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.

October 1, 2020

Abstract

The digital PCR (dPCR) is a method to quantify the DNA copies of known strains related to diseases. As a new approach to the gold-standard RT-PCR, further research is required to assess the quality and accuracy of this method. One particular area of dPCR is its novel step in droplet classification that distinguishes it from RT-PCR. As of writing, few droplet classifiers exist in literature as well as the assessment of these methods. This thesis reviews the classification methods of current dPCR quantification tools in literature, 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

Quantification of nucleic acids is a developing research field in molecular biology for the detection and quantification expression levels of genes (Huggett, O'Grady, & Bustin, 2015). These molecules are found in deoxyribonucleic acid (DNA) and ribonucleic acid (RNA), which carries genetic information and is 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).

To be able to determine the concentration of target nucleic acid, detection is naturally a pre-requisite. There are, however, nucleic acid of interests that have very low concentrations to the point that it becomes undetectable in existing detection technologies. This problem is solved by amplifying the nucleic acid 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 can multiply specific nucleic acid sequences in DNA or RNA from low concentrations to millions of copies. This method exposes the nucleic acid sequences mixed with chemical components in a series of 20 to 40 temperature cycles. In each cycle,

PCR doubles the nucleic acid molecule; theoretically producing 2^n molecules after n cycles (Quan et al., 2018).

After PCR amplification, absolute nucleic acid quantification is achieved using digital Polymerase Chain Reaction (dPCR) technique. This equally divides the nucleic acid samples into thousands of partitions; each of these partitions is evaluated as either off or on, or in this context, labeled as positive or negative, hence the term "digital" (Cao et al., 2017).

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 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. The determination of this threshold is not constant for different DNA samples and unclear for assays that produce ambiguous readouts (Trypsteen et al., 2015).

An important aspect in positioning the threshold is the presence of noise features in the data. Poor quality dPCR assays adds 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, which is also termed as 'rain'. The factors affecting these noise characteristics are further explored in the succeeding sections.

In the case of three or more fluorescence populations, this formation may be attributed to a group of inhibited targets that had varying amplification efficiency as the other targets. This may also be the case of high presence of intermediate fluorescences for unoptimized assays (Lievens et al., 2016).

"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 1.1 and 1.2 demonstrates 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 publicly available dataset from the study of Lievens et al. (2016).

Poor quality dPCR assays negatively limit the level of sensitivity and accuracy it may reach. Among the consequence of unoptimized dPCR assays is droplet misclassification, which may lead to serious incorrect conclusions.

The primary problem in misclassification lies in 'rain' droplets; these are partitions that emits an intermediate fluorescence signal that is difficult to classify as positive or negative. Figures 1.1 and 1.2 demonstrates 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 publicly available dataset from the study of Lievens et al. (2016).

Figure 1.1: Fluorescence readings of 4 repititions of DNA target cru

The estimate for DNA target concentration highly depends on the positive and negative classifications. For experiments with low copy DNA targets, the focus is

Figure 1.2: Fluorescence readings of 4 repititions of DNA target TC1507

on maximizing the sensitivity of the test for these very small number of positive droplets. Low sensitivity translates to failure in detecting lower concentrations.

For assays with large differences in the distance between the two fluorescence groups (positive and negative droplets), most quantification tools estimate the target concentrations with high sensitivity. As exhibited in an optimized assay experiment of E. amylovora (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, produces unsatisfactory analytical sensitivity of the concentration estimates.

The danger of false rates due to misclassification is expounded at the clinical level, where false rates lead 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, even for a small FPR, 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 error and sources of variations, each of which is the possible source of such noise. The dPCR workflow, as illustrated in Figure 1.3, is usually a sequential procedure of extracting the sample from an organism, concocting the sample with several chemical components into a reaction mix, distributing the reaction mix to equal partitions, amplifying and detecting the target molecules using PCR, and the concentration is then finally estimated using a Poisson correction factor. In (Jacobs, Goetghebeur, & Clement, 2014), it was emphasized that every step of the dPCR workflow inevitably allows for the introduction of different sources of variation.

These variance components within the dPCR workflow is shown in Figure 1.4.

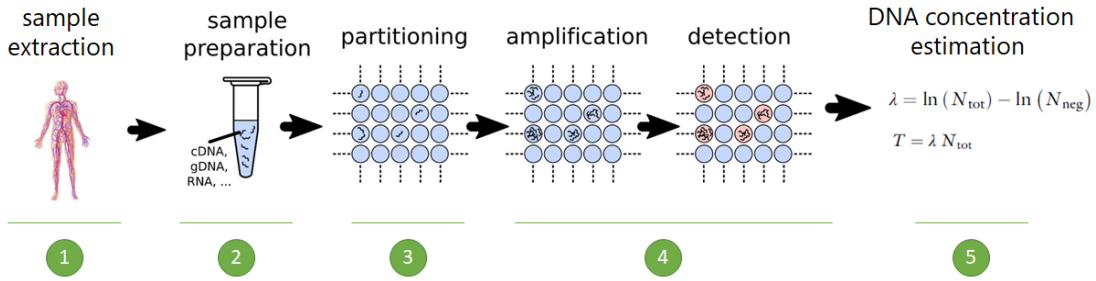

Figure 1.3: The dPCR workflow

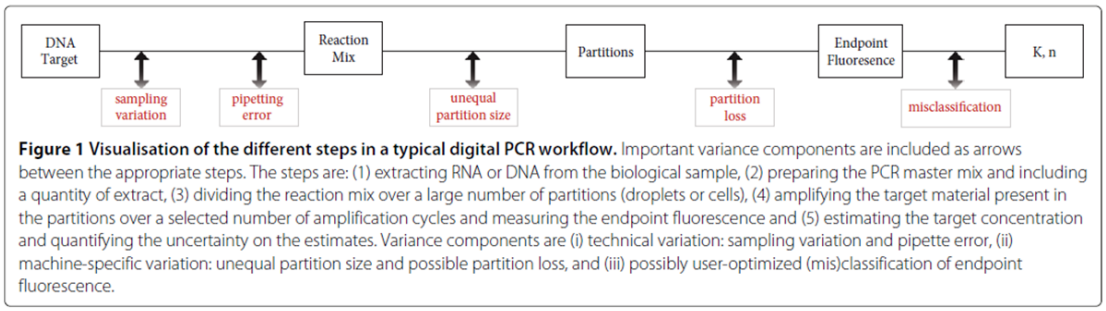

Figure 1 Visualisation of the different steps in a typical digital PCR workflow. Important variance components are included as arrows between the appropriate steps. The steps are: (1) extracting RNA or DNA from the biological sample, (2) preparing the PCR master mix and including a quantity of extract, (3) dividing the reaction mix over a large number of partitions (droplets or cells), (4) amplifying the target material present in the partitions over a selected number of amplification cycles and measuring the endpoint fluorescence and (5) estimating the target concentration and quantifying the uncertainty on the estimates. Variance components are (i) technical variation: sampling variation and pipette error, (ii) machine-specific variation: unequal partition size and possible partition loss, and (iii) possibly user-optimized (mis)classification of endpoint fluorescence.

Figure 1.4: Potential variation components between steps of the dPCR workflow

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 separates the two populations better (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 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.

Preparing the reaction mix is a delicate process that strictly requires the accuracy of the pipetting volume; and yet, technical variation still occurs that results in pipetting errors. In addition to physical variation, chemical and biological

factors play a role in the dPCR assay quality, such as target sequence variation, concentration of polymerase, MgCI2, dNTPs, and primers (Köppel & Bucher, 2015; Kramer & Coen, 2001), dye or probe quencher (Witte et al., 2016), and the 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 identification of the factor directly influencing the noise present 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 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. Even in unoptimized assays, Demeke and Dobnik (2018) were still able to produce reliable results upon the automatic threshold set that enabled for the increased repeatability. Reliable automated threshold systems eliminate the operator bias from manually adjusting the threshold. When dealing with rain, some researchers exclude it in the final droplet counts (M. Jones et al., 2014a), 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). Some researchers. In the case of droplet volume variability, correction-factor must be taken to improve the agreement of estimates (Demeke & Dobnik, 2018). For unresolved noise features, it is important to take this uncertainty into account and include it in the final estimate error.

Expounding further on the automated threshold setting approach, such 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 areas of improvement as automatic systems are said to be not yet the best option for the cases of abundant intermediate fluorescence (Demeke & Dobnik, 2018). Ideally, this system should be able to provide precise estimates regardless of dPCR assay quality, and thus, reducing the negative impact from compromising quality with time. 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).

1.3 Statement of the Problem

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

1. Find initialization parameters and fit G-component mixture densities on dPCR droplet fluorescence intensities.
2. Utilize EM clustered droplets to provide precise quantification estimates for DNA samples with varying amounts of "noise" and concentration.
3. Evaluate and compare the precision and bias of the estimates amongst the existing classification 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 to risk for false-positives. A modern approach to DNA and RNA target quantification is through the dPCR method. In one of the steps of the dPCR workflow, the classification of droplet fluorescence still has areas of 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: Quantasoft propriety software, definetherain (M. Jones et al., 2014a), manual global threshold (Dreo et al., 2014), cloudy (Lievens et al., 2016), and Umbrella (Jacobs et al., 2017). Most of the aforementioned droplet classifying 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, multimodal distributions are considered as not to restrict the possibility of multiple fluorescence populations. The method used 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 fluorescence datasets from published research papers. Only two were found and will be used for statistical analysis, namely from Lievens et al. (2016) and M. Jones et al. (2014a). 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 setting and nucleic

acid targets.

Lastly, statistical results presented may lack biological explanations which could be useful for explaining the variances of the droplet fluorescence. Such information may be utilized to further improve the estimation process.

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. Detection is crucial in analyzing low-copy molecules such as in viruses like the HIV-1 DNA and 2-LTR circles (Henricha, Gallienb, Lia, Florencia Pereyraa, & Kuritzkesa, 2012). As mentioned in section 1.2, the dPCR workflow may introduce several sources of variation, including human error in pipetting techniques, to the use of diluting DNA samples to lessen the Poisson variability in a positive partition. This section explores the statistical methods in current quantification systems of single-channel dPCR experiments. The terms ddPCR (droplet dPCR) and dPCR are synonymous and may be used interchangeably.

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 negative otherwise.

One popular tool incorporated with automatic thresholding is the QuantaSoft software. The QuantaSoft software is the dPCR analysis tool that comes with the Bio-Rad droplet dPCR System package. It allows for the setting up of sample and experiments, running and controlling the instrument, and finally, the analysis of the nucleic acid concentration (Bio-Rad, 2019). According to the Bio-Rad

Laboratories website (https://www.bio-rad.com), it has been a leading product developer for 65 years in the research fields of life science and clinical diagnostics. Among its popular focus areas, dPCR is one of its most featured technology, providing ddPCR instruments; kits, reagents and 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, 2017a; Nystrand, Ghanima, Waage, & Jonassen, 2018), food safety (J. Chen et al., 2020; Capobianco et al., 2020; Basanisi et al., 2020), up to 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; a demonstration is shown in Figure 2.1. As with other automated tools, its documentation recommends reviewing this threshold to make changes if needed; and thus, manually setting the threshold is also allowed. Unfortunately, the calculation of the automatic threshold is not publicly available.

The evaluation of the QuantaSoft software show 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. Among the several reasons for this high detection, Basanisi et al. (2020) emphasizes the need for a highly sensitive and specific test at the molecular level.

Assessing the QuantaSoft software's ability to classify droplets, it was shown that for samples exhibiting a substantial amount of intermediate fluorescence, the system fails to determine a threshold (outputs "No call"). This was the case in the bacteria study of Dreo et al. (2014), that were ran on very high concentrations. 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.

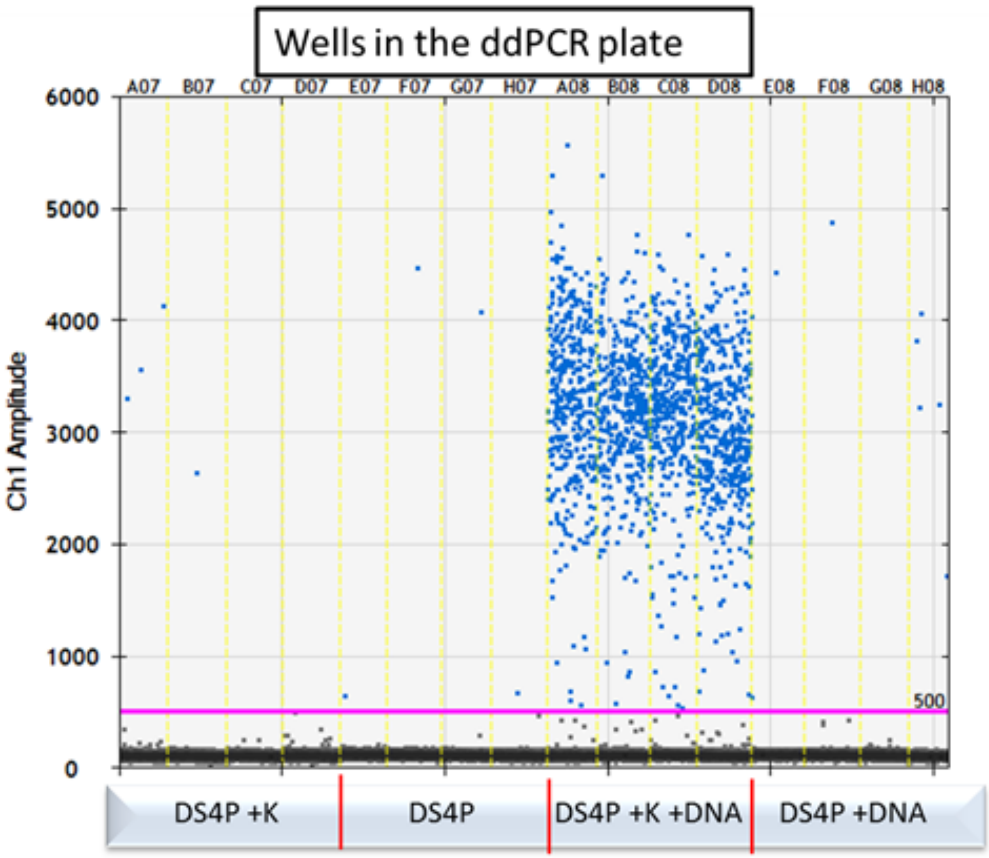

Figure 2.1: QuantaSoft threshold in pink line from (Hussain & Bowers, 2017b)

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 a well-optimised 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 external researchers.

2.1.3 definetherain

Based on the research papers curated by Peterson (2009), K-nearest-neighbor (KNN) is a 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. (2014a), called definetherain, utilizes the KNN algorithm in identifying 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 values are classified as rain. These cutoff values 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.

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 as well as optimization parameters. The

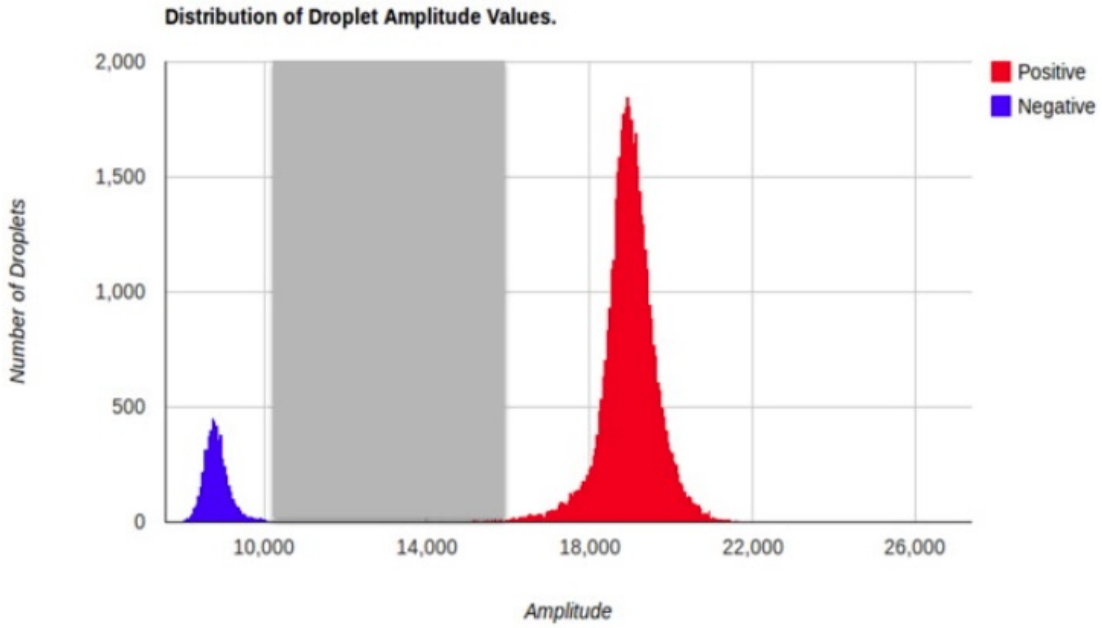

Figure 2.2: Determined thresholds calculated by definetherain, reprinted from (M. Jones et al., 2014b)

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 refered here as cloudy.

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 Guassian kernel density of the fluorescence is estimated with a minimum bandwith 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 $\hat{\sigma}$. Assuming normality, the latter is estimated as half the peak width at 60-65% of its maximum height.
4. Refine the estimates using a reiterative method, first initialized with $a = 4$.

5. Re-estimate $\hat{\mu}$ and $\hat{\sigma}$ 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 distribution

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 rain is included as a category, 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 rain categorization is not of interest, then a threshold $\theta = \hat{\mu_n} + 1.5 \cdot a_n + \hat{\sigma_n}$ is calculated; where n is a population.

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 nucleic acid targets. After determining the final population parameters, a range or a single threshold 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. Figure 2.3 left panel reveals the distribution of negative droplets to be heavily skewed to the right, thereby causing their exclusion to be labeled as negative due to the symmetry of the categorization rule (right panel).

It is noted that the context of the threshold setting here is based on in-house data, and may not necessarily conclude as a classifier for external experiments. In their study, the cloudy algorithm was able to produce estimates for differing PCR experimental factors, such as sonication, PCR enhancers, annealing conditions, and number of cycles, to achieve optimization or diminishing rain droplets for dPCR experiments.

2.1.5 Umbrella

As opposed to the distance-based clustering in Section 2.1.3, a probabilistic approach is achieved with model-based clustering. According to McNicholas (2016), a finite mixture model is a sum of weighted density components. This mixture

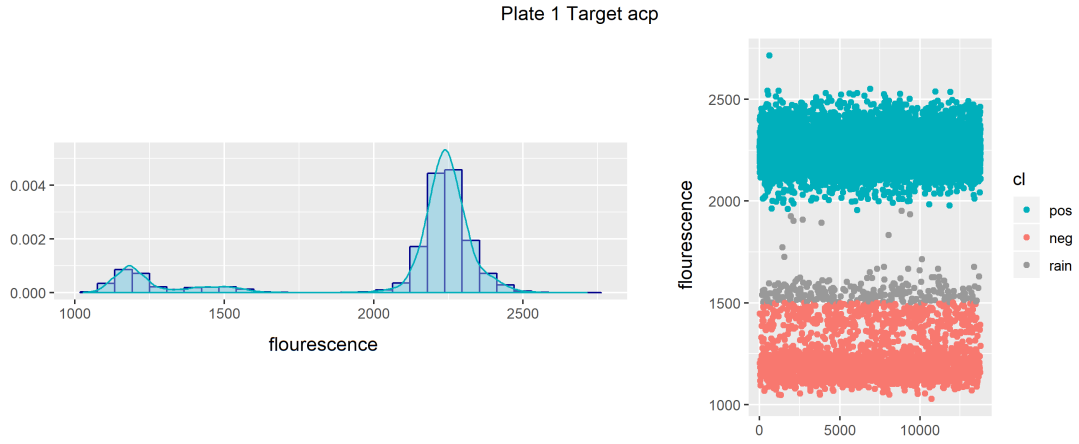

Figure 2.3: One replicate of the DNA target acp Plate 1 from Lievens et al. (2016) dataset. Left panel shows the fluorescence densities. Right panel is the result of droplet categorization using cloudy

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 are defined as a cluster; and each observation has a calculated probability of it belonging to a cluster.

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. The assumption of the Gaussian component distributions are common in model-based clustering; however, in the succeeding texts, dPCR droplet fluorescence populations are claimed to be not normally distributed.

In the application dPCR, Jacobs et al. (2017) developed Umbrella, a model-based clustering for dPCR droplets using nonparametric density estimation (https://github.com/sta Umbrella requires 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 procedure 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 partitions of a target partition set A is assumed to have a mixture density

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

   - $p_{0,A}$ = 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. Begin estimating the parameters by first aligning the modes of the null component of $f_A(x)$ and the NTC reference $f_0(x)$.

4. Discretize the aligned distributions by generating a histogram with the same bins.

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 NA 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 of the 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 nucleic acid 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 are 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 subtracting the fluorescence intensities of each sample by 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 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 calculated 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, corrects for baseline shifts. In calculating the target concentration, it does not discard any droplet as it states can underestimate the true concentration.

According to the authors' evaluation, 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, and may be caused by no baseline correction between NTC and target samples, and by its assumption of NTC being normally distributed without fitting experiments. Additionally, QuantaSoft fails to quantify concentration in some samples that result in "No call" outputs.

2.2 Expectation-Maximization (EM) Clustering

Recall from Section 2.1.5, model-based clustering refers to fitting a finite mixture model given the 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.

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. As mentioned in the model-based clustering examples in Section 2.1.5, the mixture components are frequently assumed to follow a Gaussian distribution, and is also known as Gaussian Mixture Model (GMM). Although popular, GMMs poorly fit the 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.

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 2.3, 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 are 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 lead 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 is 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.

An important metric is the λ, which denotes the mean number of copies per partition. 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 as 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 in 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 CVs of the estimated target copies/ml. 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 ranging from 10^{-4} to 10^{-9}. Their initial purpose of performing 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_0, can be estimated by finding the relationship between λ and D (serial dilution factor). The mean target copies per partition can also be expressed as $\lambda = (c_0 \times D)/N$. By taking the logarithms on both sides, $-\log -ln(\lambda) = -\log D - \log \frac{C_0}{N}$ is derived, and a linear relationship between $-\log -ln(\lambda)$ and $-\log D$ is established. When a lin-

ear regression is fitted, C_0 can be derived using the intercept. In order to compare their dPCR device amongst a commercial dPCR system, Zhu et al. (2017) has used C_0, observing that their device matches the C_0 of the commercial system. In addition, they have also recorded the CV of the replicates for each dilution step. Besides estimating C_0, 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 an R^2 of 0.9994 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 known as prepared by a researcher, these samples are known as a positive control sample. In their evaluation of 'definetherain', a dPCR quantification system with focus on low-copy counts (M. Jones et al., 2014a), 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 $\log_{10}(\text{CopyNumber} + 1)$. 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 PCR, fluorescent probes and primers attach to target sequences in a sample of DNA templates. This sample in the form of a master mix, is then partitioned into thousands of droplets. When the DNA molecules become excited, each droplet emits a fluorescence endpoint intensity F, which are then used to identify if a droplet is positive or negative. A threshold defines the demarcation line to classify positive from negative droplets, such that any intensity less than the threshold is a negative droplet, and positive otherwise (Trypsteen et al., 2015).

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). The succeeding concentrations in this paper are in terms of target molecule counts per μL, unless otherwise specified.

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, of interest is the unknown stock concentration, c_1. which can be derived from the formula above as

$$c_1 = c_2 \times \frac{1}{D}.$$

Let λ be the concentration in terms of the average molecule count per droplet. A droplet (analogous to partition, chamber well, or reaction) of the diluted sample has a known constant volume (nL), V_{drp}. Then, the unknown diluted concentration, c_2, can then be derived as

$$c_2 = \hat{\lambda} \times \frac{1000}{V_{drp}}.$$

In another representation of λ, it is defined as the stock concentration of a droplet, c_{drp}, multiplied to the dilution factor (Zhu et al., 2014)

$$\lambda = c_{drp} \times D \tag{3.1}$$

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 the 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}. According to the Poisson distribution, the expected value of the target copies X per droplet is λ. Before deriving a formula for λ, it is first noted that the droplet outcomes follow a binomial distribution with parameters n, the number of independent trials, and p, the number of successes. Using the MLE of p, the probability of getting a success for each trial is estimated as $\hat{p} = \frac{x}{n}$; where x is the observed successes in n trials. By defining trial as a droplet, and success as a positive droplet, then this consequently means that the probability of getting a positive for one droplet can be estimated as $P(x > 0) = \frac{N_{pos}}{N_{tot}}$, where N_{tot} and N_{pos} is the count of the total and positive droplets in the assay, respectively. Combining the concepts from the Poisson and binomial distributions, the expected value of X (target copies), λ, can then be derived as follows :

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

where N_{neg} is the negative droplet count (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 3.1, for a given droplet with stock conconcentration c_{drp} of dilution factor D_i, the expected target copies per droplet is λ_i. Thus, it can be said that with a fixed quantity c_{drp}, D_i is a predictor of λ_i. Its relationship in equation 3.1 can then be linearized by taking the logarithm on both sides

$$-\log \lambda = -\log c_{drp} - \log D, \tag{3.2}$$

Then, 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. Defining c_{drp} as the slope β_0, then the final model is

$$-\log \lambda_i = -\log c_{drp} - \log D_i \beta_1 \tag{3.3}$$

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

Equation 3.3 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 a 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 $beta_0$ and $beta_1$ are

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

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

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 the 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 is able to predict 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 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.

3.3 EM Clustering

3.3.1 G-component Finite Mixture Density

Denote $X = x_1, ..., x_N; x_i \epsilon \mathbb{R}^P$ as a statistically independent observation sequence where N is the number of observations and P is the dimensionality of x_i. The G-component Finite Mixture Density is

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

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

3.3.2 Model-based clustering

Denote $C = C_1, ..., C_G$ is the set of cluster mixture labels. $Z = z_1, ..., z_N; z_i \epsilon C$ is the set of "hidden" states where $z_i = C_g$ means that the gth mixture generated x_i.

$$\delta_{gi} = \delta(z_i, C_g) = \begin{cases} 1 & \text{if } x_i \text{ is generated by mixture } C_g \\ 0 & \text{otherwise} \end{cases}$$

Suppose $X = x_1, ..., x_N$ of N p-dimensional data vectors are observed and all N are unlabelled or treated as unlabelled. The likelihood is then expressed as

$$\begin{aligned} f(Z, X|\theta) &= \prod_{i=1}^{N} f(z_i, x_i|\theta) \\ &= \prod_{i=1}^{N} \sum_{g=1}^{G} \delta_{gi} f_g(z_i, x_i|\theta_g) \\ &= \prod_{i=1}^{N} \sum_{g=1}^{G} \delta_{gi} f_g(x_i, z_i = C_g|\theta_g) \\ &= \prod_{i=1}^{N} \sum_{g=1}^{G} \delta_{gi} f_g(x_i, \delta_{gi}|\theta_g) \end{aligned}$$

Given a maximized parameter set θ^*, the membership function is the posterior probabilities $h_g(x_i) = P(\delta_{gi} = 1|x_i, \theta^*)$. The membership function is the probability of x_i belonging to the g-th cluster, given x_i and the model. This can be derived using Bayesian theorem

$$\begin{aligned} h_g(x_i) &= P(\delta_{gi} = 1|x_i, \theta^*) \\ &= \frac{\pi_g f_g(x_i|\delta_{gi}=1, \psi_g)}{\sum_{k=1}^{G} \pi_k f_k(x_i|\delta_{ki}=1, \psi_k)} \end{aligned}$$

The common criteria for assigning a cluster g to x_i is by finding the component $g_i^* = \underset{g \epsilon G}{\operatorname{argmax}}(h_g(x_i))$.

3.3.3 Expectation Maximization

An approach to finding a maximized likelihood function set θ^* for a G-component mixture model is the Expectation Maximization (EM) algorithm. This procedure follow an iterative step:

1. *Initialize* : Provide an initial guess for $theta_0$. Set $t = 0$ and $Q(\theta_0|\theta_{-1}) = -\infty$.

2. $E - step$: Compute the expected value of the log-likelihood function of θ with respect to the current conditional distribution of z given x and the current estimates of the parameter θ_t.

$$\begin{aligned} Q(\theta|\theta_t) &= E_z[\log f(Z, X|\theta)|X, \theta_t] \\ &= \prod_{i=1}^{N} \sum_{g=1}^{G} E[\delta_{gi}|x_i, \theta_t] \log(f_g(x_i|\delta_{gi} = 1, \psi_g)|\pi_g) \\ &= \prod_{i=1}^{N} \sum_{g=1}^{G} h_{gt}(x_i) \log(f(x_i|\delta_{gi} = 1, \psi_g)|\pi_g) \end{aligned}$$

The posterior probability membership function of step t is also computed as

$$h_{gt}(x_i) = \frac{\pi_{gt} f_g(x_i|\delta_{gi} = 1, \psi_{gt})}{\sum_{k=1}^{G} \pi_{kt} f_k(x_i|\delta_{ki} = 1, \psi_{kt})}$$

3. $M - step$: Determine the value of θ_{t+1} which maximizes $Q(\theta|\theta_t)$. $\theta_{t+1} = \underset{g \epsilon G}{\text{argmax}} Q(\theta|\theta_t)$. Deriving this value for the parameter $\pi_{g,t+1}$ can be done by maximizing $Q(\theta|\theta_t)$ with respect to $\pi_{g,t+1}$ by $\frac{\partial Q(\theta|\theta_t)}{\pi_{gt}} = 0$, while subject to the constraint of $\sum_{g=1}^{G} \pi_g = 1$. This finally yields to $\pi_{g,t+1} = \frac{1}{N} \sum_{i=1}^{N} h_{gt}(x_i)$.

4. If $Q(\theta_{t+1}|\theta_t) - Q(\theta_t|\theta_{t-1}) \leq \xi$ (ξ is the specified termination threshold), then proceed to the last step. Otherwise, go back to Step 2.

5. The final parameter set $\theta^* = \theta_{t+1}$ is the derived maximized likelihood estimate of θ.

3.4 Generalized Hyperbolic Distribution for Data Simulation

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 less heavier 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

$$P(x; \lambda, \alpha, \beta, \delta, \mu) = a(\lambda, \alpha, \beta, \delta, \mu)(\delta^2 + (x-\mu)^2)^{1/2\lambda - 1/4} \cdot B(\lambda - 0.5, \alpha\sqrt{\delta^2 + x^2 - 2x\mu + \mu^2}) e^{\beta(x-\mu)} \tag{3.4}$$

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 was assessed in a study of Browne and McNicholas (2015) using real and simulated datasets. The real dataset (Old Faithful data (GeyserTimes, 2017)) were 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. One hundred 2-component datasets were simulated each from a mixture of Gaussian distributions and from a mixture of skew-t distributions. When a GH mixture model were fitted using EM method, 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, which may then be used to simulate data for other analyses.

Chapter 4

Methodology

4.1 Data

4.1.1 Real Dataset

4.1.2 Simulated Dataset

4.2 EM Model fitting

4.3 Performance Evaluation of Quantification tools

Chapter 5

Diagrams and Other Documentation Tools

This appendix may consist of proposed architectural design, algorithms, scientific formula for MSCS and Data Flow Diagrams, Fishbone for MSIT.

Chapter 6

Theoretical and/or Conceptual Framework

Discusses the basic framework/foundation the thesis is based on. This section is normally referred to when discussing Scope and Limitations, and Research Methodology

Chapter 7

Resource Persons

Dr. Firstname1 Lastname1
Adviser
CCS
De La Salle University-Manila
emailaddr@dlsu.edu.ph

Mr. Firstname2 Lastname2
Role2
Affiliation2
emailaddr2@domain.com

Ms. Firstname3 Lastname3
Role3
Affiliation3
emailaddr3@domain.net

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

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*. Wood-

head 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

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

Henricha, T. J., Gallienb, S., Lia, J. Z., Florencia Pereyraa, C., & Kuritzkesa, D. R. (2012). Low-Level Detection and Quantitation of Cellular HIV-1 DNA and 2-LTR Circles Using Droplet Digital PCR. *Journal of Virological Methods*, *186*(1-2), 68–72. doi: 10.1038/jid.2014.371

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. (2017a). 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

Hussain, M., & Bowers, J. (2017b). *A droplet digital pcr method for cho host residual dna quantification in biologic drugs - scientific figure on researchgate.* Retrieved from https://www.researchgate.net/figure/Droplet-fluorescence-amplitude-with-CHO-hrDNA-ddPCR-method-The-mAb-DS4P-52g-was_fig1_316088304 ([Online; accessed April 7, 2020])

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. (2014a). Low copy target detection by Droplet Digital PCR through application of a novel open access bioinformatic pipeline, 'definetherain'. *Journal of Virological 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

Jones, M., Williams, J., Gärtner, K., Phillips, R., Hurst, J., & Frater, J. (2014b). *Low copy target detection by Droplet Digital PCR through application of a novel open access bioinformatic pipeline, 'definetherain'.* Retrieved from https://www.researchgate.net/figure/Screenshots-illustrating-the-application-of-definetherain-Screenshots-illustrating-the_fig2_260441031 ([Online; accessed April 7, 2020])

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

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

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