A New Machine Learning-Based Framework for Mapping Uncertainty Analysis in RNA-Seq Read Alignment and Gene Expression Estimation

Adam McDermaid$^{1,2}$, Xin Chen$^3$, Yiran Zhang$^{1,4}$, Cankun Wang$^1$, Shaopeng Gu$^4$, Juan Xie$^{1,2}$ and Qin Ma$^{1,2}$*

$^1$Bioinformatics and Mathematical Biosciences Lab, Department of Agronomy, Horticulture, and Plant Science, South Dakota State University, Brookings, SD, United States, $^2$Department of Mathematics and Statistics, South Dakota State University, Brookings, SD, United States, $^3$Center for Applied Mathematics, Tianjin University, Tianjin, China, $^4$Department of Electrical Engineering and Computer Science, South Dakota State University, Brookings, SD, United States

One of the main benefits of using modern RNA-Sequencing (RNA-Seq) technology is the more accurate gene expression estimations compared with previous generations of expression data, such as the microarray. However, numerous issues can result in the possibility that an RNA-Seq read can be mapped to multiple locations on the reference genome with the same alignment scores, which occurs in plant, animal, and metagenome samples. Such a read is so-called a multiple-mapping read (MMR). The impact of these MMRs is reflected in gene expression estimation and all downstream analyses, including differential gene expression, functional enrichment, etc. Current analysis pipelines lack the tools to effectively test the reliability of gene expression estimations, thus are incapable of ensuring the validity of all downstream analyses. Our investigation into 95 RNA-Seq datasets from seven plant and animal species (totaling 1,951 GB) indicates an average of roughly 22% of all reads are MMRs. Here we present a machine learning-based tool called GeneQC (Gene expression Quality Control), which can accurately estimate the reliability of each gene’s expression level derived from an RNA-Seq dataset. The underlying algorithm is designed based on extracted genomic and transcriptomic features, which are then combined using elastic-net regularization and mixture model fitting to provide a clearer picture of mapping uncertainty for each gene. GeneQC allows researchers to determine reliable expression estimations and conduct further analysis on the gene expression that is of sufficient quality. This tool also enables researchers to investigate continued re-alignment methods to determine more accurate gene expression estimates for those with low reliability. Application of GeneQC reveals high level of mapping uncertainty in plant samples and limited, severe mapping uncertainty in animal samples. GeneQC is freely available at http://bmbl.sdstate.edu/GeneQC/home.html.

**Keywords:** gene expression, RNA-Seq read alignment, mapping uncertainty, machine learning, elastic-net, mixture model fitting, k-means clustering, EM-algorithm
INTRODUCTION

RNA-Seq is a revolutionary high-throughput process that allows researchers to observe the genetic makeup of a particular sample (Wang et al., 2009; Garber et al., 2011; Ozsolak and Milos, 2011) and can assist in determining regulatory mechanisms and transcription unit prediction (Chou et al., 2015; Chen et al., 2011) and analyzing regulatory mechanisms (Wang et al., 2009; Garber et al., 2011; Ozsolak and Milos, 2011). Researchers to observe the genetic makeup of a particular sample (Wang et al., 2009; Garber et al., 2011; Ozsolak and Milos, 2011) and can assist in determining regulatory mechanisms and transcription unit prediction (Chou et al., 2015; Chen et al., 2011) and analyzing regulatory mechanisms (Wang et al., 2009; Garber et al., 2011; Ozsolak and Milos, 2011).

expression profiles, in which a discrete expression value for each annotated gene for that species is identified. These gene expression profiles are extracted through computational analysis pipelines (Trapnell et al., 2009; Andrews, 2010; Wang et al., 2010; Grabherr et al., 2011; Kong, 2011; Li and Dewey, 2011; Dobin et al., 2013; Philippe et al., 2013; Wu et al., 2013, 2016; Anders et al., 2015; Bonfert et al., 2015; Chang et al., 2015; Kim et al., 2015; Pertea et al., 2015, 2016; Yuan et al., 2017), which can be analyzed further to identify differentially expressed genes between treatment groups (Robinson et al., 2010; Anders and Huber, 2012; Trapnell et al., 2012; Ritchie et al., 2015; Pimentel et al., 2017; Monier et al., 2018), enriched functional gene modules (Subramanian et al., 2005; Zhou and Su, 2007; Chen et al., 2009; Pathan et al., 2015), co-expression networks (Zhang et al., 2016; Cao et al., 2017), and to generate visualizations to assist in broad interpretations between treatment groups (Goff et al., 2013; Powell, 2015; Younesy et al., 2015; Ge, 2017; Harshbarger et al., 2017; Nelson et al., 2017; Nueda et al., 2017; McDermaid et al., 2018a; Perkel, 2018), among other applications.

One application of RNA-Seq analysis pipelines is to use the sequenced RNA-Seq reads (or reads for short) with a reference genome, if available, to estimate the expression level of each gene (Nagalakshmi et al., 2008; Miller et al., 2014). The basic process is to map these reads to the location with the best alignment score on the reference genome (Wu et al., 2014). Even though numerous methods have been developed to facilitate this analysis, some critical issues persist. The nature of DNA—long strands of millions of base-pairs created by a reordering of the four nucleotides—makes it inevitable that some similarities and duplications will occur throughout the genome. This can lead to ambiguity during read mapping, with specific reads being aligned to multiple locations across the reference genome with the same alignment scores (Li et al., 2009; Oshlack et al., 2010; Swam, 2013; Trapnell et al., 2013; Baruzzo et al., 2017).

This MMR problem can be observed in any genomic region, including, exons and transcripts. For conciseness, we refer to these genomic regions simply as “genes.” This issue has been observed in many diploid species, including human and other mammals and Arabidopsis (Albrecht et al., 2009; Cho et al., 2009; Yoder-Himes et al., 2009; Zhu et al., 2011; Network CGAR, 2018a), as well as many multiploid species (Consortium IWGS, 2014). In some species, such as Glycine max, up to 75% of the genes have the duplicated partners in its genome (Schmutz et al., 2010). For species with high levels of uncertainty, especially angiosperms, the MMR problem can have serious implications on gene expression levels and can be extremely hard to remediate due to the genes’ and chromosomes’ duplicative nature. To more fully investigate the prevalence of MMRs in current RNA-Seq analyses, we analyzed almost two terabytes of RNA-Seq data from seven plant and animal species. Upon analysis of this data, it was clear that a large amount of MMRs was present in a variety of data. Thus, mapping uncertainty is inevitably affecting the gene expression estimates and eventually causing bias in downstream analyses.

During our initial investigation into the MMR problem, 95 datasets totaling 1,951 GB were analyzed. Both paired- and single-end reads were collected from NCBI (Coordinators, 2016), URGI (https://urgi.versailles.inra.fr/), and JGI (Nordberg et al., 2013) for seven plant and animal species. These species include Arabidopsis thaliana, Vitis vinifera, Solanum Lycopersicum, Panicum Virgatum, Triticum Aestivum, Homo sapiens, and Mus musculus. The 95 datasets average 20.6 GB, with an average overall alignment rate of 81.87%. Each dataset was aligned using HISAT2 (Kim et al., 2015) against the appropriate reference genome. Alignment statistics were collected or calculated from the HISAT2 output file, as shown in Table 1. It was determined that an average of 22% of all reads were ambiguously aligned in each of the seven distinct plant and animal species. In four datasets, over 35% of the reads were ambiguously aligned, and over two-thirds of the analyzed datasets having at least 18% of the reads multi-mapped. Panicum virgatum exhibited the highest overall proportions—ranging from 17 to 33%—of MMRs over all analyzed datasets, while Arabidopsis thaliana displayed the lowest proportion, ranging from 8 to 17%. The other analyzed species had similar percentages of MMRs. More details of the MMR analyses over these 95 datasets can be found in Supplementary File 1.

The general solution of the MMR problem in previous studies is to discard or evenly distribute to all potential locations, leading to severe, biased underestimation or overestimation of the gene expression levels, respectively (Kim et al., 2013). More commonly, a proportional assignment of ambiguous reads, in which the read is segmented in smaller portions based on the number of possible mapping locations and uniquely mapped reads to each of them (Li et al., 2009). Recently, additional methods have been employed to attempt remediation of mapping uncertainty after initial alignment (Li and Dewey, 2011; Kahles et al., 2015; Bray et al., 2016). However, even these realignment strategies do not provide a thorough method for evaluation of the alignment quality. While RNA-Seq pipelines traditionally begin with read-level quality control using FastQC (Andrews, 2010), no such method currently exists for controlling the quality of gene expression estimation after read alignment.

If researchers continue processing RNA-Seq data with such high levels of mapping uncertainty, all downstream analyses will have skewed and biased results. Just as raw reads require quality control (Andrews, 2010) so do gene expression estimates based on mapping results. Even with tools that are specifically designed to address mapping uncertainty, such as MMR (Kahles et al., 2015), the quality of the derived gene expression estimates based on mapping results still requires investigation, especially in real datasets not simulated datasets. Without some quality control for gene expression estimation, researchers could potentially be using unreliable data, and blindly doing so.
One promising method for addressing the issue of gene expression-level quality control is the implementation of machine learning. It uses or relates to following concepts or algorithms including statistics, artificial intelligence, philosophy, information theory, biology, cognitive science, computational complexity and control theory to give computers and algorithms the ability to learn and improve performance on a specific task without being explicitly programmed (Mitchell, 1997). Machine learning has two main categories: supervised and unsupervised learning. The majority of practical studies use supervised learning methods to train the relationship from the input to the output, using provided category labels or resultant values to develop a mapping function for the prediction of unlabeled data. Specifically, Elastic-net regularization, a supervised method, was used in this research. Meanwhile, machine learning can also be used to train a model from unlabeled data through the unsupervised learning, aiming to model the underlying structure or distribution in the training data for clustering and association problems. Two unsupervised learning algorithms were used in this study, i.e., K-means clustering and the Expectation-Maximization algorithm (EM-algorithm).

To address issue of mapping uncertainty, we present the machine learning-based tool GeneQC (Figure 1), which uses extracted multi-level features combined with novel applications of regularized regression and mixture model fitting approaches to quantify the mapping uncertainty issue (McDermaid et al., 2018b). This tool can determine the genes having reliable expression estimates and those require further analysis, along with a statistical significant evaluation of the mapping uncertainty level. GeneQC develops a novel score, referred to as D-score, to represent the level of mapping uncertainty for each annotated gene and groups genes into several categorizations with different reliability levels, through integration and modeling of three genomic and transcriptomic features. Specifically, (i) sequence similarity between a particular gene and other genes is collected to give an insight into the genomic characteristics contributing to the MMR problem; (ii) the proportion of shared MMR between gene pairs provides information regarding the transcriptomic influences of mapping uncertainty within each dataset; and (iii) the degree of each gene, representing the number of significant gene pair interactions resulting from calculating (i) and/or (ii). More details of the procedure can be found in the Methods section.

**METHODS**

**GeneQC Implementation**

GeneQC is designed to fit into computational pipelines for RNA-Seq data immediately following read alignment, acting as a supplement to most current pipelines. GeneQC is composed of two distinct processes: feature extraction and statistical modeling. The feature extraction process is implemented using a Perl program and the statistical modeling is performed on the feature extraction output using an R package, which provides the final output for GeneQC (http://bmbl.sdstate.edu/GeneQC/download.html). More details on the implementation of GeneQC can be found at http://bmbl.sdstate.edu/GeneQC/tutorial.html.

**Required Inputs**

GeneQC takes as inputs three pieces of information that are easily found in most RNA-Seq analysis pipelines: (1) the read mapping result SAM file; (2) the fasta reference genome corresponding to the to-be-analyzed species; and (3) the species-specific annotation general feature format (gff/gff3) file (Figure 1B). Example datasets can be found on the GeneQC websolver at http://bmbl.sdstate.edu/GeneQC/result.html.

**Feature Extraction**

From input information, GeneQC first performs feature extraction, in which the three characteristics are calculated for each annotated gene (Figure 1C). The first extracted feature ($D_1$) is derived from genomic level information and involves the similarity between two genes (Figure 2A). For each gene, this is calculated as the maximum of the sequence similarity multiplied by the match length, where the match length is the longest continuous string of matching base pairs. More specifically, $D_1 = \max_y s_{i,y}l_{i,y}$, where $s_{i,y}$ is the base pair sequence similarity of gene $i$ and gene $y$ and $l_{i,y}$ is the match length of these two genes.

The second feature ($D_2$) comes from transcriptomic level information and represents the proportion of shared MMRs (Figure 2B). This value is calculated as the maximum...
FIGURE 1 | Mapping Uncertainty and GeneQC. (A) The MMR percentages for the 95 datasets across seven species. More detailed information is showcased in Table 1; (B) GeneQC takes a read alignment, reference genome, and annotation file as inputs; (C) The first step of GeneQC is to extract features related to mapping uncertainty for each annotated gene; (D) Using the extracted features, elastic-net regularization is used to calculate the D-score, which represents the mapping uncertainty for each gene; (E) A series of Mixture Normal and Mixture Gamma distributions are fit to the D-scores; and (F) The mixture models are used to categorize the D-scores into different levels of mapping uncertainty along with a statistical alternative likelihood value for each gene.

FIGURE 2 | Genomic, transcriptomic, and network feature development. (A) Genes with significant similarity are displayed, with $D_1$ being the maximum value of $s_{ss_i} \cdot l_{ij}$. In this situation, genes $y_2$, $y_3$, and $y_4$ all have the same $s_{ss_i}$ value, but gene $y_3$ has a longer consecutive string of matching base pairs ($l_i$) than the other values, making it the more similar genomic location. (B) Graphical representation of the sets of reads aligned to each gene. $D_2$ is the largest overlapping proportion of shared ambiguous or multi-mapped reads between the target gene, gene $i$, and all other genomic locations that have at least one read potentially aligned to both locations. (C) This graph displays the significant interactions of gene $i$ with other genomic locations. Each node represents a genomic location, with the red edges representing sequence similarity scores and black edges representing multi-mapping proportions. In this situation, $D_1 = 310$, $D_2 = 0.24$, and $D_3 = (3 + 1) = 0.602$.

The proportion of shared MMRs between the gene of interest and another gene. In other words, $D_2 = \frac{|G_i \cap X|}{|G_i|}$, where $G_i = \{all \ reads \ aligned \ to \ gene \ i\}$ and $X = |G_i \cap Y|$. The third feature ($D_3$) is a network factor that represents the number of alternate gene locations with significant interactions with the gene of interest based on the previous two parameters.
Thus, we approximate the true alignment as \( \hat{R}_i \) must fall somewhere between these two values, with \( |U_i| \leq |R_i| \leq |G_i| \). Thus, we approximate the true alignment as \( |\hat{R}_i| = \frac{|G_i| + |U_i|}{2} \).

Using this approximation, we calculate

\[
D_4 = 1 - \frac{|\hat{R}_i|}{|G_i|} = 1 - \frac{|G_i| + |U_i|}{2|G_i|}
\]

Elastic-Net Regularization

To develop a model evaluating the severity of mapping uncertainty and thus expression estimation quality, a regression approach is utilized. Ordinary least squares has been demonstrated to have particular issues when dealing with real world data, especially data that does not fit linearity, homoscedasticity, lack of serious multi-collinearity, or other requirements (Dempster et al., 1977). Because of this, alternative approaches were explored. Ridge regression, which develops a continuous shrinkage and variable selection (Tibshirani, 1996). While this is an appealing feature in generating a model, lasso has shortcomings when it comes to dealing with variables exhibiting high pairwise correlation (Zou and Hastie, 2005). Elastic-net regularization—sometimes referred to as lasso with potential to overcome the shortcomings of both ridge and lasso regression methods by implementing a combination of the two approaches.

Take the set of \( n \) response variables \( y = (y_1, y_2, \ldots, y_n)^T \), a set of \( p \) predictor variables \( x_i = (x_{i1}, x_{i2}, \ldots, x_{ip}) \), \( i \in \{1, \ldots, n\} \), a set of \( p \) coefficients \( \beta = (\beta_1, \beta_2, \ldots, \beta_p) \), and matrix of predictor variables

\[
X = (x_{11}, x_{21}, \ldots, x_{n1}, x_{12}, \ldots, x_{n2}, \ldots, x_{1p}, \ldots, x_{np})^T
\]

For a given \( \lambda_1, \lambda_2 \geq 0 \), elastic-net regularization uses a criterion based on

\[
L(\lambda_1, \lambda_2, \beta) = \|y - X\beta\|_2^2 + \lambda_2 \|\beta\|_2^2 + \lambda_1 \|\beta\|_1
\]

\[
\|\beta\|_2 = \sqrt{\sum_{j=1}^{p} \beta_j}
\]

\[
\|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|
\]

Thus, the set of coefficient estimates \( \hat{\beta} \) are calculated as

\[
\hat{\beta} = \arg\min_{\beta} L(\lambda_1, \lambda_2, \beta) = \arg\min_{\beta} \|y - X\hat{\beta}\|_2^2 + \lambda_2 \|\hat{\beta}\|_2^2 + \lambda_1 \|\hat{\beta}\|_1
\]

Given \( \alpha = \frac{\lambda_2}{\lambda_1 + \lambda_2} \), solving for \( \hat{\beta} \) is equivalent to optimizing

\[
\hat{\beta} = \frac{y - X\beta}{\|\beta\|_2^2} \|\beta\|_1 \|\beta\|_2^2 + (1 - \alpha) \|\beta\|_1 \leq k, \text{ for some } k
\]

In the construction of this elastic net, \( \alpha \|\beta\|_2^2 + (1 - \alpha) \|\beta\|_1 \) is considered as the elastic net penalty, representing a combination of the penalties used in ridge and lasso regression methods. In the situation where \( \alpha = 1 \), the elastic net is equivalent to basic ridge regression. For \( \alpha = 0 \), the approach becomes lasso regression (Zou and Hastie, 2005).

GeneQC utilizes the elastic-net regularization method (Zou and Hastie, 2005) with default \( \alpha = 0.5 \) to develop a regression model for the calculation of D-scores. Here, elastic-net regularization is used to properly perform the variable selection, while simultaneously fitting a sufficient model to the provided data (Figure 1D). This approach also accounts for potential serious multicollinearity issues which were detected in some of the test data and prevents overfitting of the regression model (Zou and Hastie, 2005). The set of calculated D-scores represents the mapping uncertainty for each annotated gene and is provided to give researchers an idea of how reliable their initial read mappings are. A higher D-score represents more mapping uncertainty, and thus a less reliable expression estimate.

Mixture Model Fitting

Based on the calculated sets of D-scores through above GeneQC development, there are apparent conditions for these scores, intuitively representing levels of mapping uncertainty. For this purpose, extensive mixture model fitting is included within GeneQC to best fit a mixture model distribution with three sub-distributions to each set of D-scores (Figure 1E).

Our mixture model fitting process involves \( k \)-means initialization with randomized initial grouping. Cluster means,
\(\mu_i\)

are then calculated for each of the \(k\) clusters, followed by two iterative steps: (1) reassignment of data points to the cluster with the lowest distance between a data point and cluster mean, and (2) recalculation of cluster centers. This process is continued until achieving the minimum within-cluster sum of squares:

\[
\arg\min_k \sum_{i=1}^{k} \sum_{x \in k_i} \|x - \mu_i\|^2
\]

After initialization using the \(k\)-means process defined above, the EM-algorithm is implemented to find the best fitting distributions. Based on our preliminary investigations into the D-score development, we have selected two underlying distributions for this purpose: Gamma and Gaussian. Specifically, it is assumed that each set of D-scores can be expressed as a mixture model distribution given by

\[
P(X|\theta) = \sum_k \beta_k Y_k(X|\theta_k)
\]

with \(\beta_k\) representing the weighting parameter of the \(k\)th component, \(Y_k\) representing the probability density function of the \(k\)th component of the mixture model, and \(\theta_k\) representing the parameters of the \(k\)th component. Considering the Gaussian distribution scenario, \(Y_k(X|\theta_k) = N(X|\mu_k, \sigma_k^2)\). In this case,

\[
\begin{align*}
\text{MLE}(& \mu_k) = \hat{\mu}_k = \frac{\sum_{j} x_{jk}}{N_k} \\
\text{MLE} (\sigma_k^2) &= \hat{\sigma}_k = \frac{\sum_{j} (x_{jk} - \hat{\mu}_k)^2}{N_k} \\
\beta_k &= \frac{N_k}{N}
\end{align*}
\]

\[
x = \left(-\frac{\mu_i + \sigma_i^2 - \mu_i^2}{\sigma_i^2} \pm \sqrt{\frac{2\sigma_i^2 \sigma_{i+1}^2 \ln \left( \frac{\sigma_{i+1}^2}{\sigma_i^2} \right) - \mu_i^2 \sigma_{i+1}^2 + \mu_i \sigma_{i+1}^2}{\sigma_{i+1}^2 - \sigma_i^2}} \right)^2 + \left(\frac{\mu_{i+1} \sigma_i^2 - \mu_i \sigma_{i+1}^2}{\sigma_{i+1}^2 - \sigma_i^2}\right) \quad \text{for } i \in [1, 2].
\]

These parameter estimates are then used as the parameters for the next Expectation step, through which this process iteratively continues until convergence, i.e., no significant improvement in the log-likelihood is achieved from the previous iteration. This process is implemented iteratively to quickly generate a series of mixture model distributions for both Gamma and Gaussian distributions.

The optimally fitted mixture model is determined using a Bayesian Information Criterion (BIC) with a penalization based on the number of distributions is used to determine the best-fitting distribution. The BIC for a mixture distribution \(K\) is based on the number of sub-distributions \(k\), the number of data points \(n\), and the log likelihood \(\hat{L}\).

\[
\text{BIC}(K) = -2k\log(n) - 2\hat{L}
\]

### Mapping Uncertainty Categorization

The best fitting mixture model is then used to separate each D-score into a category representing the severity of mapping uncertainty, thus indicating the mapping uncertainty categorization for each gene (Figure 1F). The categorizations are based on the intersections of the density functions representing the mixture model fitting. If the Gaussian distributions provide the minimal BIC, the categorization cutoffs are calculated as

\[
\text{mean} \left(\arg\max_x \left(\delta_i(x) > \delta_{i+1}(x)\right)\right), \text{mean} \left(\arg\max_x \left(\delta_i(x) < \delta_{i+1}(x)\right)\right)
\]

resulting in two cutoff values.
TABLE 2 | GeneQC example output.

<table>
<thead>
<tr>
<th>Gene ID</th>
<th>D1</th>
<th>D2</th>
<th>D3</th>
<th>D-score</th>
<th>Category</th>
<th>Alternative likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td>gene17958</td>
<td>1439.981</td>
<td>0.022727</td>
<td>1.041393</td>
<td>0.02765</td>
<td>Low</td>
<td>0.106445</td>
</tr>
<tr>
<td>gene299138</td>
<td>228</td>
<td>1</td>
<td>0.69897</td>
<td>0.50935</td>
<td>High</td>
<td>0.012702</td>
</tr>
<tr>
<td>gene17991</td>
<td>2560</td>
<td>1</td>
<td>0.477121</td>
<td>0.498094</td>
<td>High</td>
<td>0.015754</td>
</tr>
<tr>
<td>gene24086</td>
<td>321.9967</td>
<td>0.005017</td>
<td>2.060698</td>
<td>0.020863</td>
<td>Low</td>
<td>0.10397</td>
</tr>
<tr>
<td>gene23209</td>
<td>365</td>
<td>0.0224</td>
<td>1.78533</td>
<td>0.027916</td>
<td>Low</td>
<td>0.113361</td>
</tr>
<tr>
<td>gene429</td>
<td>157</td>
<td>0.04878</td>
<td>0.954243</td>
<td>0.031332</td>
<td>Low</td>
<td>0.120682</td>
</tr>
<tr>
<td>gene15873</td>
<td>691.9874</td>
<td>0.780952</td>
<td>0.477121</td>
<td>0.391438</td>
<td>Medium</td>
<td>2.15E-54</td>
</tr>
<tr>
<td>gene24903</td>
<td>855</td>
<td>1</td>
<td>0.477121</td>
<td>0.499807</td>
<td>High</td>
<td>0.015276</td>
</tr>
<tr>
<td>gene26458</td>
<td>4864</td>
<td>1</td>
<td>0.477121</td>
<td>0.496779</td>
<td>High</td>
<td>0.016419</td>
</tr>
</tbody>
</table>

Due to the nature of mapping uncertainty and the lack of current approaches to evaluate this concept, we have included an alternative likelihood value, for the first time, as a proposed method of evaluating the mapping uncertainty categorizations computationally. This value based on the posterior probabilities of the other distributions is provided to represent the certainty of the gene ID belonging to that category. This value \((s_d)\) is computed as the maximum posterior probability of the D-score belonging to any other categorization distribution.

\[
s_d = \max \{ 1 - F_{i-1}(d) , F_{i+1}(d) \}
\]

where \(i\) is the distribution for which \(d\) is categorized, and \(F_j\) represents the cumulative distribution function of distribution \(j\).

RESULTS

GeneQC Output

The final output of GeneQC includes the three extracted features (named \(D_1\), \(D_2\), and \(D_3\)), D-score, mapping uncertainty categorization, and alternative likelihood for each annotated gene. This information is combined into a concise table to provide users with all relevant information related to the mapping uncertainty of their read alignment data, allowing them to make informed decisions about further and continued analysis. An example of the output file from **Vitis vinifera** can be found in Table 2. For each annotated gene, the D-score indicates the severity of mapping uncertainty for that particular gene in this particular RNA-Seq data. A higher D-score indicates a higher level of mapping uncertainty with maximum levels of mapping uncertainty occurring around 0.5 for most samples. Genes with relatively high D-scores have mapping uncertainty issues resulting in potentially unreliable expression estimates (i.e., the High category). Whereas, genes with D-scores close to 0 have little to no mapping uncertainty, and therefore have reliable expression estimates (i.e., the Low and Medium categories).

Source code and implementation instructions can be found on the GeneQC web server at http://bmbl.sdstate.edu/GeneQC/home.html. Additionally, example data for seven analyzed species can be downloaded on this server, including all reference genomes, annotations, original raw data, and outputs from both the feature extraction and modeling portions of GeneQC. An in-depth tutorial for application instructions can also be found on this site.

Implementation and Application of GeneQC Results

GeneQC has four main applications in RNA-Seq analyses. (1) Users can take the D-score and categorization results from an entire dataset to evaluate the alignment quality of their data or to determine how severe the overall mapping uncertainty is within their RNA-Seq datasets. This process would involve displaying the set of D-scores in some visualization technique, such as a boxplot, violin plot, or histogram. Displaying the D-scores in this format would allow for users to determine if the overall alignment quality is sufficient to continue analysis or if it requires further evaluation using a re-alignment method. It is expected that there will be high D-scores for some genes; however, a large portion of data having high D-scores would indicate severe problems with alignment requiring further analysis. (2) Users can use D-scores and mapping uncertainty categorizations to evaluate the reliability of their downstream analyses, such as differential gene expression results. If users have identified a particular set of genes that are differentially expressed, it would be of interest to evaluate the reliability of the expression estimates from which those comparisons were made. Genes identified


<table>
<thead>
<tr>
<th>Species</th>
<th>Sample ID</th>
<th>Mean D1</th>
<th>Mean D2</th>
<th>Mean D3</th>
<th>Mean D-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. thaliana</td>
<td>SRIR3305038</td>
<td>0.02</td>
<td>0.58</td>
<td>0.01</td>
<td>0.29</td>
</tr>
<tr>
<td>V. vinifera</td>
<td>SRIR2080995</td>
<td>0.04</td>
<td>0.46</td>
<td>0.16</td>
<td>0.24</td>
</tr>
<tr>
<td>S. lycopersicum</td>
<td>SRIR5274891</td>
<td>0.06</td>
<td>0.66</td>
<td>0.04</td>
<td>0.33</td>
</tr>
<tr>
<td>P. virgatum</td>
<td>SRIR5188117</td>
<td>0.01</td>
<td>0.32</td>
<td>0.09</td>
<td>0.16</td>
</tr>
<tr>
<td>T. aestivum</td>
<td>ATW_AACSW_6.2</td>
<td>0.02</td>
<td>0.60</td>
<td>0.15</td>
<td>0.31</td>
</tr>
<tr>
<td>H. sapiens</td>
<td>SRIR6029567</td>
<td>0.05</td>
<td>0.84</td>
<td>0.32</td>
<td>0.43</td>
</tr>
<tr>
<td>M. musculus</td>
<td>SRIR611161</td>
<td>0.06</td>
<td>0.84</td>
<td>0.28</td>
<td>0.42</td>
</tr>
</tbody>
</table>

This table shows the sample ID and relevant metrics for each of the seven datasets analyzed. Mean values for \(D_1\), \(D_2\), \(D_3\), and D-score are calculated based on the genes that exhibit some level of mapping uncertainty, and \(D_1\), \(D_2\), and \(D_3\) were normalized for comparison.
as differentially expressed having high mapping uncertainty levels—either through D-scores or categorization—would be less reliable than the differentially expressed genes that have low mapping uncertainty. (3) GeneQC can be used to directly compare the severity of mapping uncertainty between samples or even between species. This application method is used in section GeneQC Application: Analysis of Seven Plant and Animal Species to demonstrate which species have relatively high levels of mapping uncertainty and to determine which characteristics or features could be affecting this issue. In particular, identification of characteristics impacting mapping uncertainty for a single species could provide information that would assist in realignment processes. (4) GeneQC can be used to perform large-scale comparisons of alignment tools using real data. Currently, comparisons of alignment tools require either simulated data which cannot accurately replicate the complexities within real RNA-Seq data, or they rely on small-scale real data, which has implicit biases that may favor one tool. GeneQC allows for the large-scale comparisons of alignment methods with complex data of any species.

**GeneQC Application: Analysis of Seven Plant and Animal Species**

In order to display the use of GeneQC, one dataset from each of the seven species were investigated for multi-mapping issues (Table 3). Based on this analysis, it is evident that plant samples tend to have higher proportions of genes with mapping uncertainty than animal samples (Figure 3A). These results correlate with the fact that plant genomes tend to have higher levels of duplication, which is a strong contributing factor to mapping uncertainty. While *H. sapiens* and *M. musculus* have lower proportions of genes with mapping uncertainty than the plant samples, the proportion of genes with high mapping uncertainty of all the genes with mapping uncertainty is much higher. Plant species exhibited mapping uncertainty in an average of 12.6% of genes across the five species, whereas animal species exhibited this issue in an average of 5% of genes (Supplementary Files S2, S3). However, over half of the genes with mapping uncertainty in the animal samples fall into the “High” categorization, while only around one-fifth of genes with mapping uncertainty from plant samples fall into this category. The contributing factors to the higher proportion of “High” categorized genes for animal samples can be seen when looking at the three extracted features for each species.

The analysis results for the three features and calculated D-scores for genes with some level of mapping uncertainty are displayed in Figures 3B,C, respectively. Both *H. sapiens* and *M. musculus* display higher levels of sequence similarity (D1), shared MMR proportion (D2), and degree (D3) than what is generally exhibited in the analyzed plant species. These relatively high values for each feature led the higher D-scores, translating to a higher measure of mapping uncertainty in the animal samples compared with the plant samples. Mean D-score for *H. sapiens* and *M. musculus* are 0.43 and 0.42, respectively. These average values are much higher than those for the analyzed plant samples, which are 0.29, 0.24, 0.33, 0.16, and 0.31 for *A. thaliana*, *V. vinifera*, *S. lycopersicum*, *P. virgatum*, and *T. aestivum*, respectively.

![Figure 3](https://example.com/figure3.png)

**FIGURE 3 | GeneQC application.** The results related to the analysis of seven datasets representing five plant and two animal species. (A) Categorizations for the level of mapping uncertainty per gene are shown relative to all categorizations. (B) Boxplots for the three extracted features of each gene are shown for each analyzed sample. D1, D2, and D3 represent the sequence similarity, proportion of shared MMR, and degree weight, respectively. Each value is shown normalized between 0 and 1. Only genes with mapping uncertainty are displayed. (C) Derived D-scores for each gene are shown by species, as calculated from the three features in (B). Higher D-scores represent higher levels of mapping uncertainty.
CONCLUSION

GeneQC is a tool used to investigate the prominent issue of mapping uncertainty in modern RNA-Seq analysis through the combination of feature extraction and machine learning methods. Oversight in the quality of derived gene expression estimates based on mapping results can have drastic consequences for all downstream analyses and read mapping uncertainty is a significant cause of problems in further analysis. While read mapping has been accepted as sufficient, entirely ignoring the possibility of poorly mapped reads used for further analysis can have detrimental effects on all manner of RNA-Seq studies. As demonstrated in our analysis of 95 RNA-Seq datasets, the problem of mapping uncertainty is prominent and is displayed directly in the gene expression estimates. GeneQC can provide insight into the severity of this issue for each annotated gene along with a statistical evaluation framework. It utilizes feature extraction, elastic-net regularization, and mixture model fitting to provide researchers with a sense of the quality of gene expression estimates resulting from the read alignment step. GeneQC provides sufficient information for researchers to make more well-informed decisions based on the results of their RNA-Seq data analysis and to plan further analyses to address mapping uncertainty.

The application of GeneQC on the seven analyzed datasets display some interesting differences between plant and animal samples. Fewer genes displayed mapping uncertainty in the animal samples, while a higher proportion of these genes were categorized as “High.” Alternatively, a much higher proportion of plant genes displayed mapping uncertainty, but more of these genes had moderate to low mapping uncertainty, relative to genes from animal samples. Both of these scenarios display the severity of mapping uncertainty in modern RNA-Seq analyses. High mapping uncertainty displayed in animal samples can lead to very biased expression estimates over fewer genes, while moderate levels of mapping uncertainty on a wider scale as displayed in plant species can cause widespread expression estimate biases on a lesser scale.

DISCUSSION

In addition to the direct provisions of GeneQC, interpretations of the coefficients allow for a further examination of the specific features contributing the mapping uncertainty. This will allow for further analysis and re-alignment strategies to be developed to the specific characteristics of the dataset. We are currently using this information to develop a computational tool capable of performing re-alignment of reads currently aligned to genes with high D-scores with the purpose of assisting researchers in the correction of mapping uncertainty. In the future, GeneQC will be integrated into a web server that applies this tool and associated re-alignment tools to perform large-scale RNA-Seq analyses on human, plant, and metagenome datasets. This application will allow for ease-of-use and collection of more data to support research with significant MMR issues.

Additionally, further exploration of machine learning approaches, both supervised and unsupervised, will be explored with respect to their applicability in detecting mapping uncertainty. Large-scale use of simulated data for multiple species will provide a direct indication of the actual expression level, which can be compared with the expression estimate from various high-performing and widely-used alignment tools. The various machine learning methods can then be used to detect mapping uncertainty for each tool, with performance comparisons being derived from the correlation between the predicted mapping uncertainty level from the machine learning algorithm and the difference between actual and estimated expression for each gene. A determination for the best-performing method will be based on the highest correlation and may be alignment tool-specific.

AUTHOR CONTRIBUTIONS

QM conceived the basic idea and designed the analysis. AM, XC, YZ, CW, and QM contributed to development of feature extraction conceptualization, methods, and implementation. AM, SG, JX, and QM contributed to machine learning modeling conceptualization, methods, and implementation. AM, YZ, CW, and QM contributed to the development and maintenance of GeneQC website. AM, SG, and QM contributed to the manuscript development and writing. All authors contributed to the manuscript revisions, read and approved the final version of the manuscript.

FUNDING

This work was supported by National Science Foundation/EPSCoR Award No. IIA-1355423, the State of South Dakota Research Innovation Center and the Agriculture Experiment Station of South Dakota State University (SDSU). This work is also supported by Hatch Project: SD00H558-15/project accession No. 1008151 from the USDA National Institute of Food and Agriculture. This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation (grant number ACI-1548562).

ACKNOWLEDGMENTS

This work has been released as a pre-print (McDermaid et al., 2018b).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2018.00313/full#supplementary-material


**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2018 McDermaid, Chen, Zhang, Wang, Gu, Xie and Ma. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.