

Kansas State University Libraries

## New Prairie Press

---

Conference on Applied Statistics in Agriculture

2017 - 29th Annual Conference Proceedings

---

# Statistical Methods for Assessing Individual Oocyte Viability Through Gene Expression Profiles

Michael O. Bishop

Utah State University, [michael.o.bishop@aggiemail.usu.edu](mailto:michael.o.bishop@aggiemail.usu.edu)


John R. Stevens

Utah State University

S. Clay Isom

Utah State University

Follow this and additional works at: <https://newprairiepress.org/agstatconference>

 Part of the [Agriculture Commons](#), and the [Applied Statistics Commons](#)



This work is licensed under a [Creative Commons Attribution 3.0 License](#).

---

### Recommended Citation

Bishop, Michael O.; Stevens, John R.; and Isom, S. Clay (2017). "Statistical Methods for Assessing Individual Oocyte Viability Through Gene Expression Profiles," *Conference on Applied Statistics in Agriculture*. <https://doi.org/10.4148/2475-7772.1516>

This Event is brought to you for free and open access by the Conferences at New Prairie Press. It has been accepted for inclusion in Conference on Applied Statistics in Agriculture by an authorized administrator of New Prairie Press. For more information, please contact [cads@k-state.edu](mailto:cads@k-state.edu).

## STATISTICAL METHODS FOR ASSESSING INDIVIDUAL OOCYTE VIABILITY THROUGH GENE EXPRESSION PROFILES

Michael O. Bishop<sup>\*1,2</sup>, John R. Stevens<sup>2</sup>, S. Clay Isom<sup>3</sup>

<sup>1</sup> University of Iowa, Department of Biostatistics, Iowa City, IA, michael-bishop@uiowa.edu

<sup>2</sup> Utah State University, Department of Mathematics and Statistics, Logan, UT

<sup>3</sup> Utah State University, Department of Animal, Dairy and Veterinary Sciences, Logan, UT

\* Denotes corresponding author

### ABSTRACT

Oocytes are the precursor cells to the female gamete, or egg. While reproduction may vary from species to species, within humans and most domesticated animals, the oocyte maturation process is fairly similar. As an oocyte matures, there are various processes that take place, all of which have an effect on the viability of the individual oocyte. Barring outside damage that may come to the oocyte, one of the primary reasons for non-viability is that of abnormal gene expression. Within this project, we focus on two oocyte maturation techniques: *in vivo* (IVV) derived oocytes (our gold-standard) and *in vitro* matured (IVM) oocytes. A great disparity exists between the viability rates of the two origination techniques, and this disparity has led to low yields and inefficiency in the fields of cloning, fertility treatments, and personalized medicine.

Within our project we use existing swine oocyte gene expression profile data as a proxy measure of viability, based on the similarity to IVV oocytes. Four statistical techniques for assessing the individual oocyte viability are proposed and compared, including: a weighted root mean squared deviation (wRMSD) approach, a distance kernel p-value approach, a distance tolerance interval approach, and a classification tree method. The relative performance of these four measures is discussed.

Keywords: Oocyte, Gene Expression Profile, Tolerance Interval, Classification Tree

## 1. INTRODUCTION

Oocytes are the precursor cells to what we often think of as the female egg cell. They operate and grow in essentially the same way within swine as they do in humans (Thomas, 2016). Currently, there are numerous methods of deriving (growing) oocytes, each with its benefits. *In vivo* (IVV) derived oocytes are held as the gold standard for viability, and other known origination methods are sub-par by comparison. However, alternatively derived oocytes do have many traits that are desirable. The *in vitro* maturation (IVM) method is one such method of oocyte derivation, as it allows the researchers to have access to the oocyte from early on. There are however, many problems associated with this method. From past studies, it is shown that the viability rate associated with IVV derived oocytes is upwards of 95%, however for IVM oocytes (and other similar methods) the viability rate is roughly only 25% (Dr. S. Clay Isom, personal communication). Similar low viability concerns arise with the SCNT (somatic cell nuclear transfer, or traditional cloning) origination method. With such a low viability rate it is only natural to try to alleviate the frustration and waste that comes from focusing time, resources and energy on oocytes that are simply not viable. But how can one know if an oocyte is viable prior to investing in it? In general, one cannot. However, methods can be developed that test individual oocytes for viability, establishing a viability-optimized general gene expression profile. As laboratory originated oocytes can be treated with a wash that promotes growth and expression patterns in chosen directions (Dr. S. Clay Isom, personal communication), it is apparent that establishing methods to test individual oocyte viability and thereby derive a general viable profile for oocytes would, in theory, increase the viability of non-IVV oocytes. But rather than stopping there, one could test not only if specific genes were differentially expressed between the different groups, but also if one could use gene expression profiles to turn viability into a classification problem.

Kwon *et al.* (2015) took embryos that were obtained from three different origination methods (IVV, IVM, and SCNT) and examined the gene expression profiles for 15 different genes of interest. In their project, a Weighted Root Mean Squared Deviation (wRMSD) was calculated based on expression level deviation from the mean expression level of the origination group. The average wRMSD for each origination group was calculated and compared. They found that the difference between the IVV group and the SCNT group was greater than the difference between the IVV group and the IVM group. These conclusions were relevant for comparing methods of origination, however we desire to compare individual oocyte viability.

Within our project, the goal has been to establish origin-independent methods of evaluating individual oocyte viability, based simply on the observed gene expression profiles. As the process of collecting the observed gene expression profiles of oocytes is destructive, making the oocytes non-viable (and thereby making it impossible to tell if the oocyte would have been viable or not, simply by observing), there are a number of assumptions that one must make. First, assume that IVV-derived oocytes are the “gold standard” for oocyte viability. This assumption does not seem to be erroneous, as historically over 95% of all IVV-derived oocytes have been observed to be viable (Dr. S. Clay Isom, personal communication). The second assumption that is made is that the closer an oocyte’s gene expression profile gets to the mean of the IVV gene expression profiles, the more likely the oocyte is to be viable. Third, assume that the IVV oocyte

gene expression profiles are a representative sample of all IVV oocyte gene expression profiles, and likewise for the IVM oocytes within the data.

Dr. S. Clay Isom provided the project dataset, which contained the gene expression profiles for 29 IVV derived oocytes as well as 29 IVM derived oocytes. Each of the gene expression profiles was represented in the form of a log<sub>2</sub> fold-change on 67 genes of interest compared to a common “housekeeping” gene. These specific 67 genes were selected, as they are affiliated with early cell growth, regulation and viability (Dr. S. Clay Isom, personal communication). Overall there were few values that were missing (about 4.4%), but computationally if the case arose that a value was missing, that value was omitted from gene level calculations in the following methods.

In this paper (from Bishop 2017), an adaption (on the single oocyte level) of the Kwon *et al.* (2015) wRMSD method is summarized in Section 2, along with three other novel methods – a Distance Kernel P-value method in Section 3, a Tolerance Interval method in Section 4, and a Decision Tree method in Section 5. Each of these methods is then compared for accuracy via simulation in Sections 6 and 7, and the methods’ performance is discussed in Section 8.

## 2. WEIGHTED ROOT MEAN SQUARED DEVIATION KERNEL DENSITY P-VALUE

We use an application of the weighted root mean squared deviation (wRMSD) approach proposed by Kwon *et al.* (2015). Here, we compute the wRMSD from the center of the oocyte viability class. In the more practical case that viability status is unknown, we would then compute the wRMSD from the center of the oocyte maturation class (center of IVV if using an IVV oocyte, or IVM if using an IVM oocyte). Again, we operate under the assumption that IVV is considered viable. Kwon *et al.* (2015) said “We considered each gene expression to be an independent event; therefore, we combined all of the expression measurements of each (gene) sample in the calculation of the wRMSD. To minimize the bias from a measurement error of a gene expression profile with a low coefficient of variation (CV), the deviation of each gene expression level from the mean was weighted with the CV of the gene in the group.” Within the wRMSD calculation there are three main parts: the reference expression level (mean expression level for the gene within group), the expression level of the specified gene within the oocyte of interest, and the weighting coefficient. The weighting coefficient was further made up of “the proportion of the CV for the expression level of the *i*<sup>th</sup> gene to the sum of CV for those of all genes in the group” (Kwon *et al.*, 2015):

$$wRMSD = \sqrt{\sum_i w_i \cdot (Em_i - E_i)^2} \quad w_i = \frac{CV_i}{\sum_j CV_j} \quad (1)$$

In Equation 1, wRMSD is defined for a given oocyte. Here,  $E_i$  refers to the expression level of the *i*<sup>th</sup> gene in the oocyte,  $Em_i$  refers to the reference expression level for the oocyte (i.e. mean of the IVV if the oocyte is IVV or IVM if the oocyte is in the IVM group) of the *i*<sup>th</sup> gene, and  $w_i$  refers to the weighting coefficient.

We set up the null hypothesis that an observed oocyte (with an accompanying wRMSD) came from the viable (or IVV) class, with the respective alternative hypothesis that it did not. P-

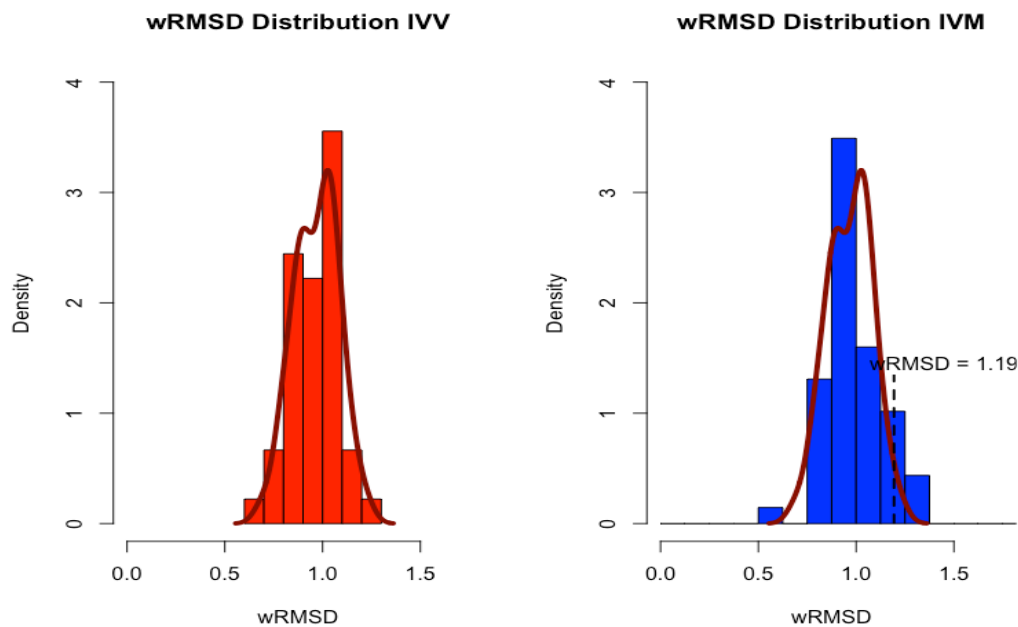


Figure 1. The wRMSD distribution of IVV and IVM oocytes with calculated kernel density for the IVV wRMSD distribution overlaid on both histograms.

values for likelihood of belonging to the “viable” class were then computed for the individual oocytes by comparing the oocyte’s observed wRMSD to the kernel density of the viable (or IVV) wRMSD distribution, and computing an upper tail area (see Figure 1). These p-values were then compared to an alpha 0.05 level for determining if there was enough evidence to reject our null hypothesis (that the observed oocyte was viable). This Kernel Density p-value portion was not utilized in the Kwon *et al.* paper, however it is a reasonable and necessary application of their published method that allows for individual oocyte classification and comparison between results from this and other methods.

In Figure 1, an example IVM oocyte has been selected and its wRMSD has been found to be 1.19. The wRMSD of the oocyte is then compared to the kernel density of the wRMSD for the IVV group; we can then set up a hypothesis test with a null hypothesis that the observed oocyte comes from the IVV group. By computing an upper tail area from the kernel density, we can observe a p-value for our example oocyte equal to 0.052. This p-value is larger than our 0.05 cutoff, so we would classify the observed oocyte as viable (by our assumption that the more closely related to the IVV group that an oocyte is, the more likely it is to be viable).

### 3. DISTANCE KERNEL DENSITY P-VALUE METHOD

We considered also a distance measurement method as a modification to the wRMSD approach presented by Kwon *et al.* (2015). It utilizes only a subset of the available genes; those that a limma eBayes approach (Ritchie, 2015) has determined are differentially expressed (between IVV and IVM groups) at an alpha 0.05 level. Briefly, the limma eBayes approach

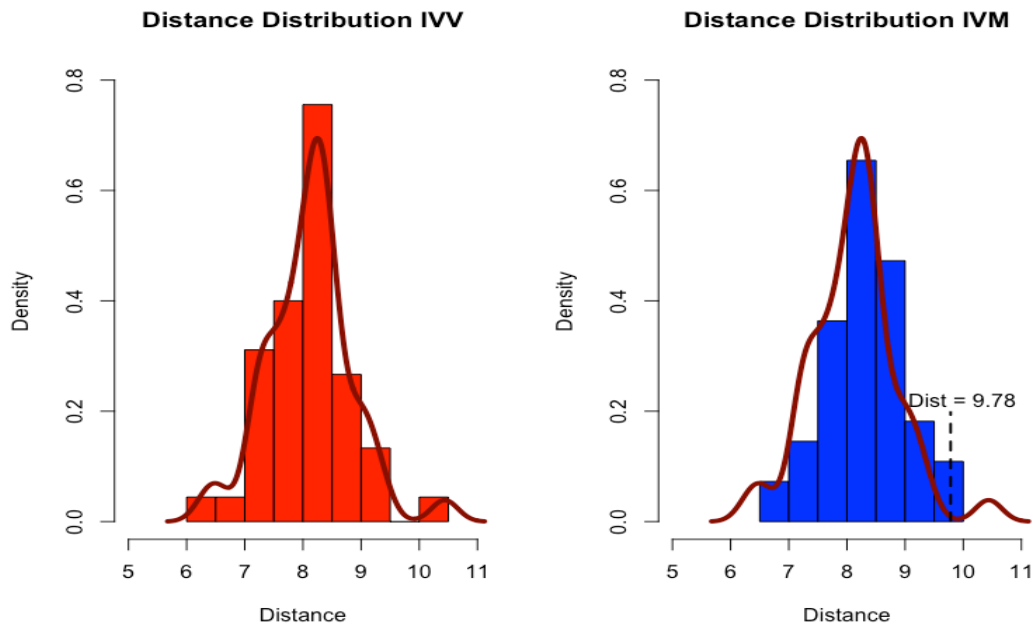


Figure 2. The distance distribution of IVV and IVM oocytes with calculated kernel density for the IVV distance distribution overlaid on both histograms.

performs a modified t-test on the expression level of each gene, testing for differential expression between two conditions (IVV and IVM here). The gene expression profile of all IVV (or viable) oocytes is computed, and the mean expression is taken for each gene in order to form an IVV group mean gene expression profile. Using only the subset of differentially expressed genes, the distance measure for each oocyte from the mean of the IVV group gene expression profile is then computed. Here, distance is measured as an adaption to the Kwon *et al.* wRMSD approach, with weighting coefficients based on the variance of the genes:

$$Distance = \sqrt{\sum_i w_i \cdot (Em_i - E_i)^2} \quad w_i = \frac{\sigma_i^2}{\sum_j \sigma_j^2}. \quad (2)$$

In Equation 2, we only use information from differentially expressed genes.  $Em_i$  represents the mean of the IVV (or viable) group for gene  $i$ .  $E_i$  represents the expression level of the  $i^{\text{th}}$  differentially expressed gene for the specified oocyte.  $W_i$  is the weight for the specified gene  $i$ ,

The distance kernel density is then calculated for the IVV (or viable) oocytes, based upon the calculated IVV distances. We then can use this kernel density distribution (see Figure 2) to set up a hypothesis test for each of the individual oocytes, using the following null and alternative hypotheses:  $H_0$  – The oocyte distance comes from the IVV (viable) distribution of distances, i.e. the observed distance for the oocyte is within an expected range if it was of the

IVV (viable) class, vs.  $H_a$  – The oocyte distance does not come from the IVV (viable) distribution of distances, i.e. the observed distance for the oocyte is outside the expected range if it is of the IVV (viable) class. In order to make a decision for the hypothesis test, we compute the upper tail probability above the distance observed in the oocyte of interest. This upper tail area is our observed p-value.

In Figure 2, the same example IVM oocyte has been selected as in Figure 1, and its calculated distance has been found to be 9.78. The distance of the oocyte is then compared to the kernel density of the distances for the IVV group; we can then set up a hypothesis test with a null hypothesis that the observed oocyte comes from the IVV group. By computing an upper tail area from the kernel density, we can observe a p-value for our example oocyte equal to 0.019. This p-value is smaller than our 0.05 cutoff, so we would classify the observed oocyte as non-viable (by our assumption that the more closely related to the IVV group that an oocyte is, the more likely it is to be viable).

#### 4. TOLERANCE INTERVAL METHOD

A tolerance interval is a numerical interval that is calculated to provide limits wherein at least a specified proportion of a sampled population falls with an indicated level of confidence. Oftentimes, tolerance intervals are constructed and applied in areas such as quality control or manufacturing to establish that certain product standards are being met by the overall bulk of the products. “More specifically, a  $100 \times p\% / 100 \times (1 - \alpha)$  tolerance interval provides limits within which at least a certain proportion ( $p$ ) of the population falls with a given level of confidence ( $1 - \alpha$ )” (Young, 2010). Tolerance intervals are based on the sampled data; however they allow us to say something about the population distribution. “A tolerance interval differs from a confidence interval in that the former encloses a proportion of the entire population distribution, while the latter is constructed to contain the value of a population parameter” (Millsap, 1988). Frequently, tolerance intervals are based on a specified distribution of the data, and more often than not, that distribution is assumed to be normal. However, we can also make the tolerance interval more general by taking a non-parametric approach (Wilks, 1941). In Wilks’ paper, he proves that there is a systematic way of calculating a confidence interval for an unknown data distribution; a tolerance interval can be calculated for a given population coverage proportion, level of confidence and minimum sample size, that guarantees at least the given population coverage proportion. This approach is outlined as follows:

- Let  $a$  be the average value which  $p$  is to have, where  $p$  is the proportion of the population to be included in the interval (the mean coverage).
- Draw a sample of size  $n$  from the population subject to the constraint that  $[(1-a)(n+1)]/2 = r$ , a positive integer.
- Order the sampled data according to increasing magnitude from  $x_1$  to  $x_n$ .
- Let  $L_1 = x_{n-r+1}$ , the upper tolerance limit of our 1-sided non-parametric tolerance interval.

As noted above, we need a minimum sample size value to ensure proper coverage of the population with the specified level of confidence. Within the framework of this project, we utilized the principles of this method to create a 95% non-parametric one-sided tolerance interval for 95% coverage of the IVV distance distribution. The distribution of distances for the IVV

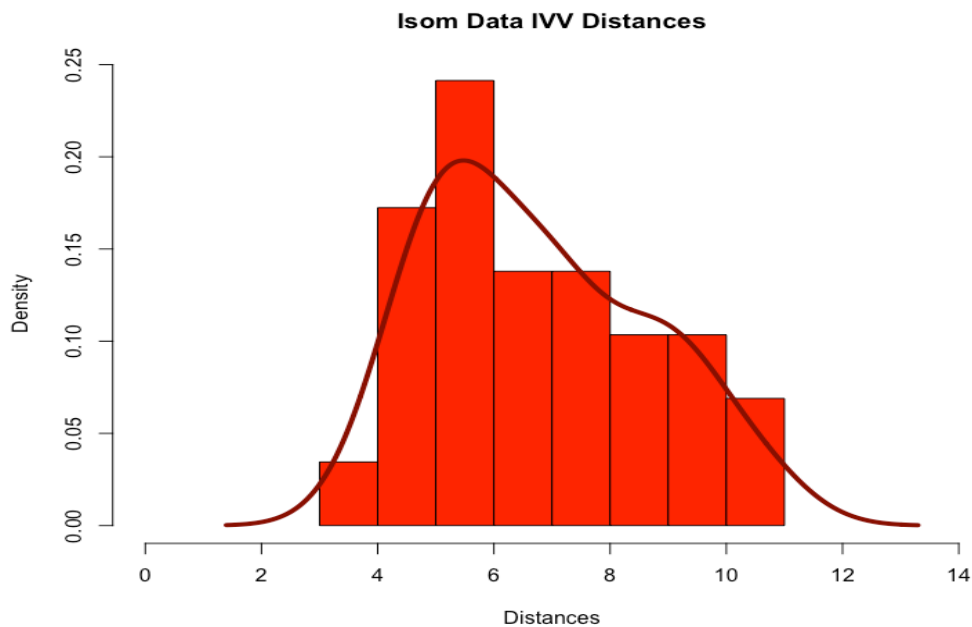


Figure 3. The distribution of IVV distances. The kernel density was overlaid to help depict the skewedness of the distribution.

oocytes is right skewed (see Figure 3), so we used an application of the non-parametric tolerance interval method.

For a non-parametric approach, the calculation of the tolerance interval can be different than when we specify a distribution for the data. As we do not restrict our interval to a specific distribution, we require a larger sample size in order to maintain the same level of confidence and coverage for our tolerance interval than that of a specified distribution (i.e. Gaussian, Weibull, etc). The calculations for the minimum required sample size of a non-parametric tolerance interval are shown below (NIST, 2012):

$$n \approx \frac{1}{4} \frac{(1+p)}{(1-p)} \chi_{1-\gamma, 4}^2 + \frac{1}{2}.$$

The above equation is an approximation for the minimum sample size  $n$ , needed for a non-parametric tolerance interval with confidence level  $\gamma$ , and population coverage proportion  $p$ . We also note that we are calling a specified value from a chi-squared distribution with 4 degrees of freedom.

For our project, the minimum required sample size for a 95% confidence and at least 95% coverage was  $n=94$ , so we ended up needing a sample size of 100. In the simulations of Section 6 below, we use  $n=100$  oocytes, and we proceed with the construction of the tolerance interval here solely for demonstration purposes (though the actual sample size in the project dataset is  $n=58$ , so the actual coverage in this demonstration example is likely less than 95%). For each oocyte in the project, the IVV distance distribution was computed using the same distance



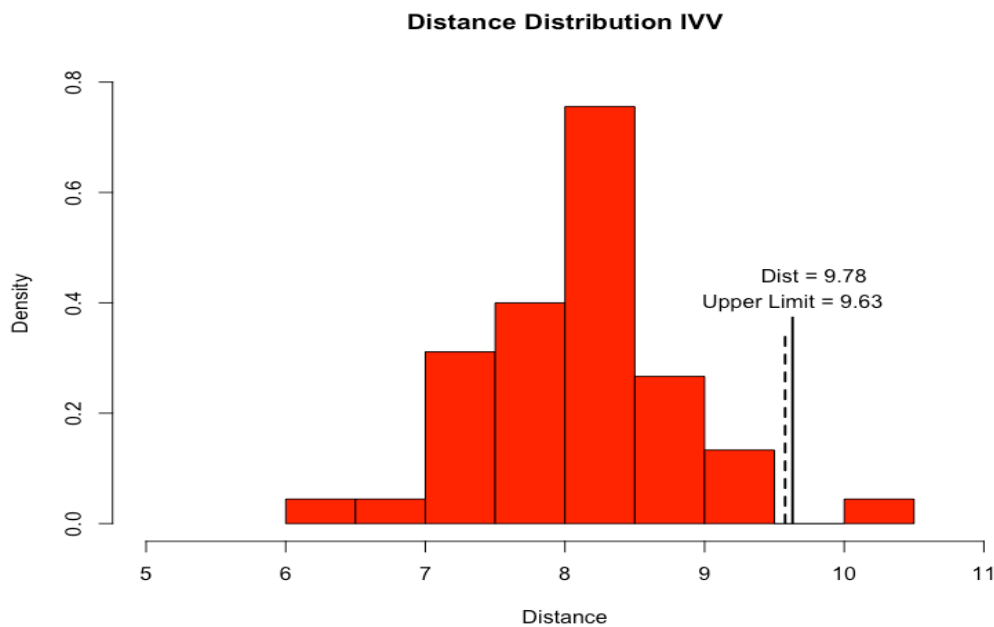


Figure 4. The distance distribution of IVV. The upper limit of the 95% confidence, 95% coverage non-parametric tolerance interval for the IVV group was calculated to be 9.63.

function as previously given in Equation 2. Those oocytes that individually have a distance statistic outside of the constructed tolerance interval for the IVV oocytes (i.e. their distance was greater than that of the upper tolerance bound) were selected as being nonviable (see Figure 4).

In Figure 4, the same example IVM oocyte as in Figures 1 and 2 is shown with distance calculated as 9.78. As the observed distance of 9.78 is larger than the upper tolerance limit of 9.63, we would classify this oocyte as non-viable.

## 5. CLASSIFICATION TREE METHOD

Classification trees or decision trees are a graphical representation of a set of rules used to classify data into categories. They are appropriate to use when one has a 2 or more level categorical variable as an outcome, with one or more variables as predictors. In general, we would use a classification tree to predict the class or outcome level of a number of observations within a dataset, based on the observed values of the predictor variables. In R, the default index for choosing the best split in the data (for classification) is the Gini index. The Gini index is a measure of impurity of a node (or a whole tree). The Gini impurity measure can be calculated in the following way (Kingsford, 2008):

$$Gini = 1 - \sum_{i=1}^m p_i^2.$$

Within the Gini calculation, we are trying to classify items into  $m$  classes using a set of training items  $E$ . Let  $p_i$  ( $i=1, \dots, m$ ) be the fraction of the items of  $E$  that belong to class  $i$ . Thus, the Gini index reaches an optimal value of zero when the set  $E$  contains items from only one class. Construction of a decision tree using this index can be broken down into a number of simple steps:

1. The tree algorithm uses a training data set to build the tree, in other words, one needs to know the true outcomes in order to build a classification tree that can be used on other datasets drawn from the same population.
2. The algorithm then uses the predictor variables to make the most optimal splits in the data. An optimal split is defined as using a variable to split the data from one group into two that provides the greatest differentiation between the different classes, separating them from one another as well as possible.
3. The algorithm uses the most important variables to make splits in the data. If the observations in a specified branch of the tree contain a diverse group of classes, then the algorithm finds “the best” rule based on a single variable/feature to split that branch into two smaller branches. The quality of the split is again measured based upon a reduction in the Gini impurity measure.
4. Only if every observation in the “branch” of the tree is from the same class, does the tree form a terminal node or leaf.
5. One can invoke the *minsplit* argument in the *rpart* function of the package *rpart* (Therneau, 2015) for R (R Core Team, 2016) to indicate the minimum number of observations needed to make a split in the data (to mitigate overfitting and nodes with single observations).

This process of actually using a tree to classify objects was aptly summarized in the following way: “In order to classify an object, we start at the root of the tree, evaluate the test, and take the branch appropriate to the outcome. The process continues until a leaf is encountered, at which time the object is asserted to belong to the class named by the leaf” (Quinlan, 1986).

Within the framework of this project, the outcome or class variable is oocyte viability. A key advantage to a classification tree is that it provides a visual rule for distinguishing between viable and non-viable classifications, based again on the most differentially expressed genes (as was the case with our previous methods using the limma eBayes method). Again, the genes are the variables on which the tree is split, and while all genes have a chance of being chosen initially, the tree will choose the genes to split on that give the best split as calculated by the Gini Index (see Figure 5). Using the same example IVM oocyte as in Figures 1, 2 and 4, we arrive at a classification of non-viable (shown in the red boxed terminal node of Figure 5).

As single trees can have a tendency to over-fit the data, often they are trimmed back slightly such that the last few splits are not used. In order to remedy any possible problems of overfitting, we fit a single tree to the data, and then used a *cp* cutoff value of 0.18, which yielded the smallest possible tree within one standard error of the optimal tree size for that specific dataset. This cutoff value was used with all trees in the simulations of Section 6.

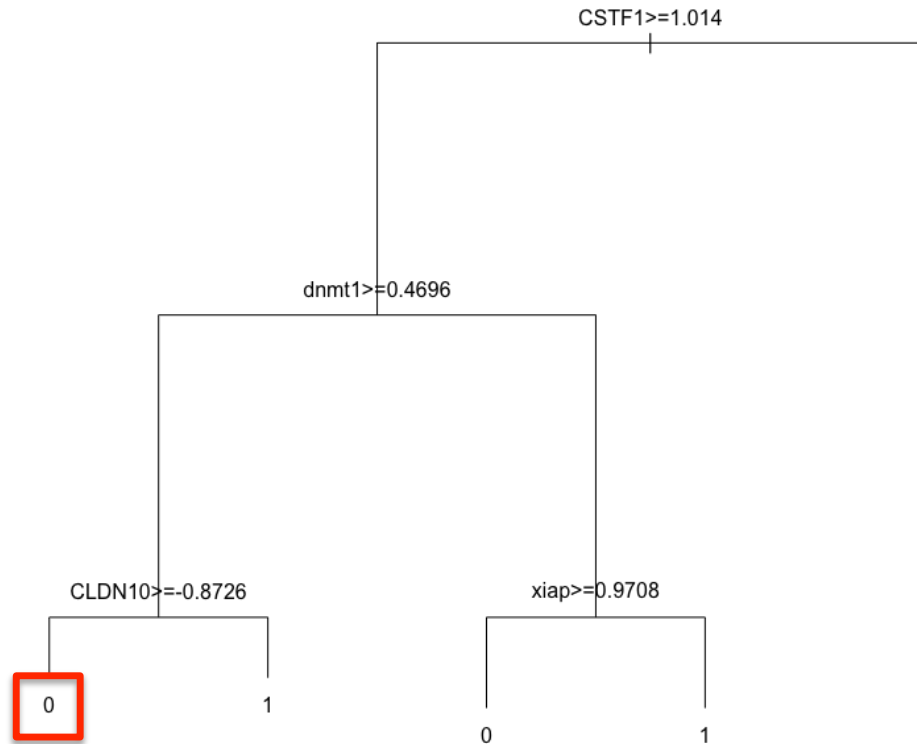


Figure 5. A classification tree that was grown for the dataset; the splits are conducted based upon the gene expression values of the oocytes in the dataset. The resulting classification values of 0 and 1 refer to non-viable and viable respectively.

## 6. SIMULATIONS

In order to adequately compare the four previously summarized methods from Sections 2 through 5, simulations were conducted, in which datasets were generated and results were gathered for the four methods. As the goal was to determine how the methods compared to one another, the simulation was constructed as a function in R, with the following five variables that could be altered to yield different situations for the simulated data:

- $n$ , the number of oocytes to be simulated
- $P_i$ , the mixing proportion for the mixture distribution of the IVM oocytes (i.e. the degree of similarity between the viable IVM group and the IVV group)
- $\delta$ , the magnitude of differential expression for genes that are differentially expressed between IVM and IVV groups
- WhichGenes, a given list of genes to be differentially expressed in the simulation
- ProbViableIVM, the prior probability that any given IVM oocyte is viable

At the start of a simulation, the function starts with the first of the  $n$  oocytes (for our project,  $n=100$ ), and decides whether it will be generated as an IVV or IVM oocyte. This process is done randomly using a binomial generator with probability of IVV equal to 0.5. After

maturation type is decided, we determine if the oocyte will be simulated as a viable oocyte or not. For the sake of the simulation, all IVV oocytes were given the viable class, while the probability that an IVM oocyte is deemed viable depends on the simulation variable “ProbViableIVM” (binomial distribution with probability of viable being equal to “ProbViableIVM”). After each of the oocytes has been assigned a maturation and viability type, the individual gene expression profiles for each oocyte are generated.

Within the process of gene expression profile generation, the simulation function takes into account specific genes that the user wants to make sure are differentially expressed between viable and non-viable classes. The list of these gene names is to be supplied to the function by the user in the simulation argument “WhichGenes” (for our project, these were derived from the Isom data). For the genes identified as being differentially expressed, we then determine the degree of differential expression. Each named gene has its individual degree of differential expression,  $\delta_i$ , which is generated randomly from a uniform distribution, with minimum equal to the user supplied variable  $\delta_{min}$ , and maximum equal to the user supplied variable  $\delta_{max}$  (in our case, these are calculated as the minimum and maximum respectively of the absolute value of the log fold-change for the differentially expressed genes from the Isom data). If the gene of interest is not included in the list of genes to be differentially expressed,  $\delta_i$  is simply set to zero. Next, for each gene within the oocyte, we generate an expression value in the following manner:

$$\text{Degree of Differential Expression for Gene } i = \delta_i \sim U(\delta_{min}, \delta_{max}),$$

$$\text{Viable IVM Expression for Gene } i = \beta * X + (1 - \beta) * Y,$$

$$\text{NonViable IVM Expression for Gene } i = Y \sim N(\delta_i, 1),$$

$$\beta \sim \text{Binom}(1, P_i).$$

The expression value for gene  $i$  of the oocyte is simply a random value from a standard normal distribution if the oocyte is an IVV oocyte. If the oocyte is a non-viable IVM oocyte, then the expression value for gene  $i$  is a random value from a normal distribution with mean equal to  $\delta_i$  and standard deviation equal to one. If the oocyte is a viable IVM oocyte, then the expression level for gene  $i$  is a mixture of the two aforementioned distributions, with the mixing proportion,  $P_i$ , determining the degree of the mixture (i.e. the probability it will be more similar to the IVV group, see Figure 6).

For the purpose of this project, we have chosen only to vary the variables  $P_i$  and ProbViableIVM using the values 0.1, 0.3, 0.5, 0.7 and 0.9; all other variables remained constant throughout the simulations. Constant values and the list of differentially expressed genes came from the provided Isom data. We simply wished to see the effects of our mixing proportion as well as the probability of viable IVM oocytes, and how they affected our methods. Note, as we increase the mixing proportion, we are simply increasing the similarity between viable IVV and

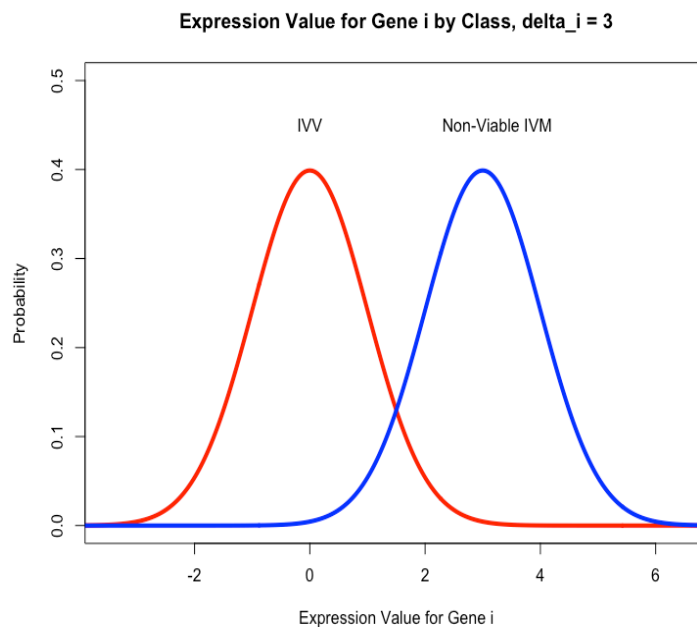


Figure 6. Example expression value distributions for IVV and Non-Viable IVM oocytes for gene  $i$ , such that  $\delta_i=3$  (for example purposes). The distribution of the Viable IVM oocyte expression value is represented as a mixture of the two above distributions, with mixing proportion  $P_i$ .

viable IVM oocytes within the simulation. When we increase the probability of viable IVMs, what we are doing is increasing the probability that a viable IVM oocyte is generated within our simulation set, as our oocyte number stayed constant across simulations. We note that this creates an imbalance in the group sizes, which we can then use to test how our methods handle such conditions. The support and resources from the Center for High Performance Computing at the University of Utah are gratefully acknowledged in providing the means of running the simulations (for a helpful tutorial, see Barton 2016). As the University of Utah system is set up for cluster computing, we were able to run 500 iterations at each simulation level in under 14 hours whereas it would have taken multiple days to compute the same amount on a traditional PC.

In addition to gene expression data generation, the functions created for the simulations also compile results from our primary measures of interest for model evaluation, as averaged over 500 simulations: proportion correctly classified (PCC), sensitivity and specificity for each of the models. In this context, we define proportion correctly classified as the count of those oocytes that were classified as either being viable or non-viable when it was truly their respective viability status, divided by the total number of oocytes that were classified within the simulation (in this case 100). This was repeated 500 times and then averaged across each simulation level combination. A similar process was done for sensitivity and specificity. Sensitivity was defined as the count of correctly classified viable oocytes divided by the total number of viable oocytes, and specificity was defined as the count of correctly classified non-viable oocytes divided by the total number of non-viable oocytes.

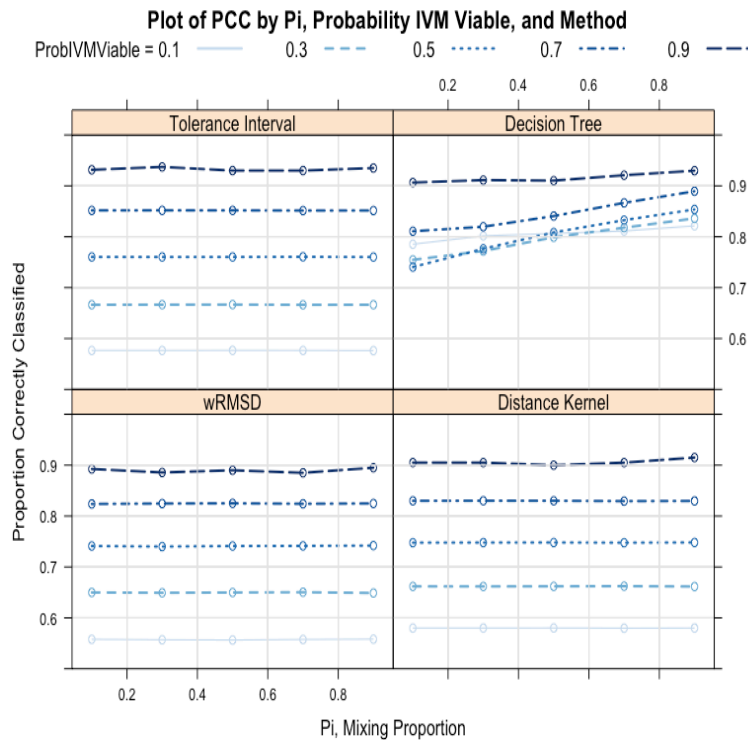


Figure 7. The Proportion Correctly Classified (PCC) of the different simulation combinations, as it specifically relates to the Mixing Proportion ( $\pi$ ). The different line types, as well as the increasing depth of the blue color signify probability of a given IVM being viable across the four different methods.

## 7. RESULTS

On examining Figure 7, we notice that the tolerance method approach tends to outclass the other methods when looking at the proportion correctly classified (PCC). A similar result is seen in Figure 8 when sensitivity (proportion of viable oocytes correctly classified) is used as our measured outcome. In both outcomes, the tolerance interval approach scores roughly two to three percentage points higher on average (after averaging across 500 simulations at each mixing proportion by probability viable IVM combination) than the relational wRMSD and distance kernel methods. In the aforementioned areas, the disparity between tolerance intervals and that of a classification tree depend on the variable ProbIVMViable, as the classification tree method doesn't follow the same exact patterns of results as the other three methods.

Also, in reviewing Figure 7, we note that it appears that the variable  $\pi$  (mixing proportion) seems to have very little effect on the method outcomes. We do see some changes in the slopes of the lines within the classification tree method in Figure 7 and Figure 8; for the other three methods it seems to be of little importance. The reason as to why the variable  $\pi$  has such little difference in the overall results, is that in the design of the simulation (see Section 6), the

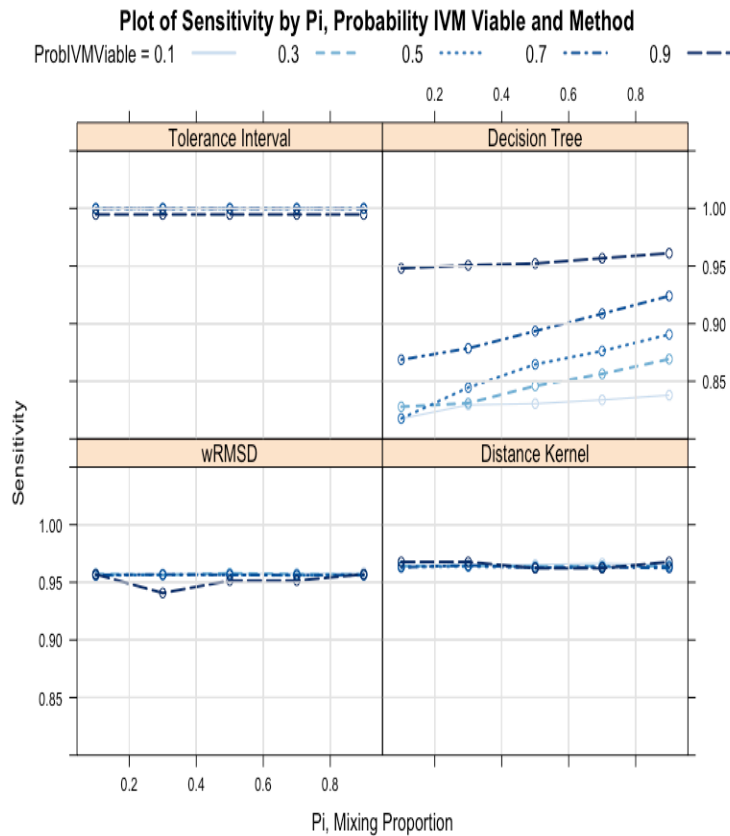


Figure 8. The Sensitivity, or ability to correctly classify the viable oocytes of the different simulation combinations, as it specifically relates to the mixing proportion  $P_i$ . The different line types, as well as the increasing depth of the blue color signify the probability of a given IVM oocyte being viable across the four different methods.

portion of the data that is changed as  $P_i$  is varied is merely a “fraction of a fraction” of the overall data generated in the dataset, as such it generally has little overall effect.

While it appears that the classification tree is generally outclassed by the other three methods (specifically tolerance intervals) in sensitivity, it does outperform all other methods when looking at specificity (the ability of the methods to correctly classify non-viable oocytes, see Figure 9). In Figure 9, it appears that as the probability of an IVM oocyte being viable increases, the harder time the decision tree has at correctly classifying the non-viable samples.

## 8. DISCUSSION

Currently in the fields of cloning and embryology, researchers constantly run into complications and unplanned procedural failures due to the effects of low viability amongst oocytes derived in a laboratory setting. Non-viability of samples is a costly and unwanted outcome that is worth evaluating and attempting to mitigate. Past focus has been on evaluating origination methods and determining their ‘closeness’ to one another; specifically, the method’s

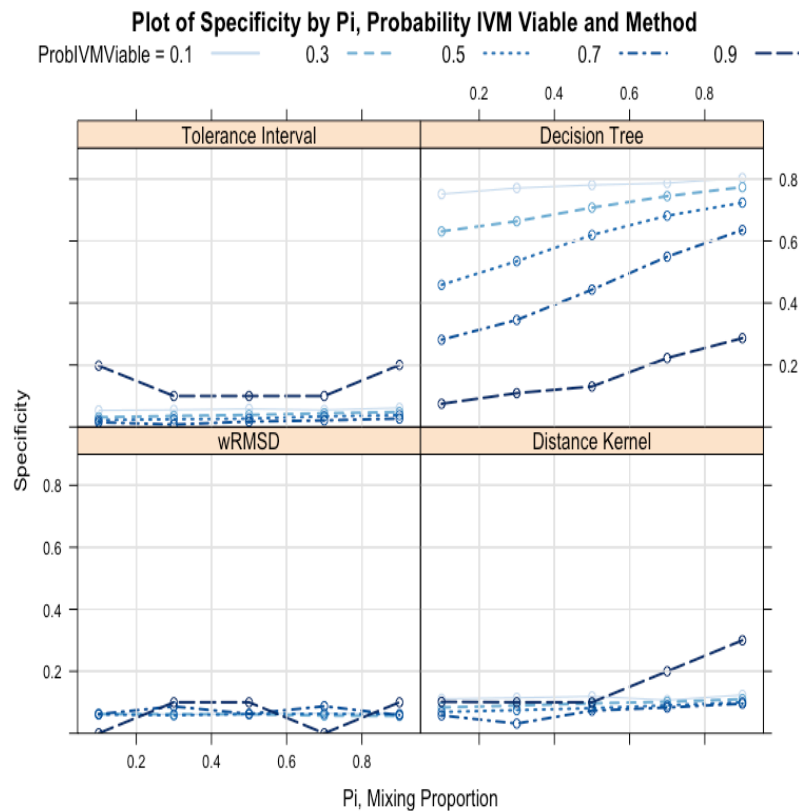


Figure 9. The Specificity, or ability to correctly classify the non-viable oocytes of the different simulation combinations, as it specifically relates to the mixing proportion  $P_i$ . The different line types, as well as the increasing depth of the blue color signify the probability of a given IVM oocyte being viable across the four different methods.

proximity to IVV derived oocytes, the gold standard (Kwon *et al.* 2015). While this may be a decent approach at evaluating origination methods, we claim that we can use alternate approaches to evaluate, and thereby later improve the quality of, individual oocytes within a given method. With recent advances in gene expression technology, nutrient washes can be created in order to encourage the oocytes to express specific genes within a given origination method, thereby improving the viability and quality of the samples, while still retaining the flexibility of using the origination method of choice.

While the goal was to identify methods that were best to use overall, we notice that we actually have a few different situations represented in the data, and as a result we get different “top methods” depending on the situation at hand. As noted above in Figures 7 and 8, when looking at PCC and sensitivity, tolerance intervals seemed to always be the method that performed the best, however the kernel density p-value approach wasn’t far behind, with our modification of the published Kwon wRMSD method following shortly behind that. However, one reason that these methods could be performing so well in these result measurements (PCC and sensitivity), is that there is an imbalance of sample sizes. The viable oocyte group is consistently larger than the non-viable group, simply due to the nature of the probabilities, as all



IVV oocytes are simulated as being viable, while an increasing number of IVM oocytes are counted as viable as the variable ProbIVMViable starts to increase. So, as a reiteration, as the variable ProbIVMViable increases, the disparity between the viable and non-viable group sizes starts to increase (note, when ProbIVMViable is set to 0.90, nearly all oocytes are in the viable group). It appears as though the classification tree method is more balanced and more immune to the effects of this imbalance in the data, as it appears to perform roughly equally across all groups, though it does still get swayed some as ProbIVMViable reaches higher values (see Figures 7, 8, and 9). So, if one is interested in a method that is more robust to the effects of imbalanced samples, or has a vested interest in the accuracy of identifying the non-viable samples (where they severely outperformed the other methods, see Figure 9), classification trees may be the method to choose. If not, we would recommend the non-parametric tolerance interval approach based upon our calculated distance measure.

At this point, we acknowledge the possible limitations of these simulations. First, we note that we have made a number of assumptions that we cannot fully prove with the resources at hand. First, we assume that oocytes that have gene expression profiles that are more like those of IVV oocytes are going to be more viable. Second, we assume the center of the observed IVV distribution to be the optimal gene expression profile. Third, we assume that the data provided to us from Dr. Isom is a representative sample of all IVV and IVM oocytes (viable and non-viable). Fourth, assumptions were made about the relative rates of viability in the different origination method groups based upon conversations with Dr. Isom, as he was our expert in the field. While these are a fair amount of assumptions, we did take precautions as to try to minimize their effects. The methods used were built or conceived in a way that they operate independent of oocyte origination method, they require no distributional assumptions of the oocyte data, and as demonstrated, do not require a complete balance of group sizes (though it may have yielded slightly different results had we forced this on the generated data).

While overall the methods seemed to perform well in specific areas, there are some procedural aspects of these simulations that can be better handled. One such example is the techniques for handling missing values in the data. With oocyte data, when obtaining the gene expression profiles of the oocyte, we destroy the oocyte, thereby making it possible only to attempt to retrieve the profile once; if we miss a few values, we cannot go back and “re-observe” them. Hence, having a set method for handling missing values is advised. One recommendation could be to impute or borrow information from other similar oocytes in the same origination group. For the purposes of these simulations, if there was a missing value in the original Isom data, that value was simply omitted from the gene level calculations. Despite not being able to go back and “re-observe” the missing values, they were fairly infrequent (about 4.4% of the specific gene-level data) in the original dataset.

Further method application and development is necessary in this field, as we showed how under the simulated circumstances that our three proposed methods consistently outperformed the previously published wRMSD method by Kwon *et al.* (2015); however, our proposed methods were quite rudimentary or basic. More advanced classifiers such as Random Forest, neural networks, and the like, may end up doing a better job on these types of classification problems. But, also bear in mind, if the dataset has a large number of variables (such as genes here), then these other methods, while more complex and possibly better classifiers, may require a large increase in processing power and computing time.

## 9. SUMMARY

To this point, viability of laboratory accessible oocytes has stood in the way of progress in numerous fields. We propose an alternate approach to the viability issue, by not only looking at individual genes as coding for viability, but by using gene expression profiles as a way to classify between possible viable and non-viable oocytes. We have proposed four methods: a wRMSD approach, a Distance Kernel Density approach, a Distance Tolerance Interval, and a Classification Tree. Of these approaches, the Distance Tolerance Interval approach was overall the most accurate, however the Classification Tree approach yielded results much quicker and was able to hold a higher specificity score (able to correctly classify the non-viable cases better). While the wRMSD and Distance Kernel Density approaches yielded high results, their percent correctly classified values were lower than the Distance Tolerance Interval approach, and they took more time to run, therefore they are not recommended for future use within this simulation setup. Future real data exploration and further simulation studies are suggested to see how these proposed methods hold up under different conditions.

## 10. ACKNOWLEDGEMENTS

This research was supported (in the form of salaries and conference travel for two authors -- JRS, MB) by the Utah Agricultural Experiment Station (UAES), Utah State University, and approved as journal paper number 9016.

## 11. REFERENCES

- Barton S.W. (2016). "Tutorial for Using the Center for High Performance Computing at The University of Utah and an example using Random Forest". Unpublished M.S. Report, Utah State University. <http://digitalcommons.usu.edu/gradreports/873>
- Bishop M.O. (2017) "Statistical Methods for Assessing Individual Oocyte Viability Through Gene Expression Profiles". Unpublished M.S. Report, Utah State University. <http://digitalcommons.usu.edu/gradreports/916/>
- Kingsford C. and Salzberg S. L. (2008). "What are decision trees?" *Nature Biotechnology*, 26(9), 1011–1013. <http://doi.org/10.1038/nbt0908-1011>.
- Kwon S., Jeong S., Jeoung Y.S., Park J.S., Cui X.S., Kim N.H., and Kang Y.K. (2015). "Assessment of Difference in Gene Expression Profile Between Embryos of Different Derivations", *Cellular Reprogramming*, VOL. 17, NO. 1. DOI: 10.1089/cel.2014.0057.
- Millsap R. E. (1988). "Tolerance Intervals: Alternatives to Credibility Intervals in Validity Generalization Research", *Applied Psychological Measurement*, March 1988, VOL. 12, NO. 1, PP.27-32.
- NIST/SEMATECH (2012). "e-Handbook of Statistical Methods", <http://www.itl.nist.gov/div898/handbook/prc/section2/prc265.htm>, 3-3-17.
- Quinlan J. R. (1986). "Induction of decision trees." *Machine learning* 1.1: 81-106. doi:10.1023/A:1022643204877. <http://link.springer.com/article/10.1023/A:1022643204877>.

- R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>.
- Ritchie M. E., Phipson B., Wu D., Hu Y., Law C.W., Shi W., and Smyth G.K. (2015). “limma powers differential expression analyses for RNA-sequencing and microarray studies”, *Nucleic Acids Research* 43(7), e47.
- Therneau T., Atkinson B., and Ripley B. (2015). rpart: Recursive Partitioning and Regression Trees. R package version 4.1-10. <https://CRAN.R-project.org/package=rpart>
- Thomas F. H. and Vanderhyden B. C. (2006). “Oocyte-granulosa cell interactions during mouse follicular development: regulation of kit ligand expression and its role in oocyte growth”, *Reproductive Biology and Endocrinology* 2006 4:19. doi: 10.1186/1477-7827-4-19. <https://rbej.biomedcentral.com/articles/10.1186/1477-7827-4-19>.
- Wilks S. S. (1941). “Determination of Sample Sizes for Setting Tolerance Limits”. *Ann. Math. Statist.* 12, no. 1, 91--96. doi:10.1214/aoms/1177731788. <http://projecteuclid.org/euclid.aoms/1177731788>.
- Young D. S. (2010). Book Reviews: "Statistical Tolerance Regions: Theory, Applications, and Computation", *Technometrics*, February 2010, VOL. 52, NO. 1, pp.143-144.