



Cairo University

**Faculty of Computers and Artificial
Intelligence**

Computer Science Department



Computational Analysis for RNA-Seq of Plant Organisms

PHD. Thesis

Submitted by

Heba Mohammed Zaki

Supervised by

Prof. Dr. Hesham Ahmed Hassan

Computer Science Department
Faculty of computers and Artificial
Intelligence, Cairo University

Prof. Amr Anwar Badr

Computer Science Department
Faculty of computers and Artificial
Intelligence, Cairo University

Dr. Mohammad Nassef

Computer Science Department
Faculty of computers and Artificial
Intelligence, Cairo University

Dr. Ahmed Farouk Al-Sadek

Agricultural
Research Center, ARC

A thesis submitted to the Department of Computer Science, Faculty of Computers and Artificial Intelligence, Cairo University, in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Computer Science.

November / 2020

Table of Contents

List of Figures	iv
List of Tables	v
List of Publication	vi
List of Abbreviations	vii
Glossary	ix
Abstract	xi
Chapter 1: Introduction	1
1.1 Problem Definition	3
1.2 Contribution	4
1.3 Thesis Layout	4
Chapter 2: Background and Related Work	5
2.1 Introduction	6
2.2 Microarrays Vs. NGS technology.....	7
2.3 RNA-seq Pipeline	9
2.4 RNA-seq Analysis	10
2.5 Computational RNA-seq Analysis	12
2.6 Analysis Method for RNA-seq data	17
2.7 Tools and Software packages	18
2.8 Reference Databases	21
2.9 Related work	22
2.10 Summary	27
Chapter 3: Computational RNA-seq Differential Expression and Analysis.	29
3.1 RNA-seq Differential Expression	30
3.2 Experimental Design	32
3.3 RNA-seq Analysis workflow	34
3.4 Results and Analysis	44
3.5 Summary	48

Chapter 4: Computational SNPs Analysis using Rough set theory	49
4.1 SNP detection in RNA-seq data	50
4.2 Rough Set Analysis	51
4.3 SNPs Detection Framework	52
4.4 SNPs Analysis using RST	56
4.5 Results and Analysis	57
4.6 Summary	62
Chapter 5: Conclusions and Future Work	63
5.1 Conclusions	64
5.2 Future Work	65
Appendices	67
Appendix A: Reference Databases	68
Appendix B: Expression Quantification	79
Appendix C: SNPs annotation and analysis	80
References	82

List of Figures

Chapter 2: Background and Related Work

Figure 2.1: RNA-seq pipeline	9
Figure 2.2: RNA-seq pre-analysis	11
Figure 2.3: RNA-seq core analysis	11
Figure 2.4: RNA-seq advanced analysis	12
Figure 2.5: RNA-seq disciplines	13
Figure 2.6: RNA-seq genome-guided approach	14
Figure 2.7: RNA-seq based Gene Expression	15
Figure 2.8: RNA-seq based variant calling	16
Figure 2.9: RST attribute reduction steps	18

Chapter 3: Computational RNA-seq Differential Expression and Analysis

Figure 3.1: RNA-seq analysis Workflow	35
Figure 3.2: Distribution of some highest F_{value} genes predicted by edgeR	41
Figure 3.3: Distribution of some highest FC_{Total} genes predicted by FC	43
Figure 3.4: Venn diagram showing the common heat-stress genes between FC, edgeR, DRASTIC and TAIR10	46

Chapter 4: Computational SNPs Analysis using Rough set theory

Figure 4.1: A framework for RNA-seq SNP detection	52
Figure 4.2: VCF file header and content	54
Figure 4.3: Header of (*.ann.vcf) file	55
Figure 4.4: Content sample of 'ANN' field	55
Figure 4.5: Rules Strength (coverage)	61
Figure 4.6: Rules Certainty (accuracy)	61
Figure 4.7: Decision Coverage (Yes)	61
Figure 4.8: Decision Coverage (No)	61

List of Tables

Chapter 3: Computational RNA-seq Differential Expression and Analysis

Table 3.1: Arabidopsis Thaliana reference genome and annotation files	32
Table 3.2: Data files of RNA-seq FASTQ raw reads	33
Table 3.3: Arabidopsis Thaliana annotation files in BED format	36
Table 3.4: some mapping statistics in 'Log.final.out'	37
Table 3.5: some results from file 'rsem.genes.results' for sample '12h_Rep2'	38
Table 3.6: some results from file 'rsem.isoforms.results' for sample '12h_Rep2'	38
Table 3.7: The content of the created data matrices for some genes at Gene-level	39
Table 3.8: Top percentage intersected genes between DEGs by edgeR and Reference Databases	41
Table 3.9: Top percentage intersected genes between DEGs by FC and Reference Databases	44
Table 3.10: The genes intersected between equal slices of the total edgeR, and FC DEGs	45

Chapter 4: Computational SNPs Analysis using Rough set theory

Table 4.1: VCF meta-data	54
Table 4.2: SNPs biological features	57
Table 4.3: The highest expressed heat-stress genes and their SNPs in all replicates	58
Table 4.4: The lowest expressed heat-stress genes and their SNPs in all replicates	58
Table 4.5: Replicates objects and their rules	59
Table 4.6: Sample of generated rules	59
Table 4.7: Evaluation measures for Rough Set Rules	60

Abstract

As a revolutionary technology for life sciences, RNA-seq has many applications and the computational pipeline has also many variations. RNA-seq combines simultaneous transcript identification and quantification of a large number of genes in a single assay. Consequently, the actual RNA-seq data analysis also has many variations, depending on the applications and studied organisms. Computational techniques are used widely to help in the RNA-seq analysis process for better understanding genes behavior regarding different biotic and abiotic stress conditions. Usage of computational methods has led to saving a lot of time, effort and money for biologists.

This thesis aims at maximizing benefit from RNA-seq data analysis, via usage of computational methods which can be applied on any organism. The work presented in this thesis is divided into two parts: the first part towards enhancing the differential expression analysis using edgeR and Fisher criterion (FC) analysis methods to obtain more reliable expressed genes, and second part investigates the relationship between the expression level of genes and the biological features of their SNPs using Rough set theory.

First part suggests a workflow for RNA-seq analysis to identify differentially expressed genes. This workflow is applied on the analysis of *A. thaliana* plant under heat-stress conditions. The identified candidate genes are validated via two popular references; DRASTIC and TAIR10. Results suggest that edgeR and FC methods can be combined to perform differential expression analysis for RNA-Seq data, without strong assumptions. Moreover, new promising (23) genes have been identified through comparison of results, which are unreported as heat-stress genes.

Second part suggests general guidelines for accurate SNP discovery from RNA-seq data. Those SNPs annotations are used to find relation between their biological features and the differential expression of the genes to which those SNPs belong via Rough set. This strategy is applied on the analysis of SNPs in *A. thaliana* plant under heat-stress conditions. Rough sets are utilized to define this kind of relationship into a finite set of rules. Set of (32) generated rules proved good results with strength, certainty and coverage evaluation terms. The result increases the amount of knowledge for SNPs discovery and analysis in functional genomics research.