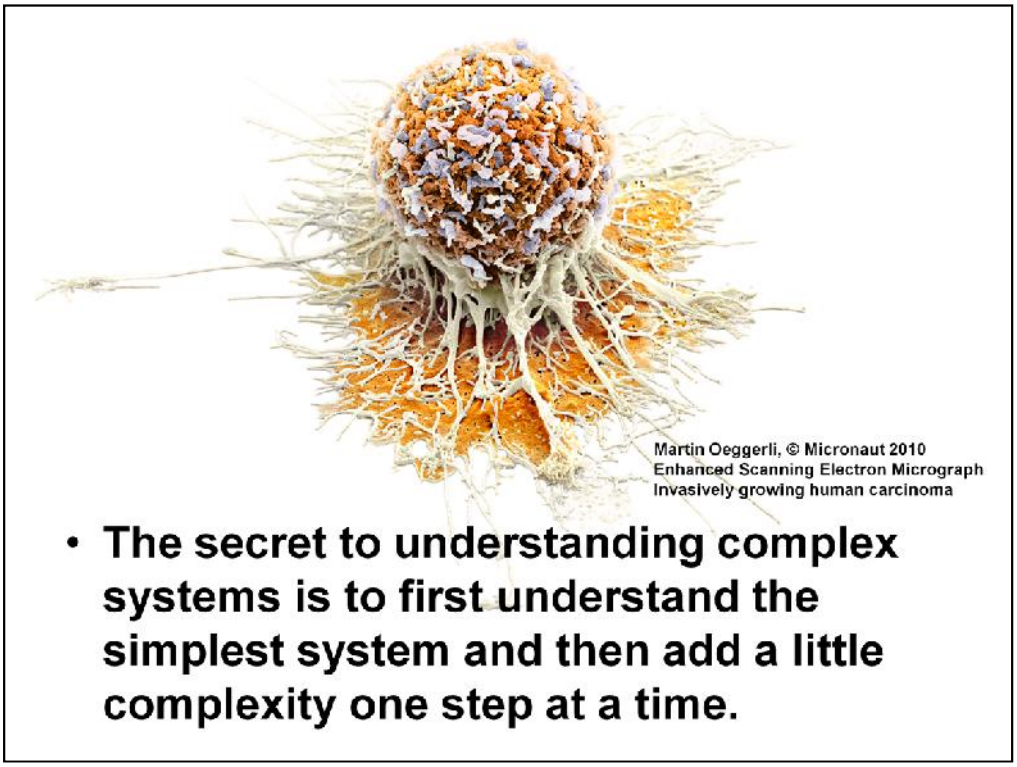
Modeling Kinetic Processes Without The Benefit Of Time As A Measurement

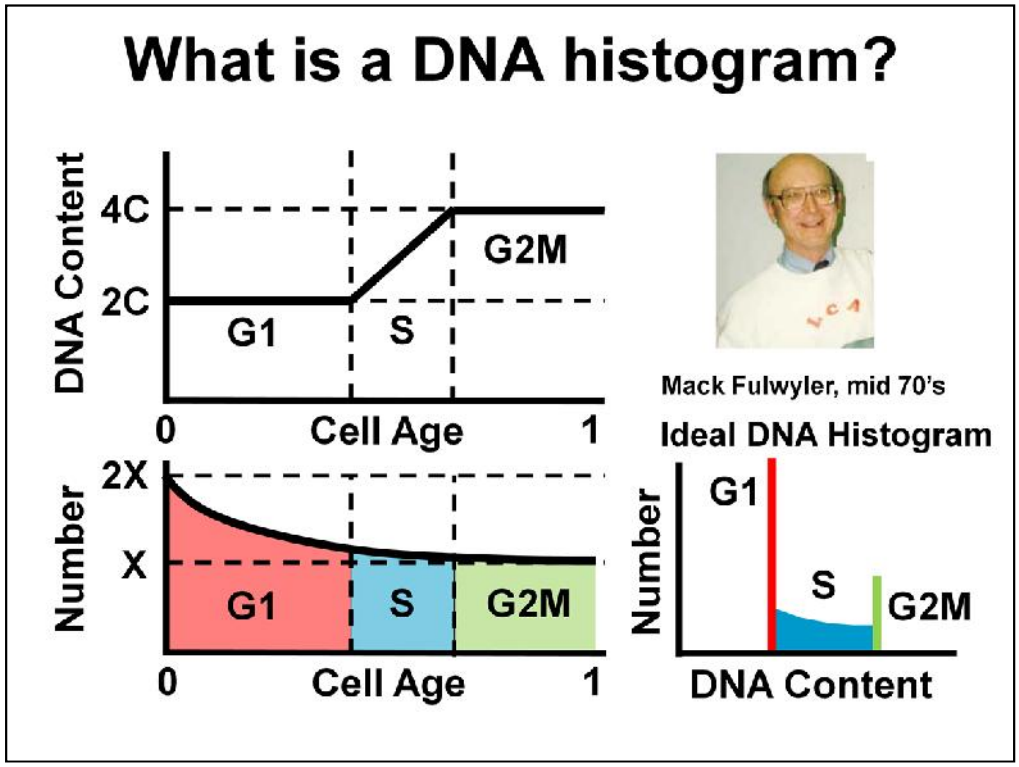
C. Bruce Bagwell MD, Ph.D.
Verity Software House

The literature is rich with articles describing mathematical models that can represent kinetic processes. Many times, however, there are systems were it is difficult to directly measure time or relative time. This lecture shows that even without the benefit of time as a measurement, complex kinetic progressions can be successfully modeled. The applications of this technology were originally intended to study cellular progressions in cytometry, but the theory is sufficiently general to handle any type of system where one can make correlated measurements on numerous objects, but just can't directly measure time.



Cells are wonderfully complex chemical machines and modeling them to reveal their hidden secrets can indeed be a challenging prospect. The secret to understanding and modeling complex systems like cells is to first understand the simplest possible system and then add a little complexity one step at a time.

One way of searching for a good simple starting point is to go back in time and examine some of the simpler systems. Let's go back to the early and middle 1970's where cytometry was still in its infancy. Most of the interest in cytometry at that time was with DNA histogram analysis. This would quickly change in the 1980's when AIDS was discovered but at this time, it was thought DNA content analysis would play a pivotal role in understanding and treating malignancies.

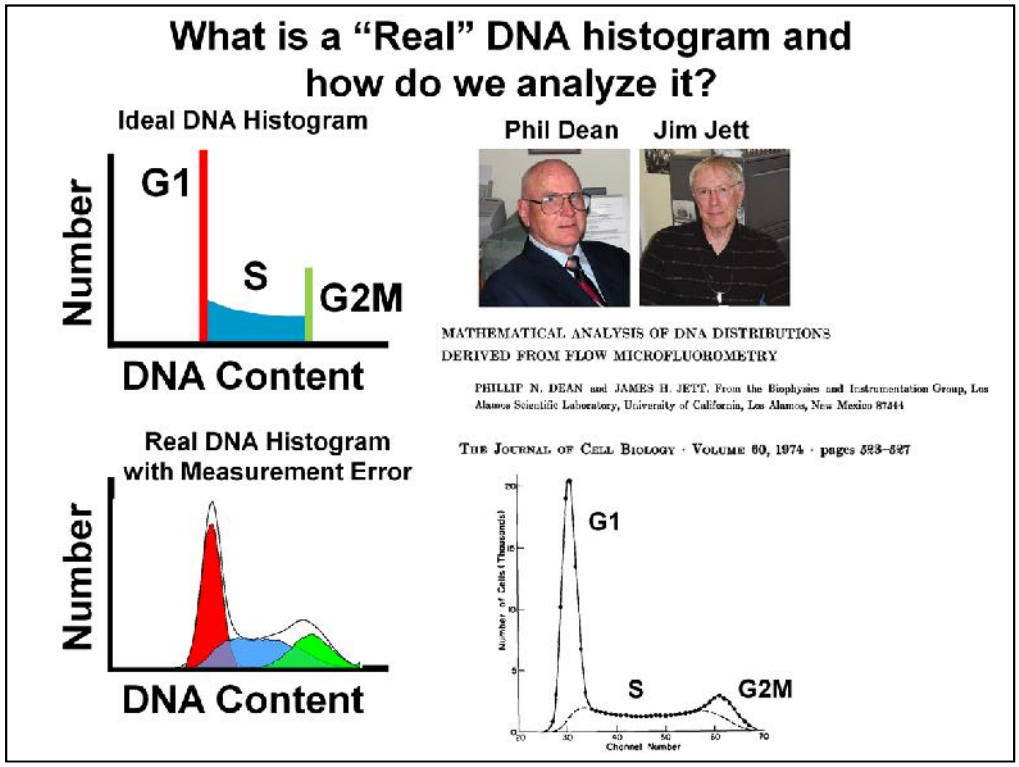


The first time I saw the story of "what is a DNA histogram" was in the mid-1970's when Mack Fulwyler, the inventor of cell sorters, was giving a black-board talk on DNA histogram analysis. From what we can gather, he was describing what he had been taught by Marv Van Dilla or Phil Dean at Los Alamos. He began by drawing two axis (1) which he labeled as Cell Age and DNA Content (2). Cell Age was a relative time scale that ranged from 0 (right after division) to 1 (just before division). Cells begin the cycle with 2C (complements) of DNA and when the DNA synthetic machinery turns on, their nuclear DNA increases until the genome is duplicated at 4C (top graph). Thus, the cycle naturally divides into G1, S, and G2M phases along the Cell Age axis.

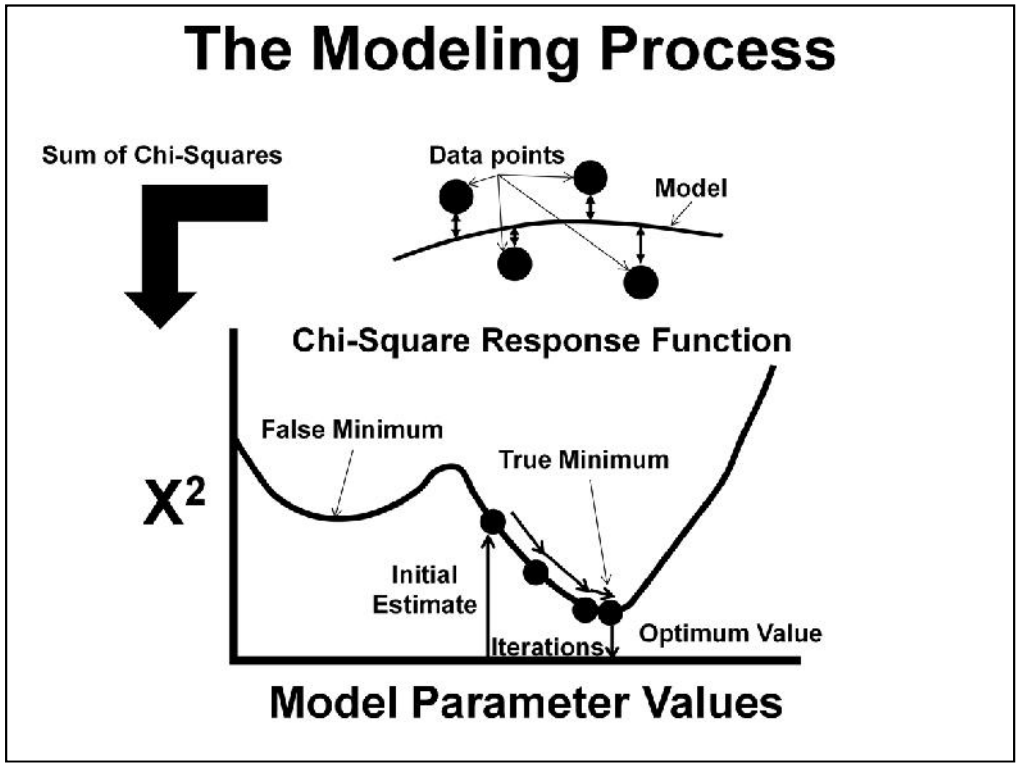
Mack then drew two more axes below the first graph. The x-axis was the same, Cell Age, but the y-axis was labeled as Number (Number of Events). He indicated that the density of events or cells immediately after cell division, Cell Age=0, for asynchronous population of dividing cells was 2X and just before the cells divide, was X. For an exponentially growing population the theoretical relationship was a decreasing exponential starting at 2X and ending at X where X was dependent on the area under the curve.

He then drew two more axes with DNA Content on the x-axis and Number on the y-axis and labeled the graph as an Ideal DNA Histogram. He showed that the area under the G1 part of the curve is represented as a spike with its height equal to the area and likewise, the G2M part of the curve was also a spike with the same height as its respective area. He reasoned that as Cell Age moved from left to right, the DNA content would increase in the DNA Histogram plot and the number of events would decrease as a truncated exponential.

What Mack really did was show that the Ideal DNA Histogram could be derived from two functions that were related to Cell Age.

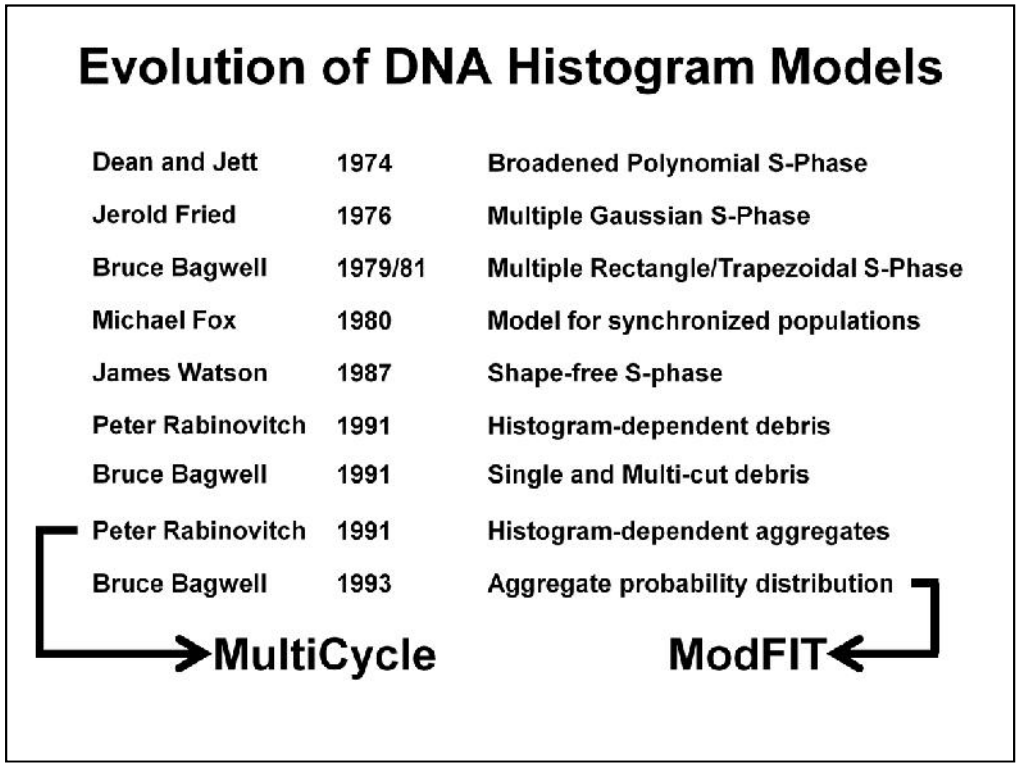


He then indicated that because there is always uncertainty in measuring things with the cytometer, each point in the Ideal DNA Histogram was really represented as a normal (Gaussian) distribution. The summation of these normal distributions along with statistical noise form the Real DNA Histogram, which looks just like the DNA histogram we measure with our cytometers. But how do we analyze this data? Phil Dean and Jim Jett were the first to tackle this problem using a modeling technique. They decided to use normal distributions for G1 and G2M along with a broadened polynomial to fit the observed DNA histogram (bottom graph). How did they find the optimal values for their model?

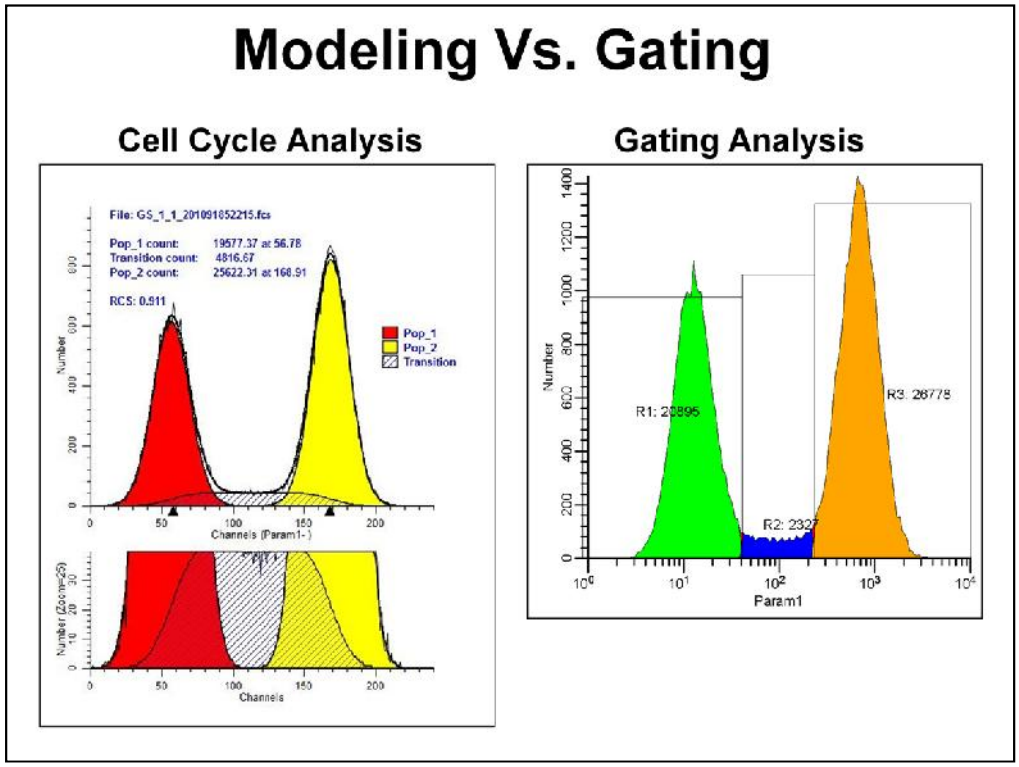


The modeling process involves mathematically quantifying the overall degree of difference between the model and the observed data points. Typically, this is done by summing the weighted squared differences known as chi-square. For a single parameter in the model, this can be visualized as a special response function that has some set of valleys, where one is deeper than all the others. This valley is called the true minimum. Modeling begins with an estimate that is close enough to the True Minimum that it can follow the gradient through an iterative process to the optimum value. This method is easily extended to many model parameters. It's important to distinguish between a model parameters and a cytometric measurement often called a parameter.

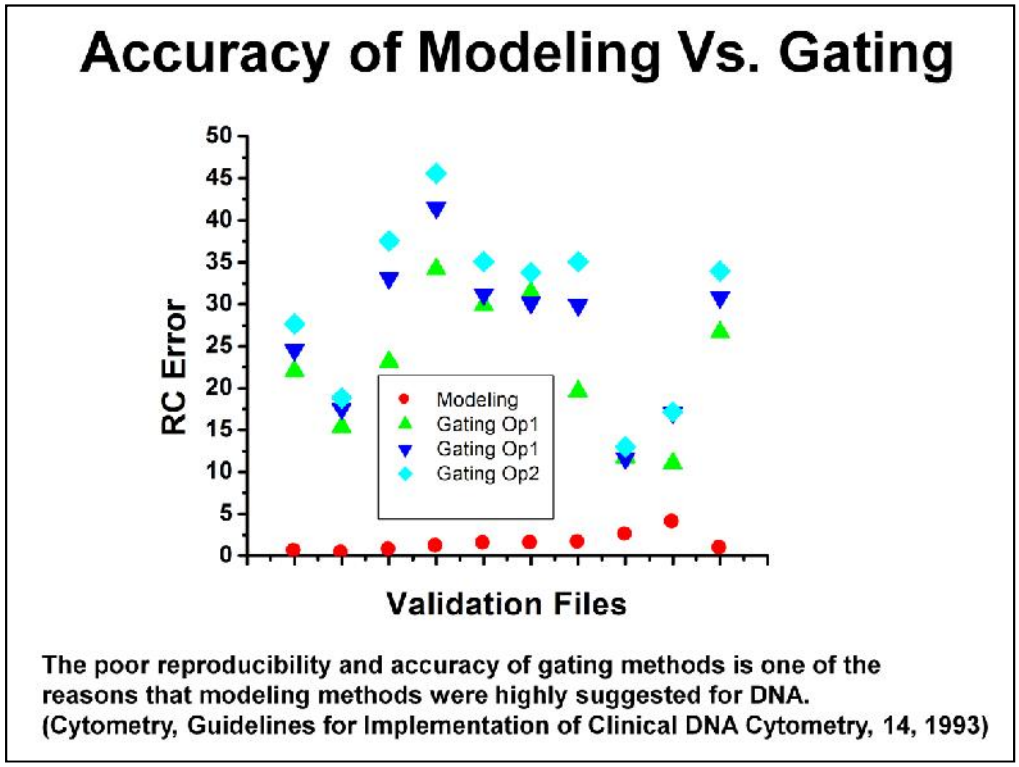
The Dean and Jett broadened polynomial required the optimization of 9 model parameters. It began a vigorous search for the optimal model for DNA histogram analysis.



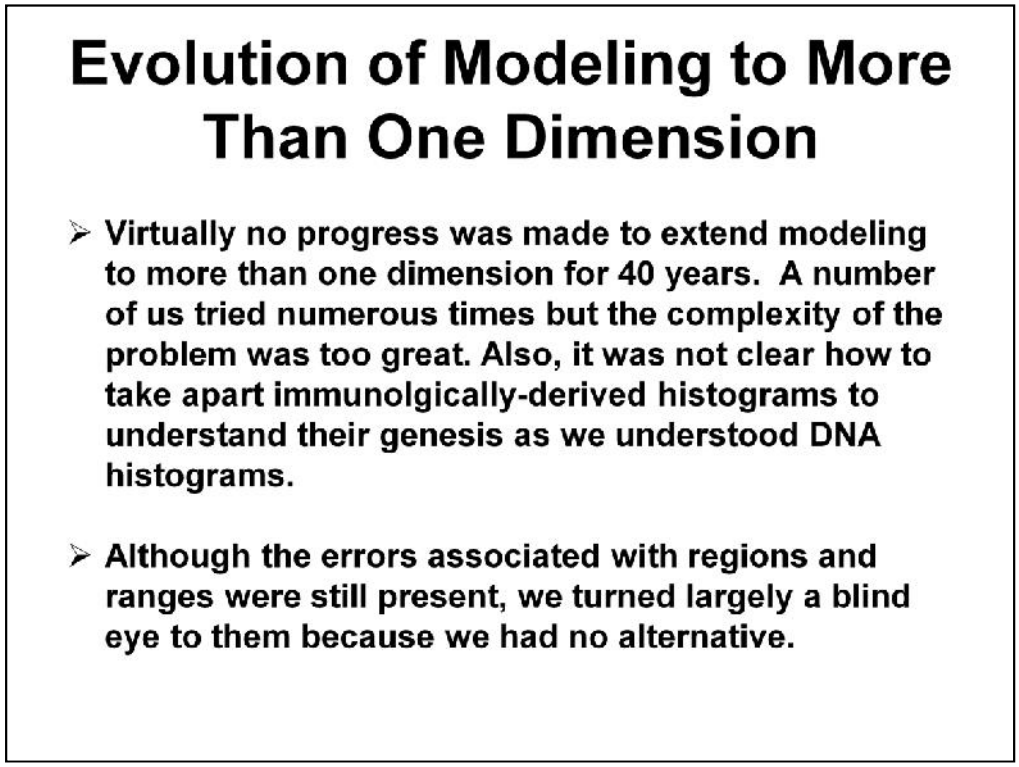
Quite a few DNA models have been proposed over the years and some of them are shown above. Sometime in the middle 1990's the estimation and modeling algorithms reached a level of sophistication where the modeling process became largely automatic no matter what the complexity of the DNA sample. The program MutliCycle evolved from principally Peter Rabinovitches work and ModFit resulted from mine.



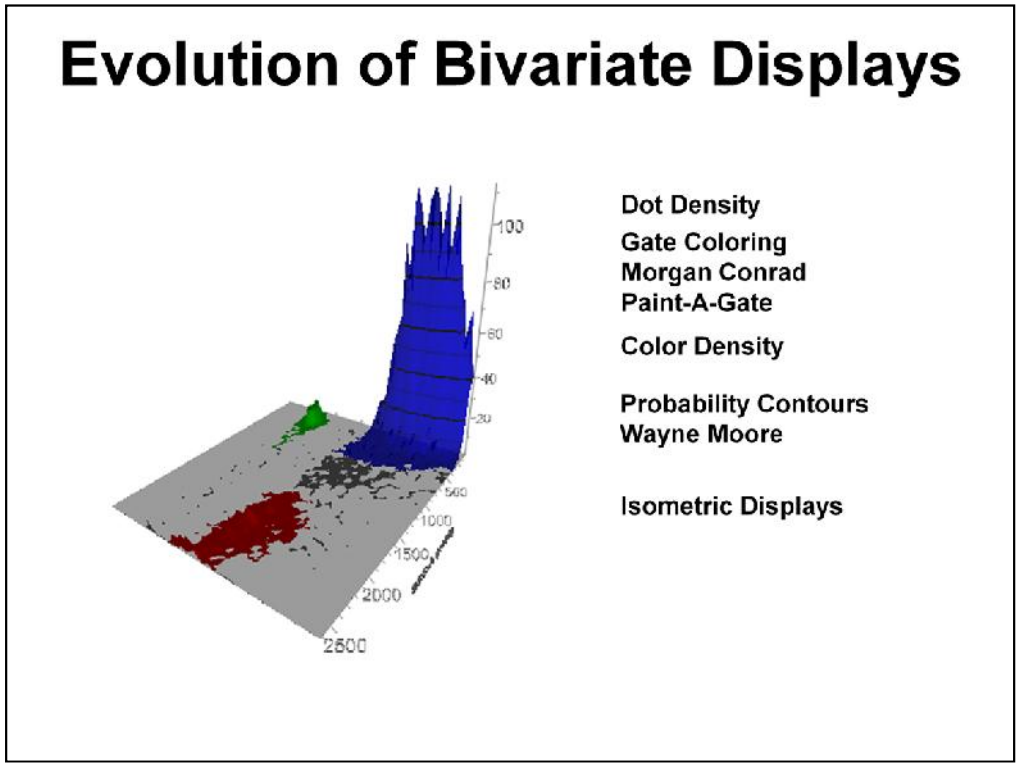
The left panel shows a typical cell cycle analysis with one of the testing data sets and the right panel shows conventional gating analysis of the same data. Cell cycle analyses are now completely automatic and therefore the results from the program are reproducible. Any operator given the same data should receive the same exact results.



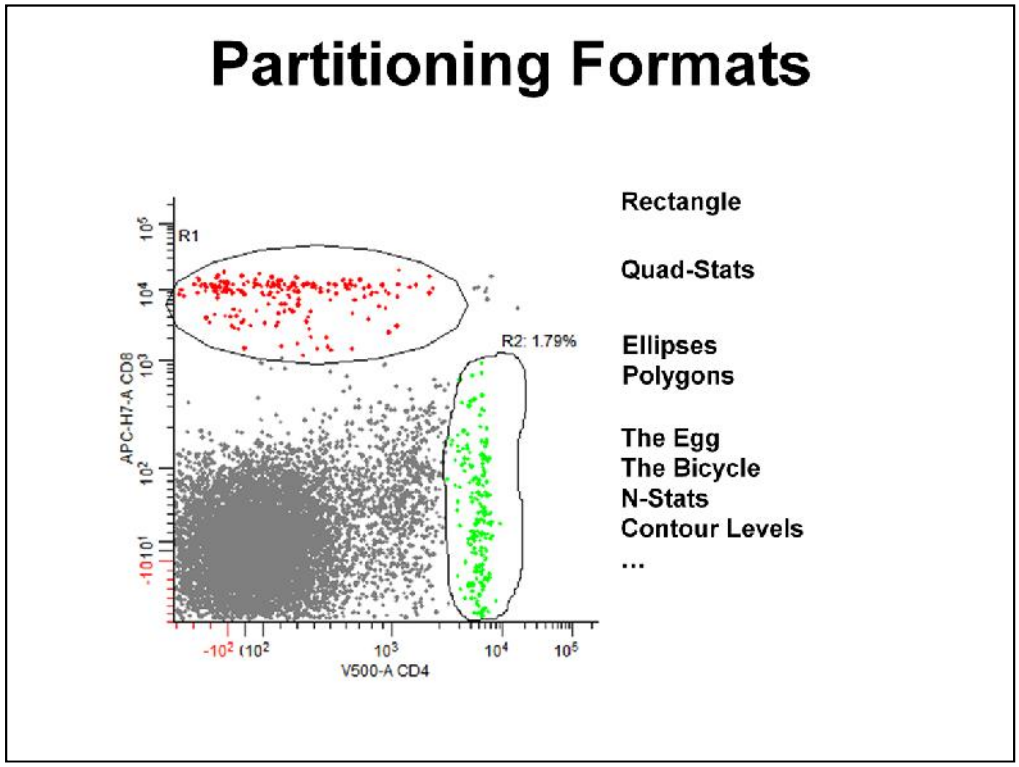
The RC Error's, a method that quantifies error based on chi-square statistics, from simple gating are found to be much greater than a modeling approach. The triangles represent data from the same operator which demonstrates that gating is neither accurate nor reproducible. Because of data like this, in 1993 it was highly suggested that DNA cell cycle phases be obtained from modeling techniques.



Unfortunately, is was very difficult to extend this modeling approach to multiple dimensions. The desire was to be able to model immunofluorescence data, but the complexity of the data and the resultant models was large enough to make computer solutions impractical. Although most cytometrists knew that gating was subjective and error prone, it was largely ignored because there was no viable alternative.

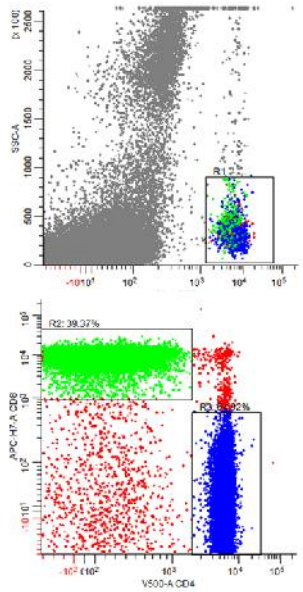


The evolution of cytometric analyses followed several pathways. The graphics associated with represented two correlated measurements evolved from simple dot plots to more sophisticated isometric displays.



The geometric shapes that were used to partition that data into populations evolved as well. It started with simple rectangles and evolved to much more complex polygons.

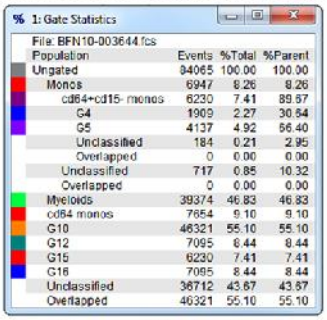
Gating Formats



Simple Gates

Boolean Gates
(R1 and R2 and not R3)
Ortho 2150, Ray Lafebvre,
LYSYS

Hierarchical/Containment Gates



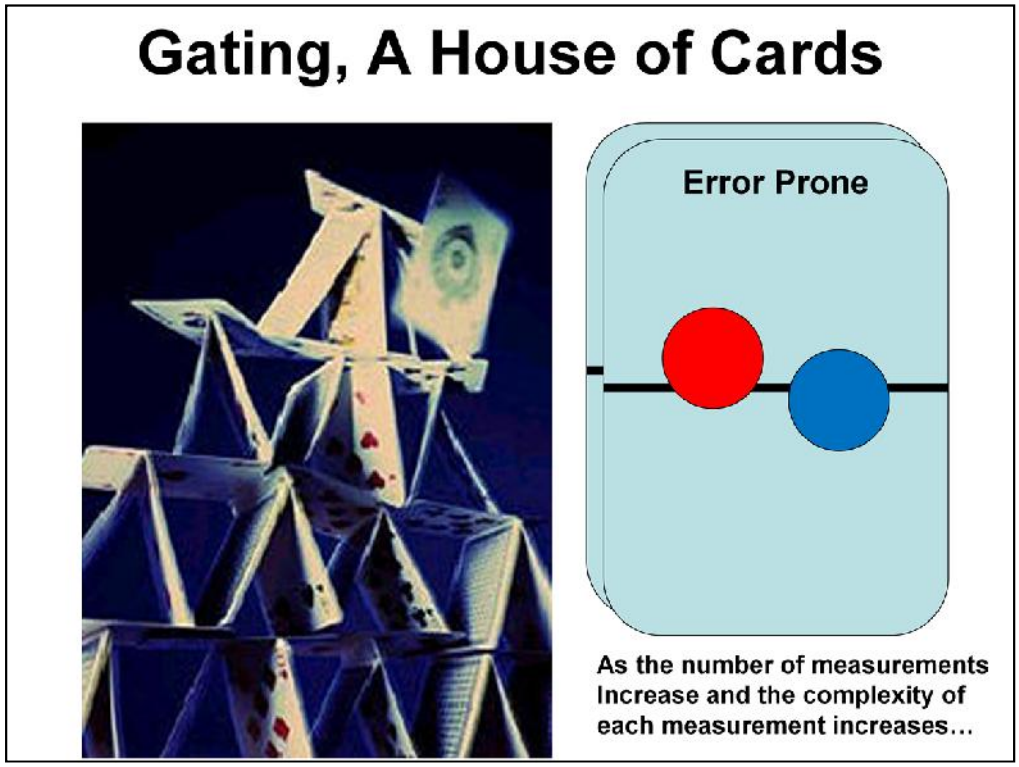
% 1: Gate Statistics

File: BFN10-003644.fcs

Population	Events	%Total	%Parent
Ungated	84065	100.00	100.00
Monos	6947	8.26	8.26
cd64+cd15- monos	6230	7.41	89.67
G4	1909	2.27	30.64
G5	4137	4.92	66.40
Unclassified	184	0.21	2.95
Overlapped	0	0.00	0.00
Unclassified	717	0.85	10.32
Overlapped	0	0.00	0.00
Myeloids	39374	46.83	46.83
cd64 monos	7654	9.10	9.10
G10	46321	55.10	55.10
G12	7095	8.44	8.44
G15	6230	7.41	7.41
G16	7095	8.44	8.44
Unclassified	36712	43.67	43.67
Overlapped	46321	55.10	55.10

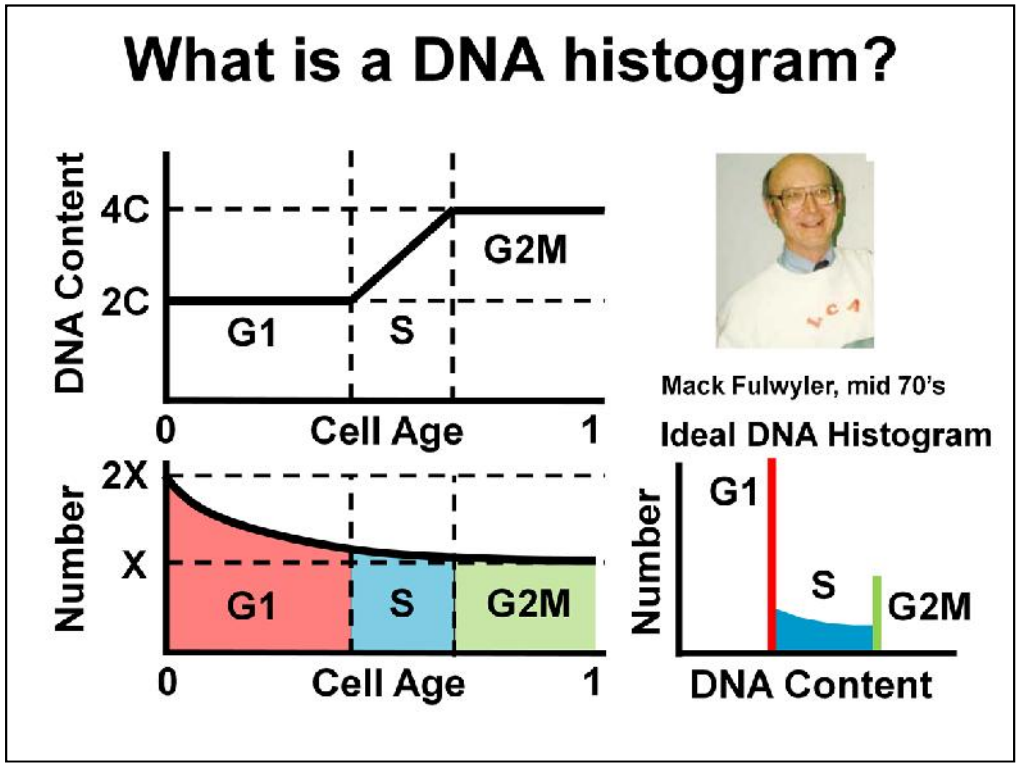
Wayne Moore, Mario Roederer

Initially we used simple gates where a histogram was gated on events that were inside a rectangle. Fairly early on, this process was abstracted in the Ortho 2150 computer system and refined by Ray Lafebvre in his development of LYSYS. Over time it was found that most immunologists wanted to divide their populations into ever smaller populations using hierarchical or containment gates. This work was pioneered by Wayne Moore and Mario Roederer.



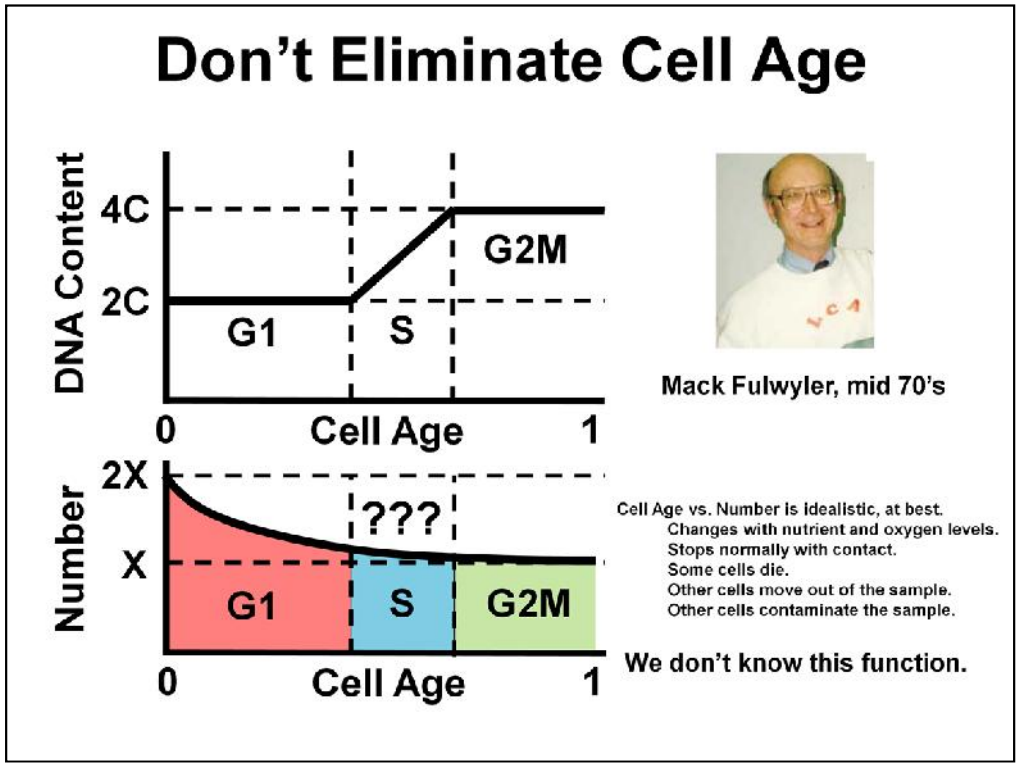
But the reality was that simple boundaries work well if the populations are separated. If the populations had overlap or contained transition events, then this methodology had the same kinds of errors that were unacceptable for DNA histogram analysis. The other issue at play was that as the number of correlated measurements grew, the number of two-dimensional surfaces to draw gates on grew geometrically. It was a house of cards that was about to fall…

Knowing what we now know about modeling complex systems, it would be wonderful to go back in time and change how we initially viewed the modeling process as it pertained to cytometric DNA histograms. We were a lot closer to the correct solution than most people realize.



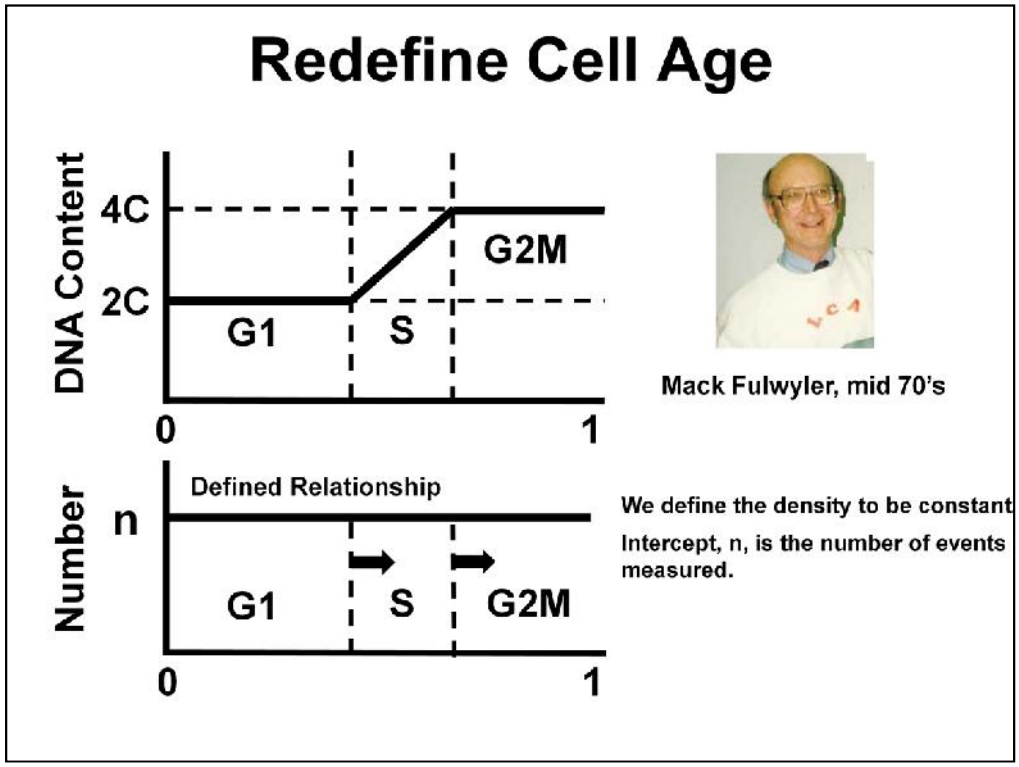
If we could go back in time, I would come up to Mack after he gave his talk on what a DNA histogram is in and talk with him. I would say something like, Mack, I enjoyed your talk, but I think I can show you a much better way of understanding and analyzing DNA histograms. Mack might say something like, “OK, let’s hear it.”

First of all, Mack, if you eliminate Cell Age and do the modeling in measurement space, modeling will essentially be relegated to the one-dimensional world of DNA histogram analysis for over 30 years. So, let’s begin by not eliminating Cell Age and get rid of this concept of the Ideal DNA Histogram. This path doesn’t lead us to where we need to go.



We don't need the concept of an Ideal DNA Histogram to create cell cycle models. The next thing I would tell Mack is that his relationship between Number and Cell Age is idealistic at best. It is much more complicated than a simple exponential. Cells slow down their dividing as nutrients and oxygen become limited. Some cells die and others move away from our sample space; while other types of cells migrate into our sample space. The reality is that since we can't normally measure Cell Age, we don't know the number of events relationship with Cell Age. Essentially, what we wish to do is model a kinetic process but not use time or relative time in our model since we can't measure it. The simplest way for us to proceed is to change our x-axis to something we can measure.

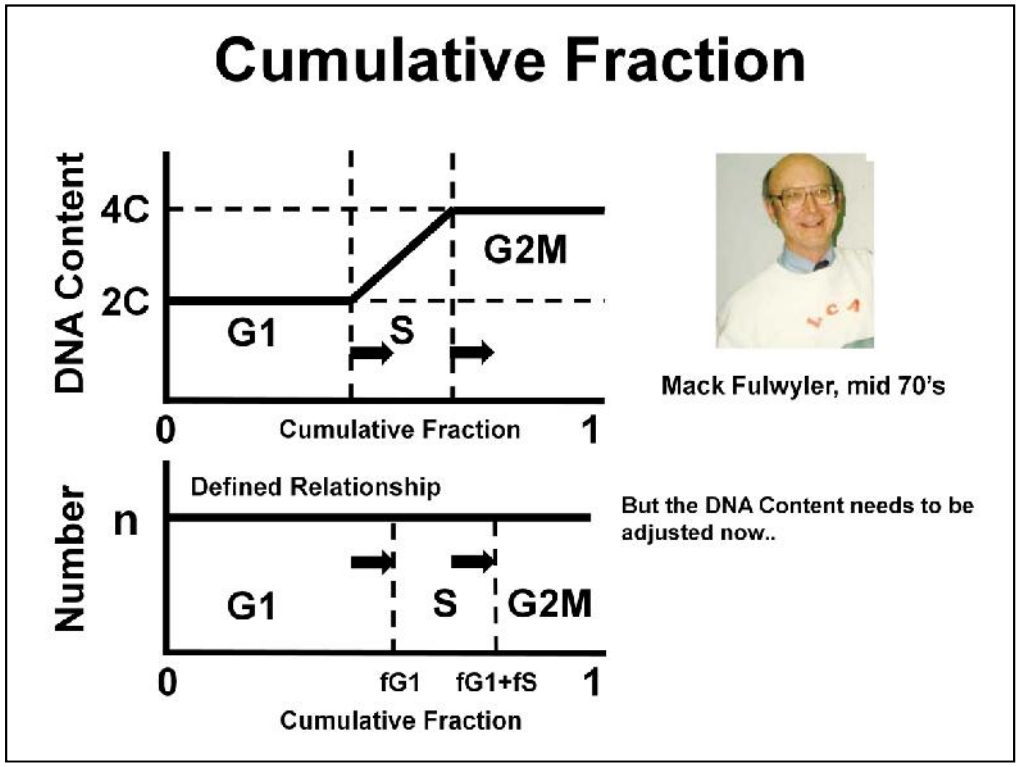
"OK, that makes sense, so what should it be?"



What we can do is define the Number relationship such that it has constant cell density all the way along the x-axis (see above). Since our boundaries are 0 to 1, this means that the y-axis intercept is simply the number of events measured or n.

Mack puts his hand on his chin and says, “well, if you are going to make that definition, then the boundaries between G1 and S and S and G2M will need to change.”

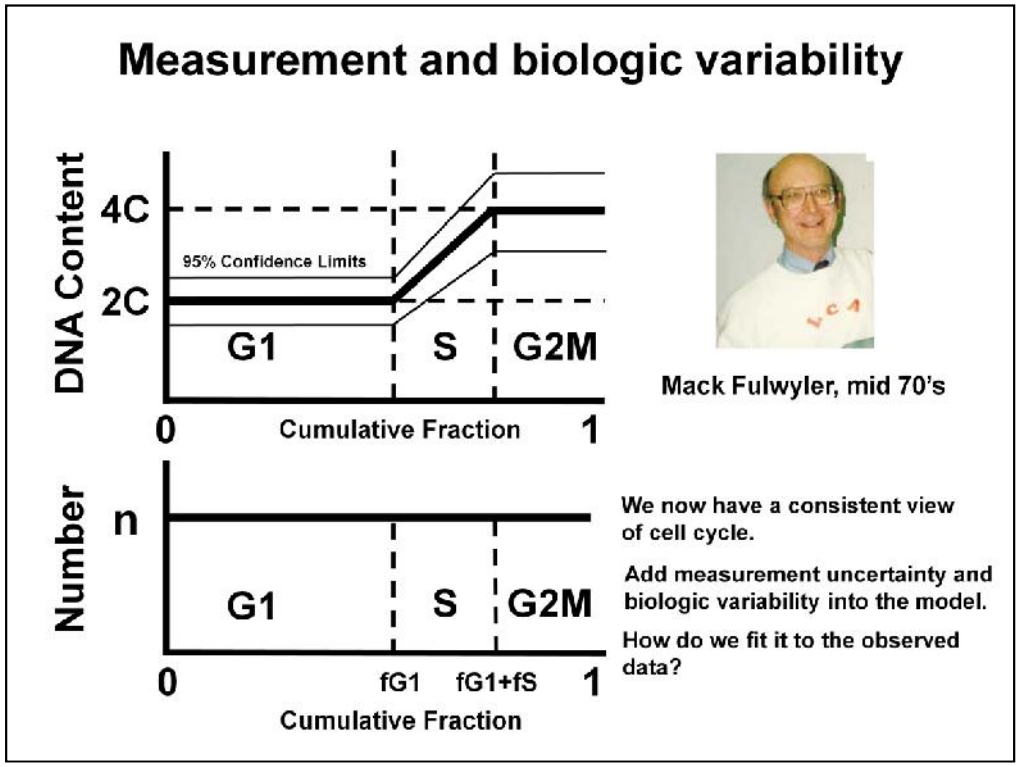
That’s right, Mack, let’s change these boundaries to where ever they need to go to be consistent with our data (arrows).



Notice, Mack, that these boundaries are now at the fraction of G1 (fG1) and the fraction of G1 plus the fraction of S (fG1+fS). In other words, as soon as we defined the constant density relationship, our x-axis changed to cumulative fraction. Just to make this concept clearer, let's use a specific example. If the fraction of G1 in our population were 0.6, then the first G1/S boundary would be placed at 0.6. If the S-phase had a fraction of 0.15, then the second S/G2M boundary would be at 0.6+0.15 or 0.75.

"OK, I see that", Mack says. "But that means that our DNA Content vs. Cumulative Fraction function needs to be changed as well for the model to be totally consistent."

That's right, Mack, let's make that change as well.



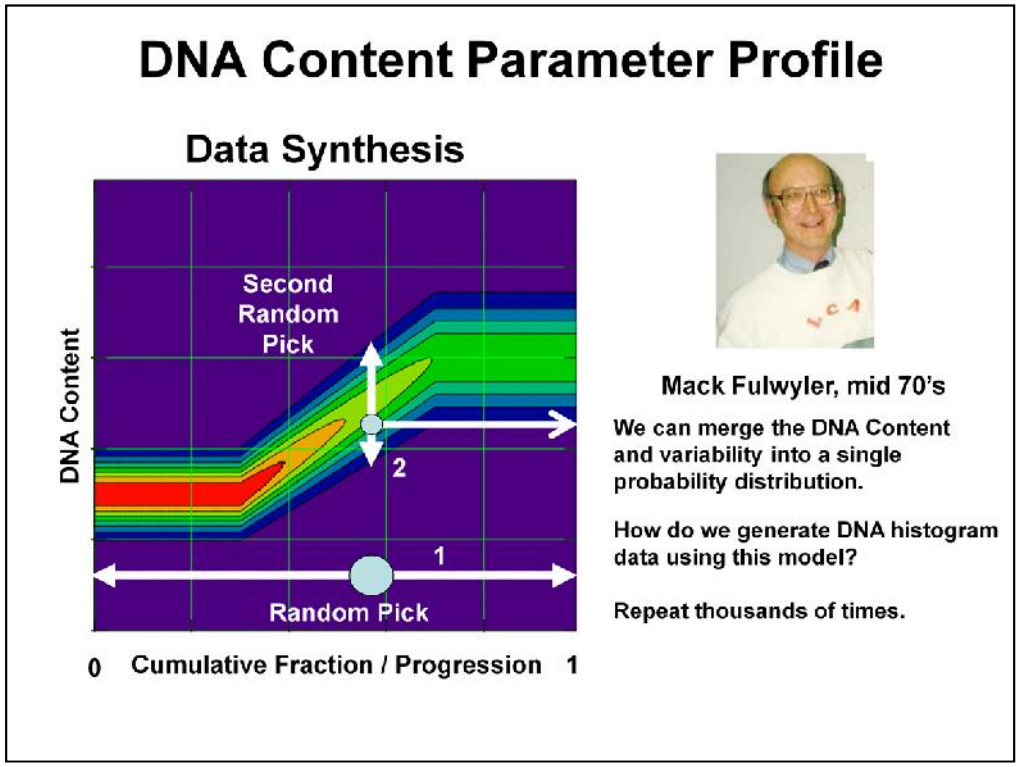
At this point, Mack, we have a consistent model of cell cycle involving two relationships with Cumulative Fraction.

“OK, Bruce, I’m liking this so far. What’s next?”

The next enhancement to your DNA histogram story involves how you introduced the variability of the measurement and biology into the model. The correct way of doing it, Mack, is to integrate that information into this model at this level. We can do this by providing our model with 95% confidence limits. In the case of DNA content measured with linear amplifiers, this means that the width of our 95% confidence limits will increase as DNA content increases.

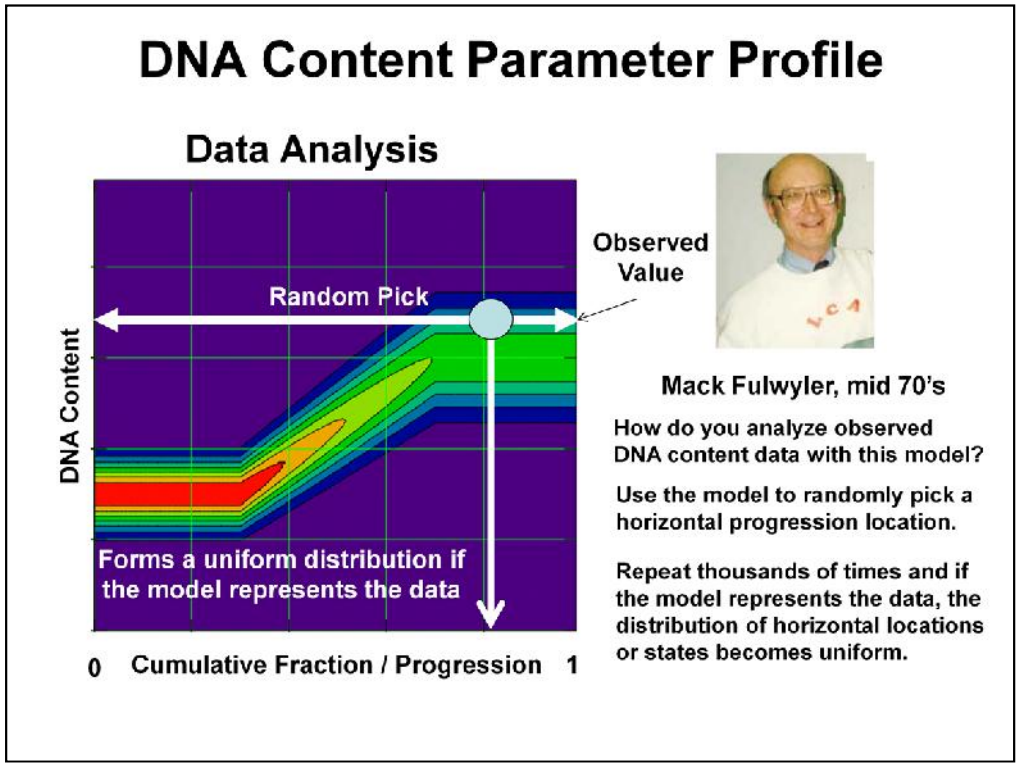
“Bruce, I can see that this model now captures all the essential information that is contained in a DNA histogram, but what is troubling me is how are you going to fit it to the observed data?”

Mack, there are a few more steps we need to take before we’re ready to do this.



The first step is that we can merge the DNA Content and Number axes together by creating a probability distribution from both of them. If we were to look at this function using colored contours, it would look as shown above. This one plot merges cumulative fraction or progression (x-axis), DNA Content (y-axis), and probability (z-axis).

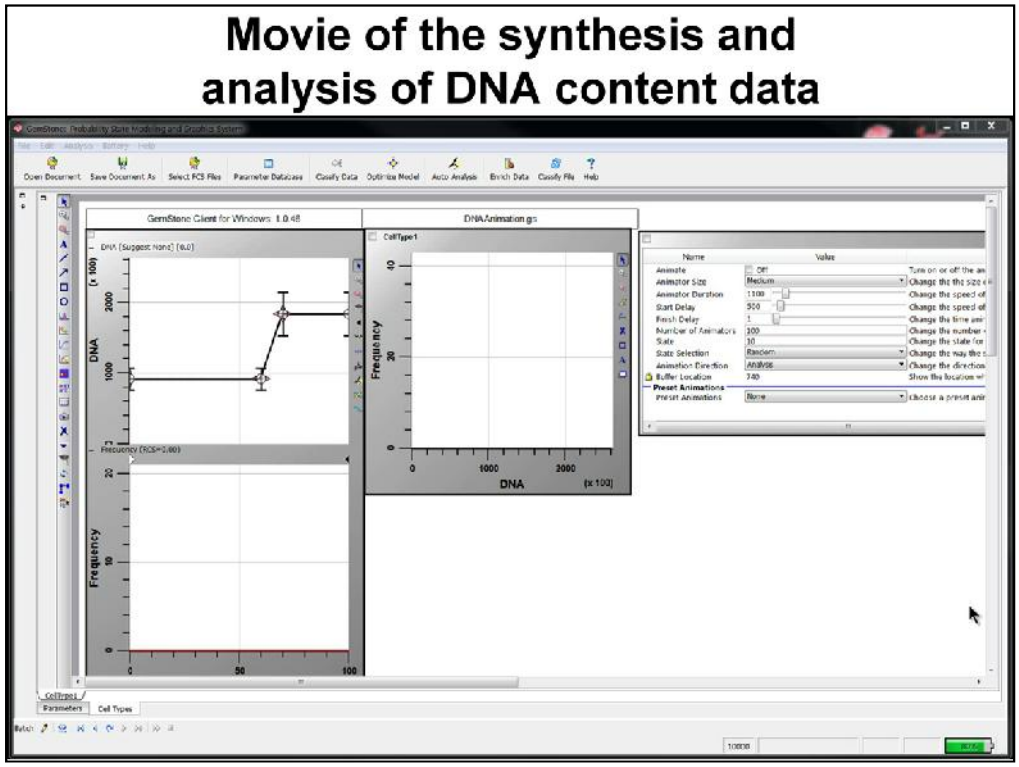
Before examining how to analyze the data, it's easier to first see how this data structure can generate DNA histogram-like data. Models should be able to generate as well as analyze data. The procedure is quite simple. First, we randomly choose a location along the progression axis (1). Given this location, we use the probability distribution at this location to randomly pick our DNA Content (2). This means that we are using the probability information in the vertical direction to appropriately choose the DNA Content value. If we repeat this procedure thousands of times, we produce a set of data that is very much like the data we obtain when we measure cells for their DNA content.



“OK, Bruce, I think I can visualize the synthesis process, but I’m still troubled by how all this allows us to fit our model to an observed DNA histogram? “

It turns out, Mack, that if you reverse this random picking process, you set the stage for analysis. By reversing, I mean that when you observe a single DNA content value (y-axis), you use the above probability distribution to randomly pick a horizontal position (1) along the progression axis. Usually we bin this axis into a set of 100 states. If you repeat this process thousands of times, you will obtain a uniform distribution of state frequencies if the model is consistent with the observed data. This is not an approximation, Mack. If the model and the data are consistent with each other, the resultant state frequencies will be as uniform as possible given counting error.

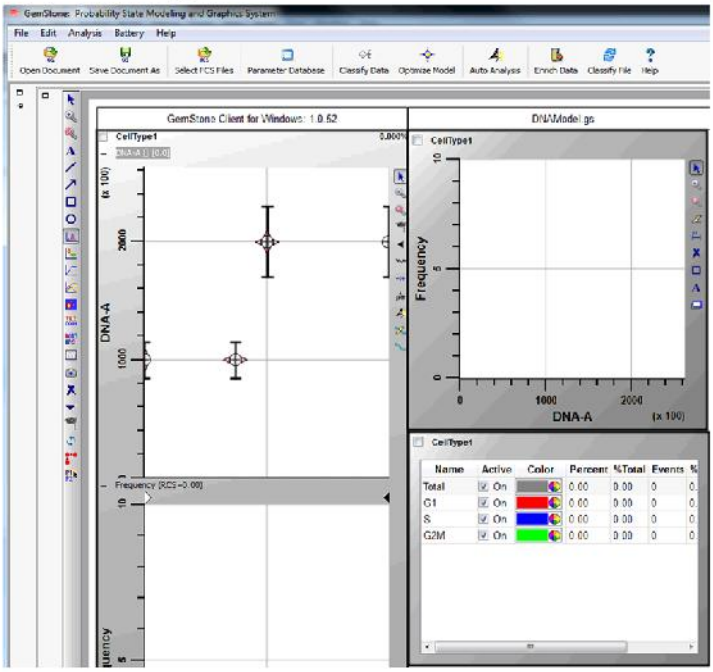
We can quantify how uniform this distribution is use this as a model response function to search for the best model parameters. That’s how the system can fit observed data to our new model. We call this approach Probability State Modeling or PSM.

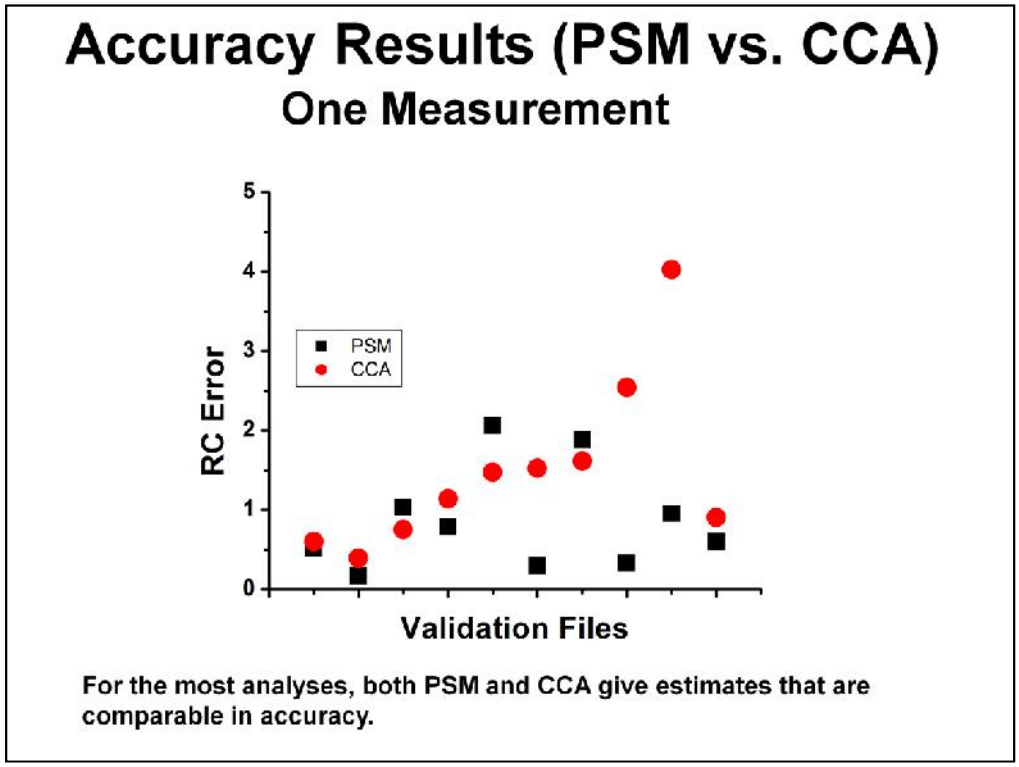


Mack scratches his head and says, “It all seems to make sense, but it would be extremely helpful to see this process in action. ”

It just so happens, Mack, that I brought a movie of both the synthesis and analysis of data to help visualize this process.

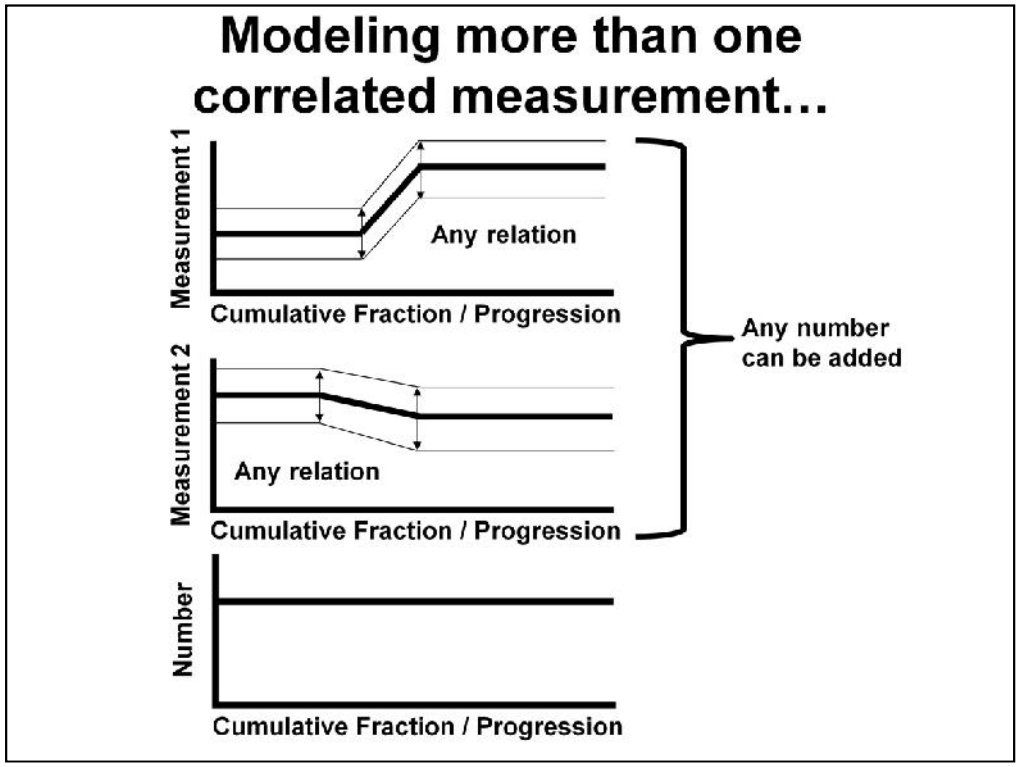
DNA Synthesis and Fitting





"I see how this works now, Bruce. Just as a point of curiosity, how accurate is it in estimating %G1, %S, and %G2M as compared to the more traditional modeling methods?"

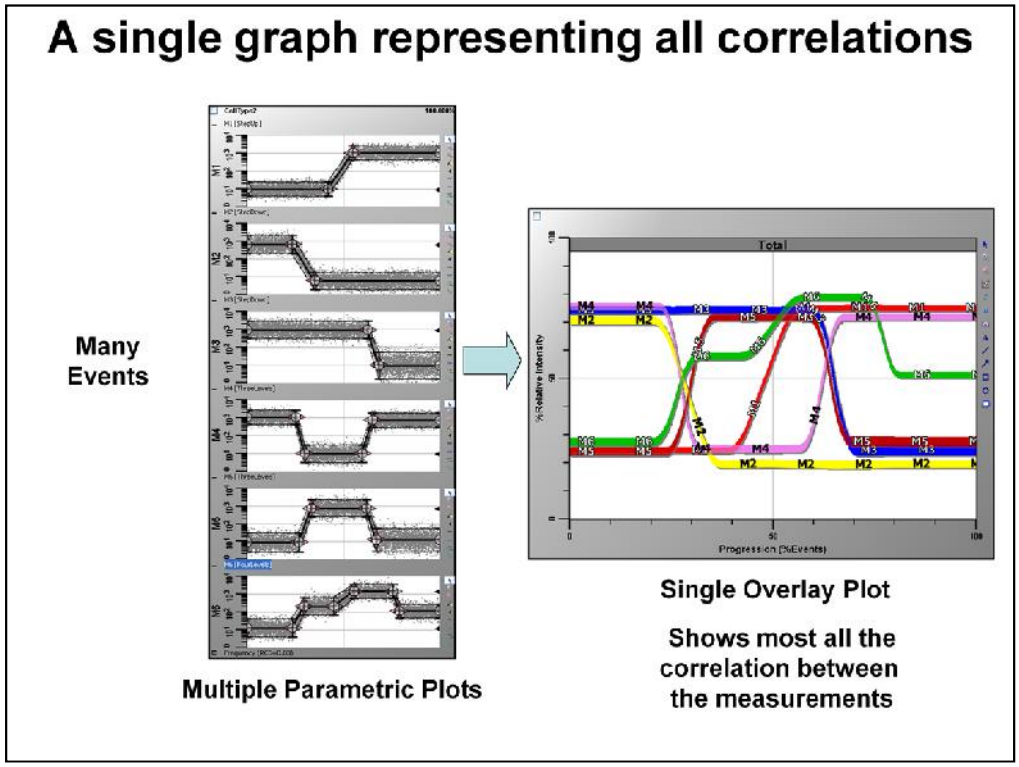
If we compare the computed errors using chi-square statistics for Probability State Modeling (PSM) and Cell Cycle Analysis (CCA), we find that in general both methods produce analysis results that are quite comparable. PSM does have a few advantages over CCA, however. With traditional modeling, if you have two model components of similar shape that are overlapped, CCA will become unstable and yield unreliable results; whereas, PSM simply assigns each model component equal numbers of events and is perfectly stable.



"I assume, Bruce, that you're showing me this new approach because it does something that traditional modeling methods can not."

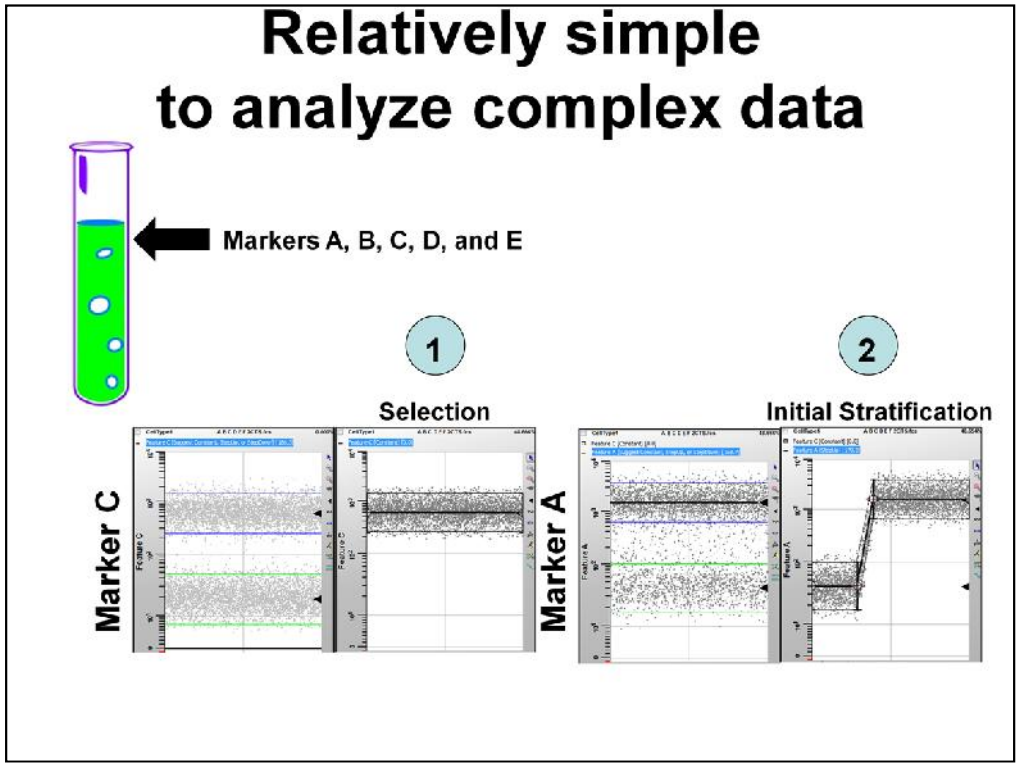
That's right, Mack. We are now in a position to model data with numerous correlated measurements. By defining our model as we have, we can potentially model any number of markers that modulate over some progression. This is something we were never able to do before.

We can theoretically add any number of markers we want to our model of progression. The first marker could be DNA, but it could be any other relation as well. It's also very easy to extend the algorithms to either synthesize data from all these parameter profiles or to analyze them.



“How do you represent all this information graphically after you’ve modeled the data?”

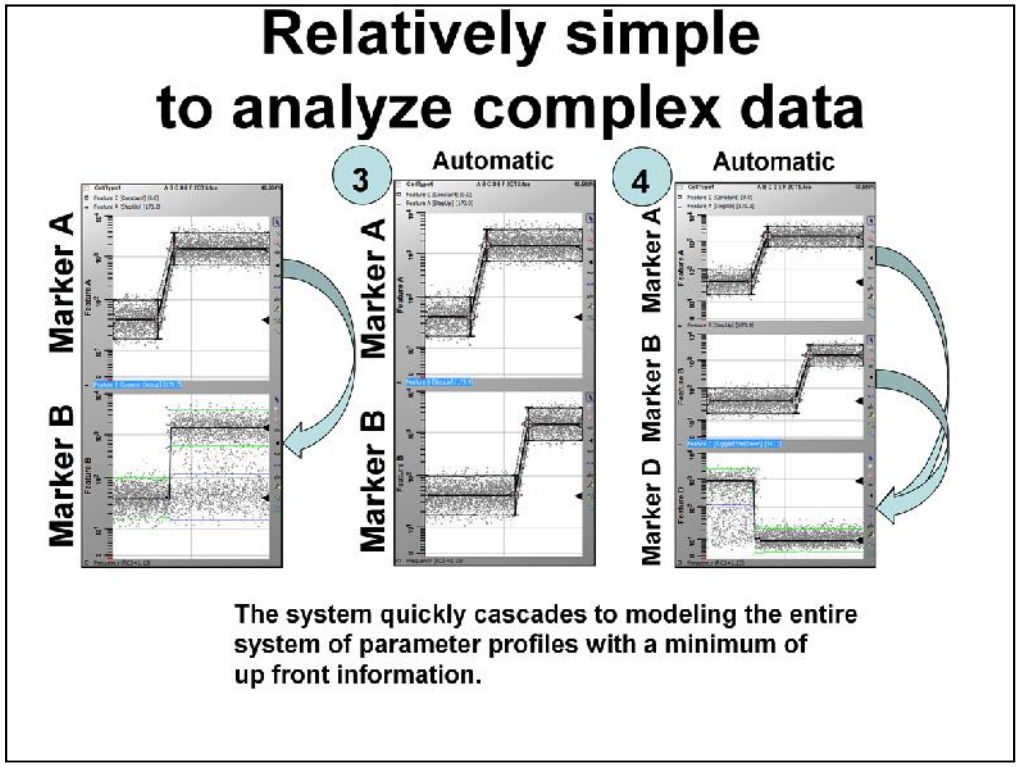
After many events are processed in this manner, we end up with tracks of events in each parameter profile plot as shown above. Each one of these tracks has statistically determined 95% confidence limits for each state, forming a data envelop. These data envelopes can be scaled-down and co-plotted on an overlay as shown to the right where the common vertical axis is %Relative Intensity and the common horizontal axis is %Cumulative Fraction or Progression. This one graph summarizes all the correlations between the measurements as well as all the important stage statistics.



"Just a few more questions, Bruce. It seems to me that it should be very complicated to figure out the proper parameter profile to use for each measurement. For DNA, it was easy since this is a well-known relationship to most biologists. But some of the new markers are not so well-known. How do you handle this complexity issue?"

It ends up that it is far easier to model these complex populations than you might think. Mack, the way to get at this answer is to use a hypothetical example and see how to apply the technique of probability state modeling to it. Let's say we have some five-color data that looks at five markers (A, B, C, D, and E) for two hypothetical populations. To get started we need to know that the population of interest for us has a lot of C marker on its surface (C+). We refer to this as our selection marker. To tell the system we're interested in C+ we use a constant parameter profile that selects for the events of interest (1).

We then choose a relation that we do know something about. In this case, suppose we know that Marker A is up-regulated in our progression much like our DNA content example and we model it as we did our DNA (2).



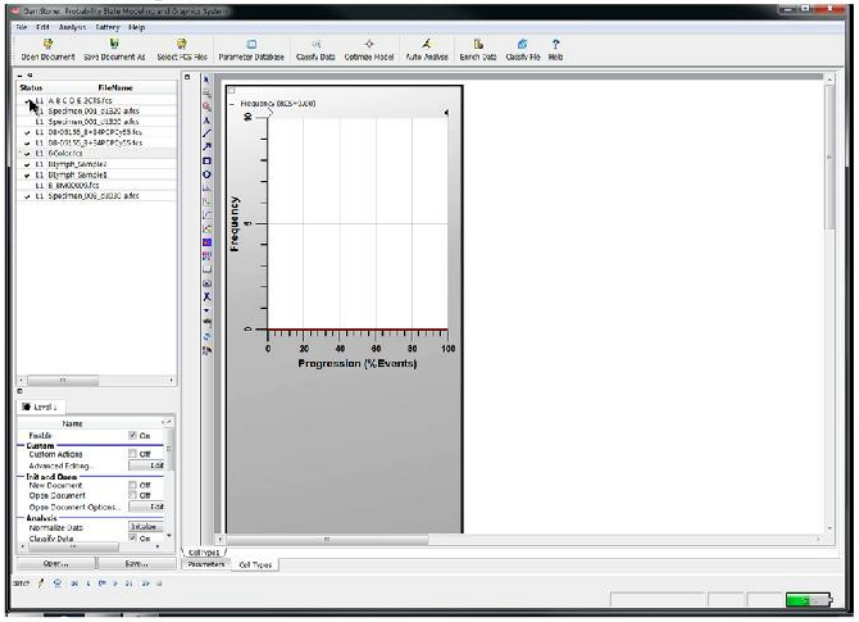
What happens next is really quite amazing. Once the system has modeled Marker A, it has partially ordered the events along the x-axis. This means that when we look at Marker B, we can usually tell how it varies with A without any *a priori* knowledge. The decision of how to model Marker B can be entirely automatic (3).

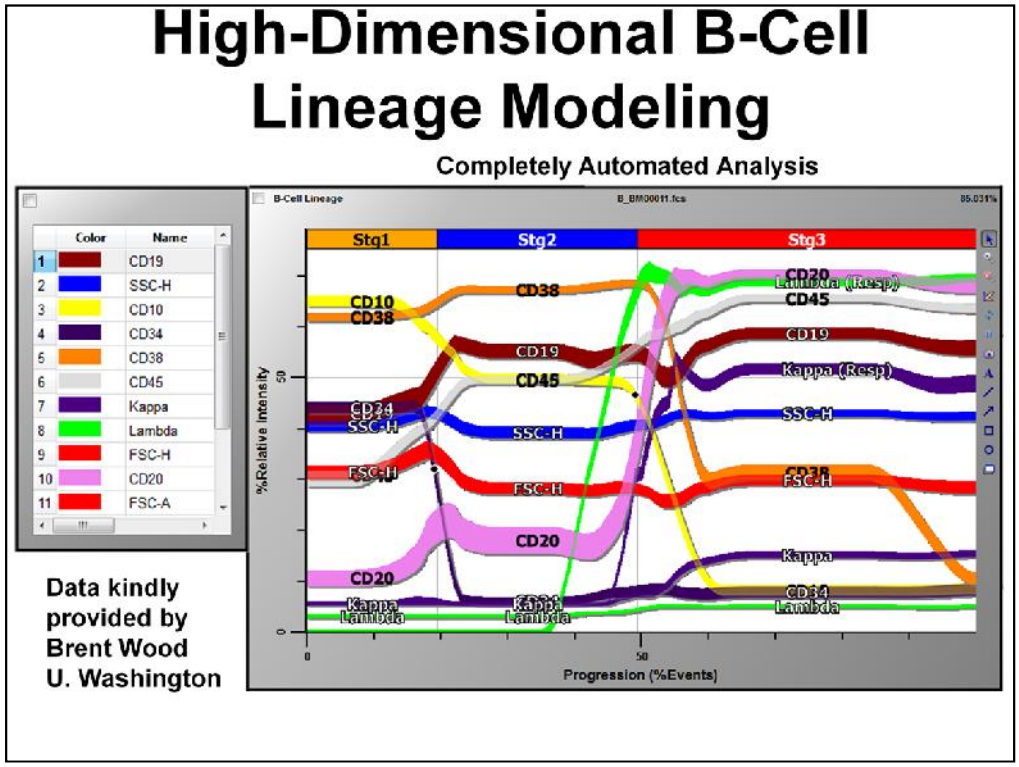
After it has finished with B, it now can use both the A and B information to better define D. An appropriate parameter profile can then be chosen for D completely automatically. This process continues until all the markers have been modeled.

“You don’t have a movie that shows this process in action do you? “

Of course I do…

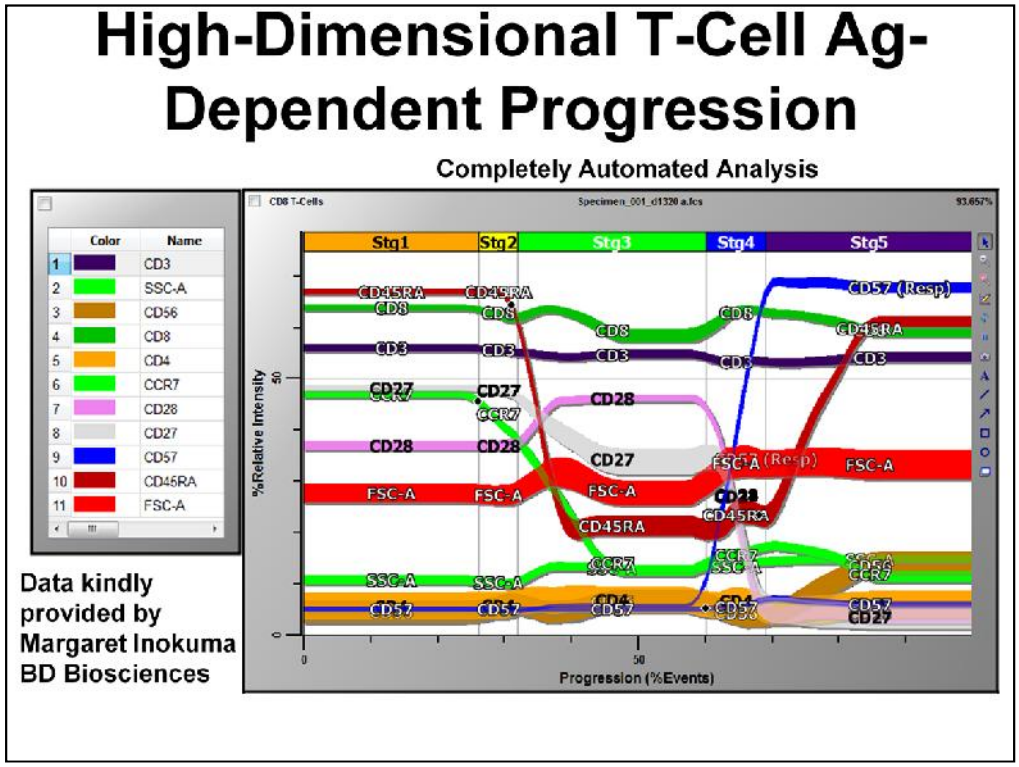
Movie of a simple example of fitting multiple correlated measurements



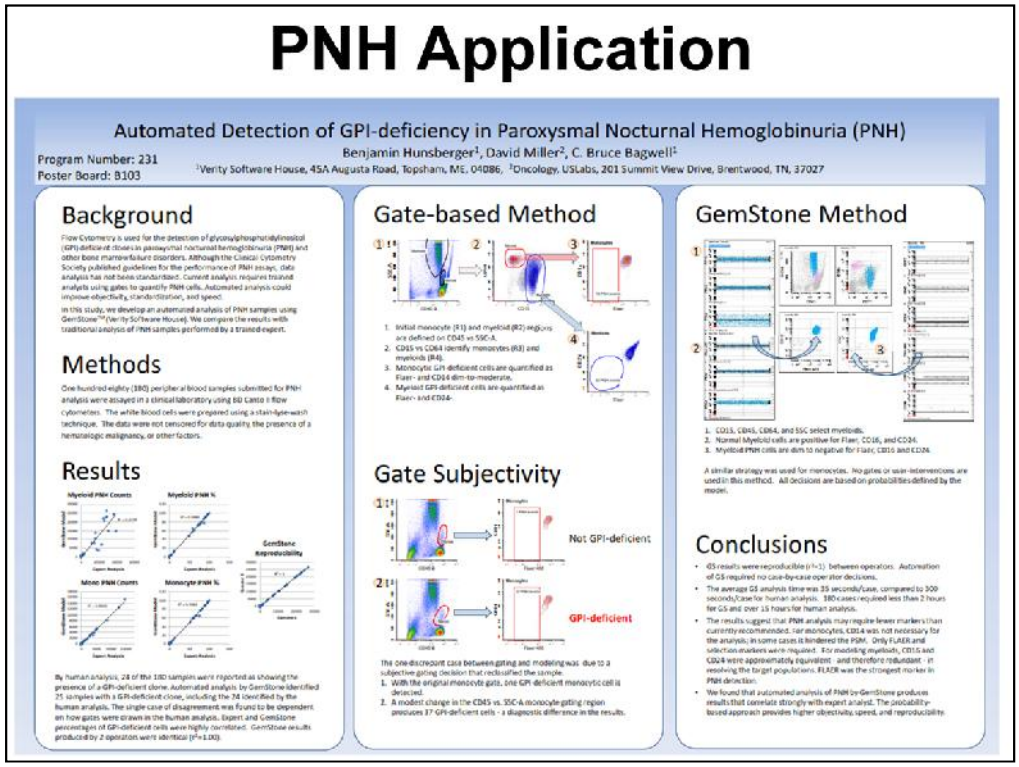


"OK, Bruce, let's see some actual applications with some of the instruments capable of taking numerous correlated measurements."

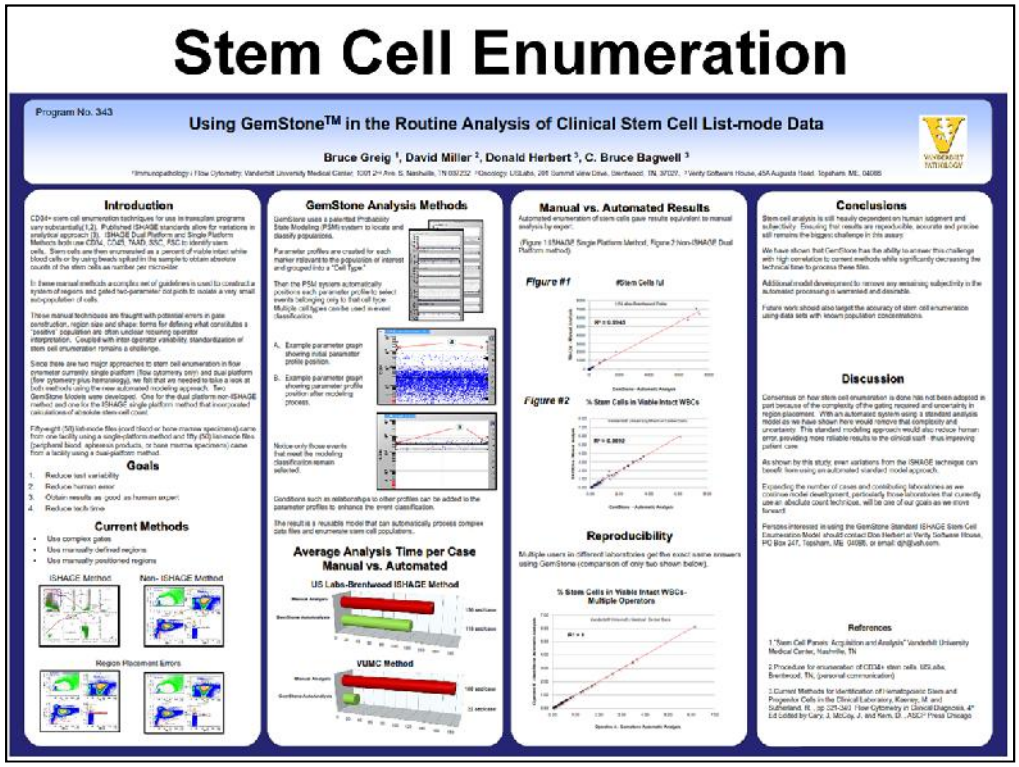
The first application we applied this technique to was modeling the B-cell lineage in human bone marrow. When examining normal B-cell lineage progression, it is quite impressive to see the sharp transitions that B-cells go through in the bone marrow. When the dimensionality issue is eliminated, it is easier to appreciate the genetic program at work.



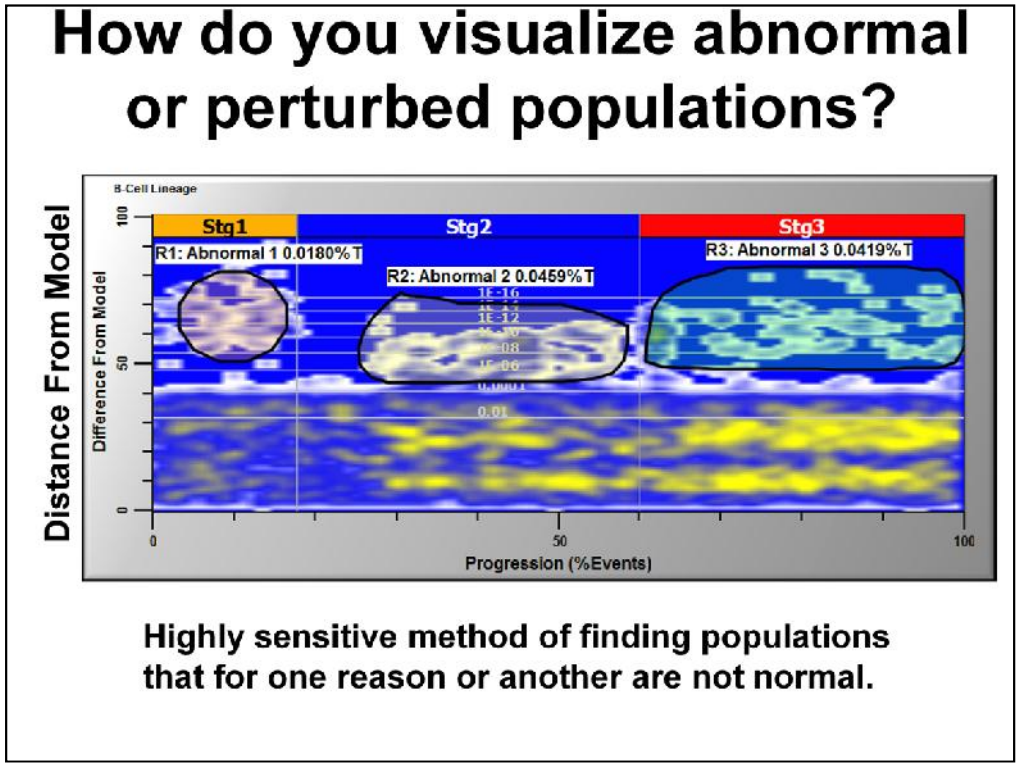
Another application is the changes in CD8 T-cells as they interact with specific antigens in the peripheral blood.



This slide shows a PNH abstract and poster shown at CYTO 2011 by Ben Hunsberger, demonstrating how PSM can automate this widely ordered test. Interestingly, these models have no progressions at all.



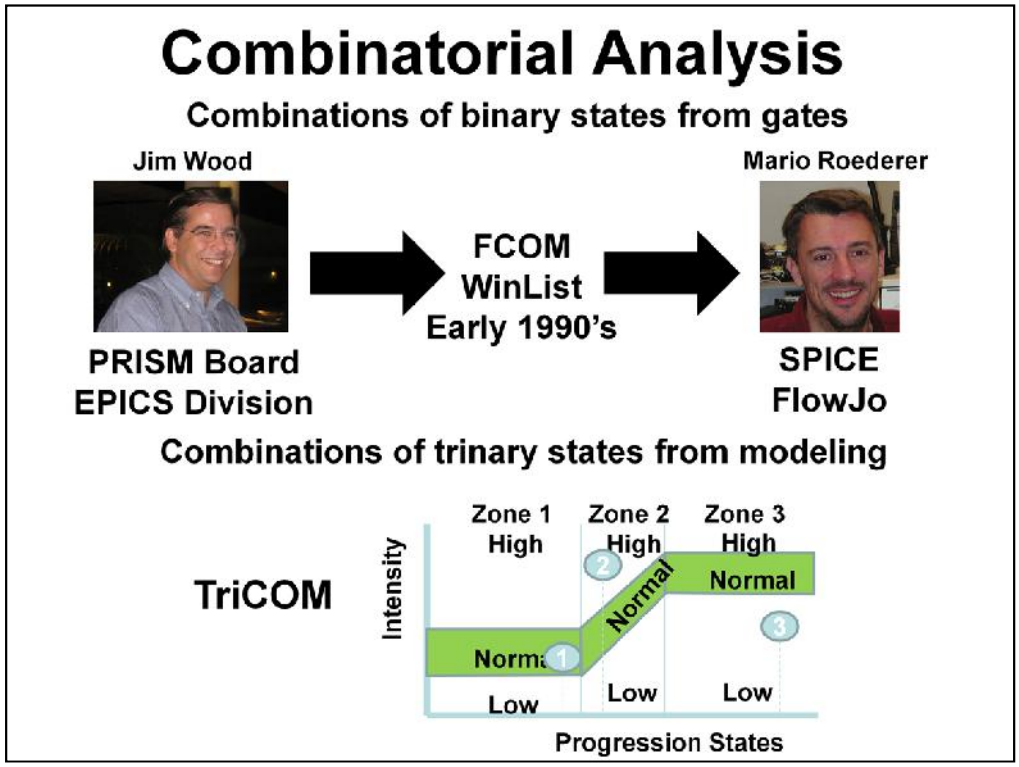
Stem cell enumeration, CYTO2011 abstract and poster by Don Herbert, is another example of the PSM automation capabilities where the models have no progressions.



"I can see that using PSM allows the modeling of any number of markers in progressions, but it is many times valuable to see perturbations from normal such as T-cell activation or B-cell malignancies. How do you visualize these processes?"

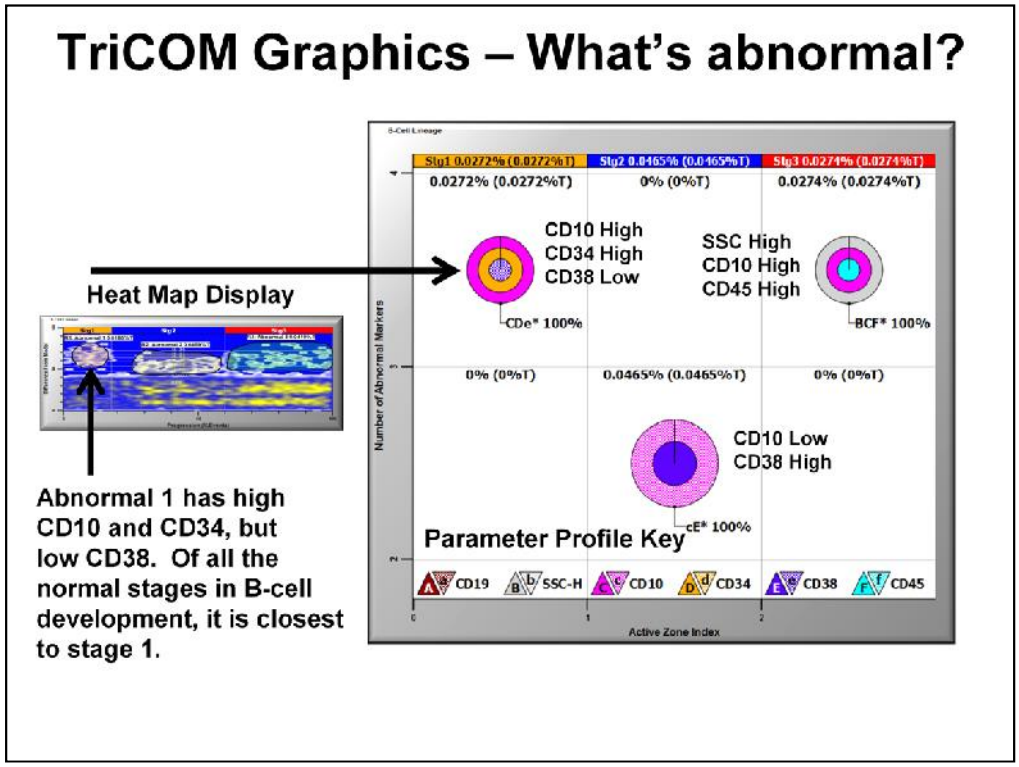
In 2010 we showed that one can visualize populations that somehow don't belong to the normal model by the use of a specially designed chi-square heat map. Essentially, an event's distance from a model can be quantified by means of sum of chi-squares. This distance along with its position along the progression axes defines the populations in the heat map display. Since the events that are within the 95% confidence limits are subtracted in the heat map, this display has the capability of finding very rare events. The three abnormal populations shown above have frequencies in the 0.02 to 0.05% range.

Of course, the next question that we want to know is what is exactly abnormal with these populations.

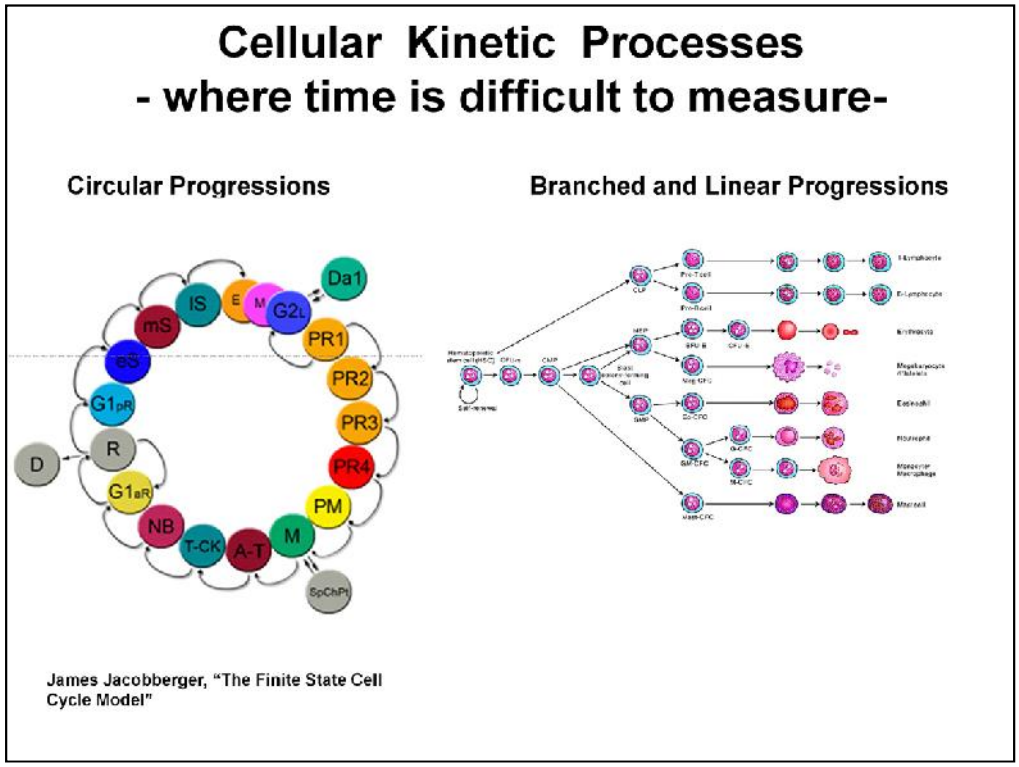


The first comprehensive method of finding all phenotype combinations was the PRISM board designed by Jim Wood and the EPICS Division team. Later, in the 1990's, this type of logic was generalized to include the results of boolean gates with WinList's FCOM logic. More recently, Mario Roederer has used the same kind of logic and generalized it to include patient group averaging and the use of pie charts to graphically represent all the phenotypes.

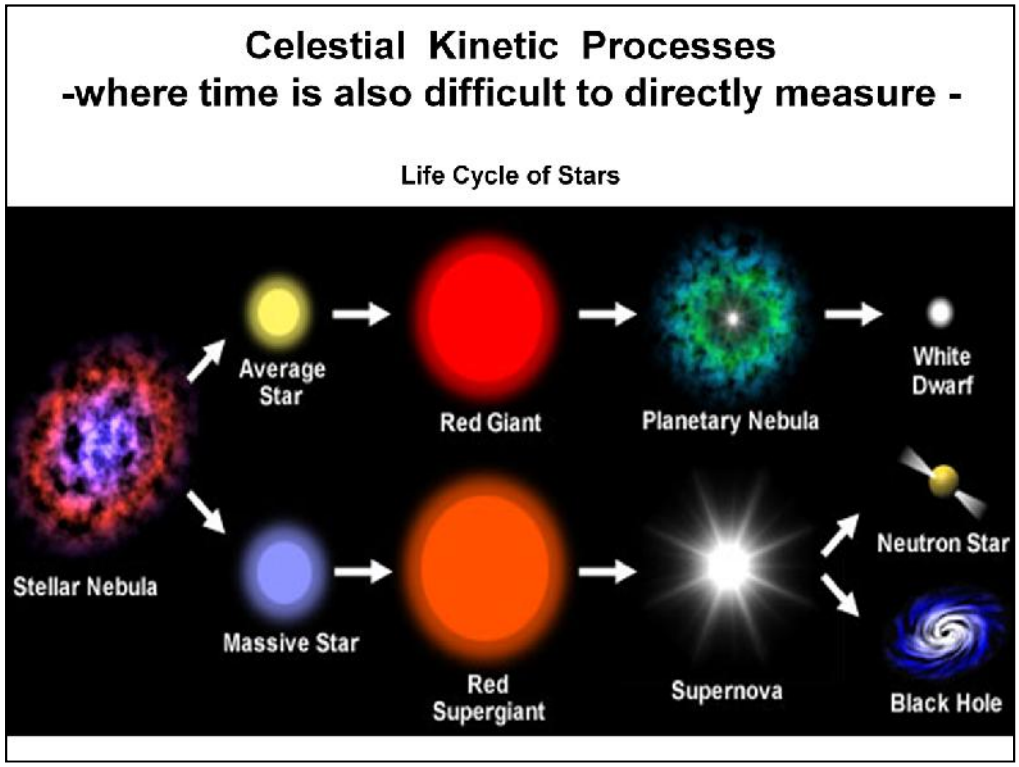
Modeling of progressions offers the opportunity to extend the concept gates (in or out) to three states (normal, lower-than-normal, and higher-than-normal). Also, these calculations can be applied within specific stages of the progressions referred to as zones (see above slide, bottom). Probably most important, however, is that since it is a result of modeling, all boundaries are positioned automatically with no human bias.



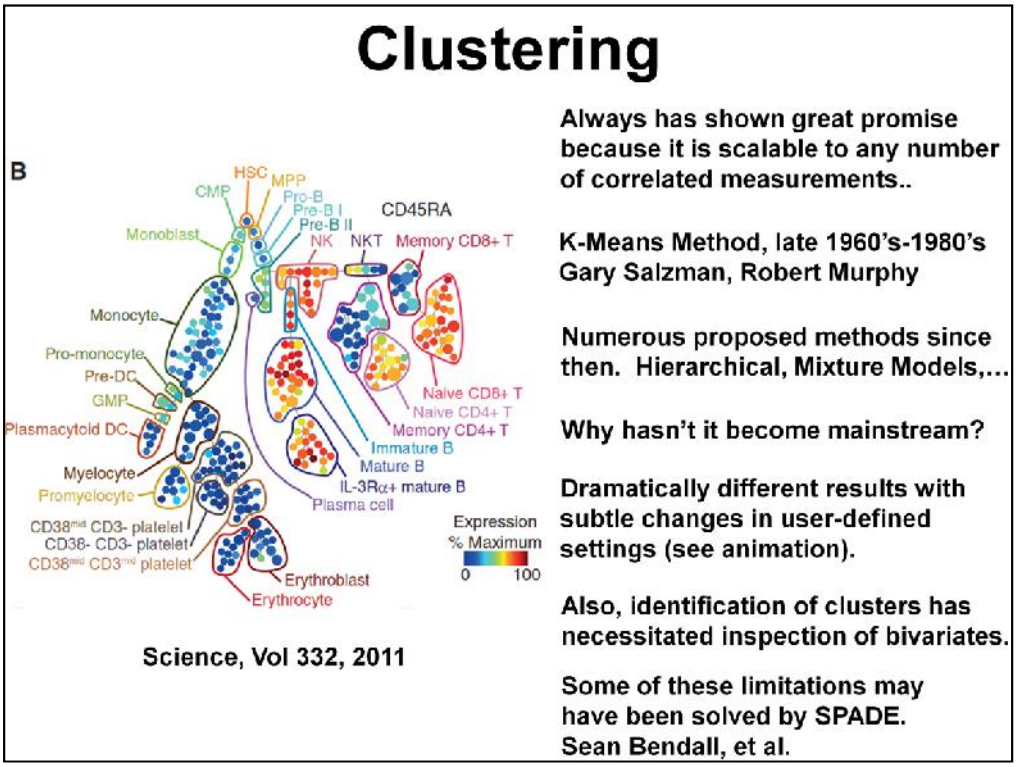
The Heat Map display shows us that there are abnormal populations present, but what are their phenotypes? The TriCOM display systems shows all the phenotypes by using multiple ring pie charts. The TriCOM x-axis depicts the stages of the modeled progression. In this B-cell example there are three stages. The y-axis quantifies the number of abnormal marker expressions in each phenotype. The key at the bottom shows how to interpret the phenotypes of the abnormal populations. For example, solid blue (E) represents higher than normal CD38; whereas, low density blue (e) represents lower than normal CD38. Reading from the outside ring to the center, the top-left pie chart shows that there is an abnormal population in Stg1 that is high for CD10 and CD34, but low for CD38. The other two abnormal populations are detected as well. The percentages above the pie charts quantify the abnormal phenotypes for B-cells and total cells. Multiple phenotypes at the same stage and number of abnormal markers are represented as slices in the pie (not shown above). This one graph allows the inspection of all abnormal phenotypes that may be present in a sample.



In our particular technology, we can use PSM to learn more about how cells do what they do. We can apply it to circular progressions such as the cell cycle (left) or linear progressions shown at the right. You can even model branched progressions now.



Any time it is difficult or impossible to measure time and you want to model the kinetic process, probability state modeling can be used. It works best when there are many measured objects with correlated measurements.



I should make a few comments about clustering before leaving this lecture. Clustering has always shown great promise since it scales very well with number of correlated measurements. Initially Gary Salzman and Robert Murphy explored its potential in cytometry starting late in the 1960's and published a number of papers and chapters through out the 1970's and 1980's. Since then, there have been numerous different strategies proposed for clustering, but it really has never become mainstream. I suspect the reason for this is three-fold. The first issue is that clustering algorithms many times divide well-known populations into multiple parts and also many times fails to find low-frequency continuums that connect the clusters. The second issue is that clustering generally has a number of user-defined parameters that can radically change the final solution. The animation shows an example of how the boundaries around clusters can suddenly and dramatically change. The third issue is more insidious. After the clustering algorithm has identified the clusters, the next question is do these cluster represent. Since most clustering algorithms are devoid of biologic information, the user or biologist must use traditional methods to identify the found clusters. If clustering is suppose to be a solution to the dimensionality problem, requiring the inspection of bivariates to understand the meaning of the clusters is not really a solution that will work for very high-dimensional data.

Recently, a new clustering method has been devised by Sean Bendall at Stanford that may obviate some of these limitations. Since it creates numerous micro-clusters via a process called agglomeration, it may not have a tendency to eliminate the continuums between populations that we normally use to understand the meaning of the data. Also, it looks as though it has some biologic constraints so that the cluster interrelationships are immediately evident by someone with some knowledge of the biologic process being

studied. However, it is new and it is not yet known how sensitive it is to subjective user decisions and how automated it can become over the next few years.

Summary

- When we decided to model DNA histograms directly in the early 1970's, modeling was too complicated to be applied to more than one correlated measurement and thus was mainly relegated to DNA histogram analysis.
- As a consequence of the early limitations of modeling, gating evolved considerably in cytometry even though it is inherently inaccurate and subjective.
- If we use cumulative fraction instead of Cell Age, we can construct a mathematical model that embodies both DNA content changes and variability as a function of cell cycle progression. This type of model is called a Probability State Model or PSM.
- It is very easy to extend PSM to any number of dimensions. Once the data is modeled, all correlations of measurements in progressions can be summarized in an easy-to-read graph.
- Once we have modeled what it means to be normal, it is possible through a chi-square heat map or TriCOM to visualize populations that are somehow not normal.

The Verity team made the creation of Probability State Modeling possible. Thanks guys!