TABLE OF CONTENTS

1. Introduction **1**

1.1 Central Dogma 1

1.2 Gene Expression 3

1.2.1 Gene Expression Data 3

1.3 Gene Expression Data Analysis 5

1.3.1 Gene Co-expression Analysis: Biclustering Approach 5

1.3.2 Differential Co-expression Analysis (DCA) 5

1.3.3 Differential Expression Analysis 6

1.3.4 Discussion 6

1.4 Motivation 8

1.5 Objectives 8

1.6 Contributions 9

1.6.1 Biclustering Analysis 9

1.6.2 Differential Expression Analysis (DEA) 10

1.6.3 Differential Co-expression Analysis (DCA) 10

1.6.4 DCA on scRNA-Seq data 11

1.6.5 Potential Biomarker Identification and Ranking 12

1.7 Organization of the Thesis 12

2. Background **14**

2.1 Statistical Methods and Measures 14

2.1.1 *p-value* 14

2.1.2 False Discovery Rate 15

2.1.3 *q-value* 16

2.1.4 Local False Discovery Rate 17

2.1.5 Heirarchical Clustering 17

2.1.6 Principal Component Analysis (PCA) 18

2.1.7 Benjamini-Hochberg 19

2.1.8 t-test 20

2.1.9 Preservation Analysis 21

2.1.10 Z-summary Statistics 21

2.2 Gene Expression Analysis Programming/Tools 22

2.2.1 R 22

2.2.2 Bioconductor 23

2.2.3 DAVID 24

2.2.4 GENIE3 . . . 26
2.3 Knowledge Repositories . . . 26
2.3.1 Entity Identifiers . . . 27
2.3.2 Gene Ontology (GO) . . . 29
2.3.3 Kyoto Encyclopedia of Genes and Genomesm(KEGG) . . . 30
2.3.4 hgu133plus2.db . . . 30
2.3.5 org.Hs.eg.db . . . 31
2.4 Biological Analysis . . . 31
2.4.1 Functional Enrichment analysis . . . 32
2.4.2 Regulatory Network Behavior Analysis . . . 33
2.5 Biomarker Criterion . . . 34
2.6 Datasets Used . . . 35
2.6.1 Microarray Data . . . 35
2.6.2 BulkRNAseq data . . . 36
2.6.3 scRNAseq data . . . 36
2.7 pre-processing of Gene Expression Data . . . 37
2.7.1 Pre-processing of Microarray Data . . . 37
2.7.2 Pre-processing of bulk RNA-Sequencing (Bulk RNA-Seq) Data 38
2.7.3 Pre-processing of Single Cell RNA-Sequencing (scRNA-Seq) Data 38
2.8 Discussion . . . 39
3. Biclustering Approach 40
3.1 Introduction . . . 40
3.1.1 Biclustering Analysis . . . 41
3.2 Related Works . . . 42
3.2.1 Metric-based Biclustering Approach . . . 43
3.2.2 Non-metric Biclustering Approach . . . 49
3.3 Some Selected Biclustering Algorithms . . . 50
3.3.1 Bimax . . . 50
3.3.2 xMOTIFs (conserved gene expression Motifs) . . . 51
3.3.3 Plaid models(PM) . . . 52
3.3.4 Iterative Signature Algorithm (ISA) . . . 53
3.3.5 Factor Analysis for BIcluster Acquisition (FABIA) . . . 54
3.3.6 QUalitative BIClustering algorithm(QUBIC) . . . 54
3.3.7 Iterative Binary Biclustering algorithm with greedy search (iBBiG) 55
3.3.8 FLexible Overlapped biClustering (FLOC) . . . 56
3.4 BicGenesis: A Method to Identify ESCC Biomarkers Using Biclustering Approach . . . 61
3.4.1 Pre-processing . . . 63
3.4.2 Bicluster Analysis . . . 63

3.4.3 DCA . . . 64
3.4.4 Identification of DEGs . . . 66
3.4.5 Validation . . . 66
3.5 Experimental Results . . . 67
3.5.1 Pre-processing . . . 67
3.5.2 Bicluster Analysis . . . 68
3.5.3 DCA . . . 70
3.5.4 Preservation Analysis . . . 75
3.5.5 Hub-gene Finding . . . 78
3.6 Validation . . . 82
3.6.1 Enrichment Analysis of Biclusters . . . 82
3.6.2 Biological Analysis . . . 83
3.6.3 Literature Trace . . . 91
3.7 Discussion . . . 100
3.8 Chapter Summary . . . 101
4. Differential Expression Analysis **103**
4.1 Introduction . . . 103
4.1.1 Differential Expression Analysis (DEA) . . . 103
4.2 Related Works . . . 105
4.2.1 Statistical Tests . . . 105
4.2.2 Microarray Methods . . . 106
4.2.3 Bulk RNA-Seq Methods . . . 108
4.3 Identification of Potential Biomarkers using Integrative Approach: Application to ESCC . . . 115
4.3.1 Pre-processing . . . 115
4.3.2 DEA . . . 116
4.3.3 DCA . . . 117
4.3.4 Identification of TEDs . . . 118
4.3.5 Validation . . . 118
4.4 Experimental Results . . . 119
4.4.1 Pre-processing . . . 119
4.4.2 Identification of DEGs . . . 120
4.4.3 DCA . . . 121
4.5 Validation . . . 128
4.5.1 Enrichment Analysis of DEGs and Modules . . . 128
4.5.2 Biological Analysis . . . 130
4.5.3 Literature Trace . . . 137
4.6 Discussion . . . 145
4.7 Chapter Summary . . . 148

5. Differential Co-expression Analysis **150**
5.1 Introduction . . . 150
5.1.1 Differential Co-expression Analysis (DCA) . . . 150
5.2 Related Works . . . 152
5.2.1 Unupervised DCA approaches . . . 152
5.2.2 Supervised DCA Approaches . . . 154
5.3 Basics of Centrality Measures . . . 156
5.4 CBDCEM: An effective Centrality Based Differential Co-Expression Method for Crucial Gene Finding . . . 161
5.4.1 Pre-processing . . . 161
5.4.2 CEN Construction . . . 162
5.4.3 Module Extraction . . . 162
5.4.4 Hub-gene Finding . . . 163
5.4.5 Validation . . . 167
5.5 Experimental Results . . . 168
5.5.1 Pre-processing . . . 168
5.5.2 CEN Construction . . . 169
5.5.3 Module Extraction . . . 169
5.5.4 Hub-gene Finding . . . 175
5.6 Validation . . . 178
5.6.1 Enrichment Analysis of Modules . . . 178
5.6.2 Biological Analysis . . . 178
5.6.3 Literature Trace . . . 186
5.7 Discussion . . . 191
5.7.1 Comparison of with four other hub-gene finding algorithm . . . 194
5.7.2 CBDCEM vs. WGS . . . 194
5.7.3 CBDCEM vs. PCO . . . 195
5.7.4 CBDCEM Vs IMC . . . 195
5.7.5 CBDCEM Vs Degree . . . 197
5.8 Chapter Summary . . . 198
6. Differential Co-expression Analysis on Single Cell RNA-Seq Data **201**
6.1 Introduction . . . 201
6.1.1 Single Cell RNA Sequencing (scRNA-Seq) . . . 202
6.2 Related Works . . . 203
6.3 Background . . . 205
6.3.1 Measures for hub gene finding . . . 205
6.3.2 Seurat . . . 208
6.3.3 High Dimensional WGCNA (hdWGCNA) . . . 208

6.4 scDiffCoAM: A Complete Framework to Identify Potential ESCC Biomarkers using ScRNA-Seq Data Analysis . . . 210
6.4.1 Pre-processing . . . 211
6.4.2 Dimensionality Reduction . . . 211
6.4.3 Partitioning into cell types . . . 212
6.4.4 CEN Construction and Module Extraction . . . 212
6.4.5 Hub-gene Finding . . . 213
6.4.6 Identification of DEGs . . . 214
6.4.7 Validation . . . 214
6.5 Experimental Results . . . 215
6.5.1 Pre-processing . . . 215
6.5.2 Dimensionality Reduction . . . 217
6.5.3 Partitioning into cell types . . . 218
6.5.4 CEN Construction and Module Extraction . . . 218
6.5.5 Preservation Analysis . . . 221
6.5.6 Hub Gene Finding . . . 226
6.5.7 Identification of DEGs . . . 228
6.6 Validation . . . 229
6.6.1 Enrichment Analysis of Modules . . . 229
6.6.2 Biological Analysis . . . 230
6.6.3 Literature Trace . . . 240
6.7 Discussion . . . 252
6.7.1 Comparison with four other hub-gene finding algorithm . . . 257
6.7.2 Biomarker Ranking . . . 260
6.8 Chapter Summary . . . 262
7. Conclusion **263**
7.1 Concluding Remarks . . . 263
7.2 Future Works . . . 266