Mixture-Model Classification in DNA Content Analysis

Huixia Wang,1 Shuguang Huang2*

Abstract
DNA abundance provides important information about cell physiology and proliferation activity. In a typical in vitro cellular assay, the distribution of the DNA content within a sample is comprised of cell debris, G0/G1-, S-, and G2/M-phase cells. In some circumstances, there may be a collection of cells that contain more than two copies of DNA. The primary focus of DNA content analysis is to deconvolute the overlapping mixtures of the cellular components, and subsequently to investigate whether a given treatment has perturbed the mixing proportions of the sample components. We propose a restricted mixture model that is parameterized to incorporate the available biological information. A likelihood ratio (LR) test is developed to test for changes in the mixing proportions between two cell populations. The proposed mixture model is applied to both simulated and real experimental data. The model fitting is compared with unrestricted models; the statistical inference on proportion change is compared between the proposed LR test and the Kolmogorov–Smirnov test, which is frequently used to test for differences in DNA content distribution. The proposed mixture model outperforms the existing approaches in the estimation of the mixing proportions and gives biologically interpretable results; the proposed LR test demonstrates improved sensitivity and specificity for detecting changes in the mixing proportions.

Key terms
DNA content; high-content imaging; likelihood ratio test; mixture model; mixing proportion

DNA abundance provides important information about cell physiology and cellular proliferation activity. Since most normal somatic cells contain two sets of N chromosomes, such a quantity of DNA is referred to as 2N and is termed diploid (1). The abundance of DNA in a cell, also referred to as DNA content, changes in processes such as cell proliferation and apoptosis. In conjunction with other cellular phenotype information, DNA content can help provide an understanding of the current physiological state of a cell. For instance, a cell with two copies of DNA (tetraploid, or 4N) and expressed phosphorylated Histone H3 may indicate that it is undergoing mitosis. In practice, a cell sample is rarely uniform, but rather is composed of several subtypes. For instance, in terms of the cell cycle, a typical cell population consists of a constellation of G0/G1-, S-, and G2/M-phase cells. For many in vitro assays, there is also a component of cell debris from dead cells. For tumor cells, there may be a component of polyploidy cells that contain even more copies of DNA (e.g., octaploid, 8N). The distribution of DNA content in a sample is therefore a mixture of various components present in the sample. A change in the DNA distribution profile may suggest an interesting biological impact. For instance, an increase in the proportion of 4N cells in an experimental sample may indicate that the treatment applied is causing mitosis arrest of the cells.

In this research, we consider a new technology, high-content imaging (HCI), which allows molecule-specific (DNA, RNA, proteins, etc) measurements to be made with remarkable sensitivity. HCI systems are capable of measuring cellular fluorescence-labeled targets through multiple channels simultaneously, where each channel represents a fluorophore. In the past few years, HCI technology has become widely
used in high-throughput screening studies. Cellomics Target Activation BioApplication systems (Cellomics Inc. http://www.cellomics.com), Acumen Explorer (http://www.ttplabtech.com/explorer/explorer.htm), and Cellumen (http://www.cellumen.com) are examples of such systems. The DNA content data analyzed in this article were generated using the Cellomics ArrayScan, in which the fluorescence-labeled DNA was used for identifying individual cells, referred to as object segmentation. In theory, the observed distribution of DNA content is a mixture of the corresponding components that are present in the sample: cell debris, G0/G1-phase, S-phase, G2/M-phase, and possibly cells with more copies of DNA (these could be multiploid or cell clumps due to segmentation errors). In the DNA content distribution obtained from high precision instruments, such as flow cytometry, the S-phase component is represented as a flat area between the 2N and 4N modes. However, the data corresponding to this transition stage are usually not observable in HCl experiments because of instrument noise. The observed distributions of HCl DNA content are typically multimodal, with the degree of overlap between modes dependent upon experimental variables such as sample handling, cell type, inherent biological variation among cells, the choice of object segmentation algorithm, as well as other factors.

One key interest in DNA content analysis is to quantitatively classify the cell subtypes by deconvoluting the overlapping distributions into estimated mixing proportions. Such quantitative classification provides very useful information about the cell sample. For example, this classification can address questions about the proportion of cells with 2N DNA (G0/G1-phase) and with 4N DNA (G2/M-phase). Gasparri et al. (2) proposed a multiparameter approach to quantify the cell cycle phases using four markers, but the subpopulations were discriminated using some arbitrary thresholds. A number of statistical methods have been developed for the deconvolution of DNA mixing proportions in the past. Mann et al. (3) used a mixture modeling approach where each bin of a histogram represented a component of the mixture population. Baldetorp et al. (4) fit flow cytometric DNA data with a mixture model of seven components corresponding to G0/G1-phase, S-phase, G2/M-phase, CRBC (chicken red blood cells, internal control), TRBC (trout red blood cells, internal control), cell debris, and random noise. The readers are referred to Vindelov and Christensen (5) and Diaz-Frances and Sprott (6) for discussion on DNA distribution model specification and parameter estimation. Eudey (7) provides an extensive review on the statistical modeling of DNA content distributions. Although previously proposed methods represent significant advances in this field, these methods share some common shortcomings. First, data are modeled on an inappropriate scale, which may result in a severe violation of modeling assumptions. Second, the previous models do not fully utilize the available biological information in the assay, so even when the parameter estimates achieve satisfactory goodness-of-fit with respect to the mathematical model, the results do not have a sensible biological interpretation. Finally, the key question of whether an experimental treatment has had significant effect on the distribution of cellular components is not addressed with statistical rigor.

All statistical models require assumptions about the data. Typical assumptions include normally (Gaussian) distributed errors with homogeneous variances across the subpopulations. The conclusions drawn from the statistical method may be in error when the model assumptions are violated. It is quite common in fluorescence-based experiments (e.g., flow cytometry, HCl, microarray) for the distribution of the intensity data to be skewed to the right (7–9); see also Figure (2a) and (2b). Additionally, higher intensity values tend to have higher variation (5). For existing methods that assume Gaussian mixtures for the raw DNA intensity data, the modeling assumptions are clearly inappropriate. A logarithmic transformation often stabilizes the variation and eliminates the skewness in the data distribution. The logarithmic transformation converts multiplicative changes into additive changes, and downweights the influence of extreme values (e.g., outliers caused by cell clumping).

The analysis of mixture models is a challenging problem within the field of statistics, and this is particularly complicated by the complexity of DNA content distributions. Certain cellular components, such as cell debris and S-phase cells, are difficult to model using a parametric form. Depending on the settings of a particular experiment, certain components may not be represented by enough cells to allow for reliable modeling. Such complexities can lead to nonconvergence of the model fitting algorithm and provide unstable parameter estimates [e.g., over estimation of cell debris (7)]. In addition, the existing methods use goodness-of-fit as the criterion for evaluating the model performance, but models with similar goodness-of-fit can yield dramatically different estimates for the fractions of cells (7). We wish to emphasize that a statistical approach is less favorable if it does not reflect the underlying biology, even though it might yield some good mathematical properties, for instance a satisfactory goodness-of-fit. In DNA content data analysis, we propose that modeling and parameter estimation be “supervised” by the underlying biological truth. Since fluorescence intensity in HCl is proportional to DNA abundance, the peak DNA intensity corresponding to G0/G1-phase cells should be twice the peak intensity corresponding to G0/G1-phase cells. Such critical information should be taken into account in the model parameterization. In particular, the expected twofold difference between these two modes for the raw intensity is one unit on a base-2 logarithmic scale. We propose a mixture model that incorporates the restriction that the modes of the neighboring subpopulations differ by one unit after log2 transformation. The study results show that the proposed parameterization constraint, though simple and straightforward, effectively improves both modeling stability and the interpretability of the DNA content data analysis.

In many biomedical research efforts, it is important to assess whether a treatment perturbs the DNA content distributions. In the context of HCl studies, the distributions may arise from cell samples in different wells on a microplate, where each well is treated differently. To answer the question
in a rigorous statistical manner, we propose a likelihood ratio (LR) test for comparing the mixing proportions of two experimental cell populations. This differs from traditional approaches, where the significance of the change in mixing proportions is determined subjectively. Our proposed method also differs from the Kolmogorov–Smirnov test (KS), which is sensitive to nonspecific artifacts. In applications to simulated data and real experimental data, our proposed LR test demonstrates both higher sensitivity and specificity than the other approaches considered.

**STATISTICAL METHODS**

**Mixture Model**

The distributional properties of DNA content data vary with the experimental settings, including the cell type used, assay technology, instrument sensitivity, and so on. Based on our experience, the DNA distribution of cancer cells typically consists of the following components: cell debris with less than one copy of DNA (<2N), G0/G1-phase cells with a single copy of DNA (2N), G2/M-phase cells with a double copy of DNA (4N), and multinuclei cells containing 8N or higher DNA. As mentioned previously, the S-phase cells are not directly observable due to instrumental noise associated with HClI technologies. Although we have no theoretical justification, we have found empirically that the difference between the mode of the cell debris and that of the 2N cells is approximately one log-unit. In light of these considerations, we propose a four-component Gaussian mixture model with the restriction that the means of neighboring components differ by one log-unit.

In this article, we focus on the comparison between two DNA distributions. Let $X_{ij}$ represent the $j$th log2DNA intensity measurement from sample $i$, where $i = 1, 2, j = 1, \ldots, n_i$. We assume that $X_{ij}$ are independent random variables following a Gaussian mixture distribution given by

$$f(x) = \sum_{k=1}^{4} p_k \phi(x; \mu_k + k - 1, \sigma_k^2), \tag{1}$$

where $\phi(x; \mu, \sigma^2)$ denotes the density function of the normal distribution with mean $\mu$ and variance $\sigma^2$, $p_{13}$, $p_{23}$, $p_{14}$, and $p_{4}$ stand for the proportions of cell debris, 2N cells, 4N cells, and 8N cells for sample $i$, respectively. Notice that we need the constraints that $p_k > 0$ and $\sum_{k=1}^{4} p_k = 1$ for each $i$. Our main interest is to test whether the treatment causes a change in the proportion of cells with DNA of 4N or more. Therefore, our hypotheses to be tested are

$$H_0 : p_{13} + p_{14} = p_{23} + p_{24} \quad \text{versus} \quad H_a : p_{13} + p_{14} \neq p_{23} + p_{24}. \tag{2}$$

A rejection of the null hypothesis signifies that there is strong evidence for a change in the mixing proportion of cells with ≥4N DNA. Such an increase suggests possible mitosis arrest caused by the treatment.

**Likelihood Ratio Test**

To test the hypotheses in [2] of a change in the proportion of polyploidy cells, an LR test procedure is proposed. Let $\theta = (\mu_k, \sigma_k^2, p_k, i = 1, 2, k = 1, \ldots, 4)$ denote the vector of unknown parameters. Let $L(\hat{\theta}; \bar{X})$ denote the likelihood function based on data $\bar{X} = (x_{i1}, \ldots, x_{in_i}), i = 1, 2$, and let $\theta$ and $\bar{\theta}$ be the maximum likelihood estimates for $\theta$ obtained under $H_0$ and under $H_a$ respectively. Applying the EM algorithm (10), the estimates $\theta$ and $\bar{\theta}$ can be easily obtained. Details of the algorithm can be found in the Supplementary Material available from the Cytometry Part A website. The LR test statistic is given as

$$L = 2 \left[ \ln \left( L(\bar{\theta}; \bar{X}) \right) - \ln \left( L(\hat{\theta}; \bar{X}) \right) \right]. \tag{3}$$

As noted in (11), since the null hypothesis [2] is specified in the interior of the parameter space, the regularity conditions for the asymptotic null distribution of the LR statistic do not break down as they do when testing the number of components. We reject the null hypothesis [2] when $L$ is greater than the $(1 - \alpha)$th quantile of the $\chi^2$ distribution with one degree of freedom.

**Bootstrap Test**

We also consider a bootstrap test based on the Wald-type test statistic $W = \frac{(p_{13} + p_{14}) - \frac{1}{n_1} (p_{23} + p_{24})}{\sqrt{\bar{\sigma}^2}}$, where the numerator is the estimated change in the proportion of polyploidy cells, and the denominator is the standard error. We resampled $n_2$ observations from $\bar{X}_1 = (x_{i1}, \ldots, x_{in_1})$, and $n_2$ observations from $\bar{X}_2 = (x_{i2}, \ldots, x_{in_2})$ with replacement. The restricted mixture model [1] is fitted to the bootstrap $\bar{X}_1$ and $\bar{X}_2$ samples separately, and the bootstrap change in mixing proportion is estimated. This procedure was repeated 1,000 times. The sample standard deviation of the bootstrap mixing proportion change was used to estimate the SE in $W$. The test can then be constructed by using the standard normal distribution as the reference.

**Kolmogorov—Smirnov Test**

The Kolmogorov–Smirnov (KS) test is frequently used to compare two DNA content distributions. The KS test assesses whether the observed difference between two distributions may be due to random chance by evaluating the maximum difference between the two cumulative distribution functions; see Gibbons (12) and Conover (13) for more details on the KS test, Young (14) and Lampariello (15) for applications of the KS test to DNA content analysis. However, the KS test has been criticized for being overly sensitive in applications, since it often detects a statistically significant difference that is not of biological interest. The primary reason for this criticism is based on the observation that the KS test is sensitive to any change in the shape of the distribution, regardless of the experimental context. For instance, a common phenomenon in HClI experiments is the existence of systematic drifts across the distributions because of compound fluorescence, instrument variation, plate edge effects, unintended differences in...
sample handling, as well as other causes. These systematic effects cause a location shift of the entire histogram, which can subsequently result in a significant KS test. Since these are experimental artifacts, they should be distinguished from changes that are biologically meaningful (e.g., perturbed mixing proportions). In this article, the KS test is considered for comparison purposes.

**DATA AND RESULTS**

**Simulation**

A simulation study was conducted to assess the performance of different methods for testing the mixing proportion change between control and treatment groups. For the control sample, the data were generated from model [1] with \((\mu_1, \sigma_1^2, \sigma_2^2, \sigma_3^2, \sigma_4^2) = (14.3, 0.4, 0.05, 0.05, 0.4)\) and \((p_{11}, p_{12}, p_{13}, p_{14}) = (0.1, 0.65, 0.2, 0.05)\). For the treated cell populations, three different cases were considered. In each case, the mixing proportions of the four components were set at \((0.1, 0.65 - p, 0.2 + p, 0.05)\) for the treatment sample, with \(p = 0\) included to examine the specificity, and \(p = 0.06\) included to examine the sensitivity of each method. In Case 1, the treatment sample was generated using the same parameters as those used for the control sample except with different mixing proportions. Case 2 mimics the situation where the treatment sample has a small location shift. Specifically, \(\mu_2\) was set to 14.35, which corresponds to a location shift of 0.05. In Case 3, the treatment samples for the 2N and 4N cells were generated using \(\sigma_{22}^2 = \sigma_{23}^2 = 0.1\). Each sample consisted of 500 observations and 500 simulation runs were conducted.

The LR test, the bootstrap test (Boot), and the KS test were performed for each run of the simulation. Table 1 summarizes the observed Type I error rates for the three methods when the nominal significance level is set to \(\alpha = 0.05\). Since the KS test is sensitive to any change in the shape of the distribution, it reflects the location shift in Case 2 and the variance change in Case 3, neither of which is of biological interest in DNA content analysis. Both the LR test and the bootstrap test maintain the Type I error around 0.05 in all three cases.

The receiver operating characteristics (ROC) curve is a useful graphical tool to investigate the specificity and sensitivity of a test procedure. An ROC curve is generated by plotting the true positive rate (sensitivity/power) against the false positive rate (1-specificity/Type I error) as the threshold of the test statistic is varied. A smaller distance between the ROC curve and the upper left corner of the plotting region indicates better performance for the test procedure. Figure 1 shows the ROC curves for all three tests in Cases 1–3. The solid, dotted, and dashed curves are for the LR, KS, and bootstrap test, respectively. In all three cases, the LR and the bootstrap test perform similarly, with both tests demonstrating higher sensitivity and specificity than the KS test. Since the bootstrap approach is computationally more expensive, we focus on the LR test in the following analyses.

<table>
<thead>
<tr>
<th>Case</th>
<th>KS</th>
<th>LR</th>
<th>Boot</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.048</td>
<td>0.046</td>
<td>0.032</td>
</tr>
<tr>
<td>2</td>
<td>0.440</td>
<td>0.046</td>
<td>0.032</td>
</tr>
<tr>
<td>3</td>
<td>0.504</td>
<td>0.062</td>
<td>0.052</td>
</tr>
</tbody>
</table>

Table 1. Type I error rates for the Kolmogorow-Smirnov test (KS), the likelihood ratio test (LR), and the bootstrap test (Boot) at nominal significance level of 0.05

**Figure 1.** The ROC curves for the LR test (solid), the KS test (dotted), and the bootstrap test (dashed) in Cases 1–3.
The purpose of this HCI study is to investigate the effect of nocodazole on HeLa cells. Nocodazole is a mitosis spindle assembly inhibitor that causes cell mitosis arrest. In terms of DNA content distribution, a cell population treated with nocodazole is expected to have an increased proportion of cells with higher DNA content ($\geq 4N$) than found in an untreated population (16). The extent of the proportion change depends on the duration and the concentration of the drug used.

HeLa cells were routinely cultured in DMEM supplemented with 10% FBS and 2 mM $L$-Glutamine (Invitrogen #25030-081). Cells were harvested for plating by washing once with PBS, then washing with trypsin-EDTA (Invitrogen #25300-054), and then incubating at room temperature for 1–2 min until detached. A total of 3,000 cells were plated per well in 100 $\mu$L complete media in poly-D-lysine coated 96-well microplates. Only the first five columns of the microplate were used; the last seven columns were left blank. Column 1 was treated with DMSO control. From column 2 to column 5, increasing concentrations of nocodazole were applied: 20, 40, 80, and 160 nm. The same concentration was applied to all the wells in a column, so each treatment condition had eight technical replicates located on each row A, B, ..., H of the plate. The last seven columns were left blank. Column 1 was treated with DMSO control. From column 2 to column 5, increasing concentrations of nocodazole were applied: 20, 40, 80, and 160 nm. The same concentration was applied to all the wells in a column, so each treatment condition had eight technical replicates located on each row A, B, ..., H of the plate. Following 24 h of drug treatment, cells were fixed and stained with 200 ng/mL Hoechst 33342 fluorescent dye (Molecular Probes #21492) for HCI analysis using Cellomics Target Activation BioApplication (17). DNA intensity for individual objects was captured in channel one and extracted for analysis. Other phenotypic data were also collected but not considered here. A 20× lens objective was used to capture the valid cells that meet the user-defined object selection parameters (size, number of nuclei, etc.). The algorithm sampled the cells in the center square, which accounts for $2/\pi \approx 64\%$ of the total well area. Because of the effect of nocodazole, the cell counts range from 1,200 to 7,500 among the 40 wells. The cell-level DNA intensity data were log$_2$-transformed in the following analyses.

Figure 2 shows the histograms of DNA intensity for the control and 80 nm nocodazole-treated samples in row G. As shown in Figure 2(c), the majority of the control cells contain 2N DNA (with a spike around 15.5), while a smaller proportion of cells contain 4N DNA (with a spike around 16.5). There were some cell debris represented in the left tail of the distribution, and some cells with 8N DNA were represented on the right tail. For the cells treated with nocodazole, 4N cells predominate, and the proportion of 8N cells is also increased. It is evident that, on the logarithmic scale, the distribution of each component is bell shaped. In addition, the adjacent modes of the 2N, 4N, and 8N components differ by approximately one unit.

The solid curve in Figure 2 represents the mixture density estimate for the model with the restriction on the adjacent modes, whereas the dashed line represents the density estimate for the model without the restriction. Both curves appear to fit the data reasonably well. Table 2 displays the parameter estimates and the $\chi^2$ goodness-of-fit statistics obtained for the two samples fit under each of the two mixture models. The unrestricted model results in smaller $\chi^2$ statistics, but the restricted mixture model yields parameter estimates with more plausible biological interpretations. From Figure 2, it is clear that the treated sample has a much higher proportion of $\geq 4N$ cells than does the control sample. However, the unrestricted mixture model fails to capture this difference ($\hat{p}_3 + \hat{p}_4 = 0.26$.

---

**Figure 2.** The first row shows the histograms for the raw DNA data, and the second row depicts the histograms for the log$_2$-transformed DNA data with fitted density curves. Figures (a) and (c) are for the DMSO-treated control sample in row G. Figures (b) and (d) are for the 80 nm nocodazole-treated sample in row G. In (c) and (d), the solid curves are the fitted mixture density functions with the location restriction; the dashed curves are the fitted mixture densities without the location restriction.
for control and 0.25 for the treated sample), since the four subpopulations are not separated in a biologically meaningful manner. The restricted mixture model gives $\hat{p}_3 + \hat{p}_4 = 0.25$ for the control and 0.58 for the treatment, and the proposed LR test indicates that these two samples have significantly different mixing proportions.

To assess the specificity of the tests, we performed all possible pair-wise comparisons on the proportion of 4N cells among the eight control samples in the first column, resulting in 28 comparisons in total. Since these eight samples are technical replicates treated with DMSO, their differences are primarily due to random noise and potential systematic biases associated with plate position. To control the overall Type I error rate, the Bonferroni correction was applied, which requires the significance level for each individual test to be $0.05/28 = 0.0018$. The results show that KS is overly sensitive, since it misidentified 20 out of 28 comparisons to have significant differences in the proportion of $\geq$4N cells. The LR test based on the unrestricted model, referred to as LR0, performed better than KS, but it still misidentified 12 out of 28 comparisons. The proposed LR test performed the best among the three methods, resulting in only three misidentifications.

To assess the sensitivity of each method for detecting the treatment effect, we performed separate pairwise comparisons between the control and the nocodazole-treated samples for each of the four concentrations. This results in a total of 64 pairwise comparisons per concentration. As before, the Bonferroni correction was applied to maintain the overall Type I error rate at 0.05 for each concentration. The ROC curves in Figure 3 demonstrate that LR achieved the highest sensitivity and specificity among the three methods. Table 3 summarizes the average of the $\geq$4N proportion changes between the treated and control samples, as well as the number of significant comparisons identified by the LR test for each concentration. These results suggest that the $\geq$4N cell proportion increased with the concentration of nocodazole. Almost all the replicate wells with concentrations

---

**Table 2.** Parameter estimates and the $\chi^2$ goodness-of-fit statistics for models with and without the location restriction in the comparison between the DMSO-treated control and 80 nm nocodazole-treated samples in row G

<table>
<thead>
<tr>
<th></th>
<th>$\hat{\mu}_1$</th>
<th>$\hat{\mu}_2$</th>
<th>$\hat{\mu}_3$</th>
<th>$\hat{\mu}_4$</th>
<th>$\hat{\beta}_1$</th>
<th>$\hat{\beta}_2$</th>
<th>$\hat{\beta}_3$</th>
<th>$\hat{\beta}_4$</th>
<th>$\chi^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Mixture model with location restriction</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Control</td>
<td>14.29</td>
<td>15.29</td>
<td>16.29</td>
<td>17.29</td>
<td>0.20</td>
<td>0.55</td>
<td>0.23</td>
<td>0.02</td>
<td>765.1</td>
</tr>
<tr>
<td>Nocodazole</td>
<td>14.54</td>
<td>15.54</td>
<td>16.54</td>
<td>17.54</td>
<td>0.24</td>
<td>0.18</td>
<td>0.49</td>
<td>0.09</td>
<td>393.7</td>
</tr>
<tr>
<td><strong>Mixture model without location restriction</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Control</td>
<td>15.23</td>
<td>15.35</td>
<td>15.47</td>
<td>16.46</td>
<td>0.46</td>
<td>0.29</td>
<td>0.13</td>
<td>0.13</td>
<td>761.5</td>
</tr>
<tr>
<td>Nocodazole</td>
<td>15.6</td>
<td>16.45</td>
<td>16.68</td>
<td>17.48</td>
<td>0.49</td>
<td>0.25</td>
<td>0.17</td>
<td>0.08</td>
<td>280.5</td>
</tr>
</tbody>
</table>

The $\hat{\mu}_k$ and $\hat{\beta}_k$ are the estimated location and proportion parameters for the $k$th component, $k = 1, \ldots, 4$.

---

**Figure 3.** The ROC curves for the LR, LR0, and KS tests in the nocodazole HCI study to detect changes in mixing proportions between control and treated samples. The solid curve is for the LR test under the mixture model with the location restriction (LR), the dashed curve is for the LR test under the mixture model without location restriction (LR0), and the dotted curve is for the KS test.
of 40 nm or higher showed significantly higher proportions of ≥4N cells relative to the control wells.

**DISCUSSION**

We have focused on two specific issues pertinent to DNA content analyses: (1) appropriate modeling of DNA content data and (2) appropriate statistical inference for testing changes in mixing proportions. For data modeling, we have highlighted three issues as important. First, it is important to model the data on an appropriate scale to ensure that the model assumptions are reasonably satisfied; otherwise the downstream analysis results may not be reliable. Since in our experience most DNA content data are right-skewed, we suggest that the logarithmic transformation be applied. Second, it is important to incorporate biological and engineering knowledge into the model to ensure that results are both biologically plausible and interpretable. We propose taking advantage of the biological fact that the neighboring modes in the DNA content distribution are expected to differ by one unit after the log₂-transformation. We restricted the parameter estimates in the mixture model to satisfy this condition. Third, it is important to choose an appropriate criterion for assessing the performance of the statistical model. Goodness-of-fit statistics judge how well a particular model fits the data, but this should not be interpreted as the definitive measure of model performance. By allowing known biological facts to play a role in supervising the modeling, we improved the method in a manner that is not reflected in general purpose measures of model performance, such as the goodness-of-fit statistic.

In the conduct of inference on treatment effects, we have emphasized several important points. First, it is important to correctly establish a hypothesis of interest. The KS test is appropriate for detecting changes in the distribution profile, while a t-test or a rank-based test may be used for detecting the location changes in the distributions. In DNA content analyses, general interest lies specifically in detecting changes in the mixing proportions. A desirable test therefore is one that is sensitive to changes in mixing proportions but at the same time is robust to other types of changes, such as location shifts and differences in variation within classes. Second, the significance of changes in the mixing proportions should be determined in a statistically stringent manner. Traditionally, the significance of a drug effect is often determined subjectively. In this article, we proposed an LR test to objectively quantify the likelihood of a treatment effect in terms of a p-value.

<table>
<thead>
<tr>
<th></th>
<th>20 nm</th>
<th>40 nm</th>
<th>80 nm</th>
<th>160 nm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average change in proportion of cells with ≥4N DNA</td>
<td>0.029</td>
<td>0.084</td>
<td>0.319</td>
<td>0.472</td>
</tr>
<tr>
<td>Number of tests called significant (out of 64 comparisons)</td>
<td>25</td>
<td>62</td>
<td>64</td>
<td>64</td>
</tr>
</tbody>
</table>

Using HCI platforms, the transitional components (e.g., S-phase) between the modes of DNA content are often not observable because of instrument noise. This problem does not arise in data generated with high-precision instruments such as flow cytometry. For this reason, we have made no attempt at this time to specifically model the S-phase component in the distribution, and we by no means intend to fix the parameterization of the model. As the HCI technology advances, however, one can expect that the molecular targets be measured with higher sensitivity and specificity, thus allowing more components to be observable with high accuracy. The methodology developed here can be extended to accommodate such additional features in the data. For instance, assuming that the raw DNA intensities of the S-phase cells follow a uniform distribution between the 2N mode and the 4N mode, i.e., the raw DNA intensity data of S-phase cells have a uniform distribution between 2^{k+1} (mode of the 2N cell population) and 2^{k+2} (mode of the 4N population), the log-transformed data could then be captured as a truncated exponential distribution in our current model. We have modeled the DNA content from cell debris using a Gaussian distribution with a mean equal to one unit less than that of the 2N cells. This has no theoretical justification, but from our experience this choice has empirical validity. A different parameterization, however, may be required for other experimental settings. For example, cell type, imaging platform, and choice of object segmentation algorithm all appear to affect the form of DNA content distribution.

Because of the large amount of data collected in such high-throughput experiments, computational cost needs to be considered in HCI DNA data analysis. It is critical to have an analytical process in place that both confers good performance (e.g., reliable estimates of mixing proportions, high sensitivity, and specificity in hypothesis testing) and is also of sufficient computational efficiency so that data analysis is not a bottleneck. In the context of our simulation study, the LR test performed similarly to the bootstrap test in terms of sensitivity and specificity, but was much more computationally efficient. For a test data set with 500 observations in each of two populations, using R (version 2.3.1) in a 3.4 GHz Dell computer with 3.0 GB of RAM, the LR test required 0.50 s of computing time, while bootstrap approach with 1,000 resampling required 155.25 s of computing time. Because of the requirement of computational efficiency, an approach that is marginally suboptimal in terms of performance is necessary.

**ACKNOWLEDGMENTS**

We would like to thank Steve Iturria, Sujit Ghosh, Bao-guang Han, and Jerry Davis for their tremendous help in editing the manuscript and in providing valuable suggestions. We would also like to thank Caiping Li for sharing the high content imaging data with us. The research of Wang H. was supported by National Science Foundation grant DMS-0706963.

**LITERATURE CITED**