EPiC Series in Computing

Volume 83, 2022, Pages 76–85

Proceedings of 14th International Conference
on Bioinformatics and Computational Biology

A Novel Approach for Mapping Ambiguous Sequences of Transcriptomes

Tamer Aldawiri[1,2*], Bindu Nanduri[1], Mahalingam Ramkumar[1] and Andy D. Perkins[1]

[1] Mississippi State University, Starkville, MS 39762, USA
[2] Temple University, Philadelphia, PA 19122, USA
aldwairi@temple.edu, bnanduri@cvm.msstate.edu,
ramkumar@cse.msstate.edu, perkins@cse.msstate.edu

Abstract

An organism's transcriptome is the set of all transcripts within a cell at a certain time. We often analyze the transcriptome by quantifying gene expression and performing subsequent analyses such as a differential expression or a network analysis. Such analysis helps us in understanding and interpreting the functional elements of the genome. Many challenges limit the accuracy and ability to map all the RNA-Seq correctly into its genome sequence. Some of these challenges are exemplified when mapping sequences fall at exon junctions, sequences containing polymorphisms, multiple insertions or deletions, and reads falling partially or wholly within introns. One of the most significant problems is the loss of data occurring from the inability to map sequences when they align to multiple genomic locations, sometimes called ambiguous sequence mappings. In this paper, we present a novel method to increase the accuracy of gene expression estimation by relying on a statistical approach to increase the accuracy of mapping the ambiguous reads to their proper locations within the genome. This approach allows us to better identify significantly expressed genomic locations so we can accurately map ambiguous reads to their most likely accurate genomic locations and to define more precisely which genes are expressed throughout the genome. Due to its statical nature the approach can be easily combined with other existing mapping tools and mechanisms as well.

1 Introduction

RNA-Seq is a technique that utilizes next-generation sequencing technologies to study transcriptomes at the nucleotide level. RNA-Seq is known for its precision in measuring transcripts levels and their

[*] Corresponding Author

identify various isoforms, yet it faces some challenges that hinder it from being the technique of choice for gene expression profiling. One of the main challenges appears when attempting to map RNA sequences to a reference genome; high percentages of short sequence reads are often assigned to multiple genomic locations. A sequence mapping is said to be "*ambiguous*" when the read sequence is mapped to multiple genomic locations within a genome. One approach to handling these "*ambiguous mappings*" has been to discard them [2,26]. This results in a loss of data, which can sometimes be as large as half of the sequenced data and also affects precise breakpoint prediction [23].

Another approach assigns them randomly to one of the locations from a set of best assignments using mapping tools like MAQ [12] -MAQ has been superseded by BWA for sequence aligment- or BWA [22]. Another approach "*rescues*" multi-reads through estimating an initial gene expression through allocating unique reads which are used to partially allocate the ambiguous reads and then a final gene expression is acquired through re-estimating the gene expression after the allocation of the multi-reads [3, 4, 15]. All of these approaches will result in a significant loss of the original data and other problems that might occur such as an overestimation of homoeolog co-regulation and the incorrect inference of subgenome asymmetry in network topology these types of problems can hinder our understanding of duplicate gene expression. These problems can be reduced by modifying the factors influencing the ambiguity sequencing and resequencing strategy and the fundamental resources[24].

Tools like Eland [14], SOAP [11], MAQ [12], BWA[22], RMAP [13], Seqmap [16], and Bowtie [1] are usually used to perform initial mapping of sequence data to a corresponding reference genome. Nevertheless, these tools do not fully address the ambiguous mappings of the sequence reads. This is an important problem since those ambiguous reads often comprise a large portion of the genomic sequences generated. Newer tools like RSEM[27], Salmon[28], and FUDA[29] provide a better estimation but are not suitable for all situations.

Sequence-based transcriptome analysis, specifically the high throughput sequencing of cDNA known as RNA-Seq, has emerged as an alternative to microarray gene expression profiling [10] which had long been the most widely used method for transcriptome analysis. The reason for RNA-Seq's emergence over microarrays was due to several reasons and is mainly motived by the limitations of microarray studies, such as access and cost. Microarray experiments require the physical disruption of the cell to get access to gene expression patterns, and the complexity and limited amount of tissue samples to be obtained can be limiting factors to the quantity and quality of RNA that can be isolated from microarray experiments. Another limiting factor in microarray experiments in medical applications is that many clinical specimen sizes are small since they are usually obtained during early diagnoses. Degraded RNA is also an issue to be concerned about since it could result in the generation of false data. The degradation usually appears due to the numerous steps that are prone to errors in a microarray experiment. Therefore microarray experiments need to be replicated to eliminate such errors. Another important issue is that although many tools are available, microarray experiments still lack standard methodologies for collecting, analyzing, and validating the data [21].

On the other hand, RNA-Seq provides many advantages over microarrays. RNA-Seq is less prone to errors due to the omission of the hybridization step used during the process of preparing microarrays samples [17, 18]. RNA-Seq studies are more suitable for discovery-based experiments, unlike microarray studies, which are pre-design-driven. pre-design-driven means that what we know about the genome guides the design of the experiment and is already built into the microarrays. RNA-Seq does not require previous knowledge about the nature of the transcriptome because of its hypothesis-free nature, and it also allows us to study species with poor or missing genomic annotations. RNA-Seq also permits the detection of lowly expressed genes, alternative splice variants, and novel transcripts. This is in contrast to microarray hybridization techniques, which can limit the accuracy of expression measurements, especially when transcripts are present in low abundance [6]. Even though comparing results across arrays can help in identifying gene expressions among samples of interest. A single sample is usually not sufficient to provide reliable information about the expression levels for different transcripts [2, 7, 9]. This paper is a continuation of our work on identifying ambiguous sequences of transcriptomes [30].

2 Methods

Our method exploits the use of ambiguous reads to provide us with a more accurate estimation of gene expression and better mapping of ambiguous reads to their most appropriate location within the genome. We start by assigning a weighted score for each position in the genome. This is done by finding the expression value for that position which represents the number of reads – including the ambiguous reads – that map to that exact single location within the genome. For each gene within the genome, the weights are averaged to provide us with an average weight score. Then we find if each gene is expressed by comparing the average weight score for that gene against a list of one hundred to one thousand random genes along with their average weight scores. If the statistical significance or the p-value of that gene is below a certain threshold for example below 0.1, 0.05, or 0.01 then that gene is considered expressed. Once we define which genes are expressed and which are not. We then revisit the mapping process but this time we assign the unique reads to their proper location and map the ambiguous reads to their most appropriate location which are the expressed genes regions within the genome. This is in contrast with the previous step where we mapped the ambiguous reads to all their probable locations. The second step provides us with a more accurate mapping of the ambiguous reads to their accurate location within the genome. We find that our estimation of gene expression provides us with more accurate results than previously used methods since we include the ambiguous reads in our estimation which are usually discarded or randomly mapped to one of the locations within the genome by the other methods.

We start by aligning the RNA-Seq reads to their reference sequence using one of the known sequence mapping tools. We had choices between different short-read alignment tools like Bowtie [1] and MAQ [12]. MAQ provides higher sensitivity in mapping unique sequences than Bowtie while Bowtie is much faster than MAQ in the mapping process [1]. When comparing the benefits of gaining a little higher sensitivity using MAQ to a much higher speed using Bowtie, we found that using Bowtie is much more beneficial for the analysis we are doing here. This is because our approach requires identifying all the unique and ambiguous reads and their locations within the genome which is more time-consuming than just finding the unique reads. For this reason, we used Bowtie to extract and find all the possible mappings for each read. Another reason for choosing Bowtie is that its results can be easily imported into other tools like Tophat and Cufflinks.

We used the following parameters: (-a) to report all the valid alignments. It is important to report all the valid alignments, and not only the best ones, which can be specified using the (--best) parameter or the unique ones which are usually specified by the (-m) parameter since we are interested in the reads that align to multiple locations. Specifying (--best) will only identify the alignments that have the least mismatches while using (-m k) will suppress any multiple reads that exceed the value of k. This means that if the read has more than three possible mappings and –m was specified to be three, then the mappings associated with the read will not be reported. We specified that (-n 2) which allows alignment of reads with only two mismatches at the most. We used multithreading by specifying (–p 10) which allows for 10 threads. Finally, we used the (-S) option, which allows us to display our output in SAM format. We also used the --sam-nohead option to remove the extra headings provided by the SAM format. Understanding the difference between these parameters is important since Bowtie does not by default report the ambiguous reads.

We then calculate the expression value for each position in the genome. This is done by finding the number of reads that map to each single nucleotide location in the genome. To facilitate this task we needed to create a file and store the locations of each position in the read along with its count in ascending order. The counts represent the weights of all the reads that map to that location. This is not a trivial task since if the number of aligned reads for our data is large, sorting them could be a time-consuming task. To prove the validity of our approach, we decided to use a small data set before incorporating our approach into a larger and more complicated data set. Figure .1 below shows the process of mapping reads to single and multiple locations based on the positive and negative strands.

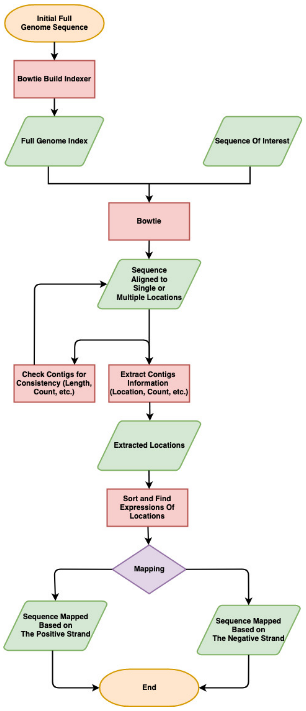
Initial Full Genome Sequence
Bowtie Build Indexer
Full Genome Index
Sequence Of Interest
Bowtie
Sequence Aligned to Single or Multiple Locations
Check Contigs for Consistency (Length, Count, etc.)
Extract Contigs Information (Location, Count, etc.)
Extracted Locations
Sort and Find Expressions Of Locations
Mapping
Sequence Mapped Based on The Positive Strand
Sequence Mapped Based on The Negative Strand
End

Figure.1 Flowchart for ambiguous sequence mapping

We chose to run Bowtie on the Escherichia coli strain K-12 sub-strain MG165 [19]. Bowtie reported 86.30 % of the reads to have at least one possible alignment location and 13.70% to have failed to align. Out of the 86.30%, only 32.60% aligned to a single unique location while 53.70% have multiple alignment locations. The number of ambiguous reads in this sample constitutes more than half of the mappable genome. Of course, this is not always the case, but the percentage presented here shows how significant the ambiguous read problem could be.

Once we have all the reads sorted according to their positions, we assign a weighted score to each position within the genome where a certain read or reads maps. For calculating the weighted score we used two different approaches. The first approach assigns a score of one to every single location to which a read maps. For example, if ten reads maps to that certain location then the weighted score of the location will be 10. The second approach is similar to the first approach but instead of assigning a score of one to each read that maps to a certain read location, it calculates the number of locations that each read maps to and assigns a fraction of that number as the score for that location. For example, if a read maps to 5 different locations then for each location a weighted score of 1/5 would be assigned to each of those different locations. This means that if 10 reads map to a single certain location then the weighted score of that location will be the sum of the individual weights for all those 10 reads. We decided to call the first approach the Individual-Count-Weight and the second approach the Partial-Count-Weight.

The steps below show the algorithm for evaluating the significance of gene expression while allowing the ambiguous reads to be included from the beginning and throughout the mapping process.

1. Enumerate all the possible mappings for every read to their different positions within the genome and assign a weighted score to each position based on their read depth.
   1a. Individual-Count approach a weighted score equal to 1 for each read that maps to a single location.
   1b. Partial-Count approach a weighted score equal to 1/r where r is the number of locations to which a read maps in the genome.
2. Create a file and store the different weight scores based on the approach used.
   2a. For the Individual-Count approach, we store and sort the locations of every single read along with its score in ascending order.
   2b. For the Partial-Count approach, we store the ambiguous reads and their locations and divide the multiple reads that map to the same location as a single group.
3. Find the coordinates (start, end) for each gene within the reference genome.
4. Calculate the mean score for each gene in the reference genome

$$Mean,\ \mu = \frac{\sum_{i=0}^{n} w_i}{n} \tag{1}$$

$w =$ *the individual weight for each location within the gene*
$n =$ *lenght of the gene*

5. Select 100 random gene locations, xi, from the expression flat-file and calculate the average score for those 100 locations.

6. Compute standard deviation (σ) of all of the random locations as compared to the gene mean score x, along with their Z-score and p-value using the following equations.

$$Standard\ Deviation,\ \sigma = \frac{\sqrt{\sum_{i=0}^{N} (x_i - \mu)^2}}{N}\ ,\ Where\ N = 100 \tag{2}$$

$$Z - score,\ z = \frac{x - \mu}{\sigma} \tag{3}$$

7. Find which genes are significantly expressed based on p-values below 0.1, 0.05 or 0.01. The p-value was calculated using the cumulative distribution formula presented in the equation below.

$$P - value = \frac{1}{2}\left[1 + erf\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right] \tag{4}$$

The total weighted score for each individual gene is the sum of all the weights that map to each individual location. If there are no reads that map to a location then the read count is zero for that specific location. Once we establish a score total for a position in the genome, we need to find the expression value of genes by calculating the average score for each gene. We do that by summing the weights of each individual position in the gene and dividing them by the length of that gene. We repeat the process until we have a total score for every gene in the genome.

To be able to establish the significance of the gene expression values calculated for each gene, we compare the score of each gene (termed the gene of interest) against the scores of several genes chosen at random. To do so we selected 100 random gene scores and found the average score for each of those genes. We also tested the possibility of choosing 1000 random locations and found that it rarely affects the accuracy of the results yet it increases the complexity and the computational running time of the program significantly. We then compute the standard deviation, the z-score, and p-value of the gene of interest with regards to the mean random gene value. The p-value calculated for each gene will give us information regarding which of the genes can be considered expressed at a significant level.

3 Results and Discussion

3.1 Defining significantly expressed genes

Determining which genes are expressed has traditionally been a difficult problem with no clearly defined solution. The reason for that could be attributed to the fact that a certain gene could appear as being expressed when compared to a certain number of genes while it may not appear to be expressed when compared to another set. This explains why some studies decided to choose the top 10% of the mean gene values as their cutoff value for deciding which genes are expressed and which are not [8, 20]. We believe that such an approach does not give us the most accurate results regarding which genes should be considered and which are not. To show that choosing the top 10% as a cut-off value does not provide us with the most accurate results. We took the top 10% of the mean gene values and compare them to the top 10% of the genes p-values. We show that both approaches yield different results and that taking the top 10% as a cut-off value will lead us into considering a number of genes as expressed when they really are not. We will do that by comparing the top 10% approach to choosing significantly expressed gene values that are less than a certain p-value like 0.1, 0.05, or 0.01.

We start by choosing the top 10% as a cut-off value for both the mean gene values and the genes p-values and tested them on both approaches that we developed previously. Using the Individual-Count-Weight approach we found that out of the 454 genes that represent the top 10% of the mean gene values, 350 genes intersect with the top 10% of the genes according to their p-values while 104 genes approximately 23% do not. When using the Partial-Count-Weight approach we found that out of the 454 genes that represent the top 10% of the mean gene values, 362 intersect with 10% of the genes according to their p-values while 92 genes approximately 20% do not.

When comparing both our Individual and Partial count approaches against the approach of choosing the top 10% of the mean values we noticed that the top 10% of the mean values do not always lead to the selection of the significantly expressed genes within the genome. Nevertheless, we also observed that during our inspection of the genes' p-values that not all the genes that fall in the highest 10% according to their p-values have a level of significance below 0.1, 0.05, or 0.01. The reason for that could be attributed to the fact that certain genes appear as expressed when compared to a certain set of genes with very low expression values. While those same genes could appear as unexpressed when compared to another set that has a very high expression value. Setting a certain percentage as a threshold will either allow for some genes that are unexpressed to be included or for some genes that are expressed to be excluded. That is why we believe that setting a percentage for a cut-off value to define gene expression does not present us with a decisive approach regarding the evaluation of gene expression.

In addition, we noticed during the analysis of the genes' expression values that there are a small number of genes that have extremely high expression values, which we believe are outliers. Those genes can cause the mean expression values and the standard deviation to be extremely large. When selecting the random locations, filtering out the top 2% of genes with the highest expression values will increase the number of significant genes that fall below the p=0.01 or 0.05 threshold by two-fold.

Using the Individual-Count-Weight approach and the Partial-Count-Weight approach, we found that the number of genes that are highly expressed varies at different p-values. We can see that there is nearly a two-fold increase in the number of genes that are expressed after filtering out 2% of the top expressed random genes.

Table.1 below shows the different numbers of genes that are expressed at different p-values. From the information in the table below we can see that the number of genes that are expressed at different p-values is less than the 454 genes represented by the top 10% of the mean genes.

Number of random genes	Before filtering 2%	After filtering 2%
Individual-Count-Weight at 0.01	94	198
Individual-Count-Weight at 0.05	112	224
Partial-Count-Weight at 0.01	106	199
Partial-Count-Weight at 0.05	126	239

Table.1 The number of genes that are expressed at different p-values.

3.2 Remapping the ambiguous reads for the expressed genes

Once we define which genes are expressed within the genome we revisit our mapping process. This time we run Bowtie with a different setting we specify (-m) = 1 to report only reads that have only one reported alignment which will give us the reads with the unique mappings only. We then find the ambiguous reads that map to an expressed gene location. If some of the ambiguous reads map to more than one expressed gene location then it is assigned to those locations and its weight will be equal to a fraction of the number of locations that the reads map to. The benefit of this revisiting approach is that it is more likely to give us a more accurate mapping of the ambiguous reads by using the estimated expression of each gene that was identified in the previous step.

4 Conclusions

We devised an approach to provide a more accurate estimation of expressed genes by providing a statistical solution to the ambiguous mapping problem that would increase the accuracy of the reads that map to multiple places. We found that considering the top 10% percent of the genes mean values as being expressed does not reflect the genes whose expression is statistically significant. It also contributes to a solution to the ambiguous mapping problem by allowing the multiple reads to be included in the mapping of the reads from the beginning of the process. We also found that the number of genes that are highly expressed nearly doubles when filtering out the top 2% of the randomly chosen genes. This indicates that there is a certain percentage of genes with mean expression values that are extremely high that could statistically affect the decision to accept certain genes as being expressed or not. Finally, there is no significant difference between the Individual-Count and the Partial-Count approach when it comes to the number of expressed genes for both p-values at 0.01 and 0.05. This indicates that the accuracy gained through placing partial weights when mapping the multiple reads does not provide us with a more accurate expression of the reads. Future work includes trying to solve the problem of working with larger sets of data. Working with larger data sets presents a major challenge since it increases the computational complexity of our approach. The complexity increases drastically during the mapping step because we need to enumerate all the possible locations for the mapping step of each ambiguous read and then we need to store the expression values and changes associated with those values for each individual location within the genome during the mapping step. One possible solution to this problem would be to sort all the reads before the mapping starts which would decrease the complexity of the mapping step of the approach but would add more time due to the addition of the sorting step to the approach. This approach can be combined with machine learning algorithms to select features/genes [25, 31, 32, 33] or genes that mostly affect the ambiguous mapping process to reduce the complexity of mapping the reads into many different locations. The approach we present here is meant to utilize the data that is usually excluded by many alignment tools to provide a more accurate gene expression and can be easily combined with other tools and algorithms to provide a more accurate mapping of the ambiguous sequences. The standard approach to verify the sequences of full-length transcripts is through laboratory-based studies through utilizing statical inferences genome alignment tools can provide more accurate mapping regarding ambiguous reads which are usually too tightly or loosely assigned if not discarded at all resulting in overestimation or underestimation of transcript expression.

References

[1] B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg, “Ultrafast and memory-efficient alignment of short DNA sequences to the human genome,” *Genome Biology*, vol. 10, no. 3, p. R25, Mar. 2009.

[2] J. C. Marioni, C. E. Mason, S. M. Mane, M. Stephens, and Y. Gilad, “RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays,” *Genome Res.*, vol. 18, no. 9, pp. 1509–1517, Sep. 2008.

[3] N. Cloonan, A. R. R. Forrest, G. Kolle, B. B. A. Gardiner, G. J. Faulkner, M. K. Brown, D. F. Taylor, A. L. Steptoe, S. Wani, G. Bethel, A. J. Robertson, A. C. Perkins, S. J. Bruce, C. C. Lee, S. S. Ranade, H. E. Peckham, J. M. Manning, K. J. McKernan, and S. M. Grimmond, “Stem cell transcriptome profiling via massive-scale mRNA sequencing,” *Nat Meth*, vol. 5, no. 7, pp. 613–619, Jul. 2008.

[4] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold, “Mapping and quantifying mammalian transcriptomes by RNA-Seq,” *Nat. Methods*, vol. 5, no. 7, pp. 621–628, Jul. 2008.

[5] M. Taub, D. Lipson, and T. P. Speed, “Methods for Allocating Ambiguous Short-reads,” *Commun. Inf. Syst.*, vol. 10, no. 2, pp. 69–82, 2010.

[6] L. Gautier, L. Cope, B. M. Bolstad, and R. A. Irizarry, “affy—analysis of Affymetrix GeneChip data at the probe level,” *Bioinformatics*, vol. 20, no. 3, pp. 307–315, Feb. 2004.

[7] D. B. Allison, X. Cui, G. P. Page, and M. Sabripour, “Microarray data analysis: from disarray to consolidation and consensus,” *Nat. Rev. Genet.*, vol. 7, no. 1, pp. 55–65, Jan. 2006.

[8] O. Wurtzel, R. Sapra, F. Chen, Y. Zhu, B. A. Simmons, and R. Sorek, “A single-base resolution map of an archaeal transcriptome,” *Genome Res.*, p. gr.100396.109, Nov. 2009.

[9] N. A. Twine, K. Janitz, M. R. Wilkins, and M. Janitz, “Whole Transcriptome Sequencing Reveals Gene Expression and Splicing Differences in Brain Regions Affected by Alzheimer’s Disease,” *PLoS ONE*, vol. 6, no. 1, p. e16266, Jan. 2011.

[10] Z. Wang, M. Gerstein, and M. Snyder, “RNA-Seq: a revolutionary tool for transcriptomics,” *Nat Rev Genet*, vol. 10, no. 1, pp. 57–63, Jan. 2009.

[11] R. Li, Y. Li, K. Kristiansen, and J. Wang, “SOAP: short oligonucleotide alignment program,” *Bioinformatics*, vol. 24, no. 5, pp. 713–714, Mar. 2008.

[12] H. Li, J. Ruan, and R. Durbin, “Mapping short DNA sequencing reads and calling variants using mapping quality scores,” *Genome Res.*, vol. 18, no. 11, pp. 1851–1858, Nov. 2008.

[13] A. D. Smith, Z. Xuan, and M. Q. Zhang, “Using quality scores and longer reads improves accuracy of Solexa read mapping,” *BMC Bioinformatics*, vol. 9, p. 128, 2008.

[14] D. R. Bentley, S. Balasubramanian, H. P. Swerdlow, G. P. Smith, J. Milton, C. G. Brown, K. P. Hall, D. J. Evers, C. L. Barnes, H. R. Bignell, “Accurate whole human genome sequencing using reversible terminator chemistry,” *Nature*, vol. 456, no. 7218, pp. 53–59, Nov. 2008.

[15] G. J. Faulkner, A. R. R. Forrest, A. M. Chalk, K. Schroder, Y. Hayashizaki, P. Carninci, D. A. Hume, and S. M. Grimmond, “A rescue strategy for multimapping short sequence tags refines surveys of transcriptional activity by CAGE,” *Genomics*, vol. 91, no. 3, pp. 281–288, Mar. 2008.

[16] H. Jiang and W. H. Wong, “SeqMap: mapping massive amount of oligonucleotides to the genome,” *Bioinformatics*, vol. 24, no. 20, pp. 2395–2396, Oct. 2008.

[17] T. Casneuf, Y. V. de Peer, and W. Huber, “In situ analysis of cross-hybridisation on microarrays and the inference of expression correlation,” *BMC Bioinformatics*, vol. 8, no. 1, p. 461, Nov. 2007.

[18] M. J. Okoniewski and C. J. Miller, “Hybridization interactions between probesets in short oligo microarrays lead to spurious correlations,” *BMC Bioinformatics*, vol. 7, p. 276, 2006.

[19] Taxonomy browser (Escherichia coli str. K-12 substr. MG1655). [Online]. Available: http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?mode=Info&id=511145. [Accessed: 25-Nov-2021].

[20] R. Kumar, M. L. Lawrence, J. Watt, A. M. Cooksey, S. C. Burgess, and B. Nanduri, “RNA-Seq Based Transcriptional Map of Bovine Respiratory Disease Pathogen ‘Histophilus somni 2336,’” *PLoS ONE*, vol. 7, no. 1, p. e29435, Jan. 2012.

[21] G. Russo, C. Zegar, and A. Giordano, “Advantages and limitations of microarray technology in human cancer,” *Oncogene*, vol. 22, no. 42, pp. 6497–6507, Sep. 2003.

[22] H. Li, R. Durbin, "Fast and accurate short read alignment with Burrows–Wheeler transform", *Bioinformatic*s, Volume 25, Issue 14, Pages 1754–1760, 15 July 2009. https://doi.org/10.1093/bioinformatics/btp324.

[23] T. Gong, V. M. Hayes, E. K. F Chan, "Detection of somatic structural variants from short-read next-generation sequencing data", *Briefings in Bioinformatics*, Volume 22, Issue 3, May 2021. https://doi.org/10.1093/bib/bbaa056

[24] G. Hu, C. E. Grover, M. A. Arick, II, M. Liu, D. G. Peterson, J. F. Wendel, " Homoeologous gene expression and co-expression network analyses and evolutionary inference in allopolyploids", *Briefings in Bioinformatics*, Volume 22, Issue 2, Pages 1819–1835, Mar. 2021. https://doi.org/10.1093/bib/bbaa035

[25] T. Aldwairi, D. Perera, M. A. Novotny, "Measuring the Impact of Accurate Feature Selection on the Performance of RBM in Comparison to State of the Art Machine Learning Algorithms". *Electronics,* 2020, *9*, 1167. https://doi.org/10.3390/electronics9071167

[26] L. C. Lim , Y. Y. Lim, and Y. S. Choong. "Data curation to improve the pattern recognition performance of B-cell epitope prediction by support vector machine." *Pure and Applied Chemistry,* 2021. https://doi.org/10.1515/pac-2020-1107

[27] B. Li, C. N. Dewey. "RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome". *BMC Bioinformatics* 12**,** 323, 2011. https://doi.org/10.1186/1471-2105-12-323

[28] R. Patro, G. Duggal, M. Love, *et al.* "Salmon provides fast and bias-aware quantification of transcript expression". *Nat Methods* 14**,** 417–419, 2017. https://doi.org/10.1038/nmeth.4197

[29] M. Chung, R. S. Adkins, J. S. A. Mattick, K. R. Bradwell, *et al*. "FADU: a quantification tool for prokaryotic transcriptomic analyses." *Msystems* 6, no. 1, 2021. https://doi.org/10.1128/mSystems.00917-20

[30] T. Aldwairi, B. Nanduri, M. Ramkumar, D. Gautam, M. Johnson, A. D. Perkins. "Statistical Methods for Ambiguous Sequence Mappings". *In Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics (BCB'13)*, Association for Computing Machinery, New York, NY, USA, 674–675, 2013. DOI:https://doi.org/10.1145/2506583.2506678

[31] T. Aldwairi, D. Perera, M. A. Novotny, “An Investigation of the Role of Feature Selection on the Classification Performance of Machine Learning Algorithms”. *In Proceedings of the 33rd International Conference on Computers and Their Applications (CATA)*, Las Vegas, Nevada, Mar. 2018.

[32] V. Dixit et al., "Training a Quantum Annealing Based Restricted Boltzmann Machine on Cybersecurity Data", *IEEE Transactions on Emerging Topics in Computational Intelligence*, doi: 10.1109/TETCI.2021.3074916.

[33] T. Aldwairi, D.J. Chevalier, A. D. Perkins, Exploring the Effect of Climate Factors on SNPs within FHA Domain Genes in Eurasian *Arabidopsis* Ecotypes. *Agriculture* **2021**, *11*, 166. https://doi.org/10.3390/agriculture11020166