Model-Based Clustering of Digital PCR Droplets using Expectation Maximization

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

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

by

GUIAO, Joyce Emlyn B.

Frumencio F. Co
Adviser

April 14, 2020

Abstract

From 150 to 200 words of short, direct and complete sentences, the abstract should be informative enough to serve as a substitute for reading the thesis document itself. It states the rationale and the objectives of the research.

In the final thesis document (i.e., the document you'll submit for your final thesis defense), the abstract should also contain a description of your research results, findings, and contribution(s).

Keywords can be found at `http://www.acm.org/about/class/class/2012?pageIndex=0`. Click the link "HTML" in the paragraph that starts with ”The **full CCS classification tree**...”.

Keywords: Keyword 1, keyword 2, keyword 3, keyword 4, etc.

Contents

List of Figures

List of Tables

Chapter 1

Research Description

1.1 Introduction

1.2 Background of the Study

Quantification of Nucleic acids (NA) is a developing research field in molecular biology for the detection and quantification expression levels of genes (Huggett, O'Grady, & Bustin, 2015). These NA 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, NA 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 (?, ?).

To be able to determine the concentration of target NAs, NA detection is naturally a pre-requisite. There are, however, NAs 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 NA sequences using Polymerase Chain Reaction (PCR), a widely-used method for NA amplification since its invention in the 1980s (?, ?). PCR can multiply specific NA sequences in DNA or RNA from low concentrations to millions of copies. This method exposes the NA sequences mixed with chemical components in a series of 20 to 40 temperature cycles. In each cycle, PCR doubles the NA molecule; theoretically producing 2^n

molecules after n cycles (Quan et al., 2018).

After PCR amplification, absolute NA quantification is achieved using digital Polymerase Chain Reaction (dPCR) technique. This equally divides the NA 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" (?, ?).

The dPCR workflow, as illustrated in Figure 1.1, 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 1.2.

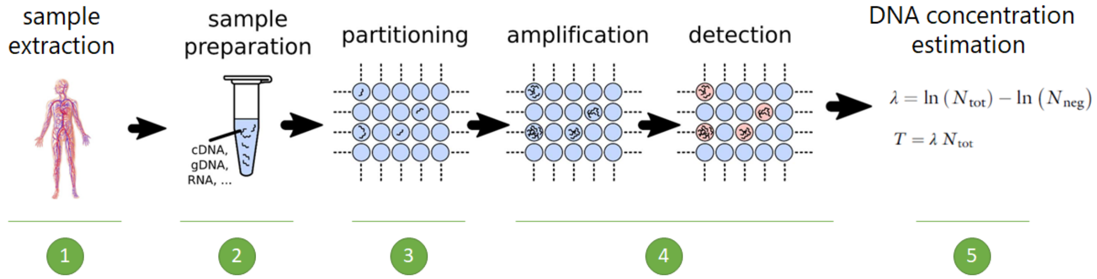
sample extraction
sample preparation
partitioning
amplification
detection
DNA concentration estimation
cDNA, gDNA, RNA, ...
$\lambda = \ln(N_{\text{tot}}) - \ln(N_{\text{neg}})$
$T = \lambda N_{\text{tot}}$
1
2
3
4
5

Figure 1.1: The dPCR workflow

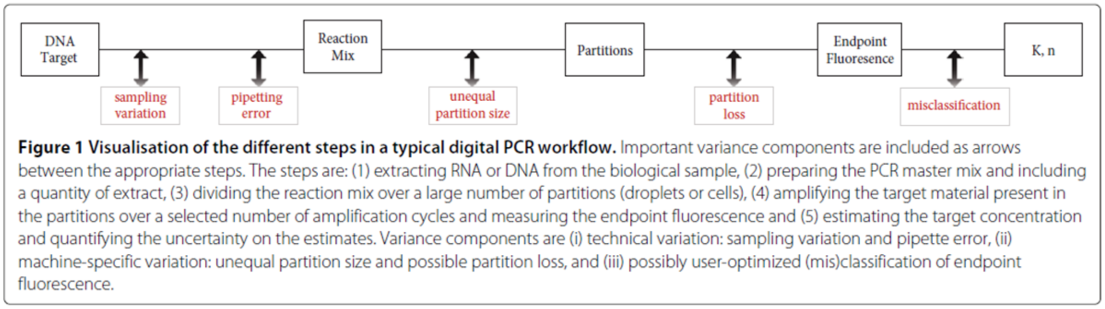

Figure 1 Visualisation of the different steps in a typical digital PCR workflow. Important variance components are included as arrows between the appropriate steps. The steps are: (1) extracting RNA or DNA from the biological sample, (2) preparing the PCR master mix and including a quantity of extract, (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 1.2: Potential variation components between steps of the dPCR workflow

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 (n.d.) 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.

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. The next variation components are the possibility of the distributed partitions to be of unequal volumes and the loss of partitions due to physical interventions. Finally, upon PCR amplification, each droplet partition emits an endpoint fluorescence that would be used to classify the partition as positive or negative. However, some partitions are difficult to classify due to inhibition, delayed reactions, primer depletion, and other biological factors.

Each variance component accumulates to the bias and variance of the final estimated target molecule 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 count. Due to the variance contributed by misclassification, Tzonev (n.d.) recommends reporting the rates of false positives (FPR) and false negatives (FPN) per partition. More importance is given to either of the two depending on the kind of test being performed. If the total negative partition is expected to be large, then there is more chance that a true negative partition may turn into a false positive reading; which in this case, FPR may be of more interest than FNR.

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 1.3 and 1.4 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).

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

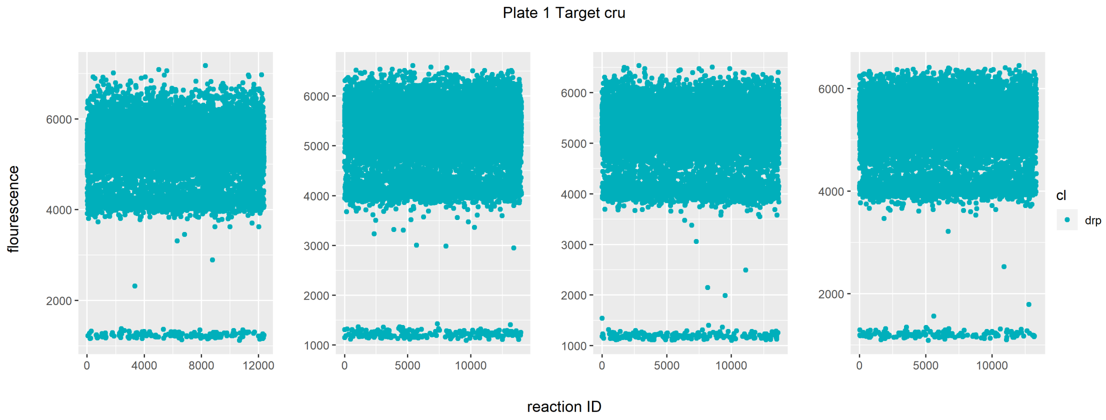

Figure 1.3: Fluorescence readings of 4 repititions of DNA target cru

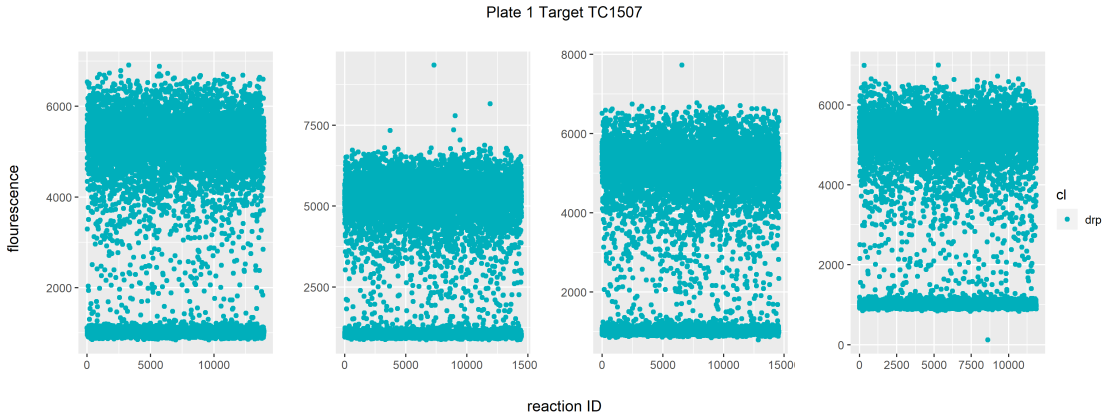

Figure 1.4: Fluorescence readings of 4 repititions of DNA target TC1507

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, n.d.). 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.

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. Fit two-component finite mixture densities on the datasets with varying amounts of rain and increasing target DNA concentration;
2. Classify partitions using the fitted models for each experiment and estimate DNA concentration using the standard formula;
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 (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. Similar to Umbrella, this study also aims to use model-based clustering for the droplet classification but with relaxing the assumptions using the Expectation-Maximization algorithm. The significance of the study will be useful in quantifying 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 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 Current dPCR Droplet Classification Methods

2.1.1 Droplet dPCR (ddPCR) System

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 (ddPCR) System package. It allows for the setting up of sample and experiments, running and controlling the instrument, and finally, the analysis of the NA 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, Zhang, Tan, & Jing, 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), upto 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.

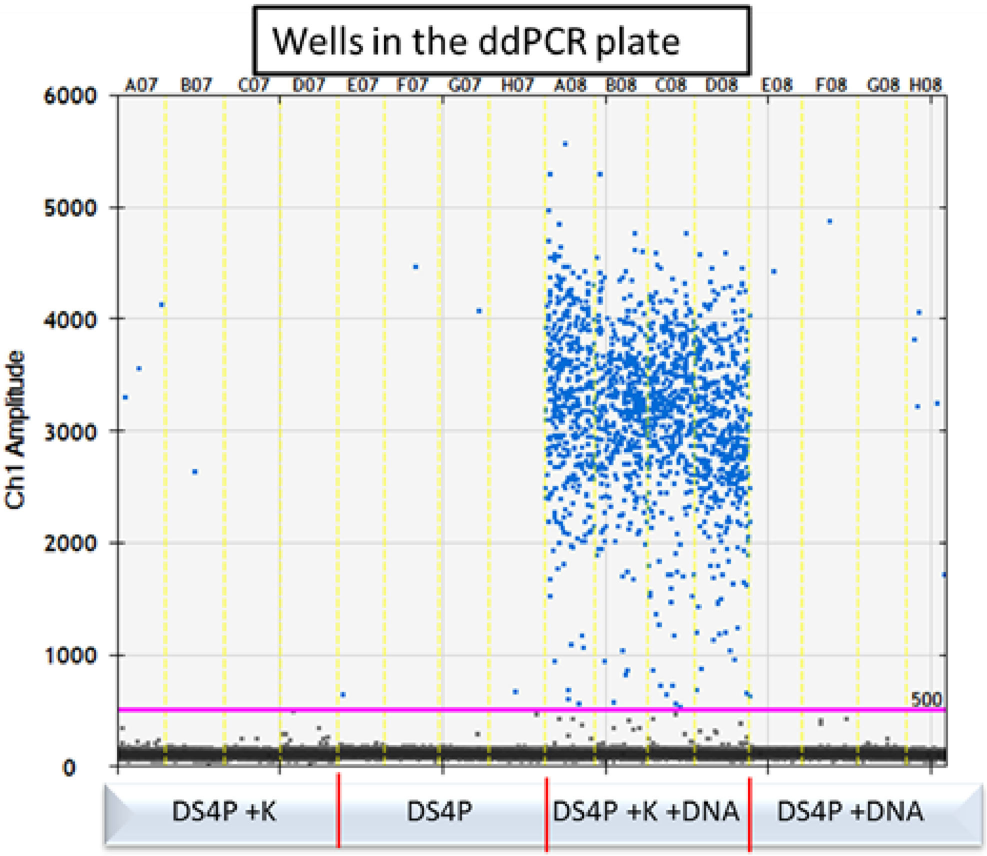

Figure 2.1: QuantaSoft threshold in pink line from (Hussain & Bowers, 2017b)

The evaluation of the QuantaSoft software show exemplary 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) suggests the need for a highly sensitive and specific test in the molecular level. Observation of NA molecules can be done through their droplet fluorescence in a dPCR assay.

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 two 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. This study concluded that the QuantaSoft software requires a well-optimized assay with good discrimination of positive and negative droplets for its threshold to be reliable.

2.1.2 Manual Global Threshold (MTg)

As opposed to the automatic threshold, Dreo et al. (2014) proposes setting a manual global threshold (MTg) determined by no template control (NTC) samples. This takes into consideration that individual assays behave differently, and could require expert intervention. For example, 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 example 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 clear definition or guidelines in setting the MTg; this consequently will cause reproducibility issues for succeeding experiments and external researchers.

2.1.3 Kernel Density Estimation (KDE)

The positive and negative droplet classification can be framed as a clustering problem. One solution is by using the popular theory of kernel density estimation (KDE); whereby each data point influence is modeled using a kernel function — frequently Gaussian distributed, and the sum of all the kernel functions would be the overall density of the data. Clusters can then be derived from the estimated densities depending on its application (Hinneburg & Keim, 2003).

KDE has been particularly helpful in spatial analysis. Clusters were detected in areas with diabetes in Berlin (Kauhl, Schweikart, Krafft, Keste, & Moskwyn, 2016), HCV hepatitis C virus in Massachusetts (Stopka et al., 2017), and crime hotspots in Brazil (Junior, Da Silva, De Queiroz Neto, De MacÊdo, & Porcino, 2019). For other data types, KDE clustering has been improved in an algorithm developed by Rodriguez and Laio (2014), called clustering by fast search and find of density peaks. This technique has piqued the interests of many researchers; thereby further improving the original limitations via DNA genetic algorithm (Zang, Ren, Zhang, & Liu, 2017), heat diffusion (Mehmood, Zhang, Bie, Dawood, & Ahmad, 2016), and entropy of data field (Wang, Wang, Li, Li, & Ding, 2016).

Although not named in the dPCR optimization paper of Lievens et al. (2016), cloudy is the name of the function in their supplementary source code file that calculates the threshold from the clusters found by using Gaussian KDE. The following steps explain this in more detail:

1. Estimate the density function of the fluorescence using a Guassian kernel density with a minimum bandwith of 50
2. Identify density peaks 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 and normality assumptions to detect peaks, which are then considered as populations.

From then population density parameters are estimated based on specified formulas. A range or a single threshold is then calculated for classifying droplets as positive, negative, or optionally, rain.

Special cases such as overlapping population boundaries and threshold placement problems are also checked in their source code. It is worth noting that in step 6, the formula for re-calculating a for density parameter estimation is based on the analysis of their in-house data, and should be used with caution when implementing for other unobserved NA targets. The rain classification rule in step 8(a) also poses a problem for fluorescence densities that are heavily skewed. Figure 2.2 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).

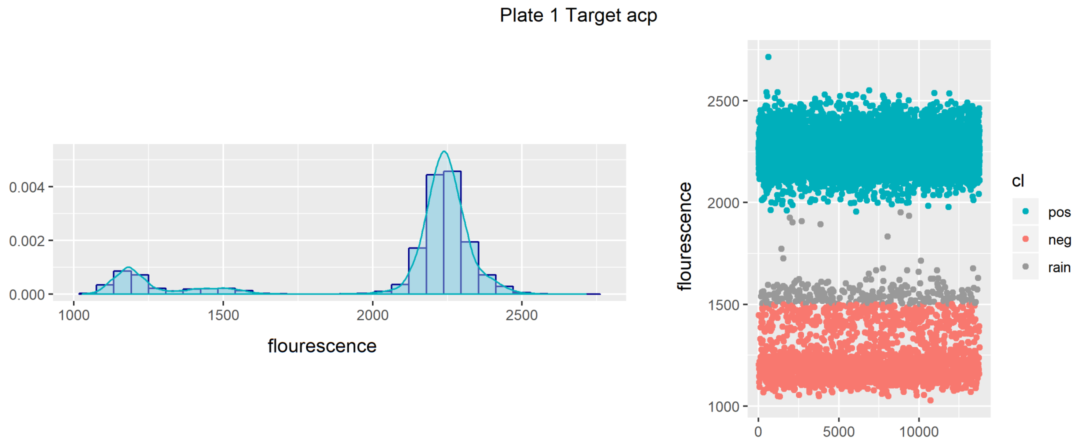

Figure 2.2: One replicate of the DNA target acp Plate 1 from Lievens et al. (2016) dataset. Left panel shows the fluorescence densities. Right panel is the result of droplet categorization using cloudy

It is noted that the context of the threshold setting here is based on in-house data for studying dPCR experiment optimization, and not necessarily as a classifier for external data. The constant threshold criterium enables the control for other PCR experimental factors, such as sonication, PCR enhancers, annealing conditions, and number of cycles, to achieve optimization or diminishing rain droplets for dPCR experiments.

2.1.4 K-Nearest-Neighbors (KNN)

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.

An open-source tool developed by 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 disadvantage of this is that the control has to be representative of the NA target; otherwise, concentration estimates would be biased.

2.1.5 Model-Based Clustering w/ Non-parametric Density Estimation

As opposed to the distance-based clustering in Section 2.1.4, a probabilistic approach is achieved with model-based clustering. According to Mcnicholas (2016),

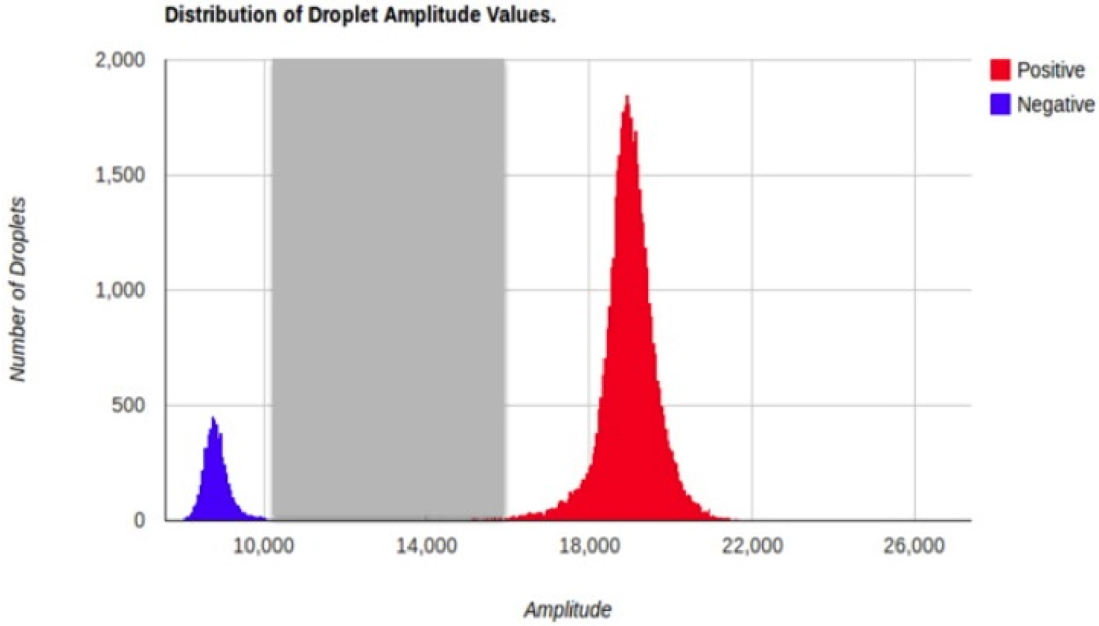

Figure 2.3: Determined thresholds calculated by definetherain, reprinted from (Jones et al., 2014b)

a finite mixture model is a sum of weighted density components. This mixture model has to be appropriate such that its parameters are flexible for fitting the characteristics of the data. In this approach, each unimodal density component are defined as a cluster; and each observation has a calculated probability of it belonging to a cluster.

In an electricity usage profiling study, Li et al. (2018) noticed several elongated ellipses in the scatterplot of electricity usage data, urging them to use 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, 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/normal component distributions are common in model-based clustering; however, in the succeeding texts, dPCR droplets is claimed to violate this assumption.

Jacobs et al. (2017) developed Umbrella, a model-based clustering for dPCR droplets using nonparametric density estimation (https://github.com/statOmics/umbrella). Umbrella requires a representative NTC sample(s), the procedure then follows a series of estimation and assumptions in deriving the final estimated concentration. An oversimplification of the Umbrella procedure, that skips the detailed steps for handling specific cases, 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, $p_{\hat{0,A}}$, can be defined from the estimated $\hat{p}_{0,A}$ from the previous step as

$$\hat{p}_{i,0,A} = \hat{p}_{0,A}\left(\begin{array}{c}\hat{f}_{0,A}(x_i)\\ \hat{f}_A(x_i)\end{array}\right)$$

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, as the location and spread estimators 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 NA 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 cautioned that Umbrella may not be precise in detection experiments for low copy samples, as classifying individual samples is not the strength of this method.

2.2 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, skew-t, multivariate power exponential, variance-gamma, generalized hyperbolic, etc. Unlike Gaussian, these component models are flexible in data of 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, there is room for research in identifying the fluorescence intensity distribution that will fit the characteristics of the positive/negative groups. Since heavy tails are observed in fluorescence densities in Figure 2.2, its distribution appears to be nonnormal, as also stated by Jacobs et al. (2017).

Chapter 3

Theoretical Framework

3.1 EM Clustering

3.1.1 Model-based clustering

3.1.2 G-component Finite Mixture Density

3.1.3 Expectation Maximization

3.2 Performance Evaluation

Chapter 4

Methodology

4.1 Data

4.1.1 Rain Experiment Dataset

Plate 2 - Primer and Probe Concentration Gradient

Plate 4 - PCR Enhancers Experiment

Plate 5 - Cycle Gradient

Plate 6 - Sonication Gradient

Plate 7 - Annealing Temperature Gradient

4.1.2 DNA Quantification Dataset

Plate 3 - Rain Dilution Series

Albumin

4.2 Model Fitting and Classification

4.3 Performance Evaluation

Appendix A

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.

Appendix B

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

Appendix C

Resource Persons

Dr. Firstname1 Lastname1
Adviser
College of Computer Studies
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

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

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

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

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

Dang, U. J., Gallaugher, M. P. B., Browne, R. P., & McNicholas, P. D. (2019). Model-based clustering and classification using mixtures of multivariate skewed power exponential distributions. (1965), 1–23. Retrieved from http://arxiv.org/abs/1907.01938

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

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

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

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

Hinneburg, A., & Keim, D. A. (2003). A General Approach to Clustering in Large Databases with Noise. *Knowledge and Information Systems*, *5*(4), 387–415. doi: 10.1007/s10115-003-0086-9

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 ofl 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). 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.2019.115213` doi: 10.1016/j.watres.2019.115213

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])

Junior, F. C., Da Silva, T. L., De Queiroz Neto, J. F., De MacÊdo, J. A. F., & Porcino, W. C. (2019). A novel approach to approximate crime hotspots to the road network. *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Prediction of Human Mobility, PredictGIS 2019*, 53–61. doi: 10.1145/3356995.3364538

Kauhl, B., Schweikart, J., Krafft, T., Keste, A., & Moskwyn, M. (2016). Do the risk factors for type 2 diabetes mellitus vary by location? A spatial analysis of health insurance claims in Northeastern Germany using kernel density estimation and geographically weighted regression. *International Journal of Health Geographics*, *15*(1), 1–12. doi: 10.1186/s12942-016-0068-2

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). Combining ddPCR and environmental DNA to improve detection capabilities of a critically endangered freshwater invertebrate. *Scientific Reports*, *9*(1), 1–9. Retrieved from `http://dx.doi.org/10.1038/s41598-019-50571-9` doi: 10.1038/s41598-019-50571-9

Mcnicholas, P. D. (2016). Model-Based Clustering. , *373*(November), 331–373. doi: 10.1007/s0035

Mehmood, R., Zhang, G., Bie, R., Dawood, H., & Ahmad, H. (2016). Clustering by fast search and find of density peaks via heat diffusion. *Neurocomputing*, *208*, 210–217. Retrieved from `http://dx.doi.org/10.1016/j.neucom.2016.01.102` doi: 10.1016/j.neucom.2016.01.102

Mittal, K., Aggarwal, G., & Mahajan, P. (2019). 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-x` doi: 10.1007/s41870-018-0233-x

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

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

Quan, P. L., Sauzade, M., & Brouzes, E. (2018). DPCR: A technology review. *Sensors (Switzerland)*, *18*(4). doi: 10.3390/s18041271

Rodriguez, A., & Laio, A. (2014). Clustering by fast search and find of density peaks. *Science*, *344*(6191), 1492–1496. doi: 10.1126/science.1242072

Shi, J., & Yang, L. (2020). A climate classification of China through k-nearest-neighbor and sparse subspace representation. *Journal of Climate*, *33*(1), 243–262. doi: 10.1175/JCLI-D-18-0718.1

Stopka, T. J., Goulart, M. A., Meyers, D. J., Hutcheson, M., Barton, K., Onofrey, S., … Chui, K. K. (2017). Identifying and characterizing hepatitis C virus hotspots in Massachusetts: A spatial epidemiological approach. *BMC Infectious Diseases*, *17*(1), 1–11. doi: 10.1186/s12879-017-2400-2

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

Tzonev, S. (n.d.). Chapter 3 Fundamentals of Counting Statistics in Digital PCR : I Just. , *1768*, 25–43.

Wang, S., Wang, D., Li, C., Li, Y., & Ding, G. (2016). Clustering by fast search and find of density peaks with data field. *Chinese Journal of Electronics*, *25*(3), 397–402. doi: 10.1049/cje.2016.05.001

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

Zang, W., Ren, L., Zhang, W., & Liu, X. (2017). Automatic Density Peaks Clustering Using DNA Genetic Algorithm Optimized Data Field and Gaussian Process. *International Journal of Pattern Recognition and Artificial Intelligence*, *31*(8), 1–28. doi: 10.1142/S0218001417500239