GeneQC: A quality control tool for gene expression estimation based on RNA-sequencing reads mapping

Adam McDermaid[1,2], Xin Chen[3], Yiran Zhang[1,4], Juan Xie[1,2], Cankun Wang[1], Qin Ma[1,2,$]

[1]Bioinformatics and Mathematical Biosciences Lab, Department of Agronomy, Horticulture and Plant Science, South Dakota State University, Brookings, SD, USA, [2]Department of Mathematics and Statistics, South Dakota State University, Brookings, SD, USA, [3]Center for Applied Mathematics, Tianjin University, Tianjin, China, [4]Department of Electrical Engineering and Computer Science, South Dakota State University, Brookings, SD, USA, and [$]To whom correspondence should be addressed.

Abstract

Motivation: One of the main benefits of using modern RNA-sequencing (RNA-seq) technology is the more accurate gene expression estimations. 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 test the reliability of gene expression estimations, thus are incapable of ensuring the validity of all downstream analyses.

Results: Our investigation into 95 RNA-seq datasets from seven species (totaling 1,951GB) indicates an average of roughly 22% of all reads are MMRs for plant and animal species. Here we present a tool called ***GeneQC*** (**Gene** expression **Q**uality **C**ontrol), which can accurately estimate the reliability of each gene's expression level. The underlying algorithm is designed based on extracted genomic and transcriptomic features through extensive use of mathematical and statistical modeling and design. GeneQC utilizes big data-driven mathematical modeling approaches and allows researchers to determine reliable expression estimations and conduct further analysis on the gene expression that are of sufficient quality. This tool also enables researchers to investigate continued analysis to determine more accurate gene expression estimates for those with low reliability.

Availability: GeneQC is freely available at http://bmbl.sdstate.edu/GeneQC/home.html.

Contact: qin.ma@sdstate.edu

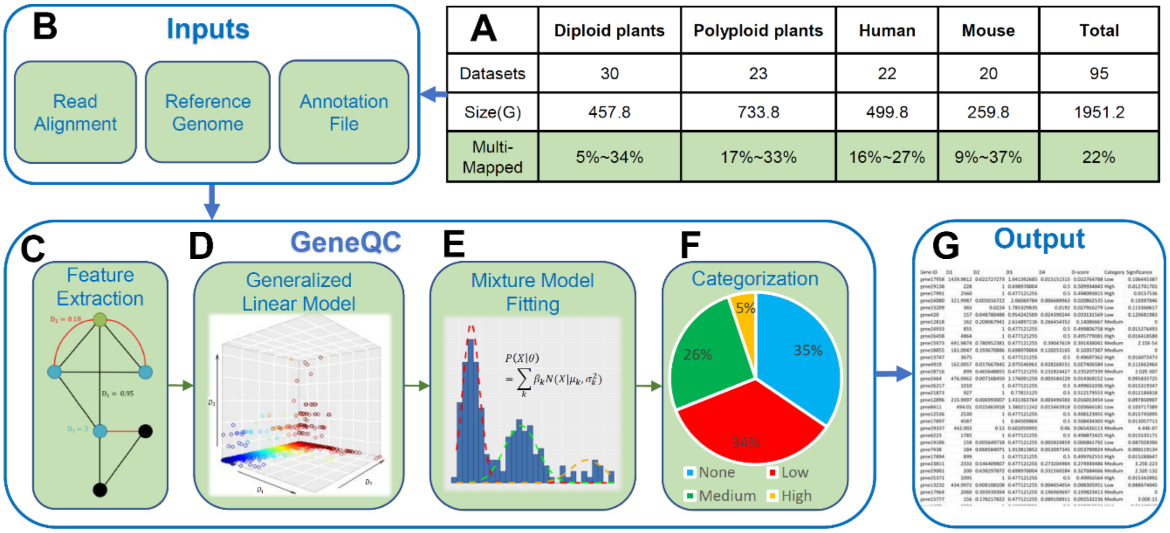
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

RNA-seq is a revolutionary high-throughput process that allows researchers to observe the genetic makeup of a particular sample (Garber, et al., 2011; Ozsolak and Milos, 2011; Wang, et al., 2009). Research involving RNA-seq data produces genetic 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 RNA-seq analysis pipelines (Anders, et al., 2015; Andrews, 2010; Bonfert, et al., 2015; Chang, et al., 2015; Dobin, et al., 2013; Grabherr, et al., 2011; Kim, et al., 2015; Kong, 2011; Li and Dewey, 2011; Pertea, et al., 2016; Pertea, et al., 2015; Philippe, et al., 2013; Trapnell, et al., 2009; Wang, et al., 2010; Wang, et al., 2009; Wu, et al., 2013; Wu, et al., 2016; Yuan, et al., 2017) and can be analyzed further to identify differentially expressed genes between treatment groups (Anders and Huber, 2012; Pimentel, et al., 2017; Ritchie, et al., 2015; Robinson, et al., 2010; Trapnell, et al., 2012), enriched functional gene modules (Chen, et al., 2009; Pathan, et al., 2015; Subramanian, et al., 2005; Zhou and Su, 2007), co-expression networks (Li, et al., 2009), among other applications.

One of the applications of RNA-seq analysis pipelines is to use the sequenced short reads with a reference genome, if available, to estimate the expression level of each gene (Miller, et al., 2014; Nagalakshmi, et al., 2008). The basic process is to map these short 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 the analysis of RNA-seq data, some important 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 (Baruzzo, et al., 2017; D'haeseleer, 2006; Li, et al., 2009; Oshlack, et al., 2010; Swan, 2013; Trapnell, et al., 2013). 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 (Anders and Huber, 2012; Anders, et al., 2015; Bonfert, et al., 2015; Garber, et al., 2011; Wang, et al., 2009), as well as many multiploid species.

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 (Chang, et al., 2015). 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). In 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 (Grabherr, et al., 2011). In some species, such as Glycine, up to 75% of the genes have the duplicated partners in its genome (Kim, et al., 2015). During our initial investigation into the MMR problem, 95 datasets totaling 1,951GB were analyzed, and it was determined that an average of 22% of all reads were ambiguously aligned over seven distinct plant and animal species (Fig. 1A). In some datasets, over 35% of the reads were ambiguously aligned, and more details are provided in *Preliminary Analysis S1* and *Table S1*.



A	Diploid plants	Polyploid plants	Human	Mouse	Total
Datasets	30	23	22	20	95
Size(G)	457.8	733.8	499.8	259.8	1951.2
Multi-Mapped	5%~34%	17%~33%	16%~27%	9%~37%	22%

Figure 1. *(A)* The MMR percentages for the 95 analyzed datasets across seven species. More detailed information can be found in Table S1; (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, a linear model is constructed for calculating 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; (F) The mixture models are used to categorize the D-scores into different levels of mapping uncertainty along with a statistical significance score for each gene; (G) GeneQC outputs a table containing the extracted features, D-score, and mapping uncertainty categorization.

To address this issue, we present GeneQC based on novel mathematical modeling approaches to quantify the mapping uncertainty issue. This tool can determine genes having reliable expression estimates and those require further analysis, along with statistical significant evaluation of the mapping uncertainty level. The basic idea is to develop a distinct score, referred to as D-score, to group genes into several categorizations with different reliability levels, through integration and modeling of 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 for each dataset; and (iii) the degree of each gene, representing the number of gene-gene interactions resulting from (i) and (ii). GeneQC provides values for each extracted feature, a D-score, and mapping uncertainty categorization for each annotated gene corresponding to a provided dataset. More details of the procedure can be found in the following section.

2 Methods

Requires Inputs for GeneQC

GeneQC takes as inputs only 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 that species; and (3) the species-specific annotation gff/gff3 file (Fig. 1B).

Extraction of Relevant Mapping Uncertainty Features

From input information, GeneQC performs feature extraction, in which the three characteristics are calculated for each annotated gene (Figure 1C). The first feature is derived from the genomic level and involves the similarity between two genes. For each gene, this value 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. The second feature comes from the transcriptomic level and represents proportion of shared MMR. Here, the value is calculated as the maximum proportion of shared MMR between the gene of interest and another gene. The third feature collected is a network factor that represents the number of alternate gene location with significant interactions with the gene of interest based on the previous two parameters. In addition to understanding the severity of the MMR problem in each sample, GeneQC provides species- or sample-specific insight into each feature's impact on mapping uncertainty. This is done by developing a linear model to determine the significance and degree of impact for each feature. To perform the linear modeling, a dependent variable is constructed, with more detailed information on this value and each feature found in *Methods S1*.

Sample-specific D-score Development using Linear Modeling

GeneQC utilizes the linear regression models to calculate the optimal coefficients with the three extracted features representing the independent variables and the constructed variable representing the dependent variable (Figure 1D). In this modeling, all possible interaction terms are considered.

$$D = \alpha_1 D_1 + \alpha_2 D_2 + \alpha_3 D_3 + \alpha_4 D_1 D_2 + \alpha_5 D_1 D_3 + \alpha_6 D_2 D_3 + \alpha_7 D_1 D_2 D_3$$

The statistically significant coefficients are then combined, generating the equation to define the D-score. This D-score represents the mapping uncertainty for each annotated gene and is provided to give researchers an idea of how reliable their initial read mappings are, with a higher D-score representing more mapping uncertainty, and thus a less reliable expression estimate. More specific details regarding this process can be found in *Methods S2*.

Mixture Model Fitting

Based on the calculated sets of D-scores through initial investigations during GeneQC development, there are clear underlying distributions for these scores, intuitively representing levels of mapping uncertainty. For this purpose, extensive mixture model fitting is included within GeneQC to best fit a variable number of distributions to each set of D-scores (Figure 1E). 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 β_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 θ_k representing the parameters of the k^{th} component. Based on our preliminary investigations into the D-score development, we have selected two underlying distributions for this purpose: Gamma and Gaussian. GeneQC fits mixture models for both the Gamma and Gaussian distributions with a variable number of distributions, ranging from two to five 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. Specific details regarding the mixture model fitting procedure are outlined in *Methods S3*.

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). In addition to the mapping uncertainty categorization, a significance value based on the posterior probabilities of the other distributions is provided to represent the certainty of the gene ID belonging to that category. Details for the categorization, cutoffs, and significance value are provided in *Methods* S4. The categorizations based on the mixture model fitting of the D-scores is then provided along with each of the three features and the D-score for each gene related to the user-provided mapping and reference genome. These items are organized in tabular form to the user to make informed decisions about further and continued analysis (Figure 1G). An example output file is provided in *Table S2*.

3 Conclusions

GeneQC is a tool used to address an issue in modern RNA-seq analysis. Oversight in the quality of RNA-seq read mapping can have drastic consequences for all downstream analyses, and mapping uncertainty is a significant cause of problems in further analysis. GeneQC can provide insight into the severity of this issue for each annotated gene in terms of three genomic and transcriptomic features. It utilizes mathematical and statistical modeling to combine these features into a distinct score representing the severity of mapping uncertainty and provides a categorization based on fitting mixture models to the derived D-scores, along with a value indicating the significance of this categorization. This information allows researchers to make more well-informed decisions based on the results of their RNA-seq data analysis and to plan further analyses to address this issue.

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 provide 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 correction of mapping uncertainty. In the future, GeneQC will be integrated into a web server that apply 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.

References

Anders, S. and Huber, W. Differential expression of RNA-Seq data at the gene level–the DESeq package. *Heidelberg, Germany: European Molecular Biology Laboratory (EMBL)* 2012.
Anders, S., Pyl, P.T. and Huber, W. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* 2015;31(2):166-169.
Andrews, S. FastQC: a quality control tool for high throughput sequence data. 2010.
Baruzzo, G., *et al.* Simulation-based comprehensive benchmarking of RNA-seq aligners. *Nature methods* 2017;14(2):135-139.
Bonfert, T., *et al.* ContextMap 2: fast and accurate context-based RNA-seq mapping. *BMC bioinformatics* 2015;16:122.
Chang, Z., *et al.* Bridger: a new framework for de novo transcriptome assembly using RNA-seq data. *Genome biology* 2015;16:30.
Chen, J., *et al.* ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic acids research* 2009;37(suppl_2):W305-W311.
D'haeseleer, P. How does DNA sequence motif discovery work? *Nature biotechnology* 2006;24(8):959-961.
Dobin, A., *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 2013;29(1):15-21.
Garber, M., *et al.* Computational methods for transcriptome annotation and quantification using RNA-seq. *Nature methods* 2011;8(6):469-477.
Grabherr, M.G., *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature biotechnology* 2011;29(7):644-652.
Kim, D., Langmead, B. and Salzberg, S.L. HISAT: a fast spliced aligner with low memory requirements. *Nature methods* 2015;12(4):357-360.
Kong, Y. Btrim: a fast, lightweight adapter and quality trimming program for next-generation sequencing technologies. *Genomics* 2011;98(2):152-153.
Li, B. and Dewey, C.N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC bioinformatics* 2011;12(1):323.
Li, B., *et al.* RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics* 2009;26(4):493-500.
Li, G., *et al.* QUBIC: a qualitative biclustering algorithm for analyses of gene expression data. *Nucleic acids research* 2009;37(15):e101-e101.
Miller, J.A., *et al.* Improving reliability and absolute quantification of human brain microarray data by filtering and scaling probes using RNA-Seq. *BMC genomics* 2014;15:154.
Nagalakshmi, U., *et al.* The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* 2008;320(5881):1344-1349.
Oshlack, A., Robinson, M.D. and Young, M.D. From RNA-seq reads to differential expression results. *Genome biology* 2010;11(12):220.
Ozsolak, F. and Milos, P.M. RNA sequencing: advances, challenges and opportunities. *Nature reviews. Genetics* 2011;12(2):87.
Pathan, M., *et al.* FunRich: An open access standalone functional enrichment and interaction network analysis tool. *Proteomics* 2015;15(15):2597-2601.
Pertea, M., *et al.* Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nature protocols* 2016;11(9):1650-1667.
Pertea, M., *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature biotechnology* 2015;33(3):290-295.
Philippe, N., *et al.* CRAC: an integrated approach to the analysis of RNA-seq reads. *Genome biology* 2013;14(3):R30.
Pimentel, H., *et al.* Differential analysis of RNA-Seq incorporating quantification uncertainty. *Nature methods* 2017.
Ritchie, M.E., *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic acids research* 2015;43(7):e47.

Robinson, M.D., McCarthy, D.J. and Smyth, G.K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 2010;26(1):139-140.

Subramanian, A., *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences* 2005;102(43):15545-15550.

Swan, M. The quantified self: Fundamental disruption in big data science and biological discovery. *Big Data* 2013;1(2):85-99.

Trapnell, C., *et al.* Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nature biotechnology* 2013;31(1):46-53.

Trapnell, C., Pachter, L. and Salzberg, S.L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 2009;25(9):1105-1111.

Trapnell, C., *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature protocols* 2012;7(3):562-578.

Wang, K., *et al.* MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic acids research* 2010;38(18):e178.

Wang, Z., Gerstein, M. and Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews genetics* 2009;10(1):57-63.

Wu, J., *et al.* OLego: fast and sensitive mapping of spliced mRNA-Seq reads using small seeds. *Nucleic acids research* 2013;41(10):5149-5163.

Wu, T.D., *et al.* GMAP and GSNAP for Genomic Sequence Alignment: Enhancements to Speed, Accuracy, and Functionality. *Methods in molecular biology* 2016;1418:283-334.

Wu, X., *et al.* Data mining with big data. *IEEE transactions on knowledge and data engineering* 2014;26(1):97-107.

Yuan, L., *et al.* GAAP: Genome-organization-framework-Assisted Assembly Pipeline for prokaryotic genomes. *BMC genomics* 2017;18(Suppl 1):952.

Zhou, X. and Su, Z. EasyGO: Gene Ontology-based annotation and functional enrichment analysis tool for agronomical species. *BMC genomics* 2007;8(1):246.