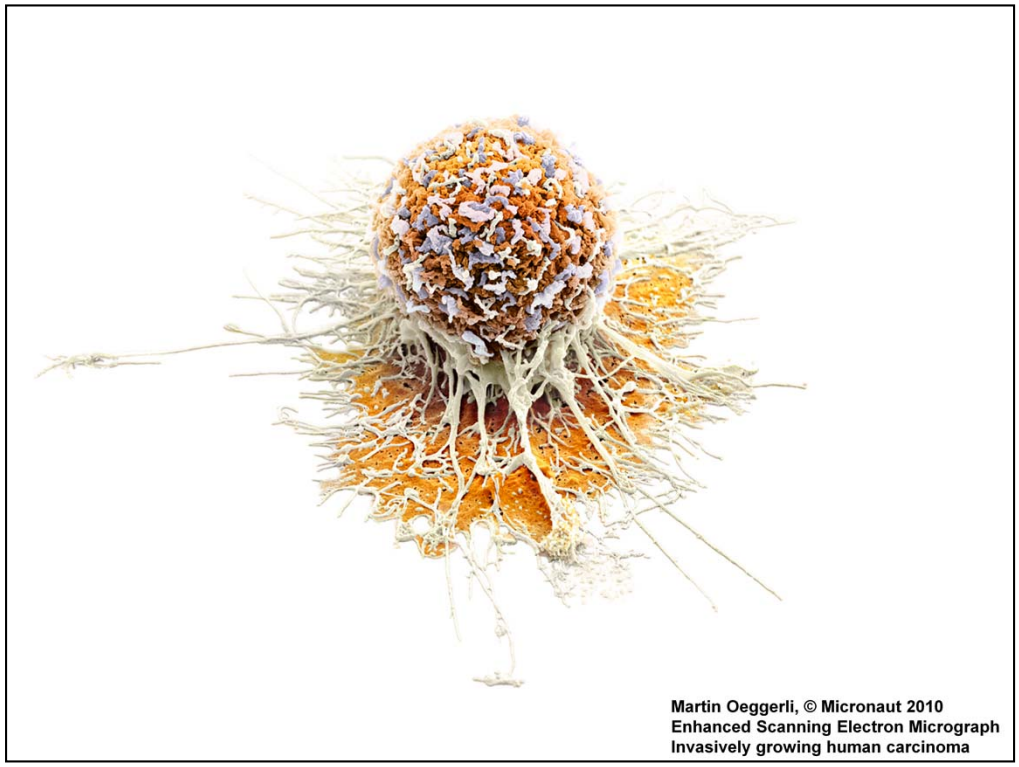
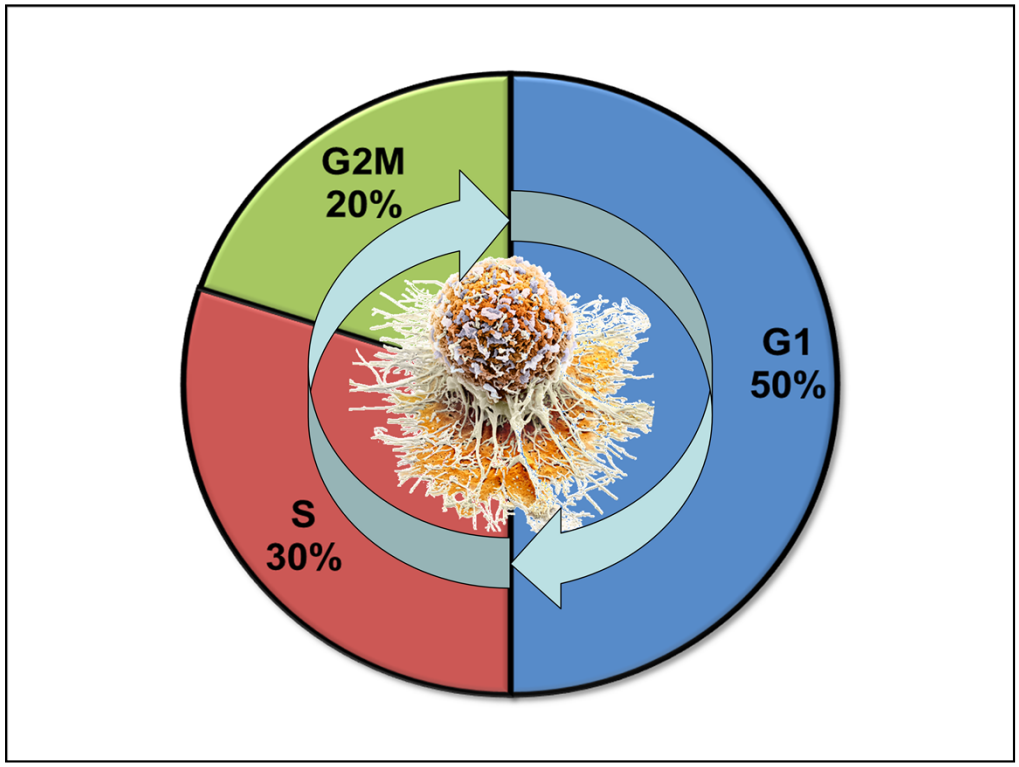
Probability State Modeling of DNA Content S Phases Using Parabolic Splines

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

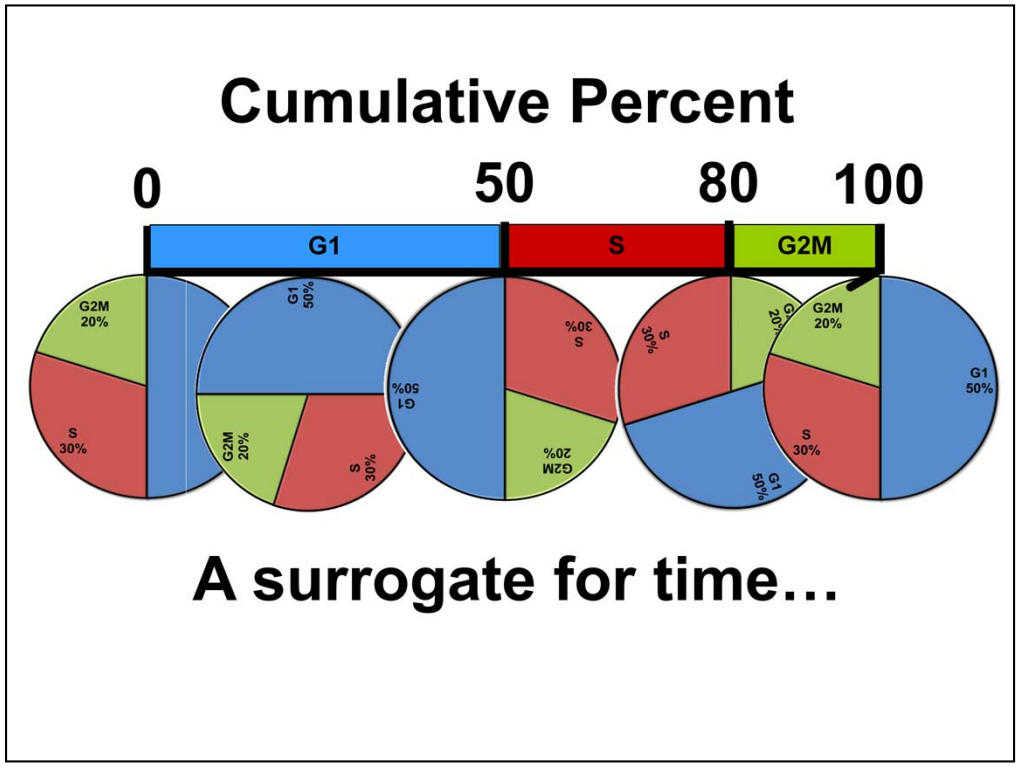
In this talk I will be talking about improving the accuracy of S phase estimation from cytometric data containing DNA content. A new method of interpolation, parabolic splines (PS), for Probability State Modeling is presented that yields extremely accurate S phase estimates.



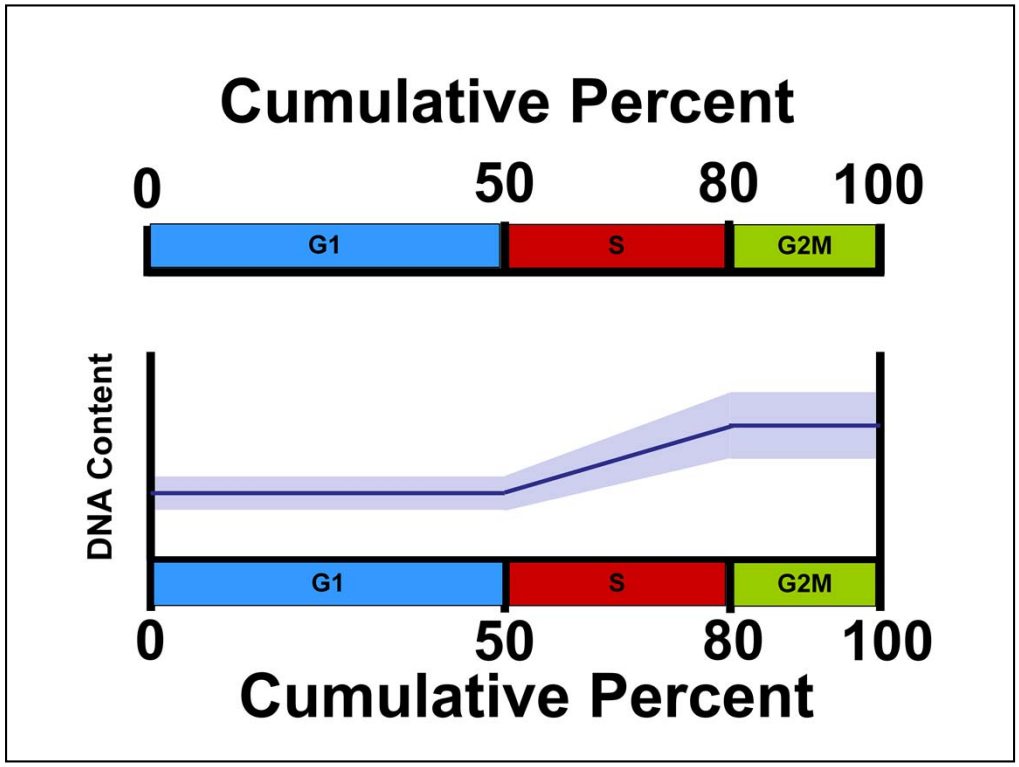
Cells are wonderfully complex chemical machines and modeling them to reveal their hidden secrets can indeed be a challenging prospect. The secret to understanding and modeling complex systems like cells is to first understand the simplest possible system and then add a little complexity one step at a time. One of the most basic attributes of a living system is its ability to divide. Cells proceed through the process of division in basic steps or stages.



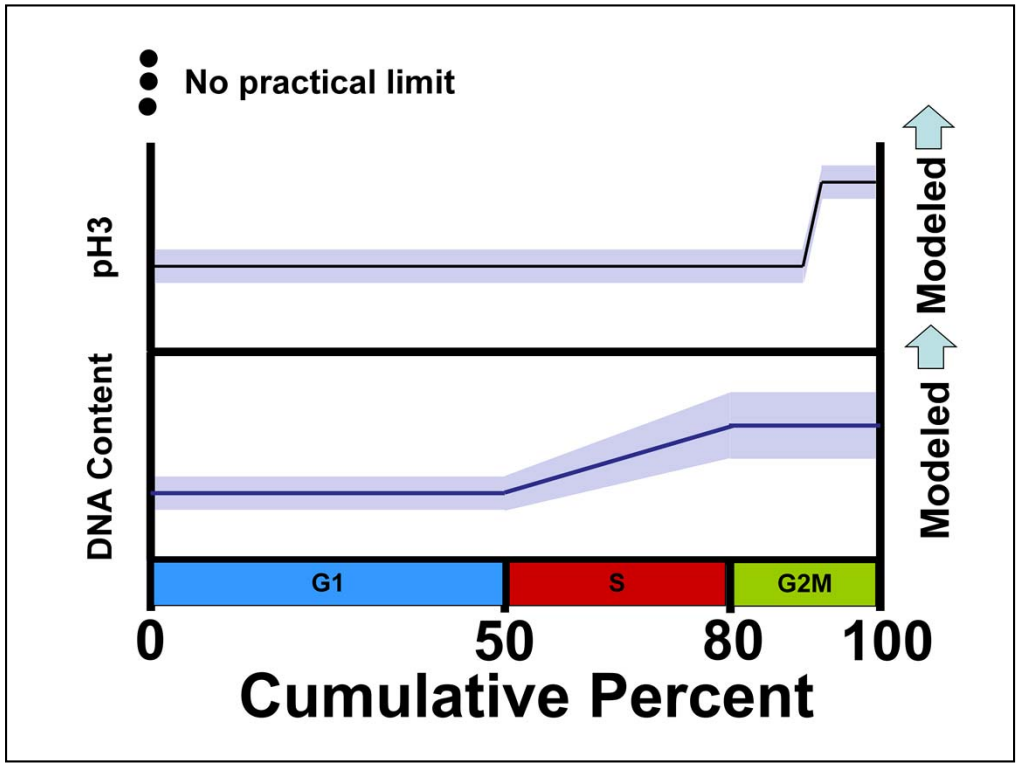
The cell cycle for a population of cells can be crudely separated into G1 (gap1), S (synthesis), and G2M (gap2 and mitosis). In the above slide, 50% of the cells are in G1, 30% in S, and 20% in G2M. The cells are moving clockwise as they go through the process of division. If we wanted to represent this progression of stages as a single line (or vector since it has direction), all we need do is roll the perimeter off the pie and demarcate the percentages as we go.



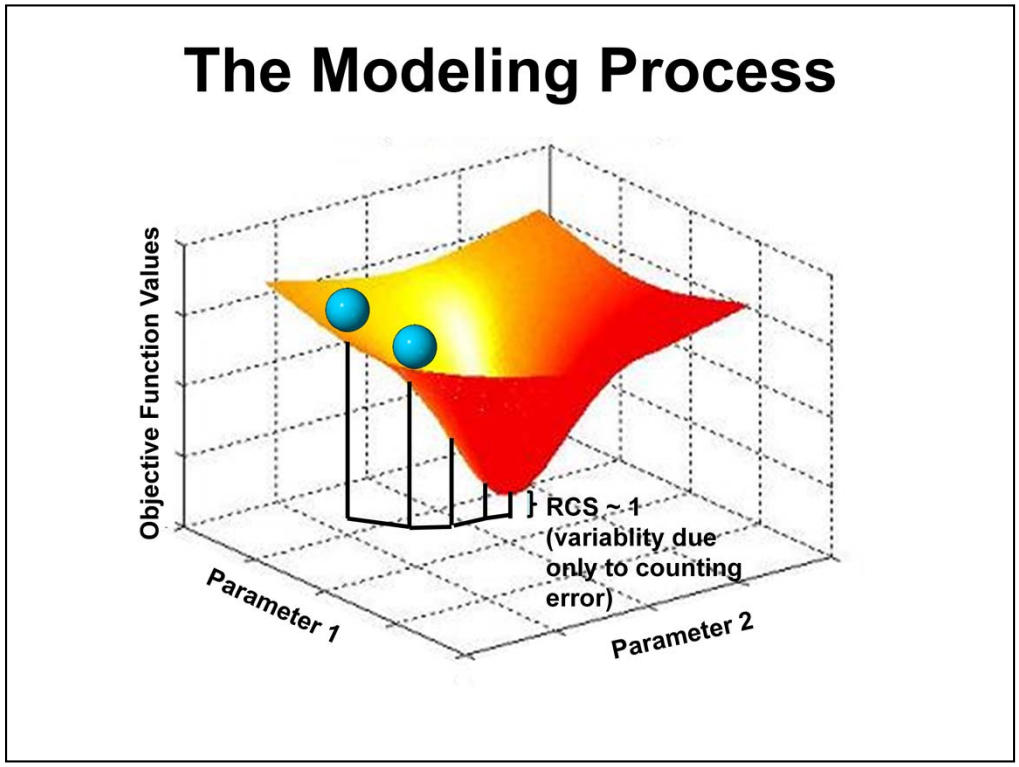
If we roll the cell cycle pie along a line, we create an axis that has direction and represents cumulative percent. In Probability State Modeling this axis is used as a surrogate for time; however, it can also serve as a means of quantifying n-dimensional mixtures (more on that later).



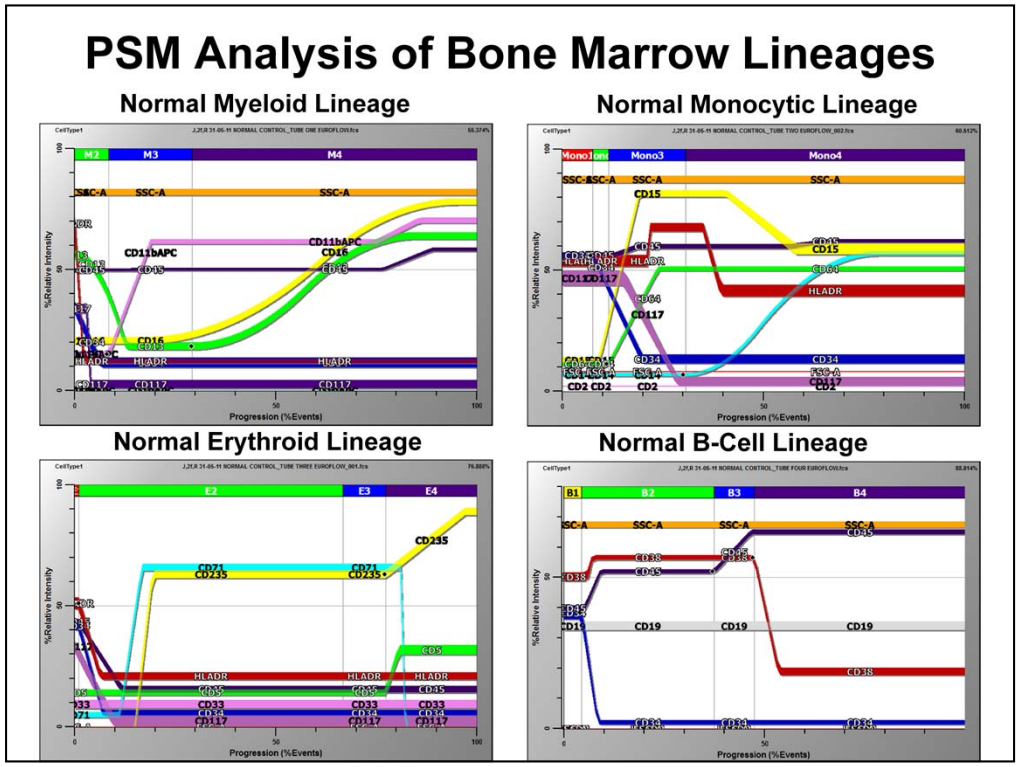
Probability state modeling uses cumulative percent as a common axis to investigate changes in measurements like DNA content as a function of progression through the cell cycle. These relationships are referred to as parameter profiles. The y-axis is the measurement intensity and the x-axis is our surrogate for time or cumulative percent. The parameter profile also defines the uncertainty and heterogeneity in the measurement or line-spread as a function of cumulative percent.



The power of probability state modeling (PSM) is that these measurement relationships with cumulative percent can be stacked with no practical limit. Each measurement adds more correlative information to the process being studied. For example, pH3 shows where the mitotic phase begins which was not evident with DNA content alone. Probability state modeling can automatically model the DNA Content parameter profile and then model pH3 and continue in this step-wise manner until all parameter profiles are modeled. There usually is an optimal sequence of parameter profiles to model. Usually one starts with what is simple and known and works towards what is more complicated and unknown.

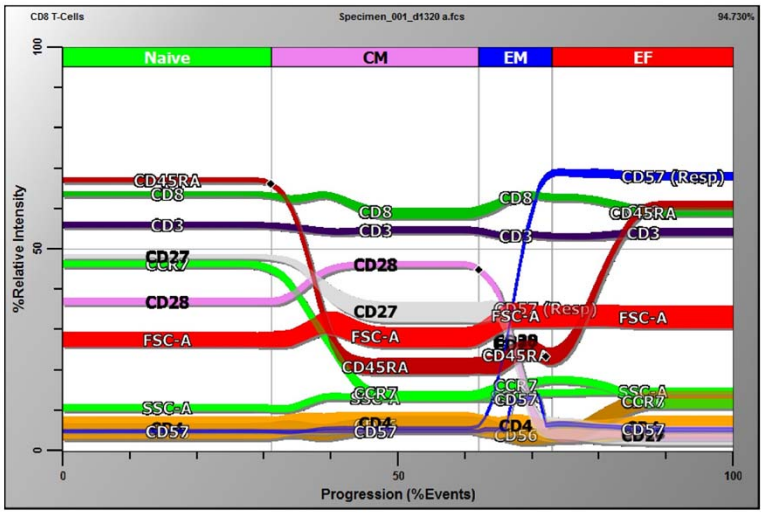


In any modeling process there is an objective function that quantifies the difference between the model and the observed data (y-axis). The other two axes represent two model parameters. Unfortunately, parameter, in the context of modeling ,has a very specific mathematical meaning. In order to avoid confusion, cytometric measurements should be called measurements or features but not parameters. Through an iterative minimization process, the system finds the lowest value of the objective function. Normally, RCS is used to quantify the magnitude of this minimum because it conveys important statistical information about the fitting process. If the RCS is near unity, then the uncertainty in the model is explained as just counting error. If it is much greater than unity, then it means that the model is not fitting some structural information in the data.



PSM has been successfully used to model all the major lineages in bone marrow. The data above represents all the major lineages in a single bone marrow specimen.

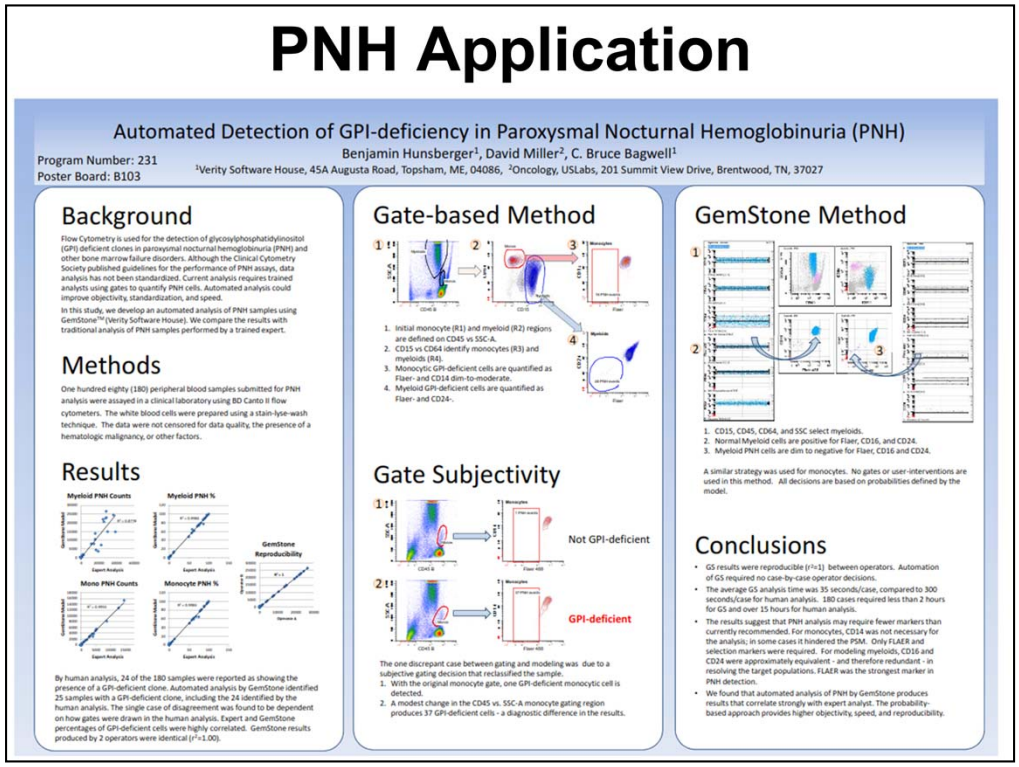
PSM Analysis of T-cell Ag-Dependent Progressions



Margaret Inokuma, BD Biosciences

PSM also has been used to show complex changes in CD8 T-cells as they mature to effector cells.

PNH Application



Automated Detection of GPI-deficiency in Paroxysmal Nocturnal Hemoglobinuria (PNH)

Program Number: 231
Poster Board: B103

Benjamin Hunsberger[1], David Miller[2], C. Bruce Bagwell[1]
[1]Verity Software House, 45A Augusta Road, Topsham, ME, 04086, [2]Oncology, USLabs, 201 Summit View Drive, Brentwood, TN, 37027

Background

Flow Cytometry is used for the detection of glycosylphosphatidylinositol (GPI) deficient clones in paroxysmal nocturnal hemoglobinuria (PNH) and other bone marrow failure disorders. Although the Clinical Cytometry Society published guidelines for the performance of PNH assays, data analysis has not been standardized. Current analysis requires trained analysts using gates to quantify PNH cells. Automated analysis could improve objectivity, standardization, and speed.

In this study, we develop an automated analysis of PNH samples using GemStone™ (Verity Software House). We compare the results with traditional analysis of PNH samples performed by a trained expert.

Methods

One hundred eighty (180) peripheral blood samples submitted for PNH analysis were assayed in a clinical laboratory using BD Canto II flow cytometers. The white blood cells were prepared using a stain-lyse-wash technique. The data were not censored for data quality, the presence of a hematologic malignancy, or other factors.

Results

Myeloid PNH Counts | Myeloid PNH % | GemStone Reproducibility | Mono PNH Counts | Monocyte PNH %

By human analysis, 24 of the 180 samples were reported as showing the presence of a GPI-deficient clone. Automated analysis by GemStone identified 25 samples with a GPI-deficient clone, including the 24 identified by the human analysis. The single case of disagreement was found to be dependent on how gates were drawn in the human analysis. Expert and GemStone percentages of GPI-deficient cells were highly correlated. GemStone results produced by 2 operators were identical (r^2=1.00).

Gate-based Method

1. Initial monocyte (R1) and myeloid (R2) regions are defined on CD45 vs SSC-A.
2. CD15 vs CD64 identify monocytes (R3) and myeloids (R4).
3. Monocytic GPI-deficient cells are quantified as Flaer- and CD14 dim-to-moderate.
4. Myeloid GPI-deficient cells are quantified as Flaer- and CD24-.

Gate Subjectivity

Not GPI-deficient

GPI-deficient

The one discrepant case between gating and modeling was due to a subjective gating decision that reclassified the sample.

1. With the original monocyte gate, one GPI-deficient monocytic cell is detected.
2. A modest change in the CD45 vs. SSC-A monocyte gating region produces 37 GPI-deficient cells - a diagnostic difference in the results.

GemStone Method

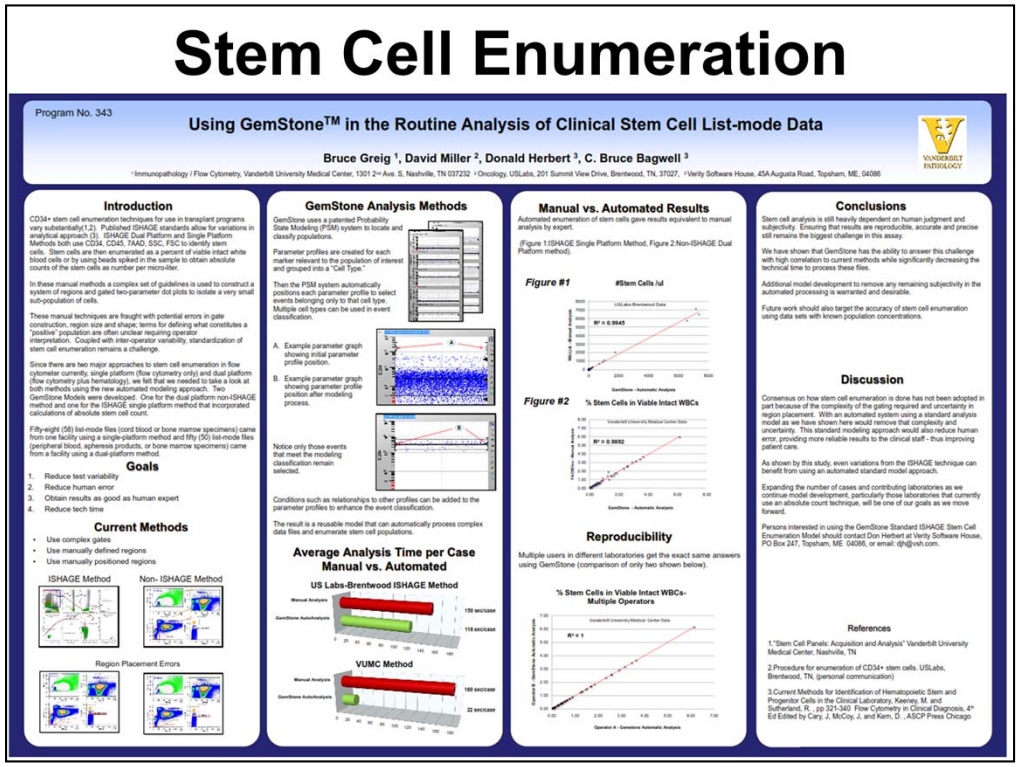
1. CD15, CD45, CD64, and SSC select myeloids.
2. Normal Myeloid cells are positive for Flaer, CD16, and CD24.
3. Myeloid PNH cells are dim to negative for Flaer, CD16 and CD24.

A similar strategy was used for monocytes. No gates or user-interventions are used in this method. All decisions are based on probabilities defined by the model.

Conclusions

- GS results were reproducible (r^2=1) between operators. Automation of GS required no case-by-case operator decisions.
- The average GS analysis time was 35 seconds/case, compared to 300 seconds/case for human analysis. 180 cases required less than 2 hours for GS and over 15 hours for human analysis.
- The results suggest that PNH analysis may require fewer markers than currently recommended. For monocytes, CD14 was not necessary for the analysis; in some cases it hindered the PSM. Only FLAER and selection markers were required. For modeling myeloids, CD16 and CD24 were approximately equivalent - and therefore redundant - in resolving the target populations. FLAER was the strongest marker in PNH detection.
- We found that automated analysis of PNH by GemStone produces results that correlate strongly with expert analyst. The probability-based approach provides higher objectivity, speed, and reproducibility.

Even if there is no progression, PSM can solve complex mixture problems. This slide shows a PNH abstract and poster presented at CYTO 2011 by Ben Hunsberger, demonstrating how PSM can automate this widely ordered test.

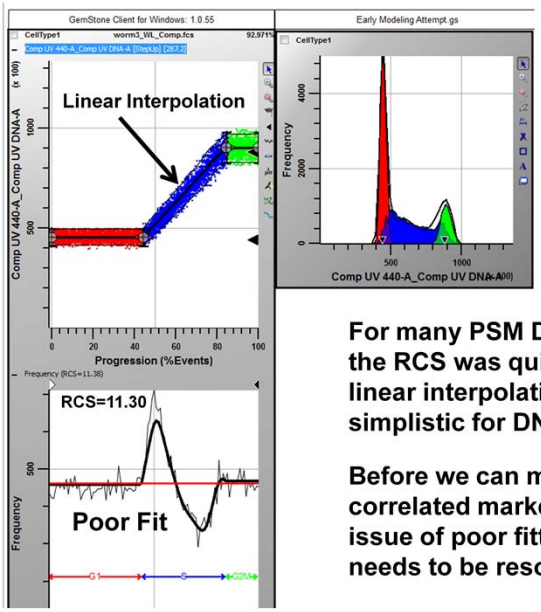


Stem cell enumeration is another example of PSM solving a complex mixture problem. This work was presented at CYTO2011 as an abstract and poster by Bruce Greig and Don Herbert.

What hasn’t worked well with PSM?

Although I didn’t like to admit it, there was one important application where PSM just didn’t work well.

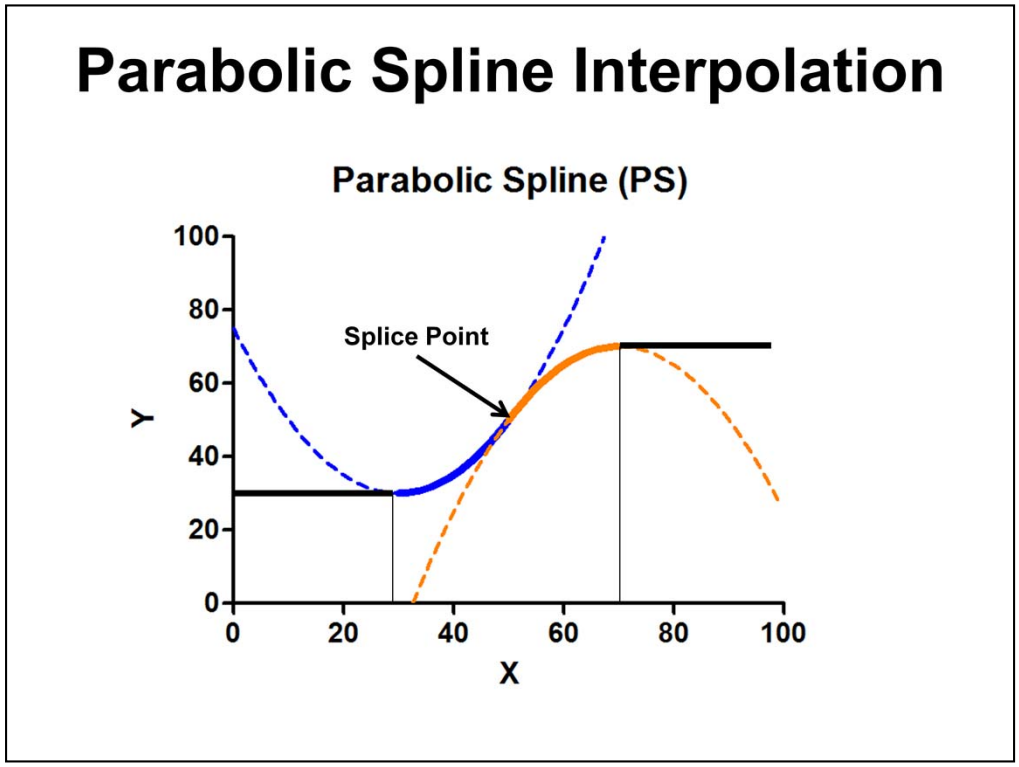
PSM And DNA Content



For many PSM DNA content models, the RCS was quite high, indicating the linear interpolation algorithm is too simplistic for DNA.

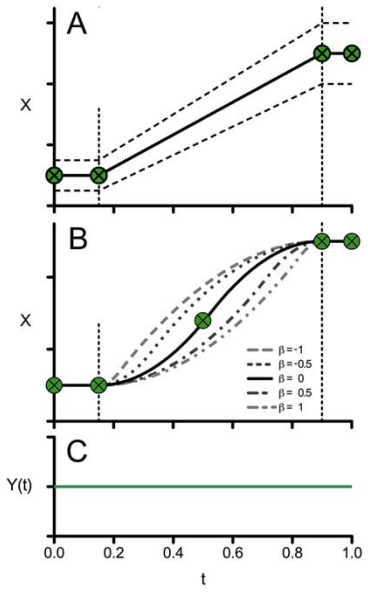
Before we can model the other correlated markers accurately, this issue of poor fitting of DNA content needs to be resolved.

For many of PSM DNA content models, the RCS was quite high, indicating the linear interpolation algorithm was too simplistic for DNA content. The reason we didn't run into this problem with the other applications was because they were immunofluorescence-based methods with relatively large line-spreads. With DNA content, however, the line-spread is narrow enough to convey information about the relative rates of DNA synthesis throughout S phase. Because the linear interpolation methods in the model restricted the rate to be constant, it did not do a good job in fitting the data through S phase. This high RCS was a problem since we wanted to correlate the DNA content measurement with cyclins and pH3. The rule of thumb in modeling high-dimensional data is that if one measurement does not model well, it will negatively affect the modeling of the other measurements. It's very much like compensation where one badly compensated measurement can affect all the other measurements.



After testing a number of possible interpolation methods, the parabolic spline was chosen because of its simplicity and how well it worked with DNA content data. In order to obtain a sigmoidal type of curve, two parabolas of opposite curvatures were spliced together. By changing the ratio of the two curvatures through a beta transition parameter, the system could change the location of the inflection point and model asymmetric sigmoidal types of transitions.

Parabolic Spline Interpolation

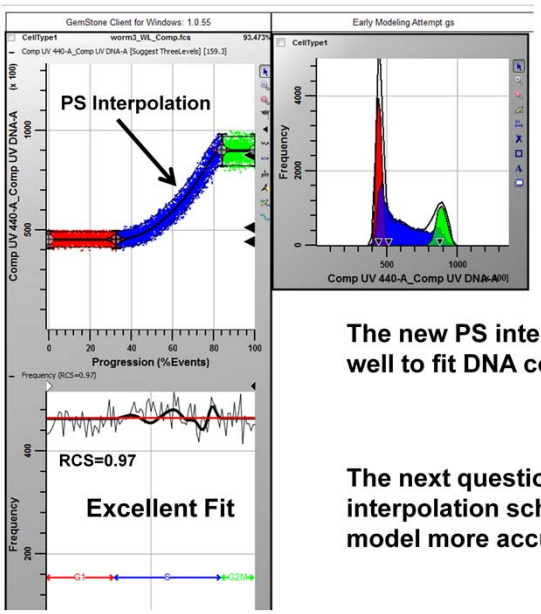


Linear

Parabolic Spline with β values.

The parabolic spline interpolation system (Panel B) adds a great deal of flexibility in modeling transitions like S phase that may have variable relative rates of synthesis throughout the transition.

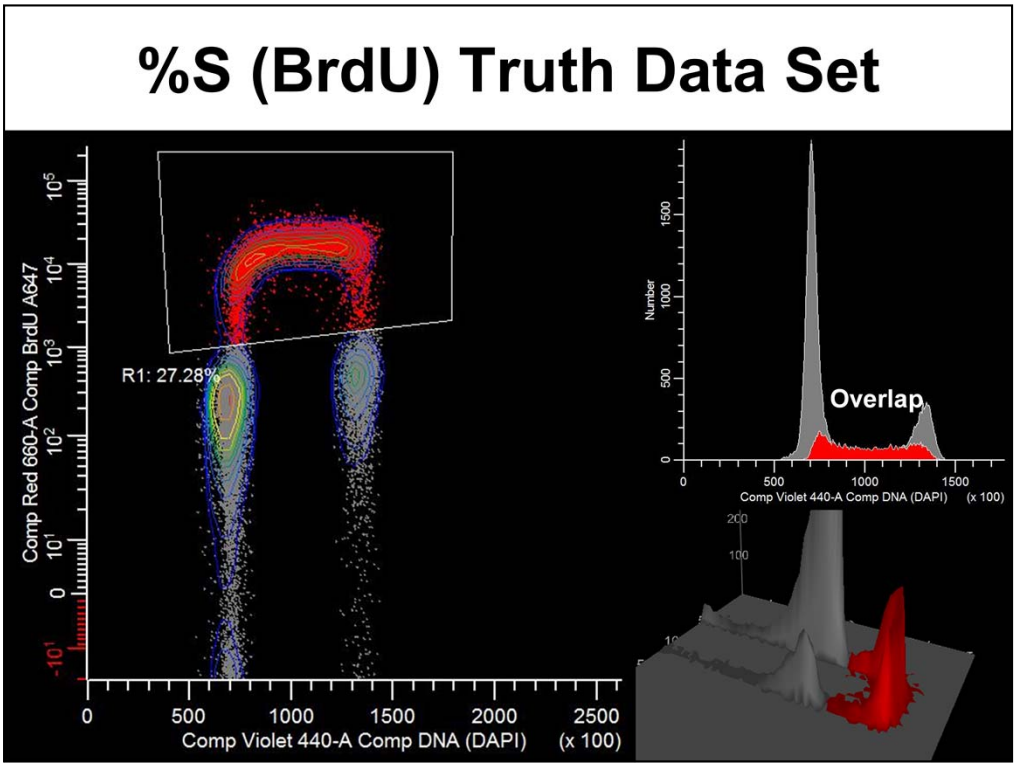
New PSM And DNA Content



The new PS interpolation scheme works well to fit DNA content data.

The next question is whether this new interpolation scheme also makes the model more accurate for quantifying S.

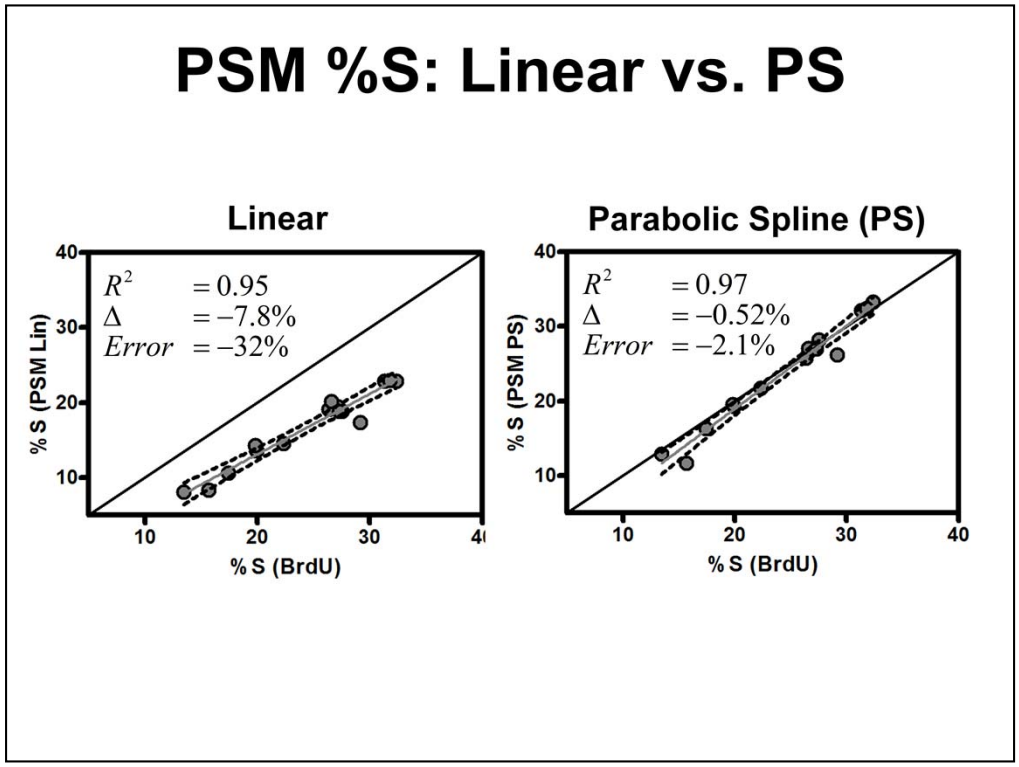
As shown above, the parabolic spline (PS) method worked well fitting sets of DNA content data. The next question to answer was whether this new interpolation scheme resulted in more accurate estimates of S phase.



In order to obtain accurate estimates of S phase, pulse labeled BrdU was used to immunofluorescently detect S phase cells in data that also had correlated DNA content as a measurement. BrdU is highly sensitive and allows the visualization of S phase overlap with both G1 and G2M populations. The isometric plot shows that most of the BrdU positive cells are well away from the G1 and G2 boundaries, which means that the number of undetected BrdU positive cells is likely to be quite small (~1%).

All modeling was completely automatic where all files were analyzed as a batch.

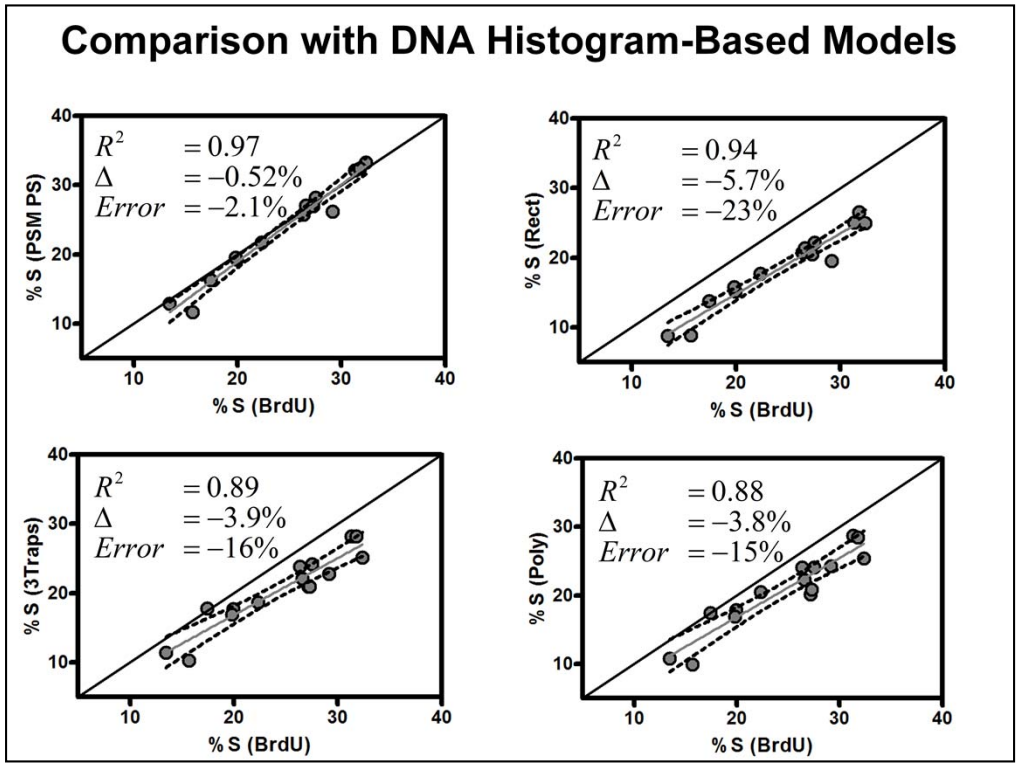
All the modeling done in these comparisons was completely automated.



When compared to the BrdU %S phase estimates, the linear PSM method had an excellent R2 of 0.95, but on the average, underestimated S phase by 7.8% which represented a -32% error. The dotted lines represent th 95% confidence limits and the light gray line is the regression line through the points. When the parabolic spline (PS) was used to interpolate the data, the R2 improved to 0.97 and underestimated S phase by only -0.52% which represented a -2.1% error. These data corroborated our proposition that allowing the model to fit variable relative rates of DNA synthesis results in more accurate S phase estimates.

How do PSM S phase estimates compare to DNA histogram-based models?

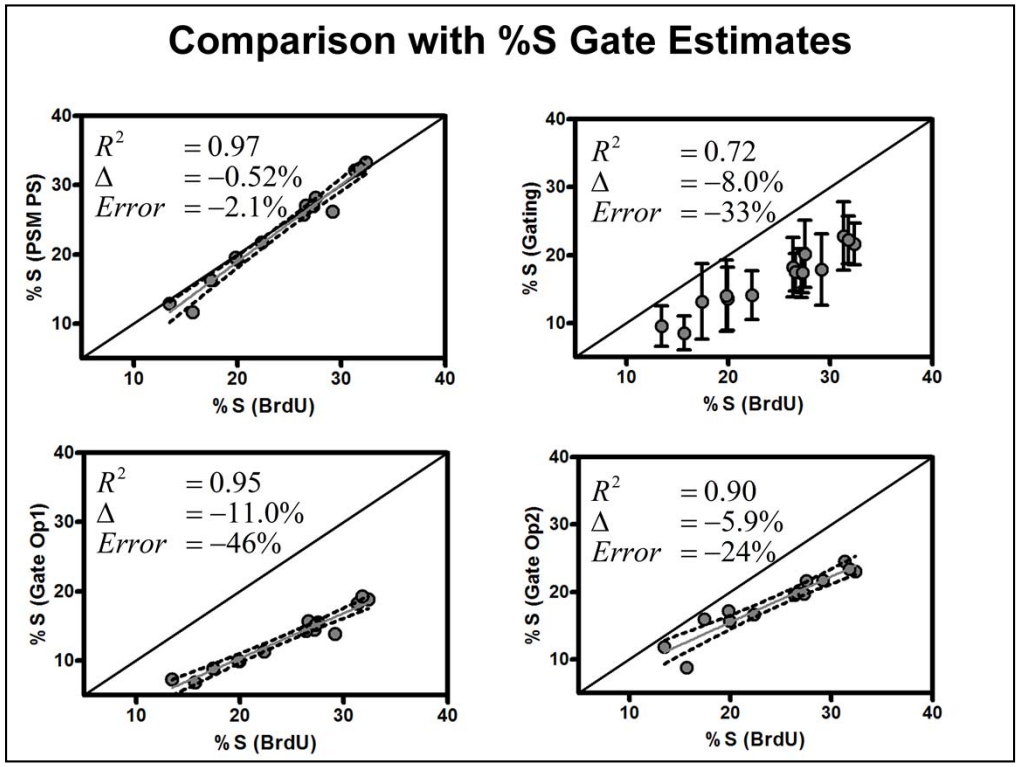
After doing this analysis we realized that the BrdU data allows the comparison of various other analysis methods. We then compared the BrdU truth S phase estimates with popular DNA histogram-based models.



The broadened rectangle method (top-right) had similar characteristics to the PSM linear method. This observation was expected since they are equivalent models. Both the broadened three trapezoids and polynomial gave similar results. Since they both could adjust their shape throughout S phase, they did not underestimate S phase to the same extent as the PSM linear and broadened rectangle methods.

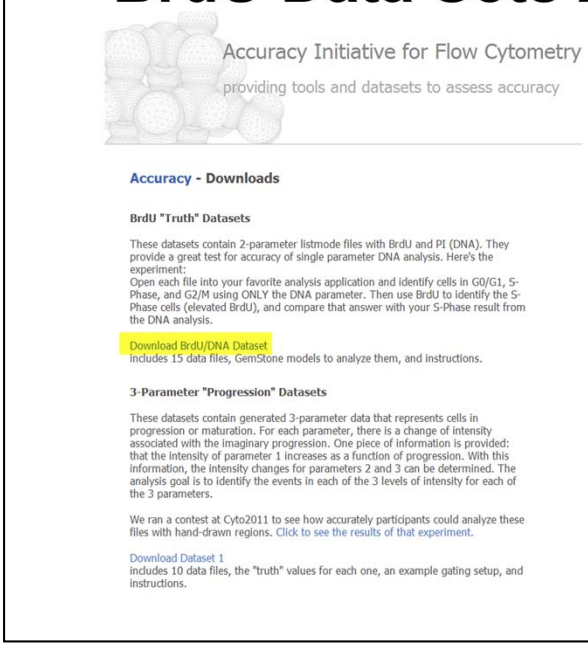
How do PSM S phase estimates compare to manual gating?

We decided to also test manual gating on this same data. Five experienced operators were asked to place regions about the S phase cells in the DNA content histograms. They were not allowed to examine the correlated BrdU data.

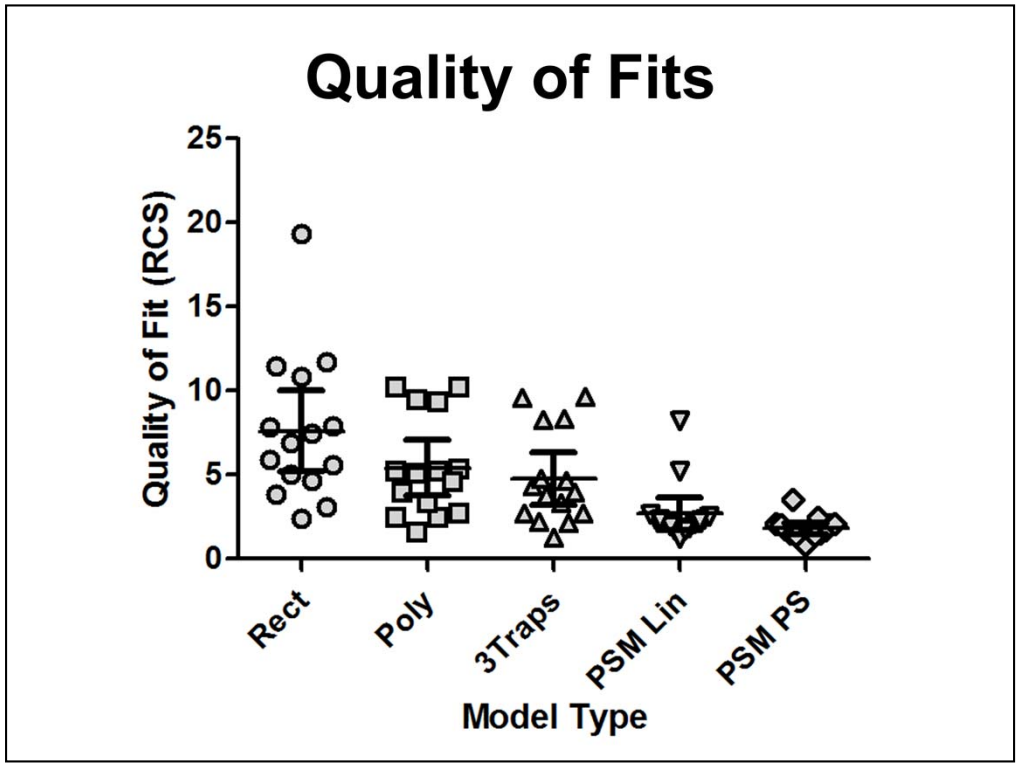


The top-right panel summarizes the gating results for all five operators. Overall, the operators underestimated S phase by 8% which represented a -33% error. Interestingly, if we examine the errors for specific users (bottom two panels) they have dramatically different trajectories through the data. Operator 1 (bottom-left) had an excellent R2 value, but greatly underestimated the true S phase estimates (-46%). On the other hand, Operator 2 was closer to the truth but had a strong tendency to underestimate S phase with larger S phases, resulting in a smaller slope for the regression. These data support the conclusions of the DNA conference held in the 1990's which strongly suggested modeling to be used for S phase estimates.

BrdU Data Sets Availability



All these data sets are available at www.vsh.com/accuracy/downloads (see above).



The above plots shows how the quality of fit, quantified by RCS, changes with different types of models. The new parabolic spline interpolation method resulted in the lowest highest quality scores for all tested methods.

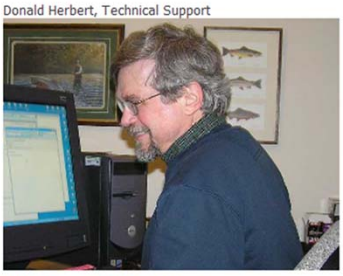
Summary

- **In Probability State Modeling, cumulative percent is used as a surrogate for time and is common to all measurements in a process.**
- **Although PSM has been successfully used for modeling a variety of biological processes, when it was initially applied to DNA content data, the quality of the fits left something to be desired.**
- **The poor DNA content fits were found to be due to PSM's linear interpolation method. Because of DNA content's relatively low line-spread, the variable rates of DNA synthesis throughout S phase were not modeled well by linear interpolation.**
- **After investigating a number of possible non-linear interpolation schemes, the parabolic spline was found to work the best.**
- **BrdU was used to investigate the accuracy of this new interpolation method. The findings were that PS interpolation provides much more accurate S phase estimates than either the linear interpolation scheme or conventional DNA histogram-based models.**

Verity Team

Benjamin Hunsberger, General Manager

Donald Herbert, Technical Support



Mark Munson, Sales

Chris Bray, Application Programmer

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

Collaborators

James Jacobberger, Ph.D.

Mike Sramkoski