Target DNA Quantification Using Expectation-Maximization Clustering of Digital PCR Droplets

Joyce Emlyn B. Guiao[1]*, Romeric F. Pobre[2], and Frumencio F. Co[1]
[1] Mathematics and Statistics Department, De La Salle University
[2] Physics Department, De La Salle University
**Corresponding Author: joyce_emlyn_guiao@dlsu.edu.ph*

Abstract: We proposed and implemented an Expectation-Maximization (EM) clustering method in gene quantification of target DNA (Deoxyribonucleic Acid) samples for digital PCR (Polymerase Chain Reaction) device. Digital PCR (dPCR) detects and quantifies target molecules, such as nucleic acid strands from bacteria, viruses, fungi, and other microbiological samples. The dPCR workflow partitions the sample into thousands of droplets that emit fluorescence upon amplification. High intensity indicates at least one target is contained in the droplet and is classified as "positive"; otherwise, low intensity indicates no target, and the droplet is classified as "negative". Droplet classification becomes challenging when several intermediate droplets called "rain" are present, causing severe misclassification. Because nonoptimal data is frequent in dPCR studies, droplet classifiers should be robust to the presence of rain, baseline shifts, multiple fluorescence populations, and poor separation of populations. Performance analysis on the EM clustering method is pitted against three well-known clustering methods of dPCR using both real and simulated data sets with varying concentration levels and rain distributions. Preliminary results showed that the EM clustering method performed better than most three cluster methods of dPCR in terms of accuracy, precision, and linearity of estimates.

Key Words: Target DNA Quantification, Digital PCR, Mixture Models, Expectation-Maximization Clustering

1. INTRODUCTION

Digital PCR (dPCR) is a method to detect and quantify target molecules found in DNA or RNA which can be used in the medical diagnosis of viral infections such as COVID-19 (Xu et al., 2020). The dPCR workflow, shown in Figure 1, consists of the steps: partitioning, amplification, and digitization. Partitioning refers to dividing the DNA sample into thousands of equal-sized droplets in an assay chip, whereby amplification by polymerase chain reaction (PCR) causes the emission of fluorescence in each droplet. Then, based on the intensity, the digitization step classifies each droplet as "positive" (high intensity; containing at least one target) or "negative" (low intensity; no target) (Cao et al., 2017). The problem arises when a substantial amount of ambiguous intermediate droplets called "rain" are present in the dPCR assay which is frequently reported in studies. Rain droplets are one of the main causes of misclassification and consequently leads to erroneous target concentration estimates (Wong et al., 2017). To improve data quality and reduce rain, design parameters in dPCR sample preparation should be optimized; however, there are times this is difficult to achieve and very time-consuming (Witte et al., 2016). A different approach is to improve quantifier tools to be robust instead of different levels of data quality and presence of rain, which shifts the focus in improving the droplet classifier method.

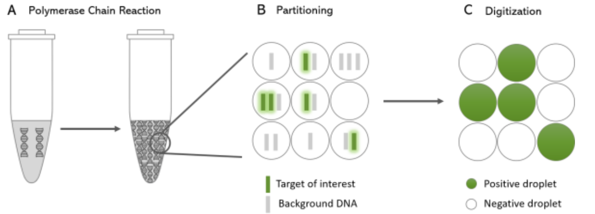

Fig. 1 dPCR workflow summary

One commonly used quantification tool is the Bio-Rad QuantaSoft system. However, despite its popularity, it is expensive to acquire, its threshold determination algorithm is undisclosed, and there have been reports that it gives imprecise estimates or fails to produce output for nonoptimal dPCR assays (Witte et al., 2016). In contrast, the following methods are freely accessible and available as an R-script. First is Cloudy (Lievens et al., 2016), it classifies droplets by first finding fluorescence populations, iteratively estimating their parameters using a combination of linear and non-linear modeling, then it classifies droplets as negative if it is below $\hat{\mu} + 1.5\alpha \cdot \sigma$ of the leftmost population, where $\hat{\mu}$ is the median and α is an in-house derived formula. The second one is Umbrella (Jacobs et al., 2017), which takes a more inferential approach to classify droplets through the use of model-based clustering. It fits a two-component mixture model (one component each for positive and negative populations) by assuming that the negative fluorescence population is approximately equal to the No-Template-Control (NTC) sample (prepared to contain no targets, thereby only producing a negative population), and classifies droplets as negative if its probability to belong to the negative population is above 0.8. And lastly, the tool ddpcRquant (Trypsteen et al., 2015) also takes advantage of NTC samples to determine a threshold value for classifying negative droplets. It fits one hundred extreme value distribution on NTC subsamples and defines the final threshold as the average of all the one hundred 0.995 quantiles.

To improve the robustness of droplet classifiers, the features listed in Table 1 may be desirable. First, it should be required that baseline shifts be considered since it has been observed in studies (Trypsteen et al., 2015). Second, including rain as part of the classification allows for quality checking, wherein a large amount should alert the researcher to re-examine the prepared sample. In addition, multiple rain populations may exist where the researcher may decide it these are primer dimers or target DNA with low amplification. Finally, it may be desirable to classify droplets based on probability as it allows flexibility in setting the threshold. To satisfy all these features, we propose classifying dPCR droplets using Expectation-Maximization (EM) clustering. This method has many applications such as in profiling and clustering (Li et al., 2018).

Table 1. Comparison of features

Features	Cloudy	ddpcRquant	Umbrella	EM Clustering
1. Considers baseline shifts	✓	✓	✓	✓
2. Allows for "rain" classification	✓		✓	✓
3. Allows for classifying more than 3 populations				✓
4. Allows probability-based classification			✓	✓

2. METHODOLOGY

2.1 Target Quantification

To be able to quantify the target concentration in a dPCR assay, it is assumed that these properties are followed: (1) Target molecules are homogeneous in a sample and are distributed randomly in partitions of equal volume, (2) At least one target molecule in a partition is necessary and sufficient for a positive signal, and (3) Target molecules are independent in a sense that there is no interaction with one another or on device surfaces (Kreutz et al., 2011). The Poisson distribution is used to estimate the mean target copies per partition as

$$\lambda = -\ln\left(\frac{N_{neg}}{N_{total}}\right),$$

where N_{neg} and N_{total} are the classified negative and total droplet counts, respectively.

2.2 Data

2.2.1 Real Dataset

A dPCR fluorescence dataset was obtained from a research study of Lievens et al. (2016) Their experiment aimed to study the design factors that would optimize the efficiency of the dPCR amplification for 12 DNA targets. In their definition, an optimized assay produces only two fluorescence populations, has a distant separation between positives and negatives, and low presence of rain. In their study, nine plates were prepared, where each plate controls for a different experimental factor. It has been observed that some factors have worsened, improved, or had no effect on the quality of the data. In one case, the assay with target M88017 in the control group has a high presence of rain, but by applying sonication, the rain droplets have been significantly reduced. On another plate, higher annealing temperature resulted in more rain droplets, and some samples even produced more than two populations. Because of the varying quality produced in this dataset, it is of interest to see how precise the droplet classification methods are for all plates and DNA targets. In addition to precision, the linearity of a method's estimates can be measured in two DNA targets where a serial dilution series was performed in one plate.

Since this real dataset does not have NTC samples, only Cloudy and Umbrella are included in the evaluation in section 3. It is also noted that Cloudy is the method developed by the authors of this dataset and has been used for their study. Thus, it is ideal to at least match the performance of Cloudy.

2.2.2 Simulated Dataset

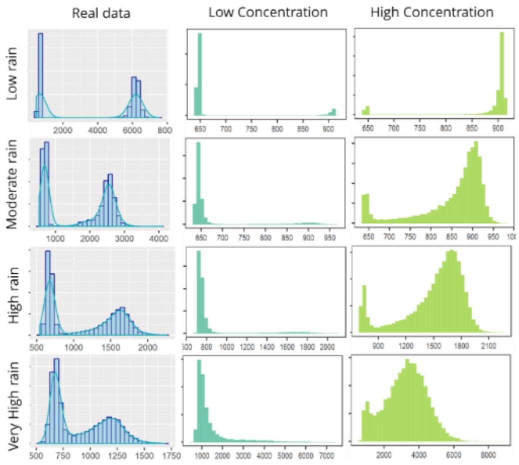

Fig. 2 Samples in the simulated dataset modeled from real samples

For this study, a dataset was simulated to assess the performance of droplet classifiers in extreme cases of dPCR assay quality. Four rain categories were considered: low, moderate, high, and very high rain. For each of these categories, a reference sample in the real dataset was chosen to be modeled by a generalized hyperbolic mixture model. This model is known to have a superior fit in skewed and heavy-tailed distributions (Browne and McNicholas, 2015), which is exhibited by the reference samples in Figure 2. The R package used here is called "MixGHD" (Tortora et al., 2019) where the only input required was the vector of droplet fluorescence for each sample and was fitted under the default settings. After modeling these four reference samples, 5 concentration levels were set ($\lambda = 0.1, 0.2236, 0.5, 1.118, 2.5$) and 15 replicates were generated for each combination of the rain and concentration settings. A sample for the lowest and highest concentration is displayed in Figure 2.

2.3 Performance Evaluation

To evaluate the performance of droplet classifiers, their precision, accuracy, and linearity will be assessed. The precision is measured by the coefficient of variation ($\mathrm{CV} = \frac{\mathrm{sd}(\lambda)}{\mathrm{mean}(\lambda)} \times 100\%$) as it is frequently used in quantitative assay studies due to its attractive property of being unitless. Accuracy is measured with percentage error ($\left|\frac{\mathrm{Actual} - \mathrm{Estimate}}{\mathrm{Actual}}\right| \times 100\%$), where its overall average is called the mean average percent error (MAPE). Finally, when samples are prepared with a geometric series of dilution factor, D_i, the following log-log regression model can be fitted

$$-\log(\lambda_i) = -\log\left(c_1 \times \frac{V_{drp}}{1000}\right) - \log(D_i)\beta_1 \ ; \qquad \text{(Eq. 1)}$$

where the intercept is usually used to estimate the stock target concentration c_1, using the droplet volume V_{drp} in nL. Using this model, the coefficient of determination (R^2) can be used to measure the linearity of λ from a dilution series.

3. RESULTS AND DISCUSSION

3.1 EM Mixture Model Fitting

Although the Gaussian mixture model is a popular method in selecting mixture models, studies show that fluorescence populations do not exhibit a normal distribution (Trypsteen et al., 2015). Thus, in this study, the T-mixture model (EM-T) was chosen due to the observed fat tails in fluorescence populations. In addition, we explore the skewed T-mixture model (EM-skewT) that adds a skew parameter to EM-T in an attempt to fit the heavy skews exhibited by fluorescence populations. The R package “EMMIXskew” was used to run these models (Wang, Ng, & McLachlan, 2013).

To perform mixture model fitting using EM, the number of components G must be determined. A peak finding algorithm was used where G was set as the number of peaks. These peaks are also set as the initial mean μ for each component; while the other initial parameters were set constant for all samples ($\pi = \frac{1}{G}, \sigma = 1000, \mathrm{df} = 30, \delta = 0$; where δ is the skew parameter for EM-skewT). The peaks detected and resulting mixture model fits using EM-T and EM-skewT is shown in Figure 2, the line in the first row represents the peaks, while in the following plots, the lines are the threshold that separates the populations. Because the peak finding algorithm is performed independently for all samples, the challenge of baseline shift between samples is resolved. Any components between the left- and right-most populations will be considered as rain populations.

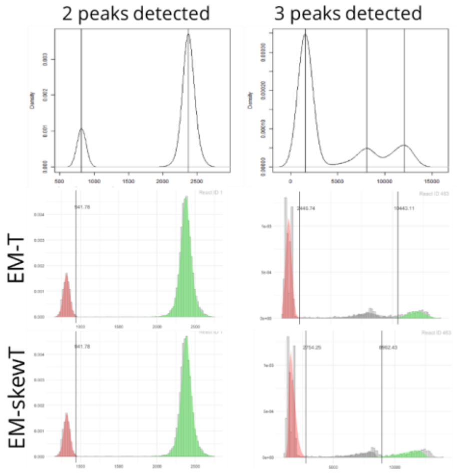

Fig. 2. Peaks and mixture model results

3.2 EM-skewT is the most precise in Real data

The real dataset consists of experimental factors of different target DNAs that produce 92 groups, and thus CVs, in total. To summarize this information, the distribution plot is shown in Figure 3. It can be seen that the methods, EM-skewT, Cloudy, and EM-T, have similar distributions and almost all CVs are within the acceptable imprecision of 25% (Vynck, Vandesompele, & Thas, 2017). In terms of the overall average, EM-skewT is the most precise.

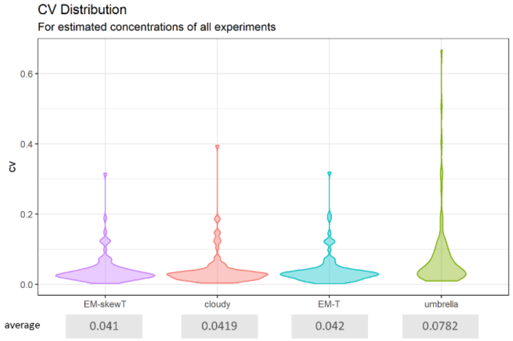

Fig. 3. Distribution of 92 CVs in the real data

3.3 EM methods have high linearity

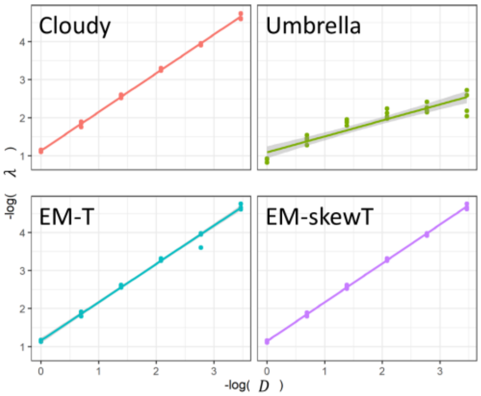

Fig. 4. Log-log regression fits for Target 1507

Table 2. Linearity and model fit estimates

Target TC1507				
	R^2	RSE	Intercept	Slope
EM-skewT	0.9990	0.0404	1.1468	1.0219
Cloudy	0.9987	0.0448	1.1361	1.0183
EM-T	0.9950	0.0890	1.1649	1.0065
Umbrella	0.8607	0.2084	1.0934	0.4190
Target M88017				
	R^2	RSE	Intercept	Slope
Cloudy	0.8800	0.4128	1.4518	0.9041
EM-skewT	0.8785	0.4134	1.4868	0.5207
EM-T	0.8767	0.4154	1.4916	0.8961
Umbrella	0.7647	0.3571	1.4770	0.8991

Two DNA targets (M88017 and TC1507) in the real dataset were prepared in a 6-step serial dilution. Then using the model in Equation 1, the linearity of λ can be measured from the resulting R^2. The results for target TC1507 are shown in Figure 4. From visual inspection, the model fits of EM-skewT, EM-T and Cloudy are almost identical, but a closer view of the model estimates in Table 2 shows that EM-skewT has the highest $R^2 = 0.9990$. Meanwhile, for target M88017, Cloudy has the highest $R^2 = 0.8800$, and EM-skewT follows with a slightly lower $R^2 = 0.8785$.

3.4 EM methods have the highest precision in Simulated data

Unlike in the real dataset where there was not much difference in the CVs of Cloudy and the EM methods, the high presence of artificial rain in the simulated data produced varying CV performances for all the methods. The most precise methods with an overall average CV of 0.0177 and 0.0214 are EM-T and EM-skewT respectively; this is followed by Cloudy, ddpcRquant, and Umbrella, with a CV of 0.0355, 0.0396, and 0.2072. In addition to the average CV, Figure 5 below provides an inspection if a method's precision is affected by rain (low, moderate, high, very high) and true concentration (0.1, 0.2236, 0.5, 1.118, 2.5), where a regression line is plotted to visualize its relationship with CV.

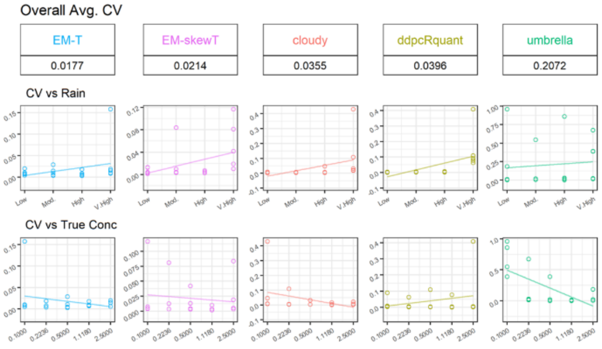

Fig. 5. Influence of rain and true concentration setting on each method's precision

It can be seen that the slope of this line is relatively low in EM-T for all rain and true concentration settings. The very high rain setting proved to be difficult for EM-skewT, Cloudy, and ddpcRquant, which increased its slope. On the other hand, the lowest concentration setting challenged the precision of EM-T, EM-skewT, Cloudy, and Umbrella. Although it seems that some CVs of the EM methods are slightly affected by these factors, all of its CVs are within the acceptable imprecision of 25%.

3.5 EM-skewT has good accuracy in Simulated data

In terms of accuracy, Cloudy has the lowest MAPE of 6.5873%, shortly followed by EM-skewT with 7.262%. These are followed by EM-T, ddpcRquant, and Umbrella, with a MAPE of 14.2557%, 23.2253%, and 141.6648%. It is then of interest to see if the percentage errors are influenced by rain and true concentration settings, as shown in Figure 6. Similar to findings in the CV assessment, the very high rain and lowest concentration setting generally increased the percentage error for all methods. However, although there is a decrease in accuracy, most of the percentage errors in Cloudy were within the acceptable bias of 25% (Vynck, Vandesompele, & Thas, 2017); the same can be said to EM-skewT, except for some challenging samples.

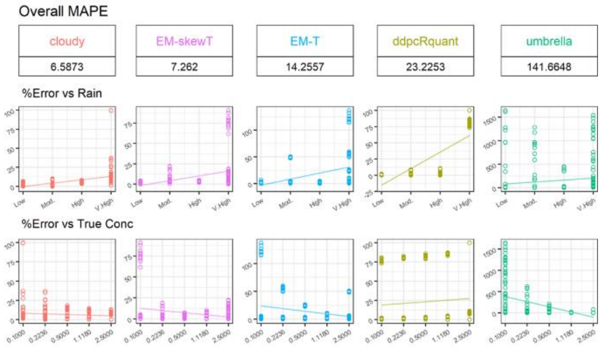

Fig. 6. Influence of rain and true concentration setting on each method's accuracy

4. CONCLUSION

This paper demonstrates the feasibility of the application of EM clustering in dPCR droplet data. This is done by determining the components using a peak finder algorithm and setting a constant initial parameter set for all samples. The T- and Skewed-T mixture models (abbreviated as EM-T and EM-skewT respectively) were explored for this study to classify droplets as positive or negative, whereas the generalized hyperbolic mixture model was used for modeling the fat-tailed and heavily skewed shapes of the real samples and is used for simulating a dataset.

When compared against other methods, the estimates and CVs of EM-T and EM-skewT were very close to Cloudy's in the real dataset. Whereas in the simulated dataset, Cloudy was the most accurate method (lowest MAPE), and EM-skewT was the most precise (low CVs for all rain and true concentration). EM-T and ddpcRquant also performed very well but only struggled in the highest rain setting. Umbrella was performing poorly for all low concentrations.

For general use, most of these methods may be used for dPCR quantification. But for the case of poor-quality assay data, the EM methods may be considered, especially EM-skewT for its high precision and accuracy. However, since the data used in this study is limited, further study is needed to assess the performance of the EM methods for different applications such as low copy target. The R package used in this study is publicly available, and its installation guideline and user manual are accessible in this link (https://zeroh729.github.io/popPCR).

5. ACKNOWLEDGMENT

This work is supported in part by the Philippine Commission on Higher Education (CHED) through the Philippine California Advanced Research Institutes (PCARI) project IHITM 2016-09. The authors would like to extend appreciation to Dr. Wilfred Espulgar and Dr. Glenn Oyong for providing technical validation.

6. REFERENCES

Browne, R. P., & McNicholas, P. D. (2015). A mixture of generalized hyperbolic distributions. Canadian Journal of Statistics, 43(2), 176–198.

Cao, L., Cui, X., Hu, J., Li, Z., Choi, J. R., Yang, Q., … Xu, F. (2017). Advances in digital polymerase chain reaction (dPCR) and its emerging biomedical applications. Biosensors and Bioelectronics, 90 (November 2018), 459–474.

Jacobs, B. K. M., Goetghebeur, E., Vandesompele, J., De Ganck, A., Nijs, N., Beckers, A., … Clement, L. (2017). Model-Based Classification for Digital PCR: Your Umbrella for Rain. Analytical Chemistry, 89(8), 4461–4467.

Li, T., Shao, Y., Fu, L., Xie, Y., Zhu, L., Sun, W., … Guo, J. (2018). Plasma circular RNA profiling of patients with gastric cancer and their droplet digital RT-PCR detection. Journal of Molecular Medicine, 96(1), 85–96.

Lievens, A., Jacchia, S., Kagkli, D., Savini, C., & Querci, M. (2016). Measuring Digital PCR Quality: Performance Parameters and Their Optimization. PloS One, 11(5), e0153317.

Tortora, C., ElSherbiny, A., Browne, R. P., Franczak, B. C., McNicholas, P. D., & Amos., D. D. (2019). MixGHD: Model Based Clustering, Classification and Discriminant Analysis Using the Mixture of Generalized Hyperbolic Distributions. Retrieved from https://cran.r-project.org/package=MixGHD

Trypsteen, W., Vynck, M., de Neve, J., Bonczkowski, P., Kiselinova, M., Malatinkova, E., … de Spiegelaere, W. (2015). ddpcRquant: threshold determination for single channel droplet digital PCR experiments. Analytical and Bioanalytical Chemistry, 407(19), 5827–5834.

Vynck, M., Vandesompele, J., & Thas, O. (2017). Quality control of digital PCR assays and platforms. Analytical and Bioanalytical Chemistry, 409(25), 5919–5931.

Wang, K., Ng, A., & McLachlan, G. (2013). The EM Algorithm and Skew Distribution Mixture. Retrieved from https://cran.r-project.org/package=EMMIXskew

Witte, A. K., Mester, P., Fister, S., Witte, M., Schoder, D., & Rossmanith, P. (2016). A Systematic Investigation of Parameters Influencing Droplet Rain in the Listeria monocytogenes prfA Assay - Reduction of Ambiguous Results in ddPCR. PLOS ONE, 11(12). https://doi.org/10.1371/journal.pone.0168179

Wong, Y. K., Tsang, H. F., Xue, V. W., Chan, C. M., Au, T. C., Cho, W. C., … Wong, S. C. (2017). Applications of digital PCR in precision medicine. Expert Review of Precision Medicine and Drug Development, 2(3), 177–186.

Xu, M., Wang, D., Wang, H., Zhang, X., Liang, T., Dai, J., … Yu, X. (2020). COVID-19 diagnostic testing: Technology perspective. Clinical and Translational Medicine, 10(4), 1–15.