Model-Based Clustering of Digital PCR Droplets using Expectation Maximization

A Thesis Proposal
Presented to
the Faculty of the College of Science
De La Salle University-Manila

In Partial Fulfillment
of the Requirements for the Degree of
Master of Science in Statistics

by

GUIAO, Joyce Emlyn B.

Frumencio F. Co
Adviser

October 18, 2020

Abstract

The digital PCR (dPCR) is an emerging technology to quantify the DNA copies of known strains related to diseases. Currently, the dPCR methodology is being further researched and improved to surpass its accuracy over the gold standard real-time qPCR. One area of study in dPCR is its "digitization" step in which droplets are classified as positives or negatives. This thesis reviews the current droplet classification methods of single-channel dPCR quantification and proposes the Expectation-Maximization Clustering method in aims to improve the accuracy of the final estimated DNA concentration.

Keywords: Quantitative PCR, Droplet Digital PCR, Expectation Maximization Clustering

Contents

List of Figures

List of Tables

Chapter 1

Research Description

1.1 Introduction

1.2 Background of the Study

Detection of target molecules found in deoxyribonucleic acid (DNA) or ribonucleic acid (RNA) is a rapidly growing field of study in the interest of quantifying gene expression levels (Huggett, O'Grady, & Bustin, 2015). These nucleic acid strands carry genetic information and are used as biomarkers for the detection of diseases. Additionally, along with the rise of bioinformatics tools, quantification methods are also utilized in rare mutation detection, copy number variation detection, single-cell gene and microRNA expression analysis, and next-generation sequencing (Quan, Sauzade, & Brouzes, 2018). Outside the scope of molecular biology, its application has also found its way in forensic research (Whale, Cowen, Foy, & Huggett, 2013), medical diagnosis, environmental monitoring, and food safety analysis (Cao et al., 2017).

Before target molecules can be quantified, the challenge of detecting these microscopic targets should be addressed first. In a DNA sample, the relative concentration of target DNA is undetectably low. One solution to this problem is to amplify the DNA sequences using polymerase chain reaction (PCR), a widely-used method for nucleic acid amplification since its invention in the 1980s (Cao et al., 2017). PCR exponentially multiplies specific targets in the DNA or cDNA strands into millions to billions of copies (Figure 1.1 (A)). The dPCR sample preparation consists of mixing the DNA sample with chemical components that will encourage the amplification, such as buffers, Taq polymerase, and intercalating dyes or

fluorescence probes. This blend of chemicals, called the reaction mix, is pipetted into an assay containing equal sized reaction wells, thereby partitioning the reaction mix into thousands of droplets (Figure 1.1 (B)). The assay is then exposed in a series of 20 to 40 thermal cycles. In each cycle, the process of denaturation, annealing, and extension are performed which doubles the target DNA copy – theoretically producing 2^n molecules after n cycles (Quan et al., 2018). A droplet that contains at least one target DNA copy will emit high fluorescence intensity; each droplet is then "digitized", or classified binarily as "positive" or "negative", based on its fluorescence amplitude (Figure 1.1 (C)).

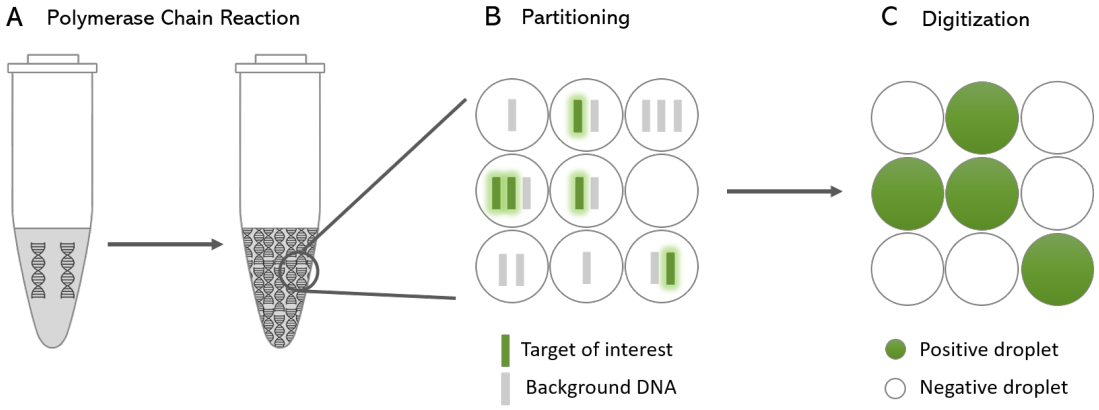

Figure 1.1: Digital PCR Quantification Steps

Since the earliest published dPCR experiment from 1988 (Saiki et al.), the advances of nanofluidic technology in biomedical instruments have continuously pushed the limits of dPCR. An increasing number of researchers have found dPCR to be reaching, and even outperforming, the precision, sensitivity, and reproducibility of real-time qPCR, which is the gold standard for molecular quantification (D. F. Chen, Zhang, Tan, & Jing, 2018; Persson, Eriksson, Lowther, Ellström, & Simonsson, 2018; Taylor, Laperriere, & Germain, 2017; Arvia et al., 2017; Blaya, Lloret, Santísima-Trinidad, Ros, & Pascual, 2016; G. M. Jones et al., 2016; Sanders et al., 2011). The nature of dPCR allows it to standardize quantitation as opposed to qPCR's use of a reference curve, resistant to inhibition, and less negatively influenced by the target sequence variability (Hall Sedlak & Jerome, 2014). The unraveling superiority of dPCR makes it a necessity for research experiments that require intensive accuracy, such as the certification or stability studies of reference materials.

Despite the optimistic performance of dPCR, several challenges are met before being able to reach the optimal results from dPCR. One of the final steps in dPCR is the threshold determination that separates the endpoint fluorescence of the dPCR assay into positives and negatives. This threshold is not constant

for all DNA samples and is unclear for assays that produce ambiguous readouts (Trypsteen et al., 2015).

An important aspect of positioning the threshold is the presence of noise features in the data. Poor quality dPCR assays add to the ambiguity of fluorescence signals that contribute to the difficulty of threshold determination. Due to the emerging demand for dPCR data analysis, Lievens, Jacchia, Kagkli, Savini, and Querci (2016) has determined a set of method performance criteria to assess the quality of a dPCR assay run. Their following criteria aims to measure the efficiency of the separation between positive and negative droplets: (i) there should only be two fluorescence populations, or in other terms, a single amplification product; (ii) there should be a good separation between positives and negatives measured in peak resolution; and (iii) there should be very minimal amounts of intermediate fluorescence, also called as 'rain'. The factors affecting these noise characteristics are further explored in the succeeding sections.

"Rain" is the term used in several studies to describe droplets that emits a fluorescence intensity settling in between the positive and negative populations (Lievens et al., 2016; Trypsteen et al., 2015; Witte et al., 2016; Dreo et al., 2014; Brink, Meskas, & Brinkman, 2018; Attali, Bidshahri, Haynes, & Bryan, 2016). Figures 1.2 and 1.3 demonstrate two different DNA targets with the former showing a visually clearer distinction of the positive and negative population than the latter, of which is possessing multiple rain droplets. The data used in these figures are sourced from the dataset made publicly available by Lievens et al. (2016). Having more than two fluorescence population is also a problem, since it is not known whether these droplets do contain the target DNA or are just fluorescence residue. Poor quality dPCR assays negatively limit the level of sensitivity and accuracy it may reach. Among the consequences of unoptimized dPCR assays is droplet misclassification, which may lead to serious misdiagnoses.

Figure 1.2: Fluorescence readings of 4 repititions of DNA target cru

Figure 1.3: Fluorescence readings of 4 repititions of DNA target TC1507

Estimating the DNA target concentration is based on the classification of positive and negative droplets. When assays produce two distant fluorescence populations, most quantification tools can estimate target concentrations with high sensitivity. As demonstrated in an optimized E. amylovora experiment (Dreo et al., 2014), slight differences of thresholds calculated from different tools had little effect on the final estimated concentration. However, for R. solanacearum, which is observed to manifest false-positive signals in qPCR experiments, produced unsatisfactory analytical sensitivity of the concentration estimates. The danger of low sensitivity is expounded at the clinical level, where inaccurate target quantification leads to the misdiagnosis of patients (Tzonev, 2018). One such case is the prenatal screening test for Down Syndrome; this test is expected to mostly result in normal pregnancies. However, many pregnancies are still falsely reported as positive for Down Syndrome. False negatives also risk the overall health of the patient that truly possesses the genetic disorder.

The whole dPCR workflow introduces multiple entry points for technical error, which may add "noise" in the amplified fluorescence. As illustrated in Figure 1.4, the workflow is usually a sequential procedure of extracting from a biological sample, preparing and partitioning the reaction mix, amplifying the assay, detecting the target molecules, and then finally, the target concentration is estimated using a Poisson correction factor. Jacobs, Goetghebeur, and Clement (2014) emphasized that every step of the dPCR workflow inevitably allows for the introduction of different sources of variation shown in Figure 1.4. Operator-specific and repeatability variations are caused by pipetting skills, aliquot mismeasurements, and sample preparation time (when prolonged, probe degradation occurs which affects the quality of emitted fluorescence). Machine-specific variation may be caused by unequal partition sizes and possible partition loss. Finally, even if the former variations were managed, sample assay replicates such as in Figure 1.4. B,

will still produce varying target concentration estimates due to sampling errors. These errors are caused by 1.) sub-sampling only a portion of an organism and 2.) uneven counts of target copies distributed in the partitions. In addition to operator and machine variations, sample preparation in itself is a challenge as the optimal parameters in designing the reaction mix need to be identified. These parameters or experimental factors in Figure 1.5. A affect the dPCR amplification efficiency. When these experimental factors are not optimal, assays generate ambiguous fluorescence readouts as shown in Figure 1.5. B. It is not clear where to set the threshold that separates positive and negative droplets, in these scenarios, droplet misclassification error is prevalent.

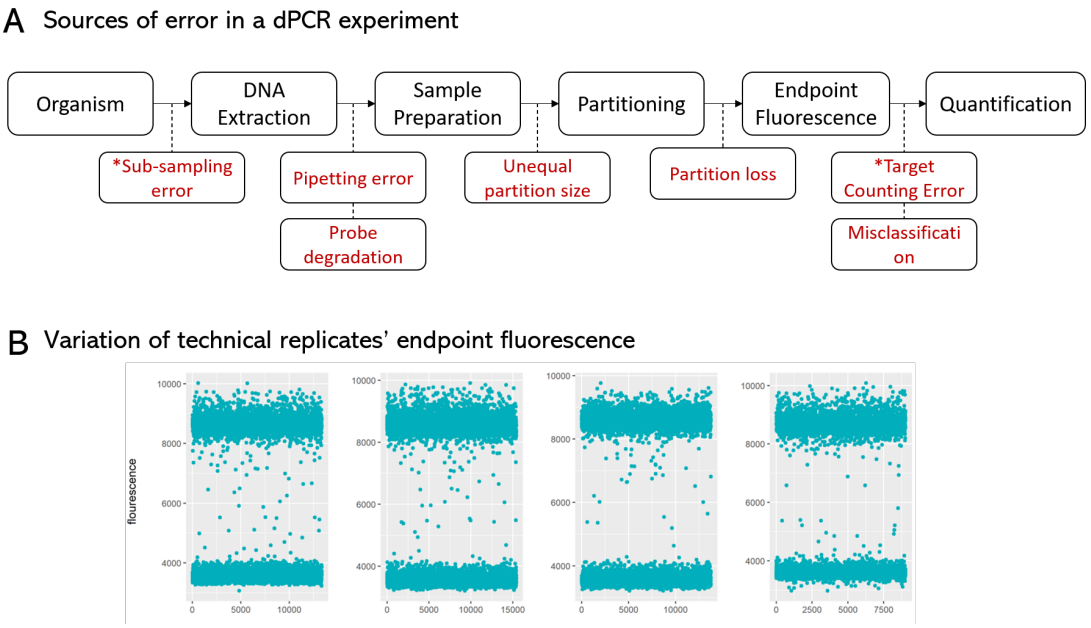

Figure 1.4: Sources of variation in the dPCR workflow- (A) Multiple sources of error can be introduced for each step in the dPCR workflow. Items with '*' indicate sampling error). (B) In effect, sample replicates obtained from the same organism will inevitably have variation in its fluorescence readouts.

Different dPCR systems do not share a common default setting and thermal profiles. Each one is a factor that has to be optimized depending on the target molecule. Increasing the number of cycles has shown to affect the amplification of dPCR droplets that increases the separation of the two populations (Köppel & Bucher, 2015). Temperature gradients are frequently performed to find the most favorable setting to reduce rain (Gerdes, Iwobi, Busch, & Pecoraro, 2016). However, the optimized parameters to improve the quality of a target's dPCR assay may not work for another target. As shown in the experiment of Witte et al. (2016), parameters that increased the efficiency for prfA did not work for

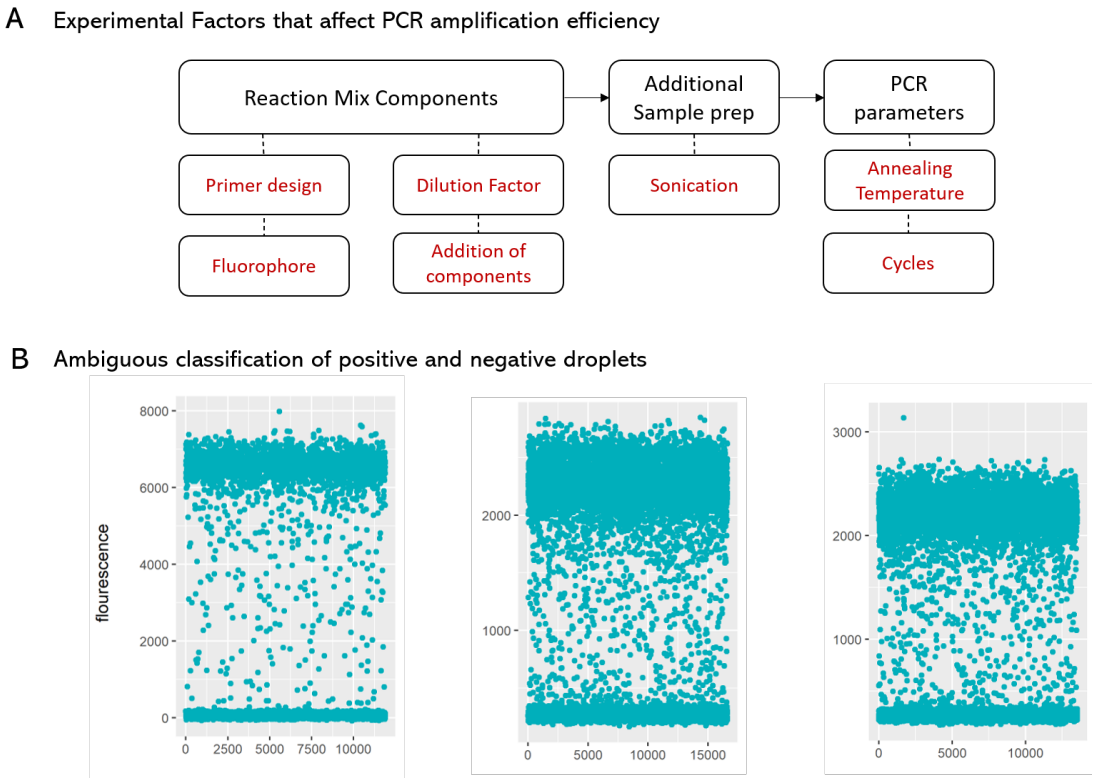

Figure 1.5: Experimental factors that affect dPCR amplification- (A) List of parameters that may be controlled to optimize dPCR amplification (B) Unoptimized assays produce ambiguous readouts, making it difficult to determine a clear demarcation line between the positive and negative droplets

ΔprfA. Besides controllable settings, different dPCR platforms also have been revealed to deviate from its claimed volume (Pinheiro et al., 2012; Dong et al., 2015; Corbisier et al., 2015; Dagata, Farkas, & Kramer, 2016; Košir et al., 2017). These discrepancies have been observed in the Bio-Rad QX100/QX200 platforms and the RainDrop platform. Unequal partition volumes may produce suboptimal PCR amplification that contributes to increased rain droplets.

In addition to physical variation, chemical and biological factors play a role in the dPCR assay quality, such as target sequence variation, amount of polymerase, MgCI2, dNTPs, and primers (Köppel & Bucher, 2015; Kramer & Coen, 2001), dye or probe quencher (Witte et al., 2016), fluorophore used (Gerdes et al., 2016), inhibition, delayed reactions, primer depletion, and other biological factors (Jacobs et al., 2014).

Sampling variation stems from the fact that only a small sample of the organism is extracted; and although there is an expected number of target molecules

per liters of a sample, drawing equally sized samples will result in different target molecules that are more or less near the average. Tzonev (2018) demonstrates the number of target molecules that can be drawn from extraction is distributed as Poisson. Besides the sampling error, samples may also exhibit imperfections, and thus have inhibited amplification. Each variance component accumulates to the bias and variance of the final estimated target concentration, and thus, this gives rise to the importance of providing solutions that would increase precision in every step. To increase the sensitivity and specificity of the estimate, the misclassification of droplet partition should be minimized as much as possible. A high presence of false-negative droplets reduces sensitivity, while specificity is lowered for high false-positive counts.

The factor directly causing noisy readouts is very difficult to pinpoint. In case of failure to optimize the design parameters, other hands-on approaches may be taken, such as running a real-time qPCR experiment, running PCR solution in gel electrophoresis, or performing dilution series in the cost of additional labor. However, preparing replicate samples are prone to pipetting and operator errors. On the other hand, the problem may also be alleviated using statistical approaches. For droplet volume variability, that means a correction-factor must be taken into account to improve the agreement of estimates. In the case of unoptimized assays, statistical methods can help automatically determine the threshold to separate the positive and negative droplets and also eliminate manual operator bias. Based on the experience of Demeke and Dobnik (2018), they were able to produce precise estimates from unoptimized assays due to statistically-based thresholds. When dealing with rain, some researchers exclude it in the final droplet counts (M. Jones et al., 2014), but this option is said to produce underestimated concentrations if rain were actually suboptimal PCR reactions; instead, the threshold algorithm should be improved (Trypsteen et al., 2015). Based on expert opinion, rain droplets may also be caused by primer dimers, which are primers that annealed together instead of the target DNA. Because of the different interpretations of rain droplets, adjusting the threshold should be a feature available for droplet classifier systems.

Expounding further on the automated threshold setting approach, current algorithms can be improved and should be robust to baseline shifts, rain droplets, multiple populations, and poor separation of populations. The baseline fluorescence of the negative population has been observed, but even the popularly used QuantaSoft systems do not take this into account (Trypsteen et al., 2015). Thus, discrepancies may occur in the number of positive droplets. The threshold setting problem may be seen as a droplet classification problem. Reducing the misclassification while being robust to different data characteristics increases the reliability of the system. There are currently many areas of improvement in the calculation

	High threshold							Low threshold						
Input	0	1.5 x 10^1	1.5 x 10^2	1.5 x 10^3	1.5 x 10^4	1.5 x 10^5	1.5 x 10^6	0	1.5 x 10^1	1.5 x 10^2	1.5 x 10^3	1.5 x 10^4	1.5 x 10^5	1.5 x 10^6
Total	17720	18188	18873	17583	16296	13310	16536	17720	18188	18873	17583	16296	13310	16536
Positive	0	17	160	1226	8952	11017	14955	0	20	173	1394	10272	12320	15763
Negative	17720	18171	18713	16359	7380	2318	1590	17719	18167	18699	16172	5973	940	722
Difference of positives	0	3	13	168	1320	1303	808							
% (difference to low threshold)	0	17	8	14	15	12	5							

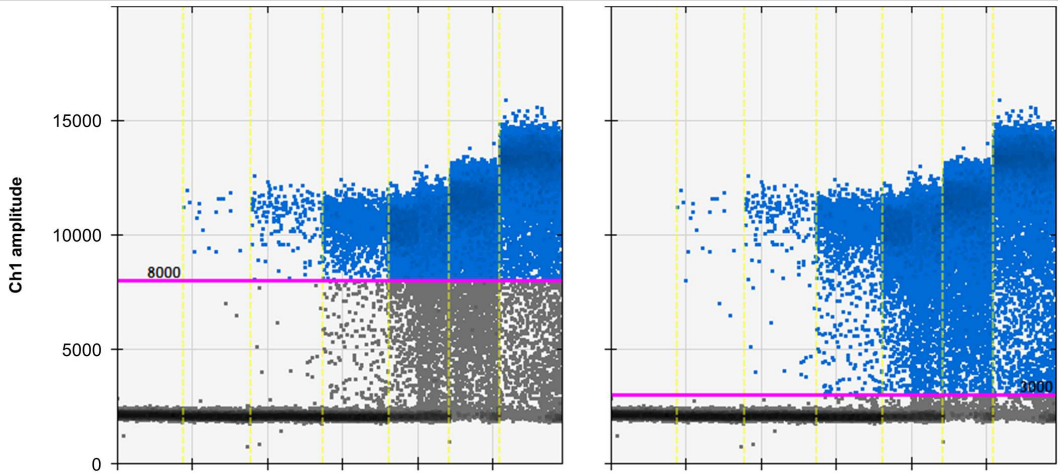

Figure 1.6: QuantaSoft High and Low Threshold Settings- Reprinted from "Systematic Investigation of Parameters Influencing Droplet Rain in the Listeria monocytogenes prfA Assay - Reduction of Ambiguous Results in ddPCR," by Witte, A. K., Mester, P., Fister, S., Witte, M., Schoder, D., & Rossmanith, P., 2016, A PLOS ONE, 11(12), https://doi.org/10.1371/journal.pone.0168179

of thresholds especially in a high presence of intermediate fluorescence (Demeke & Dobnik, 2018). In the report of Witte et al. (2016), it can be seen that the QuantaSoft's high and low threshold settings give conflicting positive droplet counts. As noted by several studies, dPCR assay quality is often traded with time. Even when an optimal setting is determined, it may be time-consuming to run the optimal parameters (Witte et al., 2016), such as when increasing cycles (Lievens et al., 2016), thermal profile variations of thermocyclers (Young, Yang, Bae, & Park, 2008). The advantage of a robust droplet classifier is then the reduction of the negative impact upon compromising quality with time.

1.3 Statement of the Problem

This thesis aims to classify dPCR droplet partitions into positive or negative by exploring Model-Based clustering via the Expectation-Maximization method. The specific objectives of this study are to:

1. fit G-component mixture models on dPCR droplet fluorescence intensities using Expectation Maximization,
2. utilize EM classified droplets to provide precise quantification estimates for DNA samples with varying amounts of "noise" and concentration, and
3. evaluate and compare the performance of the EM classification amongst existing droplet classifier methods.

1.4 Significance of the Study

Quantification of target concentrations for pathogenic bacteria, gene expression of diseases, cancer diagnostic, and other health-related applications strongly demand estimators with high sensitivity and precision, as lives are put on risk for false-positives. A modern approach to DNA target quantification is through the dPCR method. In one of the steps of the dPCR workflow, the classification of droplet fluorescence still has many areas for improvement.

The most prominent problem in classification lies in experiments exhibiting a high frequency of rain, or intermediate fluorescence values. These are experiments that have not yet been optimized. As different DNA target samples exhibit distinct structures (Lievens et al., 2016), an optimized setup for one DNA target may not be applicable for other targets. Additionally, for samples with low concentration, the total count of detected positive droplets dramatically changes the final concentration estimate, due to the greater impact of false positives in the proportion of detected over the number of true positives. The following are some tools and methodologies proposed for droplet classification: Bio-Rad Quantasoft ddPCR software, definetherain (M. Jones et al., 2014), manual global threshold (Dreo et al., 2014), Cloudy (Lievens et al., 2016), and Umbrella (Jacobs et al., 2017). Most of the aforementioned droplet classifier tools rely strongly on how representative reference samples are. According to Dreo et al. (2014), such approaches are sensitive to significant shifts in amplitude for previously unobserved factors, such as cross-reactions or the influence of inhibitors.

In an attempt to prevent the problem of representation, this study will explore the feasibility of estimating target concentrations without a reference sample. Additionally, G-components are considered as to accommodate the possibility of multiple fluorescence populations. The method in this paper uses the concept of iterative parameter estimation from Cloudy and the model-based clustering for the droplet classification from Umbrella. This is both achieved using Expectation-Maximization algorithm. The significance of the study will be useful in quantifying precise concentrations in targets that have not yet been optimized for dPCR experiments and also for quantifying targets of low concentrations.

1.5 Scope and Limitations

This study solely relied on publicly available dPCR datasets from published research papers. Datasets from Lievens et al. (2016) and M. Jones et al. (2014) were found and will be used for performance evaluation. The former dataset contains twelve DNA targets from food and feed samples ran on nine different settings by controlling for experimental factors; the latter dataset is a serial dilution of the Albumin DNA ranging from 10^0 to 10^5 copies.

The droplet classification method in this study uses model-based clustering, or the use of finite mixture models to perform clustering. However, the identification of the distribution of the mixture densities will be dependent on the observed available dataset. As a consequence of the limited dataset, the paper's methodology described here needs more study for other experimental settings and DNA targets.

Chapter 2

Review of Related Literature

2.1 ddPCR Quantification tools

Because of the partitioning nature of droplet dPCR, it is more sensitive in detecting target nucleic acid. However, as mentioned in section 1.2, the dPCR workflow is vulnerable to variance components which are reflected as "noise" in the endpoint fluorescence, leading to droplet misclassification. In this section, droplet classification methods of single-channel dPCR quantification systems are explored to shed light on the current capabilities, limitations, and possible areas of improvement of the methods.

2.1.1 Bio-Rad Quantasoft

The most common method in classifying positive and negative droplets is by enforcing a hard threshold. Generally, all droplets with a fluorescence amplitude greater than this threshold are then classified as positive, and anything lower is negative. One popular tool incorporated with automatic thresholding is the Bio-Rad QuantaSoft software. This software is the dPCR analysis tool that comes with the Bio-Rad droplet dPCR System package. This package includes all the instruments required for running dPCR experiments — from sample preparation to the analysis of the target concentration (Bio-Rad, 2019). According to the Bio-Rad Laboratories website (*Advancing Scientific Discovery and Improving Healthcare for Over 65 Years.*, n.d.), it has been a leading product developer in the research fields of life science and clinical diagnostics for over 65 years. Among its popular focus areas, dPCR is one of its most featured technology, providing ddPCR instru-

ments. kits, reagents, assays, and other consumables. Several studies in hospitals (López et al., 2016; D. F. Chen et al., 2018; Abed, Carbonneau, L'Huillier, Kaiser, & Boivin, 2017; Tagliapietra et al., 2020), public health (Hussain & Bowers, 2017; Nystrand, Ghanima, Waage, & Jonassen, 2018), food safety (J. Chen et al., 2020; Capobianco et al., 2020; Basanisi et al., 2020), and even environmental quality (Hamaguchi et al., 2018; Jahne et al., 2020; Dobnik, Štebih, Blejec, Morisset, & Žel, 2016; Mauvisseau et al., 2019) have found the Bio-Rad QuantaSoft dPCR systems useful for their analyses.

Of all the QuantaSoft software features, the focus of this section is on its threshold setting. By default, QuantaSoft sets an automatic threshold to the single-well or multiple-well amplitude data. As with other automated tools, its documentation recommends that the operator reviews the automated threshold and to make adjustments if needed (e.g. manually setting the threshold is allowed). Unfortunately, the calculation of the automatic threshold is not publicly available.

The evaluation of the QuantaSoft software shows satisfactory results from the food safety study of Basanisi et al. (2020), whereby nine pure meat samples were discriminated with 100% diagnostic accuracy, sensitivity, and specificity. However, upon checking the authenticity of twenty commercially available meat products, twelve samples were said to contain DNA traces of other animals not declared. In the bacteria analysis of Dreo et al. (2014), it was shown that for highly-concentrated samples that exhibited substantial amounts of intermediate fluorescence, the system fails to determine a threshold, where its outputs are "No call". In the case of low bacteria concentrations, droplets near the negative droplets were classified as positives. Similarly, Witte et al. (2016) has observed around 10% positive count differences between low and high threshold settings using the same data. The heavy presence of rain in their assay prevented a clear threshold value. Both these studies noted that the QuantaSoft software requires a well-optimized assay with good discrimination of positive and negative droplets for its threshold to be reliable.

2.1.2 Manual Global Threshold (MTg)

As opposed to the automatic threshold, Dreo et al. (2014) proposes setting a manual global threshold (MTg) determined by no template control (NTC) samples. This takes into consideration that individual assays behave differently, and could require expert intervention. As a standard approach, the threshold for an optimal assay was defined as the NTC mean + 6 standard deviations; on the other hand, a noisy assay had its threshold set above the highest value in NTC samples. It is expected in the latter case that the sensitivity would be lower due to its

high threshold. However, the paper claims that this resulted in high analytical sensitivity for that assay. A major disadvantage in this approach is the lack of a clear definition or guideline in setting the MTg; this consequently will cause reproducibility issues for succeeding experiments and introduce operator bias.

2.1.3 definetherain

Based on the research papers curated by Peterson (2009), K-nearest-neighbor (KNN) is an unsupervised clustering approach that should be among the first methods considered for data with little to no information about its distribution. This clustering method operates on the chosen distance measure — commonly the Euclidean distance — between the observations. Due to its simplicity, data from various fields have applied KNN such as in a movie recommendation system (Ahuja, Solanki, & Nayyar, 2019), climate classification (Shi & Yang, 2020), breast cancer diagnostics (Mittal, Aggarwal, & Mahajan, 2019), among others.

The positive and negative droplet classification can be framed as a clustering problem. An open-source tool developed by M. Jones et al. (2014), called definetherain, utilizes the KNN algorithm in detecting rain droplets. According to their research, they claim that definetherain is accurate in estimating assays with low template numbers, which is particularly applicable in research fields such as the HIV-1 cure research. definetherain follows these steps for classification:

1. Setup a positive control sample of known input copy numbers.
2. Cluster the droplets using kNN with $k = 2$. The cluster on the left is the negative cluster, and on the right is the positive cluster.
3. Observations between the range of the negative cluster's mean + 3 standard deviations and the positive cluster's mean - 3 standard deviations are classified as rain.
4. Rain droplets are not included in the final calculation of the concentration estimate.

Unlike the other methods discussed here, this tool produces two cutoff values — one for each cluster. The droplets falling between these two cutoffs are classified as rain. The determination of these thresholds are solely dependent on the control sample. The disadvantages of the use of a constant threshold for the succeeding target samples is that 1) the control has to be representative of the target, otherwise, concentration estimates would be biased; and 2) the baseline shift of the fluorescence populations are not taken into consideration.

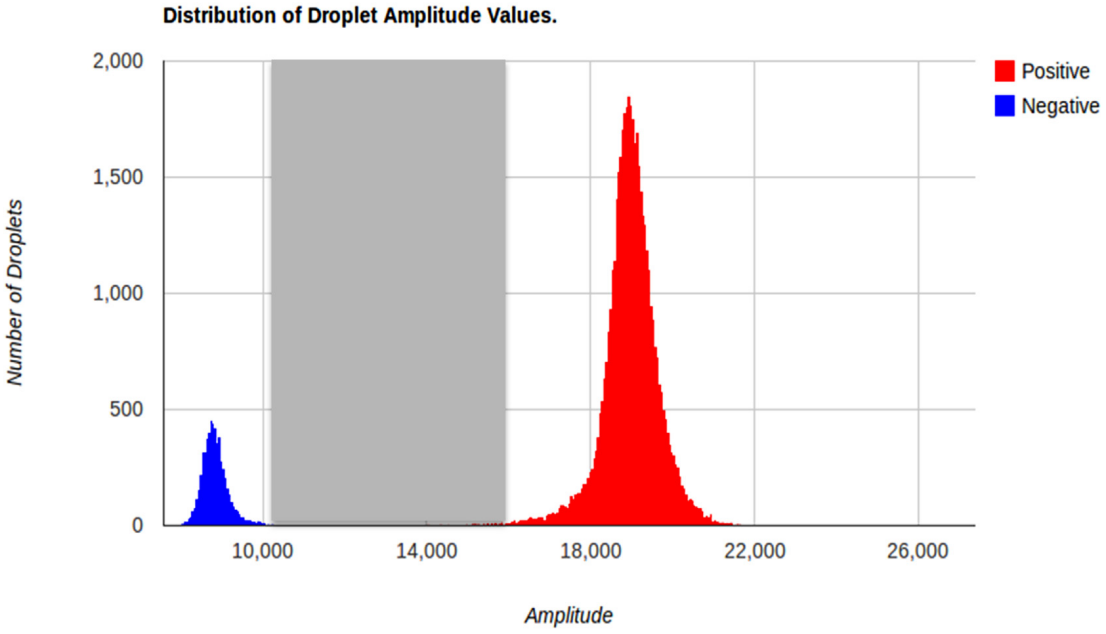

Figure 2.1: definetherain Threshold Setting- Reprinted from "Low copy target detection by Droplet Digital PCR through application of a novel open access bioinformatic pipeline, 'definetherain.'," by Jones, M., Williams, J., Gärtner, K., Phillips, R., Hurst, J., & Frater, J., 2014, Journal of Virological Methods, 202, 46–53. https://doi.org/10.1016/j.jviromet.2014.02.020

2.1.4 Cloudy

The research work of Lievens et al. (2016) has been well cited for its definition of the performance criteria for dPCR assays and experiments on several design parameters. The quantification method used in their experiments was available as a supplementary file named "S3_file.R", which is captioned as their main function to categorize droplets and quantify the concentration. Inside this R code is a function named Cloudy. Although "Cloudy" was not mentioned in their paper, their algorithm will be referred to as Cloudy in this paper.

The Cloudy algorithm first determines the fluorescence populations using density peaks, then iteratively estimates the parameters of each population. The droplet categorization depends on its standard deviation distance from a population's mean estimate. The following list summarizes the cloudy algorithm:

1. The Gaussian kernel density of the fluorescence is estimated with a minimum bandwidth of 50

2. Density peaks are identified using a sliding window approach. The subsequent steps will differ according to if one, two, or more than three peaks were found. But generally, the proceeding steps are followed.

3. For each population found through the peaks, its location and spread is initially estimated using the median $\hat{\mu}$ and standard deviation $\hat{\sigma}$. Assuming normality, $\hat{\sigma}$ is estimated as half the peak width at 60-65% of its maximum height.

4. The iterative procedure starts by initializing $a = 4$.

5. For each population, $\hat{\mu}$ and $\hat{\sigma}$ are re-estimated using only the observations within $\hat{\mu} \pm (a \cdot \hat{\sigma})$.

6. Recalculate $a = 4.55 + 0.35 \cdot \log k + 0.045 \cdot \log k^2$; where k is the kurtosis of the given population

7. Repeat steps 5-6 until stabilization.

8. After stabilizing the estimates for all the population, the last step is different when either including or excluding rain in the final categorization.

   (a) If droplets can be categorized as rain, observations within $\hat{\mu} \pm (a \cdot \hat{\sigma})$ are then classified as members of that population; observations not falling within any population are classified as rain.

   (b) If droplets can only be positive or negative, then any observation below a threshold θ is negative, and positive otherwise; where $\theta = \hat{\mu_{neg}} + 1.5 \cdot a_{neg} + \hat{\sigma_{neg}}$.

In summary, the Cloudy algorithm uses the Gaussian kernel density to detect peaks, which are then considered as populations. Population parameters are then estimated iteratively until convergence. It is worth noting that in its iterative step for estimating the population parameter (step 6), the formula for re-calculating a is based on the analysis of their in-house data, and should be used with caution when implementing for other unobserved DNA targets. After determining the final population parameters, threshold(s) is then calculated for classifying droplets as positive, negative, or optionally, rain. However, the rain classification rule in step 8(a) poses a problem for fluorescence densities that are heavily skewed. The histogram in Figure 2.2 reveals the distribution of a sample's negative droplets are heavily skewed to the right. Using Cloudy to classify the droplets with the rain option, the skewness was not taken into account, which can be observed in the scatterplot, because of the symmetric categorization rule.

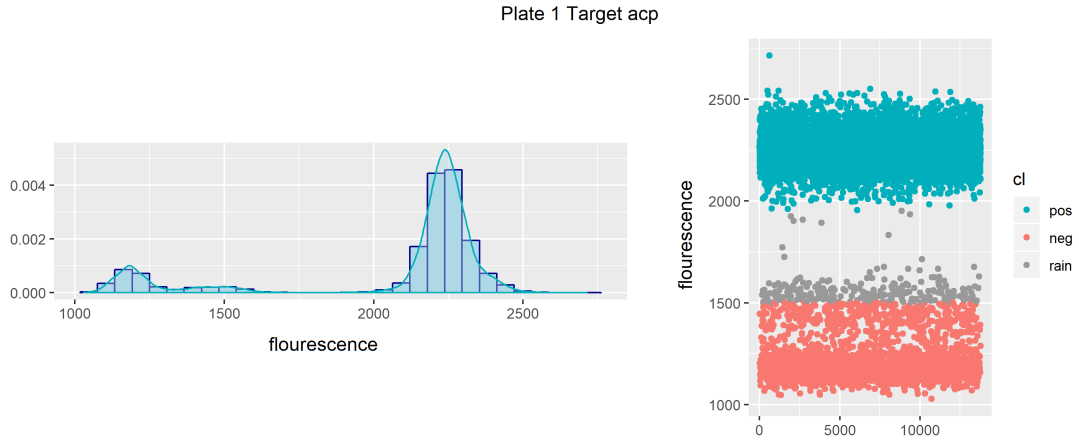

Figure 2.2: One replicate of the DNA target acp Plate 1 from Lievens et al. (2016) dataset. Left panel shows the fluorescence densities. Right panel is the result of droplet categorization using cloudy

In their study, the Cloudy algorithm was able to produce satisfactory estimates for differing PCR experimental factors, such as sonication, PCR enhancers, annealing temperatures, and cycle numbers.

2.1.5 Umbrella

As opposed to the distance-based clustering in Section 2.1.3, a probabilistic approach is achieved using model-based clustering. According to McNicholas (2016), a finite mixture model is a sum of weighted density components. This mixture model has to be appropriate such that its parameters are flexible for fitting the characteristics of the data. In this approach, each unimodal density component is defined as a cluster; and each observation has a calculated probability of it belonging to a cluster.

Umbrella, developed by Jacobs et al. (2017), is a model-based clustering tool for dPCR droplets using nonparametric density estimation. This tool takes in a set of representative NTC sample(s); the procedure then follows a series of assumptions and estimations in deriving the final estimated concentration. An oversimplification of the Umbrella algorithm are as follows:

1. The NTC distribution $f_0(x)$ is assumed to follow a unimodal distribution. The location and variation is estimated by the mode and mean of absolute deviation (MAD), respectively.

2. The fluorescence intensitities x observed in a partition of the target sample A is assumed to have a mixture density

$$f_A(x) = p_{0,A} f_{0,A}(x) + (1 - p_{0,A}) f_{1,A}(x)$$

;

- where $p_{0,A}$ = proportion of negative partitions
- $f_{0,A}(x)$ = densities of the partitions without target copy (null component of the target partition set A)
- $(1 - p_{0,A})$ = proportion of positive partitions
- $f_{1,A}(x)$ = densities of the partitions with target copy

3. These parameters are estimated by first aligning the modes of the null component of $f_A(x)$ and the NTC reference $f_0(x)$.

4. Aligned distributions are discretized by generating a histogram of equal bin widths.

5. The bin counts of the aligned distribution are modeled using a Poisson regression model, resulting in estimates for $\hat{p}_{0,A}$,$\hat{f}_0(x)$, and $\hat{f}_A(x)$.

6. The posterior probability that partition i is void of the target with fluorescence intensity x_i of partition set A, $\hat{p_{0,A}}$, can be defined from the estimated $\hat{p}_{0,A}$ from the previous step as

$$\hat{p}_{i,0,A} = \hat{p}_{0,A} \begin{pmatrix} \hat{f}_{0,A}(x_i) \\ \hat{f}_A(x_i) \end{pmatrix}$$

7. The Umbrella threshold estimator is then determined by the estimated $\hat{p}_{i,0,A}$. For intensity value i, the interpretation of the posterior probabilities are

- $\hat{p}_{i,0,A} > 80\%$ are considered negative partitions with a probability of $\leq 20\%$ to be false negatives
- $\hat{p}_{i,0,A} < 5\%$ are considered positive partitions with a probability of $\leq 5\%$ to be false positives
- $5\% \leq \hat{p}_{i,0,A} \leq 80\%$ are considered as rain

The mode and MAD, which estimates the location and spread for the null NTC distribution, $f_0(x)$, are chosen due to its robustness and insensitivity to skewed tails. Only observations within 10 deviations from mode are included for the null model.

Following the assumption of a mixture density in step 2, unlike most model-based clustering algorithms, Umbrella does not assume normal densities for $f_{1,A}(x)$ and $f_{0,A}(x)$, the partitions with and without the target, respectively. This is due to the exhibition of dPCR fluorescence intensities to be non-normal, as clusters tend to have heavy tails to the left or to the right. The solution for this is the use of non-parametric density estimation in step 3.

After estimating all the components in the mixture model from steps 3 - 5, the component of interest $\hat{p_{0,A}}$ is then used to determine $\hat{p}_{i,0,A}$ in step 6. Finally, this is used as the basis for Umbrella threshold estimator in step 7. It is warned that Umbrella may not be precise in detection experiments for low copy samples, as classifying individual samples is not the strength of this method.

2.1.6 ddpcRquant

The ddpcRquant determines a threshold for negative droplet fluorescence based on extreme value theory. It is available as an R library developed by Trypsteen et al. (2015). The extreme value theory assumes that the maxima distribution of large samples is distributed as a generalized extreme value (GEV), regardless of the original value's distribution. Hence, the extreme value theory is considered as asymptotically nonparametric provided that samples are sufficiently large. Applying this theory in dPCR droplet classification, an extreme value percentile of the merged NTC samples are used to calculate the threshold. The summarized steps of ddpcRquant are described below :

1. The required NTC sample inputs are baseline corrected. This is done by subtracting the fluorescence intensities of each sample using the Robertson-Cryer estimated mode.
2. The fluorescence of all NTC samples are merged, and then randomly assigned to equally sized k groups,
3. Using the maxima of all k groups, the generalized extreme value distribution is fitted by maximum likelihood.
4. The tentative threshold is then the 0.995 percentile of this distribution.
5. The final threshold is the average of all the tentative thresholds upon 100 repeats of steps 2-4.
6. To correct for the baseline of target samples, each sample is subtracted by its fluorescence mode below a cutoff c — calculated as the average of the NTC modes plus the final threshold.

7. Finally, to calculate the target concentration, negative and positive droplets are separated using the final threshold from the baseline-corrected target samples.

The advantages of ddpcRquant are that it doesn't assume any distribution of the droplet fluorescence, and it corrects for baseline shifts. In calculating the target concentration, it does not discard any droplet as they stated that this can underestimate the true concentration.

According to the evaluation of Trypsteen et al. (2015), ddpcRquant is superior to QuantaSoft in regards to having less false positive counts in the NTC samples, as a consequence, QuantaSoft concentration estimates are generally higher as it identifies more positive droplets. The reason was found to be that Quantasoft places its threshold too close to the negative droplet population, which may be caused by not applying baseline correction between NTC and target samples, or by its assumption of NTC being normally distributed without fitting experiments. Additionally, QuantaSoft fails to quantify concentration in some of the authors' samples that result in "No call" outputs.

2.2 Expectation-Maximization (EM) Clustering

Recall from Section 2.1.5 that model-based clustering refers to fitting a finite mixture model given a data set X; then the cluster membership of observation x_i is determined by the highest probability of it belonging to a density component. Building the mixture model $f(x|\Theta)$ requires the determination of 1) G — the number of mixture components (clusters) and 2) $f_g(x|\theta_g)$ — the distribution assumed to be followed by the mixture component g.

The use of mixture models for clustering has been found to have many applications. In an electricity usage profiling study, K. Li, Ma, Robinson, and Ma (2018) noticed several elongated ellipses in the scatterplot of electricity usage data, urging the use of Gaussian components for their mixture model. In another example, (Choy, Lam, Yu, Lee, & Leung, 2017) performed image segmentation using the generalized Gaussian density model, where each cluster formed is interpreted as an object. Their algorithm was able to segment objects such as a starfish, cat, or tree in photographs. Additionally, using genetic data, X. Li et al. (2018) discovered a potential differentially-variable microRNA (miRNA) not yet reported in literature, upon fitting a three-component multivariate normal distribution to miRNA expression levels.

Although popular, GMMs poorly fit data that exhibit skewness and different levels of kurtosis, consequently leading to overestimation on the number of clusters (Dang, Gallaugher, Browne, & McNicholas, 2019). Alternatively, the following distributions can better generalize these kinds of data: multivariate t, skewed-t, multivariate power exponential, variance-gamma, generalized hyperbolic, etc. Unlike Gaussian, these component models are flexible for data with varying tail weight, peakedness, and skewness.

A common method for determining G is by selecting the model with the lowest Bayesian Information Criterion (BIC) amongst the proposed G-component mixture models. After determining the distribution of G mixture components, the next problem is on how to estimate its corresponding parameter set $\Theta*$. Expectation-Maximization (EM), a well-known parameter estimation algorithm, is an iterative procedure that maximizes the likelihood of the parameters given the observed data (Garriga, Palmer, Oltra, & Bartumeus, 2016). In the EM procedure, the parameter set is initially guessed and is re-estimated in every iteration of the E-step and M-step, until the parameter set reaches convergence. E-step computes the likelihood weight, or the posterior probability, of each data point x_i belonging to a component g. M-step re-estimates new parameters that maximize the likelihood of these weights for each component. The result of EM guarantees to reach a local maximum for parametric distributions. The direct application of using the final EM posterior probabilities in assigning data points to groups is called EM Clustering (EMC) (Garriga et al., 2016).

For dPCR droplets classification, since the groups of interest are the positive and negative droplets, a two-component mixture model suffices. However, observations of dPCR data reveal that three or more populations may form; in this case, G will have to be determined. Additionally, there is room for research in identifying the fluorescence intensity distribution that will fit the characteristics of the positive/negative groups. Since heavy tails are observed in fluorescence densities in Figure 2.2, distributions have to be explored that may best fit the data.

2.3 Performance Evaluation

2.3.1 Essential Metrics for dPCR

As an emerging technology, the number of researchers adopting to dPCR is increasingly growing, and thus, there is a need to standardize the experimental protocols, information, and metrics that should be included in published works.

This necessity led to the proposed Minimum Information for the Publication of Digital PCR Experiments (dMIQE) Guidelines by Huggett et al. (2013). Compliance of the dMIQE guidelines allows dPCR analyses to have data comparability and reproducibility between experiments. The main categories of the dMIQE checklist that requires detailed documentation are the following: 1.) experimental design, 2.) sample, 3.) nucleic acid extraction, 4.) dPCR target information, 5.) dPCR oligonucleotides, 6.) dPCR protocol, 7.) dPCR validation, 8.) and data analysis. Since this paper focuses on the analysis of endpoint fluorescence data, only the metrics in data analysis are further explored.

The data analysis section lists the dPCR metrics that are either essential or desired. Some of the essential metrics are as follows: 1.) mean copies per partition, denoted as λ, 2.) results of positive and negative control samples, 3.) repeatability (intraassay variation), 4.) and experimental variance or confidence intervals (CI); on the other hand, the desirable metrics include 1.) reproducibility (interassay/user/lab etc. variation), and 2.) number and concordance of biological replicates.

Among these metrics, reporting λ is emphasized since this is an important factor that determines the precision of the estimate. To calculate λ with accuracy, the three assumptions must be followed (Kreutz et al., 2011):

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.
3. Target molecules are independent in a sense that there is no interaction with one another or on device surfaces.

When these assumptions are satisfied, the Poisson distribution can be used to derive the formula $\lambda = -ln(1 - k/n)$; where k is the number of successes in n trials. In this context, a positive droplet is considered a success and the total number of droplets or partitions is the number of trails.

2.3.2 Precision of Quantification Estimates

As discussed in section 1.2, there are numerous sources of variation in the dPCR workflow, including biological, chemical, and operator errors. It is therefore necessary to produce experimental replicates to measure its repeatability (intraassay

variation), reproducibility (interassay variation), and experimental variation. Intraassay samples are technical repeats of the same sample and are prepared at the same time and plate. Reproducibility includes the variation from interassay variability (variation added upon repeating an experiment, which includes variation from differing days, times, and plates the assay was prepared), variation from differing operator or laboratory. Experimental variation can be measured when biological replicates are available; if not, an error variance may be estimated using the confidence of interval from the Poisson estimate (Huggett et al., 2013).

To measure the precision between replicate measurements, the standard deviation, variance, and coefficient of variation (CV) are commonly reported in studies. In a polymavirus study (Arvia et al., 2017), repeatability and reproducibility of a ten-fold serially diluted dPCR assay were measured using the CV of each run's triplicate. Even before dPCR analyses, real-time qPCR has already used CV to measure reproducibility of target concentration estimates and threshold cycles (Cook, Atienza, Bagabag, Obrigewitch, & Jerome, 2009; Lai, Cook, Wendt, Corey, & Jerome, 2003). The CV is also used to compare the precision of dPCR against real-time qPCR for concentration estimates in several studies. In virology, Strain et al. (2013) observed a lower average CV for dPCR than qPCR for quantifying HIV DNA. A similar finding was shown on cytomegalovirus, Hall Sedlak and Jerome has reported increased dPCR precision over qPCR through the use of CVs of the estimated target copies per μL. In the comparative study of Hindson et al. (2013), they found that in serum miRNAs, dPCR consistently had the lowest average CVs within- and between-runs than qPCR; CVs were used to compare precision across preparative replicates, across RT replicates, and across PCR replicates.

2.3.3 Evaluating Unknown Target Concentrations

Although dPCR is found to be precise, it becomes poor for samples with low molecules per partition and also for samples with high target molecules per positive partition (Huggett et al., 2013). Because of this limited dynamic range of the instrument, dPCR samples are first diluted. The first use of dilution series to quantitate molecules with Poisson statistics were published by Sykes et al. (1992). Serial dilution is the stepwise dilution of a substance usually in a geometric progression. In the study of Sykes et al., a series of 10-fold dilutions of the sample were prepared that ranges from 10^{-4} to 10^{-9}. Their initial purpose of performing a dilution series was to find the optimal point at which the amplification would be distinguishable as positive or negative; recently, researchers have also found the use of dilution factors to estimate the initial target concentration of a sample prior dilution (Gou et al., 2018; Zhu et al., 2017).

The starting target concentration in the sample, denoted as c_1, can be estimated by finding the relationship between λ and the dilution factor D. The mean target copies per partition can also be expressed as $\lambda = (c_1 \times D)/N$. By taking the logarithms on both sides, a linear relationship between $-\log -ln(\lambda)$ and $-\log D$ can be established, and c_1 can be estimated from a regression intercept. The c_1 has been used by Zhu et al. (2017) to assess their proposed dPCR equipment's agreement with the Bio-Rad equipment. In addition, they have also recorded the CV of the replicates for each dilution step. Besides estimating c_1, Gou et al. (2018) has also utilized this linear model to assess the validity of their proposed dPCR devices, wherein a strong linear relationship between $-\log -ln(\lambda)$ and $-\log D$ implies good detection within a dynamic range. Gou et al. have claimed that their device is a robust tool for detection and quantification upon finding a high R^2 from a four step 10-fold dilution series (10 - 10^{-4}) of cDNA samples.

2.3.4 Evaluating Known Target Concentrations

In specific situations, the starting target concentration in a sample is a known quantity and is controlled by the researcher. In the evaluation of 'definetherain', a dPCR quantification system with focus on low-copy counts (M. Jones et al., 2014), a positive control sample was prepared for two targets, albumin and HIV-1 proviral DNA, with expected concentrations of 10^5 to 10^0 copies. The estimates of definetherain and QuantaSoft were plotted against the expected by regressing $\log_{10}(\text{Expected} + 1)$ by $\log_{10}(\text{CopyNumber} + 1)$ (CopyNumber refers to the estimated target concentration). The ideal outcome is a 1:1 correspondence between the expected and resulting estimates. In their first linear regression model, only samples with low expected values were included (<3000 copies), this resulted in a much more significant p-value (<0.01) for definetherain as compared to QuantaSoft; but when including the whole range of concentrations, the linear relationship becomes insignificant for both methods. This conclusion has supported the advantage of definetherain in low-copy samples through the use of the regression of known concentrations by the estimated target copy counts.

Chapter 3

Theoretical Framework

3.1 Target Quantification

3.1.1 Fluorescence, Concentration, and Dilution

Most molecules are in its state of lowest vibration level at room temperature, and its state becomes excited upon absorbing energy from light. This excitation elevates the molecule to higher vibration levels that cause the emission of fluorescence. The fluorescence intensity of dilute samples is related to physical variables such as the molecular extinction coefficient, quantum efficiency, intensity of incident light (Elmer, 2000).

When only a single fluorescent reporter is present, molecules emit only one type of fluorescence intensity. In the process of dPCR, fluorescent probes and primers attach to target sequences in the partitioned reaction mix. When the DNA molecules become excited, each droplet emits a fluorescence endpoint intensity, which is then used to identify if a droplet is positive or negative. An intensity threshold is then determined to classify droplets, such that any intensity less than the threshold is a negative droplet and positive otherwise (Trypsteen et al., 2015). The target concentration is then estimated using the counts of the classified droplets.

In analytical chemistry, the concentration, c, is a measurement of the amount of solute present in an amount of solution,

$$c = \frac{\text{amount of solute}}{\text{amount of solution}},$$

where the fraction can be in terms of molarity ($\frac{\text{moles solute}}{\text{liters solution}}$), weight percent ($\frac{\text{mL solute}}{\text{100 mL solution}}$), weight-to-volume percent ($\frac{\text{grams solute}}{\text{100 mL solution}}$), etc (Harvey, 2010). Unless otherwise stated, concentrations in this paper are in terms of target molecule counts per μL. The concentration formulas specific to dPCR discussed in the following sections are referenced from Kreutz et al. (2011), Zhu et al. (2014), and Gou et al. (2018).

The dilution factor, D, is the ratio of the initial volume to the final diluted volume ($V_1 : V_2$), or equivalently, the ratio of the final diluted concentration to the initial stock concentration ($c_2 : c_1$).

$$D = \frac{V_1}{V_2} = \frac{c_2}{c_1}$$

In reporting concentrations, the unknown stock concentration, c_1, is the variable of interest. From the formula above, c_1 can be obtained as

$$c_1 = c_2 \times \frac{1}{D} \tag{3.1}$$

One approach to solve for c_1 is to estimate c_2, the target concentration from the diluted sample. Suppose that the average target copies per droplet λ is given. Then, a simple unit conversion from λ to c_2 (i.e. from target copies per droplet to target copies per μL) can be derived as follows

$$c_2 = \lambda \times \frac{1000}{V_{drp}}, \tag{3.2}$$

where V_{drp} is the known constant droplet volume in terms of nL. Substituting c_2 from equation 3.2 to 3.1, and solving for λ brings

$$\lambda = c_1 \times \frac{V_{drp}}{1000} \times D, \tag{3.3}$$

which will be useful when a dilution series is available as discussed later on Section 3.1.3.

3.1.2 Poisson Distribution in Counting Target Copies

Let X be a random variable that represents the number of outcomes that appeared in either a time interval or a region of equal units h (specified as a time, distance, area, or volume). X is defined to follow a Poisson distribution, when the following assumptions are satisfied:

1. For all disjoint fixed time intervals or regions, the number of occurences in the span of h is independent from each other.

2. The probability of only one outcome happening is proportional to the specified h.

3. The probability of more than one outcome given a small h is negligible relative to the probability of only one outcome occuring in the same space.

The probability distribution function of X is defined as

$$P(x;\lambda) = \frac{e^{-\lambda}\lambda^x}{x!}, \quad x = 0, 1, 2, ...$$

where λ is the average number of outcomes per fixed time interval or region of h units (Walpole, Myers, Myers, & Ye, 2011).

In the context of DNA quantification, the outcome of interest X is the number of target copies, and the region of fixed sizes h corresponds to a droplet of equal volumes V_{drp}. The expected value of the target copies X per droplet, denoted by λ can be estimated using MLE as $\hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} x_i$; where n is the number of independent trials, and x_i is 1 or 0 for a successful trial. By defining trial as a droplet, and success as a positive droplet, it follows that the probability of getting a positive droplet (i.e. at least one target copy) is $P(x > 0) = \frac{N_{pos}}{N_{tot}}$, where N_{tot} and N_{pos} is the count of the total and positive droplets, respectively. The equation below shows how λ can be estimated using Poisson probabilities

$$\begin{aligned} 1 - P(x = 0) &= P(x > 0) \\ 1 - e^{-\lambda} &= \frac{N_{pos}}{N_{tot}} \\ e^{-\lambda} &= 1 - \frac{N_{pos}}{N_{tot}} \\ \hat{\lambda} &= -\ln(1 - \frac{N_{pos}}{N_{tot}}) \\ \hat{\lambda} &= -\ln(\frac{N_{neg}}{N_{tot}}) \end{aligned} \tag{3.4}$$

Though the last two lines are obviously equivalent, it was derived here since the latter is the commonly used formula for λ (Tzonev, 2018). For this study, the comparison between quantification methods will be in terms of $\hat{\lambda}$ concentration, rather than c_1, since the latter is just $\hat{\lambda}$ multiplied by some constants.

3.1.3 Log-log Model in Limiting Dilution

Serial dilution assays is a technique to estimate the target concentraton in a population; usually, the dilution factor (a level in the dilution series) progresses in a geometric sequence (Deng, Custer, Busch, Bakkour, & Lee, 2017). For each dilution factor, sample replicates are prepared; producing a total of n assays. Let D_i denote the dilution factor at assay sample i, where $i = 1, 2, ..., n$. Then, continuing from equation 3.3, for a given stock conconcentration c_1 diluted by D_i, the expected target copies per droplet is λ_i. Thus, it can be said that with a fixed quantity c_1, D_i is a predictor of λ_i. Its relationship in equation 3.3 can then be linearized by taking the logarithm on both sides

$$-\log(\lambda) = -\log(c_1 \times \frac{V_{drp}}{1000}) - \log(D) \tag{3.5}$$

The proportion of target copies in the droplet population can then be estimated by fitting a binomial generalized linear model (GLM) with a log-log link:

$$g(\lambda_i) = \beta_0 + \log(D_i).$$

In GLM terms, $\log D_i$ is the offset and $g()$ is the complementary log-log link function resulting in the final model

$$-\log(\lambda_i) = -\log(c_1 \times \frac{V_{drp}}{1000}) - \log(D_i)\beta_1 \tag{3.6}$$

Equation 3.6 can be formulated as a simple linear regression model $Y = \beta_0 + X\beta_1 + \epsilon$, where β_0 and β_1 denotes the slope and intercept, respectively (Walpole et al., 2011). The error term ϵ is a random variable assumed to be normally distributed with mean 0 and constant variance σ^2. Given a set of ordered pairs (x_i, y_i); $i = 1, 2, ...n$ and an estimated regression model $\hat{y}_i = b_0 + x_i b_1$, the ith residual is defined as $e_i = y_i - \hat{y}_i$. The ordinary least squares (OLS) estimator finds the values of b_0 and b_1 so as to minimize the residual sum of squares

$$\text{SSRes} = \sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

The OLS estimates of b_0 and b_1 for the regression coefficients β_0 and β_1 are

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad \text{and}$$

$$b_0 = \frac{\sum_{i=1}^{n} y_i - b_1 \sum_{i=1}^{n} x_i}{n}.$$

In assay analysis, the interpretation of a slope significantly greater than one implies that the proportion of the target sequence is hyper responsive to the diluted concentration. Otherwise, a slope less than one implies that the proportion of targets is less responsive to the diluted concentration, and suggests heterogeneity. (Hu & Smyth, 2009).

3.2 Evaluation Metrics

3.2.1 Precision of Technical Replicates

In quantitative assay studies, assay variability is typically summarized using the coefficient of variation (CV) (Reed, Lynn, & Meade, 2003). The CV of $\hat{\lambda}$ is defined as

$$CV = \frac{SD(\hat{\lambda})}{mean(\hat{\lambda})}.$$

A smaller CV implies good agreement amongst replicate estimates. The advantage of CV over standard deviation (SD) is that it takes into account the magnitude of the units, making CV comparable regardless of analyte concentration.

3.2.2 Accuracy of Regression Model

Recall that given an estimated regression line, the deviation of the fitted $\hat{y}$ from the observed y is the error term ϵ with mean 0 and variance σ^2. The deviation of ϵ measures the model's *lack of fit* (James, Witten, Hastie, & Tibshirani, 2013). The residual standard error (RSE), or $\hat{\sigma}$, is the estimated standard deviation of the error terms from fitting a regression model and is calculated as

$$RSE = \sqrt{\frac{1}{n-2} \times \text{SSRes}}.$$

When the model predicts values that are very close to the observed data, such that $\hat{y}_i \approx y_i$ for all $i = 1, 2, ..., n$, then RSE will be very small. On the other hand, predicted values that are far from the actual data will have a large RSE,

indicating a poor fit. Since the magnitude of RSE depends on the units of Y, it is not comparable between datasets, and also it is unclear what defines an acceptable RSE.

In addition to RSE for assessing model accuracy, the coefficient of determination, R^2, is unitless and is in the form of a proportion

$$R^2 = \frac{\text{SSTotal - SSRes}}{\text{SSTotal}} = 1 - \frac{\text{SSRes}}{\text{SSTotal}}.$$

where $\text{SSTotal} = \sum(y_i - \bar{y})^2$ is the total variance in the response Y. Since SSRes is equivalent to the amount of unexplained variance from the regression model, then in contrast, the interpretation of R^2 is the proportion of variability in Y that can be explained by X. An R^2 close to 1 means that the regressor X explained a large percentage in the variability in Y, and a value close to 0 means X does not explain much of the variability in the response.

3.2.3 Binary Classification Metrics

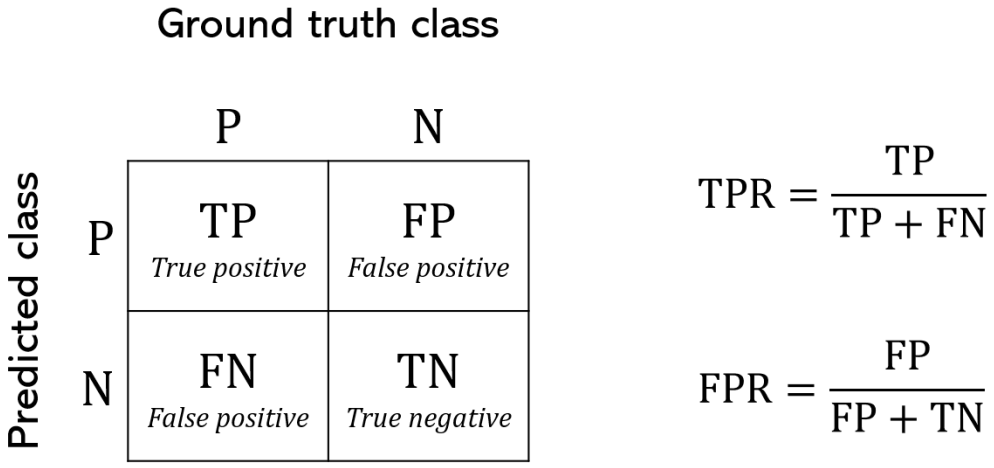

Figure 3.1: Confusion matrix with TPR and FPR calculations

Most binary classification metrics are founded on the contingency table of ground truth vs predicted conditions, resulting in a tally of the true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). This table in Figure 3.1 is also known as the confusion matrix, from which, ratios of cell counts, marginal (row or column) totals, or grand totals can be derived and are used as binary classification metrics. For this study, the classifier metrics False Positive Rate (FPR) and True Positive Rate (TPR) will be assessed. These two

statistics are commonly assessed synchronously on a TPR vs FPR plot, called the receiver operating characteristics (ROC) graph. When TPR and FPR pairs are produced from multiple sample sets (or a range of threshold settings), a curve is usually formed in the ROC graph which represents the tradeoff of detection errors (false positives) and benefits (true positives). A good classifier should have more benefits than errors in most cases; that means, in the ROC graph, it is ideal that the points lie very close on the top left corner (1,1), producing a curve that largely covers the graph (Fawcett, 2006).

3.3 EM Clustering

3.3.1 G-component Finite Mixture Model

Denote $X = \{x_1, ..., x_N;\ x_i \in \mathbb{R}^P\}$ as a statistically independent observation sequence where N is the number of observations and P is the dimensionality of x_i, then the G-component finite mixture density is defined as

$$f(X|\theta) = \sum_{g=1}^{G} \pi_g f_g(X|\psi_g),$$

where π_g is the mixing proportion ($\pi_g > 0, \sum_{g=1}^{G} \pi_g = 1$), $f_g(X|\psi_g)$ is the gth component density where ψ_g are its set of parameters, $\theta_g \in \theta = \{\theta_1, ..., \theta_G\}$ and $\theta_g = \{\pi_g, \psi_g\}$ is the unknown parameter set that defines the density function for approximating the true probability of X.

3.3.2 Parameter Estimation Using EM

In modeling a mixture of densities, the parameter estimation faces the problem that an observation's component membership is not known or is "hidden", and this complicates the structure of the likelihood function. Data with "hidden" or latent variables are considered as an incomplete-data. One approach to solving the likelihood of an incomplete-data is by introducing the complete-data, the data where the component memberships are known for each observation. The Expectation-Maximization (EM) is an iterative procedure that computes the MLE of an incomplete-data by exploiting the complete-data likelihood. The EM algorithm is further discussed in the following sections, first by introducing the dependence of the incomplete-data likelihood to the complete-data likelihood, then the iterative steps of EM and its convergence property.

Incomplete-data Likelihood

To demonstrate this dependence, the following notations are continued from the previous section:

- $X = \{x_1, ..., x_N;\ x_i \in \mathbb{R}^P\}$ denotes the set of observations
- $C = \{C^{(1)}, ..., C^{(G)}\}$ is the set of cluster labels for each component g
- $Z = \{z_1, ..., z_N;\ z_i \in C^{(g)}\}$ denotes the set of the observations' component membership
- θ_t is the parameter set estimate in iteration t

Then the parameter set θ_t can have a complete-data likelihood function defined as

$$p(Z, X|\theta_t) = \prod_{i=1}^{N} p(z_i, x_i|\theta_t), \tag{3.7}$$

then by probability theory, the likelihood of the incomplete-data can be expressed using the complete-data likelihood :

$$p(X|\theta_t) = \frac{p(Z, X|\theta_t)}{P(Z|X, \theta_t)}. \tag{3.8}$$

Based on Kung, Mak, and Lin (2004), the incomplete-data log-likelihood can be obtained as a sum of two terms derived as follows :

$$\begin{aligned}
L(\theta_t) &= \log(p(X|\theta_t)) \\
&= \log(p(X|\theta_t)) \underbrace{\sum_Z P(Z|X, \theta_t)}_{(1)} \\
&= \sum_Z \log(p(X|\theta_t)) P(Z|X, \theta_t) \\
&= \sum_Z \log\left(\frac{p(X, Z|\theta)}{P(Z|X, \theta_t)}\right) P(Z|X, \theta_t) \\
&= \sum_Z \log(p(X, Z|\theta_t)) P(Z|X, \theta_t) \ - \ \sum_Z \log(P(Z|X, \theta_t)) P(Z|X, \theta_t) \\
&= E_Z[log(p(X, Z|\theta_t)|X, \theta_t] \ - \ E_Z[log(P(Z|X, \theta_t)|X, \theta_t] \\
&= Q(\theta_t|\theta_t) \ + \ R(\theta_t|\theta_t)
\end{aligned}$$

By denoting $Q(\theta|\theta_t) \equiv E_Z[\log(p(Z,X|\theta_t))|X,\theta_t]$ as the expectation of the complete-data log-likelihood given X using the parameter set of step t, and $R(\theta|\theta_t)$ as the entropy term, then the incomplete-data likelihood maximization problem can be accomplished by maximizing $Q(\theta|\theta_t)$ or $R(\theta|\theta_t)$. Whereas the latter can be achieved by performing simulated annealing, $Q(\theta|\theta_t)$ can be maximized using the convergence property demonstrated in the following:

$$
\begin{aligned}
L(\theta) &= \log(p(X|\theta)) \\
&= \log\left(\sum_Z p(X,Z|\theta)\,\underbrace{\frac{P(Z|X,\theta_t)}{P(Z|X,\theta_t)}}_{(1)}\right) \\
&= \log\left(\sum_Z \frac{p(X,Z|\theta)}{P(Z|X,\theta_t)} P(Z|X,\theta_t)\right) \\
&= \log\left(E_Z\left[\frac{p(X,Z|\theta)}{P(Z|X,\theta_t)} \mid X,\theta_t\right]\right) \\
&\geqslant E_z\left[\log\left(\frac{p(X,Z|\theta)}{P(Z|X,\theta_t)}\right) \mid X,\theta_t\right] \qquad \text{(by Jensen's inequality)} \\
&\geqslant E_z[\log(p(X,Z|\theta)) \mid X,\theta_t] \;-\; E_z[\log(P(Z|X,\theta_t)) \mid X,\theta_t] \\
&\geqslant Q(\theta|\theta_t) + R(\theta|\theta_t)
\end{aligned}
\tag{3.9}
$$

Suppose θ^* is an element in the parameter space and it is given to maximize Q, that is:

$$\theta^* = \arg\max_\theta Q(\theta|\theta_t)$$

Continuing the inequality from 3.9,

$$L(\theta^*) \geqslant Q(\theta^*|\theta_t) + R(\theta_t|\theta_t)$$

And since $Q(\theta^*|\theta_t) > Q(\theta_t|\theta_t)$, then it follows that

$$
\begin{aligned}
L(\theta^*) &\geqslant Q(\theta_t|\theta_t) \;+\; R(\theta_t|\theta_t) \\
L(\theta^*) &\geqslant L(\theta_t)
\end{aligned}
\tag{3.10}
$$

This shows that for a given parameter set θ_t at iteration t, a θ^* can be chosen that will increase the incomplete-data log-likelihood $L(\theta_t)$.

Complete-data Likelihood

The indicator variable δ_{gi} is introduced to express the status of the hidden states as

$$\delta_i^{(g)} = \delta\left(z_i\ , C^{(g)}\right) = \begin{cases} 1 & \text{if } x_i \text{ was generated by } C^{(g)} \\ 0 & \text{otherwise} \end{cases}$$

Using the indicator variable, the complete-data likelihood is given by

$$\begin{aligned} p(X, Z|\theta) &= \prod_{i=1}^{N} \sum_{g=1}^{G} \delta_i^{(g)}\ p\left(x_i,\ z_i = C^{(g)} \mid \theta\right) \\ &= \prod_{i=1}^{N} \sum_{g=1}^{G} \delta_i^{(g)}\ p\left(x_i,\ \delta_i^{(g)} = 1 \mid \theta\right) \\ L_C(\theta) \equiv \log(p(X, Z|\theta)) &= \sum_{i=1}^{N} \log\left(\sum_{g=1}^{G} \delta_i^{(g)}\ p\left(x_i,\ \delta_i^{(g)} = 1 \mid \theta\right)\right) \\ &= \sum_{i=1}^{N} \log\left(\sum_{g=1}^{G} \delta_i^{(g)}\ p\left(x_i \mid \delta_i^{(g)} = 1, \theta\right)\ p\left(\delta_i^{(g)} = 1 \mid \theta\right)\right) \end{aligned}$$

The following observations are noted to continue the equation:

- when expanding the summation $\sum_{g=1}^{G}$, there would only be one nonzero term. And therefore δ_i^g can be pulled out from the log function without affecting the results;
- $p\left(x_i|\delta_i^{(g)} = 1,\ \theta\right) = p\left(x_i|\delta_i^{(g)} = 1,\ \phi^{(g)}\right)$
- $p\left(\delta_i^{(g)} = 1 \mid \theta\right) = \pi^{(g)}$

Putting these all together, the complete-data log-likelihood $L_C(\theta)$ is finally expressed as

$$L_C(\theta) = \sum_{i=1}^{N} \sum_{g=1}^{G} \delta_i^{(g)}\ \log\left(p\left(x_i \mid \delta_i^{(g)} = 1, \phi^{(g)}\right)\ \pi^{(g)}\right) \tag{3.11}$$

The EM algorithm is an iterative algorithm that can maximize $L(\theta)$, by maximizing $L_C(\theta)$, at each iteration of the expectation step (E-step) and the maximization step (M-step). To start the algorithm, t is initialized to 0 and values for the

initial parameter set θ_0 are given. Then E-step and M-step execute alternatively, incrementing t for each loop and updating θ_t until $L(\theta)$ converges. The E-step computes the expectation of the likelihood of the mixture model parameters by including the hidden variables. The M-step re-estimates the parameters by maximizing the expected likelihood from the E-step. These new model parameters from the M-step are then used for the next iteration until the difference of $L(\theta_{t+1})$ and $L(\theta_t)$ become arbitrarily small.

Initialization

Before starting the first iteration ($t = 0$), EM requires an initial guess of the parameter set for θ_0.

E-step

At iteration t, E-step computes the function $Q(\theta|\theta_t)$ using the complete-data log-likelihood from equation 3.11

$$\begin{aligned} Q(\theta|\theta_t) &= E_Z[L_C(\theta) \mid X,\ \theta_t] \\ &= E_Z\left[\sum_{i=1}^{N}\sum_{g=1}^{G} \delta_i^{(g)}\ \log\left(p\left(x_i \mid \delta_i^{(g)}=1, \phi^{(g)}\right)\ \pi^{(g)}\right) \mid X,\ \theta_t\right] \\ &= \sum_{i=1}^{N}\sum_{g=1}^{G} E_Z\left[\delta_i^{(g)} \mid x_i,\ \theta_t\right]\ \log\left(p\left(x_i \mid \delta_i^{(g)}=1, \phi^{(g)}\right)\ \pi^{(g)}\right) \\ &= \sum_{i=1}^{N}\sum_{g=1}^{G} P\left(\delta_i^{(g)}=1 \mid x_i\ ,\ \theta_t\right)\ \log\left(p\left(x_i \mid \delta_i^{(g)}=1, \phi^{(g)}\right)\ \pi^{(g)}\right) \end{aligned} \tag{3.12}$$

By denoting the component membership function $h_t^{(g)}(x_i)$ as the probability of the latent variable z_i to be the component label $C^{(g)}$ given the observation x_i and

parameter set θ_t, its equation can be expressed using Bayes theorem as follows

$$\begin{aligned} h_t^{(g)}(x_i) &\equiv E_Z\left[\delta_i^{(g)} \mid x_i,\ \theta_t\right] \\ &= P\left(\delta_i^{(g)} = 1 \mid x_i\ ,\ \theta_t\right) \\ &= \frac{p\left(x_i \mid \delta_i^{(g)} \ =\ 1,\ \theta_t\right) P\left(\delta_i^{(g)} \ =\ 1 \mid \theta_t\right)}{\sum_{k=1}^{G} p\left(x_i \mid \delta_i^{(k)} \ =\ 1,\ \theta_t\right) P\left(\delta_i^{(k)} \ =\ 1 \mid \theta_t\right)} \\ &= \frac{p\left(x_i \mid \delta_i^{(g)} \ =\ 1,\ \phi_t^{(g)}\right) \pi_t^{(g)}}{\sum_{k=1}^{G} p\left(x_i \mid \delta_i^{(k)} \ =\ 1,\ \phi_t^{(g)}\right) \pi_t^{(g)}} \end{aligned} \tag{3.13}$$

Given θ_t, the best guess of the component membership probability $h_t^{(g)}(x_i)$ for all observations and components is determined, and by doing so, $Q(\theta|\theta_t)$ is now then a function of θ.

M-step

At iteration t, M-step finds an adjusted θ_{t+1} that maximizes the function $Q(\theta|\theta_t)$ with respect to θ over the parameter space Ω. Recall that for a parameter set of component g, $\theta^{(g)} = \left\{\phi^{(g)}\ ,\ \pi^{(g)}\right\}$, it consists of the gth component mixing proportion $\pi^{(g)}$ and the density parameters $\phi^{(g)}$. The estimation of these two parameters are done sequentially and is described succeedingly.

The adjusted mixing proportion $\pi_{t+1}^{(g)}$ is calculated independently from $\phi_{t+1}^{(g)}$; and that if the component membership z_i was observed, then the mixing proportion can be obtained using ML as

$$\pi_{t+1}^{(g)} = \sum_{i=1}^{N} \delta_i^{(g)}/N.$$

Instead, the E-step provides a component membership probability $h_t^{(g)}(x_i)$ to estimate $\delta_i^{(g)}$

$$\begin{aligned} \pi_t^{(g)} &= \sum_{i=1}^{N} h_t^{(g)}(x_i)/N \\ &= \sum_{i=1}^{N} E\left[\delta_i^{(g)} \mid x_i\ ,\ \theta_t\right]\ /\ N \end{aligned}$$

Which means that the estimated $\pi_{t+1}^{(g)}$ is based on the posterior probabilities of gth component membership for all observations X.

Following the estimation of $\pi_{t+1}^{(g)}$, the only variable remaining to be maximally estimated in the function $Q(\theta|\theta_t)$ are the component density parameters $\phi_t^{(g)}$ for all g (based on equation 3.12). The ML estimate of $\phi_{t+1}^{(g)}$ can be obtained as an appropriate root of

$$\sum_{g=1}^{G}\sum_{i=1}^{N} h_t^{(g)}(x_i)\ \frac{\partial \log\left(p\left(x_i \mid \delta_i^{(g)}=1,\ \phi_t^{(g)}\right)\ \pi_t^{(g)}\right)}{\partial\phi} = 0$$

Finally, after obtaining the estimates θ_{t+1} that maximizes $Q(\theta|\theta_t)$, until the converge rule is satisfied, the iteration t is incremented, thereby carrying over θ_{t+1} to the calculation of the next E-step.

Convergence

After the M-step, it is checked if the incomplete-log likelihood has reached converge, that is

$$L(\theta_{t+1}) - L(\theta_t) \geqslant \epsilon;$$

where ϵ is the specified termination threshold. As stated in equation 3.10, the likelihood $L(\theta_t)$ increases at each iteration, until it reaches a point where the improvement becomes arbitrarily small and is considered negligible. However, the convergence of the likelihood value from $L(\theta_t)$ to L^* does not imply that θ_t has reached its maximum point θ^*. In most cases, the limiting value L^* is a local maximum. A common workaround for this problem is to perform EM in a variety of initial parameter sets.

3.3.3 Model-based Clustering

The common method for assigning a cluster g to x_i is by using the maximum a posteriori (MAP) classification such that

$$\text{MAP}(\delta_{gi}) = \begin{cases} 1 & \text{if } \max_g\{h_g(x_i)\} \text{ is in cluster } C_g \\ 0 & \text{otherwise} \end{cases}, \tag{3.14}$$

where $\text{MAP}(\delta_{gi})$ of 1 means that amongst all the clusters in C, x_i has the highest probability to belong to cluster C_g, and is therefore assigned to C_g.

3.4 Nonnormal Mixture Models

3.4.1 T-Mixture Models

3.4.2 Skewed T-Mixture Models

3.4.3 Generalized Hyperbolic Distribution Mixture Models

The generalized hyperbolic (GH) distribution, first introduced by Barndorff-Nielsen (Barndorff-Nielsen, 1977), is a continuous probability distribution with five parameters that describe its location, scale, asymmetry, and the decay of its tails. As the name suggests, this distribution is generalized and is a superclass of the normal inverse Gaussian distributions, scaled t-distributions, standard hyperbolic distributions, variance-gamma distributions, among others. The tails of the GH distribution can range from a Gaussian-like tail to a heavy tail of exponential behavior. Both tails can exhibit different behaviors simultaneously, where the left-hand can be less heavy than the right-hand tail. This property of the GH distribution allows the modeling of asymmetric heavy-tailed populations commonly observed in finance and econometric data (Takahashi, Watanabe, & Omori, 2016; Nwobi, 2014; Necula, 2009; Aas & Haff, 2006; Bibby & Sørensen, 2003). Its applications are in predicting risk models of exchange rates, portfolios, and stock index returns data.

Let X be a random variable that follows a generalized hyperbolic distribution with parameters for location (λ), scaling (δ), shape (α), skewness (β), and a parameter μ that influences kurtosis and the GH characterization.

$$X \sim GH(\lambda, \alpha, \beta, \delta, \mu)$$

Then the probability distribution function of X is defined as

$$\begin{aligned} P(x; \lambda, \alpha, \beta, \delta, \mu) = a(\lambda, \alpha, \beta, \delta, \mu)(\delta^2 + (x-\mu)^2)^{1/2\lambda - 1/4} \\ \cdot B(\lambda - 0.5, \alpha\sqrt{\delta^2 + x^2 - 2x\mu + \mu^2})e^{\beta(x-\mu)} \end{aligned} \tag{3.15}$$

where

$$a(\lambda, \alpha, \beta, \delta, \mu) = \frac{(\alpha^2 - \beta^2)^{1/2\lambda}}{\sqrt{2\pi}\alpha^{\lambda - 1/2}\delta^{\lambda} B(\lambda, \delta\sqrt{\alpha^2 - \beta^2})}$$

and $B(\lambda, \cdot)$ denotes the modified Bessel function of the third kind with index λ. GH distribution mixture models were assessed in a study of Browne and

McNicholas (2015) using Old Faithful data (GeyserTimes, 2017) and simulated datasets. The former dataset was observed to have skewed tails, and the resulting GH mixture model was shown to have a superior fit as compared to the scale mixture of skew-normal distributions. The simulated dataset was composed of one hundred 2-component mixtures of Gaussian and skew-t distributions. When a GH mixture model was fitted using EM algorithm, all the population parameters were very close to the true values. These demonstrate the ability of GH mixture model to closely capture real data consisting of several populations.

Chapter 4

Methodology

4.1 Data

4.1.1 Real Dataset

To evaluate the performance of dPCR droplet classifiers, a publicly available dataset from the study of Lievens et al. (2016) was analyzed. The premise of their problem is that there were no established criteria to assess dPCR assay quality; they addressed this problem in their study by proposing measures to evaluate the following criteria: i.) there should only be a single amplification product (i.e. only two fluorescence population is present), ii.) clear separation between positive and negatives, iii.) and limit the amount of intermediate fluorescence. Having set these standards, Lievens et al. designed nine plates, or sets, of experimental parameters in aims of optimizing their own dPCR samples. They have examined twelve DNA targets from stock solutions of food and feed materials. In the first plate, reaction mixes were prepared using real-time qPCR validated conditions, this resulted in poor dPCR performance metrics. To improve the results, they experimented on design parameters that were proven effective in other studies. The list of the experimental factors explored are shown in Table 4.1. The experimental factors in plates 1 and 9 were already fixed. As for the other plates, varying levels of the specified factors were ran as a means of finding the optimal parameter. It can be seen that two DNA targets (TC1507 and M88017) were closely examined for half of the plates, since both were producing significant amounts of rain in most cases. In fact, most of the samples prepared were exhibiting noise characteristics. The observed dPCR quality for each plate is summarized in Table 4.2.

Table 4.1: Levels of the design factors per plate and the DNA targets examined

Plate – Experimental Factor	**Levels**		**DNA Targets**		
1 - qPCR validated conditions	N/A		acp cru GT73 GTS4032	hmg le1 M1445 M810	M88017 M88701 M89788 TC1507
2 – Primer concentration	150 300	450 600	acp cru GT73 GTS4032	hmg le1 M1445 M810	M88017 M88701 M89788 TC1507
3 - Two-fold dilution series	Conc 8000 Conc 4000 Conc 2000	Conc 1000 Conc 500 Conc 250	M88701	TC1507	
4 - PCR Enhancers	Enhancer NA (none) Enhancer DMSO2% Enhancer Trehalose0.2M		M88701	TC1507	
5 - Two-fold dilution and Cycles	Conc 8000 Cycles 45 Conc 8000 Cycles 60 Conc 8000 Cycles 75 Conc 8000 Cycles 90 Conc 4000 Cycles 45 Conc 4000 Cycles 60 Conc 4000 Cycles 75 Conc 4000 Cycles 90		M88701	TC1507	
6 – Sonication time	0 3 6	9 12 15	M88701	TC1507	
7 - Annealing temperature	62 61.6 60.9 59.8	58.4 57.3 56.5 56	acp cru GT73 GTS4032	hmg le1 M1445 M810	M88017 M88701 M89788 TC1507
9 - dPCR optimized parameters	N/A		GT73 GTS4032 hmg le1	M1445 M810 M88017 M88701	M89788 TC1507

Plates 1 and 9 have constant design parameters for each DNA target, the other plates explore varying levels of the given design parameter(s). Conc refers to the diluted template DNA concentration (copies/μL, and not the target DNA concentration. Primer concentration is measured in nM, sonication time is measured in seconds; and annealing temperature is in degree Celsius. Plate 8 was omitted since it was a digital touchdown variation of acp from plate 7.

Table 4.2: Observed assay quality results per experimental factor

Plate – Experimental Factor	**Lievens et.al's remarks on the results**	**Additional remarks from our observations**
1 - qPCR validated conditions	- More than two populations amplified for acp and cru - High amount of rain for TC1507 and M88017	- Negligible amount of rain for few targets - Moderate amount of rain for few targets
2 - Primer concentration	- Significant increase in peak separation at higher primer concentrations	- No effect for some targets
3 - Two-fold dilution series	- Rain is concluded to contain target sequence but does not amplify at the same efficiency as the distinct positives	- Moderate amount of rain at higher concentrations for both targets
4 - PCR Enhancers	- No effect for both targets	- More than two populations amplified for one replicate of M88017
5 - Two-fold dilution and Cycles	- Rain is decreased at higher cycles for both targets	
6 - Sonication	- Significant decrease in rain for M88017 - No effect for TC1507	
7 - Annealing temperature	- Significant increase in peak separation at higher temperatures for TC1507 and GTS4032 - Increased rain at higher temperatures - No effect for cru and MON1445	- High presence of rain for some targets - More than two populations amplified for acp and cru - Two populations overlap for GTS4032
9 - dPCR optimized parameters	- Optimal peak separation, template DNA concentration, and at most 2.5% rain* for all targets.	- More than two populations amplified for one replicate of M88701 - Moderate amount of rain for some targets

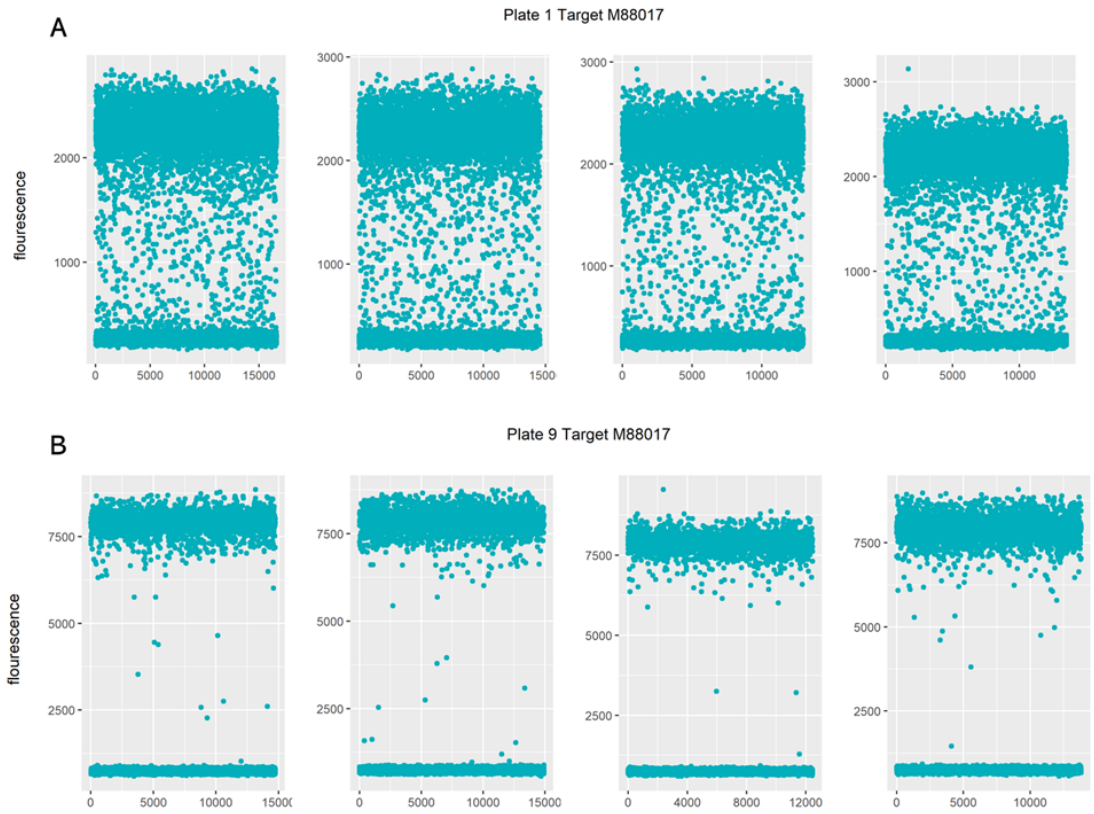

Figure 4.1: Lieven's optimization results for target M88017- (A) Using real-time qPCR validated design parameters, target M88017 displays very ambiguous read-outs. (B) Using the same target but with optimized dPCR parameters, it can be seen that there is significant reduction of rain and increased separation between the two populations.

For DNA target M88017 in Figure 4.1, it is visually clear that from the initial real-time qPCR validated conditions (subfigure A) their dPCR assay optimization (subfigure B) has been effective in removing the rain and also in separating the two populations. However, it should be noted that this is not the case for all the targets, as some still exhibit an amount of rain fluorescence in their optimized parameters. In contrast, Figure 4.2 demonstrates selected samples with noisy features. All the sample replicates in plate 1 acp (subfigure A) produced more than 2 bands of fluorescence intensities. In plate 4 where PCR Enhancers were noted to be not effective, each of the target M88017 samples with no enhancer, DMS20%, and trehalose enhancer (subfigure B) can be seen to produce heavy rain. In the case of target M88017 in plate 6 (subfigure C), it can be observed that applying sonication reduced the rain droplets. Although it is impossible to have all the λ estimate to be equal within a sample, since DNA extraction was from the organism, it is expected to some degree that the estimates should be very close to each other, regardless of the presence of noise.

4.1.2 Simulated Dataset

For this study, it is also of interest to assess the performance of droplet classifiers in extreme cases of dPCR assay quality. Since the real dataset from the previous section only has a few samples exhibiting large amounts of rain and poor separation of populations, a dataset with these characteristics were simulated. This section first reviews the simulation settings of Jacobs et al. (2017), and then describes the common and unique features of this study's simulation procedure.

In an attempt to explore the accuracy of their concentration estimator in various noise levels, Jacobs et al. simulated 9 rain settings of 4 different concentrations. For each rain and concentration setting, 15 replicates of target sample and 3 replicates of NTC samples were generated. The middle facet in figure 4.3 displays a target sample for each setting. The columns from left to right simulates the concentrations $\lambda = 0.1, 0.4, 1, 2.5$ (also indicated as blue, pink, red, and yellow, respectively). The rows from top to bottom simulates very minimal to high amounts of rain amplified by the increasing overlap of the two fluorescence populations. Results are summarized in the left and right facet in figure 4.3, where for each rain and concentration setting, the distribution of λ point estimates should ideally be very close to the true λ values (colored horizontal line).

This experiment allowed them to compare the accuracy of quantification methods in extreme cases of clean and noisy data. However, since this dataset nor the code to reproduce it is not publicly available, other researchers who may want to compare their method using the same settings may have to simulate their own.

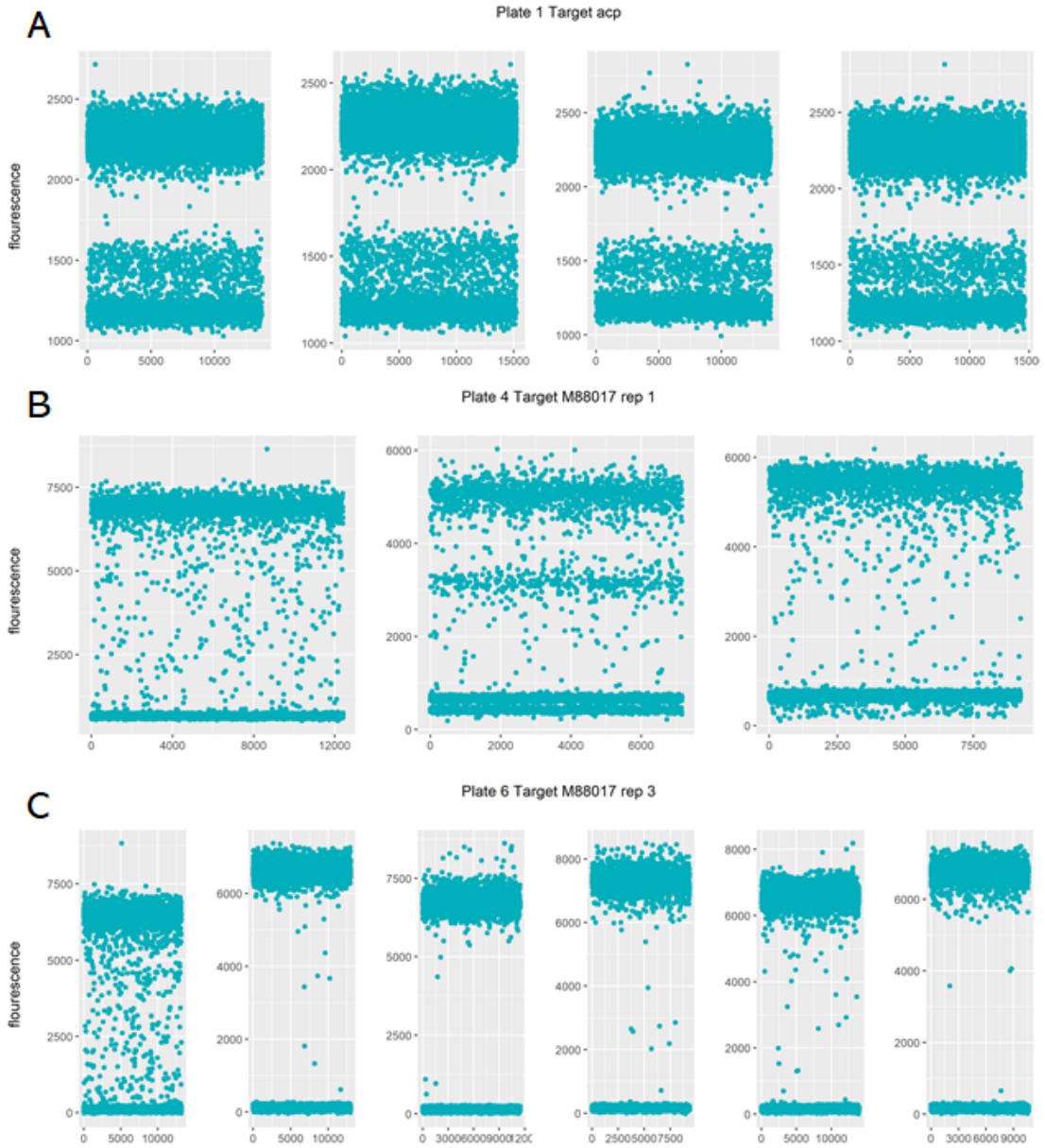

Figure 4.2: Example dPCR results using different experimental factors- (A) Target acp using real-time qPCR validated condition. (B) Target M88017 with no PCR enhancers, DMSO20% enhancer, and trehalose enhancer (one replicate each from left to right). (C) Target M88017 with sonication 0, 3, 6, 9, 12, and 15 (one replicate each from left to right).

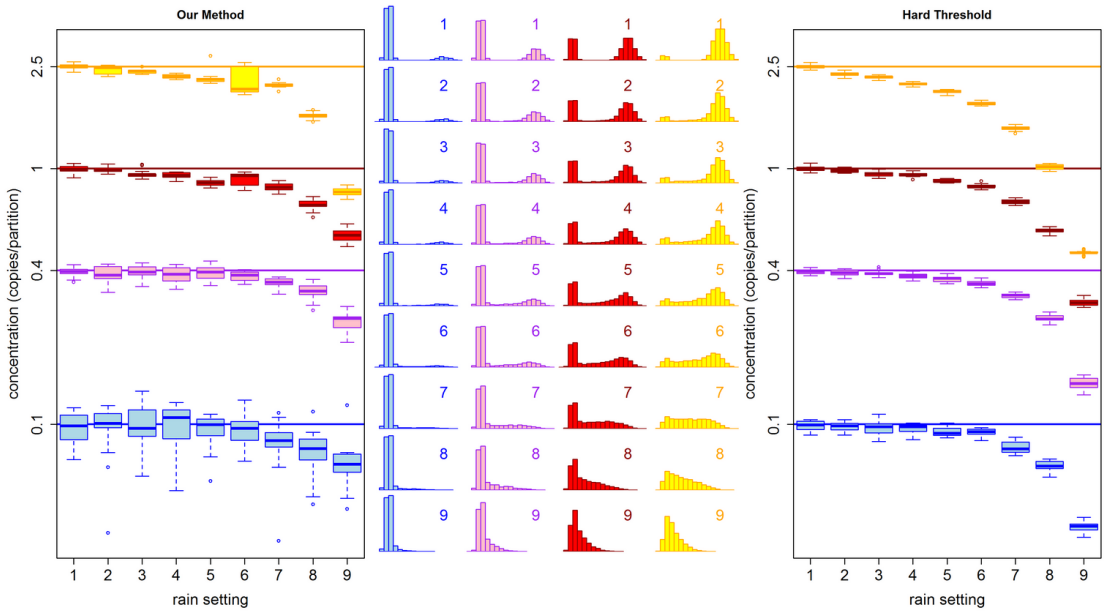

Figure 4.3: Simulated data and accuracy results from Jacobs et. al- Reprinted from Jacobs, Bart K. M.; Goetghebeur, Els; Vandesompele, Jo; De Ganck, Ariane; Nijs, Nele; Beckers, Anneleen; et al. (2017): Model-Based Classification for Digital PCR: Your Umbrella for Rain. ACS Publications. Collection. https://doi.org/10.1021/acs.analchem.6b04208

For this study, a dataset of varying rain levels and concentrations were generated with 4 rain settings (No rain, Low rain, Moderate ran, and High rain) and 5 concentration levels ($\lambda = 0.1, 0.2236, 0.5, 1.118, 2.5$). The concentration range of 0.1 - 2.5 was based from Jacobs et al.'s simulation settings, but the values in between were modified to be a geometric sequence to mimic a dilution series. Likewise, 15 replicates of target samples and 3 replicates of NTC samples were generated. Figure 4.4 shows a replicate of an artificial dPCR droplet fluorescence for each setting.

In order to closely mimic the behaviour of real world data, selected dPCR assays from the real dataset were fitted using a GH distribution mixture model. Observe in Figure 4.5 the fluorescence distribution of Plate 7 for target GTS4032. Each assay was ran in identical conditions except for differing annealing temperatures (62, 61.6, 60.9, 59.8, 58.4, 57.3, 56.5, 56 °C, labeled here as A-H). High presence of rain and poor separation of populations are eminent in sample A. These noise characteristics were slightly reduced in B; and in C, the separation of the populations improved. Lastly, sample H achieved a clear separation of the fluorescence populations and very minimal amount of rain. Because of these distinct visual comparisons of the dPCR data quality, the four samples in GTS4032,

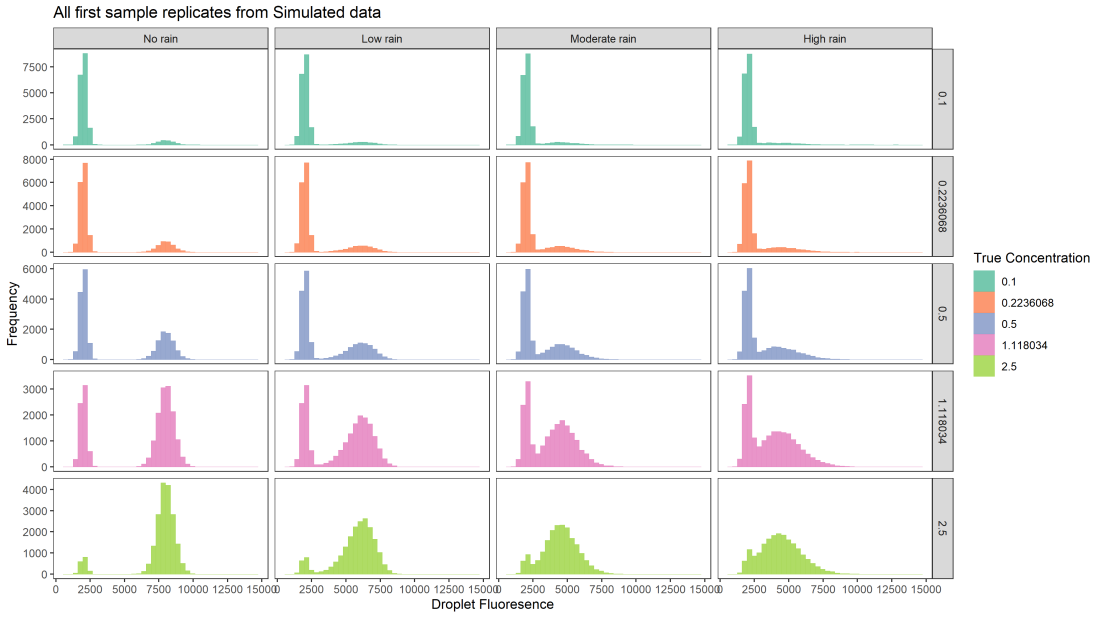

Figure 4.4: Simulated dataset for this study- Displayed are one sample replicates for each rain and concentration setting. Generated from fitting GH mixture models on real dPCR data.

marked A, B, C, and H in Figure 4.5, were used as a reference in simulating heavy, moderate, low, and no rain, respectively. A two-component GH mixture model was used to fit these samples in attempt to capture the skew and heavy tails of these observations. The R package "MixGHD" (Tortora et al., 2019) was used for fitting GH mixture models in the selected samples. The fitted models were used to generate datasets for each rain setting.

To derive the ground truth counts of positive and negative droplets, the total number of droplets were set to 20,000 for all samples; then using equation 3.4, the count of negative and positive droplets can be derived as $N_{neg} = \exp(\log(N_{tot}) - \lambda)$ and $N_{pos} = N_{tot} - N_{neg}$.

4.2 Mixture Model Fitting using EM

The mixture models considered in this study for fitting droplet fluorescence are the Gaussian distribution, T-distribution, and skewed-T distribution. The selection of these models is based on its numerous applications in literature and also in its availability in the R package "EMMIXskew" (?, ?). However, because the fluorescence population was rejected to follow a normal distribution (?, ?, ?),

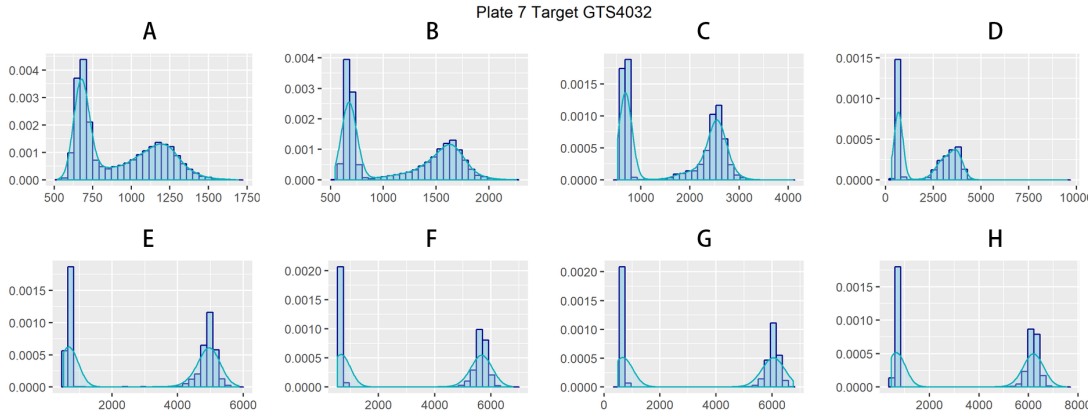

Figure 4.5: Target DNA GTS4032 in plate 7 of Lievens' dataset- Samples A to H corresponds to the annealing temperatures 62, 61.6, 60.9, 59.8, 58.4, 57.3, 56.5, and 56 °C.

Gaussian mixture model was not included in the analyses. Additionally, even though GH distribution was used to fit the reference samples for the data simulation, it was not considered here as it takes a significant amount of time to run. This downside is noted by the authors of "MixGHD" package and that code efficiency is an ongoing work. The final list of this study's proposed methods are the T-distribution and skewed-T distribution mixture models and are abbreviated as EM-T and EM-skewT, respectively.

4.3 Performance Evaluation of Quantification tools

The single-channel droplet classifier methods in literature currently includes the Bio-Rad Quantasoft software, Cloudy, ddpcRquant, definetherain, and Umbrella. The differences of these droplet classifier methods along with the EM method are summarized in Table 4.3 and has been discussed more in detail in section 2. The accessibility of these tools are listed in Table 4.4. Cloudy, ddpcRquant, and Umbrella are all publicly accessible as an R source code. Interactive web apps are also available for the ddpcRquant and definetherain analysis. Among these, only the Bio-Rad Quantasoft is not open-sourced and is costly to acquire, which is the reason it has been excluded from the analysis. In addition to this, not all of the methods can be used due to the limitations of the dataset available. The final list of methods to be compared against are Cloudy, ddpcRquant, Umbrella, EM-T, and EM-skewT.

The estimated mean concentration per droplet λ will be used as the variable

Table 4.3: Table comparison of Single-channel droplet classifier methods

Method	**Algorithm**	**Droplet classification rule**
Bio-Rad QuantaSoft	(Undisclosed)	(Undisclosed)
Cloudy	Iterative parameter estimation of μ_p and σ_p using observations within the $\hat{\mu}_p \pm a_p \cdot \hat{\sigma}_p$ of population p at each iteration.[1]	Negative if fluorescence is $< \hat{\mu}_{neg} + 1.5 \cdot a_{neg} \cdot \hat{\sigma}_{neg}$
ddpcRquant	Extreme Value theory to estimate negative population	Negative if fluorescence is less than the average of one hundred 0.995 percentiles of extreme values sampled
definetherain	K-means clustering to cluster 2 populations	Negative if fluorescence is $< \hat{\mu}_{neg} + 3 \cdot \hat{\sigma}_{neg}$; positive if fluorescence is $> \hat{\mu}_{pos} + 3 \cdot \hat{\sigma}_{pos}$; otherwise, droplet is rain.
Umbrella	Non-parametric approach to estimate a 2-component mixture density	Negative if the droplet's probability given the negative population is $> 80\%$
EM-T / EM-skewT	Expectation Maximization to estimate a G-component mixture density	Droplet population membership is where its probability given a population is highest.

[1] a_p is derived from Lievens et. al's own analysis of in-house data. It is calculated as $a_p = 4.55 + 0.35 \cdot \log(k_p) + 0.045 \cdot \log(k_p)^2$; where k_p is the kurtosis of the estimated population p.

Table 4.4: Accessibility of droplet classifier methods

Method	**Accessibility**	**Included in Peformance Evaluation?**
Bio-Rad QuantaSoft	Paid desktop software[1]	No (not available to researcher)
Cloudy	R code[2]	Yes
ddpcRquant	R code[3] and free web app[4]	Yes, only in simulated data (required NTC samples input are not available in Lievens' dataset)
definetherain	Free web app[5]	No (required positive control samples input are not available)
Umbrella	R code[6]	Yes

[1]https://www.bio-rad.com/en-ph/sku/1864011-quantasoft-software-regulatory-edition?ID=1864011

[2]https://github.com/Gromgorgel/ddPCR/blob/master/Cloudy-V2-05.R

[3]https://ddpcrquant.ugent.be/ddpcrquant_functions_qx100.R

[4]http://statapps.ugent.be/dPCR/ddpcrquant/

[5]http://www.definetherain.org.uk/

[6]https://github.com/statOmics/umbrella/blob/master/1D/Umbrella_1d_V1.R

of interest. For the real dataset, since the real value λ is unknown, it is difficult to assess the method's accuracy. For this reason, the precision of λ estimates within sample replicates will be used as a performance metric; this is measured here using the coefficient of variation (CV). Just from the overview of the experiment setup in Table 4.1, it is expected that the dPCR quality for all DNA targets will be different due to the varying parameters. Thus, the estimated λ CV of a DNA target is grouped by the experimental levels per plate. Each level per plate and DNA target in Table 4.1 consists of at least four technical replicates. The exceptions are plate 2, where each primer concentration only has two replicates, and plate 7, where each temperature only has one replicate. For these exceptions, the groupings would only be on the DNA targets, regardless of the experimental levels. In total, for all levels and DNA target groups per plate with the exceptions, 88 CVs will be produced [1], and to summarize these information more effectively, the distribution of the CVs will be assessed instead — first, within plates, and then within DNA targets per plate. In the case of plate 3 where a dilution series was prepared, the log-log regression model of equation 3.6 can be fitted; and for this case, R^2 and RSE will also be used to assess the quantification methods in addition to CV.

In the evaluation of methods using simulated data, since the true value of λ is known, then the linear regression $\lambda = \beta_0 + \hat{\lambda}\beta_1 + \epsilon$ can be fitted. The perfect scenario is that the estimated $\hat{\lambda}$ is always equal to the true value of λ, yet statistically improbable. Thus, it is ideal that a method's $\hat{\lambda}$ estimates will produce fitted coefficients with values very close to $b_0 = 0$ and $b_1 = 1$, and adequacy metrics of $R^2 = 1$ and RSE=0.

[1] Total levels x DNA target groups for all plates, where the levels of plates 2 and 7 are combined. Then this means plates 1, 2, and 7 have 1×12 groups; plates 3 and 6 have 6×2 groups; plate 4 has 3×2 groups; plate 5 has 8×2 groups; and plate 9 has 1×10 groups. Adding up to $3(12) + 2(62) + (32) + (8 * 3) + (10) = 88$ groups.

References

Aas, K., & Haff, I. H. (2006). The generalized hyperbolic skew Student's t-distribution. *Journal of Financial Econometrics*, *4*(2), 275–309. doi: 10.1093/jjfinec/nbj006

Abed, Y., Carbonneau, J., L'Huillier, A. G., Kaiser, L., & Boivin, G. (2017). Droplet digital PCR to investigate quasi-species at codons 119 and 275 of the A(H1N1)pdm09 neuraminidase during zanamivir and oseltamivir therapies. *Journal of Medical Virology*, *89*(4), 737–741. doi: 10.1002/jmv.24680

Advancing scientific discovery and improving healthcare for over 65 years. (n.d.). Retrieved from `https://www.bio-rad.com/`

Ahuja, R., Solanki, A., & Nayyar, A. (2019). Movie recommender system using k-means clustering and k-nearest neighbor. *Proceedings of the 9th International Conference On Cloud Computing, Data Science and Engineering, Confluence 2019*, 263–268. doi: 10.1109/CONFLUENCE.2019.8776969

Arvia, R., Sollai, M., Pierucci, F., Urso, C., Massi, D., & Zakrzewska, K. (2017). Droplet digital PCR (ddPCR) vs quantitative real-time PCR (qPCR) approach for detection and quantification of Merkel cell polyomavirus (MCPyV) DNA in formalin fixed paraffin embedded (FFPE) cutaneous biopsies. *Journal of Virological Methods*, *246*(November 2016), 15–20. Retrieved from `http://dx.doi.org/10.1016/j.jviromet.2017.04.003` doi: 10.1016/j.jviromet.2017.04.003

Attali, D., Bidshahri, R., Haynes, C., & Bryan, J. (2016). Ddpcr: An R package and web application for analysis of droplet digital PCR data. *F1000Research*, *5*, 1–11. doi: 10.12688/F1000RESEARCH.9022.1

Barndorff-Nielsen, O. E. (1977, March). Exponentially decreasing distributions for the logarithm of particle size. *Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences*, *353*(1674), 401–419. Retrieved from `https://doi.org/10.1098/rspa.1977.0041` doi: 10.1098/rspa.1977.0041

Basanisi, M. G., La Bella, G., Nobili, G., Coppola, R., Damato, A. M., Cafiero, M. A., & La Salandra, G. (2020). Application of the novel Droplet digital PCR technology for identification of meat species. *International Journal of*

Food Science and Technology, *55*(3), 1145–1150. doi: 10.1111/ijfs.14486

Bibby, B. M., & Sørensen, M. (2003). *Hyperbolic Processes in Finance*. Woodhead Publishing Limited. Retrieved from http://dx.doi.org/10.1016/B978-044450896-6.50008-X doi: 10.1016/b978-044450896-6.50008-x

Bio-Rad. (2019). *QX200 ™ Droplet Reader and QuantaSoft ™ Software Instruction Manual*.

Blaya, J., Lloret, E., Santísima-Trinidad, A. B., Ros, M., & Pascual, J. A. (2016). Molecular methods (digital PCR and real-time PCR) for the quantification of low copy DNA of Phytophthora nicotianae in environmental samples. *Pest Management Science*, *72*(4), 747–753. doi: 10.1002/ps.4048

Brink, B. G., Meskas, J., & Brinkman, R. R. (2018). DdPCRclust: An R package and Shiny app for automated analysis of multiplexed ddPCR data. *Bioinformatics*, *34*(15), 2687–2689. doi: 10.1093/bioinformatics/bty136

Browne, R. P., & McNicholas, P. D. (2015, jun). A mixture of generalized hyperbolic distributions. *Canadian Journal of Statistics*, *43*(2), 176–198. Retrieved from http://doi.wiley.com/10.1002/cjshttp://doi.wiley.com/10.1002/cjs.11246 doi: 10.1002/cjs.11246

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*(April), 459–474. Retrieved from http://dx.doi.org/10.1016/j.bios.2016.09.082 doi: 10.1016/j.bios.2016.09.082

Capobianco, J. A., Clark, M., Cariou, A., Leveau, A., Pierre, S., Fratamico, P., ... Armstrong, C. M. (2020). *Detection of Shiga toxin-producing Escherichia coli (STEC) in beef products using droplet digital PCR* (Vol. 319). Elsevier B.V. Retrieved from https://doi.org/10.1016/j.ijfoodmicro.2019.108499 doi: 10.1016/j.ijfoodmicro.2019.108499

Chen, D. F., Zhang, L. J., Tan, K., & Jing, Q. (2018). Application of droplet digital PCR in quantitative detection of the cell-free circulating circRNAs. *Biotechnology and Biotechnological Equipment*, *32*(1), 116–123. Retrieved from https://doi.org/10.1080/13102818.2017.1398596 doi: 10.1080/13102818.2017.1398596

Chen, J., Zhang, Y., Chen, C., Zhang, Y., Zhou, W., & Sang, Y. (2020). Identification and quantification of cassava starch adulteration in different food starches by droplet digital PCR. *PLoS ONE*, *15*(2), 1–16. Retrieved from http://dx.doi.org/10.1371/journal.pone.0228624 doi: 10.1371/journal.pone.0228624

Choy, S. K., Lam, S. Y., Yu, K. W., Lee, W. Y., & Leung, K. T. (2017). Fuzzy model-based clustering and its application in image segmentation. *Pattern Recognition*, *68*, 141–157. Retrieved from http://dx.doi.org/10.1016/j.patcog.2017.03.009 doi: 10.1016/j.patcog.2017.03.009

Cook, L., Atienza, E. E., Bagabag, A., Obrigewitch, R. M., & Jerome, K. R.

(2009). Comparison of methods for extraction of viral DNA from cellular specimens. *Diagnostic Microbiology and Infectious Disease*, *64*(1), 37–42. Retrieved from http://dx.doi.org/10.1016/j.diagmicrobio.2009.01.003 doi: 10.1016/j.diagmicrobio.2009.01.003

Corbisier, P., Pinheiro, L., Mazoua, S., Kortekaas, A. M., Chung, P. Y. J., Gerganova, T., ... Emslie, K. (2015). DNA copy number concentration measured by digital and droplet digital quantitative PCR using certified reference materials. *Analytical and Bioanalytical Chemistry*, *407*(7), 1831–1840. doi: 10.1007/s00216-015-8458-z

Dagata, J. a., Farkas, N., & Kramer, J. a. (2016). Method for Measuring the Volume of Nominally 100 μm Diameter Spherical Water-in-Oil Emulsion Droplets. *NIST Special Publication*. Retrieved from http://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.260-184.pdf doi: 10.6028/NIST.SP.260-184

Dang, U. J., Gallaugher, M. P. B., Browne, R. P., & McNicholas, P. D. (2019, jul). *Model-based clustering and classification using mixtures of multivariate skewed power exponential distributions* (No. 1965). Retrieved from http://arxiv.org/abs/1907.01938

Demeke, T., & Dobnik, D. (2018). Critical assessment of digital PCR for the detection and quantification of genetically modified organisms. *Analytical and Bioanalytical Chemistry*, *410*(17), 4039–4050. doi: 10.1007/s00216-018-1010-1

Deng, X., Custer, B. S., Busch, M. P., Bakkour, S., & Lee, T. H. (2017). Simultaneous estimation of detection sensitivity and absolute copy number from digital PCR serial dilution. *Computational Biology and Chemistry*, *68*, 1–5. Retrieved from http://dx.doi.org/10.1016/j.compbiolchem.2017.01.015 doi: 10.1016/j.compbiolchem.2017.01.015

Dobnik, D., Štebih, D., Blejec, A., Morisset, D., & Žel, J. (2016). Multiplex quantification of four DNA targets in one reaction with Bio-Rad droplet digital PCR system for GMO detection. *Scientific Reports*, *6*(September), 1–9. doi: 10.1038/srep35451

Dong, L., Meng, Y., Sui, Z., Wang, J., Wu, L., & Fu, B. (2015). Comparison of four digital PCR platforms for accurate quantification of DNA copy number of a certified plasmid DNA reference material. *Scientific Reports*, *5*(August). doi: 10.1038/srep13174

Dreo, T., Pirc, M., Ramšak, Ž., Pavšič, J., Milavec, M., Žel, J., & Gruden, K. (2014). Optimising droplet digital PCR analysis approaches for detection and quantification of bacteria: A case study of fire blight and potato brown rot. *Analytical and Bioanalytical Chemistry*, *406*(26), 6513–6528. doi: 10.1007/s00216-014-8084-1

Elmer, P. (2000). *An Introduction to Fluorescence Spectroscopy.* Post Office Lane, Beaconsfield, Buckinghamshire: PerkinElmer, Inc. Retrieved from

http://books.google.com/books?id=GgFXweh0hmQC{\&}pgis=1 doi: 10.1194/jlr.M022798

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861–874. doi: 10.1016/j.patrec.2005.10.010

Garriga, J., Palmer, J. R., Oltra, A., & Bartumeus, F. (2016). Expectation-maximization binary clustering for behavioural annotation. *PLoS ONE*, *11*(3), 1–26. doi: 10.1371/journal.pone.0151984

Gerdes, L., Iwobi, A., Busch, U., & Pecoraro, S. (2016). Optimization of digital droplet polymerase chain reaction for quantification of genetically modified organisms. *Biomolecular Detection and Quantification*, *7*(March), 9–20. Retrieved from http://dx.doi.org/10.1016/j.bdq.2015.12.003 doi: 10.1016/j.bdq.2015.12.003

GeyserTimes. (2017). *Eruptions of Old Faithful Geyser, May 2014 [online database]* . https://geysertimes.org.

Gou, T., Hu, J., Wu, W., Ding, X., Zhou, S., Fang, W., & Mu, Y. (2018). Smartphone-based mobile digital PCR device for DNA quantitative analysis with high accuracy. *Biosensors and Bioelectronics*, *120*(August), 144–152. Retrieved from https://doi.org/10.1016/j.bios.2018.08.030 doi: 10.1016/j.bios.2018.08.030

Hall Sedlak, R., & Jerome, K. R. (2014, may). The potential advantages of digital PCR for clinical virology diagnostics. *Expert Review of Molecular Diagnostics*, *14*(4), 501–507. Retrieved from http://www.tandfonline.com/doi/full/10.1586/14737159.2014.910456 doi: 10.1586/14737159.2014.910456

Hamaguchi, M., Shimabukuro, H., Hori, M., Yoshida, G., Terada, T., & Miyajima, T. (2018). Quantitative real-time polymerase chain reaction (PCR) and droplet digital PCR duplex assays for detecting Zostera marina DNA in coastal sediments. *Limnology and Oceanography: Methods*, *16*(4), 253–264. doi: 10.1002/lom3.10242

Harvey, D. (2010, October). Analytical chemistry 2.0—an open-access digital textbook. *Analytical and Bioanalytical Chemistry*, *399*(1), 149–152. Retrieved from https://doi.org/10.1007/s00216-010-4316-1 doi: 10.1007/s00216-010-4316-1

Hindson, C. M., Chevillet, J. R., Briggs, H. A., Gallichotte, E. N., Ruf, I. K., Hindson, B. J., ... Tewari, M. (2013, oct). Absolute quantification by droplet digital PCR versus analog real-time PCR. *Nature Methods*, *10*(10), 1003–1005. Retrieved from file:///C:/Users/CarlaCarolina/Desktop/Artigosparaacrescentarnaqualifica{\c{c}}{\~{a}}o/Theimpactofbirthweightoncardiovasculardiseaseriskinthe.pdfhttp://www.nature.com/articles/nmeth.2633 doi: 10.1038/nmeth.2633

Hu, Y., & Smyth, G. K. (2009, aug). ELDA: Extreme limiting dilution ana-

lysis for comparing depleted and enriched populations in stem cell and other assays. *Journal of Immunological Methods*, *347*(1-2), 70–78. Retrieved from http://dx.doi.org/10.1016/j.jim.2009.06.008https://linkinghub.elsevier.com/retrieve/pii/S0022175909001951 doi: 10.1016/j.jim.2009.06.008

Huggett, J. F., Foy, C. A., Benes, V., Emslie, K., Garson, J. A., Haynes, R., ... Bustin, S. A. (2013). The digital MIQE guidelines: Minimum information for publication of quantitative digital PCR experiments. *Clinical Chemistry*, *59*(6), 892–902. doi: 10.1373/clinchem.2013.206375

Huggett, J. F., O'Grady, J., & Bustin, S. (2015). QPCR, dPCR, NGS - A journey. *Biomolecular Detection and Quantification*, *3*(March 2007), A1–A5. doi: 10.1016/j.bdq.2015.01.001

Hussain, M., & Bowers, J. (2017). A Droplet Digital PCR Method for CHO Host Residual DNA Quantification in Biologic Drugs. *Journal of Analytical & Pharmaceutical Research*, *4*(3), 8–11. doi: 10.15406/japlr.2017.04.00107

Jacobs, B. K., Goetghebeur, E., & Clement, L. (2014). Impact of variance components on reliability of absolute quantification using digital PCR. *BMC Bioinformatics*, *15*(1), 1–13. doi: 10.1186/1471-2105-15-283

Jacobs, B. K., 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. doi: 10.1021/acs.analchem.6b04208

Jahne, M. A., Brinkman, N. E., Keely, S. P., Zimmerman, B. D., Wheaton, E. A., & Garland, J. L. (2020, feb). Droplet digital PCR quantification of norovirus and adenovirus in decentralized wastewater and graywater collections: Implications for onsite reuse. *Water Research*, *169*, 115213. Retrieved from https://doi.org/10.1016/j.watres.2019.115213https://linkinghub.elsevier.com/retrieve/pii/S004313541930987X doi: 10.1016/j.watres.2019.115213

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. Springer New York. Retrieved from https://doi.org/10.1007/978-1-4614-7138-7 doi: 10.1007/978-1-4614-7138-7

Jones, G. M., Busby, E., Garson, J. A., Grant, P. R., Nastouli, E., Devonshire, A. S., & Whale, A. S. (2016, dec). Digital PCR dynamic range is approaching that of real-time quantitative PCR. *Biomolecular Detection and Quantification*, *10*, 31–33. Retrieved from http://dx.doi.org/10.1016/j.bdq.2016.10.001https://linkinghub.elsevier.com/retrieve/pii/S2214753516300316 doi: 10.1016/j.bdq.2016.10.001

Jones, M., Williams, J., Gärtner, K., Phillips, R., Hurst, J., & Frater, J. (2014). Low copy target detection by Droplet Digital PCR through application of a novel open access bioinformatic pipeline, 'definetherain'. *Journal of Viro-*

logical Methods, *202*, 46–53. Retrieved from http://dx.doi.org/10.1016/j.jviromet.2014.02.020 doi: 10.1016/j.jviromet.2014.02.020

Köppel, R., & Bucher, T. (2015). Rapid establishment of droplet digital PCR for quantitative GMO analysis. *European Food Research and Technology*, *241*(3), 427–439. doi: 10.1007/s00217-015-2475-1

Košir, A. B., Divieto, C., Pavšič, J., Pavarelli, S., Dobnik, D., Dreo, T., ... Žel, J. (2017). Droplet volume variability as a critical factor for accuracy of absolute quantification using droplet digital PCR. *Analytical and Bioanalytical Chemistry*, *409*(28), 6689–6697. doi: 10.1007/s00216-017-0625-y

Kramer, M. F., & Coen, D. M. (2001). Enzymatic amplification of DNA by PCR: standard procedures and optimization. *Current protocols in immunology / edited by John E. Coligan ... [et al.]*, *Chapter 10*. doi: 10.1002/0471142727.mb1501s56

Kreutz, J. E., Munson, T., Huynh, T., Shen, F., Du, W., & Ismagilov, R. F. (2011, nov). Theoretical Design and Analysis of Multivolume Digital Assays with Wide Dynamic Range Validated Experimentally with Microfluidic Digital PCR. *Analytical Chemistry*, *83*(21), 8158–8168. Retrieved from https://pubs.acs.org/doi/10.1021/ac201658s doi: 10.1021/ac201658s

Kung, S. Y., Mak, M. W., & Lin, S. H. (2004). Expectation-Maximization Theory. In *Biometric authentication: A machine learning approach* (pp. 50–84). Retrieved from http://ptgmedia.pearsoncmg.com/images/0131478249/samplechapter/0131478249{_}ch03.pdf

Lai, K. K. Y., Cook, L., Wendt, S., Corey, L., & Jerome, K. R. (2003). Evaluation of real-time PCR versus PCR with liquid-phase hybridization for detection of enterovirus RNA in cerebrospinal fluid. *Journal of Clinical Microbiology*, *41*(7), 3133–3141. doi: 10.1128/JCM.41.7.3133-3141.2003

Li, K., Ma, Z., Robinson, D., & Ma, J. (2018). Identification of typical building daily electricity usage profiles using Gaussian mixture model-based clustering and hierarchical clustering. *Applied Energy*, *231*(May), 331–342. Retrieved from https://doi.org/10.1016/j.apenergy.2018.09.050 doi: 10.1016/j.apenergy.2018.09.050

Li, X., Fu, Y., Wang, X., Demeo, D. L., Tantisira, K., Weiss, S. T., & Qiu, W. (2018). Detecting Differentially Variable MicroRNAs via Model-Based Clustering. *International Journal of Genomics*, *2018*. doi: 10.1155/2018/6591634

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. doi: 10.1371/journal.pone.0153317

López, S. O., García-Olmo, D. C., García-Arranz, M., Guadalajara, H., Pastor, C., & García-Olmo, D. (2016). KRAS G12V mutation detection by droplet digital PCR in circulating cell-free DNA of colorectal cancer patients. *International Journal of Molecular Sciences*, *17*(4), 1–9. doi:

10.3390/ijms17040484

Mauvisseau, Q., Davy-Bowker, J., Bulling, M., Brys, R., Neyrinck, S., Troth, C., & Sweet, M. (2019, dec). Combining ddPCR and environmental DNA to improve detection capabilities of a critically endangered freshwater invertebrate. *Scientific Reports*, *9*(1), 14064. Retrieved from http://dx.doi.org/10.1038/s41598-019-50571-9http://www.nature.com/articles/s41598-019-50571-9 doi: 10.1038/s41598-019-50571-9

McNicholas, P. D. (2016, oct). Model-Based Clustering. *Journal of Classification*, *33*(3), 331–373. Retrieved from http://link.springer.com/10.1007/s00357-016-9211-9 doi: 10.1007/s00357-016-9211-9

Mittal, K., Aggarwal, G., & Mahajan, P. (2019, sep). Performance study of K-nearest neighbor classifier and K-means clustering for predicting the diagnostic accuracy. *International Journal of Information Technology*, *11*(3), 535–540. Retrieved from https://doi.org/10.1007/s41870-018-0233-xhttp://link.springer.com/10.1007/s41870-018-0233-x doi: 10.1007/s41870-018-0233-x

Necula, C. (2009). Modeling heavy-tailed stock index returns using the Generalized Hyperbolic Distribution. *Romanian Journal of Economic Forecasting*, *10*(2), 118–131.

Nwobi, F. N. (2014). Modeling Electricity Price Returns using Generalized Hyperbolic Distributions. *Communications in Mathematical Finance*, *3*(2), 33–50.

Nystrand, C. F., Ghanima, W., Waage, A., & Jonassen, C. M. (2018, apr). JAK2 V617F mutation can be reliably detected in serum using droplet digital PCR. *International Journal of Laboratory Hematology*, *40*(2), 181–186. Retrieved from http://doi.wiley.com/10.1111/ijlh.12762 doi: 10.1111/ijlh.12762

Persson, S., Eriksson, R., Lowther, J., Ellström, P., & Simonsson, M. (2018, nov). Comparison between RT droplet digital PCR and RT real-time PCR for quantification of noroviruses in oysters. *International Journal of Food Microbiology*, *284*(February), 73–83. Retrieved from https://linkinghub.elsevier.com/retrieve/pii/S0168160518303337 doi: 10.1016/j.ijfoodmicro.2018.06.022

Peterson, L. E. (2009). K-nearest neighbor. *Scholarpedia*, *4*(2), 1883. (revision #137311) doi: 10.4249/scholarpedia.1883

Pinheiro, L. B., Coleman, V. A., Hindson, C. M., Herrmann, J., Hindson, B. J., Bhat, S., & Emslie, K. R. (2012). Evaluation of a droplet digital polymerase chain reaction format for DNA copy number quantification. *Analytical Chemistry*, *84*(2), 1003–1011. doi: 10.1021/ac202578x

Quan, P.-L. L., Sauzade, M., & Brouzes, E. (2018, apr). DPCR: A technology review. *Sensors*, *18*(4), 1271. Retrieved from http://www.mdpi.com/1424

-8220/18/4/1271 doi: 10.3390/s18041271

Reed, G. F., Lynn, F., & Meade, B. D. (2003, nov). Use of Coefficient of Variation in Assessing Variability of Quantitative Assays. *Clinical Diagnostic Laboratory Immunology*, *10*(6), 1162–1162. Retrieved from https://cvi.asm.org/content/10/6/1162 doi: 10.1128/CDLI.10.6.1162.2003

Saiki, R., Gelfand, D., Stoffel, S., Scharf, S., Higuchi, R., Horn, G., ... Erlich, H. (1988). Primer-Directed Enzymatic Amplification of DNA with a Thermostable DNA Polymerase. *Journal of Infectious Diseases*, *158, 1154*-(9).

Sanders, R., Huggett, J. F., Bushell, C. A., Cowen, S., Scott, D. J., & Foy, C. A. (2011, sep). Evaluation of Digital PCR for Absolute DNA Quantification. *Analytical Chemistry*, *83*(17), 6474–6484. Retrieved from https://pubs.acs.org/doi/10.1021/ac103230c doi: 10.1021/ac103230c

Shi, J., & Yang, L. (2020, jan). A Climate Classification of China through k-Nearest-Neighbor and Sparse Subspace Representation. *Journal of Climate*, *33*(1), 243–262. Retrieved from https://journals.ametsoc.org/jcli/article/33/1/243/346140/A-Climate-Classification-of-China-through doi: 10.1175/JCLI-D-18-0718.1

Strain, M. C., Lada, S. M., Luong, T., Rought, S. E., Gianella, S., Terry, V. H., ... Richman, D. D. (2013, apr). Highly Precise Measurement of HIV DNA by Droplet Digital PCR. *PLoS ONE*, *8*(4), e55943. Retrieved from https://dx.plos.org/10.1371/journal.pone.0055943 doi: 10.1371/journal.pone.0055943

Sykes, P. J., Neoh, S. H., Brisco, M. J., Hughes, E., Condon, J., & Morley, A. A. (1992, sep). Quantitation of targets for PCR by use of limiting dilution. *BioTechniques*, *13*(3), 444–9. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/1389177

Tagliapietra, A., Rotondo, J. C., Bononi, I., Mazzoni, E., Magagnoli, F., Gonzalez, L. O., ... Martini, F. (2020). Droplet-digital PCR assay to detect Merkel cell polyomavirus sequences in chorionic villi from spontaneous abortion affected females. *Journal of Cellular Physiology*, *235*(3), 1888–1894. doi: 10.1002/jcp.29213

Takahashi, M., Watanabe, T., & Omori, Y. (2016). Volatility and quantile forecasts by realized stochastic volatility models with generalized hyperbolic distribution. *International Journal of Forecasting*, *32*(2), 437–457. doi: 10.1016/j.ijforecast.2015.07.005

Taylor, S. C., Laperriere, G., & Germain, H. (2017, dec). Droplet Digital PCR versus qPCR for gene expression analysis with low abundant targets: from variable nonsense to publication quality data. *Scientific Reports*, *7*(1), 2409. Retrieved from http://www.nature.com/articles/s41598-017-02217-x doi: 10.1038/s41598-017-02217-x

Tortora, C., ElSherbiny, A., Browne, R. P., Franczak, B. C., McNicholas, P. D., & Amos., D. D. (2019). Mixghd: Model based clustering, classifica-

tion and discriminant analysis using the mixture of generalized hyperbolic distributions [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=MixGHD (R package version 2.3.2)

Trypsteen, W., Vynck, M., de Neve, J., Bonczkowski, P., Kiselinova, M., Malatinkova, E., ... de Spiegelaere, W. (2015, jul). ddpcRquant: threshold determination for single channel droplet digital PCR experiments. *Analytical and Bioanalytical Chemistry*, *407*(19), 5827–5834. Retrieved from http://link.springer.com/10.1007/s00216-015-8773-4 doi: 10.1007/s00216-015-8773-4

Tzonev, S. (2018). Fundamentals of Counting Statistics in Digital PCR: I Just Measured Two Target Copies–What Does It Mean? In *Digital pcr* (Vol. 1768, pp. 25–43). Retrieved from http://link.springer.com/10.1007/978-1-4939-7778-9{_}3 doi: 10.1007/978-1-4939-7778-9_3

Walpole, R. E., Myers, R. H., Myers, S. L., & Ye, K. (2011). *Probability and Statistics for Engineers and Scientists (9th Edition)* (9th ed.). Retrieved from http://www.tandfonline.com/doi/full/10.1080/09332480.2013.845456

Whale, A. S., Cowen, S., Foy, C. A., & Huggett, J. F. (2013). Methods for Applying Accurate Digital PCR Analysis on Low Copy DNA Samples. *PLoS ONE*, *8*(3). doi: 10.1371/journal.pone.0058177

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

Young, H. K., Yang, I., Bae, Y. S., & Park, S. R. (2008). Performance evaluation of thermal cyclers for PCR in a rapid cycling condition. *BioTechniques*, *44*(4), 495–505. doi: 10.2144/000112705

Zhu, Q., Qiu, L., Yu, B., Xu, Y., Gao, Y., Pan, T., ... Mu, Y. (2014). Digital PCR on an integrated self-priming compartmentalization chip. *Lab Chip*, *14*(6), 1176–1185. Retrieved from http://xlink.rsc.org/?DOI=C3LC51327K doi: 10.1039/C3LC51327K

Zhu, Q., Xu, Y., Qiu, L., Ma, C., Yu, B., Song, Q., ... Mu, Y. (2017). A scalable self-priming fractal branching microchannel net chip for digital PCR. *Lab on a Chip*, *17*(9), 1655–1665. Retrieved from http://xlink.rsc.org/?DOI=C7LC00267J doi: 10.1039/C7LC00267J