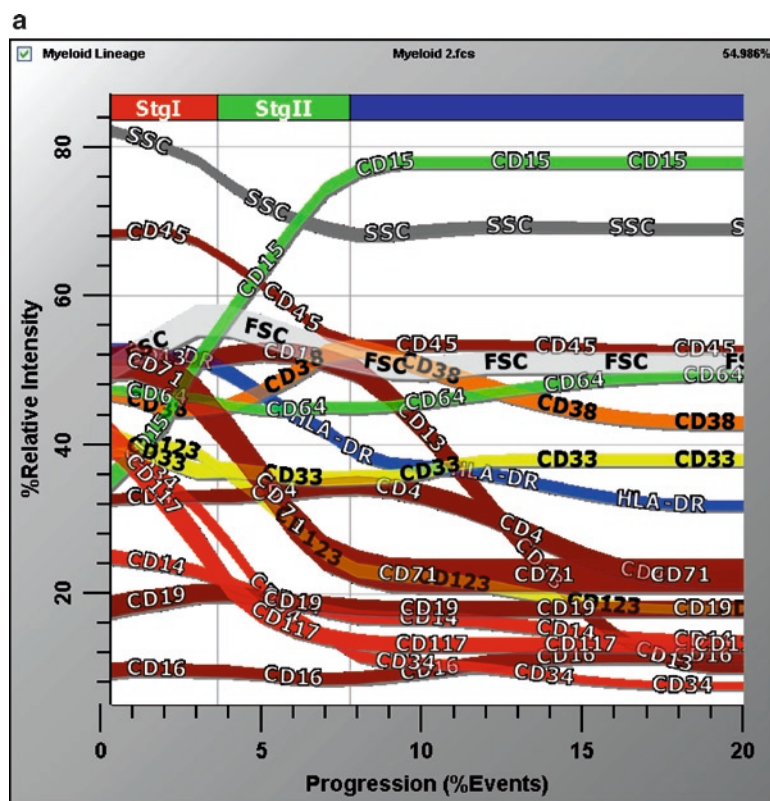


Fig. 4. Seventeen-dimensional model of human myeloid lineage in bone marrow from two twelve-dimensional listmode files. File 1 contains the measurements CD15, CD33, CD19, CD117, CD13, HLA-DR, CD38, CD34, CD71, CD45, FSC, and SSC. File 2 contains the measurements CD64, CD123, CD4, CD14, CD13, HLA-DR, CD38, CD34, CD16, CD45, FSC, and SSC. Myeloid lineage appears to also occur in distinct steps or transitions, suggesting a similar type of controlling mechanism as hypothesized for B-cells. (a) is zoomed in on the first 20% of the progression axis to better visualize early stages. (b) is the complete progression.

b

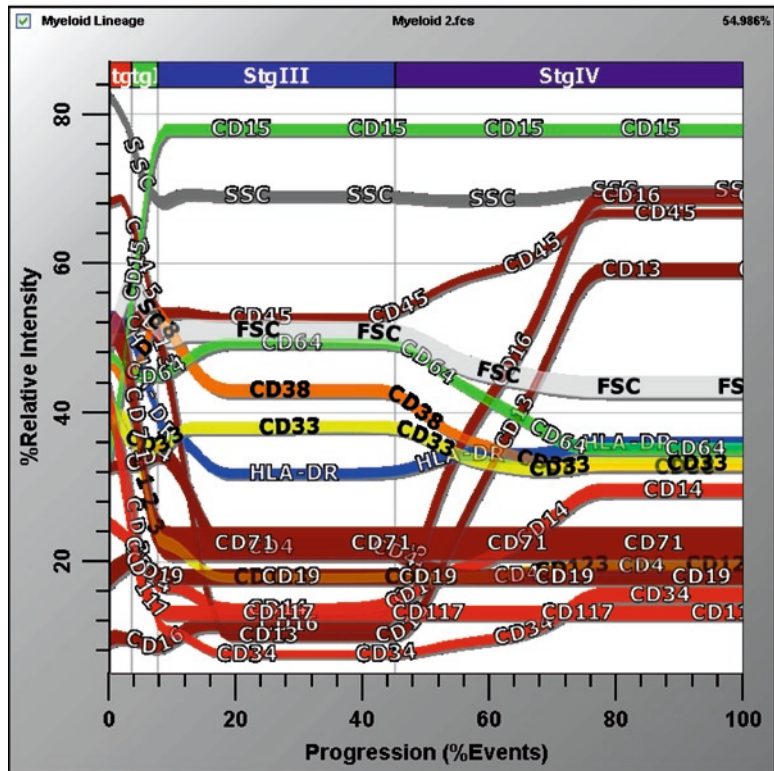


Fig. 4. (continued)

CD19 PE-TR, CD117 PE-Cy5, CD13 PE-Cy7, HLA-DR Pacific Blue, CD38 A594, CD34 APC, CD71 APC-A700, CD45 APC-Cy7, FSC, and SSC. The second twelve-dimensional file's markers are CD64 FITC, CD123 PE, CD4 PE-TR, CD14 PE-Cy5.5, CD13 PE-Cy7, HLA-DR Pacific Blue, CD38 A594, CD34 APC, CD16 APC-A700, and CD45 APC-Cy7, FSC, and SSC. The colors of the fluorescence Parameter Profiles are the emission colors of the above fluorochromes. FSC and SSC are light and dark gray respectively. The stage boundaries for this myeloid data are defined as

End of stage I (Fig. 4a): HLA-DR↓↓, CD34↓↓, and CD15↑↑

End of stage II (Fig. 4a): CD13↓↓, CD123↓↓, CD4↓, CD45↓, CD38↓, FSC↑, and SSC↑

End of stage III, start of stage IV (Fig. 4b): CD16↑↑, CD13↑↑, CD45↑, CD38↓, CD33↓, and CD14↑

The above patterns represent only neutrophil maturation and the relatively high intensity of HLA-DR is due to autofluorescence.

Figure 4a is zoomed in on the first 20% of the progression axis to better visualize early stages. Figure 4b is the complete

progression. When all measurement correlations are viewable by PSM, the myeloid lineage also appears to occur in distinct steps or transitions, suggesting a similar type of controlling mechanism as hypothesized for B-cells.

3.2.2. CD4 T-Cell Antigen-Dependent Development in Human Peripheral Blood

Figure 5 shows a ten-dimensional PSM analysis of CD4 T-cell's antigen-dependent progression in normal human peripheral blood data from Margaret Inokuma at BD Biosciences, San Jose, CA. The measurements and fluorochromes are CD3 Pacific Blue, CD8 APC-Cy7, CD4 AmCyan, CD27 APC, CD28 PerCP-Cy5.5, CD57 FITC, CCR7 PE, CD45RA PE-Cy7, SSC, and FSC. The colors of the fluorescence Parameter Profiles are the emission colors of the above fluorochromes. FSC and SSC are light and dark gray respectively. The stage boundaries for this data are defined as

End of Naïve stage: $\text{CD45RA}\downarrow\downarrow$, $\text{CD28}\uparrow$, $\text{CCR7}\downarrow$, and $\text{CD3}\downarrow$

End of T1 stage: $\text{CCR7}\downarrow\downarrow$, and $\text{CD45RA}\downarrow$

End of T2 stage: $\text{CD27}\downarrow\downarrow$

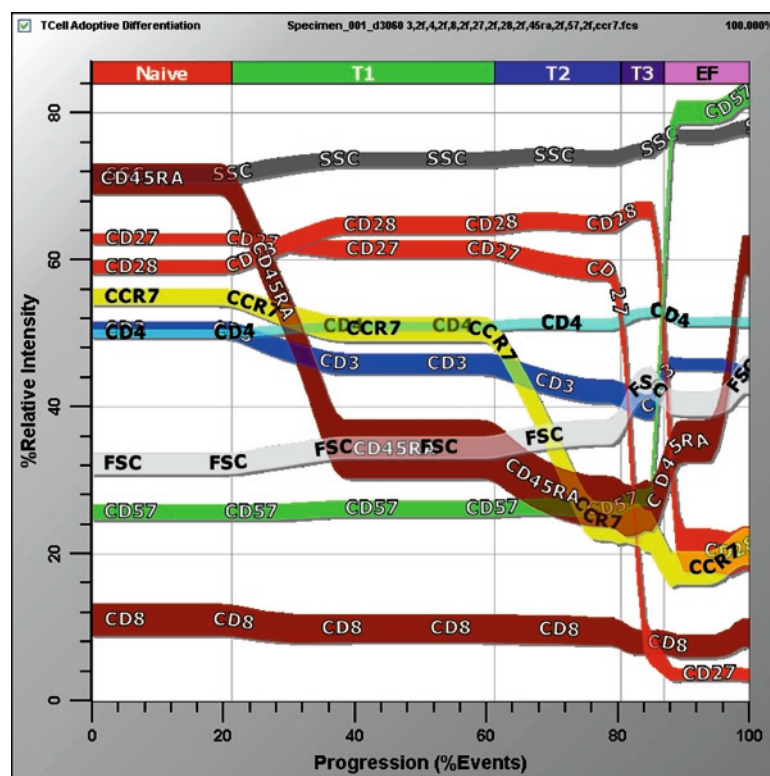


Fig. 5. Ten-dimensional model of CD4 T-cells antigen-dependent progression in human normal peripheral blood – smoothed. The listmode file contains the measurements: CD3, CD8, CD4, CD27, CD28, CD57, CCR7, CD45RA, SSC, and FSC. The stage boundaries for this data are defined as end of Naïve stage: $\text{CD45RA}\downarrow\downarrow$, $\text{CD28}\uparrow$, $\text{CCR7}\downarrow$, and $\text{CD3}\downarrow$; end of T1 stage: $\text{CCR7}\downarrow\downarrow$, and $\text{CD45RA}\downarrow$; end of T2 stage: $\text{CD27}\downarrow\downarrow$; and end of T3 stage, start of EF (effector) stage: $\text{CD28}\downarrow\downarrow$, $\text{CD57}\uparrow\uparrow$, and $\text{CD45RA}\uparrow\uparrow$.

End of T3 stage, start of EF (effector) stage: $\text{CD28}\downarrow\downarrow$, $\text{CD57}\uparrow\uparrow$, and $\text{CD45RA}\uparrow\uparrow$

The cascading nature of the measurement intensity changes in the antigen-dependent T-cell maturation suggests a different type of underlying mechanism of development than bone marrow B-cells and myeloid cells.

3.2.3. Erythroid Progression in Mouse Bone Marrow

Figure 6 shows a nine-dimensional PSM analysis of erythroid development in mouse bone marrow data from Kathleen McGrath at University of Rochester, Rochester, NY. The measurements and when appropriate, fluorochromes, from an Amnis Image Stream X listmode file are Area Cell, Area Nucleus, Mean Vybrant Violet (Mean VV), CD44 PE-Cy5.5, CD71 PE, c-Kit PE-Cy7, Ter119 APC, Mean Ter119 APC, and Thiazole Orange (TO).

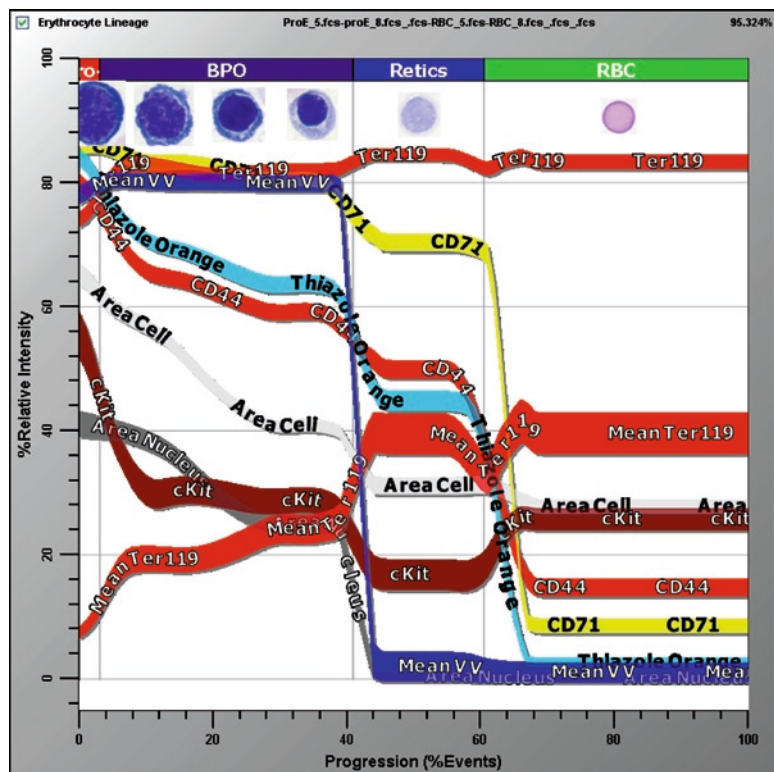


Fig. 6. Nine-dimensional model of erythroid development in mouse bone marrow. The listmode file contains the measurements: Area Cell, Area Nucleus, Mean Vybrant Violet (Mean VV), CD44, CD71, c-Kit, Ter119, Mean Ter119, and Thiazole Orange (TO). The morphology of the erythroid lineage stages are shown at the top. In the beginning of the erythroid progression, c-Kit is positive for the proerythroblasts. As progression continues, Ter119 increases slightly, while TO (RNA), Area Cell, Area Nucleus, and CD44 decrease. The point of nucleus extrusion is easily identified by a dramatic decrease in Mean VV (DNA) and Area Nucleus. At the end of the reticulocyte stage, TO, CD71, and CD44 decrease.

The colors of the fluorescence Parameter Profiles are the emission colors of the above fluorochromes. Area Cell and Area Nucleus are light and dark gray respectively. Mean VV detects double-stranded DNA in viable cells and TO detects RNA.

The basophilic, polychromatic, and orthochromatic stages of erythroid development were not clearly separated by these measurements and were grouped together as the BPO stage. The stage boundaries for this data are defined as

End of Proerythroblasts: c-Kit↓↓, TO↓, CD44↓, Area Cell↓, Area Nucleus↓, and Ter119↑

End of BPO: Mean VV↓↓, Area Nucleus↓↓, TO↓, CD71↓, CD44↓, and Area Cell↓

End of Reticulocytes, start of RBC: TO↓, CD71↓↓, and CD44↓

The morphology of the erythroid lineage stages are shown at the top of Fig. 6. In the beginning of the erythroid progression, c-Kit is positive for the proerythroblasts. As progression continues, Ter119 increases slightly, while TO (RNA), Area Cell, Area Nucleus, and CD44 decrease. The Mean Ter119 has been shown to have structural information in it that helps delineate the basophilic, polychromatic, and orthochromatic stages (41). The point of nucleus extrusion is easily identified by a dramatic decrease in Mean VV (DNA) and Area Nucleus. At the end of the reticulocyte stage TO, CD71, and CD44 decrease.

4. Summary

Cytometry has been seriously limited by the overwhelming complexity in viewing and analyzing high-dimensional data. The key to solving this problem was the design of a mathematical model system capable of representing complex measurement correlations normally found in living tissues. Later, it was discovered that stochastic selection coupled with a general minimization algorithm could be leveraged to construct and optimize these models to diverse high-dimensional listmode data. Since this method divides progressions into individual states and searches for a solution that makes these states equally probable for event selection, it was called Probability State Modeling or PSM. PSM breaks through the complexity barrier by creating a new progression probability-based variable that all measurements can relate to. Since all measurements relate to the same progression axis, a single graphical overlay with progression on the x -axis can represent all the correlations present in high-dimensional data.

When PSM was applied to complex bone marrow populations, these graphs revealed a rich tapestry of measurement changes secondary to underlying biochemical processes. In the

case of bone marrow B-cells, the data shows only three transitions where measurement intensities change in a tightly correlated manner consistent with the hypothesis that B-cell ontogeny is controlled by master regulatory genes. A sampling of other lineages and progressions showed that the PSM approach is quite general and can be very revealing in deducing possible biological mechanisms underlying cellular ontogeny.

Enabling the visualization of high-dimensional data is only a small part of the advantages of PSM. Regions or zones defined along the progression axis automatically account for population overlap due to measurement error. A single representation of a sample can be constructed from numerous files if there are proper common measurements present in each file. PSM can also objectively quantify how well specific measurements might work as common measurements.

The major advantage, however, is that PSM forces scientists to create unambiguous models of biological systems. Model building and hypothesis testing are central to the scientific method. True understanding of these processes will ultimately lead to solutions for important problems that threaten our society.

References

1. Chattopadhyay, P. K., Price, D. A., Harper, T. F., Betts, M. R., Yu, J., Gostick, E., Perfetto, S. P., Goepfert, P., Koup, R. A., DeRosa, C., Bruchez, M. P., and Roederer, M. (2006) Quantum dot semiconductor nanocrystals for immunophenotyping by polychromatic flow cytometry. *Nat. Med.* **12**, 972.
2. Perfetto, S. P., Chattopadhyay, P. K., and Roederer, M. (2004) Unravelling the immune system. *Nat. Rev. Immunol.* **4**, 648–55.
3. Bandura, D. R., Baranov, V. I., Ornatsky, O. I., Antonov, A., Kinach, R., Lou, X., Pavlov, S., Voroviev, S., Dick, J. E., and Tanner, S. D. (2009) Mass cytometry: technique for real time single cell multitarget immunoassay based on inductively coupled plasma time-of-flight mass spectrometry. *Anal. Chem.* **81**, 6813–22.
4. Ornatsky, O. I., Lou, X., Nitz, M., Schafer, S., Sheldrick, W. S., Baranov, V. I., Bandura, D. R., and Tanner, S. D. (2008) Study of cell antigens and intracellular DNA by identification of element-containing labels and metal-lointercalators using inductively coupled plasma mass spectrometry. *Anal. Chem.* **80**, 2539–47.
5. Tanner, S. D., Bandura, D. R., Ornatsky, O., Baranov, V. I., Nitz, M., and Winnik, M. A. (2008) Flow cytometer with mass spectrometer detection for massively multiplexed single-cell biomarker assay. *Pure Appl. Chem.* **80**, No. 12, 2627–41.
6. Tan, P., Steinback, M. and Kumar, V. (2006) *Introduction to Data Mining*, Pearson Education, Boston, MA, pp. 51–63.
7. Baggerly, K. A. (2001) Probability binning and testing agreement between multivariate immunofluorescence histograms: extending the chi-squared test. *Cytometry* **45**, 141–50.
8. Roederer, M., Moore, W., Treister, A., Hardy, R. R., and Herzenberg, L. A. (2001) Probability binning comparison: a metric for quantifying multivariate distribution differences. *Cytometry* **45**, 47–55.
9. Rogers, W. T., Moser, A. R., Holyst, H. A., Bantly, A., Mohler II, E. R., Scangas, G., and Moore, J. S. (2008) Cytometric fingerprinting: quantitative characterization of multivariate distributions. *Cytometry* **73A**, 430–41.
10. Kosugi, Y., Sato, R., Genka, S., Shitara, N., and Takakura, K. (1988) An interactive multivariate analysis of FCM data. *Cytometry* **9**, 405–8.
11. Lugli, E., Pinti, M., Nasi, M., Troiano, L., Ferraresi, R., Mussi, C., Salviololi, G., Patsekina, V., Robinson, J. P., Djurante, C., Cocchi, M., and Cossarizza, A. (2007) Subject classification obtained by cluster analysis and

- principal component analysis applied to flow cytometric data. *Cytometry Part A* **71A**, 334–44.
12. Bagwell C. B., Horan P., and Lovett, E. (1985) A method for displaying multiparameter flow cytometric listmode data. International Conference Analytical Cytology XI, November, 17–22.
 13. Leary, J. F., Ellis, S. P., McLaughlin, S. R., Corio, M. A., Hespelt, S., Gram, J. G., and Burde, S. (1991) High-resolution separation of rare-cell types, in *Cell Separation Science and Technology* (Kompala, P. and Todd, P. F., eds.) American Chemical Society Press, Washington, DC, Series No. **464**, pp. 26–40.
 14. Murphy, R. (1985) Automated identification of subpopulations in flow cytometric list mode data using cluster analysis. *Cytometry* **6**, 302–9.
 15. Wegman, E.J. and Luo, Q. (1997) High dimensional clustering using parallel coordinates and the grand tour. *Comput. Sci. Stat.* **28**, 352–60.
 16. Preffer, I. F., Dombkowski, D., Sykes, M., Scadden, D., and Yang, Y.-G. (2002) Lineage-negative side-population (SP) cells with restricted hematopoietic capacity circulate in normal human adult blood: immunophenotypic and functional characterization. *Stem Cells* **20**, 417–27.
 17. Preffer, F. and Dombkowski, D. Advances in complex multiparameter flow cytometry technology: applications in stem cell research (2009) *Cytometry Part B* **76B**, 295–314.
 18. Dean, P. (1990) Data processing, in *Flow Cytometry and Sorting* (Melamed, M.R., Lindmo, T., and Mendelsohn, M.I., eds.), Wiley-Liss, Hoboken, NJ, pp. 438–40.
 19. Crowell J. M., Hiebert, R. D., Salzman, G. B., Price, M. J., Cram, L. S., and Mullaney, P. F. (1978) A light-scattering system for high-speed cell analysis. *IEEE Trans. Biomed. Eng.* **BME-25**, 519–26.
 20. Finn, W. G., Carter, K. M., Raich, R., Stoolman, L., and Hero, A. O. (2009) Analysis of clinical flow cytometric immunophenotyping data by clustering on statistical manifolds: treating flow cytometry data as high-dimensional objects. *Cytometry Part B* **76B**, 1–7.
 21. Chan, C., Feng, F., Ottinger, J., Foster, D., West, M., and Kepler, T. B. (2008) Statistical mixture modeling for cell subtype identification in flow cytometry. *Cytometry Part A* **73A**, 693–701.
 22. Irish, J. M., Hovland, R., Krutzik, P. O., Perez, O. D., Bruserud, O., Gjertsen, B. T., and Nolan, G. P. (2004) Single cell profiling of potentiated phosphor-protein networks in cancer cells. *Cell* **118**, 217–28.
 23. Boedigheimer, M. J. and Ferbas, J. (2008) Mixture modeling approach to flow cytometry data. *Cytometry Part A* **73A**, 421–29.
 24. Bagwell, C. B. (2007) Probability State Models. Utility Application No. US11/897,148, 19 Sep.
 25. Bagwell, C. B. (2008) Breaking the Dimensionality Barrier, in *Laboratory Hematology Practice* (Kottke-Marchant, K. and Davis, B.H., eds.), Wiley-Blackwell, Hoboken, NJ, **Ch 12**.
 26. Bagwell, C. B. (2009) Probability State Modeling – a new paradigm for cytometric analysis, in *Flow Cytometry In Drug Discovery and Development* (Litwin, V. and Marder, P., eds.), John Wiley and Sons, Inc., Hoboken, NJ, **Ch 15**.
 27. Shapiro, H. M. (2003) *Practical Flow Cytometry*, 4th edition, Wiley-Liss, Hoboken, NJ, pp. 465–7.
 28. Loken, M. R., Shah, V. O., Dattilio, K. L., and Civin, C. I. (1987) Flow cytometric analysis of human bone marrow. II. Normal B lymphocyte development. *Blood* **70**, 1316–24.
 29. Loken M. R. and Wells, D. A. (2000) Normal antigen expression in hematopoiesis, in *Immunophenotyping* (Stewart, C. C. and Nicholson, J. K., eds.), Wiley-Liss, Hoboken, NJ, pp. 138–142.
 30. Wood, B. (2004) Multicolor immunophenotyping: human immune system hematopoiesis. *Methods Cell Biol* **75**, 559–76.
 31. Gentle, J. E. (2003) Transformations of uniform deviates: general methods, in *Random Number Generation and Monte Carlo Methods*, 2nd edition, Springer Science + Business Media, LLC, New York, NY, pp. 101–9.
 32. Bevington, P. R. (1969) Data reduction and error analysis for the physical sciences. McGraw-Hill Book Company, New York, NY, p. 89.
 33. Bevington, P. R. (1969) Data reduction and error analysis for the physical sciences. McGraw-Hill Book Company, New York, NY, p. 245.
 34. Press, W. H., Vetterling, W. T., Teukolsky, S. A., and Flannery, B. P. (1992) Numerical recipes in C, 2nd edition, Cambridge University Press, New York, NY, pp. 408–12.
 35. Bagwell, C. B. and Adams, E. G. (1993) Fluorescence spectral overlap compensation for any number of flow cytometry parameters. *Ann NY Acad Sci* **677**, 167–84.
 36. Roederer, M. (2001) Spectral compensation for flow cytometry: visualization artifacts, limitations, and caveats. *Cytometry* **45**, 194–205.

37. Lawrence, H. J., Savageau, G., Largman, C., and Humphries, R. K. (2001) Homeobox gene networks and the regulation of hematopoiesis, in *Hematopoiesis : A developmental approach* (Zon, L. I., ed.), Oxford University Press, New York, NY, pp. 404–5.
38. Argiropoulos, B. and Humphries, R. K. (2007) Hox genes in hematopoiesis and leukemogenesis. *Oncogene* **26**, 6766–76.
39. Kim, S. I. and Bresnick, E. H. (2007) Transcriptional control of erythropoiesis: emerging mechanisms and principles. *Oncogene* **26**, 6777–94.
40. Pronk, C. J., Rossi, D. J., Mansson, R., Attema, J. L., Norrdahl, G. L., Chan, C. K. et al. (2007) Elucidation of the phenotypic, functional, and molecular topography of a myeloerythroid progenitor cell hierarchy. *Cell Stem Cell* **1**, 428–42.
41. McGrath, K. E., Bushnell, T. P., and Palis, J. (2008) Multispectral imaging of hematopoietic cells: where flow meets morphology. *J. Immunol. Methods* **336**, 91–7.



<http://www.springer.com/978-1-61737-949-9>

Flow Cytometry Protocols

Hawley, T.S.; Hawley, R.G. (Eds.)

2011, XII, 486 p., Hardcover

ISBN: 978-1-61737-949-9

A product of Humana Press