South Dakota State University
Open PRAIRIE: Open Public Research Access Institutional Repository and Information Exchange

Electronic Theses and Dissertations

2018

Statistical Algorithms and Bioinformatics Tools Development for Computational Analysis of High-throughput Transcriptomic Data

Adam McDermaid
South Dakota State University

Follow this and additional works at: https://openprairie.sdstate.edu/etd

 Part of the Biometry Commons, and the Mathematics Commons

Recommended Citation
McDermaid, Adam, "Statistical Algorithms and Bioinformatics Tools Development for Computational Analysis of High-throughput Transcriptomic Data" (2018). *Electronic Theses and Dissertations*. 2645.
https://openprairie.sdstate.edu/etd/2645

This Dissertation - Open Access is brought to you for free and open access by Open PRAIRIE: Open Public Research Access Institutional Repository and Information Exchange. It has been accepted for inclusion in Electronic Theses and Dissertations by an authorized administrator of Open PRAIRIE: Open Public Research Access Institutional Repository and Information Exchange. For more information, please contact michael.biondo@sdstate.edu.

STATISTICAL ALGORITHMS AND BIOINFORMATICS TOOLS DEVELOPMENT FOR COMPUTATIONAL ANALYSIS OF HIGH-THROUGHPUT TRANSCRIPTOMIC DATA

BY

ADAM MCDERMAID

A dissertation submitted in partial fulfillment of the requirements for the

Doctor of Philosophy

Major in Computational Science & Statistics

South Dakota State University

2018

STATISTICAL ALGORITHMS AND BIOINFORMATICS TOOLS DEVELOPMENT FOR COMPUTATIONAL ANALYSIS OF HIGH-THROUGHPUT TRANSCRIPTOMIC DATA

ADAM MCDERMAID

This dissertation is approved as a creditable and independent investigation by a candidate for the Doctor of Philosophy in Computational Science & Statistics degree and is acceptable for meeting the dissertation requirements for this degree. Acceptance of this dissertation does not imply that the conclusions reached by the candidate are necessarily the conclusions of the major department.

Qin Ma, PhD.
Dissertation Advisor Date

Kurt Cogswell, PhD.
Head, Mathematics & Statistics Date

Dean, Graduate School Date

ACKNOWLEDGEMENTS

I would like to thank my research advisors, Dr. Qin Ma and Dr. Anne Fennell, for their continued support throughout my advancement toward this degree. Both have been invaluable in helping me get to this point. Their support and guidance has allowed for me to comfortably convert from a student to competent researcher.

I would also like to thank the members of the Bioinformatics and Mathematical Biosciences Lab, especially Jinyu Yang, Juan Xie, Cankun Wang, Anjun Ma, and Yiran Zhang, as their assistance with numerous aspects throughout the last three years has been very much appreciated.

Finally, I would like to thank my family. Without their support, I would not be where I am today.

CONTENTS

ABSTRACT

STATISTICAL ALGORITHMS AND BIOINFORMATICS TOOLS DEVELOPMENT FOR COMPUTATIONAL ANALYSIS OF HIGH-THROUGHPUT TRANSCRIPTOMIC DATA

ADAM MCDERMAID

2018

Next-Generation Sequencing technologies allow for a substantial increase in the amount of data available for various biological studies. In order to effectively and efficiently analyze this data, computational approaches combining mathematics, statistics, computer science, and biology are implemented. Even with the substantial efforts devoted to development of these approaches, numerous issues and pitfalls remain. One of these issues is mapping uncertainty, in which read alignment results are biased due to the inherent difficulties associated with accurately aligning RNA-Sequencing reads. GeneQC is an alignment quality control tool that provides insight into the severity of mapping uncertainty in each annotated gene from alignment results. GeneQC used feature extraction to identify three levels of information for each gene and implements elastic net regularization and mixture model fitting to provide insight in the severity of mapping uncertainty and the quality of read alignment. In combination with GeneQC, the Ambiguous Reads Mapping (ARM) algorithm works to re-align ambiguous reads through the integration of motif prediction from metabolic pathways to establish co-regulatory gene modules for re-alignment using a negative binomial distribution-based probabilistic approach. These two tools work in tandem to address the issue of mapping

uncertainty and provide more accurate read alignments, and thus more accurate expression estimates.

Also presented in this dissertation are two approaches to interpreting the expression estimates. The first is IRIS-EDA, an integrated shiny web server that combines numerous analyses to investigate gene expression data generated from RNA-Sequencing data. The second is ViDGER, an R/Bioconductor package that quickly generates high-quality visualizations of differential gene expression results to assist users in comprehensive interpretations of their differential gene expression results, which is a non-trivial task. These four presented tools cover a variety of aspects of modern RNA-Seq analyses and aim to address bottlenecks related to algorithmic and computational issues, as well as more efficient and effective implementation methods.

CHAPTER 1: Introduction

1.1 Next-Generation Sequencing and RNA-Sequencing Analysis

The advent of much improved biotechnology and the decreased associated costs have increased the amount of biological data. One of the most modern approaches is Next-Generation Sequencing (NGS) [1, 2], which has higher resolution, better accuracy, lower technical variation, and other advantages, compared with array-based counterparts [3-5]. NGS allows for a much faster-paced generation of larger volumes of biological information than ever before. The generated big data, which refers to the complex and large volumes of data collected from different sources, has changed the way research is conducted in biology [6, 7]. Although the availability of data has increased, utilizing and interpreting it requires new advances in interdisciplinary sciences, namely in mathematics, statistics, and computer science. RNA-sequencing (RNA-Seq) and Chromatin Immunoprecipitation followed by sequencing (ChIP-Seq) have arisen and been used for the interpretation of transcriptional regulation. The RNA-Seq technology measures the abundance of RNA transcripts in samples or individual cells, giving rise to the genome-scale transcriptomic (also termed as *gene expression*) data [8].

ChIP-Seq technologies provide massive amounts of information related to protein-DNA interactions and have been applied successfully to many genome-wide analyses, including transcription factor binding, polymerase binding, and histone modification markers [9, 10]. This type of data is especially useful for determination of transcriptional regulatory signals (TRSs), such as transcription factors (TFs), miRNAs, lncRNAs, and epigenomic regulators. TFs are known to play an important role in

controlling gene expression by binding to specific DNA sequences, with their TF binding sites (TFBSs) are referred to as *cis*-regulatory motifs (motifs for short).

RNA-Seq is a revolutionary technology for gene expression profiling [11, 12] and promises to provide a comprehensive picture of the transcriptome for a biological process [11]. It aims to extract usable information from the mature mRNA within a biological source and generates a huge number of short segments (*reads*, 100-250 bps), which enable the discrete quantification of all genes expressed in a cell [11, 13]. Currently, researchers can analyze a large sample of cells from a single organism in the form of bulk RNA-Seq data or can discover individual cells from complex organisms one at a time through single-cell RNA-Sequencing (scRNA-Seq), which uses optimized NGS technologies and acquires the transcriptomic information from individual cells to provide a better understanding of cell functions at genetic and cellular levels [14]. These biotechnologies have generated large-scale transcriptomic data and genome-scale gene expression data in the public domain, and their tremendous values have been confirmed in many research areas such as elucidation of cell-type-specific regulatory networks [15, 16] and cancer & complex diseases studies [17-19]. Although numerous algorithms and tools have been developed for transcriptomic data analysis, both in the public [20-46] and private sectors [47-54], the reality is that some of the most widely-used methods suffer from particular issues (e.g., cannot provide accurate gene expression estimates [55, 56]) and construction of applicable combinations of these tools is an ongoing challenge.

1.2 Analysis Tools and Pipelines

RNA-Seq analyses begins with data collection from biological samples. During this process, mature mRNA is extracted from single or multiple cells of a particular

sample with specific characteristics. This mRNA is reverse transcribed into cDNA, which is then broken apart into small segments, referred to as reads. These short reads are generally 80 to 250 base pairs (bps) in length. There are also emerging third-generation sequencing technologies that generate reads in the several mbp lengths; although these approaches can suffer from high error rates during sequencing, limiting their current application power [1, 57, 58]. The set of these reads—generally in the range of millions of reads—is referred to as the library of raw reads for analysis in an RNA-Seq experiment.

Analyzing raw reads requires numerous steps, and thus requires numerous tools (Figure 1). To effectively use these tools in combination, a pipeline is generally established with the user's tools of choice. Initially, a read level quality control is conducted on the raw reads. FastQC [20] is almost universally used for this purpose and provides information related to sequencing depth, reads duplication rates, GC bias, coverage uniformity, among other features. Any serious issues detected in this initial process are then corrected through read trimming. This process trims the end segments off the raw reads, which tend to have remnants of the sequencing process. For this purpose, numerous tools have been developed and are widely implemented in application, including Btrim [59], the Fastx toolkit [60], Trimmomatic [61], and Cutadapt [62]. To verify successful data trimming, read-level quality control can be used again.

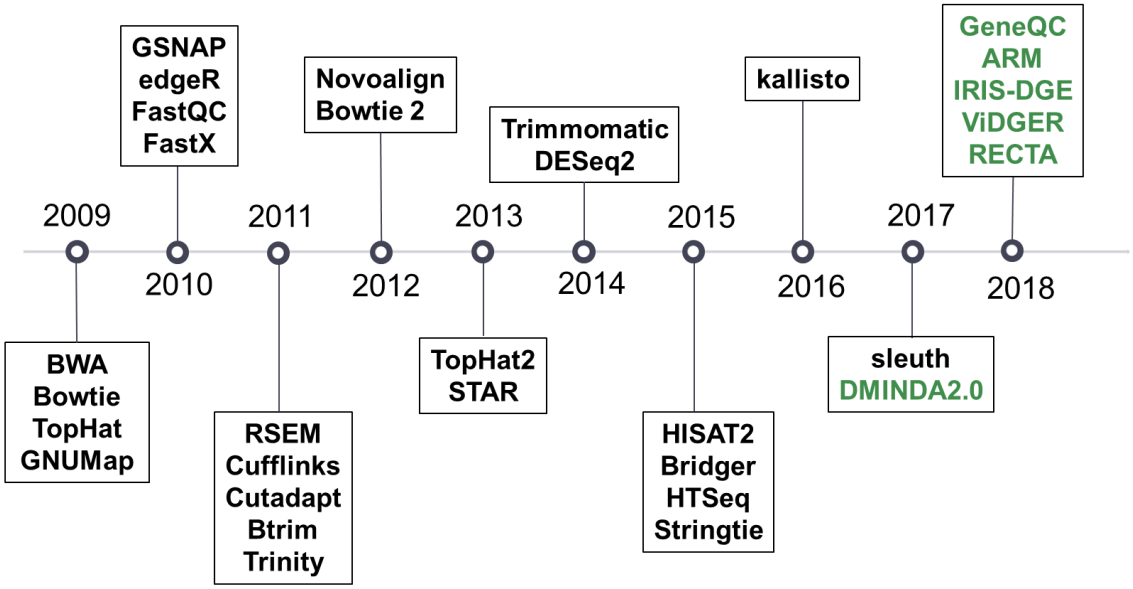


Figure 1: High-performing and widely used RNA-Seq tools developed since 2009. Green lettering indicates tools that are covered in this dissertation, between Chapters 2 & 3.

After verifying the integrity of the raw RNA-Seq data, multiple steps are conducted to quantify the read counts for each gene, which provides insight into the expression level for each gene of each sample. If a reference genome is available for the given species, reference-based read alignment (also referred to as read mapping) of raw or trimmed reads determines where along the genome each read came from. While time consuming and computationally demanding, this step is one of the most important processes used in most RNA-Seq analyses. Due to the importance, numerous tools have been developed for this purpose, including TopHat [35], BWA [63], Bowtie [64, 65], and HISAT [40], among many others [29, 31-33, 37, 42, 44].

Read alignment results still require further analysis to quantify the number of reads estimated at each gene. Two distinct pathways can be pursued at this point. The first is direct quantification of gene expression through read counts. Using a species-specific annotation file, quantification tools take the read alignment results and determine

to which gene each read is aligned. Based on this information, a discrete count of the expression for each gene is generated. Again, there are many tools that can perform this purpose, with HTSeq [45] being one common and efficient method. Alternatively, the second path requires another extensive computational approach in what is referred to as assembly. Assembly tools, such as StringTie [38, 39] and Cufflinks [34], take the aligned reads and assemble transcripts from these segments. The abundance of these transcripts is then quantified, providing an expression estimate. The assembly step is increasingly useful to determine novel transcripts that have not been annotated in a particular species and for addressing the issues presented by alternative splicing. Both of these two approaches result in an estimate of the expression level for each gene.

Having a reference genome for RNA-Seq analysis is not always possible. Some species being analyzed may not have a reference genome sequences at the time of analysis, requiring a different approach. *De novo* assembly is a process that can develops a transcriptome through alignment of the reads themselves. In this process, the reads are taken and assembled together based on overlapping sequences of various lengths. A De Bruijn graph approach is most commonly used for this purpose by most *de novo* assembly tools, such as Trinity [66, 67] and Bridger [43]. The assembly can then be used to functionally annotate the regions within the transcriptome.

Using the expression estimations generated through the reference-based approaches, numerous additional analyses can be performed. One such analysis is differential gene expression analysis, in which gene expression levels are compared between samples of particular conditions. This approach can provide insight into the genetic differences that are affecting or correlated with observed phenotypic differences.

Functional annotation is a process using expression estimates that look for highly expressed functional groups of genes within particular samples. This process can also include comparison of functional group expressions across two or more conditions. Traditional clustering approaches, such as k-means [68] or hierarchical clustering [69], can also be directly applied to expression estimates through grouping of similarly expressed samples. This method can provide insight into which samples or conditions have expression-wide similarities. Biclustering is a two-dimensional clustering approach [70] that, when applied to expression matrices, groups samples together based on subsets of the expression estimates [71]. Since it can be expected that genetic similarities can be exhibited in only a small portion of the expression estimates, this approach captures these similarities and groups sample together, as opposed to requiring high similarity throughout all expression estimates. Particularly, biclustering has the special application power in scRNA-Seq analyses [72, 73]. In addition to these defined approaches, there are virtually endless other analyses that can be performed using the expression estimates, including a wide range of network analyses and other modeling approaches.

Although substantial efforts have been made to accurately and efficiently quantify genetic expression levels, the performance of these tools is not always adequate. Many of the tools have been shown to underperform on real or synthetic RNA-Seq datasets [55, 56]. TopHat [34, 35], one of the most widely used read alignment tools, has even been demonstrated as one of the poorest performing, having less than 20% of reads correctly aligned in some cases [55]. Even combinations of tools that have excellent individual performance can result in suboptimal or even poor performance levels [56]. Hence,

further investigation into optimized approaches for high-throughput data analysis is required.

1.3 IRIS Pipeline Framework

All tools related to RNA-Seq analysis fit into a three-tier framework based both on the placement they fit into and analysis function, referred to as the Integrated RNA-Seq data analysis and Interpretation System (IRIS). This framework consists of tiers representing preprocessing, expression estimation, and end-stage analysis (Figure 2).

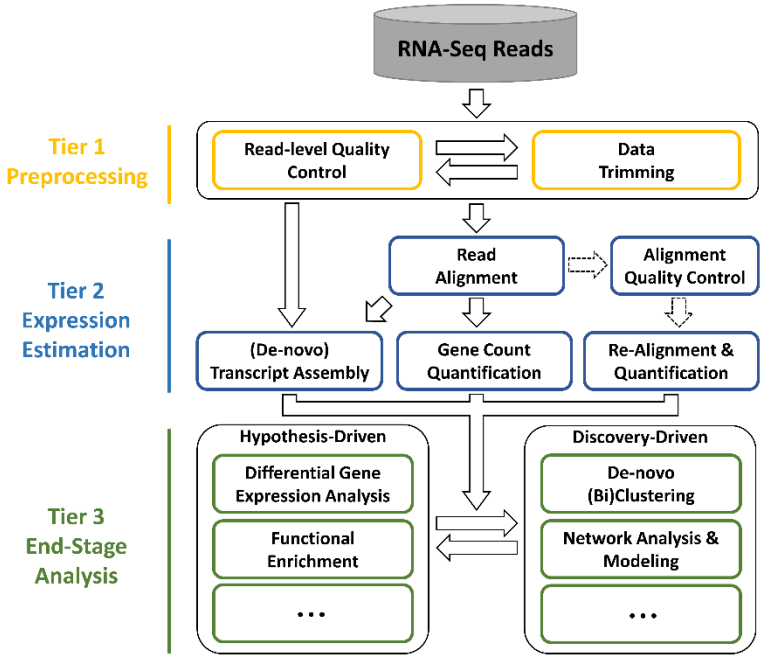


Figure 2: The IRIS Pipeline. The IRIS pipeline consists of three tiers designed to analyze and interpret RNA-Seq data. Tier 1 involves preprocessing, Tier 2 determines expression estimates, and Tier 3 provides end-stage analyses

1.3.1 Preprocessing

Preprocessing consists of tool related to quality control for the raw RNA-Seq reads. There are two analyses in this tier, the first being read-level quality control. This process involves investigation of the raw reads to determine if any abnormalities exist,

including detection of primers used to sequence the raw reads. FastQC [20] is almost universally used for this process, and provides statistics related to per base sequence quality, per sequence quality scores, per base sequence content, per base and per sequence GC content, Kmer content, among other important measures. Users can make decisions about the quality of their raw reads based on the provided information and determine if they need additional measures, such as data trimming. Data trimming involves modification of the raw reads to remove poor sequences and sequence segments, including primers remaining on the ends of reads from previous steps. A wide variety of tools can be utilized for this purpose [28, 60-62]. The results of Tier 1 used for further analysis are either the raw reads—in the case that there are no serious issues found during quality control—or trimmed reads generated using one of the read trimming tools.

1.3.2 Expression Estimation

Using the raw or trimmed reads from Tier 1, Tier 2 contains tools that convert the reads to expression estimates, generally in conjunction with additional genomic information in the form of a reference genome and annotation. This tier is the core of RNA-Seq data analysis and can proceed through multiple unique paths. Which path is pursued is determined by which data is being analyzed, availability of a reference genome, and investigative purposes. If a reference genome is not available, the reference-based approaches are not applicable. In these cases, *De novo* assembly is used and is commonly combined with annotation of sequences to determine which genes are present and to some degree a measure of the expression level.

The alternative pathway, one in which a reference genome is available, involves alignment of reads against the reference genome. This process is generally time

consuming and computationally demanding. Numerous approaches have been developed for this purpose [29, 31-33, 35, 37, 40, 42, 44, 63-65], with key emphasis on reducing the time and computational requirements.

After read alignment, there are another level of pathways that can be followed. A straightforward quantification of read counts based on the read alignment can generate a discrete estimation of the expression level for each annotated gene. Alternatively, reference-based transcript assembly can be used to generate transcripts for quantification. A third approach that is much more recent has to do with mapping uncertainty, which results when a read can be aligned to multiple locations. To address this issue, new approaches have been developed for quality control and read re-alignment.

From these pathways, users generally determine an estimation of the genetic expression levels from their samples. Depending on the tools and methods used, the measurement used for expression level can vary. Some methods generate read counts, with a discrete count of the number of reads aligned to each location is provided. Others provide normalized measures based on the gene length or raw read library size. Four commonly used normalized measures are Reads Per Kilobase per Million (RPKM), Fragments Per Kilobase per Million (FPKM), Transcripts Per kilobase per Million (TPM), and Counts Per Million mapped reads (CPM). RPKM and FPKM are calculated similarly, with the former being used for single-end reads and the latter for paired-end reads. The calculations for normalized counts for a given gene *i* are given below, with *L* representing library size (i.e. number of reads analyzed), g_i representing the length of gene *i*, and c_i representing the number of reads or fragments aligned to gene *i*.

$$FPKM_i = RPKM_i = \frac{c_i}{L/10^6} \div g_i = \frac{c}{L * g_i} * 10^6$$

$$TPM_i = \frac{\frac{c_i}{g_i}}{\sum_j \frac{c_j}{g_j}} \div \left(\frac{L}{10^6}\right) = \frac{10^6 * \frac{c_i}{g_i}}{L * \sum_j \frac{c_j}{g_j}}$$

$$CPM_i = \frac{c_i}{\frac{L}{10^6}} = \frac{c_i}{L} * 10^6$$

Frequently, all of these measures are represented in using a logarithm base-10 transformation, since measures can vary greatly.

1.3.3 End-Stage Analysis

From the expression estimates generated in Tier 2, a wide range of analyses can be performed to make biologically meaningful interpretations from the data. Tier 3 contains analysis tools related to this conversion of expression estimates to practical interpretations and is divided into two categories: Hypothesis-driven interpretations and Discovery-driven interpretations. Hypothesis-driven analyses are generally conducted following previously established hypotheses and concepts. Included in this category are differential gene expression analysis and functional enrichment analysis, among many other processes. Differential gene expression analysis is one of the most common analyses used in the analysis of RNA-Seq data and uses statistical techniques to find meaningful differences in expression levels between comparable conditions. This process uses raw or normalized read counts from replicates of the same condition to identify which genes are statistically differentially expressed between two or more conditions. One common use of this method is to determine which genes have differing expression levels for two different strains of the same species that exhibit important

phenotypic differences. This investigation can lead to further understanding of specific relationships between genotype and phenotype.

Discovery-driven analyses follow a more purely exploratory approach, one aimed at discovering interesting features from the data, as opposed to being directed at a specific hypothesis. Included in this category are clustering and biclustering methods and a wide range of network analyses. Rapidly growing in the analysis of RNA-Seq data is the use of biclustering approaches [70, 74], which isolate similarities between conditions and samples using only a subset of the gene expression estimates. It has been widely shown that most plant and animal life on earth has high genetic similarity due to commonalities in cellular structure and function [75], meaning the genetic differences in a single species, regardless of their phenotypic differences, will be relatively mild. Because of this, clustering samples based on total genetic expression may miss significant expression patterns. While traditional clustering looks for conditions or samples that have similar expression levels across all genes, biclustering can identify similarities that exist in only a fraction of the total genetic expression profile.

While the analyses included in Tier 3 generally represent the end-stage analyses, there are many times overlaps and feedback loops within this stage. For instance, cell type classification of single-cell RNA-Seq data may involve initial clustering or biclustering combined with additional graph modeling to identify which cells belong to the same cell type. This means that an end-stage analysis may not necessarily be the final analysis step in an RNA-Seq pipeline, since end-stage analyses can be layered for a specific purpose. However, most experiments using RNA-Seq data will have a well-defined design relying on direct results from Tier 3.

CHAPTER 2: Algorithms and Tools Development for RNA-Seq Data

While there have been great amounts of effort done towards designing optimized RNA-Seq analysis tools, this area of research is by no means complete. The nature of dealing with big data analysis always means a never-ending striving for increased efficiency, both in terms of the time and computational requirements. Additionally, dealing with data and results that frequently consist of tens-of-thousands of measures of statistical significance and an equal number of measures of magnitude leads to challenges with interpreting results on a global scale. Even more challenging are prominent issues within analysis pipelines that arise from biological complexities, such as the determination of the correct alignment location for a single RNA-Seq read. All of these challenges combined promote the need for continued development of analysis tools for RNA-Seq data. In this chapter, I present four tools develop to address specific pitfalls within RNA-Seq pipelines.

2.1 GeneQC: Gene Expression Estimation Quality Control

2.1.1 Mapping Uncertainty

Even though numerous methods have been developed to facilitate read alignment, 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 (Figure 3), with specific reads being aligned to multiple locations across the reference genome with the same alignment scores [7, 27, 55, 76-78]. When this issue occurs, it results in what is referred to as *mapping uncertainty*.

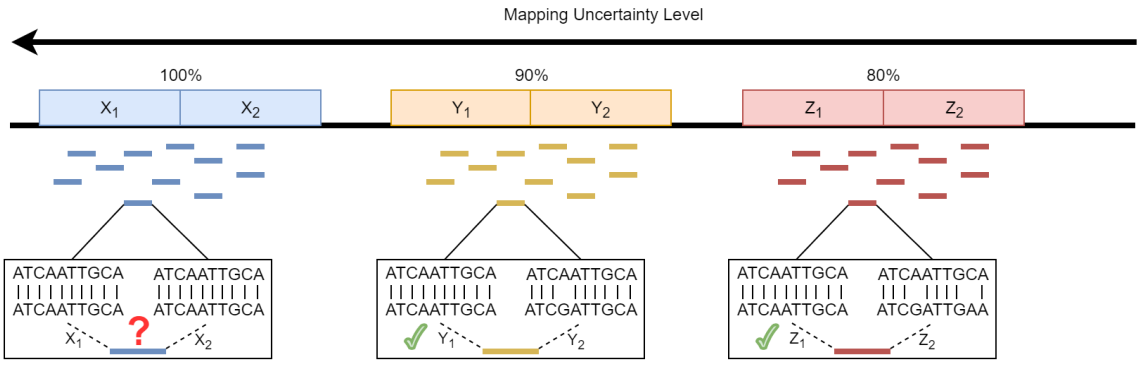


Figure 3: Mapping Uncertainty. Mapping uncertainty occurs when a single read can be mapped to two or more locations along the reference genome with equal or nearly equal confidence.

This mapping uncertainty problem can be observed in any genomic region, including, exons and transcripts. For conciseness, these genomic regions are simply referred to as "genes." This issue has been observed in many diploid species, including human and other mammals and Arabidopsis [79-83], as well as many multiploid species [84]. In some species, such as *Glycine max*, up to 75% of the genes have the duplicated partners in its genome. For species with high levels of uncertainty, especially angiosperms, mapping uncertainty can have serious implications on gene expression levels and can be extremely hard to remediate due to the genes' and chromosomes' duplicative nature [41].

To more fully investigate the issue of mapping uncertainty, 95 datasets totaling almost two terabytes of RNA-Seq data was analyzed from seven plant and animal species with respect to their alignment statistics, including the percentages of uniquely-mapped reads, ambiguously-mapped reads, and non-mapped reads (Table 1). This analysis was done using HISAT2 [40] for read alignment, which automatically generates alignment statistics.

Both paired- and single-end reads were collected from NCBI [85], URGI (https://urgi.versailles.inra.fr/), and JGI [86] 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 83 paired-end datasets and 12 single-end datasets average 20.6 GB, with an average overall alignment rate of 81.87%. Each dataset was aligned using HISAT2 [40] 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 multi-mapped reads over all analyzed datasets, while *Arabidopsis thaliana* displayed the lowest proportion, ranging from 8% to 17%. The other analyzed species had similar percentages of multi-mapped reads.

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 [20] so do gene expression estimates based on mapping results. Even with tools that are specifically designed to address mapping uncertainty, such as *MMR* [87], the quality of the derived gene expression estimates based on mapping results still requires investigation, especially in real datasets not simulated

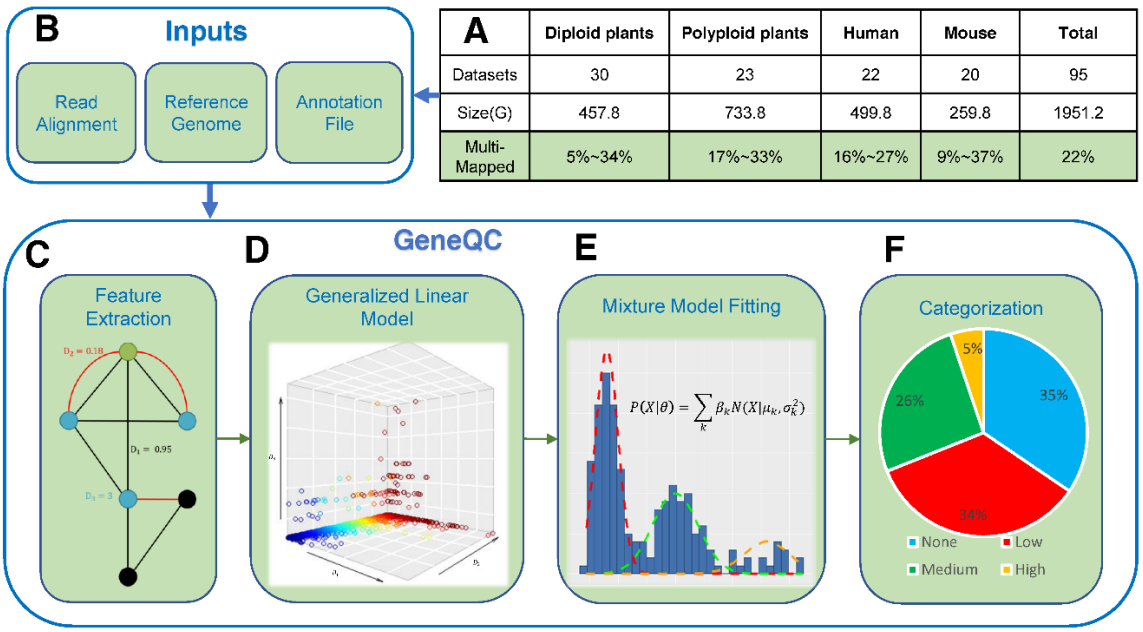
datasets. Without some quality control for gene expression estimation, researchers could potentially be using unreliable data, and blindly doing so.

Table 1: Alignment Statistics. Seven species, five plant and two animal, were aligned using HISAT2. Alignment statistics were collected and presented based on the percent of reads falling into each categorization. Also included are number of datasets and overall data size for each species.

Species	*Arabidopsis thaliana*	*Vitis vinifera*	*Solanum lycopersicum*	*Panicum virgatum*	*Triticum aestivum*	*Homo sapiens* Genome	*Homo sapiens* Transcriptome	*Mus musculus* Genome	*Mus musculus* Transcriptome	**Total**
Datasets	10	10	10	10	13	11	11	10	10	95
Size(GB)	153.7	152.3	151.8	385.7	348.1	249.9	249.9	129.9	129.9	1951
Unique-Mapped	69%~89%	55%~82%	52%~88%	47%~66%	61%~69%	56%-71%	59%~70%	41%~73%	41%~75%	55%
Multi-Mapped	8%~17%	9%~25%	5%~34%	17%~33%	17%~25%	16%-27%	15%-24%	9%~37%	9%~36%	22%
Un-mapped	2%~17%	8%~23%	4%~16%	13%~25%	9%~18%	12%-21%	12%-22%	3%~31%	2%~31%	23%
(Multi-mapped) ÷ (Total mapped)	8%-18%	10%-31%	6%-39%	22%-39%	21%-28%	19%-32%	19%-28%	11%~47%	11%~47%	29%

2.1.2 Methods

To address this issue, I present GeneQC [88] based on novel applications of regularized regression and mixture model fitting approaches to quantify the mapping uncertainty issue (Figure 4). This tool can determine the genes having reliable expression estimates and those requiring further analysis, along with a statistical 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 mapping uncertainty problem; (ii) the proportion of shared multi-mapped reads 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).



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 4: GeneQC Workflow. (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.

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. 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/gtf/gff3) file (Figure 4B).

From input information, GeneQC first performs feature extraction, in which the three characteristics are calculated for each annotated gene (Figure 4C). The first

extracted feature (D_1) is derived from genomic level information and involves the similarity between two genes (Figure 5A). 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,

$$D_1 = \max_{y}\{ss_{i,y} * l_{i,y}\}$$

where $ss_{i,y}$ is the base pair sequence similarity of gene i and gene y and $l_{i,y}$ is the match length of these two genes. Additionally, to minimize negligible interactions, some default criteria are required for determination of D_1: (1) $ss_{i,y} * l_{i,y} > 100$; (2) $mismatch\ count < 5$; (3) $\max\{gap\} < 5$; and (4) $e - value < 10^{-6}$ as determined in using BLAST [89].

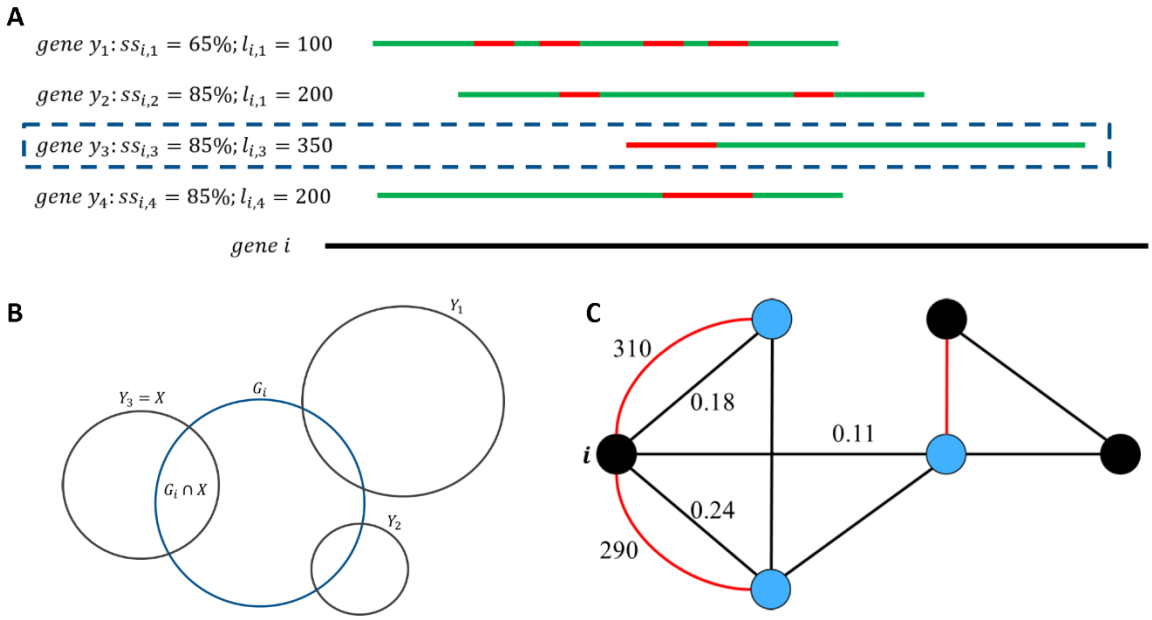


Figure 5: (A) Genes with significant similarity are displayed, with D_1 being the maximum value of $ss_{i,y} * l_{i,y}$. In this situation, genes $y_2, y_3, \& y_4$ all have the same ss_i value, but gene y_3 has a longer consecutive string of matching base pairs (l_i) than the other values, making it 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 $D_3 = log_{10}(3 + 1) = 0.602$.

The second feature (D_2) comes from transcriptomic level information and represents the proportion of shared MMRs (Figure 5B). This value is calculated as the maximum 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 $X = \underset{Y}{\operatorname{argmax}} |G_i \cap Y|$.

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 (Figure 5C) and is calculated as

$$D_3 = \log_{10}(|S \cup M| + 1)$$

where $S = \{genomic\ locations\ with\ D_1 > 0\}$ and $M = \{genomic\ locations\ with\ D_2 > 0\}$.

To perform the modeling, a dependent variable is constructed. The dependent variable *D4* is an approximation of the proportion of ambiguous reads based on the two most extreme approaches to dealing with multi-mapped reads, the unique alignment approach and the all-matches approach. If we consider $G_i = \{reads\ mapped\ to\ gene\ i\}$ and $U_i = \{reads\ uniquely\ mapped\ to\ gene\ i\}$, the true alignment 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 $\left|\hat{R}_i\right| = \frac{|G_i|+|U_i|}{2}$. Using this approximation,

$$D_4 = 1 - \frac{\left|\hat{R}_i\right|}{|G_i|} = 1 - \frac{|G_i| + |U_i|}{2|G_i|}$$

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 [90]. Because of this, alternative approaches were explored. Ridge regression, which develops a model based on an *L2-norm* penalization, has better predictive results than ordinary least squares regression [90, 91]. However, this approach tends to retain all included variables to achieve such high predictive power, in turn reducing the interpretability of the model [92]. Another approach with potential application in GeneQC is the least absolute shrinkage and selection operator, also known as lasso. This method uses an *L1-norm* penalization, while simultaneously performing continuous shrinkage and variable selection [93]. While this is an appealing feature in generating a model, lasso has shortcomings when it comes to dealing with variables exhibiting high pairwise correlation [92]. Elastic-net regularization—sometimes referred to simply as elastic net—has the 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 $\boldsymbol{y} = (y_1, y_2, \dots, y_n)^T$, a set of *p* predictor variables $\boldsymbol{x_i} = \left(x_{i,1}, x_{i,2}, \dots, x_{i,p}\right), i \in \{1, \dots, n\}$, a set of *p* coefficients $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_p)$, and matrix of predictor variables

$$\boldsymbol{X} = (\boldsymbol{x_1}, \boldsymbol{x_2}, \dots, \boldsymbol{x_n})^T = \begin{pmatrix} x_{1,1} & \cdots & x_{1,p} \\ \vdots & \ddots & \vdots \\ x_{n,1} & \cdots & x_{n,p} \end{pmatrix}$$

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

$$L(\lambda_1, \lambda_2, \boldsymbol{\beta}) = \|\boldsymbol{y} - \boldsymbol{X\beta}\|_2^2 + \lambda_2\|\boldsymbol{\beta}\|_2^2 + \lambda_1\|\boldsymbol{\beta}\|_1$$

$$\|\boldsymbol{\beta}\|_2 = \sqrt{\sum_{j=1}^{p} \beta_j}$$

$$\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{p} |\beta_j|$$

Thus, the set of coefficient estimates $\widehat{\boldsymbol{\beta}}$ are calculated as

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\{L(\lambda_1, \lambda_2, \boldsymbol{\beta})\} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\{\|\boldsymbol{y} - \boldsymbol{X\beta}\|_2^2 + \lambda_2\|\boldsymbol{\beta}\|_2^2 + \lambda_1\|\boldsymbol{\beta}\|_1\}$$

Given $\alpha = \frac{\lambda_1}{\lambda_1 + \lambda_2}$, solving for $\widehat{\boldsymbol{\beta}}$ is equivalent to optimizing $\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\|\boldsymbol{y} - \boldsymbol{X\beta}\|_2^2$, for $\alpha\|\boldsymbol{\beta}\|_2^2 + (1-\alpha)\|\boldsymbol{\beta}\|_1 \leq k, for\ some\ k$. In the construction of this elastic net, $\alpha\|\boldsymbol{\beta}\|_2^2 + (1-\alpha)\|\boldsymbol{\beta}\|_1$ 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 [92].

GeneQC utilizes the elastic-net regularization method [92] 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 4D). 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 [92]. 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.

Based on the calculated sets of D-scores through above investigations during GeneQC development, there are apparent 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 mixture model distribution with three sub-distributions to each set of D-scores (Figure 4E).

GeneQC's mixture model fitting process involves *k*-means initialization with randomized initial grouping. Cluster means, μ_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:

$$\underset{K}{\operatorname{argmin}} \sum_{i=1}^{k} \sum_{x \in K_i} \|x - \mu_i\|^2$$

After initialization using the *k-means* process defined above, the *EM-algorithm* is implemented to find the best fitting distributions. Based on the preliminary investigations into the D-score development, two underlying distributions were selected 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 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. Considering the Gaussian distribution scenario, $Y_k(X|\theta_k)$ is $N(X|\mu_k, \sigma_k^2)$. In this case,

$$MLE(\mu_k) = \hat{\mu}_k = \frac{\sum_j^{N_k} x_{j,k}}{N_k}$$

$$MLE(\sigma_k^2) = \hat{\sigma}_k^2 = \frac{\sum_j^{N_k} (x_{j,k} - \mu_k)^2}{N_k}$$

$$\beta_k = \frac{N_k}{N}$$

where $x_{j,k}$ is the j^{th} data point in component k, N_k is the number of data points in cluster k and N is the total number of data points (i.e. $\sum_k N_k = N$). After this initialization step, the algorithm proceeds to the Expectation (E) step. In this step, for each data point (i.e. each D-score from this dataset) the posterior probability of containment within each cluster k_i is generated by

$$P(x_j \in k_i | x_j) = \frac{P(x_j | x_j \in k_i)\, P(k_i)}{P(x_j)} = \frac{N(x_j | \hat{\mu}_k, \hat{\sigma}_k^2)\left(\frac{N_k}{N}\right)}{\sum_k \beta_k N(x_j | \hat{\mu}_k, \hat{\sigma}_k^2)} = \frac{\beta_k N(x_j | \hat{\mu}_k, \hat{\sigma}_k^2)}{\sum_k \beta_k N(x_j | \hat{\mu}_k, \hat{\sigma}_k^2)}$$

After this Expectation step, the Maximization step again calculates parameters $\hat{\mu}_k, \hat{\sigma}_k^2$ for each component k. Based on the previous step,

$$\hat{\mu}_k = \frac{\sum_{j=1}^{N} P(x_j \in k_i | x_j) x_j}{\sum_{j=1}^{N} P(x_j \in k_i | x_j)}$$

$$\hat{\sigma}_k^2 = \frac{\sum_{j=1}^{N} P(x_j \in k_i | x_j)(x_j - \hat{\mu}_k)^2}{\sum_{j=1}^{N} P(x_j \in k_i | x_j)}$$

$$\beta_k = \frac{\sum_{j=1}^{N} P(x_j \in k_i | x_j)}{N}$$

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 $\hat{L}$.

$$BIC(K) = 2klog(n) - 2\hat{L}$$

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 (Fig 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

$$x = -\left(\frac{\mu_{i+1}\sigma_i^2 - \mu_i\sigma_{i+1}^2}{\sigma_{i+1}^2 - \sigma_i^2}\right) \pm \sqrt{\left(\frac{2\sigma_i^2\sigma_{i+1}^2 \cdot \ln\left(\frac{\sigma_{i+1}^2}{\sigma_i^2}\right) - \mu_i^2\sigma_{i+1}^2 + \mu_{i+1}^2\sigma_i^2}{\sigma_{i+1}^2 - \sigma_i^2}\right) + \left(\frac{\mu_{i+1}\sigma_i^2 - \mu_i\sigma_{i+1}^2}{\sigma_{i+1}^2 - \sigma_i^2}\right)^2}$$

for $i \in \{1,2\}$.

For Gamma distributions providing the minimal BIC, a closed form solution of the density function intersections does not exist. To accommodate this, an estimation approach is utilized. The cutoffs are calculated as the mean value of the maximum sequence element for which sub-distribution i has a higher probability density value than it does for sub-distribution $i+1$ and the minimum sequence element for which sub-distribution $i+1$ has a higher probability density value than it does for sub-distribution i, i.e.

$$mean\left(\underset{x}{\text{argmax}}\{f_i(x) > f_{i+1}(x)\}, \underset{x}{\text{argmin}}\{f_i(x) < f_{i+1}(x)\}\right)$$

$$x \in \{a_n |\ \underset{x}{\text{argmax}}\, f_i(x) \leq a_n \leq a_{n+1} \leq \underset{x}{\text{argmax}}\, f_{i+1}(x)\}$$

resulting in two cutoff values.

Due to the nature of mapping uncertainty and the lack of current approaches to evaluate this concept, GeneQC also calculates and provides an alternative likelihood value, as a proposed method of evaluating the mapping uncertainty categorizations

computationally. This value based on the posterior probabilities of the other distributions and 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.

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).

Table 2: GeneQC Example Output. The output of GeneQC from a *Vitis vinifera* sample, providing the extracted features, calculated D-score, mapping uncertainty categorization, and alternative likelihood value.

Gene ID	D1	D2	D3	D-score	Category	Alternative Likelihood
gene17958	1439.981	0.022727	1.041393	0.022765	Low	0.106445
gene29138	228	1	0.69897	0.509935	High	0.012702
gene17991	2560	1	0.477121	0.498094	High	0.015754
gene24080	321.9987	0.005017	2.060698	0.020863	Low	0.10397
gene23209	365	0.0224	1.78533	0.027916	Low	0.113361
gene420	157	0.04878	0.954243	0.033132	Low	0.120682
gene15973	691.9874	0.7809523	0.47712125	0.39143804	Medium	2.15E-54
gene24933	855	1	0.477121	0.499807	High	0.015276
gene26458	4864	1	0.477121	0.495779	High	0.016419

2.1.3 Application on Real Data

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 6). 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. 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.

Table 3: GeneQC Analysis of Seven Species. *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.*

Species	Mean D_1	Mean D_2	Mean D_3	Mean D-score
A. thaliana	0.02	0.58	0.01	0.29
V. vinifera	0.04	0.46	0.16	0.24
S. lycopersicum	0.06	0.66	0.04	0.33
P. virgatum	0.01	0.32	0.09	0.16
T. aestivum	0.02	0.60	0.15	0.31
H. sapiens	0.05	0.84	0.32	0.43
M. musculus	0.06	0.84	0.28	0.42

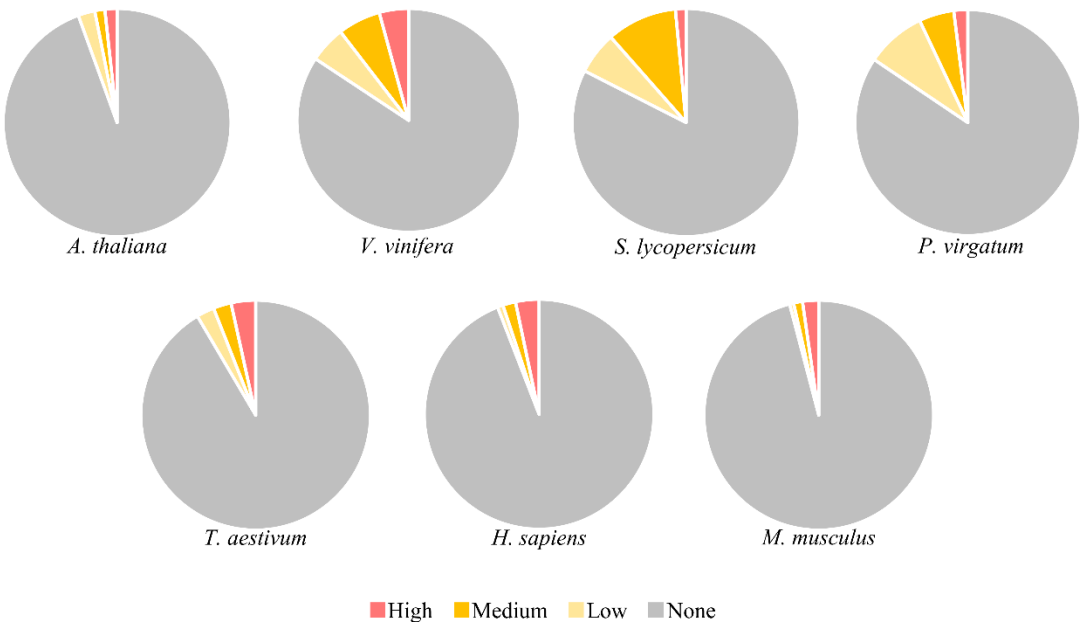


Figure 6: The categorization results related to the analysis of seven datasets representing five plant and two animal species indicating level of mapping uncertainty per gene are shown relative to all categorizations.

The analysis results for the three features and calculated D-scores for genes with some level of mapping uncertainty are displayed in Figures 7 and 8, 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.

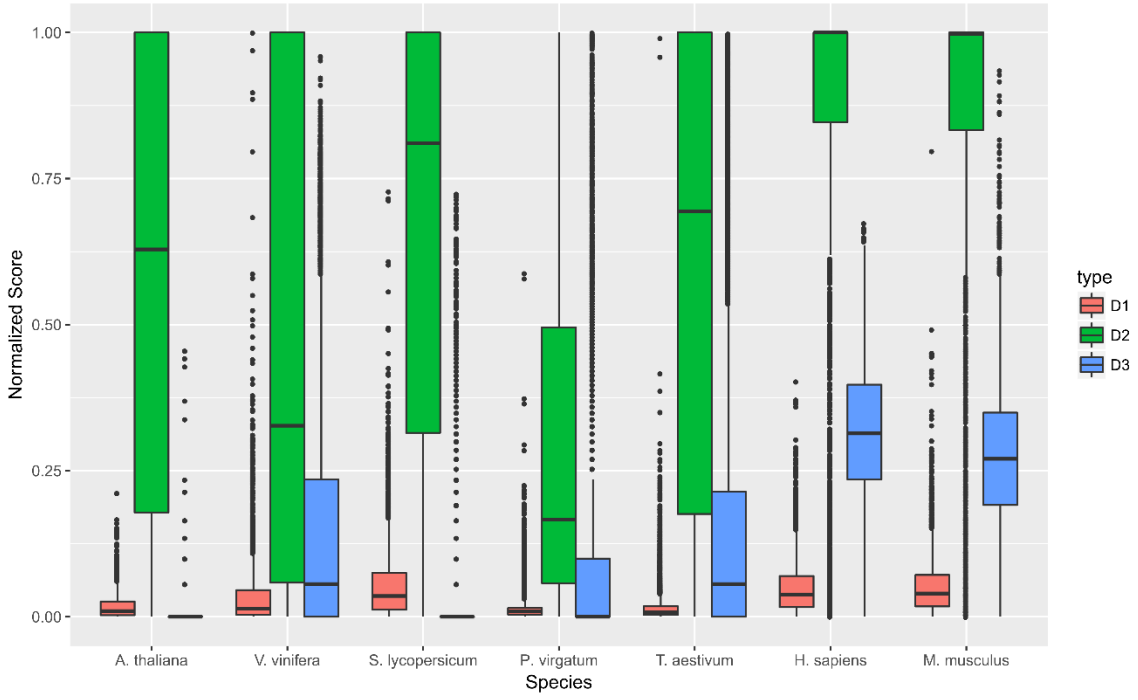


Figure 7: Boxplots results of the seven analyzed species using GeneQC for the three extracted features of each gene. D_1, D_2, and D_3 represent the sequence similarity, proportion of shared MMR, and degree weight, respectively. Each value is shown normalized between 0 and 1. Only genes with mapping uncertainty are displayed.

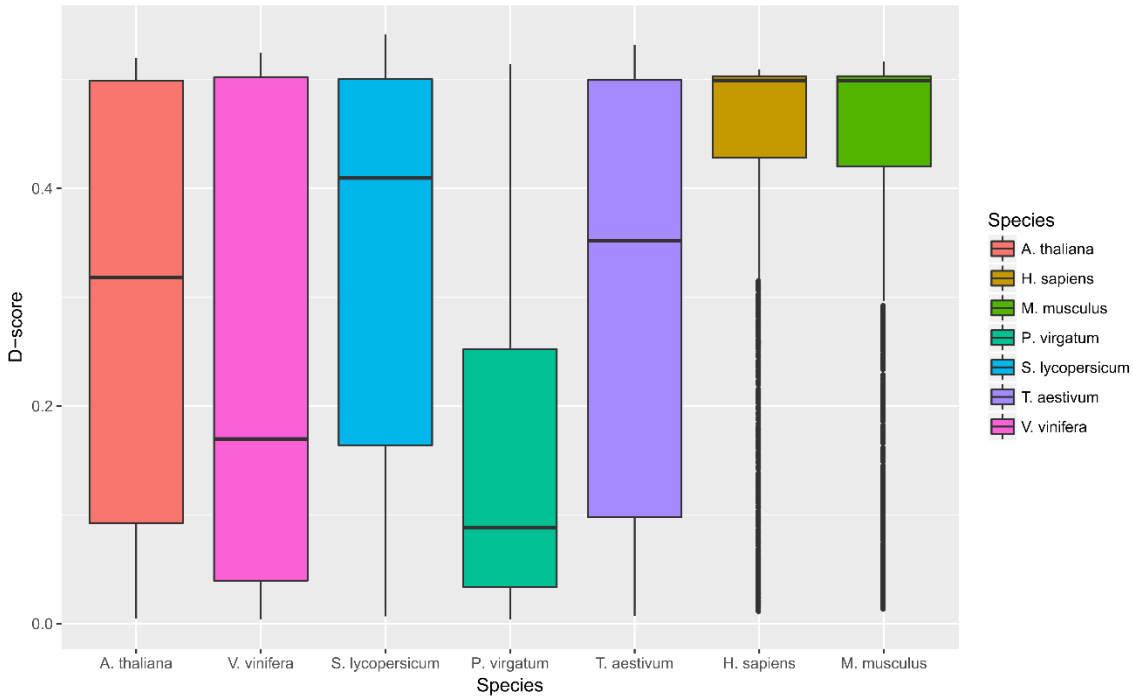


Figure 8: Derived D-scores for each gene are shown by species for each of the seven analyzed datasets, as calculated from the three features in Figure 7. Higher D-scores represent higher levels of mapping uncertainty.

2.1.4 Summary

GeneQC is a tool used to investigate the prominent issue of mapping uncertainty in modern RNA-Seq analysis. 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.

Not only does GeneQC provide a method for analyzing the severity of mapping uncertainty in analyzed data, it also enables researchers to directly compare the expression estimates generated by various alignment tools using real world data. While current comparisons rely on large-scale simulated data—which fails to accurately capture the biological complexities of real RNA-Seq data—or small-scale real data using qPCR or the limited validated gene sets, GeneQC allows for any type of real data to be used to directly compare alignment strategies through the use of D-scores and categorization percentages.

2.2 ARM: Ambiguous Read Mapping Algorithm

While GeneQC provides a direct framework to determine the severity of mapping uncertainty and reliability of expression estimates, addressing these issues involves application of a different approach. Current alignment tools mainly consider local information in the context of the reads and reference genomes. While the strategies implemented by these tools are of high quality relative to the information used, they are still not suitable to provide optimal alignment results, since there are still serious issues related to the reliability of alignment results as demonstrated in Section 2.1. One approach that could rectify this issue is to consider a wider scope of information. In particular, using pathway and regulatory information can provide a new level of information to consider when aligning reads.

Transcription factors are proteins that bind to specific DNA sequences and play important roles in controlling the expression levels of their target genes. *Cis*-regulatory motifs are short, conserved segments of DNA and are typically binding sites for these transcription factors [94]. These binding sites play significant roles in regulating the rate of transcription for nearby genes. Hence, prediction of transcription factor binding sites provides a solid foundation for inferring gene regulatory mechanisms and building regulatory networks for a genome [95-98].

In order to determine more accurate expression estimations, I present an algorithm for ambiguous reads mapping (ARM). ARM integrates information in the form of metabolic pathways, regulatory networks, alignment locations, and reads counts to provide negative binomial distribution-based re-alignment leading to more accurate expression estimates from RNA-Seq data (Figure 9).

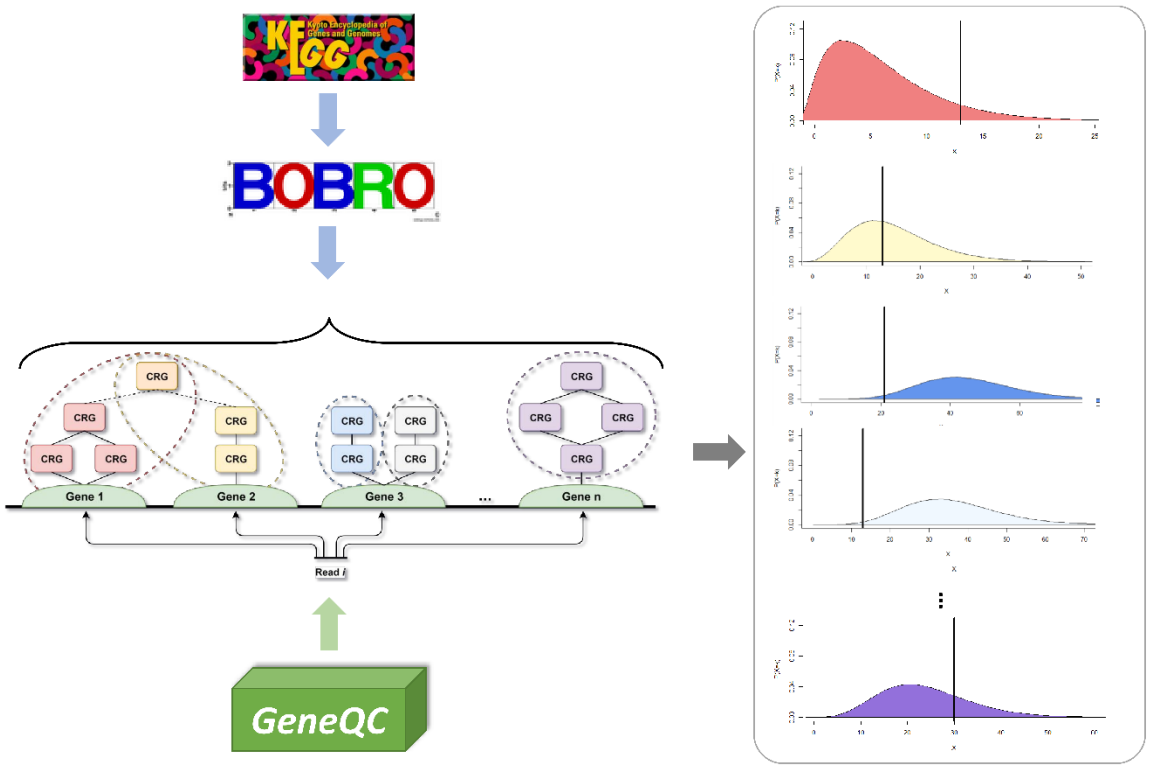


Figure 9: ARM Algorithm Framework. KEGG Pathways are analyzed using the BOBRO motif prediction tool to develop networks of co-regulated genes (CRGs). Simultaneously, GeneQC extracts information related to potential alignment locations for each read, along with unambiguous read counts. The unambiguous read counts are used along with proportional ambiguous read counts for each CRG network to generate a negative binomial-based distribution for each potential alignment location. Based on the current read count of the potential gene location, a probabilistic alignment for each ambiguous read is determined.

2.2.1 Methods

ARM relies on key pieces of information from multiple sources to determine a sounder alignment of ambiguous reads. First, ambiguous reads are determined through GeneQC as any reads belonging to genes with particular levels of mapping uncertainty. By default, any reads aligned to genes falling into the “High” or “Medium” mapping uncertainty categorizations are considered ambiguous reads; although, reads from genes falling into the “Low” categorization could be considered also.

In addition to the qualification of ambiguous reads, GeneQC also provides information related to potential alignment locations and read counts. For each ambiguous read, a modified version of GeneQC provides a list of potential alignment locations based on the initial alignment results. Furthermore, GeneQC extracts ambiguous and unambiguous read counts for each potential alignment location. The unambiguous read counts are calculated as the total number of reads that are uniquely mapped to that particular location, while the unambiguous read counts are the total number of reads that are mapped to that location but could be mapped to another location.

Co-regulatory networks are determined by integration of pathway information and motif prediction. First, KEGG metabolic pathways [99] are collected for the specific species of interest. Each of these pathways are separately analyzed using DMINDA2.0 [100] with the backend algorithm being BOBRO [101] for motif prediction. The genes that are regulated or targeted by these predicted motifs create a single co-regulatory network, as co-regulated gene modules tend to have more similar expression patterns; hence, these modules can be used to train the re-alignment model.

For each ambiguous read, the potential alignment locations are isolated with their corresponding co-regulatory networks to develop a series of distributions. Read count distributions have widely been understood to follow negative binomial distributions [34, 36, 102, 103]. Following this framework, the distribution for read counts of gene *j* can then be represented using a negative binomial distribution denoted as $X_j \sim NB(r, p)$, following the probability mass function of

$$P(X = k) = \binom{k + r - 1}{k} p^k (1 - p)^r$$

This formulation represents the probability of achieving the k^{th} success on the $r + k = n^{th}$ attempt, with the independent probability of a success being p. While this does not have direct applicability or interpretability within the scope of read counts, a conversion can shed more light. The expected value and variance of the read count of gene *j* are respectively calculated as $\mu_j = E(X_j) = \frac{pr}{1-p}$ and $\sigma_j^2 = Var(X_j) = \frac{pr}{(1-p)^2}$ [104]. Thus, with some basic algebra, we obtain the following:

$$\mu_j = \frac{pr}{1-p} \rightarrow (1-p)\mu_j = \mu_j - p\mu_j = pr \rightarrow \mu_j = pr + p\mu_j = p(r + \mu_j)$$

$$\rightarrow p = \frac{\mu_j}{r + \mu_j}$$

$$\rightarrow 1 - p = 1 - \frac{\mu_j}{r + \mu_j} = \frac{r}{r + \mu_j}$$

Using this information, an alternative formulation of the probability mass function can be derived as:

$$P(X = k) = \binom{k + r - 1}{k} p^k (1-p)^r = \binom{k + r - 1}{k} \left(\frac{\mu_j}{r + \mu_j}\right)^k \left(\frac{r}{r + \mu_j}\right)^r$$

$$= \frac{(k + r - 1)!}{k!\,(r-1)!} \left(\frac{\mu_j}{r + \mu_j}\right)^k \left(\frac{r + \mu_j}{r}\right)^{-r} = \frac{\Gamma(k + r)}{k!\,\Gamma(r)} \left(\frac{\mu_j}{r + \mu_j}\right)^k \left(1 + \frac{\mu_j}{r}\right)^{-r}$$

where Γ is the gamma function defined as

$$\Gamma(y) = \int_0^\infty x^{y-1} e^{-x} dx$$

This value is equivalent to $(y-1)!$ when y is a positive integer.

From this formula, we can estimate μ_j using $\hat{\mu}_j = \bar{x}$ and r using $\hat{r} = \frac{\bar{x}^2}{s^2 - \bar{x}}$, where $\bar{x}$ is the sample mean and s^2 is the sample variance [105]. With this estimation, we can represent the probability mass function of read counts as

$$P(X = k) = \binom{k + \hat{r} - 1}{k} \left(\frac{\hat{\mu}_j}{\hat{r} + \hat{\mu}_j}\right)^k \left(\frac{\hat{r}}{\hat{r} + \hat{\mu}_j}\right)^{\hat{r}} =$$

$$\binom{k + \left(\frac{\bar{x}^2}{s^2 - \bar{x}}\right) - 1}{k} \left(\frac{\bar{x}}{\frac{\bar{x}^2}{s^2 - \bar{x}} + \bar{x}}\right)^k \left(\frac{\frac{\bar{x}^2}{s^2 - \bar{x}}}{\frac{\bar{x}^2}{s^2 - \bar{x}} + \bar{x}}\right)^{\frac{\bar{x}^2}{s^2 - \bar{x}}}$$

Using this distribution framework, ARM calculates the sample mean $\bar{x}$ and sample variance s^2 for each co-regulatory network.

For a given read *i*, a set of *n* potential alignment locations is provided through GeneQC. Each of the *n* potential locations has a co-regulatory network with a calculated $\bar{x}$ and s^2. ARM calculates the alignment value of read *i* to gene location *j* as

$$A_{i,j} = P(X \leq \bar{x}) - P(X \leq k_j + 1)$$

where $k_j = u_j + round(c_j a_j)$, with u_j representing the unique read count, a_j representing the ambiguous read counts, $c_j = \max\{0, 1 - 2D_j\}$ representing the ambiguous count weighting factor, and D_j representing the D-score calculated using GeneQC. The weighting factor is used to give partial credit for ambiguously aligned reads for genes that have relatively low D-scores. Genes with high D-scores—those close to 0.5—will be given little to no credit for ambiguously aligned reads. Read *i* will then be aligned to the location with the highest alignment value. Based on this

alignment, the unique reads count u_j for potential location *j* is updated. This process will be repeated for each ambiguous read.

2.2.2 Application on Real Data

In order to investigate the effectiveness of the ARM algorithm on re-alignment of ambiguous reads, GeneQC was used. In particular, the pre- and post-ARM D-scores were evaluated for *Vitis vinifera*, *Arabidopsis thaliana*, *Homo sapiens*, and *Mus musculus* to determine if ARM had any appreciable or statistical effect on mapping uncertainty. D-scores for each gene with some level of mapping uncertainty were calculated based on the re-alignment using the ARM algorithm. Since D_1 represents sequence similarity that would not change with re-alignment, only D_2 and D_3 values changed. The same model used to determine D-scores for the initial alignment was used to reflect an accurate change in the alignment quality. D-score distributions for genes with original non-zero D-scores are shown in Figure 10.

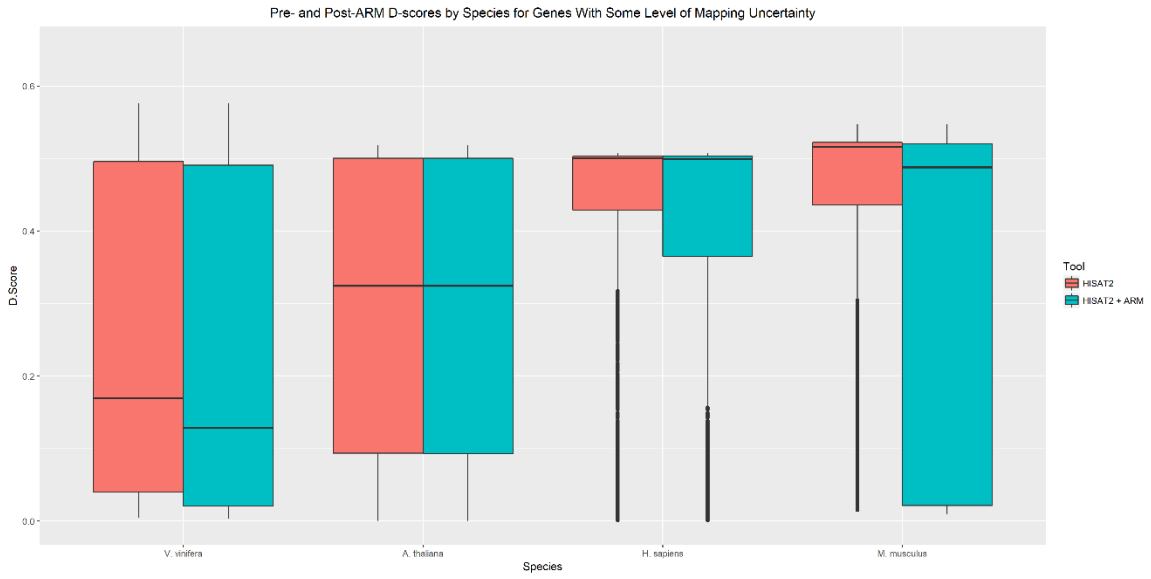


Figure 10: D-scores for the Pre- and Post-ARM algorithm for *V. vinifera*, *A. thaliana*, *H. sapiens*, and *M. musculus*. Genes included in the generation of this figure had Pre-ARM D-scores greater than zero, indicating some level of mapping uncertainty existing after initial alignment.

Based on Figure 10, the effect of ARM on D-scores appears to be relatively minor overall. To more rigorously evaluate the effectiveness of the ARM algorithm, a paired Wilcoxon signed-rank test was used. This test acts as a nonparametric version of a paired t-test to determine if there is a difference in the pre- and post-ARM D-score pairings. A significance level of $\alpha = 0.10$ was chosen to determine if significant improvements are observed. This analysis generated $p - value < 2.2e^{-16}$ for *V. vinifera*, *H. sapiens*, and *M. musculus* samples, thus indicating a statistically significant difference in D-scores due to the ARM algorithm. For *A. thaliana*, the generated p-value is 0.0596. Based on this, it is safe to conclude that the ARM re-alignment algorithm significantly improves D-scores. Figure 11 displays the percent of genes that observed improvements in D-score through the use of the ARM algorithm. Overall, *V. vinifera* saw an improvement in D-scores for 2.08%, *A. thaliana* saw an improvement in 0.02%, *H. sapiens* saw an improvement in 0.35%, and *M. musculus* saw an improvement in 1.06%. However, since

the ARM algorithm is specifically for re-alignment of ambiguous reads, it is more appropriate to view the performance of ARM relative to only the genes with some level of mapping uncertainty (i.e. D > 0). Based on these metrics, an improvement in 13.25%, 0.33%, 5.93%, and 25.93% of genes for *V. vinifera*, *A. thaliana*, *H. sapiens*, and *M. musculus* was observed, indicating a relatively large proportion of improvement. The ARM algorithm appears to be less effect for the *A. thaliana* sample than the others, which is most likely due to the relatively limited network information generated through motif prediction.

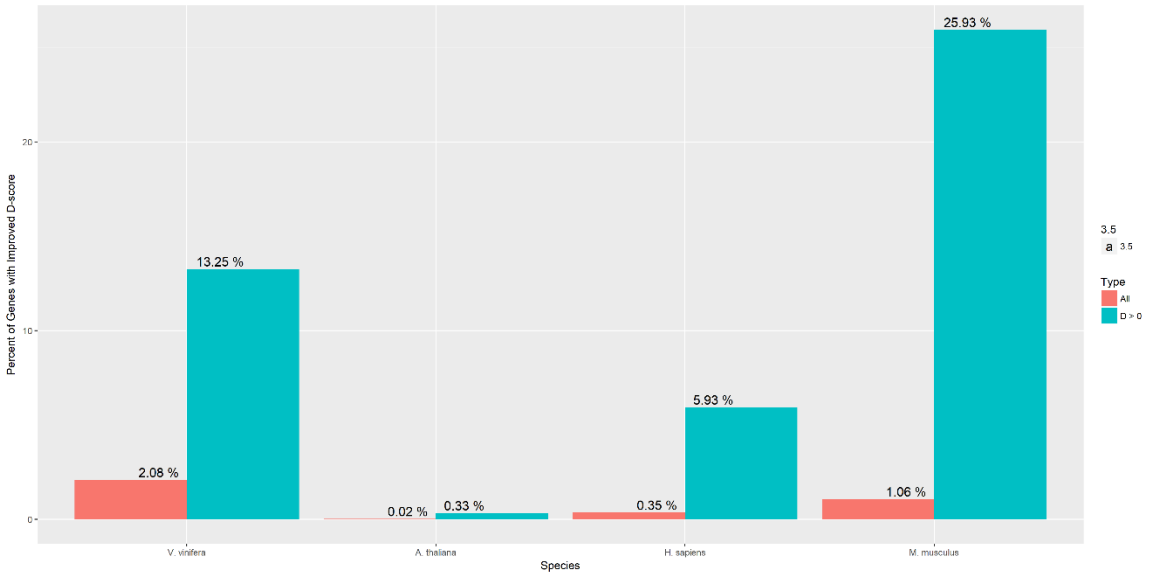


Figure 11: Percent of Genes with Improved D-scores by species. The percent of genes that observed an improved D-score through the ARM algorithm are displayed here. The red bar indicates the percentage relative to all genes, while the blue bar is with respect to the genes that had some level of mapping uncertainty to begin with (i.e. D > 0).

Additionally, of some importance is the degree to which the D-scores changed. If D-scores improved for 25% of *M. musculus* genes but that change was very minor, the impact of the ARM algorithm could be questioned. To methods were used to determine

the magnitude of impact, mean percent change and percent of genes that changed mapping uncertainty categorization as a result of the ARM algorithm. Figure 12 displays the mean percent change of D-score for the four species. Overall, the mean change for *V. vinifera*, *A. thaliana*, *H. sapiens*, and *M. musculus* are 9.77%, 0.28%, 5.61%, and 24.42%, respectively. When considering the mean percent change only for the genes that exhibited some change in D-score as a result of the ARM algorithm, these numbers increased to 75.64%, 85.4%, 95.47%, and 94.18%, respectively. Again, *A. thaliana* has a lower overall metric than the other species, which is potentially due to the limited network information. This theory is supported by the similar mean percent difference when considering only genes that showed some difference in post-ARM D-score.

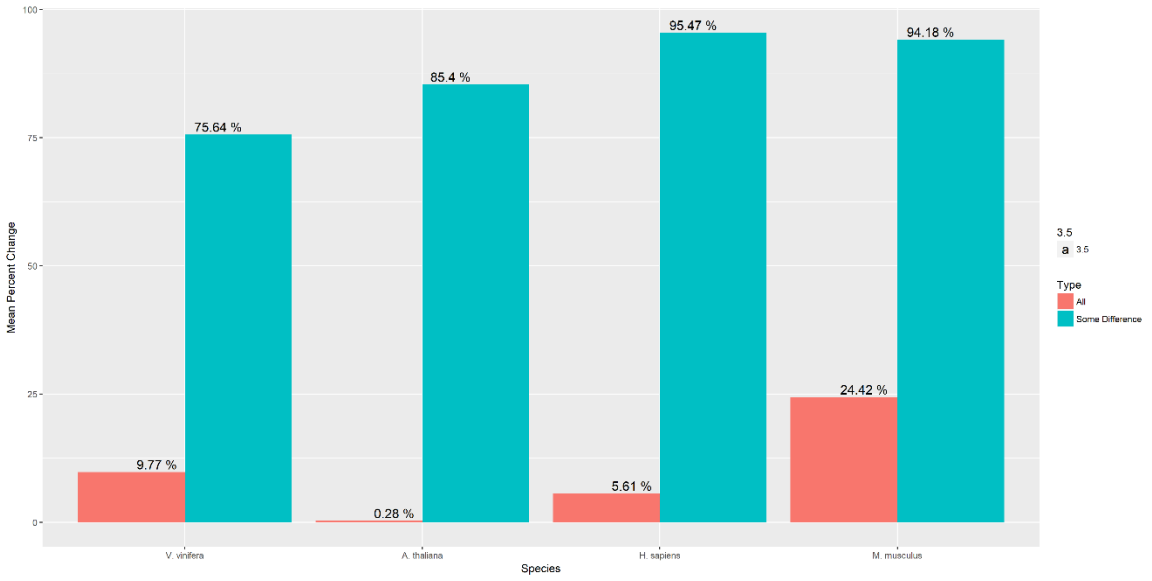


Figure 12: Mean percent change in D-score by species. The red bar indicates percent change overall genes, while the blue indicates mean percent change for genes that exhibited some change in D-score.

Impact for the ARM algorithm can also be observed through the percent of genes that changed mapping uncertainty categorization (Figure 13). 8.79% of *V. vinifera* genes,

0.07% of *A. thaliana* genes, 5.6% of *H. sapiens* genes, and 22.47% of *M. musculus* genes saw a change in mapping uncertainty categorization, all of which was a reduction in the categorization level. Similarly with the other metric, *A. thaliana* saw a relatively low result. This only strengthens the need for further investigation of network generation methods, as discussed in Chapter 4.

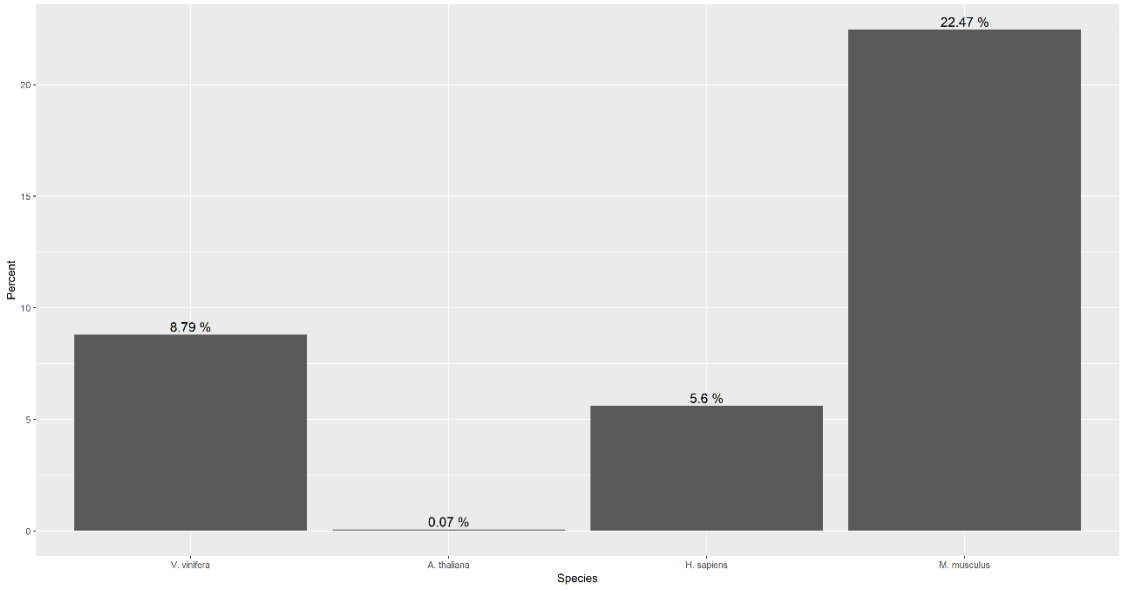


Figure 13: Percent of genes that changed mapping uncertainty categorization as a result of ARM re-alignment by species.

2.2.3 Summary

The ARM algorithm integrates the use of external information to provide a sound method for re-alignment of ambiguous reads. Information collected from GeneQC combined with predicted motifs and their target genes enables a probabilistic alignment strategy that does not rely solely on the local information from the read-level and reference genome. A negative binomial distribution is used to determine an alignment score for each potential gene location for every ambiguous read. Based on this alignment

score, the location with the highest likelihood is selected, with read counts being updated continuously throughout the process.

As demonstrated in the application of the ARM algorithm on data from *V. vinifera, A. thaliana*, *H. sapiens*, and *M. musculus*, this re-alignment strategy can significantly improve the quality of alignment, as determined through a statistically significant change observed in the pre- and post-ARM D-scores. This indicates the algorithm has applicability in reducing the impact of mapping uncertainty in reference-based RNA-Seq studies. The results also indicate a significant portion of the genes with some levels of mapping uncertainty can achieve improved alignment quality through the use of the ARM algorithm.

When considering the mapping uncertainty categorizations, the ARM algorithm also demonstrates the capacity for improving alignment qualities. In the *M. musculus* sample, over 20% of genes exhibiting mapping uncertainty saw a significantly enough reduction in mapping uncertainty to reduce their mapping uncertainty level, while no genes increased in mapping uncertainty categorization with over 25% of genes having a D-score reduction.

2.3 IRIS-EDA: Integrated RNA-Seq Interpretation System for Gene Expression Data Analysis

2.3.1 Gene Expression Data Analysis and Bottlenecks

One common investigation of RNA-Seq data is through analysis of estimated gene expression data. Analysis of the gene expression data is facilitated by computational

experience in appropriately designing the methods and experiments and conducting the analysis processes using one of many computing languages. This creates an obstacle for users with limited computational experience who want to analyze their RNA-Seq studies, thus there is an increased need for easy-to-use interactive expression analyses and results visualization [106].

While a wide variety of computational methods can be applied to expression data to determine particular qualities of the data on a sample or condition level [70, 107-112], differential gene expression (DGE) analysis is the most commonly used one. It allows researchers to identify differentially expressed genes (DEGs) across two or more conditions and can provide a meaningful way to attribute differences in gene expression levels to observed phenotypical and treatment differences. Many tools have been developed and optimized, such as: DESeq [46], DESeq2 [26], edgeR [36], limma [113], Cuffdiff [34], Cuffdiff2 [27], sleuth [114], and many others. While there have been substantial efforts in DGE analysis and visualization of DGE results [115-122], numerous pitfalls and bottlenecks persist, including experimental design implementation difficulties, a need for comprehensive integrated discovery-driven analyses and DGE tools, and the lack of functionalities and interactivity related to visualizing the analysis results.

To address these bottlenecks, we have created IRIS-EDA, which is an **I**nteractive **R**NA-Seq **I**nterpretation **S**ystem for **E**xpression **D**ata **A**nalysis. It provides a user-friendly interactive platform to analyze gene expression data comprehensively and to generate interactive summary visualizations readily. In contrast to other analysis platforms, IRIS-EDA provides the user with a more comprehensive and multi-level

analysis platform. IRIS-EDA outperforms other tools in several critical areas related to efficiency and versatile applicability: 1) Single-cell and bulk RNA-Seq analysis capabilities, 2) GEO submission compatibility, 3) six useful discovery-driven and DGE analyses, 4) experimental design approaches through three integrated tools for DGE analysis, and 5) seven interactive visualizations (Figure 14A).

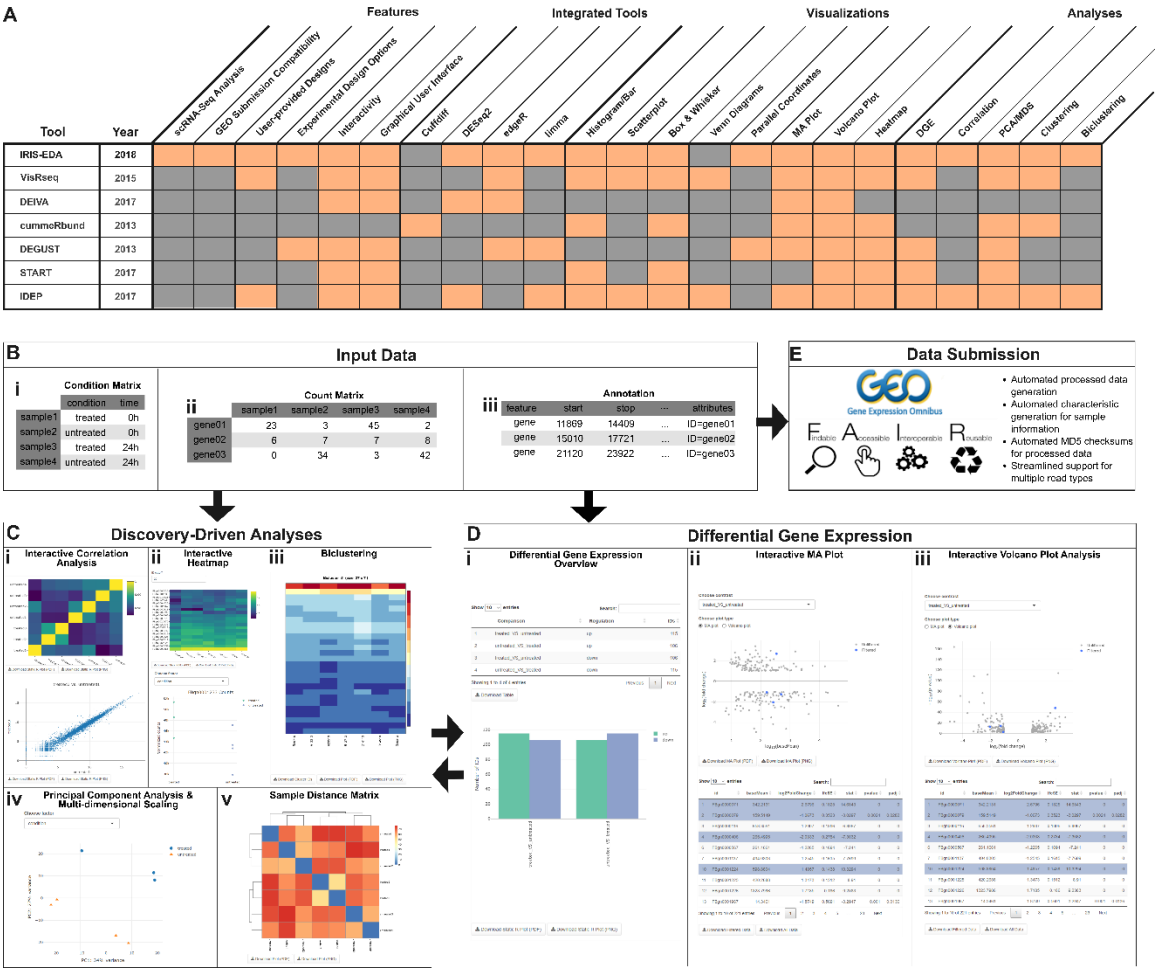


Figure 14: IRIS-EDA integrated functions. (A) Comparison of IRIS-EDA and six other DGE analyses and visualization tools; (B) Required Input Data for IRIS-EDA: (i) Condition Matrix indicating factor levels for each sample, (ii) Count Matrix consisting of gene expression values for each sample, with corresponding sample IDs matching those in the condition matrix, and (iii) the appropriate annotation file, which is required when using scRNA-Seq data; (C) Discovery-driven Analyses conducted by IRIS-EDA utilizing the Condition and Count matrices, including (i) Interactive Correlation Analysis with pairwise expression scatterplot, (ii) Interactive heatmap with parallel coordinate plot, (iii) Biclustering, (iv) Principal Component Analysis and Multi-dimensional Scaling, and (v) Sample Distance Matrix with clustering dendrogram; (D) Integrated

Differential Gene Expression analysis with visualizations: (i) Differential Gene Expression Overview with table and bar charts corresponding to up- and down-regulated gene counts, (ii) Interactive MA Plot with DGE results table, and (iii) Interactive Volcano Plot with DGE results table; and (E) Data submission compatibility to Gene Expression Omnibus following the FAIR guiding principles.

Focusing on these areas, IRIS-EDA provides comprehensive RNA-Seq data processing and analysis in a seamless workflow. This investigative approach uses expression quality control and discovery-driven analyses integrated with DGE analysis through one of the three most common R-based DGE tools (Table 4), *DESeq2*, *edgeR*, and *limma*, all of which have demonstrated capacities for differential gene expression analysis It provides users with a choice of intuitive experimental design options, as well as, the option to upload a custom design matrix in the DGE analysis. IRIS-EDA includes numerous interactive visualizations for each analysis type, enabling users to gain an immediate global view of their data and results or download as a high-resolution static image for publications. For the first time, this tool implements a framework based on the FAIR Data Principles [123] to assist users with the submission of their data and results to NCBI's Gene Expression Omnibus (GEO) [124].

Table 4: A comparative overview of citation counts for differential gene expression tools and servers as of March 1, 2018. Differential gene expression analytical tools (Tool) are compared based on the following criteria: Current number of citations (Citations), percentage of total citations from the analytical tools presented (Citation %), year the analytical tool was published (Year), approximate citations per year based on data accrued through 2017 (Citations/Year), and if the analytical tool has an R-based application (R-based).

Tool	Citations	Citation %	Year	Citations/Year (through 2017)	R-based?
edgeR [36]	7175	32.30090488	2010	1025	Yes
Cuffdiff [34]	4578	20.60955296	2012	915.6	No
Cuffdiff2 [27]	1525	6.86534912	2013	381.25	No
DESeq2 [46]	4355	19.60563634	2014	1451.666667	Yes
limma [25]	2451	11.03407914	2015	1225.5	Yes

DEGseq [22]	1244	5.600324135	2009	155.5	Yes
baySeq [24]	567	2.552559312	2010	81	Yes
SAMseq [21]	279	1.256021249	2013	69.75	Yes
NOIseq [23]	39	0.175572863	2012	7.8	Yes
sleuth [114]	45	0.202584072	2017	45	Yes

2.3.2 Methods and Implementation

IRIS-EDA was designed to provide a comprehensive platform for gene expression data analysis, which includes applicable analysis of both bulk and single-cell RNA-Seq data. Single-cell RNA-Seq (scRNA-Seq) data analysis is a growing area of study within RNA-Seq analyses and can provide unique insights into genetic occurrences within single cell types [125, 126]. The methods used for traditional DGE analysis have demonstrated applicability to scRNA-Seq DGE analysis, under certain conditions [126]. Thus, while designed for bulk RNA-Seq data analysis, IRIS-EDA can also facilitate discovery-driven and DGE analysis for scRNA-Seq data with few modifications. Namely, analysis of single-cell data can be appropriately carried out by using a stringent filter cutoff based on a default setting of transcripts per million (TPM) > 1, especially when combined with either edgeR or limma, which have both been shown to have high performance on scRNA-Seq data [126].

IRIS-EDA requires two or three user-provided input files, depending on the type of data used (Figure 14B): (i) a gene expression estimation matrix (EEM, also referred to as sample count data), (ii) a condition matrix with factor levels corresponding to the provided samples in the EEM, and (iii) a gene length matrix indicating the base-pair length of each gene to be used for filtering of scRNA-Seq data only. When uploading their data, users will select whether they are uploading bulk or single-cell RNA-Seq gene

expression data. If using scRNA-Seq data, the additional requirement for gene length matrix will be shown. Also, default parameterizations for optimized analysis for single-cell data will be populated throughout the server.

Once users have uploaded required data, IRIS-EDA provides two distinct analysis approaches. First, users can explore their data through a comprehensive discovery-driven analysis approach. This method provides users with tools and analyses for exploratory analysis of their expression data. Second, users can perform differential gene expression (DGE) analysis on their submitted data. In this method, users can determine which genes are differentially expressed using one of the three integrated DGE tools and can visualize the results through interactive visualizations. Whether users choose to first analyze their expression data using the discovery-driven analyses or through DGE analysis, they can continue to investigate their data with the other approach as well, in order to provide a comprehensive view of their RNA-Seq expression data.

After data upload, the two or three input files are first analyzed by IRIS-EDA quality control. Input data quality is evaluated using boxplots and histograms of the read count distributions. The purpose of the quality control process is to enable exploration of the submitted data and to verify that there are no unexpected or unexplainable abnormalities in the data, such as low total read counts or individual samples displaying strange distribution behavior. Once users have established proper data quality, they can proceed to the investigative analyses provided in IRIS-EDA.

IRIS-EDA discovery-driven analyses (Figure 14C) are various tools and algorithms designed to provide an investigative approach of expression data, especially for the situation where users do not have a strong direction or hypothesis for their data

analysis procedures. These algorithms assist users in analyzing and visualizing their EEM input information and discovering trends in their data that may provide additional hypotheses for downstream analyses. In particular, discovery-driven analyses can help users define a specific hypothesis within their RNA-Seq study, which can assist in development of experimental design methods for DGE analysis. Discovery-driven analyses processes that can be performed in IRIS-EDA include: sample correlation analysis and pairwise expression scatterplots (Figure 14Ci), expression heatmaps (Figure 14Cii), biclustering (Figure 14Ciii), principal component analysis and multidimensional scaling (Figure 14Civ), and sample distance matrix with clustering (Figure 14Cv). The figures generated through the discovery-driven analysis feature of IRIS-EDA are provided in an interactive manner, allowing users to select specific samples or pairwise comparisons to further evaluate. One such example is with the sample correlation analysis and pairwise scatterplots shown in Figure 14Ci. Users can choose one cell of the sample correlation matrix corresponding to a comparison between two samples. This will display the pairwise scatterplot for that specific comparison. The user can then scroll over the scatterplot and display the gene ID for an indicated data point.

After submitting data, users can move onto the DGE phase of IRIS-EDA. This analysis is performed using any one of the three provided tools: *DESeq2* [46], *edgeR* [36], and *limma* [113]. The default tool is *DESeq2*, based on independent evidence supporting its performance [56] and RNA-Seq analysis experience, but users can also select one of the other two tools based on their own preference. There are other high-performing commonly-used DGE tools available; however, their compatibility with IRIS-EDA excludes their use in IRIS-EDA. For example, tools that do not utilize sample

count data, e.g., *Sleuth,* [114] or are not R-based, e.g., *Cuffdiff* [34], are not included due to compatibility issues.

In addition to the DGE tool, the experimental design can also be specified by the user. The designs provided in IRIS-EDA include two-group comparisons for analysis of selected pairwise comparisons, multiple factorial comparisons, classic interaction design, additive models for pairing or blocking of data, main effect testing (testing time-series data) and blocked main effect testing. Additionally, IRIS-EDA provides a method for users to specify their own experimental design, for the instances when the user needs a design not already included in IRIS-EDA. Each of these methods has unique parameters to specify by the user, typically including which factors are intended for analysis and which specific comparisons are required. After analyzing the data, IRIS-EDA provides an overview displaying the number of up- and down-regulated IDs for each indicated comparison, along with a histogram displaying this information (Figure 14Di). The results table is also available through IRIS-EDA, along with interactive MA (Figure 14Dii) and Volcano plots (Figure 14Diii).

Similar to the figures generated in the Discovery-Driven Analysis section of IRIS-EDA, the plots in the DGE section are also highly interactive. Discovery-Driven Analysis features allows users to gain more specific information from their plots, including highlighting individual or regions of data points on the plot. These features highlight the corresponding row of the DGE results table, showing users gene information identifying them as outliers or falling within a certain region. Conversely, users can select specific gene IDs from the results table, resulting in the highlighting of

that gene ID's or set of gene IDs' data points on the corresponding plot. This feature can be used to easily determine the relative location of specific genes or gene sets in the plot.

Results obtained from the DGE analysis section of IRIS-EDA are often not the end of the analysis procedures. Based on the information collected, users may choose to further investigate their expression data using additional analyses provided in the Discovery-Driven Analyses section, such as the clustering or biclustering. When DGE and Discovery-Driven analyses are combined, the analyses provide a more comprehensive data interpretation.

IRIS-EDA provides users with methods for extracting content based on discovery-driven and DGE analyses. All figures in the QC, Discovery-Driven Analysis, and DGE Analysis sections have the option for users to download as a static image in PDF or PNG format. Additionally, all tables in the DGE Analysis section are downloadable as CSV files, with the final results table being downloaded in its entirety or filtered based on user-provided or default-adjusted p-value and log fold-change cutoffs. As part of the biclustering analysis, users can also download a list of gene IDs contained within the specified cluster.

Many users are eventually interested in submitting their RNA-Seq data to a public repository for accessibility, but this process can be tedious and troublesome. NCBI's GEO database has specific requirements related to the data, results, and accompanying metadata file. To assist users in their preparation of documents for GEO submission, IRIS-EDA offers an optional GEO page. In following with the standard of set forth by the FAIR Data Principles [123], this page asks users to provide a limited amount of information that will be used, along with the previously provided condition matrix

information, to populate the metadata file required for GEO submission. This populated metadata file will then be available for download with reformatted processed data files extracted from the EEM. These two pieces of information can later be submitted with the original raw FASTQ-formatted RNA-Seq data to the GEO submission page.

2.3.3 Summary

IRIS-EDA is a platform developed for comprehensive expression data analysis, visualization and interpretation of both bulk and single-cell RNA-Seq data. It is designed to address current bottlenecks and issues in existing expression analysis and DGE analysis packages. This interactive tool implements numerous features including EEM quality control, discovery-driven analyses, and DGE analysis utilizing the most commonly used R-based DGE tools in a user-friendly, comprehensive platform. It is noteworthy that IRIS-EDA provides advanced experimental design options in an intuitive format, while also allowing users to provide their own design matrix to facilitate efficient DGE analysis for a broad spectrum of users. Each analysis section within IRIS-EDA provides relevant information in a highly-interactive visual format. To further facilitate compatibility with the FAIR Data Principles, IRIS-EDA also provides a framework that will greatly assist users in formatting their results and metadata for GEO submission. It is our belief that this tool will support users of all computational experience levels and with all DGE requirements.

2.4 ViDGER: Visualization of Differential Gene Expression Results Using R

2.4.1 Interpreting Differential Gene Expression Results

While some users can benefit from an integrated web server such as IRIS-EDA, others have long been using traditional methods for analyzing expression data and generating differential gene expression results. *Cuffdiff* [27, 34], *edgeR* [36], and *DESeq2* [26] are three widely-used tools to determine which genes are differentially expressed, based on quantifications of expressed genes derived from computational analyses of raw RNA-seq reads (e.g., mapping [29, 31-33, 35, 37, 40, 42, 44] and assembly [30, 38, 41, 43, 127, 128]). Each of the three has been shown to be among the highest performing tools for DGE analysis of RNA-seq data [56, 129, 130] and contribute to the highest number of citations for DGE tools (Table 4), representing roughly 80% of all cited DGE tools. However, interpreting the format and content of results files from each program is not entirely intuitive, especially for researchers who have limited computational backgrounds. One of the best ways to provide a summary of the DGE results is to generate figures, giving a global representation of the expression changes across multiple conditions. The three tools create output files sharing some information, such as mean gene expression across replicates for each sample, log_2 fold change (*lfc*), and adjusted *p*-value. However, these output files have many differences in content and structure, which makes generating comprehensive visualizations time-intensive and potentially challenging task. *cummeRbund* [131] is an available tool to generate visualizations for *Cuffdiff* outputs but has no functionality for users of *edgeR* and *DESeq2*. Additionally, many differential gene expression tools have integrated methods to generate a limited number, variety, and quality of visualizations (Table 5). This

limited functionality leaves many researchers with no readily available method to create visualizations for their DGE results. To remediate this issue, the developed R/Bioconductor [132] package ViDGER [133] assists users in generating publication-quality visualizations from *Cuffdiff*, *edgeR*, and *DESeq2* capable of providing valuable insight into their generated DGE results.

Table 5: Nine functions for differential gene expression analysis results and their implementation in commonly-cited differential expression tools.

Function	edgeR	cummeRbund	DESeq2	limma	DEGseq	baySeq	SAMseq	sleuth	NOIseq
Treatment distrs.	No	Yes	No	Yes	No	No	No	Yes	Yes
FPKM/CPM scatterplot	No	Yes	No	No	No	No	No	Yes	No
FPKM/CPM matrix	No	Yes	No	No	No	No	No	No	No
DEG counts	No	Yes	No	No	No	No	No	No	No
MA plot	Yes	Yes	Yes	Yes	No	Yes	No	Yes	Yes
MA plot matrix	No	No	No	No	No	No	No	No	No
Volcano plot	No	Yes	No	Yes	No	No	No	Yes	No
Volcano plot matrix	No	Yes	No	No	No	No	No	No	No
Four-way plot	No	No	No	No	No	No	No	No	No

This package integrates six different types of expression-based visualizations: boxplots, scatterplots, DEG counts, MA plots, volcano plots, and Four-way plotsas shown in Figures 15 & 16. Additionally, matrices of all pair-wise comparisons can be generated with scatterplots, MA plots, and volcano plots. All the visualizations can be

classified into two tiers, with the Tier 1 functions (Figure 15) representing more basic information, whereas the Tier 2 functions (Figure 16) being used to derive more advanced information with *p*-values, fold changes, and mean expression values. All generated figures and extracted data can then be saved and used for further purposes, including reports and publications.

2.4.2 Methods and Implementation

ViDGER is a package developed for the R environment (>= 3.3.2) and is freely available at https://www.bioconductor.org/packages/3.7/bioc/html/vidger.html. Several package dependencies are required, i.e., *ggplot2* [134], *ggally* [135], *dplyr* [136], and *tidyr* [137]. Currently, it is compatible with three commonly used DGE analysis packages, which are *Cuffdiff*, *edgeR*, and *DESeq2*. Function efficiency varies depending on what type of RNA-seq package is used. Functions used for *Cuffdiff* and *edgeR* objects complete in < 1s and while *DESeq2* objects can take up to 5s to complete. *DESeq2* objects take longer to process due to the nature of the object, which contains more stored information than the relatively simple objects for *Cuffdiff* and *edgeR*. One exception is the volcano plot matrix function **(vii)**. *Cuffdiff* and *edgeR* objects took < 10s to complete while *DESeq2* objects took >10s. Calculations were performed on three toy data sets from *Cuffdiff*, *DESeq2*, and *edgeR* outputs. Additionally, we tested the robustness of this package on multiple large-scale RNA-seq datasets from human and plant samples. All computations were performed on a computer with a 64-bit Windows 10 operating system, 8 GB of RAM, and an Intel Core i5-6400 processor running at 2.7 GHz.

Nine functions are included in ViDGER, each of which is capable of using *Cuffdiff*, *DESeq2*, and *edgeR* objects. Included in the ViDGER package are three example

datasets representing the three DGE tool object types. Specifically, *df.cuff* is based on *Cuffdiff* data from the *cummeRbund* package [131]; *df.deseq* is a *DESeqDataSet* object based on gene expression data from the *pasilla* package [138]; *df.edger* is an example *DGEList* object derived from the *edgeR* package. In addition to the example data sets, ViDGER was tested on five real-world data sets, consisting of one *H. sapiens*, one *M. domestica*, and three *V. riparia* datasets, although these are not provided with the package. It is important to note that the input data for this package should be the direct output and of one of the classes corresponding to the specific tool used (DESeqDataSet, DGEList or other edgeR objects, or Cuffdiff object) and not a basic matrix or data frame containing the results of these tools. The following examples are illustrated using the *df.deseq* object, with full demonstrations with the *Cuffdiff, DESeq2, and edgeR* objects found in the supplementary file.

2.4.2.1 Tier 1 Functions

(i) vsBoxPlot visualizes log_{10} distributions for treatments in an experiment as box and whisker diagrams (Figure 15A), where only the data frame and analytical type are needed unless using a DESeq2 object where the factor is also required. This figure is useful for determining the distribution of mapped read counts for each treatment in an experiment and can highlight specific samples that have distributions differing significantly from what is expected or what is displayed with the other samples. Visualizing this information can provide insight into the base quality of the read distributions to ensure semi-consistent sample-based quality levels. The *DESeq2* object (*df.deseq*) is used in the following example, and the factor variable, *d.factor*, for the treatments need to be specified. The generated visualization is shown in Figure 15A.

```
vsBoxPlot(data = df.deseq, d.factor = 'condition', type = 'deseq')
```

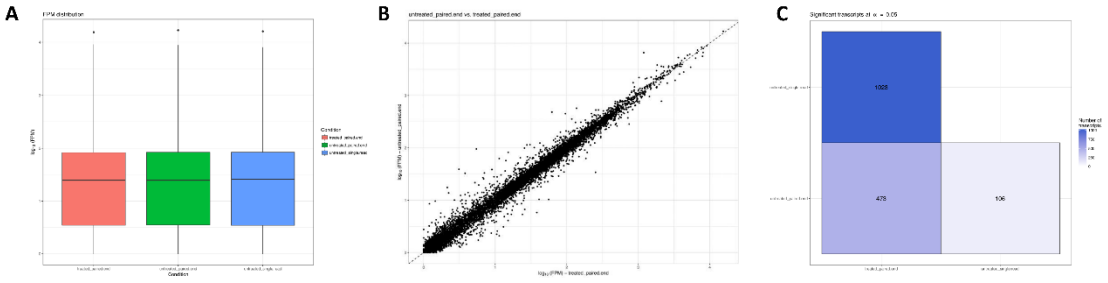


Figure 15: Tier 1 Functions. (A) Visualization generated by the vsBoxPlot function from the ViDGER package using a DESeq2 dataset, requiring a dataset, factor type, and appropriate tool type. Optional parameters include inclusion/exclusion of the main title, legend, and grid; (B) Visualization generated by the vsScatterPlot function from the ViDGER package using a DESeq2 dataset, requiring a dataset, factor type, two factor levels, and appropriate tool type. Optional parameters include inclusion/exclusion of the main title and grid; (C) Visualization generated by the vsDEGMatrix function from the ViDGER package using a DESeq2 dataset, requiring a dataset, factor type, and appropriate tool type. Optional parameters include inclusion/exclusion of the main title, legend, and grid and specification of adjusted p-value cutoff (default is 0.05).

(ii) *vsScatterPlot* creates a scatterplot of log_{10} comparison of either FPKM (Reads Per Kilobase of transcript per Million mapped reads) or CPM (cost per thousand impressions) measurements for two treatments, depending on the user-provided object format (Figure 15B). This function can be used to compare measurements of mapped reads to transcripts from two treatments, which allows for a global view of the expression similarity between the two selected treatments. Scatterplots that generate most data points falling along the diagonal indicate more similar expression patterns for the two treatments, whereas data points falling further from the diagonal would indicate relatively less similar expression levels. By stating *x* and *y* treatment variables and/or the data

source, we can generate a scatterplot of the pairwise *x* vs. *y* comparison. The generated visualization is shown in Figure 15B.

```
vsScatterPlot (x = 'treated_paired.end', y = 'untreated_paired.end', data = df.deseq, type ='deseq', d.factor = 'condition')
```

(iii) *vsScatterMatrix* generates a matrix of scatterplots for all possible treatment combinations with additional distribution information. In addition to the scatterplots which are generated as with the *vsScatterPlot* function, the matrix option provides FPKM/CPM distributions for each sample and correlation values for each pairwise comparison. This approach allows for a view of each relative expression pattern and correlation all in one visualization.

```
vsScatterMatrix(data = df.deseq, d.factor = 'condition', type = 'deseq')
```

(iv) *vsDEGMatrix* visualizes the number of DEGs at a specified adjusted *p*-value for each treatment comparison (Figure 15C). It can be utilized to quantify the number of significantly DEGs for each comparison and provides a heatmap-based color scheme with a gradient to represent the relative magnitude of DEGs for each comparison. Like the other matrix functions, data specification and analytical type are required. The user can also specify an adjusted *p*-value which defaults to 0.05. Methods for extracting the DEGs for each comparison can be found in *Data Extraction*. The generated visualization is shown in Figure 15C.

```
vsDEGMatrix(data = df.deseq, d.factor = 'condition', type='deseq')
```

2.4.2.2 Tier 2 Functions

(v) *vsMAPlot* creates an MA plot, which is a scatter plot with M (log ratio) and A (mean average) scales, of *lfc* versus normalized mean counts (Figure 16A). In addition to the basic plotting of the data points relative to the mean expression values and *lfc*, the *vsMAPlot* function also integrates visualization features that allow for a better understanding of the data. Data points in the MA plot are colored based on thresholds for the adjusted *p*-value and *lfc* of the gene in the indicated comparison to provide valuable global interpretability. Additionally, it is inevitable with most datasets that some points will be extreme relative to the majority of the data, which caused problems when generating visualizations. To address this issue, *vsMAPlot* scales the window based on the bulk of the data and represents outliers with distinct data points, indicating the magnitude of the outlier based on the size of the point. This process allows for the visualization to present the majority of the information in a viewable, usable format that is robust to outliers. Visualizing the data through this approach allows for the comparison of two treatment groups relative to the mean expression value and *lfc*. The *x* and *y* parameters specify how the fold changes are generated (e.g., $FC = log_2(\text{sample y}/\text{sample x})$). The generated visualization is shown in Figure 16A.

```
vsMAPlot(x='treated_paired.end', y='untreated_paired.end', data=df.deseq,
d.factor='condition', type='deseq')
```

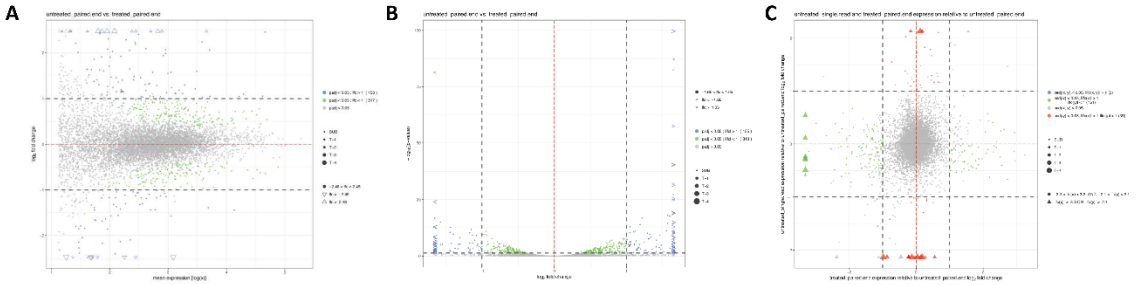


Figure 16: Tier 2 Functions. (A) Visualization generated by the vsMAPlot function from the ViDGER package using a DESeq2 dataset, requiring a dataset, factor type, two factor levels, and appropriate tool type. Optional parameters include inclusion/exclusion of the main title, legend, and grid, manual specification of the y-axis limits, lfc threshold (default is 1), and adjusted p-value cutoff (default is 0.05), and specification of returning data in tabular form; (B) Visualization generated by the vsVolcano function from the ViDGER package using a DESeq2 dataset, requiring a dataset, factor type, two factor levels, and appropriate tool type. Optional parameters include inclusion/exclusion of the main title, legend, and grid, manual specification of the x-axis limits, lfc threshold (default is 1), and adjusted p-value cutoff (default is 0.05), and specification of returning data in tabular form; (C) Visualization generated by the vsFourWay function from the ViDGER package using a DESeq2 dataset, requiring a dataset, factor type, two factor levels, reference factor level, and appropriate tool type. Optional parameters include inclusion/exclusion of the main title, legend, and grid, manual specification of the x- and y-axis limits, lfc threshold (default is 1), and adjusted p-value cutoff (default is 0.05), and specification or returning data in tabular form.

(vi) *vsMAMatrix* generates a matrix of MA plots for all possible pairwise treatment comparisons. This process, as with the other matrix options, allows users to visualize all their treatment-based comparisons in one figure. This matrix option also includes counts for each figure based on *lfc* and adjusted *p*-value thresholds, which can be specified by the user or revert to the default 1 and 0.05, respectively.

```
vsMAMatrix(data = df.deseq, d.factor = 'condition', type ='deseq')
```

(vii) *vsVolcano* creates a volcano plot for two treatments comparison by plotting the $-log_{10}$(*p*-value) against the *lfc* (Figure 16B). As with the *vsMAPlot* function, the

vsVolcano function utilizes coloring schemes to indicate the significance of magnitude of differential expression for the individual data points. Additionally, this function integrates the same data point and sizing structure to focus the plot window on the majority of the data, indicating outliers in this format. The generated visualization is shown in Figure 16B.

```
vsVolcano(x = 'treated_paired.end', y = 'untreated_paired.end', data = df.deseq, d.factor = 'condition', type = 'deseq')
```

(viii) ***vsVolcanoMatrix*** generates a matrix of volcano plots for all possible pairwise treatment comparison. This process, as with the other matrix options, allows users to visualize all their treatment-based comparisons in one figure. Additionally, to provide a more comprehensive view with a single figure, we included a count for each separate Volcano plot based on the number of data points in each section as specified by the *lfc* and adjusted *p*-value thresholds. Although this option may have experience limited use, it would be useful in situations where users wish to show mass similarity across all comparisons, highlight the individual or limited deviations, or display situations where the comparisons vary widely.

```
vsVolcanoMatrix(data = df.deseq, d.factor = 'condition', type ='deseq')
```

(**ix**) ***vsFourWay*** creates a scatter plot comparing the *lfc* between two samples and one control (Figure 16C). This approach is most useful when there are multiple comparisons being made against a specific control or relative sample. Using this function, a plot can be generated for visualizing the expression scatterplots, relative to another

expression scatterplot. As with the other two main Tier 2 functions, *vsFourWay* integrates data point features to highlight significant adjusted *p*-values, over-threshold *lfc*, and outliers. In this function, *x* and *y* arguments are needed, and a *control* level is also required. Although it is possible to generate a matrix option for the FourWay plot, the authors decided against this because of two main issues. First, the *vsFourWay* function generates a significant amount of information in a single figure, with nine distinct sections representing nine distinct combinations of relative *lfc*. Creating a matrix visualization with this figure would then force each FourWay plot to be too small to collect meaningful interpretations from, thus counteracting the purpose of the package. Secondly, the *vsFourWay* function already requires three factor levels for comparison—one reference level and two comparison levels. A matrix option for this functionality would then require a minimum of four factor levels, with at least five factor levels being preferred to generate a fully-informative matrix option. This requirement would potentially put most applications out of the scope of the matrix option for the *vsFourWay* function. The generated visualization is shown in Figure 16C.

```
vsFourWay(x = 'treated_paired.end', y = 'untreated_single.end', control =
'untreated_paired.end', data = df.deseq, d.factor = 'condition', type = 'deseq')
```

It is noteworthy that functions **(v)**, **(vii)**, and **(ix)** can return interpreting results shown in the visualizations for further analysis and interpretation (Table 6). The data extracted contains all relevant information used to generate the specified figure, including mean expression for the *x*, *y*, and *control* (in the *vsFourWay* function) factor levels, x- and y-axis values for the relevant figure, an ‘isDE’ column indicating whether the gene ID is differentially expressed based on the adjusted *p*-value threshold, ‘color’ indicating

the color of the data point in the figure—which corresponds to the *lfc* and adjusted *p*-value thresholds—and 'size' indicating whether the data point is on the plot or an outlier and magnitude of that outlier. The data extraction is accomplished by setting the *data.return* parameter to *TRUE*.

Table 6: ViDGER Data Extraction. Data extraction from the vsVolcano function from the ViDGER package using a DESeq2 dataset. This is the same parameterization as used in Figure 16B, except data.return = TRUE. This modification will allow the user to extract relevant data from the figure. In this case, the extracted data frame includes mean expression values for the x and y factor levels, $\log_2$ fold change (logFC), p-value (pval), adjusted p-value (padj), 'isDE' which represents whether the differential expression is significant, 'color' which signifies the color of the data point corresponding to the adjusted p-value and lfc thresholds, and 'size' which indicates whether the data point is within the plot frame or an outlier of a particular magnitude.

	x	y	logFC	pval	padj	isDE	color	size
FBgn0000008	7.92227	8.32225	0.07105	0.828806	0.974685	FALSE	grey	sub
FBgn0000017	318.957	383.285	0.26505	0.090161	0.467683	FALSE	grey	sub
FBgn0000018	30.2586	31.2699	0.04743	0.801233	0.971289	FALSE	grey	sub
FBgn0000032	72.3419	72.9032	0.01115	0.949842	0.993072	FALSE	grey	sub
FBgn0000037	1.53958	0.81229	-0.9224	0.231142	0.700057	FALSE	grey	sub
FBgn0000042	7928.52	5600.30	-0.5015	0.000611	0.013572	TRUE	green	sub
FBgn0000043	3273.93	1943.28	-0.7525	7.96E-08	5.68E-06	TRUE	green	sub
FBgn0000044	2.22202	1.59958	-0.4741	0.456526	0.872166	FALSE	grey	sub
FBgn0000046	2.23561	1.53025	-0.546	0.439278	0.865892	FALSE	grey	sub
FBgn0000052	187.154	201.437	0.10610	0.498756	0.889058	FALSE	grey	sub
FBgn0000053	200.419	161.082	-0.3152	0.03254	0.260826	FALSE	grey	sub
FBgn0000054	50.2460	52.843	0.0727	0.675076	0.949335	FALSE	grey	sub
FBgn0000057	56.8849	55.5293	-0.0347	0.831612	0.974685	FALSE	grey	sub
FBgn0000063	34.4397	27.5858	-0.3201	0.084865	0.453512	FALSE	grey	sub
FBgn0000064	738.380	597.975	-0.3042	0.010905	0.125567	FALSE	grey	sub
FBgn0000071	54.9849	9.35883	-2.5546	1.98E-27	1.17E-24	TRUE	blue	t4
FBgn0000077	17.9897	17.5863	-0.0327	0.898181	0.985072	FALSE	grey	sub

FBgn0000078	1.74364	3.47347	0.99427	0.058949	0.37159	FALSE	grey	sub
FBgn0000079	9.72273	21.8755	1.16988	3.45E-06	0.000156	TRUE	blue	sub

2.4.3 Summary

Differentially expressed genes are frequently used to determine genotypical differences between two or more conditions of cells in support of specific hypothesis-driven studies. Interpretation of this information can benefit significantly from the graphical representation of results files. The ViDGER R/Bioconductor package to assists in the process of generating publication quality figures of DGE results files from *Cuffdiff*, *DESeq2*, and *edgeR*. Through the use of the nine integrated functions, this package will greatly assist biologists and bioinformaticians in their interpretations of DGE results. Utilizing this package will provide a straightforward method for comprehensively viewing differentially expressed genes between samples of interest and allows researchers to generate usable figures for furthered dissemination of their differential gene expression studies.

CHAPTER 3: Collaborative Efforts

3.1 Computational Tool Collaborations

3.1.1 Review of Motif Prediction Methods and DMINDA2.0

Cis-regulatory motifs—motifs for short—are short, conserved DNA sequences, typically 8-20 bps long [94]. Often times, motifs act as transcription factor binding sites (TFBSs) and play significant roles in the rate of transcription regulation of nearby target genes and further control their expression levels. Hence, *de-novo* motif prediction and

related analyses, such as motif scan and comparison, provide a solid foundation for the inference of gene transcriptional regulatory mechanisms in both prokaryotic and eukaryotic organisms [139, 140]. Specifically, these techniques can also contribute substantially to system-level studies (e.g. regulon modeling, regulatory network construction such as that used in the ARM algorithm, etc.) [139, 141, 142]. Due to the rapid increase in size and availability of sequenced genomes combined with improvements in advanced biotechnologies, numerous computational methods for identification of motifs have been developed to extract information from query DNA sequences. Even with the substantial efforts in this area, motif characteristics (high variation and short length) still pose a great challenge [143].

Identification of motifs from provided promoters has been one of the most prevalent methods since the 1980s, with various tools and algorithms having been developed for this purpose [144-151]. Developed tools for this purpose include AlignACE, BioProspector, CONSENSUS, MDscan, MEME, CUBIC, MDscan, and BOBRO [148, 149, 151-162], some of which have been implemented successfully for construction of regulatory networks [139, 142]. Even with these variety of methods and approaches, motif prediction still suffers from high false positive rates [147, 163-165]. To address this specific issue, algorithms utilizing phylogenetic footprinting [166, 167] were also developed, including PhyloGibbs, Footprinter, PhyloCon and MicroFootprinter [152, 168-172]. However, the lack of leveraging the phylogenetic relationship between genome and query sequences led to less-than-stellar performance of many of these tools [159], which resulted in many motif instances being not conserved enough to properly carry out motif prediction [173-175].

With the increased development of high-throughput biotechnologies [9, 176-184], in particular ChIP-Seq data, a new level of information is available for motif prediction and analyses. Utilization of this data has potential benefits for motif prediction based on peak-calling methods [185-197], like those found in tools such as SPP [185], MACS[198], CisGenome [199], FindPeaks [200], QuEST [186] and PeakRanger [201]. While the use of larger-scale data and improved methods have benefited motif prediction, an algorithmic analysis of current algorithms (FMotif [202], DREME [197], RSAT peak-motifs [203], SIOMICS [204, 205], and Discrover [206]) shows that there are still areas for improvement. In particular, an integrated web server for analysis of ChIP-Seq data related to motif prediction and analyses is essential [143].

One such tool that addresses the issue of an integrated web server for motif prediction is DMINDA2.0 [100], which is an updated version of the DMINDA web server [207]. This tool integrates *de-novo* motif finding using BOBRO [101] or phylogenetic footprinting tool MP3 [208], scanning, comparison, and co-occurrence analysis in a web server format (Figure 17). DMINDA2.0 allows users to upload DNA sequences or select species-specific sequences from one of the linked databases. Motif prediction is performed on the loaded sequences using BOBRO or MP3 to identify statistically significant motifs from a set of provided promoters. BOBRO has been demonstrated to have higher performance in terms of both efficiency and accuracy than any other high-performing motif prediction tool [209]. Motif scanning searches provided genomic sequences for all instances of a query motif. Motif comparison performs a statistical comparison of the similarity of queried motifs and clusters similar motifs into groups. Motif co-occurrence analysis identifies motifs that co-occur in the provided

sequences to determine motifs that potentially regulate the same set of genes. The information obtained from motif prediction and analyses for prokaryotic genomes can then be used to predict regulons, which are co-regulated groups of genes which contribute to transcriptional regulation. In addition to the provided analysis results, the predicted motifs and regulons are presented using motif logos and Cytoscape-like visualizations, respectively.

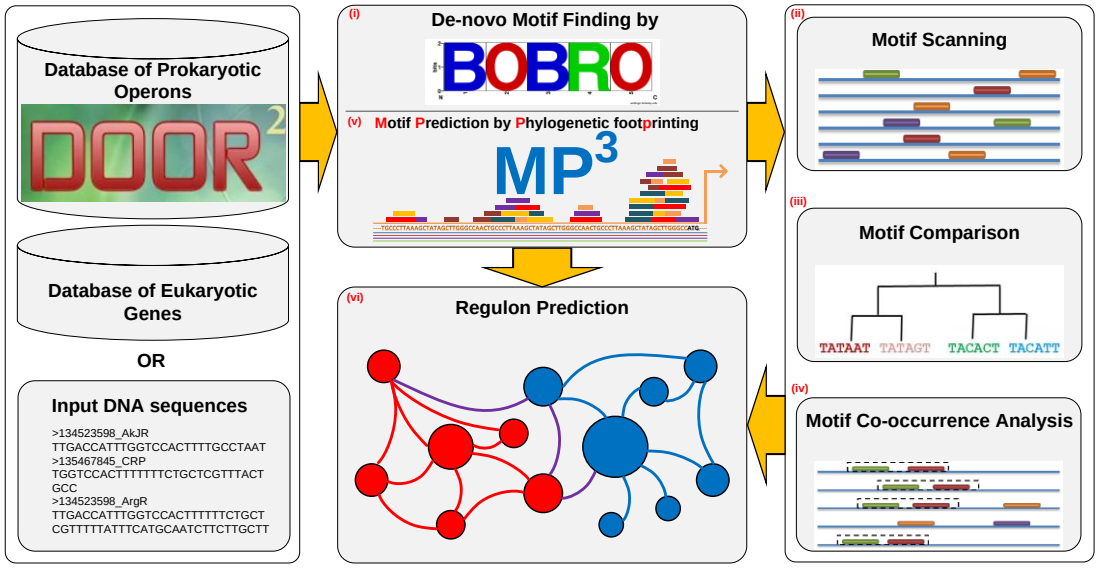


Figure 17: DMINDA2.0 Web server. Workflow of DMINDA2.0, including (i) de-novo motif finding using BOBRO, (ii) motif scanning, (iii) motif comparison, (iv) motif co-occurrence analysis, (v) de-novo motif finding based on phylogenetic footprinting strategy, and (vi) regulon prediction

3.1.2 RECTA: Regulon Identification Based on Comparative Genomics and Transcriptomics Analysis

Elucidation of gene regulatory network hierarchies offers understanding into the coordination of stress response capabilities for microbial species [210-213]. One specific way to investigate these hierarchies is through regulon prediction. Regulons are

co-regulated gene groups that contribute to transcription regulation in microbial genomes. The ability to detect and understand these gene groups has the potential to aid in the deeper understanding of regulatory mechanisms within prokaryotic cells.

There are three main ways to predict regulons. The first method combines a comparative genomic strategy with motif profiling to identify related regulon members for existing regulons, followed by a study of systematic regulation [214, 215]. The second method integrates motif analysis strategies, namely motif comparison and co-occurrence analysis. This approach identifies significantly enriched motif candidates which are then assembled into regulons [162, 216]. The third approach, *ab initio* regulon prediction through *de novo* motif finding methods, uses phylogenetic footprinting strategies combined with reference verification [166, 169, 217]. This process utilizes a parallel search of known regulons or transcription factors from relevant species to predict regulons in the target organism.

In utilization of these methods, a regulon prediction pipeline was developed. RECTA, regulon identification based on comparative genomics and transcriptomics analysis, provides a framework to determine gene regulatory networks in microbial species [218]. This framework integrates six steps: (1) co-expressed gene modules and differentially expressed genes are generated from expression data using hierarchical clustering and a Wilcoxon test, respectively. Simultaneously, the DOOR2 database [219] is used to predict operons from respective genome sequences, with operons being assigned to each co-expression module; (2) 300bp upstream of the promoter for each co-expression module is used to identify motifs using DMINDA2.0 [100]; (3) Clustering and similarity comparisons are used to reassemble the top five most significant motifs in

each co-expression module; (4) the MEME suite [220] is used to compare known transcription factor binding sites with predicted motifs, while BLAST [89] is used to map transcription factor binding sites to the appropriate genome; (5) experimentally validated functional-specific genes from similar organisms are mapped to the same genome using BLAST; and (6) the relationship between functional gene modules and identified regulons is established to determine an overall functional mechanism.

To fully develop the regulon prediction pipeline and test its application power, RECTA was used to develop and acid stress response regulatory network for *Lactococcus lactis* (Figure 18). This species has demonstrated capabilities for vaccine and protein delivery in immunological treatments of diabetes [221], malaria [222], tumors [223, 224] and various infections [225]. The relatively high acid stress response for *L. lactis* provides the result of protecting the cell against destruction inside animal bodies, something that is beneficial for oral drug therapies [226]. Its dynamic evolved stress response system has led to *L. lactis* being a promising species to study with respect to microbial response to harsh environments [210, 227, 228]. In particular, acid stress response is an area of specific interest due to its connection to alarmones [229], leading to a detectible change in cellular regulation [230].

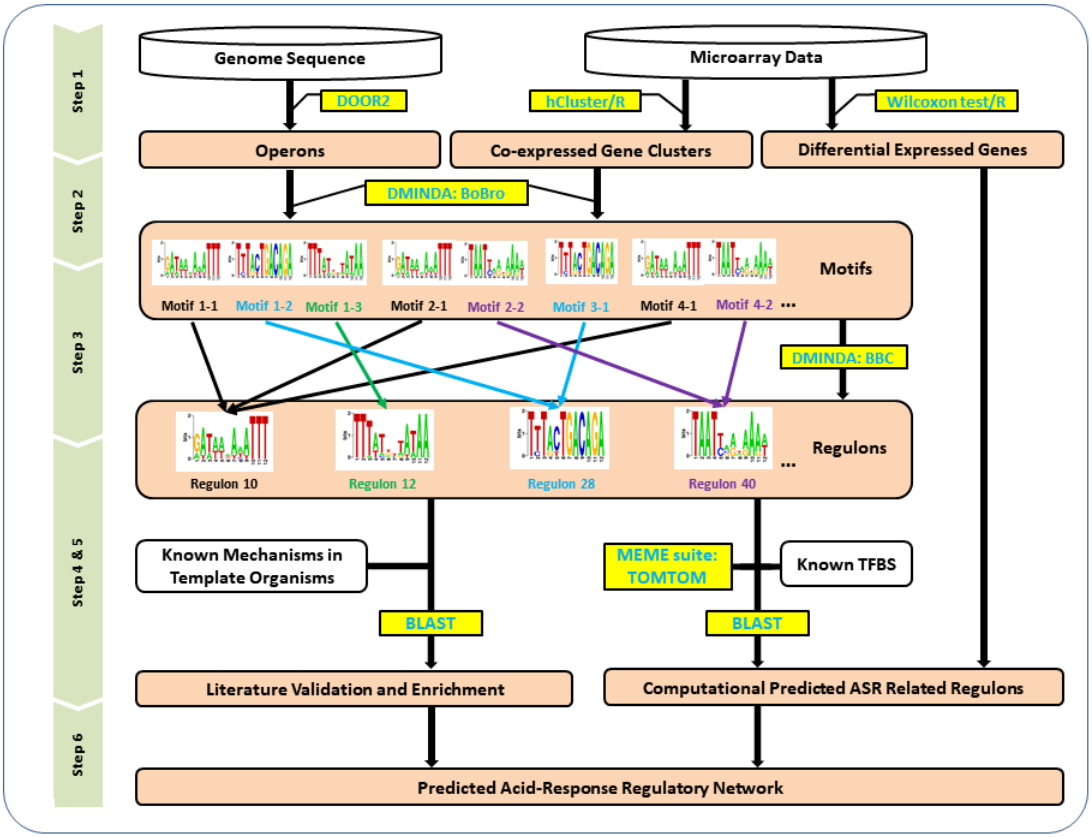


Figure 18: RECTA Framework. The flowchart of constructing global ASR transcriptional network in MG1363. Step 1: microarray data was used to generate co-expressed gene clusters and DEGs, and MG1363 genome sequence was used to find operons. Step 2: a motif finding progress was carried out to identify all statistically significant motifs in each of the CEMs. Step 3: a regulon finding procedure was designed to identify all the possible regulon candidates encoded in the genome based on motif comparison and clustering. Step 4: the motifs of each of these regulons were compared to known TFBSs, and DGE analysis between low pH condition and normal condition was used to figure out the ASR-related regulons. Step 5: regulon validation based on literature information verified the significant putative regulons and expanded the results to some insufficiently significant regulons. Step 6: the ASR-related GRN in MG1363 was predicted and described with eight regulons, nine functional modules, and 33 genes. The combination of the above information forms a genome-scale regulatory network constructed for ASR.

To investigate this acid response system in terms of regulon prediction, RECTA was applied to the *L. lactis* MG1363 genome sequence from NCBI's GenBank [231]. Microarray from eight varying acid response conditions was collected from NCBI's GEO [124]. DOOR2 was applied to the MG1363 genome sequence, resulting in 1565

identified operons consisting of 2439 coding genes. Co-expression analysis was used to group the operons into 124 co-expressed clusters. Of the 124 clusters, the two with more than 200 operons were removed to reduce the false positive rate. The BOBRO algorithm was used through the DMINDA2.0 server to analyze 300bps upstream of the start sites of each operon. The top five most significant motifs were selected from each cluster, resulting in 610 identified motifs. Using a similarity cutoff of 0.8, motif comparison was used to identify 51 motif clusters. These 51 motif clusters indicate 51 predicted regulons.

Of the 51 predicted regulons, 14 contained motifs matching known TFBSs through TOMTOM from the MEME suite. The transcription factors corresponding to these known TFBSs were mapped to the MG1363 genome using BLAST to determine the transcription factors that have been identified to regulate the respective regulons. Consequently, eight known transcription factors (spo0A, lhfB, GAL80, CovR, c4494, ihfA, CovR, and RHE_PF00288) were successfully mapped to the MG1363 genome. Considerations of the differentially expressed genes obtained from the microarray data and their containment within particular regulons, five regulons were determined to have involvement to the gene regulatory network in MG1363. Additionally, literature was used to verify the identified regulons, resulting in eight total regulons being linked to the acid stress response mechanism for MG1363.

3.1.3 Metagenomic and Metatranscriptomic Analysis & the Integrated Meta-Function Pipeline

Microbial communities are found in numerous environments, including the human gut, oceans, soils, and other animals [232]. Even within the same environment,

microbial communities can be quite diverse in their complexity and competition. Studying microbes and their respective environments has become increasingly common, especially due to the connection of microbial communities with human diseases such as obesity, inflammatory bowel disease, and lean or obese twins [233, 234] and observed evidence connecting microbial communities with human physiology [235, 236]. The use of sequencing technologies to study microbial genomes, referred to as microbiomes, provides a unique angle in which to view microbial communities and to study their underlying mechanisms in response to and affecting environmental changes.

In attempting to understand microbiomes, multiple levels of information are collected, including 16S ribosomal RNA analysis, whole-genome shotgun (metagenome) analysis, and whole-transcriptome shotgun (metatranscriptome) analysis. These analyses use rRNA to identify microbes within a microbial community, use genetic information to detect microbial identities—sometimes even down identification of particular strains—and observe gene expression patterns and functional differences in communities, respectfully.

Numerous studies utilize complex levels of information to gain a broad understanding of the interactions between microbial communities and their environments. Studies such as the Human Microbiome Project (HMP) [237], Interactive HMP [238], Metagenomics of the Human Intestinal Tract [233] investigate microbiomes with respect to human hosts, generally in one particular context such as the intestinal tract. The Earth Microbiome Project (EMP) similarly analyzes microbial ecosystems, specifically studying the distribution, diversity, and structure of the communities. So far, EMP has collected over 30,000 samples from various ecosystems and hosts around the world

[239]. The increased understanding of and interest in microbial communities and their respective microbiomes has directly led to the more widespread application of sequencing procedures in metagenomics and metatranscriptomics [240-243].

Numerous tools have been developed for the purpose of analyzing metagenomic and metatranscriptomic data, especially in the areas of species-level [244-249] and strain-level metagenomics analysis [250-254]and metatranscriptomic analysis [255-259]. While these tools can individually identify microbial composition or gene expression information, they cannot simultaneously perform both functions. Incorporation of both approaches allows for a better understanding of the mechanisms of the microbial community from a gene expression-level and species and/or strain composition-level. In pursuit of this approach, the Integrated Meta-Function (IMF) pipeline was developed (Figure 19A) [260]. This framework takes input metagenomic and metatranscriptomic sequencing data and incorporates various functional databases to efficiently and effectively map the input data together, generating a comprehensive view of a particular microbiome. Databases integrated into this framework include The Comprehensive Antibiotic Resistance Database (CARD) [261], Antibiotic Resistance Genes Database (ARDB) [262], DrugBank [263], and the Human genome.

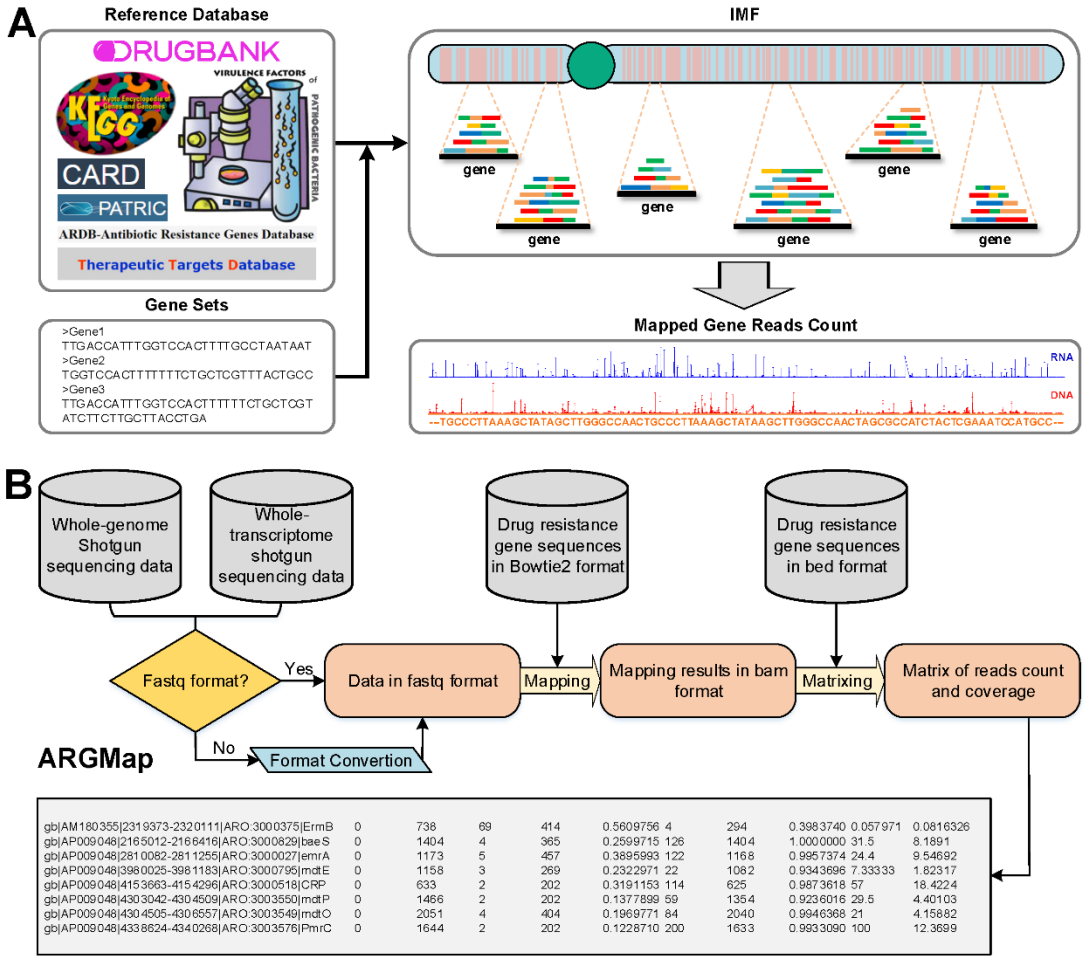


Figure 19: (A) Workflow of the IMF pipeline. IMF utilizes reference databases, e.g., DrugBank, KEGG, CARD, PATRIC, VFDB, ARDB and TTD, to map with input gene sets. It can produce mapped DNA and RNA read counts for each of the given genes, in support of other downstream analyses. (B) Flow chart of pipeline construction of ARGMap. It takes metagenomic or metatranscriptomic sequencing data pair-ended file in fastq format as input files. If the input files are not in fastq format, user should convert them into the fastq format. For example, if the original formats are in BAM format, user should use the function "bamToFastq" in Bedtools to convert them into fastq formats. Our pipeline will download CARD database by default. User will obtain CARD reference database in fasta format in the CARD directory. Then, it will utilize Bowtie2 tool to map between the CARD reference database and input files to generate mapping results in BAM format. Finally, it will use Bedtools to generate read counts tables.

Application of the IMF pipeline with respect to antibiotic resistance genes

resulted in the generation of a process-specific Antibiotic Resistance Gene Mapping

(ARGMap) pipeline (Figure 19B). This particular pipeline integrates antibiotic resistance databases, such as CARD, to analyze input metagenomic and metatranscriptomic data. The databases are used to identify antibiotic resistance genes, which are in turn used to identify the particular expression level and coverage of these genes in the provided data using optimized mapping tools. The analysis pipeline results in a table of read counts and coverage for the respective antibiotic resistance genes for the microbial community of interest. The application of this tool has the potential to greatly impact pharmacogenetic studies. While ARGMap is a pipeline specifically designed for analysis of metagenomic and metatranscriptomic data analysis with respect to antibiotic resistance genes, the IMF pipeline can be used as a framework for any other functional gene sets, such as drug targets, virulence factors, human homologs, among others.

3.2 Applications of Data Analysis in Collaborations

3.2.1 Human Cancer Cells

Analysis of human cancer cells to develop a deeper understanding of the genetic and transcriptomic mechanisms that make cancers so difficult to prevent and treat have been a popular area of interest for a wide variety of researchers [264-271]. One particular method of using RNA-Seq data on cancer samples is to analyze the gene expression differences observed through various treatments. BIO, which is a small molecule inhibitor of the glycogen synthase kinase GSK3 [272], was used to treat HCT116 cells—a colorectal cancer cell line. Of interest in this study was the gene-level differences observed based on the BIO dosage over time, specifically the regulation of the L1 promoter that is prominent in numerous cancer types [273-280].

Two BIO dosages (0.4 µM and 1 µM) were used on the cancerous cells, with samples collected at 6 hours (6h) and 12 hours (12h), with a set of control samples of 0 µM collected at 6 hours. Overall, 10 samples were analyzed: 2 replicates each of control at 6h, 0.4 µM at 6h, 1 µM at 6h, 0.4 µM at 12h, and 1 µM at 12h. To identify transcriptomic differences in the T166 versus the M26 strains, a computational pipeline was used consisting of: (1) read quality check using FastQC [20]; (2) data trimming using Trim Galore! [281]; (3) alignment of trimmed reads to indexed reference genome—collected from the HISAT2 website—using HISAT2 [40]; (4) read count quantification using HTSeq [45]; and (5) differential expression analysis using DESeq2 [46] in R.

Two levels of comparison were made to determine transcriptomic differences from the data: (1) pairwise dosage effect and (2) pairwise time effect. Dosage effects were determined as control vs. 0.4 µM at 6h, control vs. 1 µM at 6h, 0.4 µM vs. 1 µM at 6h, and 0.4 µM vs. 1 µM at 12h. Time effects were determined as 6h vs. 12h of 0.4 µM and 6h vs. 12h of 1 µM. Each comparison was a pairwise analysis using a Wald Test approach, which performs a parametric significance test of the selected factor level using a negative binomial distribution.

DESeq2 compiles a results file of the gene ID, mean expression value, $\log_2$ fold-change & standard error, statistical test value, p-value, and adjusted p-value. DESeq2 adjusts the p-values to account for multiple testing using an FDR method. For this study, genes were considered differentially expressed if their adjusted p-value was below 0.05. To account for significant statistical differences resulting from low sample variances, an additional measure was considered to identify genes that are statistically differentially expressed. This designation requires both a $|\log_2 \text{fold-change}| > 1$ and

adjusted p-value < 0.05. This classification provides genes that have a large fold-change and a statistically significant difference.

Results of the analysis are shown in Table 7. Comparison (1) provides insight into the transcript-level differences based on BIO dosage at both time points. As expected, there were a fair number of transcripts differentially expressed between the control group and 0.4 µM and roughly twice as many between the control and 1 µM. Interestingly, after 6 hours, there were no differentially expressed transcripts between two dosage levels. However, at 12 hours, there were a relatively large number of differentially expressed transcripts between 0.4 µM and 1 µM.

Comparison (2) is a time comparison for the two dosages. 0.4 µM showed a large number of transcripts that are differentially expressed between 6 hours and 12 hours, with far fewer transcripts exhibiting differential expression for the 1 µM dosage.

Table 7: Human HCT116 Cancer Cell Results. RNA-Seq analysis results for differentially expressed transcripts of the HCT116 cancer cell in human based on two dosage levels, two time points, and one control. Up- and down-regulated transcript counts are provided based on $\log_2$ fold-change > 1 and $\log_2$ fold-change < -1, respectively, for transcripts with adjusted p-value < 0.05.

Comparison			**Up-regulated**	**Down-regulated**	**Total**
(1) Dosage	**6h**	**Control vs. 0.4 µM**	34	8	42
		Control vs. 1 µM	69	18	87
		0.4 µM vs. 1 µM	0	0	0
	12h	**0.4 µM vs. 1 µM**	110	143	253
(2) Time	**0.4 µM**	**6h vs. 12h**	306	174	480
	1 µM	**6h vs. 12h**	17	3	20

3.2.2 *Malus domestica*

Malus domestica is the domesticated apple tree and is grown worldwide. This species has resulted from a hybridization between its primary wild ancestor, *Malus sieversii*, the European crab apple, *Malus sylvestris*, and minor contributions from other wild *Malus* species [282, 283]. As a popular crop within the United States, it is of specific interest to the United States Department of Agriculture's Agricultural Research Service (USDA-ARS). Data collected through the USDA Risk Management Agency shows that insured losses for apple crops were $157,177,390. Much of these claimed losses occurred in the spring time. In particular years (2007, 2010, 2012, 2014, 2016, & 2017), spring freezes killed off large amounts of apple crops. Particularly, 2007 and 2017 spring freezes each resulted in $1 billion in losses from all crops [284]. In these scenarios, unseasonably warm temperatures in early spring induce apple trees to exit dormancy and de-acclimate, i.e., lose cold hardiness. Subsequently, low temperature events several weeks later arrive when flowers and early vegetative growth have little to no cold hardiness or frost tolerance [285]. A modified strain of *M. domestica* T166 has been bred to achieve improvements in cold-hardiness. To investigate the genetic processes that may be related to cold hardiness and dormancy, specific apple crop samples were taken, sequenced and analyzed.

The data analyzed consists of 24 datasets, 12 for the M26 wild-type strain and 12 for the T166 transgenic strain of *M. domestica*. Each strain was sampled three times each during February, March, April, and July. After the 24 datasets were sequenced, each was run through an optimized RNA-Seq pipeline to determine statistically significant

differences in gene expression for particular comparisons. The v1.0 reference genome and annotation obtained from Phytozome [286] were used.

To identify genetic differences in the T166 versus the M26 strains, a computational pipeline consisting of optimized tools was developed for this purpose. The pipeline consists of: (1) read quality check using FastQC [20]; (2) data trimming using Btrim [28]; (3) reference genome indexing using HISAT2-build [40]; (4) alignment of trimmed reads to indexed reference genome using HISAT2 [40]; (5) read count quantification using HTSeq [45]; and (6) differential expression analysis using DESeq2 [46] in R.

Four distinct comparisons were considered for differential expression: (1) pairwise comparisons of M26 versus T166 at each time point; (2) time main effect for each M26 and T166 separately; (3) Pairwise comparisons of each consecutive time point for M26 and T166; and (4) Interaction effect of strain and time. The four comparison levels provide a total of 13 comparisons, with comparison (1) being responsible for four, comparison (2) being responsible for two, comparison (3) being responsible for six, and comparison (4) being responsible for one. These four levels of comparisons provide a comprehensive view of the changes in expression corresponding to strain (comparison 1) and time differences (comparisons 2 & 3) and which genes have expression patterns that differ due to strain over the course of the entire study (comparison 4).

The specific results for differential gene expression were developed using DESeq2, which implements a Wald Test or Likelihood Ratio Test to determine which genes have different expressions for the respective comparison. The pairwise comparisons (1 & 3) utilize the Wald Test approach. Significant p-values result from the

factor being determined as significant in the Wald Test. The more complex comparisons (2 & 4) utilize a Likelihood Ratio Test, which compares a full linear model considering appropriate additive and interactive effects and compares the fit against a reduced linear model with the selected factor(s) removed. Significant p-values result from a significant fitted improvement in the full model over the reduced model.

Results of this analysis are shown in Table 8. The four distinct comparisons each provide a different level of information. Comparison (1), M26 vs. T166 by month, gives a direct view of the genetic differences in the strains at particular time points. The comparisons during late winter (February and March) are similar with between one- and two-thousand differentially expressed genes each. The April comparison indicates the highest level of difference between the two strains, with over four-thousand differentially expressed genes. This comparison indicates that particular genetic differences are high at the time when temperatures return to below freezing, which may attribute to the differences observed in crop survival.

Table 8: *Malus domestica* Results. Results for the analysis of *M. domestica* from two distinct strains. Comparisons include strain-strain, time main effect, month-to-month by strain, and time-strain interaction comparisons. Up- and down-regulated gene counts are provided based on $\log_2$ fold-change > 1 and $\log_2$ fold-change < -1, respectively, for genes with adjusted p-value < 0.05.

Comparison			**Up-regulated**	**Down-regulated**	**Total**
(1) M26 vs. T166	**February**		498	1189	1687
	March		734	384	1118
	April		1834	2177	4011
	July		146	56	202
(2) Time Main Effect	**M26**		7075	4394	11469
	T166		7090	4007	11097
(3) Month-to-Month	**M26**	**Feb-Mar**	1004	1809	2813
		Mar-Apr	3399	2336	5735
		Apr-Jul	3906	2043	5949
	T166	**Feb-Mar**	255	275	530
		Mar-Apr	3096	2419	5515

		Apr-Jul	5918	4225	10143
(4) Time-Strain Interaction			1888	895	2783

Comparison (2) tests for the effect of time on expression level over time for each individual strain. Genes being identified as differentially expressed would be any that have expression levels that significantly change over time. The number of differentially expressed genes for each strain are similar (11469 and 11097). This does not mean similar expression patterns over time, just that a similar number of genes have significant changes in expression over this time period, which is expected.

Comparison (3) provides insight into the expression changes from month to month for each strain. One of the most striking differences of this comparison lies in the number of differentially expressed genes in the Feb-Mar comparison for each strain relative to the Apr-Jul comparison. The M26 wild-type strain has higher levels of differentially expressed genes in the earlier months, indicating more changes in genetic expression earlier in the spring. The T166 transgenic strain has relatively low activity in the early spring, while it shows more activity in the changes to genetic expression in early summer. This may indicate particularities contributing to T166's resilience to temperature fluctuations in early spring.

Comparison (4) details which genes have significant interaction effects between strain and month. Any genes that are differentially expressed in this comparison indicate a significant difference in the expression over time between strains. In other words, these genes have different expression patterns, depending on the strain. These genes would be the most important to investigate further, as they have been indicated to differ over the course of time between the M26 wild-type and T166 transgenic strains.

CHAPTER 4: Discussion and Further Research

The developed algorithms and tools discussed in Chapter 2 of this dissertation address particular deficiencies in current RNA-Seq analysis approaches. GeneQC is a first-of-its-kind tool used to analyze the quality of read alignment, particularly with respect to the severity of mapping uncertainty for each annotated gene. This tool provides a method for researchers to evaluate their expression estimates through integration of multi-level features and machine learning approaches. Without evaluation on this level, potential biases may be inserted into analyses, directly affecting the end-stage analyses. If severe issues are detected using GeneQC, the ARM algorithm provides a foundation for re-alignment of ambiguous reads through integration of potential alignment locations collected from GeneQC and co-regulatory networks generated through motif prediction by DMINDA2.0. These co-regulatory networks provide a background distribution for each alignment location, creating a probabilistic method for determining the most likely alignment location. These two tools work in tandem to address the particular issue of mapping uncertainty in modern RNA-Seq data analysis pipelines. However, GeneQC itself can be used to evaluate the quality of read alignment from any alignment tool, and thus has the potential application in comparing and evaluating the performance of read alignment tools.

IRIS-EDA and ViDGER perform a slightly different purpose than GeneQC and ARM, while still aiming to address bottlenecks in RNA-Seq data analysis. While GeneQC and ARM focus on addressing a computational problem (i.e. mapping uncertainty), IRIS-EDA and ViDGER work to improve the usability and interpretability

of analysis tools. IRIS-EDA is a server-based shiny application used for a variety of analyses performed on gene expression estimation data. This tool allows users to provide their expression matrix with some accompanying information related to each sample to conduct various end-stage analyses, including correlation analysis, heatmap generations, principal component analysis, multidimensional scaling, clustering, biclustering, and differential gene expression analysis. In doing so, results are provided in an interactive interface to improve the interpretability of each functionality. To perform these analyses, IRIS-EDA includes methods for analyzing both bulk and single-cell RNA-Seq data, a feature that is currently lacking in all other comparative tools. Additionally, IRIS-EDA integrates a page to assist users in generating requisite metadata for data submission to NCBI's GEO server.

ViDGER, on the other hand, has a much more limited yet highly important purpose: generating high-quality visualizations for interpretation of differential gene expression results. Nine unique visualizations can be generated with ViDGER, including three matrix functionalities that display all possible pairwise comparisons. This tool allows users multiple functionalities to visualize various features of their differential gene expression analysis results from one of the three most highly-cited differential gene expression tools (DESeq2, edgeR, and Cuffdiff). Compatibility with these three tools allows compatibility of ViDGER with over 80% of cited studies involving differential gene expression analysis.

While these tools have current applicability in RNA-Seq data analysis, there are still improvements that could benefit the functionality of each tool. The first important improvement that would achieve more widespread application power is the integrated of

all four methods into a single server-based analysis pipeline following the IRIS framework proposed in Chapter 1. Not all functionalities of the framework are covered between GeneQC, ARM, IRIS-EDA and ViDGER. This would require integrating state-of-the-art tools to fill these gaps. In particular, high-performing tools such as FastQC, Cutadapt, and HISAT2 would be included to provide read-level quality control, data trimming, and reference-based read alignment, respectively. Reference-based and *de-novo* assembly tools would be integrated as well. Following these tools, GeneQC and ARM would fall into Tier 2 for alignment quality control and re-alignment and quantification, respectively. IRIS-EDA and ViDGER belong to Tier 3, covering some aspect of both discovery- and hypothesis-driven analyses. This framework would allow users to perform high-end RNA-Seq analyses with relatively limited computational experience. Following a similar approach as with IRIS-EDA, default parameterizations and tools would allow users to analyze their data almost immediately. Various tools and methods will be provided as alternative options, allowing for user-preferred methods to be implemented as well.

In addition to improved implementations, GeneQC and ARM also have areas for methodological improvements. While the ARM algorithm itself has demonstrated performance, it still has two main places for improvement: (1) improved efficiency and (2) alternative development of gene networks. The efficiency of the ARM algorithm is currently a large concern. Determination of the co-regulatory networks requires manual download of the KEGG pathways, followed by motif prediction. In order for ARM to be widely used, a seamless method for download and prediction of required information is

necessary. This improvement will allow for a fully established computational tool to be released.

The second improvement for the ARM algorithm is further investigation of gene network development methods. Currently, the ARM algorithm can improve the estimation of expression estimates related to mapping uncertainty. However, the limitation of availability of KEGG pathway information for the particular species of interest is required. In certain situations, this information may not be readily available, effectively rendering the ARM algorithm useless in this instance. Thus, more robust methods for determination of gene networks needs to be explored. One particular approach that has more widespread applicability is the use of co-expression networks rather than co-regulatory networks. Co-expression networks can be calculated from either established expression patterns using microarray or RNA-Seq data or from user-provided data. In this scenario, biclustering—in particular the QUBIC biclustering tool—can be used to establish clusters of genes and conditions having similar expression patterns. By properly establishing parameters, co-expression networks can be used to generate the background distributions similarly to how the co-regulatory networks are used. The one main drawback of this approach is the invalidation of some downstream analyses. If co-expression networks are established from user-provided data and then are used for re-alignment, particular biases would affect the interpretability of co-expression analysis results from the data. While this is of concern in certain applications, co-expression networks may provide a method for differential gene expression or other studies. Regardless, a larger analysis of the impact of various gene network generation

methods would improve applicability and potentially the accuracy and reliability of the ARM algorithm.

GeneQC is also not immune from improvements. While the described methods for GeneQC provide useful information, there are many more approaches that could generate more reliable results. In particular, various machine learning algorithms may have applicability in this tool. Approaches like self-organizing maps [287], neural gas [288], and ensemble averaging may provide a better method for predicting the severity of mapping uncertainty, and thus better quality control from expression estimates. In evaluating these methods and the current GeneQC approach, a robust study will be undertaken. First, large-scale simulated data from various species will be generated using Flux Simulator [289]. The data generated will have known true expression values, which can be directly compared to the expression estimates generated from various alignment tools (HISAT2 [40], RSEM [29], kallisto [290], and TopHat [35]). GeneQC will be modified to analyze the alignment results from each of these tools using multiple methods, including the current algorithm, PCA, MDS, self-organizing maps, neural gas, ensemble averaging, among other approaches. Correlation between the generated D-score or categorization—depending upon the generated results from each method—and the difference in true expression and estimated expression will be used to determine the quality of method used for quality control. This approach will allow for a determination of which machine learning method most accurately predicts a significant difference between the true and estimated expressions. Additionally, this analysis will establish which alignment tool generates expression estimates closest to the true expression levels.

The best method or methods will be integrated into a new tool, called GeneQC2.0, to be used for a more robust quality control method.

Current versions of GeneQC, the ARM algorithm, IRIS-EDA, and ViDGER will continue to be used directly and indirectly in applied and computational collaborations, such as those discussed in Chapter 3. These collaborative efforts help to develop areas in need of further improvement and alternative approaches for methods, such as the application of the DMINDA2.0 server in network generation for the ARM algorithm. While these methods have demonstrated applicability to current pipelines, the proposed future directions of each tool will only further their capabilities, in terms of reach and reliability.

REFERENCES

1. Van Dijk, E.L., et al., *Ten years of next-generation sequencing technology.* Trends in genetics, 2014. **30**(9): p. 418-426.
2. Goodwin, S., J.D. McPherson, and W.R. McCombie, *Coming of age: ten years of next-generation sequencing technologies.* Nature Reviews Genetics, 2016. **17**(6): p. 333-351.
3. Marioni, J.C., et al., *RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays.* Genome research, 2008. **18**(9): p. 1509-1517.
4. 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**: p. 154.
5. Nagalakshmi, U., et al., *The transcriptional landscape of the yeast genome defined by RNA sequencing.* Science, 2008. **320**(5881): p. 1344-1349.
6. Wu, X., et al., *Data mining with big data.* IEEE transactions on knowledge and data engineering, 2014. **26**(1): p. 97-107.
7. Swan, M., *The quantified self: Fundamental disruption in big data science and biological discovery.* Big Data, 2013. **1**(2): p. 85-99.
8. Wang, Z., M. Gerstein, and M. Snyder, *RNA-Seq: a revolutionary tool for transcriptomics.* Nat Rev Genet, 2009. **10**(1): p. 57-63.
9. Park, P.J., *ChIP-seq: advantages and challenges of a maturing technology.* Nature Reviews Genetics, 2009. **10**(10): p. 669-680.

10. Collas, P. and J.A. Dahl, *Chop it, ChIP it, check it: the current status of chromatin immunoprecipitation.* Frontiers in Bioscience A Journal & Virtual Library, 2008. **13**(4): p. 929-943.

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

12. Ozsolak, F. and P.M. Milos, *RNA sequencing: advances, challenges and opportunities.* Nature reviews genetics, 2011. **12**(2): p. 87-98.

13. Garber, M., et al., *Computational methods for transcriptome annotation and quantification using RNA-seq.* Nature methods, 2011. **8**(6): p. 469-477.

14. Saliba, A.-E., et al., *Single-cell RNA-seq: advances and future challenges.* Nucleic Acids Research, 2014. **42**(14): p. 8845-8860.

15. Kiselev, V.Y., et al., *SC3: consensus clustering of single-cell RNA-seq data.* Nat Methods, 2017. **14**(5): p. 483-486.

16. Aibar, S., et al., *SCENIC: single-cell regulatory network inference and clustering.* Nat Methods, 2017. **14**(11): p. 1083-1086.

17. Prince, M.E., et al., *Identification of a subpopulation of cells with cancer stem cell properties in head and neck squamous cell carcinoma.* Proc Natl Acad Sci U S A, 2007. **104**(3): p. 973-8.

18. Navin, N., et al., *Tumour evolution inferred by single-cell sequencing.* Nature, 2011. **472**(7341): p. 90-4.

19. Xu, X., et al., *Single-cell exome sequencing reveals single-nucleotide mutation characteristics of a kidney tumor.* Cell, 2012. **148**(5): p. 886-95.

20. Andrews, S., *FastQC: a quality control tool for high throughput sequence data.* 2010.

21. Li, J. and R. Tibshirani, *Finding consistent patterns: a nonparametric approach for identifying differential expression in RNA-Seq data.* Statistical methods in medical research, 2013. **22**(5): p. 519-536.

22. Wang, L., et al., *DEGseq: an R package for identifying differentially expressed genes from RNA-seq data.* Bioinformatics, 2009. **26**(1): p. 136-138.

23. Tarazona, S., et al., *NOIseq: a RNA-seq differential expression method robust for sequencing depth biases.* EMBnet. journal, 2012. **17**(B): p. pp. 18-19.

24. Hardcastle, T.J. and K.A. Kelly, *baySeq: empirical Bayesian methods for identifying differential expression in sequence count data.* BMC bioinformatics, 2010. **11**(1): p. 422.

25. Ritchie, M.E., et al., *limma powers differential expression analyses for RNA-sequencing and microarray studies.* Nucleic acids research, 2015. **43**(7): p. e47-e47.

26. Love, M.I., W. Huber, and S. Anders, *Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.* Genome biology, 2014. **15**(12): p. 550.

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

28. Kong, Y., *Btrim: a fast, lightweight adapter and quality trimming program for next-generation sequencing technologies.* Genomics, 2011. **98**(2): p. 152-153.

29. Li, B. and C.N. Dewey, *RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome.* BMC bioinformatics, 2011. **12**(1): p. 323.

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

31. Wu, T.D., et al., *GMAP and GSNAP for Genomic Sequence Alignment: Enhancements to Speed, Accuracy, and Functionality.* Methods Mol Biol, 2016. **1418**: p. 283-334.

32. Wu, J., et al., *OLego: fast and sensitive mapping of spliced mRNA-Seq reads using small seeds.* Nucleic Acids Res, 2013. **41**(10): p. 5149-63.

33. Wang, K., et al., *MapSplice: accurate mapping of RNA-seq reads for splice junction discovery.* Nucleic Acids Res, 2010. **38**(18): p. e178.

34. Trapnell, C., et al., *Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks.* Nat Protoc, 2012. **7**(3): p. 562-78.

35. Trapnell, C., L. Pachter, and S.L. Salzberg, *TopHat: discovering splice junctions with RNA-Seq.* Bioinformatics, 2009. **25**(9): p. 1105-11.

36. Robinson, M.D., D.J. McCarthy, and G.K. Smyth, *edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.* Bioinformatics, 2010. **26**(1): p. 139-40.

37. Philippe, N., et al., *CRAC: an integrated approach to the analysis of RNA-seq reads.* Genome Biol, 2013. **14**(3): p. R30.

38. Pertea, M., et al., *StringTie enables improved reconstruction of a transcriptome from RNA-seq reads.* Nat Biotechnol, 2015. **33**(3): p. 290-5.

39. Pertea, M., et al., *Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown.* Nat Protoc, 2016. **11**(9): p. 1650-67.

40. Kim, D., B. Langmead, and S.L. Salzberg, *HISAT: a fast spliced aligner with low memory requirements.* Nat Methods, 2015. **12**(4): p. 357-60.

41. Grabherr, M.G., et al., *Full-length transcriptome assembly from RNA-Seq data without a reference genome.* Nat Biotechnol, 2011. **29**(7): p. 644-52.

42. Dobin, A., et al., *STAR: ultrafast universal RNA-seq aligner.* Bioinformatics, 2013. **29**(1): p. 15-21.

43. Chang, Z., et al., *Bridger: a new framework for de novo transcriptome assembly using RNA-seq data.* Genome Biol, 2015. **16**: p. 30.

44. Bonfert, T., et al., *ContextMap 2: fast and accurate context-based RNA-seq mapping.* BMC Bioinformatics, 2015. **16**: p. 122.

45. Anders, S., P.T. Pyl, and W. Huber, *HTSeq--a Python framework to work with high-throughput sequencing data.* Bioinformatics, 2015. **31**(2): p. 166-9.

46. Anders, S. and W. Huber, *Differential expression of RNA-Seq data at the gene level–the DESeq package.* Heidelberg, Germany: European Molecular Biology Laboratory (EMBL), 2012.

47. *NextGENe*.

48. *ERGO 2.0*.

49. *Illumina BaseSpace*.

50. *Strand NGS*.

51. *DNASTAR Lasergene*.

52. Kartashov, A.V. and A. Barski, *BioWardrobe: an integrated platform for analysis of epigenomics and transcriptomics data.* Genome biology, 2015. **16**(1): p. 158.

53. Afgan, E., et al., *The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update.* Nucleic acids research, 2016. **44**(W1): p. W3-W10.

54. Workbench, C.G., *'Version 6.5. 1.* CLC bio A/S Science Park Aarhus Finlandsgade: p. 10-12.

55. Baruzzo, G., et al., *Simulation-based comprehensive benchmarking of RNA-seq aligners.* Nat Methods, 2017. **14**(2): p. 135-139.

56. Sahraeian, S.M.E., et al., *Gaining comprehensive biological insight into the transcriptome by performing a broad-spectrum RNA-seq analysis.* Nature communications, 2017. **8**(1): p. 59.

57. Schadt, E.E., S. Turner, and A. Kasarskis, *A window into third-generation sequencing.* Human molecular genetics, 2010. **19**(R2): p. R227-R240.

58. Liu, L., et al., *Comparison of next-generation sequencing systems.* BioMed Research International, 2012. **2012**.

59. Kong, Y., *Btrim: a fast, lightweight adapter and quality trimming program for next-generation sequencing technologies.* Genomics, 2011. **98**(2): p. 152-3.

60. Gordon, A. and G. Hannon, *Fastx-toolkit. FASTQ/A short-reads pre-processing tools.* Unpublished http://hannonlab. cshl. edu/fastx_toolkit, 2010.

61. Bolger, A.M., M. Lohse, and B. Usadel, *Trimmomatic: a flexible trimmer for Illumina sequence data.* Bioinformatics, 2014. **30**(15): p. 2114-2120.

62. Martin, M., *Cutadapt removes adapter sequences from high-throughput sequencing reads.* EMBnet. journal, 2011. **17**(1): p. pp. 10-12.

63. Li, H. and R. Durbin, *Fast and accurate short read alignment with Burrows–Wheeler transform.* Bioinformatics, 2009. **25**(14): p. 1754-1760.

64. Langmead, B. and S.L. Salzberg, *Fast gapped-read alignment with Bowtie 2.* Nature methods, 2012. **9**(4): p. 357.

65. Langmead, B., *Aligning short sequencing reads with Bowtie.* Current protocols in bioinformatics, 2010: p. 11.7. 1-11.7. 14.

66. Grabherr, M.G., et al., *Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data.* Nature biotechnology, 2011. **29**(7): p. 644.

67. Haas, B.J., et al., *De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis.* Nature protocols, 2013. **8**(8): p. 1494.

68. Hartigan, J.A. and M.A. Wong, *Algorithm AS 136: A k-means clustering algorithm.* Journal of the Royal Statistical Society. Series C (Applied Statistics), 1979. **28**(1): p. 100-108.

69. Johnson, S.C., *Hierarchical clustering schemes.* Psychometrika, 1967. **32**(3): p. 241-254.

70. Hartigan, J.A., *Direct clustering of a data matrix.* Journal of the american statistical association, 1972. **67**(337): p. 123-129.

71. Cheng, Y. and G.M. Church. *Biclustering of expression data*. in *Ismb*. 2000.

72. Ntranos, V., et al., *Fast and accurate single-cell RNA-seq analysis by clustering of transcript-compatibility counts.* Genome biology, 2016. **17**(1): p. 112.

73. Wang, B., et al., *Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning.* Nature methods, 2017. **14**(4): p. 414.

74. Xie, J., et al., *It is time to apply biclustering: a comprehensive review of biclustering applications in biological and biomedical data.* Briefings in bioinformatics, 2018.

75. Spencer, G. *Background on Comparative Genomic Analysis*. 2002; Available from: https://www.genome.gov/10005835/.

76. D'haeseleer, P., *How does DNA sequence motif discovery work?* Nature biotechnology, 2006. **24**(8): p. 959-961.

77. Li, B., et al., *RNA-Seq gene expression estimation with read mapping uncertainty.* Bioinformatics, 2009. **26**(4): p. 493-500.

78. Oshlack, A., M.D. Robinson, and M.D. Young, *From RNA-seq reads to differential expression results.* Genome biology, 2010. **11**(12): p. 220.

79. Zhu, J.-Y., Y. Sun, and Z.-Y. Wang, *Genome-wide identification of transcription factor-binding sites in plants using chromatin immunoprecipitation followed by microarray (ChIP-chip) or sequencing (ChIP-seq),* in *Plant Signalling Networks*. 2011, Springer. p. 173-188.

80. Network, C.G.A.R., *Comprehensive genomic characterization defines human glioblastoma genes and core pathways.* Nature, 2008. **455**(7216): p. 1061.

81. Cho, B.-K., et al., *The transcription unit architecture of the Escherichia coli genome.* Nature biotechnology, 2009. **27**(11): p. 1043.

82. Albrecht, M., et al., *Deep sequencing-based discovery of the Chlamydia trachomatis transcriptome.* Nucleic acids research, 2009. **38**(3): p. 868-877.

83. Yoder-Himes, D., et al., *Mapping the Burkholderia cenocepacia niche response via high-throughput sequencing.* Proceedings of the National Academy of Sciences, 2009. **106**(10): p. 3976-3981.

84. Consortium, I.W.G.S., *A chromosome-based draft sequence of the hexaploid bread wheat (Triticum aestivum) genome.* Science, 2014. **345**(6194): p. 1251788.

85. Coordinators, N.R., *Database resources of the national center for biotechnology information.* Nucleic acids research, 2016. **44**(Database issue): p. D7.

86. Nordberg, H., et al., *The genome portal of the Department of Energy Joint Genome Institute: 2014 updates.* Nucleic acids research, 2013. **42**(D1): p. D26-D31.

87. Kahles, A., J. Behr, and G. Rätsch, *MMR: a tool for read multi-mapper resolution.* Bioinformatics, 2015. **32**(5): p. 770-772.

88. McDermaid, A., et al., *GeneQC: A quality control tool for gene expression estimation based on RNA-sequencing reads mapping.* bioRxiv, 2018: p. 266445.

89. Altschul, S.F., et al., *Basic local alignment search tool.* Journal of molecular biology, 1990. **215**(3): p. 403-410.

90. Dempster, A.P., M. Schatzoff, and N. Wermuth, *A simulation study of alternatives to ordinary least squares.* Journal of the American Statistical Association, 1977. **72**(357): p. 77-91.

91. Hoerl, A.E. and R.W. Kennard, *Ridge regression: Biased estimation for nonorthogonal problems.* Technometrics, 1970. **12**(1): p. 55-67.

92. Zou, H. and T. Hastie, *Regularization and variable selection via the elastic net.* Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2005. **67**(2): p. 301-320.

93. Tibshirani, R., *Regression shrinkage and selection via the lasso.* Journal of the Royal Statistical Society. Series B (Methodological), 1996: p. 267-288.

94. D'Haeseleer, P., *What are DNA sequence motifs?* Nature Biotechnology, 2006. **24**(4): p. 423-5.

95. Baumbach, J., *On the power and limits of evolutionary conservation—unraveling bacterial gene regulatory networks.* Nucleic acids research, 2010. **38**(22): p. 7877-7884.

96. Davidson, E. and M. Levin, *Gene regulatory networks*. 2005, National Acad Sciences.

97. Liu, B., et al., *Bacterial regulon modeling and prediction based on systematic cis regulatory motif analyses.* Scientific reports, 2016. **6**: p. 23030.

98. Brohée, S., et al., *Unraveling networks of co-regulated genes on the sole basis of genome sequences.* Nucleic acids research, 2011. **39**(15): p. 6340-6358.

99. Kanehisa, M. and S. Goto, *KEGG: kyoto encyclopedia of genes and genomes.* Nucleic acids research, 2000. **28**(1): p. 27-30.

100. Yang, J., et al., *DMINDA 2.0: integrated and systematic views of regulatory DNA motif identification and analyses.* Bioinformatics, 2017. **33**(16): p. 2586-2588.

101. Li, G., et al., *A new framework for identifying cis-regulatory motifs in prokaryotes.* Nucleic acids research, 2010. **39**(7): p. e42-e42.

102. Anders, S. and W. Huber, *Differential expression analysis for sequence count data.* Genome Biol, 2010. **11**(10): p. R106.

103. Love, M.I., W. Huber, and S. Anders, *Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.* Genome Biol, 2014. **15**(12): p. 550.

104. Fisher, R.A., *The negative binomial distribution.* Annals of Human Genetics, 1941. **11**(1): p. 182-187.

105. Savani, V. and A.A. Zhigljavsky, *Efficient estimation of parameters of the negative binomial distribution.* Communications in Statistics—Theory and Methods, 2006. **35**(5): p. 767-783.

106. Perkel, J.M., *Data visualization tools drive interactivity and reproducibility in online publishing.* Nature, 2018. **554**(7690): p. 133-134.

107. Kruskal, J.B., *Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis.* Psychometrika, 1964. **29**(1): p. 1-27.

108. Abdi, H. and L.J. Williams, *Principal component analysis.* Wiley interdisciplinary reviews: computational statistics, 2010. **2**(4): p. 433-459.

109. Saelens, W., R. Cannoodt, and Y. Saeys, *A comprehensive evaluation of module detection methods for gene expression data.* Nature communications, 2018. **9**(1): p. 1090.

110. Eisen, M.B., et al., *Cluster analysis and display of genome-wide expression patterns.* Proceedings of the National Academy of Sciences, 1998. **95**(25): p. 14863-14868.

111. Zhang, Y., et al., *QUBIC: a bioconductor package for qualitative biclustering analysis of gene co-expression data.* Bioinformatics, 2016. **33**(3): p. 450-452.

112. Li, G., et al., *QUBIC: a qualitative biclustering algorithm for analyses of gene expression data.* Nucleic acids research, 2009. **37**(15): p. e101-e101.

113. Ritchie, M.E., et al., *limma powers differential expression analyses for RNA-sequencing and microarray studies.* Nucleic Acids Res, 2015. **43**(7): p. e47.

114. Pimentel, H., et al., *Differential analysis of RNA-Seq incorporating quantification uncertainty.* Nature Methods, 2017.

115. Nelson, J.W., et al., *The START App: a web-based RNAseq analysis and visualization resource.* Bioinformatics, 2017. **33**(3): p. 447-449.

116. Powell, D. *Degust: Visualize, explore and appreciate RNA-seq differential gene-expression data*. in *COMBINE RNA-seq workshop*. 2015.

117. Goff, L., C. Trapnell, and D. Kelley, *cummeRbund: Analysis, exploration, manipulation, and visualization of Cufflinks high-throughput sequencing data.* R package version, 2013. **2**(0).

118. Harshbarger, J., A. Kratz, and P. Carninci, *DEIVA: a web application for interactive visual analysis of differential gene expression profiles.* BMC genomics, 2017. **18**(1): p. 47.

119. Younesy, H., et al., *VisRseq: R-based visual framework for analysis of sequencing data.* BMC bioinformatics, 2015. **16**(11): p. S2.

120. McDermaid, A., et al., *ViDGER: An R package for integrative interpretation of differential gene expression results of RNA-seq data.* bioRxiv, 2018.

121. Ge, S.X., *iDEP: An integrated web application for differential expression and pathway analysis.* bioRxiv, 2017.

122. Nueda, M.J., et al., *Identification and visualization of differential isoform expression in RNA-seq time series.* Bioinformatics, 2017.

123. Wilkinson, M.D., et al., *The FAIR Guiding Principles for scientific data management and stewardship.* Scientific data, 2016. **3**.

124. Barrett, T., et al., *NCBI GEO: archive for functional genomics data sets—update.* Nucleic acids research, 2012. **41**(D1): p. D991-D995.

125. McCarthy, D.J., et al., *Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R.* Bioinformatics, 2017. **33**(8): p. 1179-1186.

126. Soneson, C. and M.D. Robinson, *Bias, robustness and scalability in single-cell differential expression analysis.* Nature methods, 2018.

127. Ye, C., et al., *DBG2OLC: efficient assembly of large genomes using long erroneous reads of the third generation sequencing technologies.* Scientific reports, 2016. **6**: p. 31900.

128. Goodwin, S., et al., *Oxford Nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome.* Genome Res, 2015. **25**(11): p. 1750-6.

129. Seyednasrollah, F., A. Laiho, and L.L. Elo, *Comparison of software packages for detecting differential expression in RNA-seq studies.* Briefings in bioinformatics, 2013. **16**(1): p. 59-70.

130. Kvam, V.M., P. Liu, and Y. Si, *A comparison of statistical methods for detecting differentially expressed genes from RNA-seq data.* American journal of botany, 2012. **99**(2): p. 248-256.

131. Goff L, T.C.a.K.D., *cummeRbund: Analysis, exploration, manipulation, and visualization of Cufflinks high-throughput sequencing data.* 2013.

132. Gentleman, R.C., et al., *Bioconductor: open software development for computational biology and bioinformatics.* Genome biology, 2004. **5**(10): p. R80.

133. McDermaid, A., et al., *ViDGER: An R package for integrative interpretation of differential gene expression results of RNA-seq data.* bioRxiv, 2018: p. 268896.

134. Wickham, H., *ggplot2: elegant graphics for data analysis*. 2016: Springer.

135. Schloerke, B., et al., *Ggally: Extension to ggplot2*. 2011.

136. Wickham, H. and R. Francois, *dplyr: A grammar of data manipulation.* R package version 0.4, 2015. **1**: p. 20.

137. Wickham, H., *tidyr: Easily Tidy Data with spread () and gather () Functions.* R package version 0.2. 0, 2014.

138. Huber, W.R.A., *pasilla: Data package with per-exon and per-gene read counts of RNA-seq samples of Pasilla knock-down by Brooks et al., Genome Research 2011*. 2017.

139. Brohee, S., et al., *Unraveling networks of co-regulated genes on the sole basis of genome sequences.* Nucleic Acids Res, 2011. **39**(15): p. 6340-58.

140. Davidson, E. and M. Levin, *Gene regulatory networks.* Proc Natl Acad Sci U S A, 2005. **102**(14): p. 4935.

141. Liu, B., et al., *Bacterial regulon modeling and prediction based on systematic cis regulatory motif analyses.* Scientific reports, 2016. **6**.

142. Baumbach, J., *On the power and limits of evolutionary conservation--unraveling bacterial gene regulatory networks.* Nucleic Acids Res, 2010. **38**(22): p. 7877-84.

143. Liu, B., et al., *An algorithmic perspective of de novo cis-regulatory motif finding based on ChIP-seq data.* Briefings in bioinformatics, 2017: p. bbx026.

144. Lawrence, C.E., et al., *Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment.* Science, 1993. **262**(5131): p. 208-14.

145. Pevzner, P.A. and S.H. Sze, *Combinatorial approaches to finding subtle signals in DNA sequences.* Proc Int Conf Intell Syst Mol Biol, 2000. **8**: p. 269-78.

146. Nakaki, R., J. Kang, and M. Tateno, *A novel ab initio identification system of transcriptional regulation motifs in genome DNA sequences based on direct comparison scheme of signal/noise distributions.* Nucleic Acids Res, 2012. **40**(18): p. 8835-48.

147. Zambelli, F., G. Pesole, and G. Pavesi, *Motif discovery and transcription factor binding sites before and after the next-generation sequencing era.* Brief Bioinform, 2013. **14**(2): p. 225-37.

148. Li, G., et al., *A new framework for identifying cis-regulatory motifs in prokaryotes.* Nucleic Acids Res, 2011. **39**(7): p. e42.

149. Chen, X., et al., *W-AlignACE: an improved Gibbs sampling algorithm based on more accurate position weight matrices learned from sequence and gene expression/ChIP-chip data.* Bioinformatics, 2008. **24**(9): p. 1121-8.

150. Sinha, S., *PhyME: a software tool for finding motifs in sets of orthologous sequences.* Methods Mol Biol, 2007. **395**: p. 309-18.

151. Das, M.K. and H.K. Dai, *A survey of DNA motif finding algorithms.* BMC Bioinformatics, 2007. **8 Suppl 7**: p. S21.

152. Wang, T. and G.D. Stormo, *Combining phylogenetic data with co-regulated genes to identify regulatory motifs.* Bioinformatics, 2003. **19**(18): p. 2369-80.

153. Liu, X., D.L. Brutlag, and J.S. Liu, *BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes.* Pac Symp Biocomput, 2001: p. 127-38.

154. Hertz, G.Z. and G.D. Stormo, *Identifying DNA and protein patterns with statistically significant alignments of multiple sequences.* Bioinformatics, 1999. **15**(7-8): p. 563-77.

155. Liu, X.S., D.L. Brutlag, and J.S. Liu, *An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments.* Nat Biotechnol, 2002. **20**(8): p. 835-9.

156. Bailey, T.L., et al., *MEME SUITE: tools for motif discovery and searching.* Nucleic Acids Res, 2009. **37**(Web Server issue): p. W202-8.

157. Olman, V., D. Xu, and Y. Xu, *CUBIC: identification of regulatory binding sites through data clustering.* J Bioinform Comput Biol, 2003. **1**(1): p. 21-40.

158. Li, X. and W.H. Wong, *Sampling motifs on phylogenetic trees.* Proc Natl Acad Sci U S A, 2005. **102**(27): p. 9481-6.

159. Blanchette, M. and M. Tompa, *Discovery of regulatory elements by a computational method for phylogenetic footprinting.* Genome Res, 2002. **12**(5): p. 739-48.

160. Blanchette, M. and M. Tompa, *FootPrinter: A program designed for phylogenetic footprinting.* Nucleic Acids Res, 2003. **31**(13): p. 3840-2.

161. Li, G., B. Liu, and Y. Xu, *Accurate recognition of cis-regulatory motifs with the correct lengths in prokaryotic genomes.* Nucleic Acids Res, 2010. **38**(2): p. e12.

162. Ma, Q., et al., *An integrated toolkit for accurate prediction and analysis of cis-regulatory motifs at a genome scale.* Bioinformatics, 2013. **29**(18): p. 2261-8.

163. Tompa, M., et al., *Assessing computational tools for the discovery of transcription factor binding sites.* Nat Biotechnol, 2005. **23**(1): p. 137-44.

164. McCue, L.A., et al., *Factors influencing the identification of transcription factor binding sites by cross-species comparison.* Genome Res, 2002. **12**(10): p. 1523-32.

165. Simcha, D., N.D. Price, and D. Geman, *The limits of de novo DNA motif discovery.* PLoS One, 2012. **7**(11): p. e47836.

166. Katara, P., A. Grover, and V. Sharma, *Phylogenetic footprinting: a boost for microbial regulatory genomics.* Protoplasma, 2012. **249**(4): p. 901-7.

167. Tagle, D.A., et al., *Embryonic epsilon and gamma globin genes of a prosimian primate (Galago crassicaudatus). Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints.* J Mol Biol, 1988. **203**(2): p. 439-55.

168. Siddharthan, R., E.D. Siggia, and E. van Nimwegen, *PhyloGibbs: a Gibbs sampling motif finder that incorporates phylogeny.* PLoS Comput Biol, 2005. **1**(7): p. e67.

169. Blanchette, M., B. Schwikowski, and M. Tompa, *Algorithms for phylogenetic footprinting.* J Comput Biol, 2002. **9**(2): p. 211-23.

170. Neph, S. and M. Tompa, *MicroFootPrinter: a tool for phylogenetic footprinting in prokaryotic genomes.* Nucleic Acids Res, 2006. **34**(Web Server issue): p. W366-8.

171. Carmack, C.S., et al., *PhyloScan: identification of transcription factor binding sites using cross-species evidence.* Algorithms Mol Biol, 2007. **2**: p. 1.

172. Zhang, S., et al., *Genome-wide de novo prediction of cis-regulatory binding sites in prokaryotes.* Nucleic Acids Res, 2009. **37**(10): p. e72.

173. Borneman, A.R., et al., *Divergence of transcription factor binding sites across related yeast species.* Science, 2007. **317**(5839): p. 815-819.

174. Odom, D.T., et al., *Tissue-specific transcriptional regulation has diverged significantly between human and mouse.* Nature genetics, 2007. **39**(6): p. 730-732.

175. Boyle, A.P., et al., *Comparative analysis of regulatory information and circuits across distant species.* Nature, 2014. **512**(7515): p. 453-456.

176. Lingyun Song, G.E.C., *DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells.* Cold Spring Harbor Protocols, 2010. **2010**(2).

177. Wang, Z., et al., *RNA-Seq: a revolutionary tool for transcriptomics.* Nature Reviews Genetics, 2008. **10**(1): p. 57-63.

178. Tsankov, A.M., et al., *Transcription factor binding dynamics during human ES cell differentiation.* Nature, 2015. **518**(7539): p. 344-9.

179. Wu, F., B.G. Olson, and J. Yao, *DamID-seq: Genome-wide Mapping of Protein-DNA Interactions by High Throughput Sequencing of Adenine-methylated DNA Fragments.* Journal of Visualized Experiments Jove, 2015(107).

180. Maragkakis, M., et al., *CLIPSeqTools-a novel bioinformatics CLIP-seq analysis suite.* RNA (New York, N.Y.), 2015. **22**(1).

181. Hafner, M., et al., *PAR-CliP - A Method to Identify Transcriptome-wide the Binding Sites of RNA Binding Proteins.* Journal of Visualized Experiments, 2010. **41**(41): p. e2034-e2034.

182. Ingolia, N.T., *Ribosome profiling: new views of translation, from single codons to genome scale.* Nature Reviews Genetics, 2014. **15**(3): p. 205-13.

183. Giresi, P.G., et al., *FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin.* Genome Research, 2007. **17**(6): p. 877-85.

184. Nutiu, R., et al., *Direct measurement of DNA affinity landscapes on a high-throughput sequencing instrument.* Nature Biotechnology, 2011. **29**(7): p. 659-64.

185. Kharchenko, P.V., M.Y. Tolstorukov, and P.J. Park, *Design and analysis of ChIP-seq experiments for DNA-binding proteins.* Nature biotechnology, 2008. **26**(12): p. 1351-1359.

186. Valouev, A., et al., *Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data.* Nature Methods, 2008. **5**(9): p. 829-834.

187. Kuan, P.F., et al., *A statistical framework for the analysis of ChIP-Seq data.* Journal of the American Statistical Association, 2011. **106**(495): p. 891-903.

188. Mathelier, A. and W.W. Wasserman, *The next generation of transcription factor binding site prediction.* PLoS Comput Biol, 2013. **9**(9): p. e1003214.

189. Cheng, C., R. Min, and M. Gerstein, *TIP: a probabilistic method for identifying transcription factor target genes from ChIP-seq binding profiles.* Bioinformatics, 2011. **27**(23): p. 3221-3227.

190. Wu, S., et al., *ChIP-PaM: an algorithm to identify protein-DNA interaction using ChIP-Seq data.* Theoretical biology and medical modelling, 2010. **7**(1): p. 1.

191. van Heeringen, S.J. and G.J.C. Veenstra, *GimmeMotifs: a de novo motif prediction pipeline for ChIP-sequencing experiments.* Bioinformatics, 2011. **27**(2): p. 270-271.

192. Machanick, P. and T.L. Bailey, *MEME-ChIP: motif analysis of large DNA datasets.* Bioinformatics, 2011. **27**(12): p. 1696-7.

193. Kulakovskiy, I.V., et al., *Deep and wide digging for binding motifs in ChIP-Seq data.* Bioinformatics, 2010. **26**(20): p. 2622-3.

194. Jothi, R., et al., *Genome-wide identification of in vivo protein–DNA binding sites from ChIP-Seq data.* Nucleic acids research, 2008. **36**(16): p. 5221-5231.

195. Mercier, E., et al., *An integrated pipeline for the genome-wide analysis of transcription factor binding sites from ChIP-Seq.* PLoS One, 2011. **6**(2): p. e16432.

196. Hu, M., et al., *On the detection and refinement of transcription factor binding sites using ChIP-Seq data.* Nucleic Acids Res, 2010. **38**(7): p. 2154-67.

197. Bailey, T.L., *DREME: motif discovery in transcription factor ChIP-seq data.* Bioinformatics, 2011. **27**(12): p. 1653-9.

198. Zhang, Y., et al., *Model-based analysis of ChIP-Seq (MACS).* Genome Biol, 2008. **9**(9): p. R137.

199. Jiang, H., et al., *CisGenome Browser: a flexible tool for genomic data visualization.* Bioinformatics, 2010. **26**(14): p. 1781-2.

200. Fejes, A.P., et al., *FindPeaks 3.1: a tool for identifying areas of enrichment from massively parallel short-read sequencing technology.* Bioinformatics, 2008. **24**(15): p. 1729-30.

201. Xin, F., R. Grossman, and L. Stein, *PeakRanger: a cloud-enabled peak caller for ChIP-seq data.* Bmc Bioinformatics, 2011. **12**(10): p. 139-139.

202. Jia, C., et al., *A new exhaustive method and strategy for finding motifs in ChIP-enriched regions.* Plos One, 2014. **9**(9): p. e86044.

203. Thomas-Chollier, M., et al., *RSAT peak-motifs: motif analysis in full-size ChIP-seq datasets.* Nucleic Acids Res, 2012. **40**(4): p. e31.

204. Jun Ding, H.H., Xiaoman Li, *SIOMICS: a novel approach for systematic identification of motifs in ChIP-seq data.* Nucleic Acids Research, 2014. **42**(5): p. 1635-1645.

205. Ding, J., et al., *Systematic discovery of cofactor motifs from ChIP-seq data by SIOMICS.* Methods, 2015. **79-80**: p. 47-51.

206. Maaskola, J. and N. Rajewsky, *Binding site discovery from nucleic acid sequences by discriminative learning of hidden Markov models.* Nucleic Acids Res, 2014. **42**(21): p. 12995-3011.

207. Ma, Q., et al., *DMINDA: an integrated web server for DNA motif identification and analyses.* Nucleic acids research, 2014. **42**(W1): p. W12-W19.

208. Liu, B., et al., *An integrative and applicable phylogenetic footprinting framework for cis-regulatory motifs identification in prokaryotic genomes.* BMC genomics, 2016. **17**(1): p. 578.

209. Bailey, T.L., et al., *MEME SUITE: tools for motif discovery and searching.* Nucleic acids research, 2009: p. gkp335.

210. Carvalho, A.L., et al., *Metabolic and transcriptional analysis of acid stress in Lactococcus lactis, with a focus on the kinetics of lactic acid pools.* PLOS ONE, 2013. **8**(7): p. e68470.

211. Locke, J.C.W., et al., *Stochastic pulse regulation in bacterial stress response.* Science, 2011. **334**(6054): p. 366-369.

212. Levine, J.H., Y. Lin, and M.B. Elowitz, *Functional roles of pulsing in genetic circuits.* Science, 2013. **342**(6163): p. 1193-1200.

213. Arnoldini, M., et al., *Evolution of Stress Response in the Face of Unreliable Environmental Signals.* PLoS Comput Biol, 2012. **8**(8): p. e1002627.

214. Kumka, J.E. and C.E. Bauer, *Analysis of the FnrL regulon in Rhodobacter capsulatus reveals limited regulon overlap with orthologues from Rhodobacter sphaeroides and Escherichia coli.* BMC Genomics, 2015. **16**: p. 895.

215. Tan, K., et al., *A comparative genomics approach to prediction of new members of regulons.* Genome Res, 2001. **11**(4): p. 566-84.

216. Gupta, S., et al., *Quantifying similarity between motifs.* Genome Biol, 2007. **8**(2): p. R24.

217. Liu, B., et al., *An integrative and applicable phylogenetic footprinting framework for cis-regulatory motifs identification in prokaryotic genomes.* BMC Genomics, 2016. **17**: p. 578.

218. Chen, X., et al., *RECTA: Regulon Identification Based on Comparative Genomics and Transcriptomics Analysis.* bioRxiv, 2018: p. 261453.

219. Mao, X., et al., *DOOR 2.0: presenting operons and their functions through dynamic and integrated views.* Nucleic acids research, 2013. **42**(D1): p. D654-D659.

220. Bailey, T.L., et al., *MEME Suite: tools for motif discovery and searching.* Nucleic Acids Research, 2009. **37**(suppl 2): p. W202-W208.

221. Ma, Y., et al., *Oral administration of recombinant Lactococcus lactis expressing HSP65 and tandemly repeated P277 reduces the incidence of type I diabetes in non-obese diabetic mice.* PLoS One, 2014. **9**(8): p. e105701.

222. Ramasamy, R., et al., *Immunogenicity of a malaria parasite antigen displayed by Lactococcus lactis in oral immunisations.* Vaccine, 2006. **24**(18): p. 3900-8.

223. Bermudez-Humaran, L.G., et al., *A novel mucosal vaccine based on live Lactococci expressing E7 antigen and IL-12 induces systemic and mucosal immune responses and protects mice against human papillomavirus type 16-induced tumors.* J Immunol, 2005. **175**(11): p. 7297-302.

224. Zhang, B., et al., *Recombinant Lactococcus lactis NZ9000 secretes a bioactive kisspeptin that inhibits proliferation and migration of human colon carcinoma HT-29 cells.* Microb Cell Fact, 2016. **15**(1): p. 102.

225. Hanniffy, S.B., et al., *Mucosal delivery of a pneumococcal vaccine using Lactococcus lactis affords protection against respiratory infection.* J Infect Dis, 2007. **195**(2): p. 185-93.

226. Hols, P., et al., *Conversion of Lactococcus lactis from homolactic to homoalanine fermentation through metabolic engineering.* Nat Biotechnol, 1999. **17**(6): p. 588-92.

227. Hutkins, R.W. and N.L. Nannen, *pH homeostasis in lactic acid bacteria.* Journal of Dairy Science, 1993. **76**(8): p. 2354-2365.

228. van de Guchte, M., et al., *Stress responses in lactic acid bacteria.* Antonie van Leeuwenhoek, 2002. **82**(1): p. 187-216.

229. Hauryliuk, V., et al., *Recent functional insights into the role of (p)ppGpp in bacterial physiology.* Nat Rev Micro, 2015. **13**(5): p. 298-309.

230. Rallu, F., et al., *Acid- and multistress-resistant mutants of Lactococcus lactis : identification of intracellular stress signals.* Molecular Microbiology, 2000. **35**(3): p. 517-528.

231. Benson, D.A., et al., *GenBank.* Nucleic acids research, 2012. **41**(D1): p. D36-D42.

232. Shade, A., et al., *Fundamentals of microbial community resistance and resilience.* Front Microbiol, 2012. **3**: p. 417.

233. Qin, J., et al., *A human gut microbial gene catalogue established by metagenomic sequencing.* Nature, 2010. **464**(7285): p. 59-65.

234. Turnbaugh, P.J., et al., *A core gut microbiome in obese and lean twins.* Nature, 2009. **457**(7228): p. 480-4.

235. Human Microbiome Project, C., *Structure, function and diversity of the healthy human microbiome.* Nature, 2012. **486**(7402): p. 207-14.

236. Human Microbiome Jumpstart Reference Strains, C., et al., *A catalog of reference genomes from the human microbiome.* Science, 2010. **328**(5981): p. 994-9.

237. Aagaard, K., et al., *The Human Microbiome Project strategy for comprehensive sampling of the human microbiome and why it matters.* FASEB J, 2013. **27**(3): p. 1012-22.

238. Integrative, H.M.P.R.N.C., *The Integrative Human Microbiome Project: dynamic analysis of microbiome-host omics profiles during periods of human health and disease.* Cell Host Microbe, 2014. **16**(3): p. 276-89.

239. Larsen, P.E., D. Field, and J.A. Gilbert, *Predicting bacterial community assemblages using an artificial neural network approach.* Nat Methods, 2012. **9**(6): p. 621-5.

240. Handelsman, J., *Metagenomics: application of genomics to uncultured microorganisms.* Microbiol Mol Biol Rev, 2004. **68**(4): p. 669-85.

241. Riesenfeld, C.S., P.D. Schloss, and J. Handelsman, *Metagenomics: genomic analysis of microbial communities.* Annu Rev Genet, 2004. **38**: p. 525-52.

242. Streit, W.R. and R.A. Schmitz, *Metagenomics--the key to the uncultured microbes.* Curr Opin Microbiol, 2004. **7**(5): p. 492-8.

243. Handelsman, J., et al., *Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products.* Chem Biol, 1998. **5**(10): p. R245-9.

244. Meyer, F., et al., *The metagenomics RAST server–a public resource for the automatic phylogenetic and functional analysis of metagenomes.* BMC bioinformatics, 2008. **9**(1): p. 386.

245. Silva, G.G.Z., et al., *SUPER-FOCUS: a tool for agile functional analysis of shotgun metagenomic data.* Bioinformatics, 2015. **32**(3): p. 354-361.

246. Silva, G.G.Z., et al., *FOCUS: an alignment-free model to identify organisms in metagenomes using non-negative least squares.* PeerJ, 2014. **2**: p. e425.

247. Ounit, R., et al., *CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers.* BMC genomics, 2015. **16**(1): p. 236.

248. Wood, D.E. and S.L. Salzberg, *Kraken: ultrafast metagenomic sequence classification using exact alignments.* Genome biology, 2014. **15**(3): p. R46.

249. Truong, D.T., et al., *MetaPhlAn2 for enhanced metagenomic taxonomic profiling.* Nature methods, 2015. **12**(10): p. 902.

250. Truong, D.T., et al., *Microbial strain-level population structure and genetic diversity from metagenomes.* Genome research, 2017. **27**(4): p. 626-638.

251. Cleary, B., et al., *Detection of low-abundance bacterial strains in metagenomic datasets by eigengenome partitioning.* Nature biotechnology, 2015. **33**(10): p. 1053.

252. Ahn, T.-H., J. Chai, and C. Pan, *Sigma: strain-level inference of genomes from metagenomic analysis for biosurveillance.* Bioinformatics, 2014. **31**(2): p. 170-177.

253. Luo, C., et al., *ConStrains identifies microbial strains in metagenomic datasets.* Nature biotechnology, 2015. **33**(10): p. 1045.

254. Scholz, M., et al., *Strain-level microbial epidemiology and population genomics from shotgun metagenomics.* Nature methods, 2016. **13**(5): p. 435.

255. Martinez, X., et al., *MetaTrans: an open-source pipeline for metatranscriptomics.* Sci Rep, 2016. **6**: p. 26447.

256. Westreich, S.T., et al., *SAMSA: A comprehensive metatranscriptome analysis pipeline*. bioRxiv, 2016.

257. Meyer, F., et al., *The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes.* BMC Bioinformatics, 2008. **9**: p. 386.

258. Abubucker, S., et al., *Metabolic reconstruction for metagenomic data and its application to the human microbiome.* PLoS Comput Biol, 2012. **8**(6): p. e1002358.

259. Leimena, M.M., et al., *A comprehensive metatranscriptome analysis pipeline and its validation using human small intestine microbiota datasets.* BMC genomics, 2013. **14**(1): p. 530.

260. Niu, S.-Y., et al., *Bioinformatics tools for quantitative and functional metagenome and metatranscriptome data analysis in microbes.* Briefings in bioinformatics, 2017.

261. McArthur, A.G., et al., *The comprehensive antibiotic resistance database.* Antimicrobial agents and chemotherapy, 2013. **57**(7): p. 3348-3357.

262. Liu, B. and M. Pop, *ARDB—antibiotic resistance genes database.* Nucleic acids research, 2008. **37**(suppl_1): p. D443-D447.

263. Wishart, D.S., et al., *DrugBank: a comprehensive resource for in silico drug discovery and exploration.* Nucleic acids research, 2006. **34**(suppl_1): p. D668-D672.

264. Li, Y., et al., *Circular RNA is enriched and stable in exosomes: a promising biomarker for cancer diagnosis.* Cell research, 2015. **25**(8): p. 981.

265. Costa, V., et al., *RNA-Seq and human complex diseases: recent accomplishments and future perspectives.* European Journal of Human Genetics, 2013. **21**(2): p. 134.

266. Khoury, J.D., et al., *The landscape of DNA virus associations across human malignant cancers using RNA-Seq: an analysis of 3775 cases.* Journal of virology, 2013: p. JVI. 00340-13.

267. McPherson, A., et al., *deFuse: an algorithm for gene fusion discovery in tumor RNA-Seq data.* PLoS computational biology, 2011. **7**(5): p. e1001138.

268. Ren, S., et al., *RNA-seq analysis of prostate cancer in the Chinese population identifies recurrent gene fusions, cancer-associated long noncoding RNAs and aberrant alternative splicings.* Cell research, 2012. **22**(5): p. 806.

269. Nagaraj, N., et al., *Deep proteome and transcriptome mapping of a human cancer cell line.* Molecular systems biology, 2011. **7**(1): p. 548.

270. Alexandrov, L.B., et al., *Signatures of mutational processes in human cancer.* Nature, 2013. **500**(7463): p. 415.

271. Patel, A.P., et al., *Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma.* Science, 2014: p. 1254257.

272. Meijer, L., et al., *GSK-3-selective inhibitors derived from Tyrian purple indirubins.* Chem Biol, 2003. **10**(12): p. 1255-66.

273. Lee, E., et al., *Landscape of somatic retrotransposition in human cancers.* Science, 2012. **337**(6097): p. 967-71.

274. Solyom, S., et al., *Extensive somatic L1 retrotransposition in colorectal tumors.* Genome Res, 2012. **22**(12): p. 2328-38.

275. Iskow, R.C., et al., *Natural mutagenesis of human genomes by endogenous retrotransposons.* Cell, 2010. **141**(7): p. 1253-61.

276. Shukla, R., et al., *Endogenous retrotransposition activates oncogenic pathways in hepatocellular carcinoma.* Cell, 2013. **153**(1): p. 101-11.

277. Helman, E., et al., *Somatic retrotransposition in human cancer revealed by whole-genome and exome sequencing.* Genome Res, 2014. **24**(7): p. 1053-63.

278. Tubio, J.M., et al., *Mobile DNA in cancer. Extensive transduction of nonrepetitive DNA mediated by L1 retrotransposition in cancer genomes.* Science, 2014. **345**(6196): p. 1251343.

279. Rodic, N., et al., *Retrotransposon insertions in the clonal evolution of pancreatic ductal adenocarcinoma.* Nat Med, 2015. **21**(9): p. 1060-4.

280. Ewing, A.D., et al., *Widespread somatic L1 retrotransposition occurs early during gastrointestinal cancer evolution.* Genome Res, 2015. **25**(10): p. 1536-45.

281. Krueger, F., *Trim Galore!: A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files*. 2015.

282. Velasco, R., et al., *The genome of the domesticated apple (Malus× domestica Borkh.).* Nature genetics, 2010. **42**(10): p. 833.

283. Cornille, A., et al., *Postglacial recolonization history of the European crabapple (Malus sylvestris Mill.), a wild contributor to the domesticated apple.* Molecular Ecology, 2013. **22**(8): p. 2249-2263.

284. (NCEI), N.N.C.f.E.I. *U.S. Billion-Dollar Weather and Climate Disasters*. Available from: https://www.ncdc.noaa.gov/billions/.

285. Wisniewski, M., T. Artlip, and J. Norelli, *Dealing with Frost Damage and Climate Change in Tree Fruit Crops.* New York State Fruit Quarterly, 2016. **24**(3): p. 25-28.

286. Goodstein, D.M., et al., *Phytozome: a comparative platform for green plant genomics.* Nucleic acids research, 2011. **40**(D1): p. D1178-D1186.

287. Kohonen, T., *The self-organizing map.* Neurocomputing, 1998. **21**(1-3): p. 1-6.

288. Martinetz, T.M., S.G. Berkovich, and K.J. Schulten, *'Neural-gas' network for vector quantization and its application to time-series prediction.* IEEE transactions on neural networks, 1993. **4**(4): p. 558-569.

289. Griebel, T., et al., *Modelling and simulating generic RNA-Seq experiments with the flux simulator.* Nucleic acids research, 2012. **40**(20): p. 10073-10083.

290. Bray, N.L., et al., *Near-optimal probabilistic RNA-seq quantification.* Nature biotechnology, 2016. **34**(5): p. 525.

291. Monier, B., et al., *IRIS-DGE: An integrated RNA-seq data analysis and interpretation system for differential gene expression.* bioRxiv, 2018: p. 283341.

292. Chen, X., et al., *SeqTU: a web server for identification of bacterial transcription units.* Scientific reports, 2017. **7**: p. 43925.

APPENDIX 1: Grant proposal to South Dakota Competitive Research Grant Program

Project Description

Research Objectives. Innovations in genomic sequencing technologies have transformed the landscape of biological and genetic research. Encompassed in this emerging area of study is RNA-sequencing (RNA-seq), which provides a view of the genome-scale gene expressions. *The two objectives of this proposal are (i) Construct a novel computational pipeline for RNA-seq data analysis and (ii) Correct an intrinsic computational bottleneck in RNA-seq data analysis using a novel statistical model.* Through integrating existing computational techniques and developing novel methods and approaches to large-scale RNA-seq data in the public domain, we will enable a wide range of research areas to benefit and contribute to training a new generation of scientists with the capacity to elucidate biological systems by computational techniques.

Background and Significance. The advent of much-improved biotechnology and the decreased associated costs have increased the amount of biological data, including Next-Generation Sequencing (NGS) [1, 2], which has higher resolution, better accuracy, lower technical variation, and other advantages, compared with array-based counterparts [3-5]. One of the predominant data types that has arisen from NGS technologies is RNA-sequencing (RNA-seq) data, which promises to provide a comprehensive picture of the transcriptome for a biological process. RNA-seq is a revolutionary technology for gene expression profiling [11, 12]. Modern RNA-seq analyses involve computations to estimate gene expression and related biological interpretation. Numerous methods have been developed—both in the public [20-46] and private sectors [47-54]—and formed into

"pipelines" to facilitate the analysis of RNA-seq data. While numerous tools are available, many suffer from particular issues that affect analysis results, and construction of applicable combinations of these tools is an ongoing challenge. Even for the tools that have sufficient individual performance, implementation in a sequence or entire pipeline can result in decreased overall performance and biased or unreliable results [56]. This fact makes establishing a reliable computational pipeline for RNA-seq data a non-trivial task.

Although substantial mathematical modeling and computational algorithms & tools have been specifically developed for RNA-seq analysis, the reality is that some of the most widely-used methods cannot provide accurate information related to gene expression estimates [55, 56]. Even though some tools can perform RNA-seq analyses acceptably on some datasets, prominent issues are found within each step of the pipeline. One such issue is referred to as mapping uncertainty [27, 77, 78], in which similarities within a genome or across multiple genomes (i.e., metagenome) can cause difficulties in determining an accurate estimation of gene expression levels. We have conducted the analysis of almost 2TB of data from seven different plant and animal species and found that an average of 20% of RNA-seq data exhibits mapping uncertainty using the current state-of-the-art computational tools. This uncertainty has the potential to drastically impact the quality of genetic expression estimates that are used in downstream analyses, leading to misinterpretation of results and negatively affecting the understanding of biological insights for agriculture, animal sciences, and human health. Hence, it is critical to improving existing bioinformatics tools using more effective algorithms to improve performance related to mapping uncertainty.

Proposed Work.

Objective 1: Construct a novel computational pipeline for RNA-seq data analysis

The PI will establish a framework for developing new RNA-seq data analysis approaches through a four-tier integrative pipeline (Figure 1). This pipeline involved preprocessing (Tier 1), basic analysis (Tier 2), hypothesis-driven interpretation (Tier 3), and discovery-driven interpretation (Tier 4). This framework will provide a comprehensive analysis of RNA-seq data for all purposes.

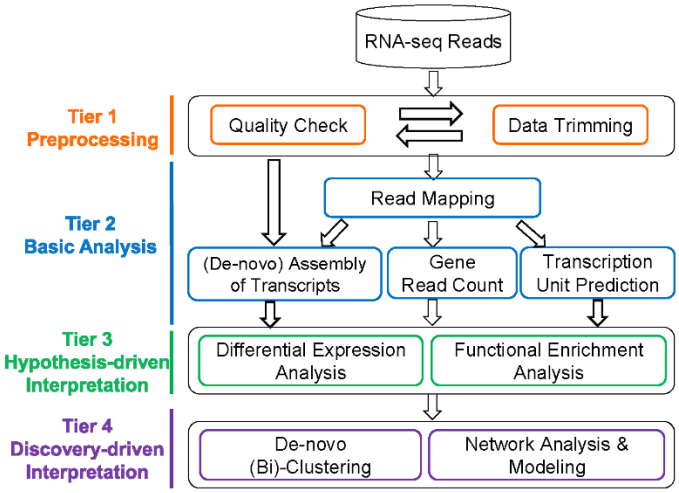


Figure 1: The four-tier RNA-seq analysis pipeline in the PI's lab, with data preprocessing, basic analysis, hypothesis-driven interpretation, and discovery-driven interpretation.

While the general framework has been clearly defined, specific pipelines designed for the species of interest require more investigation and optimization. Objective 1 focuses on the discovery of which high-performing tools should be implemented in this framework for the plant (Arabidopsis, Soybean, and Grape) and animal species (human and mouse) to provide optimized results for RNA-seq studies. Substantial RNA-seq datasets of these species can be freely downloaded from the SRA database of NCBI (https://www.ncbi.nlm.nih.gov/sra). Several existing in-house tools [88, 100, 111, 112, 133, 207, 291, 292] in the PI's lab can fully support the pipeline construction.

Objective 2: Correcting mapping uncertainty using the Ambiguous Reads Mapping (ARM) tool

Current methods for addressing mapping uncertainty are underperforming and potentially affecting the accuracy of downstream analyses [77, 87]. Therefore, a rigorous statistical model for accurate gene expression estimation is required for all downstream expression-

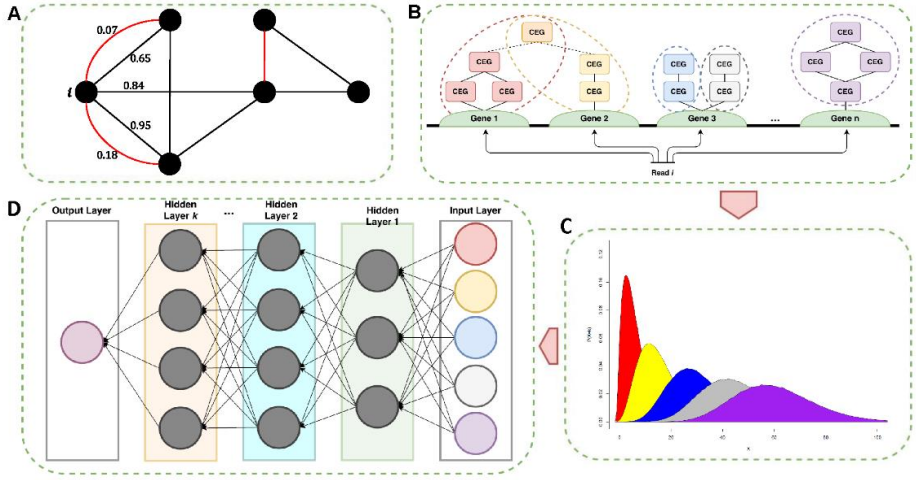


*Figure 2. (A) Gene-gene interaction established within GeneQC. (B) The algorithm, **ARM,** for re-alignment of ambiguous genes based on the information collected from part A. Previously established KEGG pathways and regulatory motifs are used to generate networks for potential gene alignment locations, (C) which are then used to generate probability distributions for each gene location. These distributions can be used independently or as prior distributions in a (D) neural network and hidden Markov model to determine the optimal alignment for ambiguous reads.*

based analysis and interpretation. To achieve Objective 2, the PI proposes the combined use of transcriptomic, genomic, and network information to establish a more biologically applicable determination of correct read alignments. Currently, the PI-developed tool GeneQC[88] is capable of extracting transcriptomic and genomic features of the sample and species information (Figure 2A). This information will then be utilized with gene regulatory information sourced from pre-existing networks (Figure 2B) to determine a probability distribution for each potential alignment (Figure 2C). These distributions will be applied in a straightforward manner or as a prior distribution for advanced machine learning processes (Figure 2D) to provide a higher-likelihood alignment.

Outcome and Assessment: The proposed algorithms in above two Objectives will be implemented within computational tools called IRIS (An integrated RNA-seq data analysis and interpretation system) and ARM (Ambiguous Reads Mapping). Once

thoroughly developed, IRIS and ARM will be tested against state-of-the-art read alignment tools and compared using a previously developed D-score metric, which indicates the level of mapping uncertainty for each gene expression estimate[88]. Significant improvement of read alignment related to mapping uncertainty will be assessed by lower D-scores, indicating lower mapping uncertainty, and more accurate expression estimates.

Broader impacts: The new computational techniques developed will enable a large community of biological researchers to conduct a broad range of RNA-seq data analyses that are currently infeasible. The new tools will enhance the understanding of how gene expression is controlled by the underlying regulatory systems. The application of the proposed methods will facilitate the elucidation of the gene regulatory network encoded in a cell. Hence, the research has the potential to transform the rapidly-developing biotechnology and bioinformatics fields yielding innovative analytical tools that enhance new biological discoveries. Through the development of the proposed pipeline and novel computational tools, numerous undergraduate and graduate students will be engaged. These activities also provide excellent opportunities for the involved students to receive much-needed experience related to bioinformatics data analysis and make them be better prepared in the rapidly expanding biotech industry, meeting the demands of interdisciplinary academic training. For example, within the region, there are numerous institutions that are actively searching for qualified bioinformatics analysts, including Sanford Health, Avera Health, Monsanto, and many other private organizations.

Description of Facilities and Resources. The PI (1) has substantial computational resources by the XSEDE clusters (https://www.xsede.org/); (2) is a member of the Biochemical Spatio-temporal NeTwork Resource (BioSNTR; http://biosntr.org) and has access to all BioSNTR resources; and (3) has full access and accounts for the High-Performance Computing (HPC) computer clusters at SDSU and IPLANT Cyberinfrastrucure.

Besides the above computational resources, the PI has established a new computational laboratory (~300 sq. ft.) in the McFadden Biostress Laboratory building at South Dakota State University (SDSU). His lab has 10 separate benches/desks and a computer studio, which currently houses eight individuals and has one Linux cluster (6 CPUs, 64GB RAM and 3TB hard disk), five desktops, two workstations, one iMac, and one MacBook pro. All computers are connected to SDSU network and have access to the Internet. The programming environment includes UNIX, C/C++, PERL/BioPERL and R. The lab is familiar with all kinds of bioinformatics databases and resources (GenBank, RefSeq, etc.); various genome annotation resources (NCBI, GO, etc.); and general bioinformatics tool packages for sequence analysis (BLAST, MCL, Cytoscape, etc.).

Capacity Building and Commercial Potential of the proposed computational software. The proposed Objectives demonstrate a method for providing significant improvements to RNA-seq data analysis in the form of optimized computational pipelines and improved analysis tools. Commercial applications of RNA-seq pipelines have been successfully demonstrated numerous times, including CLC Genomics Workbench[54] and Galaxy[53], which started as a free software but expanded into a commercial tool. While these and other commercial RNA-seq tools pipelines have

demonstrated success, they are not immune to the previously mentioned issues that plague all RNA-seq tools. Because of this, optimization of top RNA-seq tools and remediation of mapping uncertainty provides a promising potential for widespread use, especially considering the modern movement and importance of interactivity and graphical interfaces in reproducible RNA-seq data analysis[106].

In addition to these proposed objectives being implemented into a server framework, the PI has recently-developed high-performance RNA-seq tools that can be implemented in this pipeline, including GeneQC for read alignment quality control[88], IRIS-EDA for Differential Gene Expression (DGE) analysis[291], and ViDGER for visualizing differential gene expression results[133]. Integration of these novel-feature tools with the prospective ARM tool into the optimized pipeline framework in objective 1 provides a promising method for commercialization of these RNA-seq tools, which will be achieved through collaboration with the SDSU Office of Technology Transfer & Commercialization.

APPENDIX 2: Curriculum vitae

Adam McDermaid

Adam.McDermaid@sdstate.edu

EDUCATION

PhD in Computational Science & Statistics (3.89 GPA) **August 2015-June 2018**

South Dakota State University, Brookings, SD

- Coursework: Bioinformatics, Regression Analysis, Statistical Inference, Multivariate Analysis, Measure & Probability Theory
- Bioinformatics emphasis

MA in Mathematics (3.92 GPA) **August 2013-May 2015**

University of South Dakota, Vermillion, SD

- Coursework: Real & Complex Analysis, Measure Theory, Operations Research, Abstract Algebra & Algebraic Number Theory, Partial Differential Equations

BS in Mathematics (Chemistry) **August 2008-May 2013**

University of South Dakota, Vermillion, SD

- Coursework: Real Analysis & Advanced Calculus, Organic & Environmental Chemistry, Biology

EMPLOYMENT

Graduate Research Assistant **March 2016-June 2018**

Bioinformatics and Mathematical Biosciences Lab, South Dakota State University

- Duties include researching RNA-seq pipeline tools, development and applications of RNA-seq pipelines, applications of statistical techniques to bioinformatics problems, development of novel bioinformatics algorithms and softwares
- Collaborations include US Department of Agriculture, Ohio State University, NSF Plant Genome Research Projects, and SD EPSCoR/BioSNTR

Graduate Teaching Assistant **August 2015-May 2017**

Department of Mathematics & Statistics, South Dakota State University

- Duties include Introduction to Statistics Recitation instruction, College Algebra help sessions, Logic, Sets & Proofs help sessions, development of new calculus sequence teaching methods

Graduate Teaching Assistant **August 2013-May 2015**

Department of Mathematical Sciences, University of South Dakota

- Duties included Finite Mathematics and College Algebra instruction, creation and assessment of evaluation materials for Finite Mathematics

PUBLICATIONS

1. Liu, B., Yang, J., Li, Y., McDermaid, A., & Ma, Q. (2017). An algorithmic perspective of de novo cis-regulatory motif finding based on ChIP-seq data. *Briefings in Bioinformatics*. doi:10.1093/bib/bbx026
2. Niu, S., Yang, J., McDermaid, A., Zhao, J., Kang, Y., & Ma, Q. (2017). Bioinformatics tools for quantitative and functional metagenome and metatranscriptome data analysis in microbes. *Briefings in Bioinformatics*. doi:10.1093/bib/bbx051
3. Yang, J., Chen, X., McDermaid, A., & Ma, Q. (2017). DMINDA 2.0: Integrated and systematic views of regulatory DNA motif identification and analyses. *Bioinformatics*. doi:10.1093/bioinformatics/btx223
4. McDermaid, A., Chen, X., Zhang, Y., Xie, J., Wang, C., & Ma, Q. A new computational framework for mapping uncertainty analysis in RNA-Seq read alignment and gene expression estimation. (Under review in *Frontiers in Genetics*)
5. McDermaid, A., Monier, B., Zhao, J., Liu, B., & Ma, Q. Interpretation of differential gene expression results of RNA-seq data: review and integration. (Under review in *Briefings in Bioinformatics*)

6. Chen, X., Ma, A., McDermaid, A., Zhang, H., Cao, L., Cao, H., & Ma, Q. RECTA: Regulon Identification Based on Comparative Genomics and Transcriptomics Analysis. (Accepted for publication in *Genes*)
7. Monier, B., McDermaid, A., Zhao, J., Fennell, A., & Ma, Q. IRIS-EDA: A web server for user-friendly, design-robust gene expression data analysis, interpretation, & visualization. (Under review in *Bioinformatics*)
8. McDermaid, et al., ARM: A tool for comprehensive ambiguous reads mapping. (In preparation)
9. Migicovsky, Z., Harris, Z., Klein, L., Li, M., McDermaid, A., Chitwood, D., Fennell, A., Kovacs, L., Kwasniewski, M., Londo, J., Ma, Q., & Miller, A. Roostock effects on scion phenotypes in a 'Chambourcin' experimental vineyard. (In preparation)
10. McDermaid, A., Artlip, T., Ma, Q., & Wisniewski, M. Strain effect on gene expression in *M. domestica*. (In preparation)
11. Xia, Y., McDermaid, A., Wang, C., & Ma, Q. Genomic analysis of *Bacillus* sp. YF23. (In preparation)
12. McDermaid, A., Gu, S., & Ma, Q. A review of machine learning applications on the prediction of mapping uncertainty. (In preparation)
13. McDermaid, A., Gu, S., & Ma, Q. GeneQC2.0: An R package for quality control of gene expression estimation through novel application of machine learning (In preparation)

PRESENTATIONS

- Gene Expression Analysis of Transgenic Apples. June 17, 2016, University of South Dakota, Erliang Zeng Lab, Vermillion, SD. (Poster Presentation)
- Principal Component Analysis & Network Component Analysis. July 1, 2016, University Center, Sioux Falls, SD. (Zeng Lab/BMBL Inter-lab meeting presentation)
- RNA Sequencing Analysis, Applications, & Modeling. November 10, 2016, SDSU-Sanford Research Symposium, Brookings, SD. (Poster Presentation)
- Computational Techniques & Algorithm Design in RNA Sequencing Analyses. January 30, 2017, SDSU Department of Agronomy, Horticulture and Plant Science

USDA-ARS North Central Agricultural Research Laboratory, Brookings, SD. (Departmental seminar)

- Addressing Multimapping Uncertainty in RNA Sequencing Alignment. May 23, 2017, All Investigator Meeting, South Dakota Experimental Program to Stimulate Competitive Research, Oacoma, SD. (Poster Presentation)
- RNA Sequencing Analyses and the Multimapping Uncertainty Issue. June 10, 2017, It's All About Science Festival, Sioux Falls, SD. (Poster Presentation)
- RNA Sequencing Analyses and Multi-Mapping Uncertainty. August 25, 2017, Plant Genome Research Program Project Year 1 Meeting, Davis, CA.

SKILLS

- Next-Generation Sequencing Analyses
- Hypothesis- & Discovery-driven analyses
- Large-scale data management and analysis
- Mathematical & Statistical Modeling
- R programming
- Python, Perl, SQL, & SPSS experience
- Strong written & oral communication

PROFESSIONAL ASSOCIATIONS

- Mathematical Association of America, Member, 2012
- Institute of Mathematical Statistics, Member, 2016
- *BMC Genomics*, Reviewer, 2016
- *Mathematical Biosciences*, Reviewer, 2016
- *Journal of Bioinformatics and Computational Biology*, Reviewer, 2016
- International Conference on BioInformation and BioMedicine, Reviewer, 2016
- *Frontiers in Young Minds*, Understanding Mathematics, Review Editor, 2017
- *Nucleic Acids Research*, Reviewer, 2017